Université de Montréal

Le rôle des rétroactions écologiques et évolutives dans la structure des microbiomes

Par Naïma Madi

Département de sciences biologiques, Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de Philosophiae Doctor (Ph.D.)
en sciences biologiques

Avril 2023

Université de Montréal

Département de sciences biologiques, Faculté des Arts et sciences

---

Cette thèse intitulée

Le rôle des rétroactions écologiques et évolutives dans la structure des microbiomes

Présentée par

Naïma Madi

A été évaluée par un jury composé des personnes suivantes

Dr. Matthew Regan

Président-rapporteur

Dr. Jesse Shapiro

Directeur de recherche

Dr. Corinne Maurice

Membre du jury

Dr. Frédérique Le Roux

Examinateur externe

Résumé

Les communautés bactériennes sont constituées d'un grand éventail d'espèces pouvant interagir entre elles dans des environnements spatialement hétérogènes tels que le sol, les plantes ou l'intestin humain. À quel point ces interactions stimulent ou entravent la diversité du microbiome demeure inconnu. Historiquement, deux hypothèses ont été proposées pour expliquer comment les interactions interespèces pourraient influencer la diversité. L'hypothèse 'l'écologie contrôle' (EC) prédit une relation négative, dans laquelle l'évolution ou la migration de nouvelles espèces est freinée à mesure que les niches se saturent. En revanche, l'hypothèse 'la diversité engendre la diversité' (DBD) prédit une relation positive, où la diversité existante favorise l'accumulation d'une plus grande diversité à travers des interactions telles que la construction de niche.

De nombreuses études ont investigué ces modèles chez les vertébrés ou les plantes, et certaines les ont testés sur des bactéries en culture ; mais le modèle qui régit les communautés bactériennes naturelles demeure inconnu. En utilisant les données du gène ARN ribosomique 16S provenant d'un large éventail de microbiomes, j'ai montré une relation positive générale entre la diversité des taxons et la diversité des communautés de niveaux taxonomiques plus élevés. Cette observation est conforme à l'hypothèse du DBD, mais cette tendance positive plafonne à des niveaux élevés de diversité en raison des limites physiques de la niche.

Ensuite, j'ai observé que le modèle DBD restait valide à une résolution plus fine, en analysant la variation génétique intra espèce dans les métagénomes des microbiomes intestinaux humains. Conformément au DBD, j'ai observé que le polymorphisme génétique ainsi que le nombre de souches intra espèces étaient positivement corrélés avec la diversité Shannon de la communauté.

Dans le chapitre 3, j'ai examiné les interactions antagonistes entre *V. cholerae* et ses phages virulents et la manière dont ces interactions affectaient le cours de l'infection et la diversité génétique de *V. cholerae* chez les patients infectés.

J'ai quantifié les abondances relatives de *V. cholerae* et des phages virulents associés dans plus de 300 métagénomes provenant de selles de patients atteints de choléra, tout en tenant compte de leur exposition aux antibiotiques. Les phages et les antibiotiques ont supprimé *V.*

cholerae et ont été associés à une déshydratation légère chez les patients. J'ai également investigué les mécanismes de défense contre les phages dans *V. cholerae* et découvert que les éléments connus de résistance aux phages (integrative conjugative elements, ICEs) étaient associés à de faibles rapports phage: *V. cholerae*. J'ai pu montrer aussi que lorsque les ICEs ne sont pas détectés, la résistance aux phages semble être acquise par l'accumulation de mutations ponctuelles non synonymes.

Mes résultats valident que les phages virulents sont un facteur qui protège contre le choléra tout en sélectionnant la résistance dans le génome de *V. cholerae*.

Mots-clés : Diversité, interactions entre espèces, interactions phages-bactéries, communautés bactériennes, microbiome intestinal, microbiome de la terre, 16S, métagénomique, *Vibrio cholerae.*

Abstract

Bacterial communities harbor a broad range of species interacting within spatially heterogeneous environments such as soil, plants or the human gut. The extent to which these interactions drive or impede microbiome diversity is not well understood. Historically, two hypotheses have been suggested to explain how species interactions could influence diversity. The 'Ecological Controls' (EC) hypothesis predicts a negative relationship, where the evolution or migration of novel species is constrained as niches become filled. In contrast, the 'Diversity Begets Diversity' (DBD) hypothesis predicts a positive relationship, with existing diversity promoting the accumulation of further diversity via niche construction and other interactions.

Many studies investigated these models in vertebrates or plants, some focused on cultured bacteria, but we still lack insights into how natural communities are assembled in the context of these two hypotheses. Using 16S RNA gene amplicon data across a broad range of microbiomes, I showed a general positive relationship between taxa diversity and community diversity at higher taxonomic levels, consistent with DBD. Due to niche' limits, this positive trend plateaus at high levels of community diversity.

Then, I found that DBD holds at a finer resolution by analyzing intra-species strain and nucleotide variation in sampled metagenomes from human gut microbiomes. Consistent with DBD, I observed that both intra-species polymorphism and strain number were positively correlated with community Shannon diversity.

In Chapter 3, I investigated the antagonistic interactions between *V. cholerae* and its virulent phages and how these interactions affect the course of the infection and the within V*. cholerae* genetic diversity in natural infections.

I quantified relative abundances of *Vibrio cholerae* (*Vc*) and associated phages in 300 metagenomes from cholera patients stool, while accounting for antibiotic exposure. Both phages and antibiotics suppressed *V. cholerae* and were inversely associated with severe dehydration. I also looked at *V. cholerae* phage-defense mechanisms and found that known phage-resistance elements (integrative conjugative elements, ICEs) were associated with lower phage:*V. cholerae* ratios. In the absence of detectable ICEs, phages selected for nonsynonymous point mutations in the *V. cholerae* genome.

My findings validate that phages may protect against severe cholera while also selecting for resistance in the *V. cholerae* genome within infected patients.

Keywords: Diversity, biotic interactions, phage-bacteria interactions, bacterial communities, gut microbiome, earth microbiome, 16S, metagenomics, *Vibrio cholerae*.

Table of contents

List of Figures

Introduction

Thesis structure

Chapter I

Chapter 2

Chapter 3

List of Tables

Introduction

Chapter 1

Chapter 3

List of acronyms and abbreviations

AIC: Akaike Information Criterium

ARG: Antimicrobial Resistance Gene

ASV: Amplicon Sequence Variant

AZI: azithromycin

BH: Benjamini-Hochberg

BQH: Black Queen Hypothesis

BREX: BacteRiophage EXclusion

CIP: ciprofloxacin

CNV: gene Copy Number Variant

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

DBD: Diversity Begets Diversity

DEC : detected effective

DNA: DeoxyriboNucleic Acid

DNEC : detected not effective

DOBP : detected over the breakpoint

DOX: doxycycline

EC: Ecology Controls

EMP: Earth Microbiome Project

EMPO: Earth Microbiome Project Ontology

GAM: Generalized Additive Model

GLMM: Generalized Linear Mixed Model

HMP: Human Microbiome Project

icddr,b: International Centre for Diarrheal Disease Research, Bangladesh

ICE: Integrative Conjugative Element

ICP1, ICP2, ICP3: the International Centre for diarrhoeal disease research, Bangladesh cholera Phage 1, 2, 3.

KEGG: Kyoto Encyclopedia of Genes and Genomes

LC-MS: Liquid Chromatography-Mass Spectrometry

LRT: Likelihood-Ratio Test

MAG: Metagenome-Assembled Genome

MFA: Multiple Factor Analysis

MIC: minimum inhibitory concentration

ND: not detected

PCA: Principle Component Analysis

RDA: Redundancy Analysis

rRNA: ribosomal RiboNucleic Acid

SNV: Single Nucleotide Variant

SXT: SulfamethoXazole and Trimethoprim

Vc: *Vibrio cholerae*

To Elyas and Tala

Acknowledgements

I would like to express my sincere gratitude to everyone who supported me during all these years, particularly to my supervisor and to Pierre Legendre.

This work was supported by a CIHR project grant to BJS.

Introduction

Prokaryotes, the first cellular life forms, were active on Earth for more than 3.0 billion years before the evolution of multicellular life forms (Schopf et al. 2018). They have adapted to nearly every habitat on Earth and evolved novel metabolic strategies – such as oxygenic photosynthesis, which releases oxygen as a by-product of energy generation (Dunlap 2001). This process, first carried out by Cyanobacteria, allowed more complex aerobic organisms to evolve and provided a protective shield of ozone against ultraviolet radiation for terrestrial and aquatic organisms (Dunlap 2001). All of Earth's global biogeochemical cycles of major elements (i.e., carbon, nitrogen, sulfur and iron), rely on microbes (Falkowski, Fenchel, and Delong 2008). Beside the degradation of complex organic compounds, microbes have a huge range of metabolic diversity: sulfate reduction, methanogenesis, iron oxidation, denitrification, nitrite oxidation and nitrate reduction, and hydrogen and methane oxidation, just to name a few. In carrying out these processes, microbes serve humans and other higher organisms as environmental recyclers and bioremediators, at the foundation of food webs (Dunlap 2001).

Beside their involvement in ecological processes, bacteria can influence the health and wellbeing of their hosts. The explosion of microbial genome sequence data and increasingly detailed analyses of the gut microbiome has yielded insights into how several diseases are now thought to be influenced by the gut bacterial communities, including cancer (Dolgin. 2020), autoimmune disorders such as multiple sclerosis (De Luca and Shoenfeld 2019), chronic pain syndrome (Guo et al. 2019) and autism spectrum disorder (Svoboda 2020).

Microbes represent the majority of Earth's genetic diversity (Hug et al. 2016). Global estimates of the Earth's biodiversity include 5 million to 7.7 million unique species of animals, 500,000 plants, 6 million to 8 million terrestrial fungi and up to 1 trillion species of prokaryotes (Averill et al. 2022). In addition to being abundant, bacteria are ubiquitous, they colonize an extraordinary array of habitats and ecosystems, ranging from the gut to extreme environments (Sayed et al. 2020; Madigan 2000), such as hyperarid desert (Kurapova. et al. 2010), permafrost soil (Mackelprang et al. 2017) and deep-sea sediments (Ulanova and Goo 2015), as well as acidic (Chen et al. 2016) and high-temperature environments (Valverde, Tuffin, and Cowan 2012).

High-throughput 16S rRNA gene amplicon sequencing studies continue to yield unprecedented insights into the taxonomic richness of microbiomes (Thompson et al. 2017). Whereas, advances in sequencing technologies such as shotgun metagenomics (Shaffer et al. 2022) and nanopore based-DNA-sequencing (Leidenfrost et al. 2020; Chapman et al. 2023), combined with developments in computational approaches provided additional insights into fine-resolution bacterial taxonomic and functional diversity, including within-population diversity.

The role of standardized data repositories in bacterial studies is important. An increasingly high number of studies have advanced the understanding of microbial community dynamics, functional diversity and evolution using public shared datasets such as the Earth Microbiome Project (EMP) (Thompson et al. 2017) and the Human Microbiome Project (HMP) (Human Microbiome Project 2012).

1. How do bacteria diversify?

The relatively high population sizes, rapid generation times and capacity for horizontal gene transfer (HGT) make bacteria able to undergo rapid evolution over timescales ranging from days to months (Travisano. et al. 1995). To track this evolution in real time and in natural environments, metagenomic sequencing and technical advances are enabling culture-free, high-resolution strain and subspecies analyses at high throughput.

Bacterial genetic diversity is generated through de novo mutations (substitutions, deletions and inversions), and horizontal gene transfer (HGT) by self-transmissible mobile genetic elements like plasmids, integrative conjugative elements (ICEs) and bacteriophages. Mutations arise continuously in the genome due to errors in the DNA replication process, damages caused by mutagens, or errors in the DNA repair and recombination mechanisms (Van Rossum et al. 2020). Gene flow by HGT can cause rapid and large-scale additions and rearrangements of genomic regions. ICEs transfer via conjugation and integrate into and replicate along with the host chromosome (Wozniak et al. 2009). These elements allow bacteria to adapt to new environmental conditions and mediate the transfer of virulence determinants. Like ICEs, prophages can integrate into the host chromosome. Toxin encoding prophages from numerous

pathogenic bacteria have been demonstrated to transfer into non-pathogenic bacteria, converting them to virulent strains (Shah M. Faruque 1999). Many species have been found to have both pathogenic and commensal and/or environmental strains. A classic example is *Escherichia coli* strains, which can be pathogenic, commensal, host associated or environmental (Van Rossum et al. 2020); and toxigenic *Vibrio cholerae* strain that originated from a nontoxigenic environmental ancestor that acquired the filamentous bacteriophage CTXphi (Shah M. Faruque 1999). Furthermore, the co-evolutionary arms race between virulent phages and bacteria where bacteria develop resistance to phage and phage counter-adapt, is a major driving force of bacterial genetic diversification (Tamar and Kishony 2022; Brockhurst, Buckling, and Rainey 2005).

2. Phage-bacteria co-evolution

Bacteria can adapt rapidly to new environmental conditions and evolve particularly rapidly in response to predators such as phages. This selective pressure has forced bacteria to adapt with multiple defense strategies to evade virulent phage predation and prophage acquisition (Labrie, Samson, and Moineau 2010)**.** Most bacterial defense mechanisms are at the individual level but others involve multicellular (collective) defense mechanisms such as biofilm formation or abortive infection. Phage adsorption to cell receptors is the initial step of infection. To prevent this key process bacteria can alter or occlude phage-binding sites or shed outer membrane vesicles (OMVs) (Reyes-Robles et al. 2018), bubble-like extracellular vesicles separated from the membrane, as decoys to spare intact cells from infection. Once inside the cell, phages encounter a diverse set of defenses that reduce phage replication or degrade phage DNA. Ongoing research, that continues to discover new mechanisms, has revealed that these defenses tend to be encoded in genomic regions named ‘defense islands’ (Vassallo et al. 2022). The list below summarizes the most common defense mechanisms but is not exhaustive; new defense systems, whose mechanisms still need to be elucidated, are continually discovered (Doron et al. 2018):

1. Restriction-Modification (RM) systems: these encode restriction endonucleases, which bind to and cleave phage DNA at specific recognition sites. Recognition sequences on bacteria DNA are modified via methylation to protect them from degradation, whereas unmodified phage DNA is destroyed by the endonuclease. Recently discovered antiviral defenses, such as DISARM (defense island system associated with restriction-modification) (Ofir et al. 2018) and Dnd (DNA phosphorothioation) (Wang et al. 2019) systems, work similarly, respectively attacking foreign DNA that lacks methyl or sulfur modification.

2. CRISPR-Cas (clustered regularly interspaced short palindromic repeats; CRISPR-associated) (Abedon 2012; Broniewski et al. 2020): these provide bacteria with adaptive immunity against phages whose genomic signatures have previously been encountered. These systems store fragments of foreign DNA in the bacterial genome, which then guide Cas restriction enzymes to degrade DNA in the cell that resembles that of past phage infection. The prokaryotic Argonaute (pAgo) proteins operate on a similar principle, providing guided DNA interference against phages, plasmids and transposons (Smith et al. 2023).

3. Prophages and mobile genetic elements (MGEs): Prophages and MGEs, sequences that encode mobile genes of phage origin that enable them to adapt and disseminate, are common reservoirs and distributors of anti-phage defense systems (Vassallo et al. 2022) (Rousset 2022). The most important classes of MGEs are:
    a) Phage satellites: They are phages parasites. The term satellite is due to their intimate relationship with certain phages whose life cycle they parasitize for mobilization and were recently demonstrated to encode hotspots of anti-viral systems (Ibarra-Chavez et al. 2021; Rousset et al. 2022). Phage-inducible chromosomal islands (PICIs) (Penades and Christie 2015) are a class of phage satellites that provide a fitness advantage to their host bacterium by limiting phage proliferation upon infection and by carrying virulence genes (Mckitterick et al. 2018). Epidemic *Vibrio cholerae* encodes PICI-like elements (PLEs) to inhibit

ICP1 replication (McKitterick and Seed 2018). Based on current data, PLEs are phage parasites restricted to epidemic *Vibrio cholerae* and appear to exclusively parasitize the virulent phage ICP1 (LeGault et al. 2022).

b) ICEs (Integrative and Conjugative Elements) (Johnson and Grossman 2015): they are MGEs that carry genes that confer advantageous phenotypes to the bacterial host, like pathogenesis and resistance to antibiotics and phages (Johnson, Harden, and Grossman 2022). A recent study of a large dataset of phage-bacteria interactions in endemic *Vibrio cholerae* revealed how a specific ICE, the SXT ICE, carries anti-phage systems and antibiotic resistance genes (Wozniak et al. 2009) (LeGault et al. 2021). SXTs are ~100kb islands that were first discovered in *V. cholerae* (Waldor, Tschäpe, and Mekalanos 1996). SXTs have conserved 'core' genes along with variable 'hotspots' encoding different genes, hotspot 5 is the one associated with phage resistance. ICEVchInd5 and ICEVchind6 were the two most prevalent ICEs in Bangladesh. These ICEs differ in their anti-phage systems: ICEVchInd5 encodes a type 1 bacteriophage exclusion (BREX) system while ICEVchInd6 encodes several other restriction-modification systems (LeGault et al. 2021). Many other variants of SXT ICEs have been isolated from *Vibrio cholerae* (LeGault et al. 2021).

In response to bacterial defense mechanisms, phages have evolved anti-defense strategies, like anti-CRISPR genes that hinder the binding or cleavage of the phage genome by the CRISPR-Cas system (Borges, Davidson, and Bondy-Denomy 2017; Boyd. et al. 2021). In the same way, some ICP1 phages have acquired a CRISPR-Cas system to target PLEs, which allows ICP1 to persist in spite of them (LeGault et al. 2022). Vibriophages have also developed two mechanisms to overcome SXT ICEs: epigenetic escape (genetic modification during the course of infection) from RM systems and an anti-BREX inhibitor protein (OrbA) (LeGault et al. 2021).

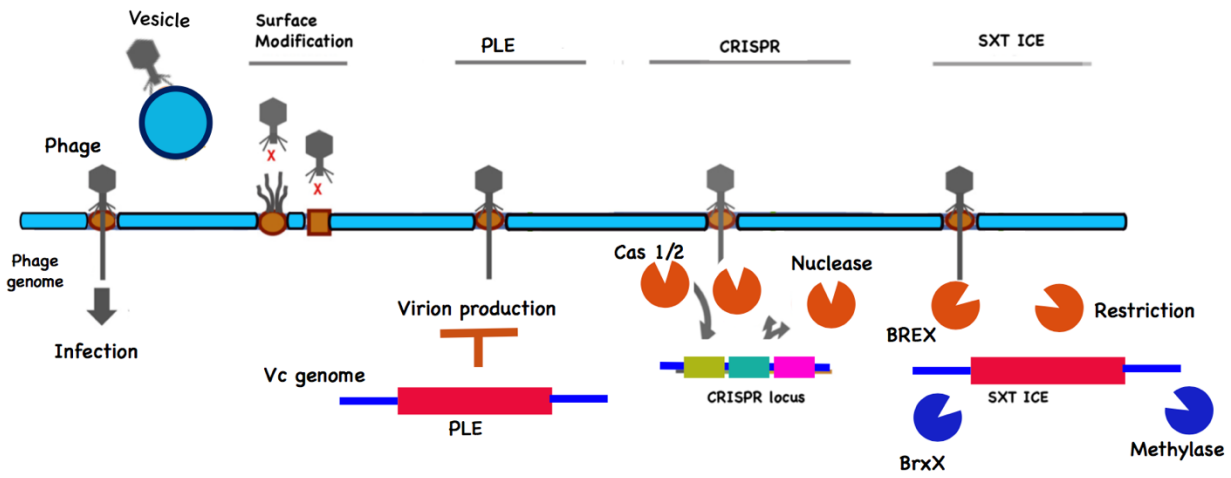

Figure 1. Vibrio cholerae defense systems against virulent phages. V. cholera protects itself from virulent phages *by phage receptor modification or occlusion, the release of outer membrane vesicle decoys; phage DNA degradation by CRISPR-Cas systems (the Cas enzymes cleave the phage DNA) or restriction-modification (RM) systems and bacteriophage exclusion (Brex) systems which are on the SXT ICEs, restriction and BREx cleave the phage DNA while methylation and BrxX protect the bacteria DNA from the cleavage. Upon infection, V. cholerae uses phage-inducible chromosomal island-like element (PLE) mediated restriction of virion production (restricted to the virulent phage ICP1). Adapted from (van Houte, Buckling, and Westra 2016) and (Boyd. et al. 2021).*

3. Factors affecting bacterial diversity

Microbial communities are shaped by both abiotic and biotic factors. Several studies have reported how abiotic factors can shape the composition of microbial communities, at both between-species and within-species levels through community assembly and evolution: pH, temperature, latitude, elevation (Delgado-Baquerizo et al. 2018), oxygen concentration (Baez and Shiloach 2014), nutrient availability (Shehata and Marr 1971) and the presence of stress-inducting xenobiotics such as drugs or heavy metals (Sobolev and Begonia 2008).

Microbes can be involved in positive interactions like cross-feeding of metabolites among different microbes (Seth and Taga 2014; Culp and Goodman 2023), and commensalism, when the association is beneficial for one organism and neither beneficial nor harmful for the other one; as well as negative interactions such as antagonism, when one microbial population produces substances that are inhibitory to other microbial populations, competition for resources or space (Hibbing et al. 2010) or predation when an organism (like *Bdellovibrio*) engulfs or attacks another

organism. Although it is known that these interactions can be important in determining community composition, little is known about how such interactions shape genetic diversity within species.

Understanding the role of interspecific interactions in shaping within-species diversity enhances our knowledge of microbial community dynamics and the interplay between ecological and evolutionary processes. Furthermore, finer-scale strain-level variation may also have important functional and ecological consequences; among other things, strains are known to engage in interactions that cannot be predicted from their species identity alone (Goyal et al. 2022.). Although closely-related bacteria are expected to have broadly similar niche preferences, finer-scale niches may differ below the species level (Martiny et al. 2015). For example, the acquisition of a carbohydrate-active enzyme by *Bacteroides plebeius* allows it to exploit a new dietary niche in the guts of people consuming nori (seaweed) (Hehemann et al. 2010), and single nucleotide adaptations permit *Enterococcus gallinarum* translocation across the intestinal barrier resulting in inflammation (Yang et al. 2022).

Historically, two hypotheses have been described to address the relationship between biotic interactions and biodiversity: The 'Ecological Controls' (EC) hypothesis predicts a negative relationship, where the evolution or migration of novel species is constrained as niches become filled (Rabosky and Hurlbert 2015; Schluter and Pennell 2017). The 'Diversity Begets Diversity' (DBD) hypothesis predicts a positive relationship, with existing diversity promoting the accumulation of further diversity via niche construction and other interactions (Whittaker 1972; Calcagno et al. 2017) (**Figure 2**). An alternative to EC or DBD id the Neutral Theory of Biodiversity and Biogeography (Hubbell, 2001), in which all species are functionally equivalent and communities assemble randomly, via speciation, ecological drift and dispersal limitation, with no effect of either biotic or abiotic factors. The neutral model provides quantitative null models for assessing the role of adaptation and natural selection (**Figure 2**).

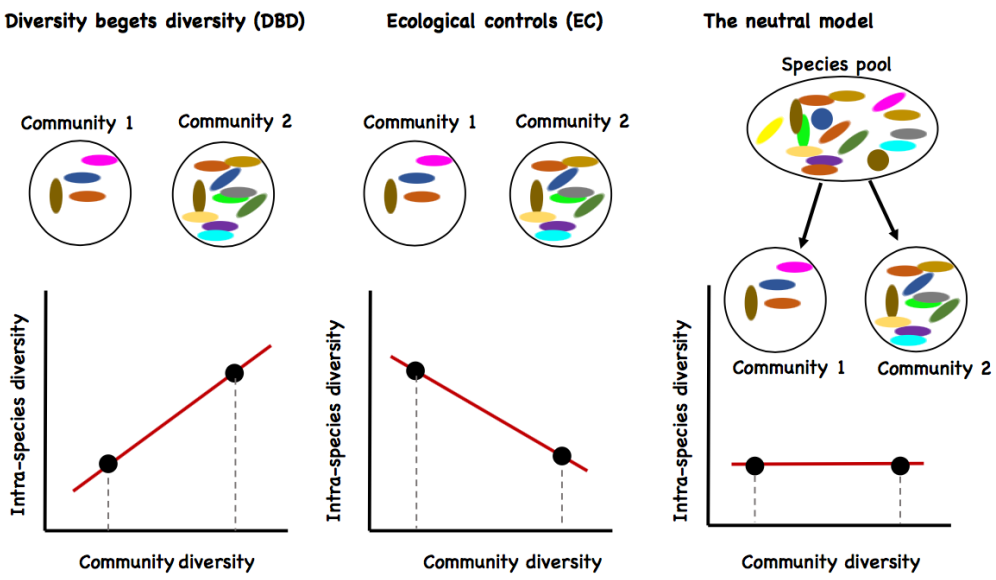

Figure 2. Illustration of community assembly models: DBD, EC and a neutral mode.

Many studies have reported conflicting observations on natural plant and animal communities (Emerson and Kolm 2005; Palmer and Maurer 1997; Price et al. 2014; Rabosky et al. 2018; Calcagno et al. 2017) (**Table 1**). Laboratory evolution experiments tracking the diversification of a focal bacterial lineage in communities of varying complexity have also yielded contradictory results, with support for EC (Gomez and Buckling 2013), DBD, or intermediate scenarios (Brockhurst et al. 2007) (**Table 1**). In **Chapter 1** and **Chapter 2** of this thesis, I investigated whether these laboratory-based findings can be generalized to natural bacterial communities that are far more complex and span more ecological and evolutionary diversity.

***Table1. Community diversity* may promote or impede focal species diversity depending on the ecological context.**

Reference	Animal model	Diversity model	Underlying mechanism
Species diversity can drive speciation, B. C. Emerson and N. Kolm. 2005.	Plants and arthropods on Islands	DBD	Not addressed
Does diversity beget diversity? A case study of crops and weeds, Palmer, Michael W. & Maurer, Teresa A. 1997.	Weed communities in monoculture versus multiculture crop	DBD	Not resolved
Niche filing slows the diversification of Himalayan songbirds. T. D. Price et al. 2014.	Birds	EC	Niche filling
An inverse latitudinal gradient in speciation rate for marine fishes, Rabosky et al. 2018.	Marine fishes	EC	Not addressed

Diversity spurs diversification in ecological communities, Calcagno et al. 2017.	Mathematical models	DBD	Not addressed
The effects of competition and predation on diversification in a model adaptive radiation, J. R. Meyer and R. Kassen. 2007.	*Pseudomonas fluorescens* cultured in a King's B microcosm with and without the predator	DBD	Predation by a protist (new genotypes that exploit predator-free space)
Niche occupation limits adaptive radiation in experimental microcosms. Brockhurst et al. 2007	*Pseudomonas fluorescens cultured in* communitites of variable of variable diversity	EC (depending on the strain)	Niches filling (competition for niche space)
High functional diversity stimulates diversification in experimental microbial communities, Jousset et al. 2016	*Pseudomona fluorescens* cultured in bacterial communitites of variable diversity	DBD	Ressource competition : the evolved phenotype showed a better use of underexploited ressources
Real-time microbial adaptive diversification in soil, P. Gómez and A. Buckling. 2013.	*Pseudomonas fluorescens* in soil microcosm with/without the microbial community over 48 days	EC	Niche filling
Diversity spurs diversification in ecological communities, Calcagno et al. 2017.	Mathematical models	DBD	Not addressed

4. Evolution and ecology within the human gut microbiome

The animal gut is among the most densely populated systems on Earth, with the microbial cells residing in the gut of a healthy human can outnumber the human cells by more than an order of magnitude (Sommer and Backhed 2013). This community is mainly composed of bacteria but also include archaea, fungi, viruses and protozoa. These organisms exist in a complex consortium of ecological and metabolic interactions that ultimately influence the taxonomic and functional profile of the community, as well as host health (Loftus, Hassouneh, and Yooseph 2021). Culture-based approaches, animal models, and advanced sequencing methodologies have unveiled the complexity of these interactions. A recent experiment based on a synthetic minimal microbiome (Shetty, Smidt, and de Vos 2019) consisting of ten core intestinal species showed that functional interactions of species in the gut span from competition and cross-feeding, where compounds released into the extracellular environment are harvested by non-degraders, to interspecies metabolic interactions leading to the production of key short-chain fatty acids (Shetty et al. 2022). It has also been reported that positive interactions (i.e., cooperation) dominate over negative associations (i.e., competition) at the species level in the human gut (Loftus, Hassouneh, and Yooseph 2021).

Cross-feeding relationship between bacteria is consistent with the black queen hypothesis (BQH) which predicts that microbes will lose functions that are costly but can be obtained from other genotypes because of their leaky nature (Morris and Lenski 2012) (**Figure 3**).

These interactions are known to be highly dynamic and governed by both biotic and abiotic factors (Shetty et al. 2022). For example, studies have revealed large variation among healthy individuals and showed associations between gut microbial composition and several factors like diet (David et al. 2014), alcohol consumption (Fan et al. 2018), lifestyle (Allen et al. 2018), age (Yatsunenko et al. 2012), gender (Kim et al. 2020) and host genetics (Bonder et al. 2016). Despite these variations, diverse gut microbial communities generally play important roles in host health. For example, the microbiome facilitates the metabolism of otherwise indigestible polysaccharides and produces essential vitamins; the microbiome is required for the development and differentiation of the host's intestinal epithelium and immune system; it confers protection against invasion by opportunistic pathogens and takes a key role in maintaining tissue homeostasis (Sommer and Backhed 2013).

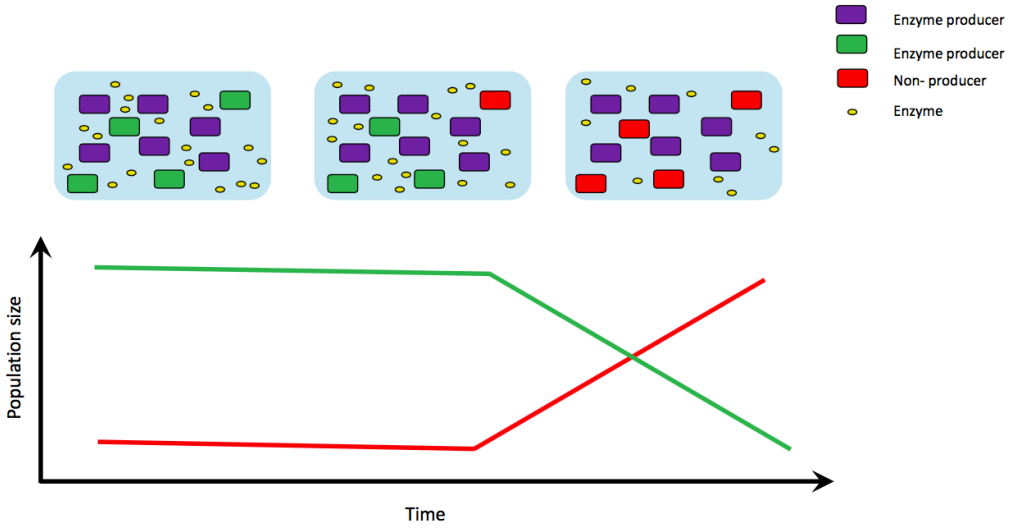

Figure 3. Example of evolution via the Black Queen Hypothesis, where the function lost is the production of an extracellular enzyme for the degradation of a complex molecule. *Initially, all bacteria produce the enzyme. Mutation or gene loss produce an individual unable to synthesize this enzyme (giving it a growth advantage via energy saving). The mutant proportion increases until all population is a non-producer and become dependent on the producer. The gene function needs to be retained in at least one member of the community, adapted from (Smith. et al. 2019).*

5. Factors that disturb the gut microbiome

Despite the huge variation among healthy individuals, gut microbial communities are able to maintain homeostatic equilibrium and are resistant against perturbations. However several factors may disturb the gut microbial equilibrium, such as health disorders like inflammatory bowel disease (IBD) (Frank et al. 2007), severe acute malnutrition (SAM) (Ghosh et al. 2014), cardiovascular diseases (Wang et al. 2011), periodontitis (Hajishengallis 2015), colon cancer (Arthur et al. 2012), depression (Koopman, El Aidy, and consortium 2017) and Parkinson's disease (Anderson et al. 2016).

Furthermore, the disequilibrium of the gut microbiome may be due to an increase in the abundance of pathogens. Pathogens may drive physiological alterations in the gut that favor pathogen growth and long-term colonization. For example, diverse Gram-negative bacteria express type III secretion systems (T3SS), which enable bacteria such as *Salmonella*, *Shigella*, and *Chlamydia* species to invade host cells and induce inflammation to promote their long-term colonization. In the same way, *V. cholerae* possesses, in addition to the two main virulence genes (the cholera toxin, CT and the toxin-coregulated pilus, TCP), several secretion systems like the type VI secretion system (T6SS) which can target eukaryotic host cells causing intestinal inflammation. The T6SS can also target prokaryotic organisms providing means of interspecies competition to enhance environmental survival (MacIntyre et al. 2010).

6. Cholera: history, interactions between cholera-phages.

V. cholerae causes the severe dehydrating diarrheal disease cholera and remains a major public health concern in many countries, principally in Africa and Asia, due to inadequate sanitation and safe drinking water resources. The Bay of Bengal is known as the epicenter from which cholera outbreaks are seeded across the globe (Verma et al. 2019).

Descriptions of a disease thought to be cholera are found in Sanskrit back to the 5th century BC, and the disease has existed on the Indian subcontinent for centuries. In 1817, cholera spread beyond the Indian subcontinent, and six worldwide cholera pandemics occurred between 1817 and 1923. Between 1849 and 1854, London physician John Snow proposed that cholera was a

transmissible disease and that stool contained infectious material. He suggested that this infectious material could contaminate drinking water supplies, resulting in transmission of cholera. Filippo Pacini, working independently in Italy in 1854, first observed comma-shaped forms under a microscope in cholera stools. In 1884, Robert Koch first isolated *V. cholerae* in pure culture in work that began in Egypt and continued in Calcutta, India (Harris et al. 2012).

Since 1817, seven cholera pandemics have spread from Asia to much of the world. The seventh pandemic began in Indonesia in 1961 and spread through Asia to Africa, Europe and Latin America (Harris et al. 2012). Between 1970 and 2011, several European countries reported cholera outbreaks of a few to more than 2000 cases (Oprea et al. 2020). After more than 60 years, the World Health Organization (WHO) estimates that 1.3 to 4.0 million cases and 21 000 to 143 000 deaths still occur worldwide annually (WHO 2022), with periodic major epidemics including those in Yemen in 2009 (more than 1 million death and 3000 death to date (Xiang Ng et al. 2020) and Haiti in 2010 (Piarroux et al. 2011).

V. cholerae is a member of the *Vibrionaceae*, a gram-positive family found in aquatic environments. It is classified into more than 200 serogroups based on the O antigen of the lipopolysaccharide (LPS) (**Figure 4**) (Lerouge and Vanderleyden 2002). Of these, only O1 and O139 serogroups cause epidemic cholera. *V. cholerae*. O1 is further classified into two biotypes, classical and El Tor. The O139 type derived from *V. cholerae* O1 El Tor by lateral transfer of a genomic island substituting the O139 for the O1 antigen, but is otherwise almost identical to *V. cholerae* O1 El Tor. Although classical *V. cholerae* O1 caused the fifth and sixth pandemics (and presumably the earlier pandemics), the seventh pandemic is attributed to the El Tor biotype, which has now replaced the classical biotype (Harris et al. 2012).

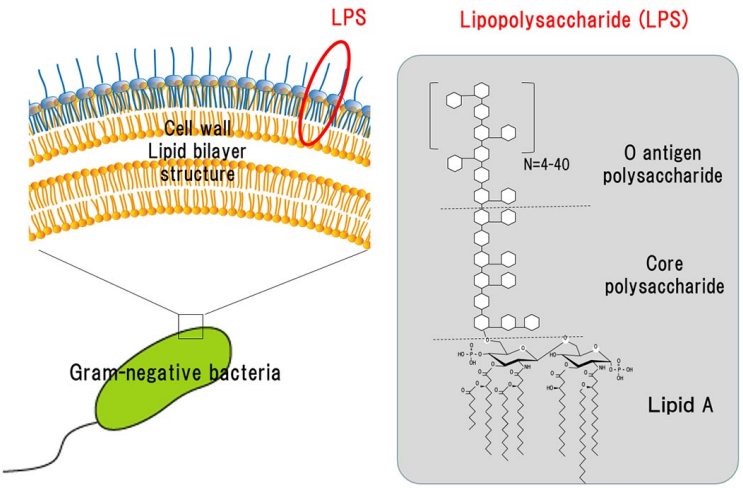

Figure 4. lipopolysaccharide (LPS) structure of gram-negative bacteria (https://www.macrophi.co.jp).

Numerous ecological factors contribute to the dynamics of cholera outbreaks; among these factors, virulent phages are thought to play an important role (Nelson et al. 2009). Three virulent phages have been isolated and sequenced from stool samples from cholera patients in Bangladesh: ICP1, a member of the *Myoviridae* family, was present in all stool samples and two other less prevalent phages ICP2 and T7-like ICP3, which are members of the *Podoviridae* family (**Figure 5**) (Seed et al. 2011).

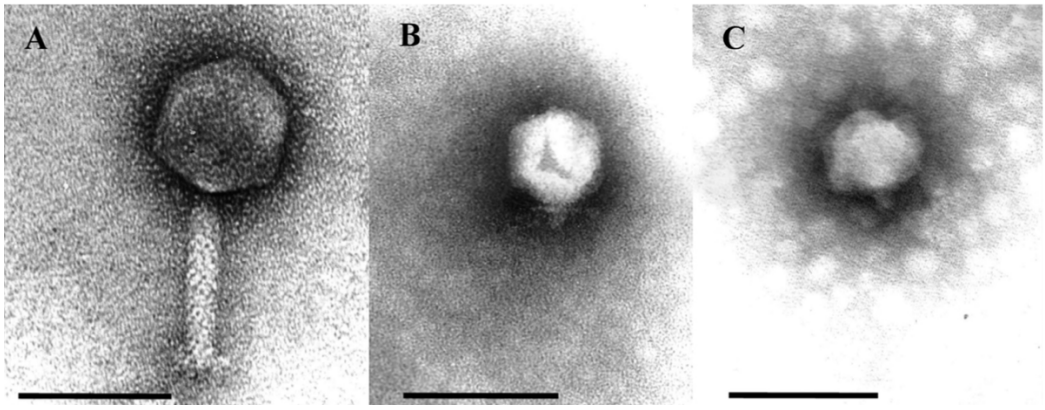

Figure 5. Electron micrograph images of V. cholerae virulent phages isolated from stool samples from cholera patients. *(A) ICP1, (B) ICP2, and (C) ICP3 (Seed et al. 2011).*

It has been demonstrated that the O1 antigen of the *V. cholerae* lipopolysaccharide (LPS) serves as an ICP1 receptor (Boyd. et al. 2021) and the outer membrane porin OmpU as an ICP2 receptor (Seed et al. 2014). The receptor for ICP3 remained uncharacterized until a recent work based on laboratory experiments and isolate sequencing suggested that the O1-antigen might also be an ICP3 receptor and a secondary receptor for ICP2 (**Figure 6**) (Beckman and Waters 2023).

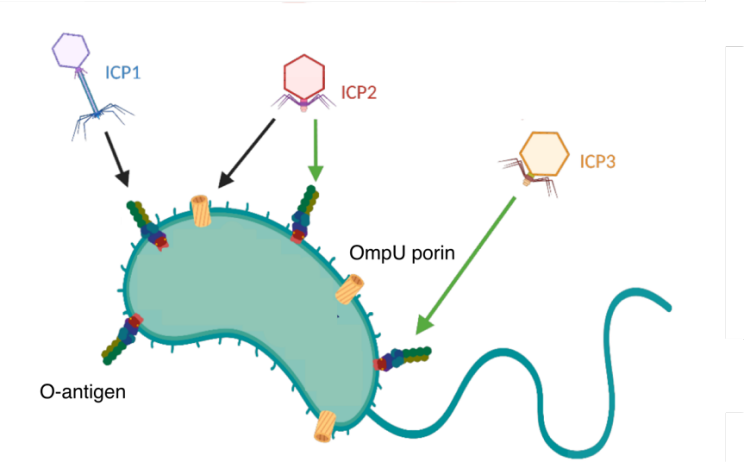

Figure 6. ICP1, ICP2 and ICP3 receptors (Beckman and Waters 2023).

Despite over a century of study, there is still much to learn about *V. cholerae*-phages interactions and co-evolution. The first study that associated cholera disease and phages was carried out in 1927 by D'Herelle. He observed a correlation between cholera patient disease outcome and the behavior of the bacteriophage ('strong' or 'weak'), and hypothesized that the course of disease is governed by the behavior of virulent phages that prey on *V. cholerae* (D'Herelle and Malone 1927). Shortly thereafter, another study in India, found that cholera cases were positively correlated with the isolation of phages from the aquatic environment (Pasricha, MDe Monte, and Gupta 1931). A century later, an inverse correlation was observed between virulent phages and *V. cholerae* isolated from aquatic environments in Dhaka, Bangladesh between 2001 and 2004 (Faruque et al. 2005a). These apparently conflicting findings, may be evidence of predator-prey dynamics which oscillate between positive and negative correlations as predicted by the Lotka-Volterra model (Carr et al. 2019a) (**Figure 7**). The role of virulent phages in the control of cholera outbreak was also observed in animal models (Zahid et al. 2008; Jaiswal

et al. 2013a; Yen, Cairns, and Camilli 2017) and using modelization (Jensen et al. 2006a). All these findings suggest that phages play an important role in *V. cholerae* epidemiology, but how do *V. cholerae* interact with its virulent phages during host infection remains an open question. I addressed this question in **Chapter3** of this thesis.

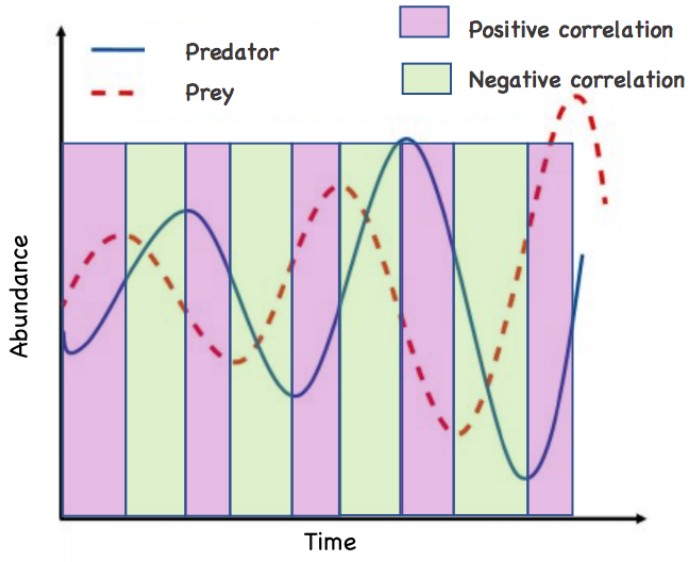

Figure 7. Lotka-Volterra predator-prey oscillatory dynamics. *This model predicts oscillations in both predator and prey abundance as a function of time. As the prey population grows, the predator population has more food and also increases in abundance. However, predation eventually out-paces the growth of the prey population and drives the prey toward near-extinction, until there are too few prey to sustain the predator population. Once the predator population crashes, the few remaining prey can recover, and the cycle begins anew. Over the course of time, predator and prey populations transition between windows of positive correlation and negative correlation (Carr et al. 2019a).*

Thesis structure

Although it is generally recognized that biotic interactions can be as important as abiotic factors in driving community composition and biodiversity, they have received less attention than the abiotic drivers of diversity, and this is true particularly for bacteria. With the exception of a few experimental studies that tracked the diversification of a focal species in communities of varying complexity (Brockhurst et al. 2007; Calcagno et al. 2017; Jousset et al. 2016a; Gómez and Buckling 2013), this question has not been further studied in natural environments. My PhD thesis contributed to investigate how biotic interactions in the communities affect species diversity. In the two first chapters, I considered species-species interactions in a range of environments. In the last chapter, I looked at the antagonistic interactions between a pathogenic bacterium and its virulent phages in the human gut and how these interactions affect the course of infection and the within-bacterium genetic diversity (**Figure 1**).

Two hypotheses have been proposed to describe the relationship between biotic interactions and biodiversity: The 'Ecological Controls' (EC) hypothesis predicts a negative relationship, where the evolution or migration of novel species is constrained as niches become filled (Rabosky and Hurlbert 2015; Schluter and Pennell 2017). The 'Diversity Begets Diversity' (DBD) hypothesis predicts a positive relationship, with existing diversity promoting the accumulation of further diversity via niche construction and other interactions (Whittaker 1972; Calcagno et al. 2017).

In **Chapter 1**, my objective was to test whether patterns of diversity in natural communities conform to Ecological Controls (EC) or Diversity Begets Diversity (DBD) dynamics. Using the Earth Microbiome Project (EMP) 16 S dataset with 16S rRNA gene amplicon sequence variants (ASVs) as the finest-grained taxonomic unit, I estimated diversity within a focal genus as an ASV:Genus ratio (as in Elton 1946), then I looked at the relationship between this ratio and the number of non-focal genera with which the focal genus could interact. This method was then repeated at higher taxonomic levels (up to the Class:Phylum level). My observations were consistent with the predictions of DBD: taxa diversity was positively correlated with community diversity at the higher taxonomic level. However, diversity hits a plateau as niches become increasingly filled due to niche limits. Then, I aimed to investigate whether DBD or EC dynamics

differ between taxa that are able to adapt to diverse habitats (generalists) and those that adapted to a specific habitat (specialists) (Van Tienderen 1991). My hypothesis was that the DBD-EC balance could differ between generalist lineages found in many environments and specialists with a more restricted distribution, as generalists are known to have a higher speciation rate than specialists (Sriswasdi, Yang, and Iwasaki 2017). I found that the effect of DBD was strongest among habitat specialists, suggesting that the establishment of niche adapted taxa is selected over that of generalist taxa. Then I hypothesized that bacteria with bigger genomes would have higher DBD than smaller ones because bacteria with larger repertoires of accessory genes are known to occupy a wider range of niches (Barberan et al. 2014). Taxa with larger genomes might therefore be hypothesized to better survive and thrive when they disperse into a new location, exhibiting stronger DBD. Testing this hypothesis confirmed that larger genomes exhibit a higher DBD slope. Overall, this study gives insights into the general trend between community diversity and taxa diversity at high taxonomic levels in a broad range of environments (Madi et al. 2020).

The results in **Chapter 1** most likely pertain to ecological community assembly rather than in-situ diversification because of the limited resolution of 16S rRNA gene sequences. In **Chapter 2**, I wanted to investigate the DBD/EC hypotheses at a finer resolution by looking at within-species genetic diversity and how it is related to the surrounding diversity. I analyzed intra-species strain and nucleotide variation in static and temporally sampled metagenomes from the human gut microbiome (HMP). Consistent with DBD, I found that both intra-species polymorphism and strain number were positively correlated with community Shannon diversity. I also investigated how community diversity correlated with gene variation (i.e., gain and loss) and found that higher community diversity was positively correlated with gene loss at a future time point; consistent with the Black Queen Hypothesis (BQH) (Morris and Lenski 2012). My work in this chapter shows that a mixture of DBD and Black Queen can operate simultaneously in the human gut microbiome (Madi et al. 2023a).

In **Chapter 3**, I used metagenomic data to investigate the antagonistic interactions between *V. cholerae* and its virulent phages. The objective of this chapter was to look at: 1) how *V. cholerae* and its virulent phages interact within infected patients; 2) how these interactions

affect the infection course (do phages control *V. cholerae* during natural infection?); and 3) how does *V.cholerae* adapt to phage selective pressure under natural conditions.

I used metagenomes from cholera patients stool and found evidence for non-linear predator-prey dynamics between *V. cholerae* and ICP1 that were suppressed by azithromycin, a commonly used antibiotic in cholera cases. These dynamics end up with high phages:*V. cholerae* ratios. This finding demonstrated that both phages and antibiotics may have an effect in reducing cholera load.

Under phage and antibiotic pressure, *V. cholerae* may evolve diverse mechanisms to defend itself. I looked at the antibiotic resistance genes in the gut of cholera patients and found that ciprofloxacin exposure was associated with many resistance genes but not azithromycin, that validates *V.cholerae* sensitivity to this drug. Then, I investigated phage-resistance within *V. cholerae* and detected the ICE phage-resistance mechanism in 75% of the data. I quantified genetic diversity within the infecting *V. cholerae* population and found that over one third of samples were hyper-mutators, and that these mutators were under phage pressure. I also observed that higher abundance of ICP1 was associated with more non-synonymous mutations in the *V. cholerae* genome when ICE was not detected. This study validates that phage and antibiotics may protect against severe cholera while also selecting for resistance in the *V. cholerae* genome within infected patients (Madi et al. 2023b).

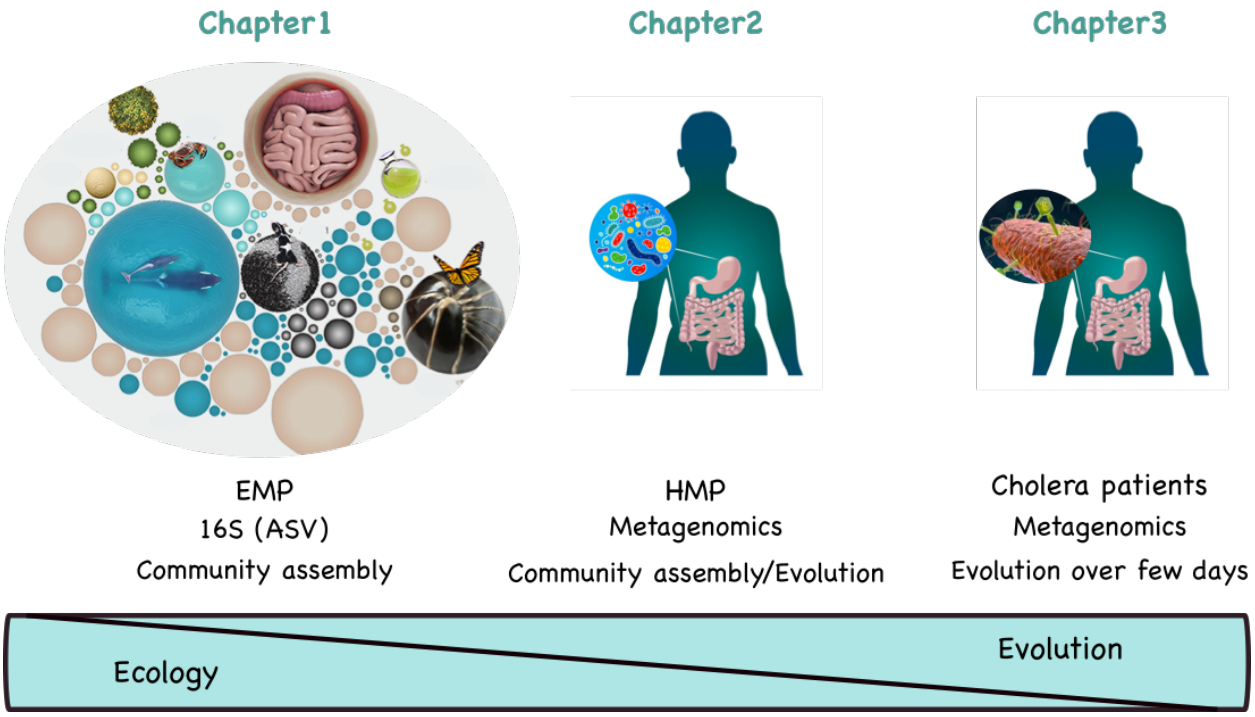

Figure 1. Illustration presenting the themes of the three chapters of the thesis. *In Chapter 1, I used amplicon sequence variants (ASVs) from the Earth Microbiome Project (EMP) to study the relationship between community diversity and focal taxa diversity at higher taxonomic levels corresponding to community assembly (ecology). In Chapter2, I further studied this relationship at higher genetic resolution using human microbiome metagenomes from the Human Microbiome Project (HMP). In Chapter3, I used metagenomes from cholera patient stool to study Vibrio cholera-phage interaction (evolution over few days).*

Chapter 1 – Does diversity begets diversity in microbiomes?

Authors: Naïma Madi[1], Michiel Vos[2], Carmen Lia Murall[1], Pierre Legendre[1] and B. Jesse Shapiro[1,3,4*]

1. Département de sciences biologiques, Université de Montréal, Canada
2. European Centre for Environment and Human Health, University of Exeter, Penryn, UK
3. Department of Microbiology and Immunology, McGill University, Canada
4. McGill Genome Centre, McGill University, Canada
   *correspondence: jesse.shapiro@mcgill.ca

keywords: microbiome, evolution, ecology, Earth Microbiome Project, 16S rRNA

Abstract

Microbes are embedded in complex communities where they engage in a wide array of intra- and inter-specific interactions. The extent to which these interactions drive or impede microbiome diversity is not well understood. Historically, two contrasting hypotheses have been suggested to explain how species interactions could influence diversity. 'Ecological Controls' (EC) predicts a negative relationship, where the evolution or migration of novel types is constrained as niches become filled. In contrast, 'Diversity Begets Diversity' (DBD) predicts a positive relationship, with existing diversity promoting the accumulation of further diversity via niche construction and other interactions. Using high-throughput amplicon sequencing data from the Earth Microbiome Project, we provide evidence that DBD is strongest in low-diversity biomes, but weaker in more diverse biomes, consistent with biotic interactions initially favoring the accumulation of diversity (as predicted by DBD). However, as niches become increasingly filled, diversity hits a plateau (as predicted by EC).

Impact statement

Microbiome diversity favors further diversity in a positive feedback that is strongest in lower-diversity biomes (*e.g.* guts) but which plateaus as niches are increasingly filled in higher-diversity biomes (*e.g.* soils).

Introduction

The majority of the genetic diversity on Earth is encoded by microbes (Hug et al. 2016; Lapierre and Gogarten 2009; Sunagawa et al. 2015) and the functioning of all Earth's ecosystems is reliant on diverse microbial communities (Falkowski, Fenchel, and Delong 2008). High-throughput 16S rRNA gene amplicon sequencing studies continue to yield unprecedented insight into the taxonomic richness of microbiomes (e.g. (Louca et al. 2019; Sogin et al. 2006)), and abiotic drivers of community composition (e.g. pH; (Lauber et al. 2009; Power et al. 2018)) are increasingly characterized. Although it is known that biotic (microbe-microbe) interactions can also be important in determining community composition (Needham and Fuhrman 2016), comparatively little is known about how such interactions, either positive (e.g. cross-feeding; (Seth and Taga 2014)) or negative (e.g. toxin-mediated interference competition; (Czaran, Hoekstra, and Pagie 2002; Hibbing et al. 2010)), shape microbiome diversity as a whole.

The dearth of studies exploring how microbial interactions could influence diversity stands in marked contrast to a long research tradition on biotic controls of plant and animal diversity (Elton 1946; Gause 2003). In an early study of 49 animal (vertebrate and invertebrate) community samples, Elton plotted the number of species versus the number of genera and observed a ~1:1 ratio in each individual sample, but a ~4:1 ratio when all samples were pooled (Elton 1946). He took this observation as evidence for competitive exclusion preventing related species, more likely to overlap in niche space, to co-exist. This concept, more recently referred to as niche filling or Ecological Controls (EC) , predicts speciation (or, more generally, diversification) rates to decrease with increasing standing species diversity because less niche space is available (Rabosky and Hurlbert 2015). In contrast, the Diversity Begets Diversity (DBD) model predicts that when species interactions create novel niches, standing biodiversity favors further diversification (Calcagno et al. 2017; Whittaker 1972). For example, niche construction (i.e. the physical, chemical or biological alteration of the environment) could influence the evolution of the species constructing the niche, as well as that of co-occurring species (Laland, Odling-Smee, and Feldman 1999; San Roman and Wagner 2018). An alternative to either EC or DBD is The Neutral Theory of Biodiversity and Biogeography, in which all species are functionally equivalent and communities assemble via random sampling (Hubbell 2001). Neutral Theory

serves as a null hypothesis of community assembly in animals and plants (Sandro et al. 2016; Gotelli and Colwell 2001), and more recently in microbiome research (Harris et al. 2017; Li and Ma 2016).

Empirical evidence for the action of EC vs. DBD in natural plant and animal communities has been mixed (Emerson and Kolm 2005; Palmer and Maurer 1997; Price et al. 2014; Rabosky et al. 2018). Laboratory evolution experiments tracking the diversification of a focal bacterial lineage in communities of varying complexity have also yielded contradictory results, with support for EC, DBD, or intermediate scenarios (Brockhurst et al. 2007; Meyer and Kassen 2007). For example, diversification of a focal *Pseudomonas* clone was favored by increasing community diversity in the range of 0-20 other strains or species within the same genus (Jousset et al. 2016b; Calcagno et al. 2017) but diversification was inhibited in highly diverse communities (*e.g.* hundreds or thousands of species in compost; (Gómez and Buckling 2013)). These experiments are consistent with interspecific competition initially driving (Bailey et al. 2013), but eventually inhibiting diversification as niches are filled.

Most laboratory experiments are restricted to relatively short evolutionary time scales and include only a small number of taxa; it is therefore unclear if they can be generalized to natural communities consisting of many more taxa evolving and assembling over much longer periods, spanning more environmental change, greater evolutionary diversification, and frequent migration events. Although the absence of a substantial prokaryotic fossil record hinders deconvoluting speciation and extinction rates (Louca and Pennell 2020; Marshall 2017), Louca et al. (Louca et al. 2018) recently estimated that bacterial diversity has mostly increased over the past billion years, with speciation rates slightly exceeding extinction rates. However, because many free-living microbes have high migration rates ("everything is everywhere, but the environment selects" (de Wit and Bouvier 2006)), we expect that the majority of diversity present within a typical microbiome sample is selected from a pool of migrants rather than having evolved *in situ*. As such, here we broadly define "diversity begets diversity" (DBD) to include the combined effects of community assembly from a migrant pool ('ecological species sorting') and *in situ* evolutionary diversification (**Fig. 1**).

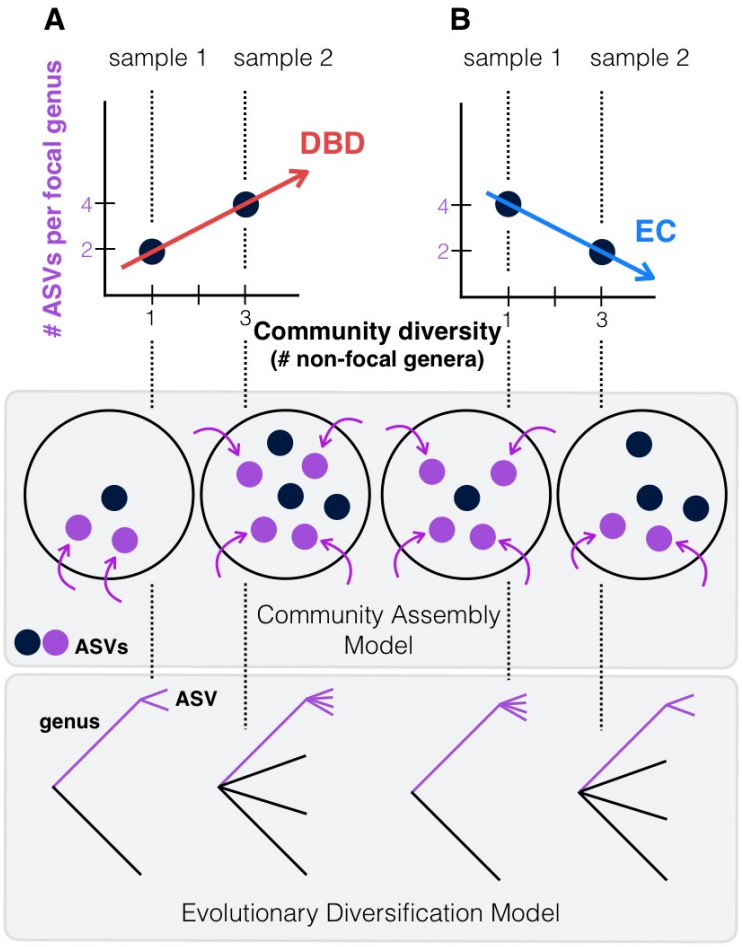

Figure 1. Contrasting the Diversity Begets Diversity (DBD) and Ecological Controls (EC) models. *(A) In this hypothetical scenario, microbiome sample 1 contains one non-focal genus, and two amplicon sequence variants (ASVs) within the focal genus (point at x=1, y=2 in the plot). Sample 2 contains three non-focal genera, and four ASVs within the focal genus (point at x=3, y=4). Tracing a line through these points yields a positive diversity slope, supporting the DBD model (red). (B) Alternatively, a negative slope would support the Ecological Controls (EC) model (blue line). In the middle panel, we consider a community assembly model to explain the hypothetical data of the top panel, in which standing diversity (black points) in a community selects (for or against) new types (referred to here as ASVs) which arrive via migration (purple points & arrows). In the bottom panel, we consider an evolutionary diversification model of a focal lineage (genus) into ASVs as a function of initial genus-level community diversity present at the time of diversification.*

To test whether patterns of diversity in natural communities conform to EC or DBD dynamics, we used 2,000 microbiome samples from the Earth Microbiome Project (EMP), the largest available repository of biodiversity based on standardized sampling and sequencing protocols, with 16S rRNA gene amplicon sequence variants (ASVs) as the finest-grained taxonomic unit (Thompson et al. 2017). Following Elton (Elton 1946), we use the equivalent of Species:Genus ratios, calculating a range of taxonomic diversity ratios (up to the Class:Phylum level) as proxies for diversity within a focal taxon, from shallow to deep evolutionary time. We

then plot each ratio as a function of the number of non-focal taxa (Genera, Families, Orders, Classes, and Phyla, respectively) with which the focal taxon could interact. We refer to the slope of these plots as the "diversity slope", with negative slopes supporting EC and positive slopes supporting DBD (**Fig. 1**). As a null, we compare these slopes to the expectation under Neutral Theory. To avoid a trivially positive diversity slope due to variation in sequencing effort, all samples were rarefied to 5,000 observations (counts of 16S rRNA gene sequences), as diversity estimates are highly sensitive to sampling effort (Gotelli and Colwell 2001). As 16S evolves at a rate of roughly 1-2 substitutions per million years (Kuo and Ochman 2009b), evolutionary diversification within individual EMP samples cannot be uncovered using this marker; rather our data represent mainly a record of community assembly.

Results

Quantifying the DBD-EC continuum in prokaryote communities compared to neutral null models

We used generalized linear mixed models (GLMMs) to estimate the diversity slope at each taxonomic level in the EMP data, which revealed a tendency toward positive slopes with significant variation explained by the random effects of lineage, environment, and their interaction (**Table 1, Figure 2, Figure 2 supplements 1-6, Supplementary Data file 1 Section 1**). All models reported here provide significantly better fits compared to models without the fixed effect of community diversity, and coefficients of determination (R^2) are higher with the inclusion of random effects, showing their importance (**Supplementary Data file 2**). Examples of how the diversity slope varies across lineages and environments are shown in **Figure 2** and **Figure 2 supplements 2-6**. To assess the significance of these slope estimates in light of potential sampling bias and data structure (Gotelli and Colwell 2001; Jarvinen 1982), we considered null models, all of which randomize the associations between ASVs within a sample, thus randomizing any true biotic interactions. Models 1 and 2 are based on draws from the zero-sum multinomial (ZSM) distribution, which arises from the standard Neutral Theory of Biodiversity (**Methods**). Model 1, in which each microbiome sample is drawn from the same ZSM distribution, produces a significantly negative diversity slope (**Figure 2 supplement 7; Table 2**). Model 2, in which each

environment draws from a separate distribution, is effectively a composite of Model 1 in which different environments, each with a negative slope, are 'stacked' to yield an overall positive slope (**Figure 2 supplement 7**). However, the Model 2 slope is not significant in a GLMM accounting for variation across environments (**Table 2, Supplementary Data file 3 Section 1.2**). In the real EMP data, most individual environments tend toward a positive slope (**Figure 2 supplement 8**). The tendency toward positive diversity slopes in the EMP is therefore not straightforwardly explained by neutral processes.

To estimate the power to detect either DBD or EC, we specifically added each of these effects to data simulated under a null model. As expected, adding DBD reversed the negative slope and rendered it positive (**Table 2; Figure 2 supplement 7, Supplementary Data file 3 Section 2.1**), suggesting reasonable power to detect DBD when truly present. In contrast, the addition of EC had little effect on the slope, suggesting low power to detect EC under some null models. Taken together, these modelling results suggest that positive diversity slopes observed in the EMP are more readily explained by DBD than by Neutral Theory, whereas negative slopes could be explained by EC, Neutral Theory, or some combination of the two.

Because taxonomic labels can be unavailable or inconsistent with phylogenetic relationships (Parks et al. 2018; Vos 2011) we repeated the analyses using nucleotide sequence identity in the 16S rRNA gene instead of taxonomy, and again recovered generally positive diversity slopes (**Methods**). As a final sensitivity analysis, we repeated the GLMMs using unrarefied community Shannon diversity instead of richness (**Methods**) and obtained similar results, with generally positive diversity slopes that could in some cases be reversed depending on the lineage or environment (**Table 3, Supplementary Data file 1 Section 2**). The Shannon diversity metric is robust to sampling effort, suggesting that the results are not biased by undersampling in diverse biomes. Even if undersampling could bias the diversity slope downward in more diverse samples, the effect is unlikely to be large at a rarefaction to 5,000 sequences, and only to occur at the extremes of diversity (*e.g.* very many genera and high ASV:genus ratios) and not at higher taxonomic levels (*e.g.* Class:Phylum) (**Figure 2 supplement 9**).

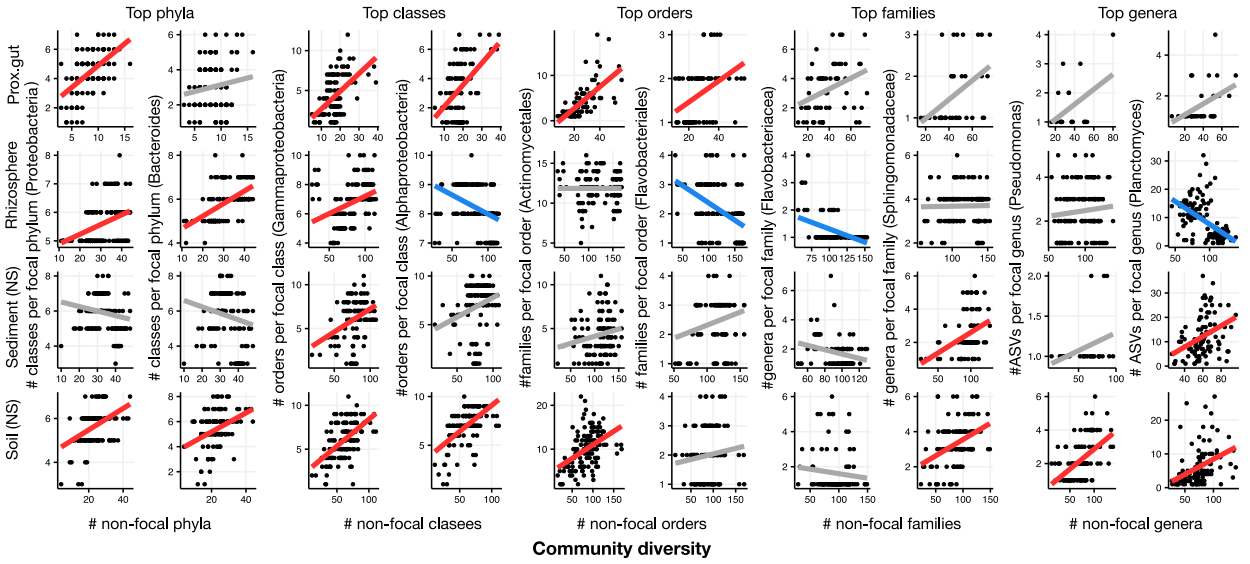

Figure 2. Focal lineage diversity as a function of community diversity in the top two most prevalent taxa at each taxonomic level. *As in* ***Fig. 1****, the x-axes show community diversity in units of the number of non-focal taxa (e.g. the number of non-Proteobacteria phyla for the left-most column), and the y-axes show the taxonomic ratio within the focal taxon (e.g. the number of classes within Proteobacteria). Significant positive diversity slopes are shown in red, negative in blue (linear models, P <0.05, Bonferroni corrected for 17 tests), and non-significant in grey. Note that linear models are distinct from GLMMs, and are for illustrative purposes only. Four representative environments are shown (see* ***Figure 2 supplements 2-6*** *for plots in all 17 environments).*

DBD reaches a plateau at high diversity

It is expected from theory and experimental studies that a positive DBD relationship should eventually reach a plateau, giving way to EC as niches become saturated (Brockhurst et al. 2007; Gómez and Buckling 2013). This expectation is borne out in our dataset, particularly in the nucleotide sequence-based analyses which support quadratic or cubic relationships over linear diversity slopes (**Figure 2 supplement 10**). For example, in the animal distal gut, a relatively low-diversity biome, we observed a strong linear DBD relationship at most phylogenetic depths; in contrast, the much more diverse soil biome clearly reaches a plateau (**Figure 2 supplement 11**).

To comprehensively test the hypothesis that more diverse microbiomes experience weaker DBD due to saturated niche space, we used a GLMM including the interaction between diversity and environment as a fixed effect. We considered this model only for taxonomic ratios with significant diversity slope variation by environment (**Table 1**): Family:Order, Order:Class, and Class:Phylum.

Diversity slopes were significantly higher in less diverse (often host-associated) biomes, suggesting that niche filling leads to a plateau of DBD in more diverse biomes (**Fig. 3, Supplementary Data file 1 Section 3**). The interaction observed in the real EMP data between community diversity and biome type in shaping focal lineage diversity was not observed under a neutral null (Model 2, in which each environment has its own characteristic level of diversity) (**Supplementary Data file 3 Section 1.2**). The DBD plateau observed in more diverse biomes is thus not readily explained by a neutral model, nor is rarefaction expected to bias the diversity slope estimates, particularly at the Class:Phylum level (**Figure 2 supplement 9**). This suggests that the plateau of DBD at higher levels of community diversity is not an artefact of data structure or sampling effort. Finally, we considered whether variation along the EC-DBD continuum could be explained by differential cell density across environments, which could affect both the frequency of cell-cell interactions (a biological effect) or the sampling depth (a technical artefact). Although precise estimates of cell densities in all EMP biomes are not available, we extracted plausible ranges for eight biomes from the literature (Kennedy and de Luna 2005; Lindow and Brandl 2003; Sender, Fuchs, and Milo 2016; Whitman, Coleman, and Wiebe 1998) and annotated these in **Figure 3**. It is clear from this figure that relatively high- and low-density samples are found along the range of community taxonomic diversities, demonstrating that cell density is unlikely to drive the trend of decreasing diversity slopes with increasing community diversity.

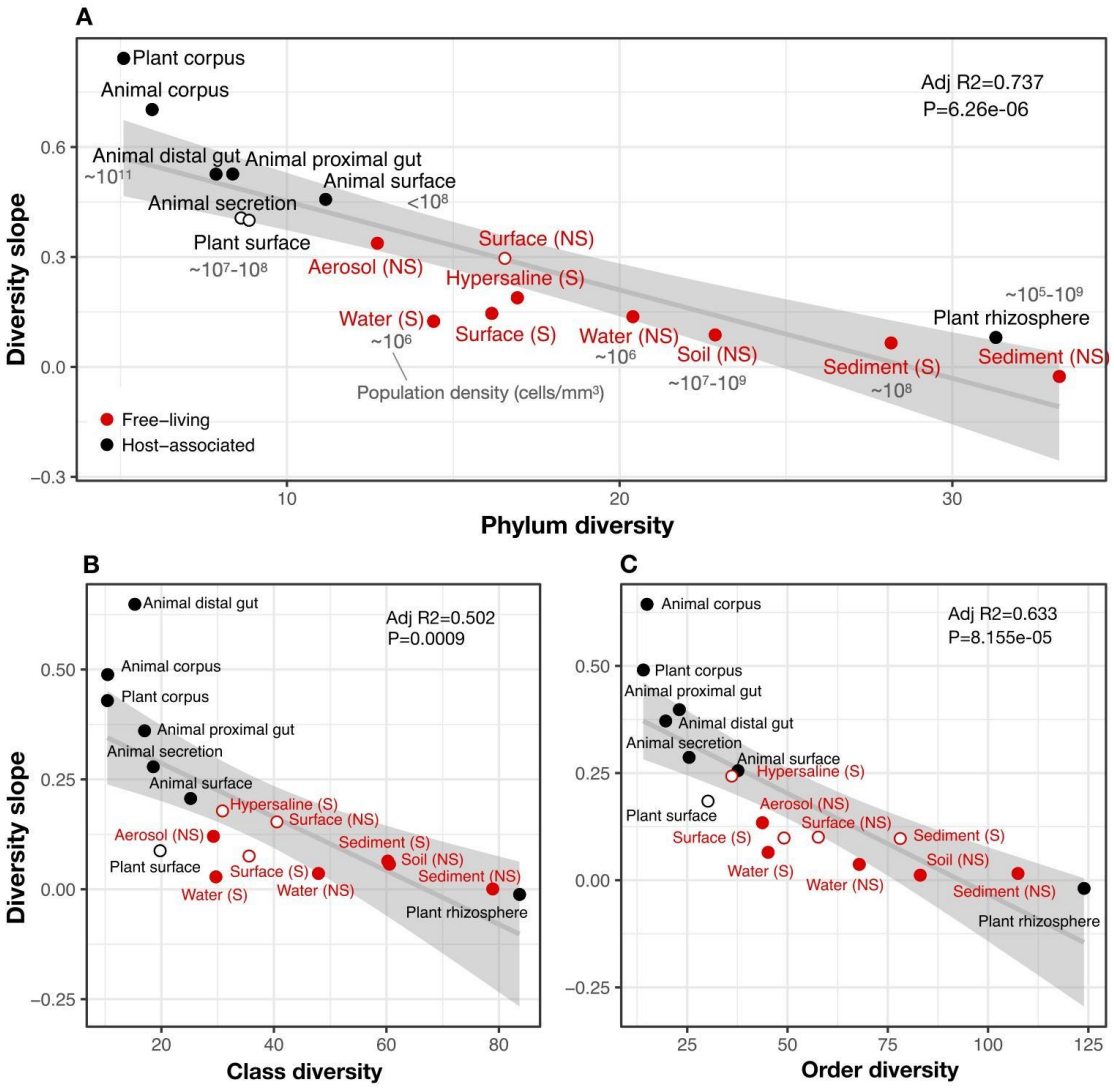

Figure 3. The diversity slope of focal taxa is higher in low-diversity (often host-associated) microbiomes. *The x-axis shows the mean number of non-focal taxa: (A) phyla, B) classes, and C) orders in each biome. On the y-axis, the diversity slope was estimated by a GLMM predicting focal lineage diversity as a function of the interaction between community diversity and environment type at the level of A) Class:Phylum, B) Order:Class, and C) Family:Order ratios (**Supplementary Data file 1 Section 3**). The line represents a linear regression; the shaded area depicts 95% confidence limits of the fitted values. Adjusted R^2 and P-values from the linear fits are shown at the top right of each panel. See **Supplementary Data file 2** for model goodness of fit. Slopes not significantly different from zero are shown as empty circles. Estimates of bacterial cell density from the literature are indicated in grey text, in units of bacteria/mm^3. For animal (skin) and plant surface, units of bacteria/mm^2 were converted to mm^3 assuming layers of bacteria 1 micron thick. For rhizosphere samples we assume a density of 1-2g/cm^3 (Kennedy and de Luna 2005).*

Abiotic drivers of diversity

Our results thus far suggest that community diversity is a major determinant of the EC-DBD continuum, and by extension that biotic interactions may override abiotic factors in determining where a community lies on the continuum. To formally test for the additional role abiotic drivers might play in generating the observed EC-DBD continuum, we analyzed two data sets in more detail.

First, we analyzed a subset of 192 EMP samples with measurements of four key abiotic factors shown to affect microbial diversity (pH, temperature, latitude, and elevation; (Delgado-Baquerizo et al. 2018; Lauber et al. 2009; Power et al. 2018; Schluter and Pennell 2017)). We fitted a GLMM with focal lineage-specific diversity as the dependent variable, and with the number of non-focal lineages, the four abiotic factors and their interactions as predictors (fixed effects). As in the full EMP dataset (**Table 1**), focal lineage diversity was positively associated with community diversity at all taxonomic ratios in the EMP subset (**Table 4**). As expected, certain abiotic factors, alone or in combination with diversity, had significant effects on focal lineage diversity (**Table 4**). However, the effects of abiotic factors were always weaker than the effect of community diversity (**Table 4**; **Supplementary Data file 1 Section 4**).

Second, we used a global 16S sequencing dataset of 237 soil samples associated with more detailed environmental metadata (Delgado-Baquerizo et al. 2018) which we reprocessed to yield ASVs comparable to those in the EMP (**Methods**). This dataset revealed weaker evidence for DBD and stronger effects of abiotic variables on diversity. Community diversity generally had significant positive effects on focal-lineage diversity, but the effect was weak and not detectable at all taxonomic ratios (**Table 5**). Known abiotic drivers of soil bacterial diversity such as pH (Lauber et al. 2009) and latitude (Delgado-Baquerizo et al. 2018) had effects of similar or stronger magnitude compared to the effect of community diversity (**Table 5, Supplementary Data file 4**). The relatively weak effect of DBD and strong effect of abiotic drivers on diversity in this soil dataset can be explained by the fact that soils generally are highly diverse and have relatively low diversity slopes (**Figure 3**).

We note that it remains possible that unmeasured abiotic effects could explain some of the DBD effects observed in the EMP. Although only a small subset of abiotic factors was

considered, the generally positive diversity slopes in the EMP are not likely to be driven by these factors in the abiotic environment (**Table 4**). Specifically, we consider it unlikely that unmeasured abiotic factors would always act similarly, and in the same direction across multiple different environments, to drive DBD. However, as demonstrated in soil (**Table 5**), abiotic factors may become increasingly important in highly diverse biomes with weak DBD.

DBD is more pronounced in resident taxa than in migrant- or generalist taxa

A recent meta-analysis of 16S sequence data from a variety of biomes suggests there is an important distinction between generalist lineages found in many environments, compared to specialists with a more restricted distribution (Sriswasdi, Yang, and Iwasaki 2017). Generalists were inferred to have higher speciation rates, suggesting that the DBD-EC balance might differ between generalists and specialists (Sriswasdi, Yang, and Iwasaki 2017). To further investigate this difference, we defined 'residents', taxa with a strong preference for a specific biome, in addition to generalists without a strong biome preference in the EMP dataset. We first clustered environmental samples by their genus-level community composition using fuzzy *k*-means clustering (**Fig. 4a**), which identified three major clusters: 'animal-associated', 'saline', and 'non-saline'. The clustering included some outliers (*e.g.* plant corpus grouping with animals), but was generally consistent with known distinctions between host-associated vs. free-living (Thompson et al. 2017), and saline vs. non-saline communities (Auguet, Barberan, and Casamayor 2010; Lozupone and Knight 2007). Resident genera were defined as those with a strong preference for a particular environment cluster (whether due to dispersal limitation or narrow niche breadth) using indicator species analysis (permutation test, $P<0.05$; **Fig. 4a**; **Figure 4 supplement 1**; **Supplementary Data file 5**), and genera without a strong preference were considered generalists. When residents of one environmental cluster were (relatively infrequently) observed in a different cluster, we defined them as "migrants" in that sample. For each environment cluster, we ran a GLMM with resident genus-level diversity (the number of non-focal genera) as a predictor of focal-lineage diversity (the ASV:Genus ratio) for residents, generalists, or migrants to that sample (**Supplementary Data file 1 Section 5**).

Resident community diversity had no significant effect on the diversity of generalists in animal-associated, saline and non-saline clusters (GLMM, Wald test, *P*>0.05), but was positively correlated with lineage-specific resident diversity (GLMM, Wald test, z=7.1, ***P*=** 1.25e-12; z=3.316, *P*=0.0009; z=7.109, *P*=1.17e-12, respectively). Resident community diversity significantly decreased migrant diversity in saline (GLMM, z=-3.194, *P*=0.0014) and non-saline environment clusters (GLMM, z=-2.840, *P*=0.0045), but had no significant effect in the animal-associated cluster (GLMM, *P*>0.05) (**Fig. 4b**). These results suggest that, although generalist lineages may have higher speciation rates and colonize more habitats than specialists (Sriswasdi, Yang, and Iwasaki 2017), they have lower diversity slopes. Migrants to the "wrong" environment experience even less DBD, and are even subject to EC in two out of three environment types (**Fig. 4b**). The accumulation of diversity via successful establishment of migrants may thus be limited, presumably because most niches are already occupied by residents.

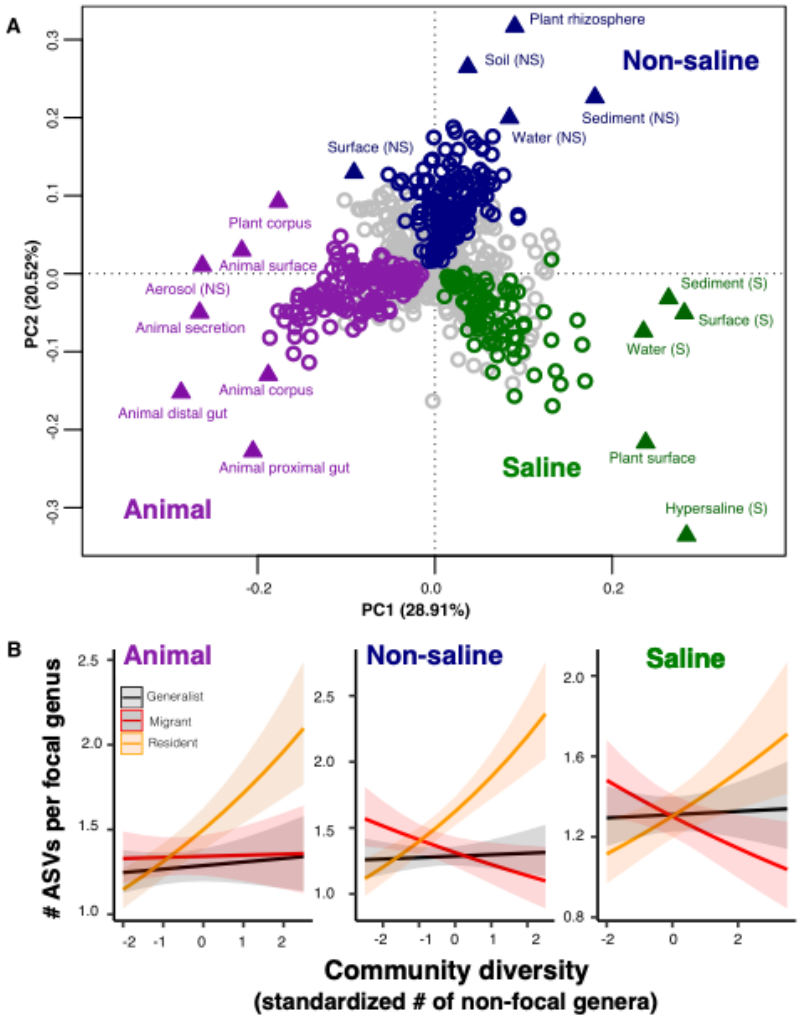

Figure 4. The DBD relationship varies between resident and non-resident genera*. (A) Ordination showing genera clustering into their preferred environment clusters. The matrix of 1128 genera (rows) by 17 environments (columns), with the matrix entries indicating the percentage of samples from a given environment in which each genus is present, was subjected to principal components analysis (PCA). Circles indicate genera and triangles indicate environments (EMPO 3 biomes). Colored circles are genera inferred by indicator species analysis to be residents of a certain environmental cluster, and grey circles are generalist genera. The three environment clusters identified by fuzzy k-means clustering are: Non-saline (NS, blue), saline (S, green) and animal-associated (purple). Triangles of the same color indicate EMPO 3 biomes clustered into the same environmental cluster. (B) DBD in resident versus non-resident genera across environment clusters. Results of GLMMs modeling focal lineage diversity as a function of the interaction between community diversity and resident/migrant/generalist status. The x-axis shows the standardized number of non-focal resident genera (community diversity); the y-axis shows the number of ASVs per focal genus. Resident focal genera are shown in orange, migrant focal genera in red, and generalist focal genera in black. Red stars indicate a significantly positive or negative slope (Wald test, P<0.005). See Supplementary Data file 2 for model goodness of fit.*

Discussion

Using ~10 million individual marker sequences from the EMP, we demonstrate an overall trend for diversity in focal lineages to be positively associated with overall community diversity, albeit with significant variation across lineages and environments. The strength of the DBD relationship dissipates with increasing microbiome diversity, which we hypothesize is caused by niche saturation. In more diverse biomes such as soil, abiotic factors therefore may become relatively more important in driving focal-lineage diversity. The effect of DBD is strongest among habitat specialists (residents), suggesting that long-term niche adaptation tends to select against the establishment of migrant diversity.

While most of the DBD literature considers a model of evolutionary diversification (Schluter and Pennell 2017; Whittaker 1972), our results pertain mainly to ecological community assembly dynamics. At the limited resolution of 16S rRNA gene sequences, we do not expect measurable diversification within an individual microbiome sample (Kuo and Ochman 2009b); however, community diversity could still select for (as in DBD) or against (as in EC) increasing diversity in a focal lineage, even if this lineage diversified before the sampled community assembled. Future work with higher resolution genomic or metagenomic data will enable testing if and how DBD arises in microbial communities via evolutionary diversification, and also how prokaryote diversification is affected by other community members including phages (Brockhurst, Buckling, and Rainey 2005), protists (Meyer and Kassen 2007), and fungi (Kastman et al. 2016). Predator-prey, cross-feeding, and other biotic interactions with these non-prokaryotic community members could explain some of the unaccounted variation we observed in diversity slopes across environments.

Our dataset also provides an opportunity to explore how DBD relates with genome size evolution. Bacteria with larger repertoires of accessory genes, and thus larger genomes, are able to occupy a wider range of niches (Barberan et al. 2014). Taxa with larger genomes might therefore be hypothesized to better survive and thrive when they disperse into a new location, exhibiting stronger DBD. Although a comprehensive test of this hypothesis will require higher resolution genomic or metagenomic data, as a preliminary exploration we assigned genome sizes to 576 focal genera for which at least one whole genome sequence was available (using the

largest recorded genome size for each genus) and added an interaction term between genome size and diversity as a fixed effect in the GLMM (**Methods**). Consistent with our expectation, we observed a significant positive effect of genome size on the diversity slope (GLMM, Wald test, z=2.5, *P*=0.01; **Fig. 5, Supplementary Data file 1 Section 6**). This effect was not observed in null models, in which the interaction between community diversity and focal genus genome size was never significant (**Supplementary Data file 3 Section 1.3 and 2.2**) and so this effect of genome size cannot be trivially explained by data structure. The positive relationship between genome size and DBD is likely even stronger than estimated, because assigning genome sizes to entire genera is imprecise (*i.e.* there is variation in genome size within a genus, or even within species), therefore weakening the correlation.

The positive correlation between genome size and DBD observed here could be driven by larger metabolic repertoires encoded by larger genomes (Barberan et al. 2014), potentially creating more opportunities to benefit from cross-feeding, niche construction (San Roman and Wagner 2018), and other interspecies interactions. This tendency appears to be at odds with the Black Queen hypothesis, which predicts that social conflict between interacting species leads to the inactivation and loss of genes involved in shareable metabolites (public goods), eventually resulting in reduced genome size (Morris and Lenski 2012). Such a process would produce a negative correlation between the degree of species interactions (*i.e.* community diversity) and genome size (Morris and Lenski 2012). The interaction between genome size, biotic interactions and diversification thus deserves further study.

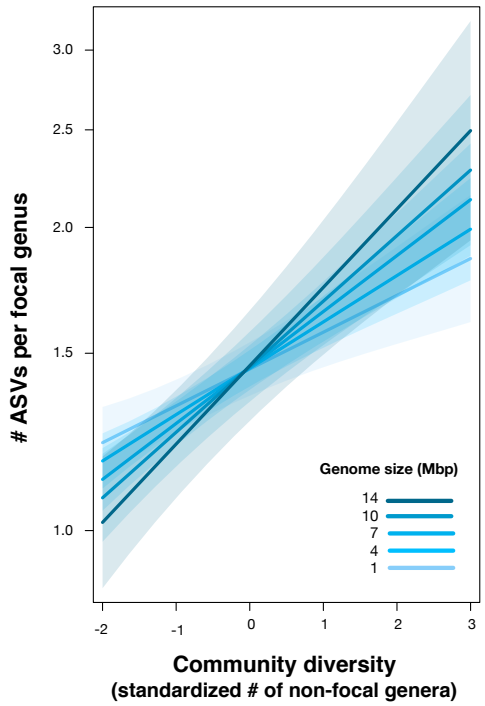

Figure 5. Positive effect of genome size on DBD. *Results are shown from a GLMM predicting focal lineage diversity as a function of the interaction between community diversity and genome size at the ASV:Genus ratio (**Supplementary Data file 1 Section 6**). The x-axis shows the standardized number of non-focal genera (community diversity); the y-axis shows the number of ASVs per focal genus. Variable diversity slopes corresponding to different genome sizes are shown in a blue color gradient; the shaded area depicts 95% confidence limits of the fitted values. See **Supplementary Data file 2** for model goodness of fit.*

Alongside theory and experimental data, the EMP survey data provide a window into the biotic drivers of microbial diversity in nature. In particular, our correlational results support previous experimental and theoretical results showing that DBD is strong when community diversity is low (Jousset et al. 2016b; Calcagno et al. 2017), driving the accumulation of diversity in a positive feedback loop until niches are filled and EC starts to predominate (Bailey et al. 2013; Brockhurst et al. 2007; Gómez and Buckling 2013; Meyer and Kassen 2007). However, due to the correlational nature of the EMP data, it is not possible to test whether DBD is primarily due to the creation of novel niches via biotic interactions and niche construction (Laland, Odling-Smee, and Feldman 1999), or due to increased competition leading to specialization on underexploited resources (Hibbing et al. 2010; Jousset et al. 2016b). We hope future higher resolution genomic studies, and complementary experiments, will be able to elucidate the types of biotic interactions that promote microbiome diversity. Regardless of the underlying mechanisms, our results

demonstrate a general scaling between different levels of community diversity, which has important implications for modeling and predicting community function and stability in response to perturbations (Coyte, Schluter, and Foster 2015; Pennekamp et al. 2018). The answer to the question 'why are microbiomes so diverse?' might in a large part be because microbiomes are so diverse (Emerson and Kolm 2005).

Materials and Methods

Earth Microbiome Project dataset

We used the EMP '2000 subset' of 16S rRNA gene sequences, rarefied to 5000 sequences per sample. This subset contains 155,002 ASVs from 2,000 samples with an even distribution across 17 natural environments (EMP Ontology level 3). Data were downloaded from the EMP FTP server (ftp.microbio.me), on February 9, 2018.

Specifically, 16S rRNA-V4 region reads (90 bp, GreenGenes 13.8 taxonomy) along with environmental data and EMPO3 designations (http://press.igsb.anl.gov/earthmicrobiome/protocols-and-standards/empo/) were downloaded from the EMP FTP server (ftp.microbio.me), on February 9, 2018. Sequence summaries were downloaded from:

ftp://ftp.microbio.me/emp/release1/otu_distributions/otu_summary.emp_deblur_90bp.subset_2k.rare_5000.tsv, environmental data from:

ftp://ftp.microbio.me/emp/release1/mapping_files/emp_qiime_mapping_release1.tsv, and EMPO3 designations from :

ftp://ftp.microbio.me/emp/release1/mapping_files/emp_qiime_mapping_subset_2k.tsv.

The list of the associated 97 studies and 61 corresponding principal investigator identities were downloaded from https://www.nature.com/articles/nature24621#s1.

Based on the ASV annotations across samples, we estimated the taxonomic ratio for each focal lineage (ASV:Genus, Genus:Family, Family:Order, Order:Class and Class:Phylum), along with the number of non-focal lineages (dbd_analys_input.py, glmm_analys_input.py, Python Version 2.7). Unclassified ASVs were removed from the analyses.

Generalized linear mixed model (GLMM) analyses

We used GLMMs to determine how focal lineage diversity (*e.g.* its ASV:Genus ratio) is affected by community diversity (*e.g.* non-focal genera). The effects of environment (as defined by the EMP Ontology 'level 3 biomes') and the focal lineage identity were included as random effects on the slope and intercept. We also controlled for the submitting laboratory (identified by the principal investigator) and the EMP unique sample identifier (i.e. if two taxa were part of the same sample).

All models were fitted in Rstudio (Version 1.1.442, R Version 3.5.2) using the glmer function of the lme4 package (Bates et al. 2015). Data standardization (transformation to a mean of zero and a standard deviation of one) was applied to all predictors to get comparable estimates. In models with only one predictor, applying standardization resolved convergence warnings and considerably sped up the optimization. We first tested the significance of random effects, by using likelihood-ratio tests (LRTs, implemented in the anova function in the R stats package) on nested models where each random effect was dropped one at a time. We then assessed the significance of fixed effects using drop1 function from stats package with the likelihood-ratio test option (this function drops individual terms from the full model and compares models based on the AIC). We calculated the Akaike information criterion (AIC) of each significant model and a null model including all random effects but no fixed effects other than the intercept. We then report the difference in AIC between the full and null models (ΔAIC), along with a likelihood ratio test *p*-value to assess the significance of the full model relative to the null. Only significant models ($P<0.05$) are reported.

As an additional test of the goodness of fit for the significant models, we estimated the coefficient of determination (R^2) using the r.squaredGLMM function from the MuMIn R package. This function implements a method developed by Nakagawa and Schielzeth and its extension for random slopes (Johnson 2014; Nakagawa and Schielzeth 2013). Two values were estimated: the marginal R^2, as a measure of the variance explained only by fixed effects, and the conditional R^2 as a measure of the variance explained by the entire model (both fixed effects and random effects). Only results from R^2 estimation based on lognormal and trigamma methods were reported because they are specific to the logarithmic link function used in all GLMMs.

Diagnostic plots (plot and qqnorm R functions in base and stats packages) were checked for each model to ensure that residual homoscedasticity (homogeneity of variance) was fulfilled: no increase of the variance with fitted values and residuals were symmetrically distributed tending to cluster around the 0 of the ordinate, but with an expected pattern due to count data. Normality plots were imperfect, but they generally showed that the residuals were close to being normally distributed. The assumption of normality is often difficult to fulfill with high numbers of observations, as is the case in our models (https://www.statisticshowto.datasciencecentral.com/shapiro-wilk-test/), and non-normality is less of concern than heteroscedastic for the validity of GLMMs (https://bbolker.github.io/mixedmodels-misc/ecostats_chap.html#diagnostics).

We tested for overdispersion using the overdisp_fun R function available at https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html, and found that all the models were not overdispersed, but rather were underdispersed : the ratio of the sum of squared Pearson residuals to residual degrees of freedom was < 1 and non-significant when tested with a chi-squared test. The only exception was Shannon diversity-based GLMMs. In case of underdispersion and given that underdispersion leads to more conservative results, we retained the GLMMs with Poisson error distribution, despite the underdispersion. (GLMM FAQ; Ben Bolker and others; 25 September 2018; https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#underdispersion). For Shannon diversity-based GLMMs, we accounted for overdispersion by adding an observation-level random effect to the GLMMs (Elston et al. 2001).

Taxonomy-based GLMMs

To test how focal lineage diversity (*e.g.* its ASV:Genus ratio) is affected by community diversity (*e.g.* non-focal genera richness), for different environment types and lineages across all taxonomic ratios, we used generalized linear mixed models (GLMMs) fitted on the EMP dataset. As the dependent variable (focal lineage diversity, defined as taxonomic ratios, ASV:Genus, Genus:Family, Family:Order, Order:Class, and Class:Phylum) was a count response, we used a Poisson error distribution with a log link function. Community diversity (number of non-focal lineages: non-focal Genera, Families, Orders, Classes, and Phyla), standardized to a mean of zero

and a standard deviation of one, was specified as the predictor (fixed effect). We included the following random effects on the slope and intercept: lineage (Lin), environment (Env), environment nested within lineage (a lineage may be present in different environments) and lab (the principal investigator who conducted the EMP study) nested within environment (different labs sampled and sequenced a given environment) (as suggested in http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html). Defining random effects on the slope enabled us to test slope variation across groups of each categorical variable (*e.g.* slope variation between different environments or different lineages). We included the EMP unique sample ID as a random effect to control for dependencies between observations (if two taxa were part of the same sample) (**Table 1**, **Supplementary file 1 section 1)**.

Shannon diversity-based GLMMs

We also tested whether ASV diversity in a focal taxon is dependent on the diversity of all other ASVs in that sample (rather than the diversity at only the focal taxonomic level, as in the taxonomy-based GLMMs above). We used the Shannon diversity index, which is robust to differences in sampling effort, and generally reaches a plateau at 5,000 sequences or fewer (48, 49). To do so, we fitted a GLMM with the number of ASVs per focal taxon as the response variable, and the Shannon diversity based on ASVs across all non-focal taxa (z-standardized) as the predictor (fixed effect), the random effects were kept as in the taxonomy-based GLMMs, but we added an observation-level random effect to account for overdispersion (**Table 3**, **Supplementary file 1 section 2**). To avoid dependence between the response and predictor variables, we used the rarefied ASV dataset (5,000 ASVs/sample as above) as the response variable, and the Shannon diversity calculated on unrarefied data from the same samples as the predictor.

Null models

We considered three null models, all of which randomize the associations between ASVs within a sample, thus breaking any true biotic interactions. These null models were randomly generated matrices of the same size as the real EMP dataset, but based on a distribution that arises from

the Neutral Theory of Biodiversity. Neutral Theory postulates that the biodiversity of a metacommunity is governed by independent random population dynamics across species. The aggregate behaviour is quantified by the fundamental biodiversity number ϑ, such that $\theta = 2\,J_M\,\upsilon$, where J_M is the size of the metacommunity and v is the speciation rate. Parametrized by ϑ, the metacommunity zero-sum multinomial distribution (mZSM) was developed to obtain random samples of size J (Alonso and McKane 2004). We used this mZSM distribution (implemented with the *sads* package in R; http://search.r-project.org/library/sads/html/dmzsm.html) to generate the counts of the ASVs for each dataset in models 1 and 2. Model 1 assumes that the whole dataset follows the same species abundance distribution (SAD), characterized by a mZSM with ϑ = 50. Model 2 assumes that each environment has its own SAD and thus all the samples of a single environment are assigned the same ϑ but are distinct across environments (ϑ was chosen uniformly between 1 and 100). The number of samples per environment were the same as the EMP dataset. To obtain similar mean counts as the real dataset, we set J = 1000 for both models 1 and 2, in order to vary ϑ from 1 to 100. These values are reasonable based on previous studies that estimated these parameters from microbiome data (Li and Ma 2016). We included a down-sampling step to replicate the zero-inflated nature of the real dataset (on average there were only 96 ASVs per sample while there was a total of 22,014 ASVs in the entire EMP dataset). To replicate the sampling effect due to rarefaction, we first created a vector of all individuals from a single sample. We then selected 5000 individuals at random whose identities determined which ASVs were found in that sample. These neutrally-derived random matrices, null models 1 and 2, were plotted using the same plots (ASV:Genus vs number of genera) as the real EMP dataset and were then analyzed using GLMMs with community diversity as a predictor of focal lineage diversity (fixed effect), with lineage identity and EMP sample ID as random effects. For Model 1, the slope was significantly negative (GLMM, Wald test, z=-9.807, P<2e-16). For Model 2, the null GLMM (including the intercept only) was significant, meaning that the community diversity has no significant effect on focal lineages diversity (Likelihood-ratio test between the model with the predictor and the intercept-only model, P=0.9399).

To generate a null model for a metacommunity assembled by niche processes, null model 3 was made by sampling from a single Poisson distribution ($\lambda = 0.01$) for each element of the data matrix. We used the Poisson distribution as a sensitivity analysis compared to the ZSM, and found the two behave quite similarly (*i.e.* Model 1 and 3 produce qualitatively similar results). The probability of size zero was sufficiently large that the down-sampling step was not needed for this model. Next, DBD and EC effects were added to null model 3 according to the following procedure. An element was chosen at random in a sample and tested if it is empty or full (*i.e.* checks the presence/absence of a particular ASV). If the element is full then the DBD algorithm fills an empty element chosen at random in the same sample, while the EC algorithm empties a filled element in the same sample. This is to mimic the effect of DBD creating a niche for a new ASV, or EC removing a niche based on the existing diversity. The strength of DBD or EC effects were determined by the percent of elements tested. These data were analyzed with GLMMs to test the power of our models to detect DBD or EC (**Table 2**, Supplementary Data file 3 Section 2.1).

Rarefaction simulation

We constructed a simple simulation in which each microbiome sample may differ in total diversity (*e.g.* in the observed range of genera) while maintaining a constant taxonomic ratio (*e.g.* ASV:genus ratio = 2). To mimic rarefaction, we then sampled a set number of sequencing reads from each synthetic community, assuming ASVs are sampled with equal probability and plotted the observed taxonomic ratio (**Fig. 2 supplement 9**). This simple simulation is implemented in permute_ASVs_synthetic.pl.

Nucleotide sequence-based analysis

We clustered ASVs at decreasing levels of nucleotide identity, from 100% identical ASVs down to 75% identity (roughly equivalent to phyla (Konstantinidis and Tiedje 2005)). We estimated focal cluster diversity as the mean number of descendants per cluster (e.g. number of 100% clusters per 97% cluster) and plotted this against the total number of clusters (97% identity in this example). This approach has the advantage of including sequences even if they come from

unnamed taxa. For each of the six nucleotide divergence ratios tested, the relationship between total number of clusters and focal cluster diversity was positive (**Fig. 2 supplement 10**), consistent with DBD and suggesting that the taxonomic analyses were qualitatively unbiased.

Fasta files with all ASVs per sample were produced by a python script (Construct_fasta_per_sample.py, Python Version 2.7) from the sequences summary file (otu_summary.emp_deblur_90bp.subset_2k.rare_5000 from EMP ftp server). We clustered sequences from each sample using USEARCH V9.2 and estimated sample diversity as the total number of clusters at a given level (*e.g.* 97% identity) and focal cluster diversity as the mean number of descendent clusters (*e.g.* number of 100% clusters per 97% cluster). To describe the putative DBD or EC relationships, we tested three models: linear, quadratic and cubic (lm function in R). Model comparisons were based on the adjusted R^2 (**Figure 2 supplement 10**).

We note that diversity at level i (d_i) and at level i+1 (d_{i+1}/d_i) are not independent in this analysis because d_{i+1} must be greater than or equal to d_i. To assess the effects of this non-independence on the results, we conducted permutation tests by randomizing the associations between d_i and d_{i+1}. Using 999 permutations, *P*-values were calculated based on how many times we observed a correlation greater than that seen in the real data (cor.test R function with kendall method). In each permutation, we recalculated the significance test (Wald z) for the correlation in the randomized data, and then computed the *P*-value based on how many times we observed a z value greater than that of the original data. At all six levels of nucleotide identity, the real data always showed a significantly stronger positive correlation when compared to permuted data ($P = 0.001$), indicating that the DBD patterns was not an artefact of the dependence structure in the data.

The effect of community diversity on focal cluster diversity was also tested across different environments analyzed separately. We modelled this relationship with linear, quadratic and cubic fits, and compared those models based on the adjusted R^2 (**Figure 2 supplement 11**).

DBD variation across environments

We tested the variation of focal lineage diversity slopes across different environments by including EMPO 3 biome type as a fixed effect. We fitted a GLMM with the interaction between

community diversity and environment type as a predictor of focal lineage diversity. All other random effects on intercept and slope were kept as in the previous GLMMs (**Figure 3**, **Supplementary Data file 1 Section 3**). DBD variation across environments was tested for Family:Order, Order:Class and Class:Phylum taxonomic ratios, as diversity slope variation by environment was statistically significant (likelihood-ratio test, $P<0.05$) for these ratios in the taxonomy based models (**Table 1**).

Abiotic effects

To test for the relative effect of biotic and abiotic environmental variables on focal lineage diversity across different taxonomic ratios, we used a separate GLMM, with Poisson error distribution and a log link function, for every ratio. We fitted the GLMM on a subset (~10%) of the whole dataset, 192 samples (from water: saline (19) and non-saline (44), surface: saline (42) and non-saline (19), sediment: saline (22) and non-saline (31), soil (8) and plant rhizosphere (7)), for which measurements of four key abiotic variables (temperature, pH, latitude and elevation) were available. As predictors of focal lineage diversity (fixed effects), we included non-focal community diversity and abiotic variables, as well as their interactions. All predictors were standardized to a mean of zero and a standard deviation of one to obtain comparable estimates. The GLMM had the same random effects as in the previous analysis, but only on the intercept for simplicity (**Table 4**, **Supplementary file 1 section 4**).

Soil dataset analysis

We used the Delgado-Baquerizo et al. 2018 soil microbiome survey (237 samples from 18 countries) to further test the relative impacts of biotic versus abiotic drivers of diversity. Raw data and abiotic measurements were downloaded from Figshare (https://figshare.com/s/82a2d3f5d38ace925492; DOI: 10.6084/m9.figshare.5611321). 16S bioinformatic processing was performed using QIIME2 and Deblur with the same protocol as in Thompson et al. 2017. Raw data 16S rRNA gene (V3-V4 region), were processed by trimming the primers (341F/805R primer set) with qiime cutadapt trim-paired, then merged using qiime vsearch join-pairs. Sequences were quality filtered and denoised using Deblur with a trimming

length of 400bp. The resulting 400-bp Deblur BIOM table was filtered to keep only ASVs with at least 25 reads total over all samples and rarefied to a depth of 5000. Taxonomy was assigned with a Naive Bayes classifier trained on the V4-V3 region of 99% OTU Greengenes 13.8 sequences with qiime feature-classifier. We obtained a final dataset of 186 samples and 24,252 ASVs which was used as input for all statistical analysis as in the EMP dataset analysis. This data set included 14 environmental factors: aridity index (Aridity_Index), minimum and maximum temperature (MINT and MAXT), precipitation seasonality (PSEA), mean diurnal temperature range (MDR), ultra-violet (UV) radiation (UV_Light), net primary productivity (NPP2003_2015), soil texture (Clay_silt), pH; total C (Soil_C), N (Soil_N) and P (Soil_P) concentrations, C:N ratio (Soil_C_N_ratio) and Latitude.

We used a separate GLMM with Poisson error distribution and a log link function to test for the effect of biotic (non-focal community diversity) and abiotic environmental variables on focal lineage diversity (*e.g.* the ASV:Genus ratio for a focal genus), across different taxonomic ratios. We defined non-focal taxa diversity and abiotic variables as predictors (fixed effects) and the lineage identity as a random effect.

We also fitted the same model but with the first three principal components (PCs) from the principal component analysis (PCA, rda function, vegan R package) of the abiotic variables (a matrix of 237 samples (rows) by 14 abiotic variables (columns)), as well as the interactions between diversity and each PC, and the interaction between PCs as predictors (fixed effects).

Because of possible non-linear relationships between abiotic variables and diversity, GLMMs were fitted with a linear and a quadratic term for every abiotic variable. The quadratic terms were not significant, except for the ASV:genus ratio (**Table 5**; likelihood-ratio test, $P < 2.2e\text{-}16$). The interaction terms were not significant except the interaction between diversity and PCs at Family:Order ratio (likelihood-ratio test, $P = 2.182e\text{-}05$; **Table 5**, **Supplementary file 4**).

Defining residents, generalists, and migrants

We defined a genus-level community composition matrix as a matrix of 1128 genera (rows) by 17 environments (columns), with the matrix entries indicating the percentage of

samples from a given environment in which each genus is present. We clustered the environmental samples based on their genus-level community composition using fuzzy *k*-means clustering. The clustering (cmeans function, package e1071 in R) was done on the 'hellinger' transformed data (decostand function, vegan R package). To identify resident genera to each cluster, we used indicator species analysis (Dufrene and Legendre 1997) as implemented in the indval function (labdsv R package). We defined residents as genera with indval indices between 0.4 and 0.9, with permutation test $P < 0.05$. Genera not associated with any cluster were considered generalists. We used principal component analysis (PCA) on the community composition matrix to visualize the clustering and the indicator genera (rda function, vegan R package) (**Figure 4**). We then ran a separate GLMM for each environmental cluster, with resident genus-level diversity (number of non-focal genera) as a predictor of focal genus diversity (ASV:Genus ratio) for resident, migrant (residents of one cluster found in a different cluster) and generalist genera. The fixed effect was specified as the interaction between diversity and a factor defining the genus-cluster association (with three levels: resident, migrant and generalist). Random effects on intercept and slope were kept as in the GLMMs described above.

Genome size analysis

We chose a subset of genera represented by one or more sequenced genomes in the NCBI microbial genomes database (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/). For these genera, a representative genome size was assigned by selecting the genome with the lowest number of scaffolds (if no closed genomes were available) (**Supplementary file 6**). If multiple genomes were available with the same level of completion, the largest genome size was used, as smaller genomes could be artefacts of incomplete assembly which would bias the mean and median downward. Moreover, given the deletional bias in bacterial genomes (Kuo and Ochman 2009a), the largest genome is likely more reflective of the ancestral genome size of the genus. Only genera with two or more ASVs in at least one sample were included in the analysis. Intracellular symbionts were excluded. We fitted a GLMM on the subset of data with known genome size (576 genera, ranging from ~1 to 15 Mbp) with the interaction between community diversity and genome size as a predictor of focal lineage diversity at the ASV:Genus level. All the

other random effects on intercept and slope were kept as in the previous GLMMs (**Supplementary file 1 section 6)**.

Acknowledgements

We thank Luke Thompson for assistance obtaining EMP data and Zofia Ecaterina Taranu, Vincent Fugère and Guillaume Larocque for advice on GLMMs. We are also grateful to Steven Kembel, Tom Battin, the reviewers Eric Kemen and Benjamin E. Wolfe, and the editor Detlef Weigel for critical comments that improved the manuscript. **Funding:** This project was made possible by an NSERC Discovery Grant and Canada Research Chair to BJS.

Competing interests

None to declare.

Data and materials availability

All data is available from the Earth Microbiome Project (ftp.microbio.me), as detailed in the Methods. All computer code used for analysis are available at https://github.com/Naima16/dbd.git.

Tables

Table 1. Effects of community diversity on focal lineage diversity across taxonomic ratios. *The GLMMs showed statistically a significant positive effect of community diversity on focal lineage diversity. Each row reports the effect of community diversity on focal lineage diversity (Div), as well as its standard error, Wald z-statistic for its effect size and the corresponding P-value (left section), or standard deviation on the slope for the significant random effects (right section). SE=standard error, Env=environment type, Lin=lineage type, Lab=Principal Investigator ID, Sample=EMP Sample ID. Interactions are denoted as '*'. n.s.=not significant (likelihood-ratio test). All models provide a significantly better fit than null models without fixed effects (ΔAIC > 10 and P < 0.05; Supplementary Data file 2).*

	Slope (fixed effects)				**Standard deviation on the slope (random effects)**				
	Div	**SE**	**z**	***P***	**Env**	**Lin**	**Lin*Env**	**Env*Lab**	**Sample**
ASV:Genus	0.091	0.016	5.792	6.95e-09	n.s.	0.074	0.142	0.114	0.067
Genus:Family	0.047	0.008	5.911	3.41e-09	n.s.	0.071	0.07	0.039	n.s.
Family:Order	0.119	0.017	7.001	2.54e-12	0.023	0.094	0.092	0.106	n.s.
Order:Class	0.109	0.020	5.447	5.13e-08	0.05	0.141	0.078	0.051	n.s.
Class:Phylum	0.272	0.043	6.341	2.29e-10	0.119	0.174	0.119	0.114	n.s.

Table 2. GLMMs applied to data simulated under null models. *Null models 1 and 2 were generated under the ZSM distribution, with a single distribution for the whole dataset (Model 1) or one distribution per environment (Model 2). Model 3 is similar to Model 1, except with a single Poisson distribution for the whole dataset, and +DBD or +EC refer to adding these effects to 100% of ASVs (see* ***Methods*** *and* ***Figure 2 supplement 7****). Each row reports the effect of community diversity on focal lineage diversity (Div), as well as its standard error, Wald z-statistic for its effect size and the corresponding P-value (Wald test) (left section), or standard deviation on the slope for the significant random effects (right section). SE=standard error, Env=environment type, Lin=lineage type, Sample=EMP Sample ID. n.s.=not significant (likelihood-ratio test), n.t.= not tested, because separate environments were not included in Models 1 or 3.*

	Slope (fixed effects)				**Stand dev on the slope (random effects)**			
	Div	**SE**	**z**	***P***	**Env**	**Lin**	**Lin*Env**	**Sample**
Model 1	**-0.005**	0.000	-9.807	**<2e -16**	n.t.	0.639	n.t.	n.s.
Model 2	n.s.							
Model 3	**-0.012**	0.002	-6.552	**5.69e-11**	n.t.	0.021	n.t.	n.s.
Model3 + DBD	**0.016**	0.001	11.48	**<2e-16**	n.t.	0.008	n.t.	n.s.
Model3 + EC	**-0.011**	0.002	-6.14	**8.26e-10**	n.t.	ns	n.t.	n.s.

Table 3. GLMMs with community diversity measured using Shannon diversity. *Results are shown from GLMMs with Shannon diversity of non-focal taxa (Div) as a predictor of ASVs richness of focal taxa. Each row reports the estimate (Div), as well as its standard error, Wald z-statistic for its effect size and the corresponding P-value (Wald test) (left section), or standard deviation on the slope for the significant random effects (right section). SE=standard error, Env=environment type, Lin=lineage type, Lab=Principal Investigator ID, Sample=EMP Sample ID. n.s.=not significant (likelihood-ratio test).*

	Fixed effects				**Random effects**				
	Div	**SE**	**z**	***P***	**Env**	**Lin**	**Env*Lin**	**Env*Lab**	**Sample**
Genus	0.055	0.013	4.33	1.49e-05	n.s.	0.08	0.15	0.085	0.054
Family	0.148	0227	6.491	8.51e-11	n.s.	0.184	0.268	0.16	0.134
Order	0.378	0.038	9.864	<2e-16	n.s.	0.34	0.417	0.258	0.202
Class	0.398	0.05	7.973	1.54e-15	n.s.	0.369	0.46	0.326	0.262
Phylum	0.319	0.088	3.614	0.0003	0.169	0.316	0.5	0.495	0.378

Table 4. Community diversity has a stronger effect than abiotic factors on focal lineage diversity (EMP dataset). *Results are shown from GLMMs with community diversity, four abiotic factors (temperature, elevation, pH, and latitude), and their interactions with community diversity, as predictors of focal lineage diversity. Random effects on the intercept included environment, lineage, lab ID and sample ID. Each row reports the taxonomic ratio, the predictors used in the GLMM (fixed effects only), their estimate (Est), standard error (SE) and P-value (P) (Wald test). Interactions are denoted as '*'. Random effects are not shown.*

	Predictor	**Est**	**SE**	**P**
ASV:Genus	**Div**	0.128	0.013	< 2e-16
	Temperature	0.04	0.014	0.00479
	Div*Temperature	0.043	0.014	0.00175
	Div*Latitude	0.031	0.013	0.02119
	Div*Elevation	-0.031	0.014	0.02829
Genus:Family	**Div**	0.094	0.009	< 2e-16
	Temperature	0.026	0.009	0.00268
	pH	-0.042	0.009	5.88e-06
Family:Order	**Div**	0.131	0.01	< 2e-16
Order:Class	**Div**	0.184	0.01	< 2e-16
	Div*Temperature	0.032	0.009	0.000827
	Div*Latitude	0.023	0.008	0.005403
Class:Phylum	**Div**	0.236	0.011	< 2e-16
	Div*Temperature	0.059	0.014	2.15e-05
	Div*Latitude	0.03	0.011	0.00884

Table 5. GLMMs applied to a soil dataset. *Each row reports the taxonomic ratio, the predictors used in the GLMM (fixed effects only), their estimate (Est), standard error (SE) and P-value (P) (Wald test). Left columns: GLMM with community diversity (Div) and all abiotic variables considered separately, as predictors of focal lineage diversity. Right columns: GLMM with community diversity (Div) and the three first principle components (PCs) representing abiotic variables, as predictors of focal lineage diversity. n.s., non-significant (LRT test). All models provide a significantly better fit than null models without fixed effects (ΔAIC > 10 and P < 0.05;* ***Supplementary Data file 2****), except for the GLMM with abiotic factors at the Family:Order level, where latitude has a significant effect on focal lineage diversity but its effect is nearly null, with a ΔAIC between full and null model of 4 and a null marginal* R^2*.*

	GLMMs with abiotic variables				**GLMMs with the 3 first PCs**			
	Predictor	**Est**	**SE**	**P**	**Predictor**	**Est**	**SE**	**P**
ASV:Genus	**Div**	n.s.			**Div**	0.064	0.016	9.47e-05
	Latitude	0.294	0.025	< 2e-16	**PC1**	-0.065	0.007	< 2e-16
	UV_light	-0.177	0.016	< 2e-16	**PC2**	-0.03	0.006	1.98e-05
	MDR	0.028	0.006	7.12e-06				
	NPP2003_201	-0.066	0.005	< 2e-16				
	Latitude^2	-0.3	0.029	< 2e-16				
	Clay_silt^2	-0.012	0.004	0.003				
	Soil_N^2	-0.007	0.001	1.66e-06				
	Soil_C_N_rati	0.003	0.001	0.004				
	PSEA^2	0.01	0.002	4.84e-06				
	MDR^2	0.017	0.003	2.40e-08				
	NPP2003_201	-0.016	0.004	0.0001				
Genus:Family	**Div**	0.032	0.01	0.0011	**Div**	0.033	0.01	0.001
	Latitude	-0.035	0.006	2.04e-09	**PC1**	-0.016	0.006	0.02
					PC2	0.02	0.006	0.00089
Family:Order	**Div**	n.s.			**Div**	n.s.		
	Latitude	-0.0005	0.0002	0.0105	**PC1**	-0.026	0.007	0.00032
					Div*PC1	0.04	0.006	2.14e-12
					Div*PC3	0.023	0.005	1.68e-06
Order:Class	Null model with no predictor was significant							
Class:Phylum	**Div**	0.032	0.01	0.00174	**Div**	0.032	0.01	0.003
	pH	0.074	0.01	4.37e-13	**PC1**	-0.051	0.01	3.54e-07
					PC2	-0.028	0.01	0.006

Supplementary Figures

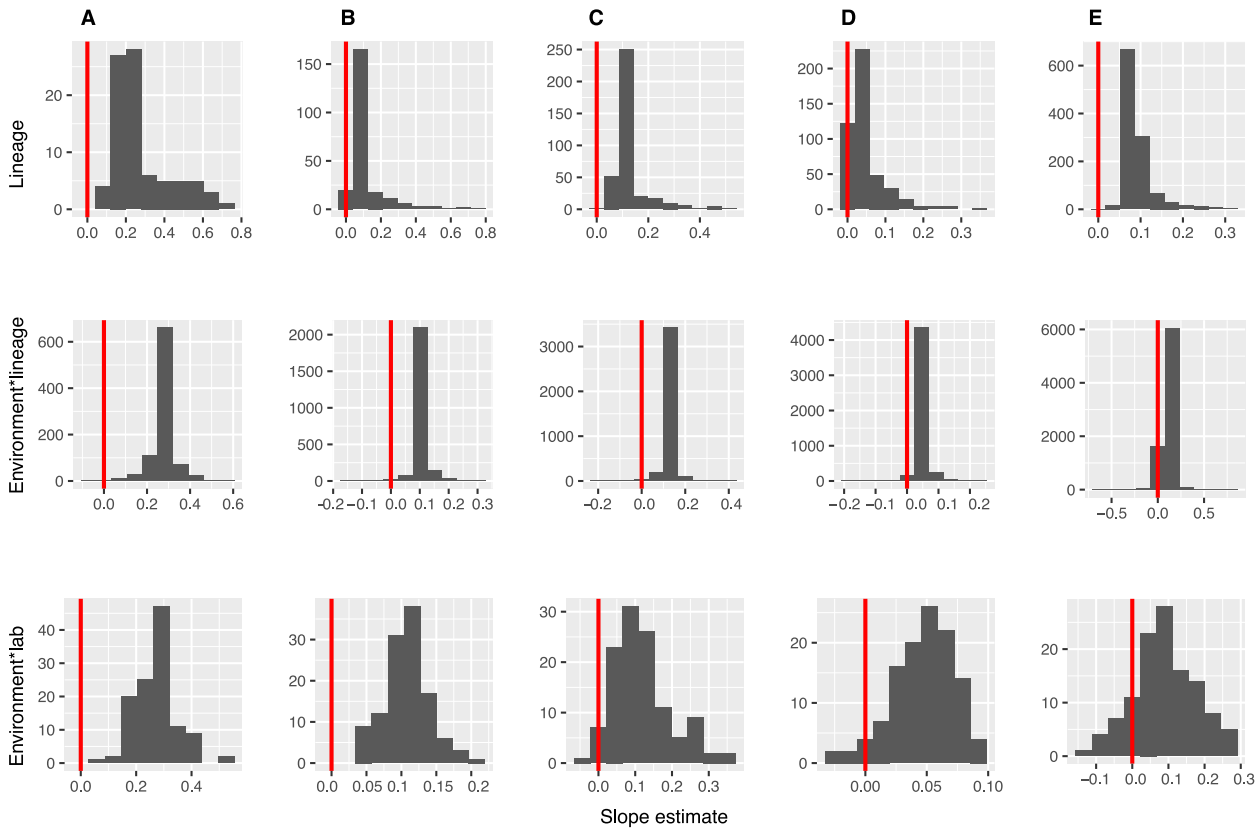

Figure S1. Distributions of diversity slope estimates across different random effects, *from the GLMMs predicting focal lineage diversity as a function of community diversity.* (**A**) Class:Phylum, (**B**) Order:Class, (**C**) Family:Order, (**D**) Genus:Family, and (**E**) ASV:Genus. Estimation of random effect coefficients from the GLMMs (Table S1), shows that the effect of diversity on focal lineage diversity (slope estimates) are generally positive but could be negative in some lineages or combinations of environment, lineage (Environment*Lineage), and the laboratory that submitted the dataset (Environment*Lab).

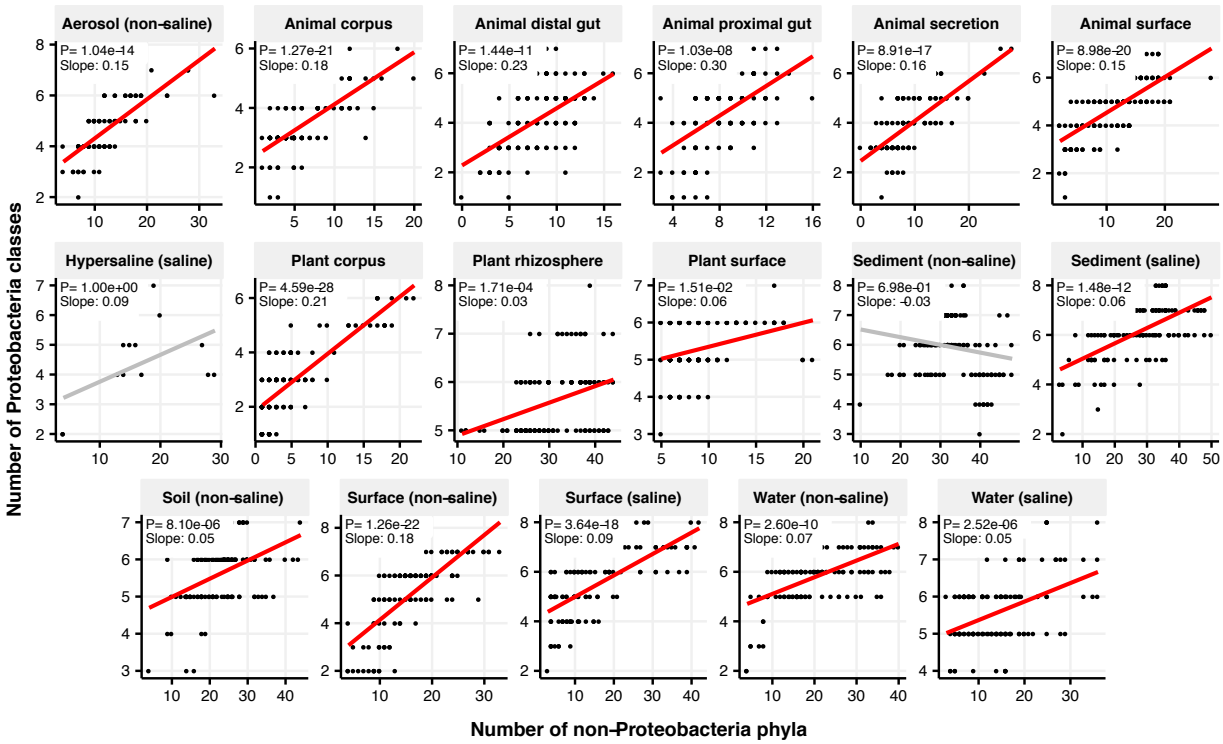

Figure S2. Focal-lineage diversity as a function of community diversity across biomes in Proteobacteria. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

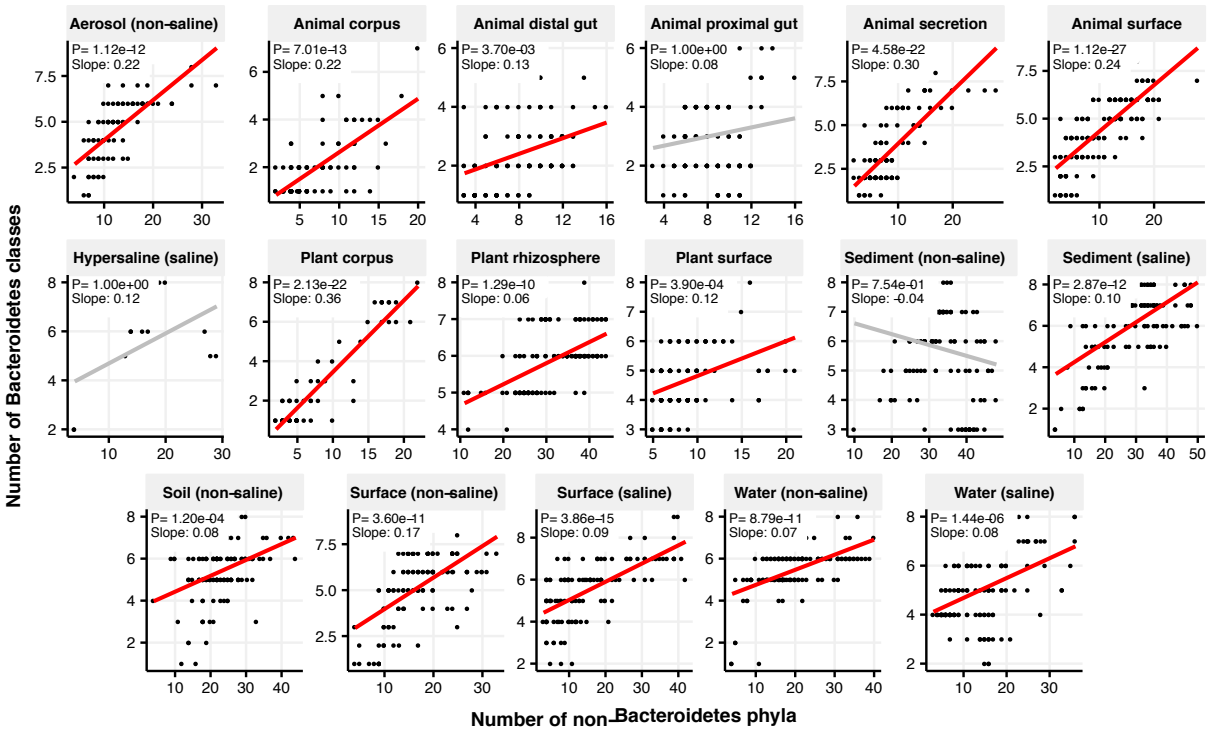

Figure S3. Focal-lineage diversity as a function of community diversity across biomes in Bacteroidetes. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

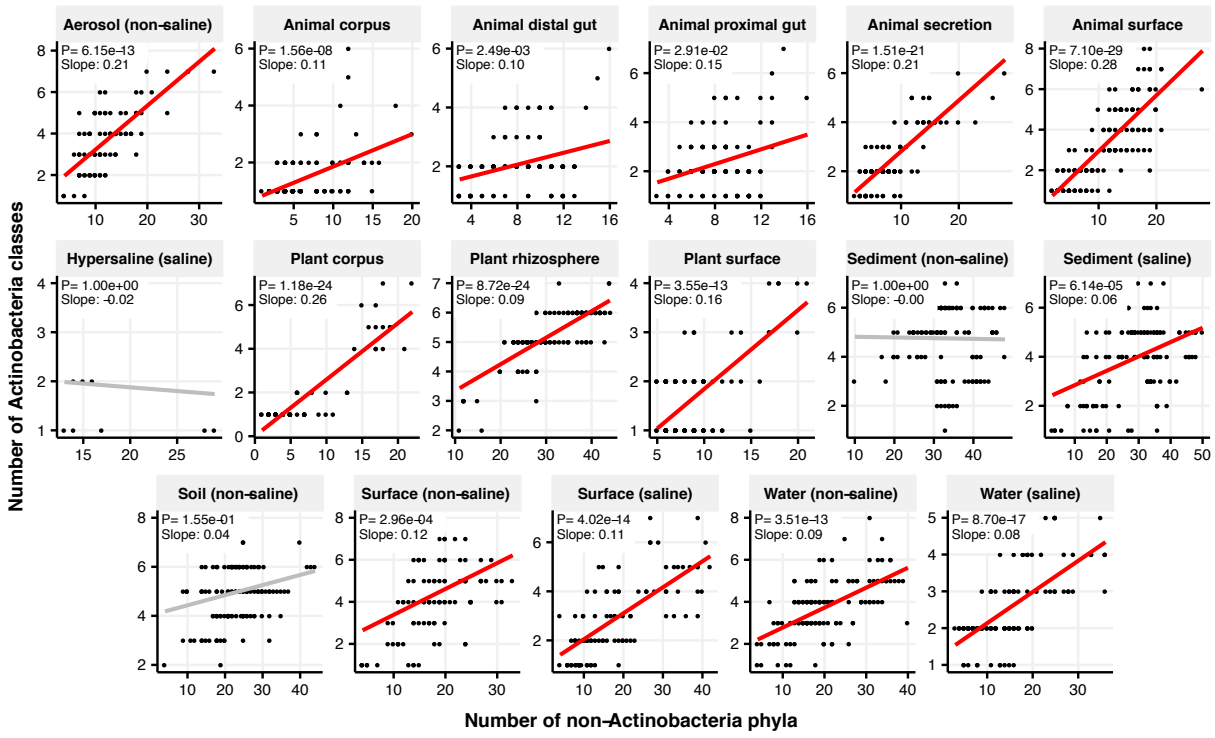

Figure S4. Focal-lineage diversity as a function of community diversity across biomes in Actinobacteria. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

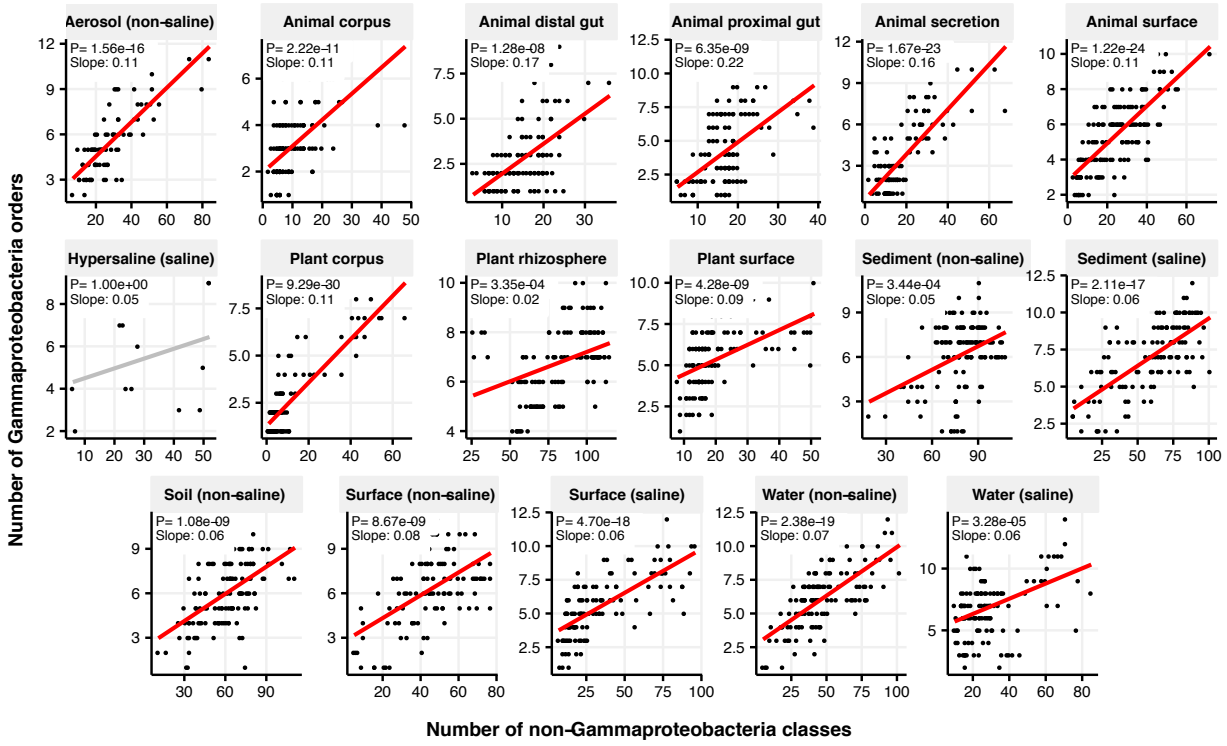

Figure S5. Focal-lineage diversity as a function of community diversity across biomes in Gammaproteobacteria.*Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

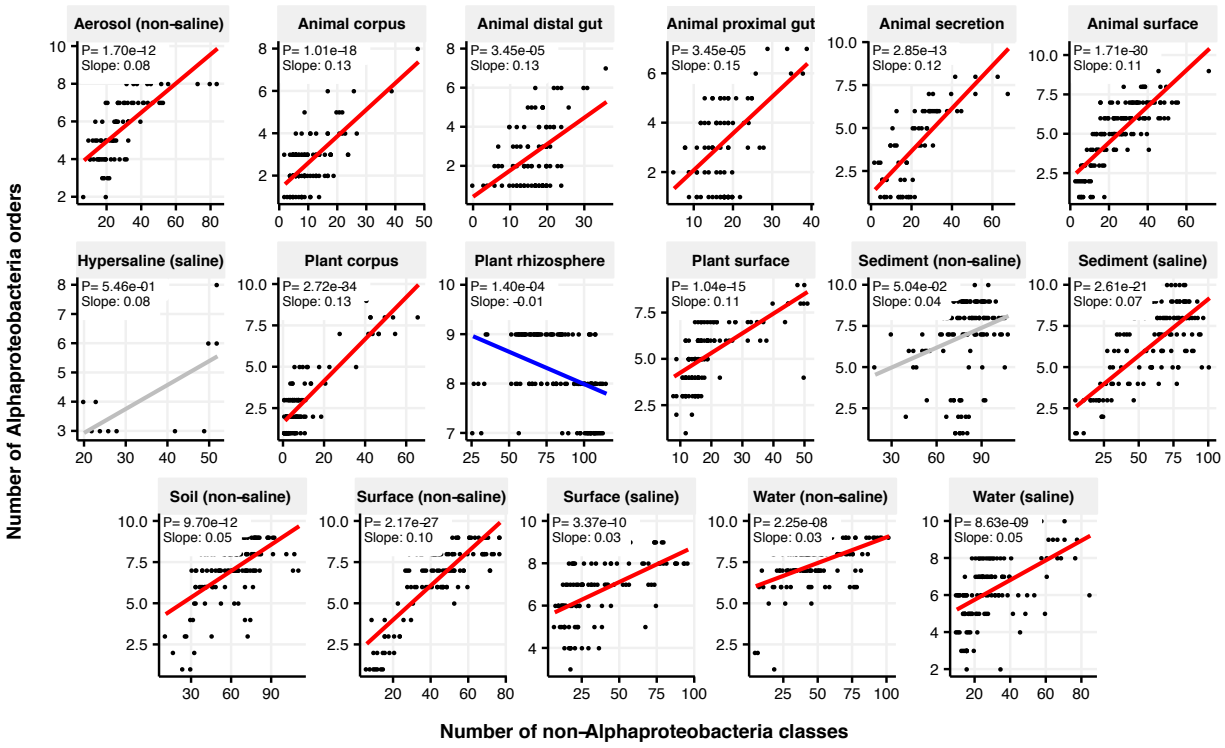

Figure S6. Focal-lineage diversity as a function of community diversity across biomes in Alphaproteobacteria. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

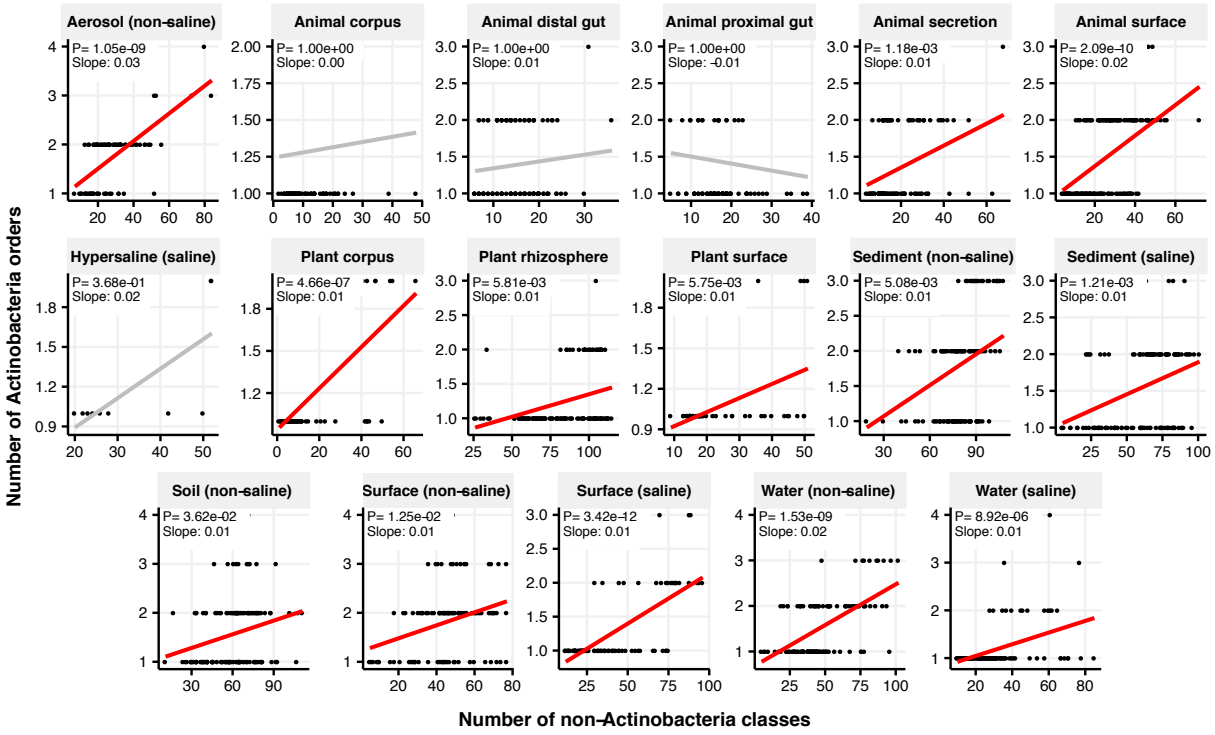

Figure S7. Focal-lineage diversity as a function of community diversity across biomes in Actinobacteria. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

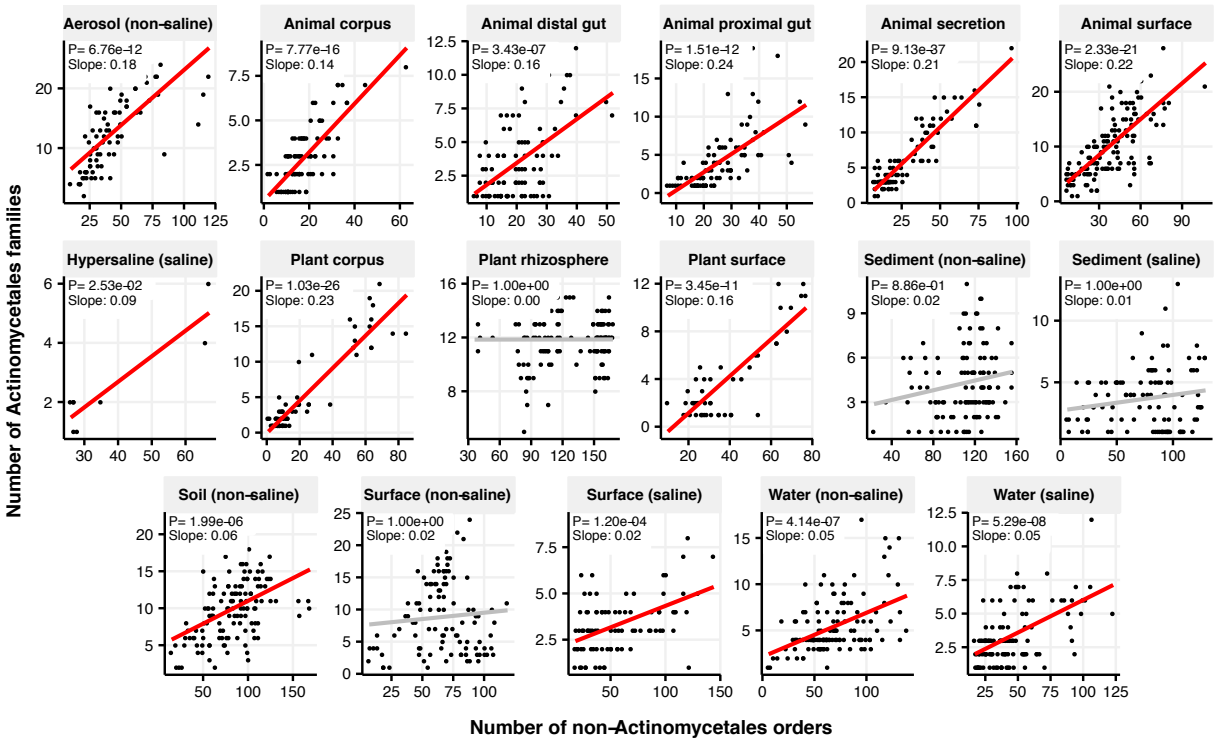

Figure S8. Focal-lineage diversity as a function of community diversity across biomes in Actinomycetales. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

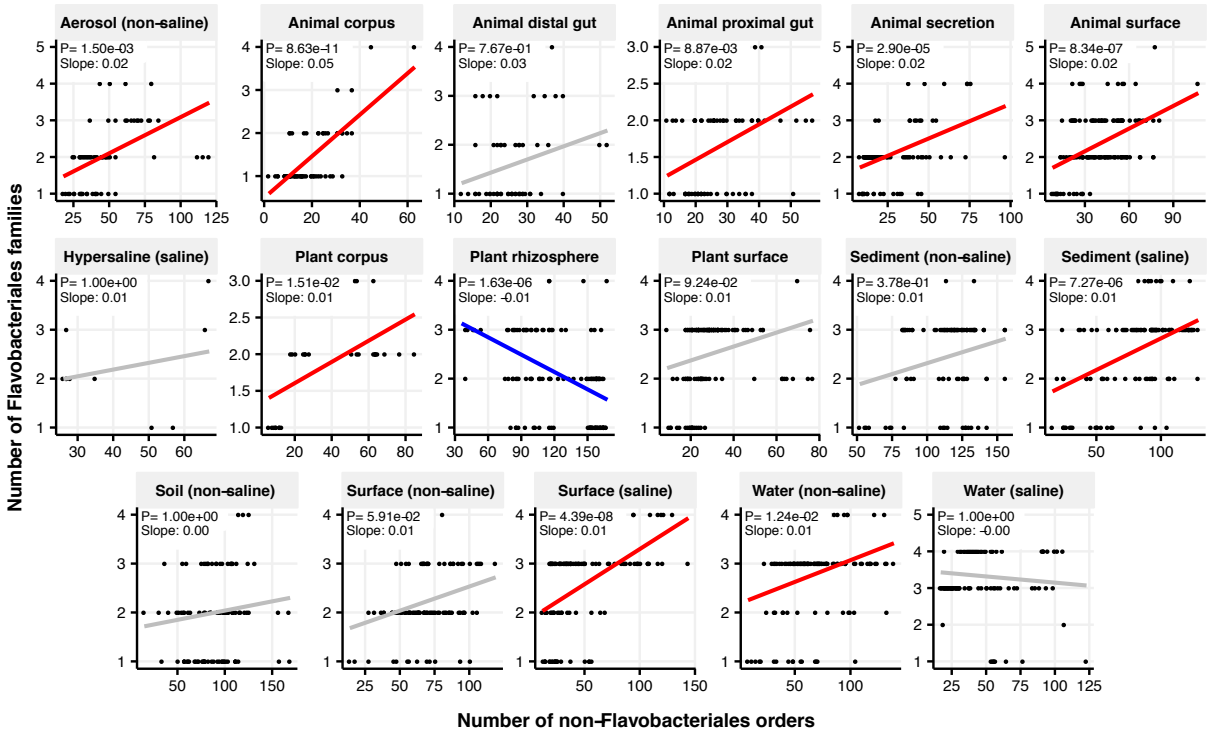

Figure S9. Focal-lineage diversity as a function of community diversity across biomes in Flavobacteriales. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

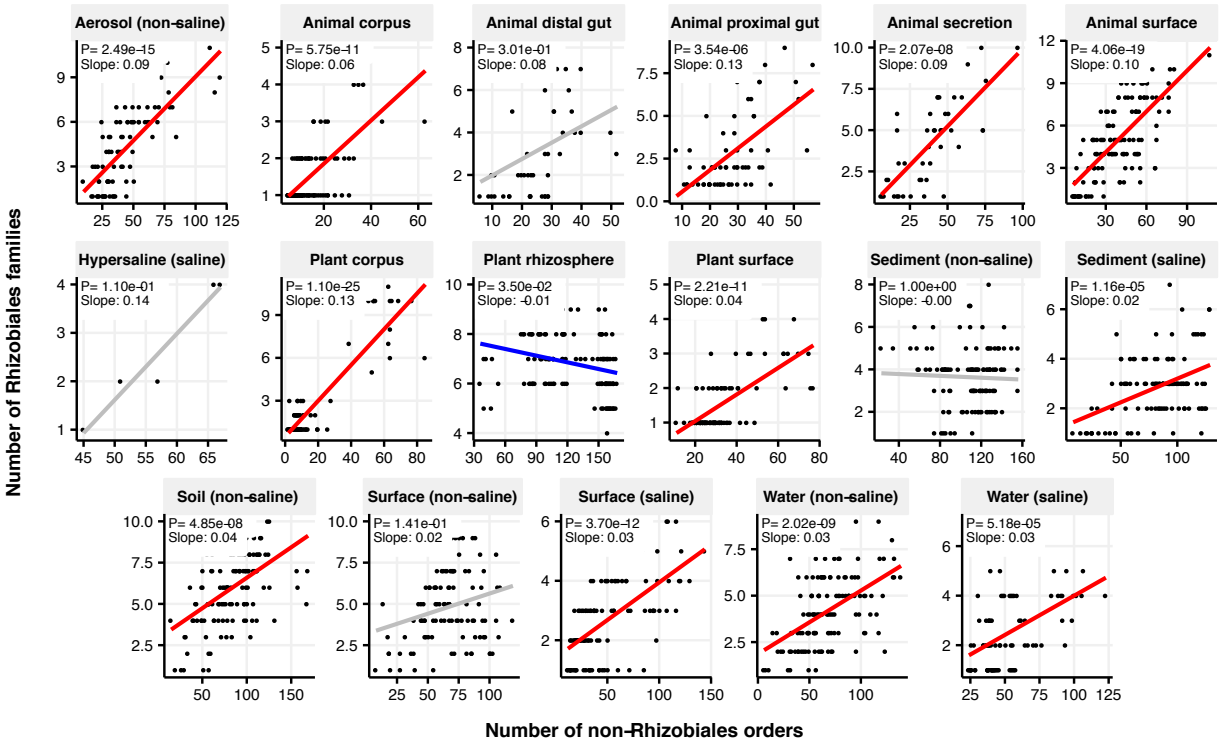

Figure S10. Focal-lineage diversity as a function of community diversity across biomes in Rhizobiales. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

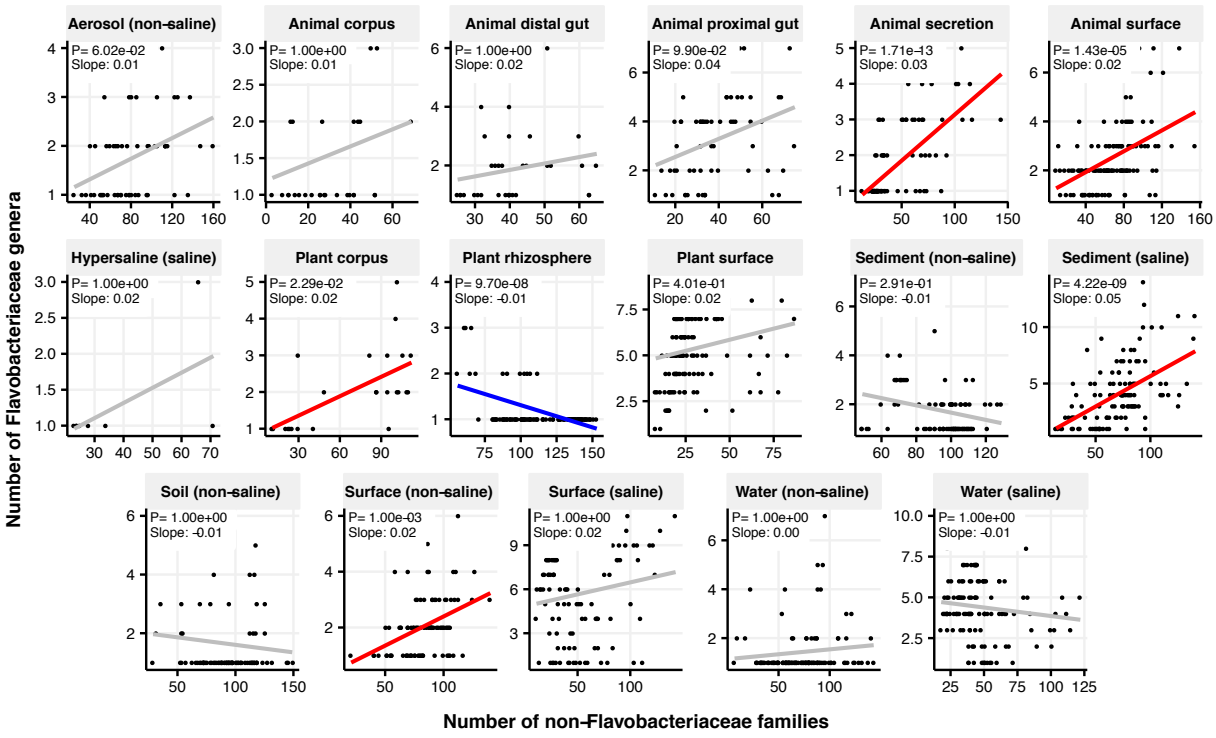

Figure S11. Focal-lineage diversity as a function of community diversity across biomes in Flavobacteriaceae.*Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

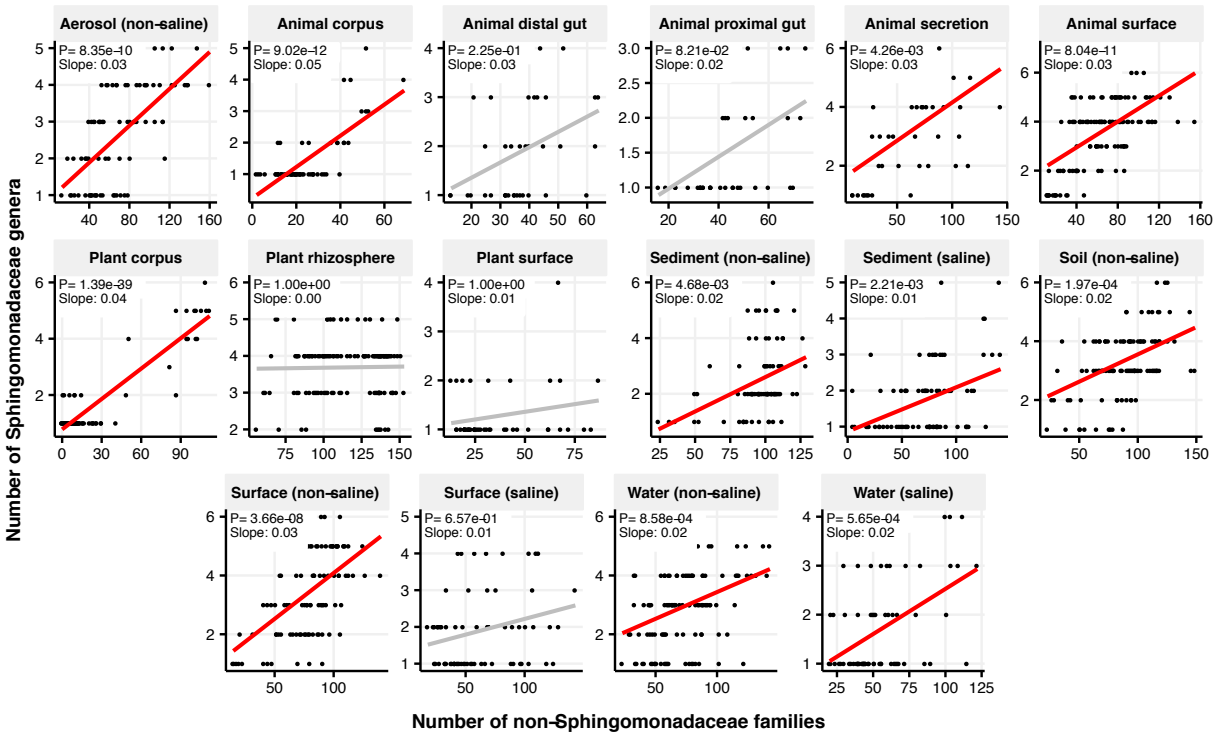

Figure S12. Focal-lineage diversity as a function of community diversity across biomes in Sphingomonadaceae.*Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

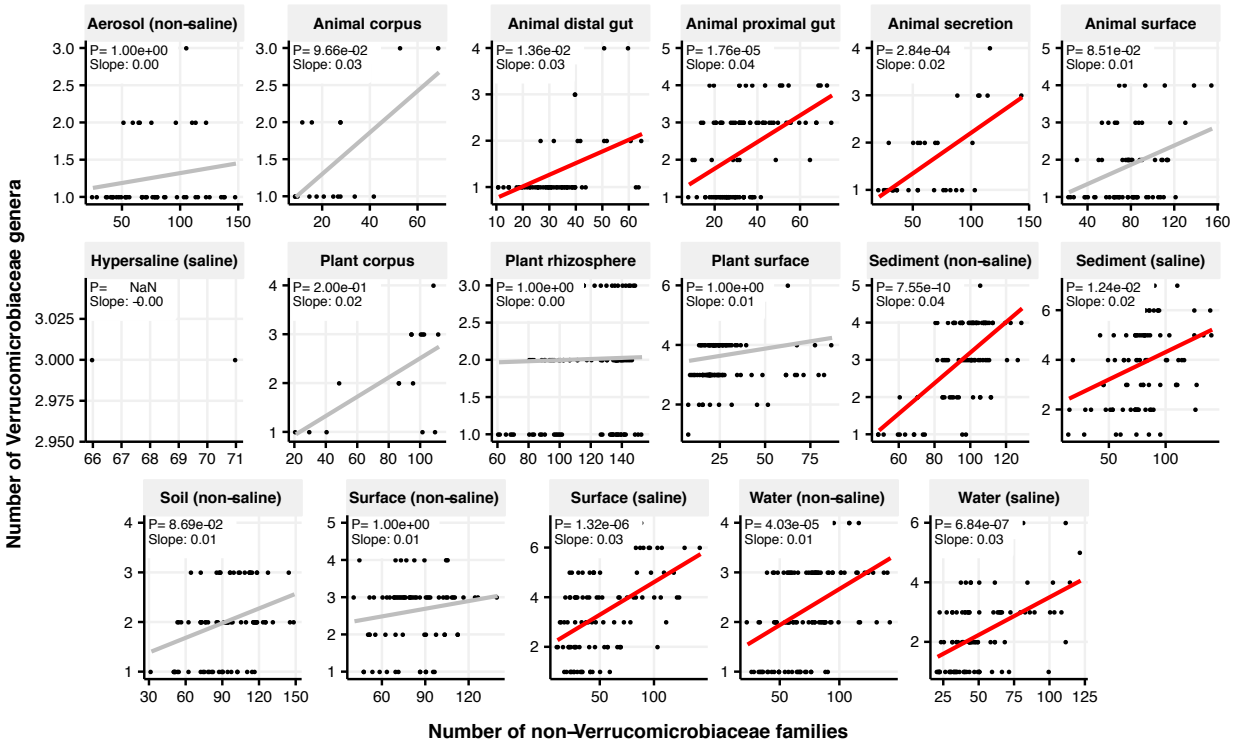

Figure S13. Focal-lineage diversity as a function of community diversity across biomes in Verrucomicrobiaceae.*Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

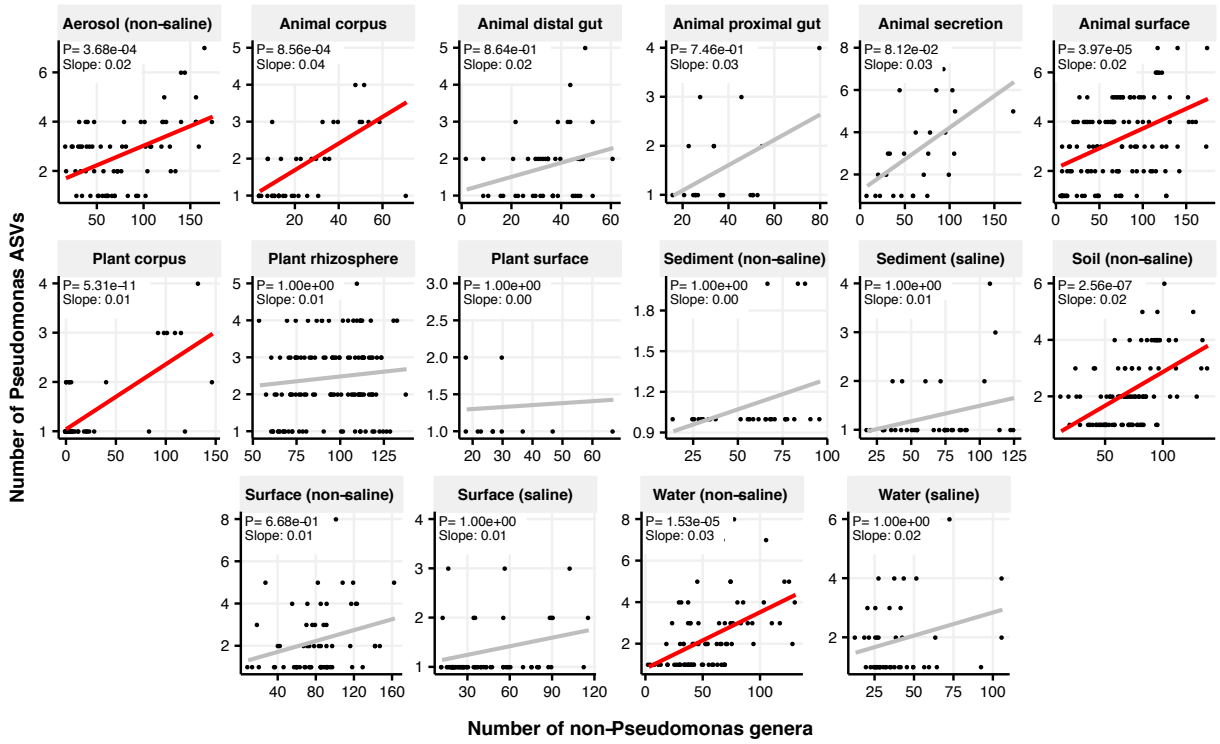

Figure S14. Focal-lineage diversity as a function of community diversity across biomes in Pseudomonas. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

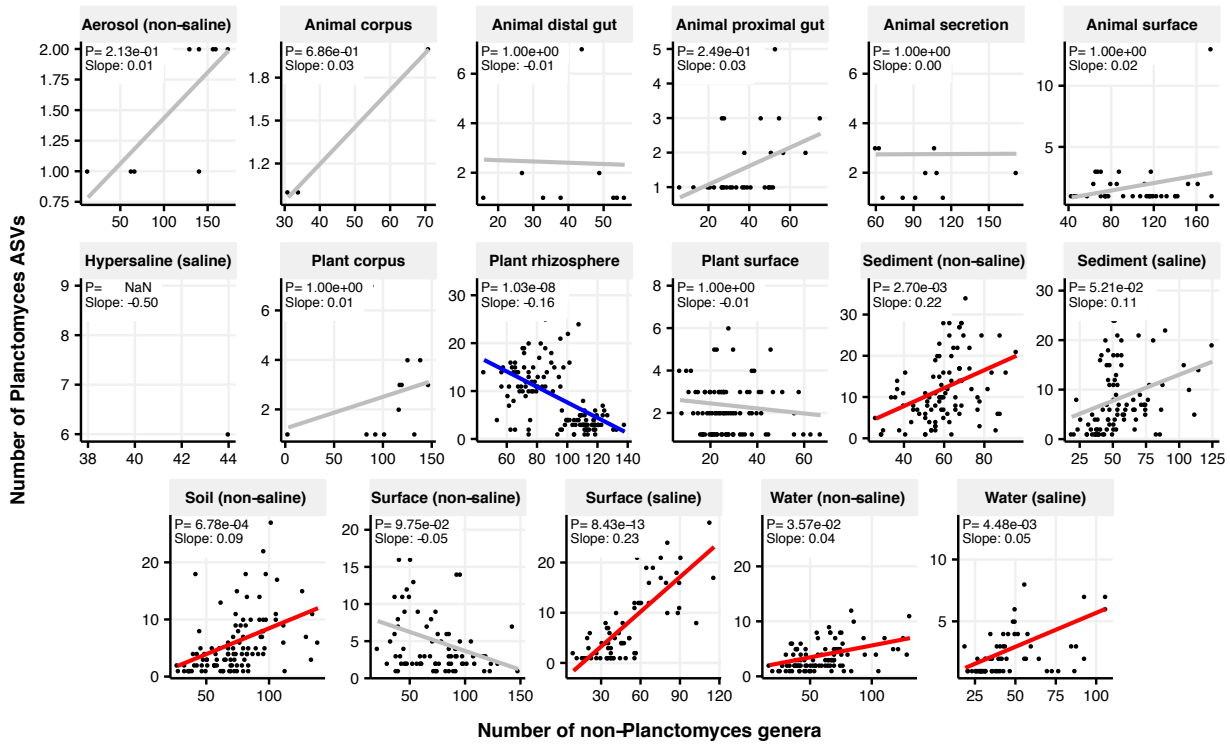

Figure S15. Focal-lineage diversity as a function of community diversity across biomes in Planctomyces. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

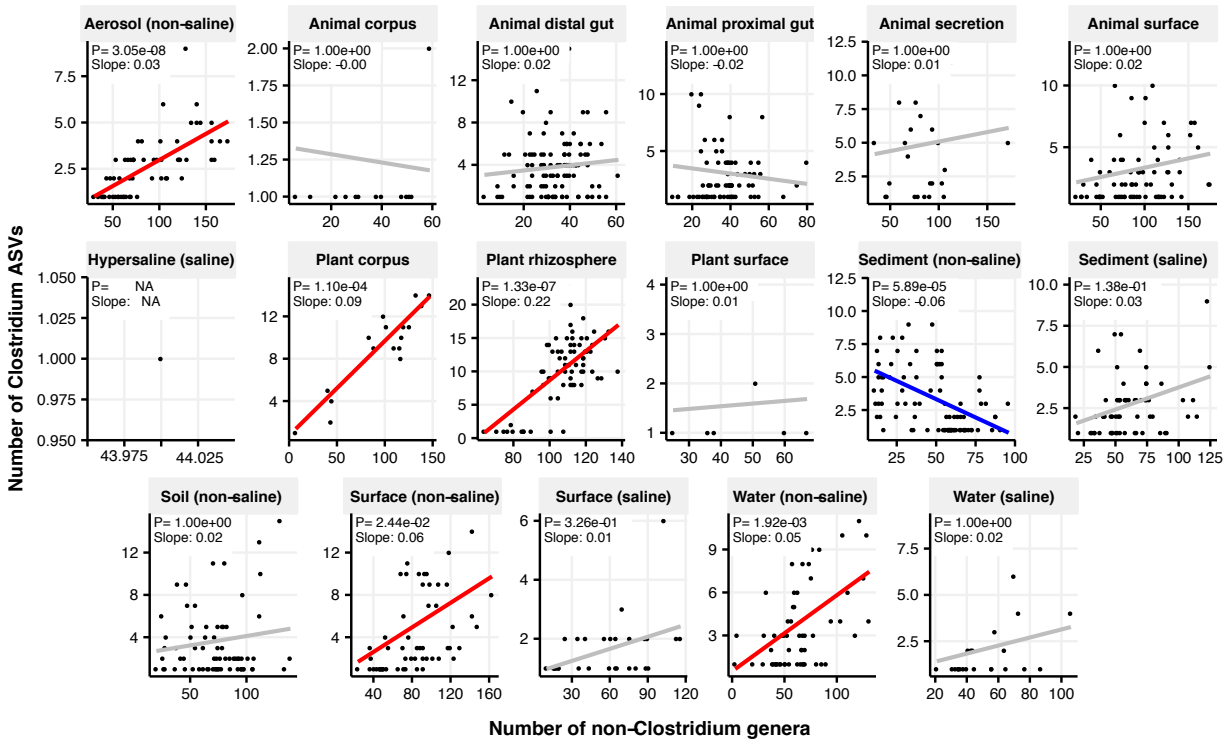

Figure S16. Focal-lineage diversity as a function of community diversity across biomes in Clostridium. *Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, p<0.05, Bonferroni corrected for 17 tests), and non-significant in grey.*

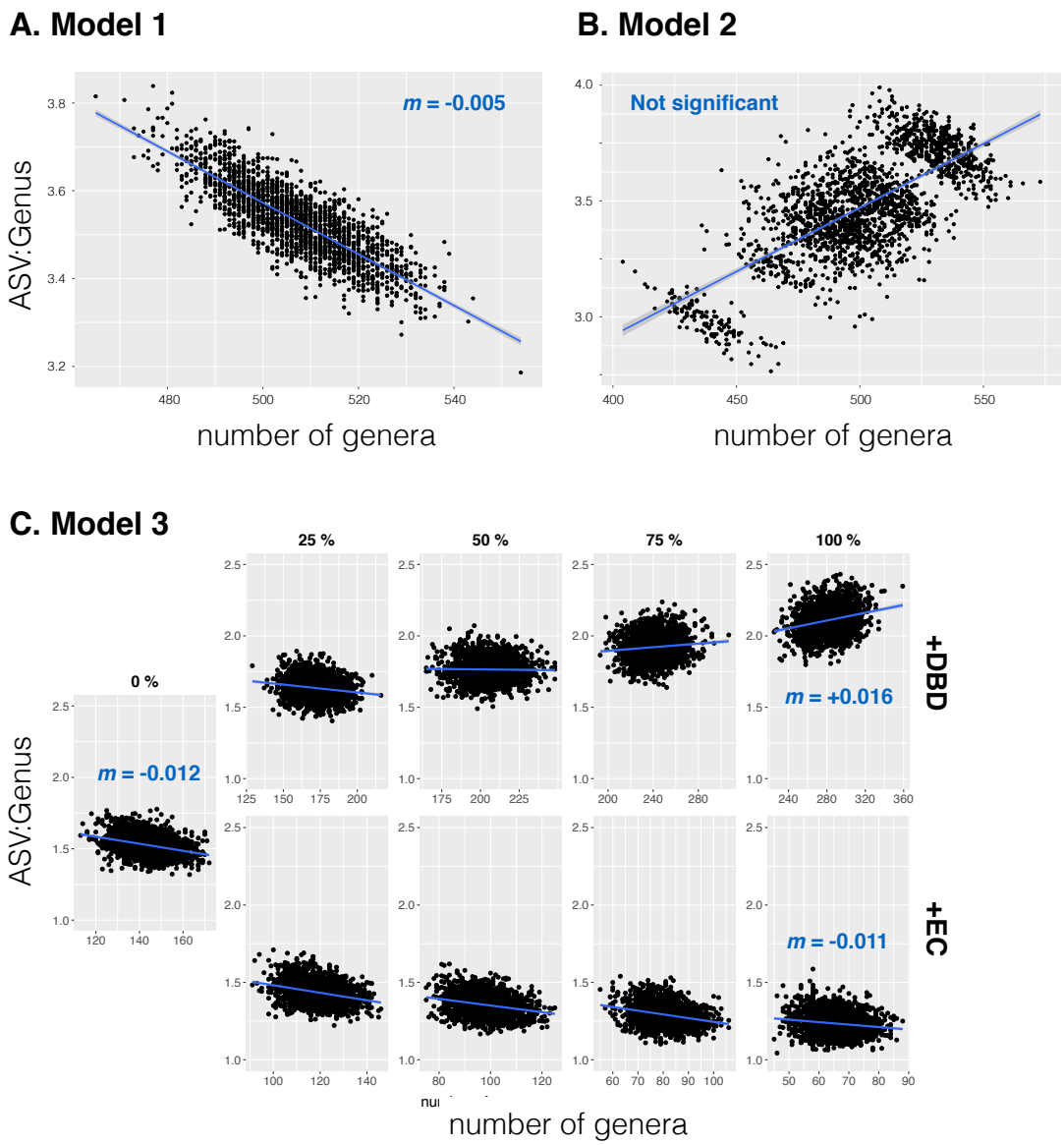

Figure S17. Null models based on Neutral Theory. *Results are shown from data simulated under (**A**) neutral Model 1, (**B**) neutral Model 2, or (**C**) neutral Model 3. Model 1 is sampled from the zero-sum multinomial distribution with a single distribution for the whole dataset, while Model 2 includes a separate distribution for each of the 17 different environments (EMPO 3 biomes). In Model 3 (C), the effect of DBD (top rows) or EC (bottom rows) are 'spiked in' at different levels, ranging from 0 to 100% of ASVs in a sample. Blue lines show a linear fit, with slopes (m) estimated by GLMM in selected panels. See Methods for model details, and Table 2 and Supplementary file 3, Section 1.2 for full GLMM results.*

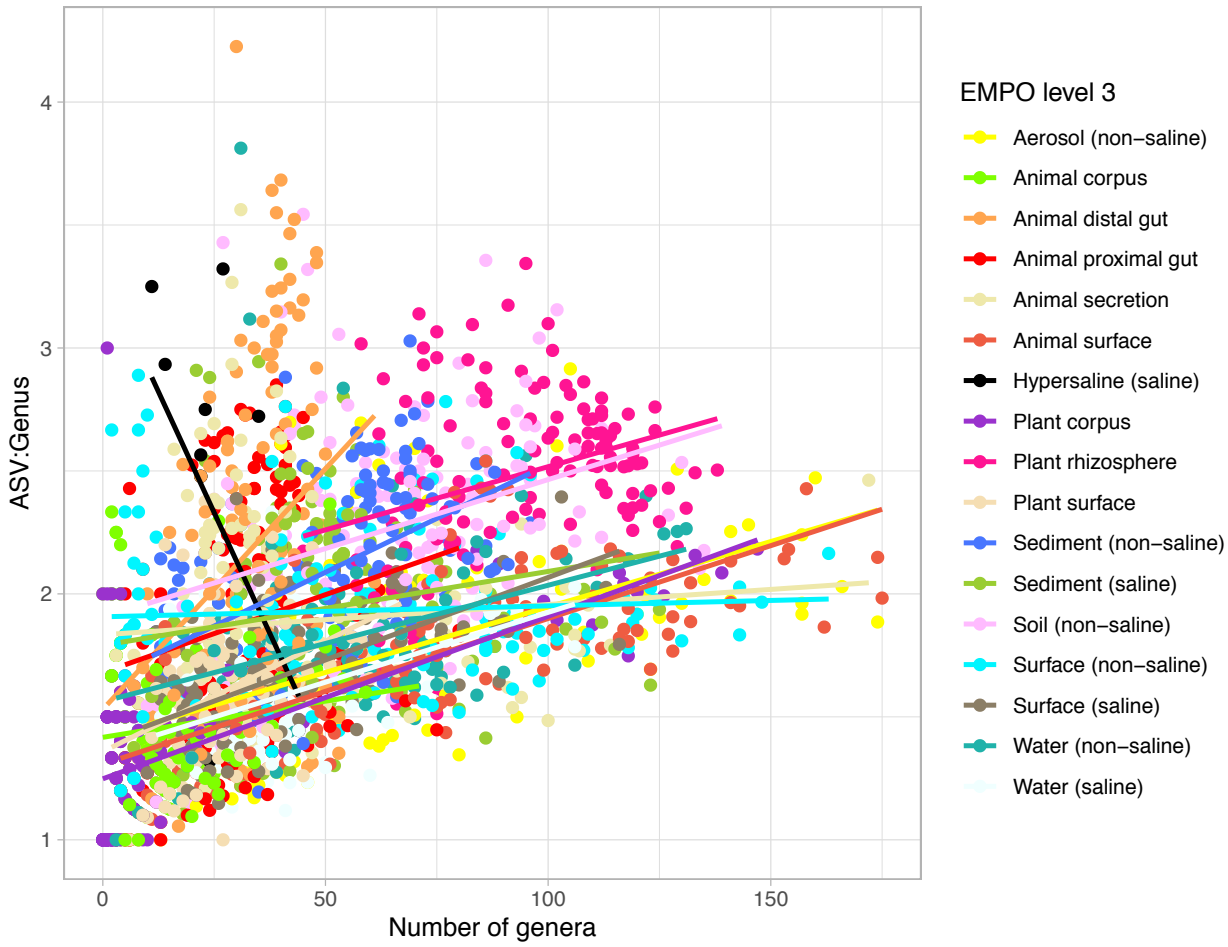

Figure S18. Lineage diversity (mean ASV:Genus ratio among all lineages) as a function of community diversity (number of genera) in the EMP data. *Samples from different environments (EMPO level 3) are shown in different colours, each with their corresponding linear model fit.*

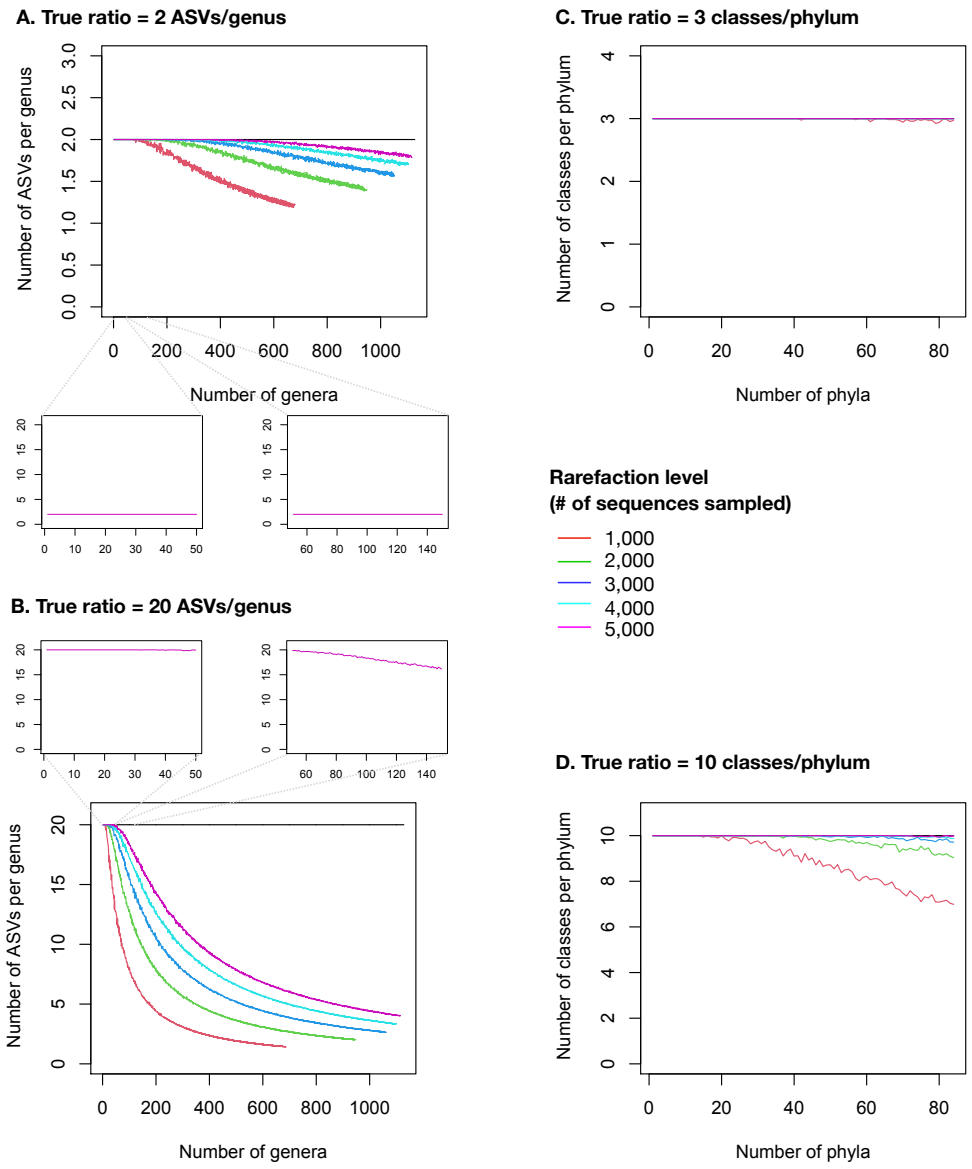

Figure S19. Taxonomic ratios estimated from simulated rarefied sequence data. *Each panel simulates a set of microbiome samples that differ in their diversity (number of genera in left panels A and B, number of phyla in right panels C and D) while maintaining a set true taxonomic ratio (horizontal black line). (**A**) True ratio set to 2 ASVs/genus, close to the per-sample mean and median in the real EMP data, in a range of samples between 1 and 1128 named genera, as observed in the real EMP data. (**B**) True ratio set to 20 ASVs/genus, equal to the overall mean of 22,014 named ASVs in 1128 named genera, and close to the maximum ratios observed in individual samples (Figure 2—figure supplement 5). Insets show the ranges of 1–50 and 51–150 genera, approximating observations from lower- or higher-diversity samples such as gut and soil, respectively (Figure 2—figure supplement 5). The insets only show the rarefaction to 5000 sequences, as used in the real EMP dataset. (**C**) True ratio set to three classes/phylum, close to the per-sample mean and median in the real EMP data, in a range of samples between 1 and 84 named phyla, as observed in the real EMP data. (**D**) True ratio set to 10 classes/phylum, close to the maximum ratios observed in individual samples (Figure 2—figure supplements 2–4). Different rarefaction levels are shown as different coloured lines.*

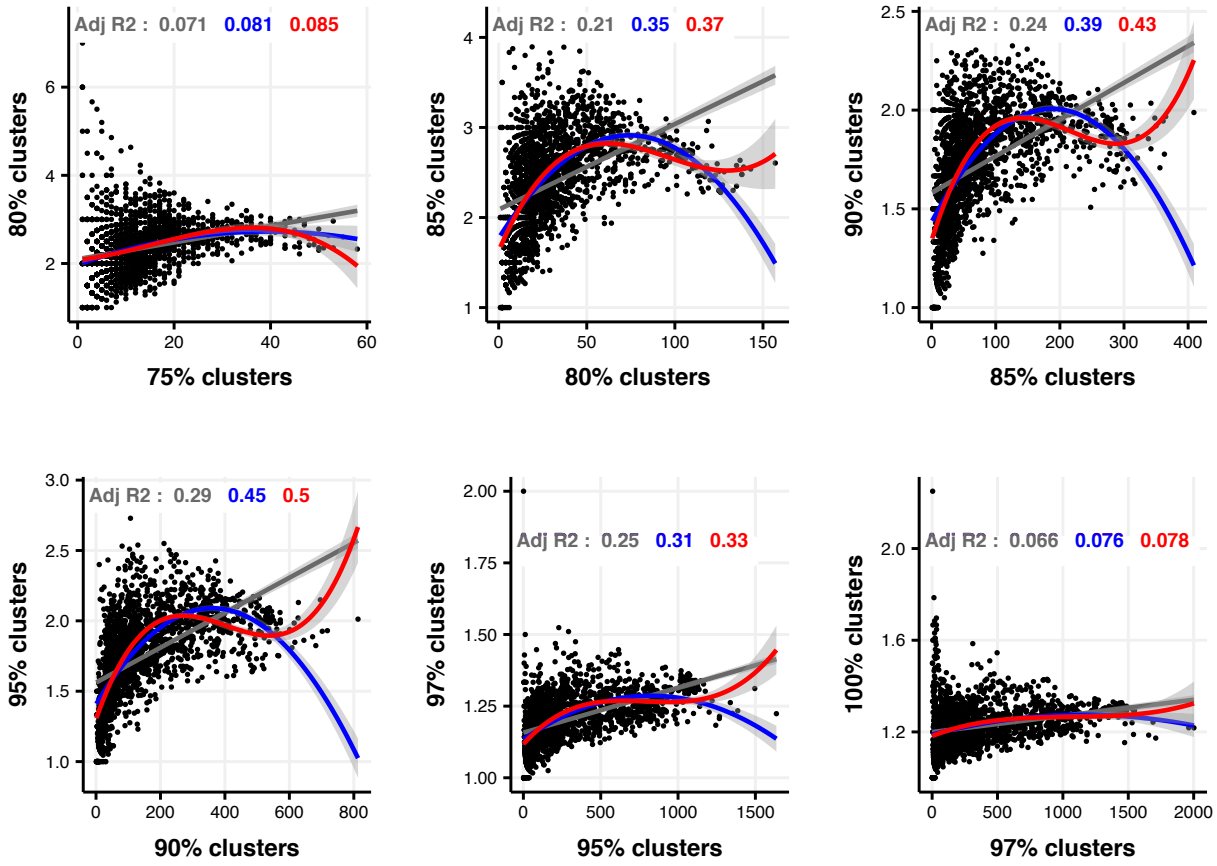

Figure S20. Linear, quadratic, and cubic models for the relationship between focal-lineage diversity and community diversity for varying levels of % nucleotide identity.*Community diversity was estimated as the number of clusters at a focal level (d_i) and focal-lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). All P-values are <0.001. Linear fit (grey); quadratic fit (blue), cubic fit (red); same colours for the associated adjusted R^2. The x-axis (diversity) shows the number of clusters at the focal percent-identity level (d_i), and the y-axis (diversification) is the mean of the clusters at the rank above (d_{i+1}/d_i).*

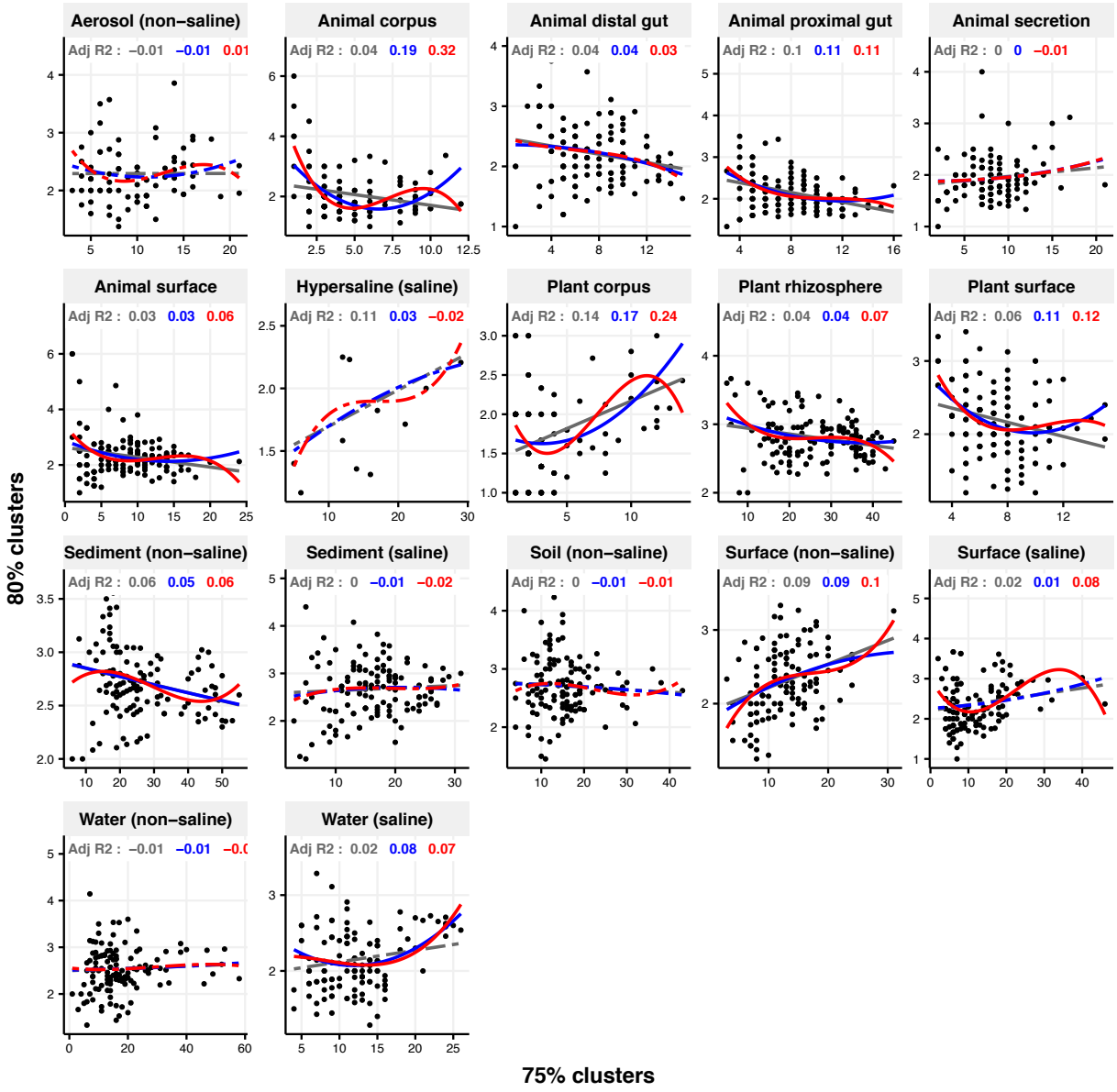

Figure S21. Focal clusters at 80% nucleotide identity.*Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. P-values are Bonferroni corrected for 17 tests. Significant, p<0.05 (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).*

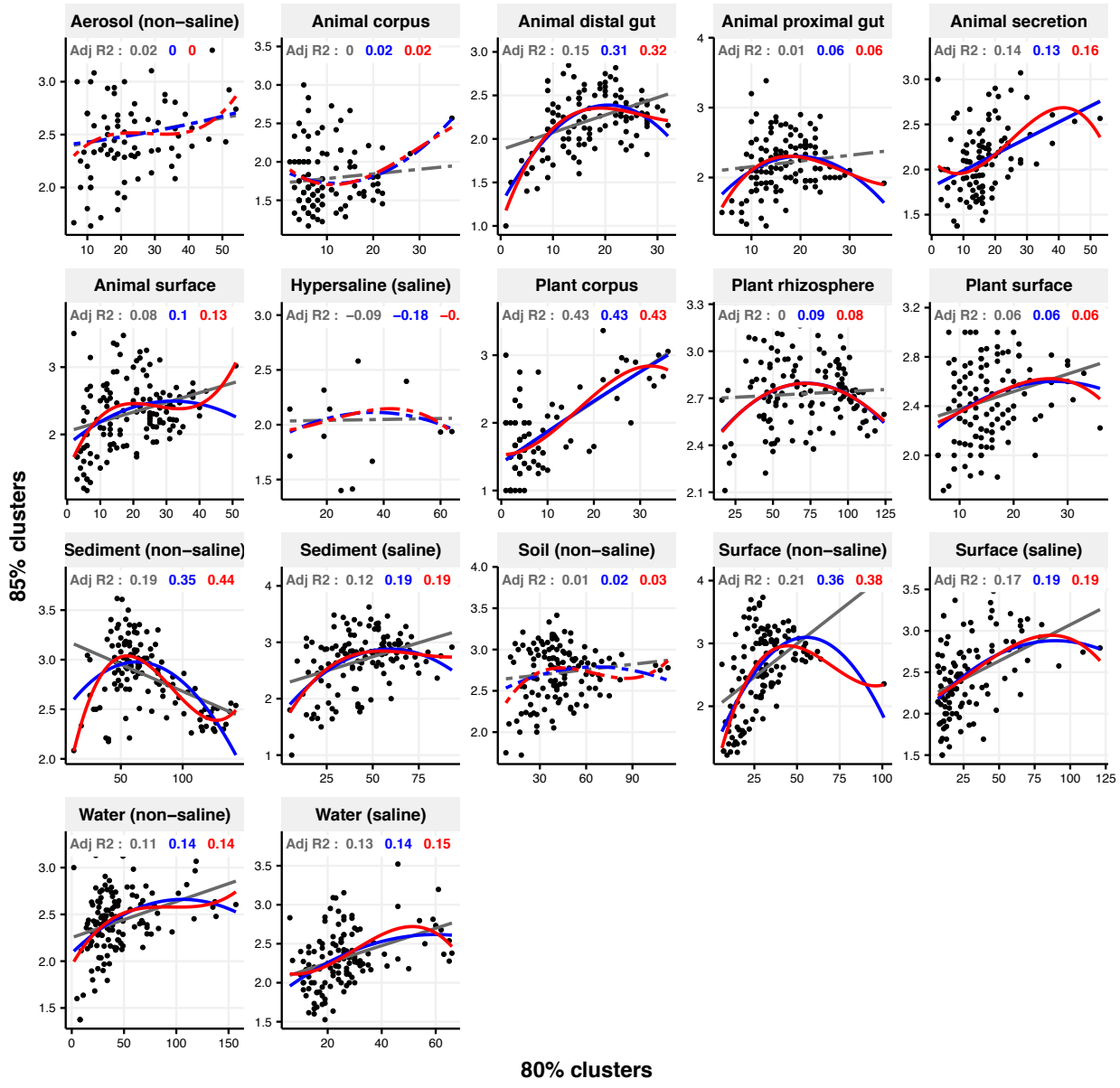

Figure S22. Focal clusters at 85% nucleotide identity.*Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. P-values are Bonferroni corrected for 17 tests. Significant, p<0.05 (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).*

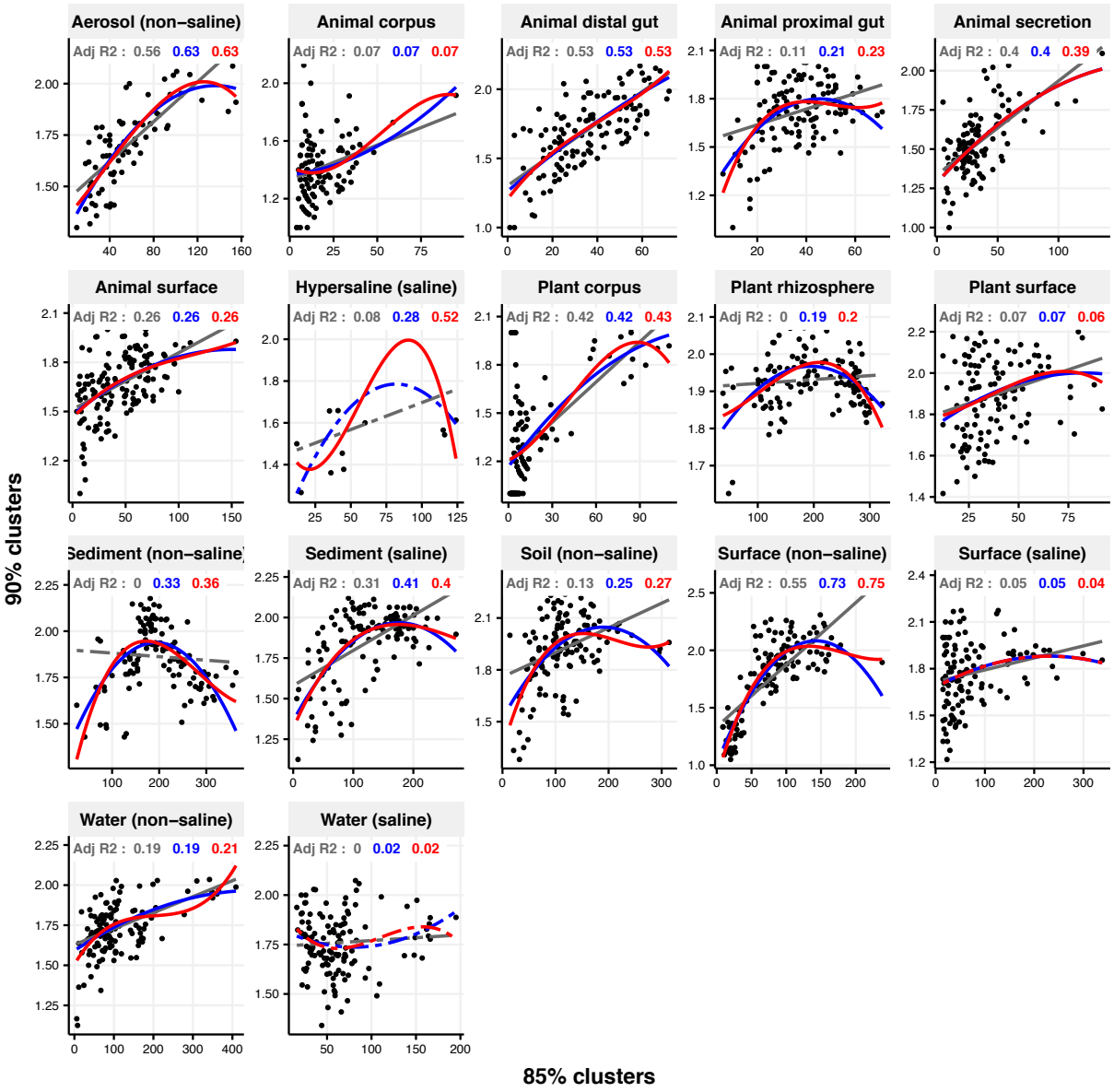

Figure S23. Focal clusters at 90% nucleotide identity. *Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. P-values are Bonferroni corrected for 17 tests. Significant, p<0.05 (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).*

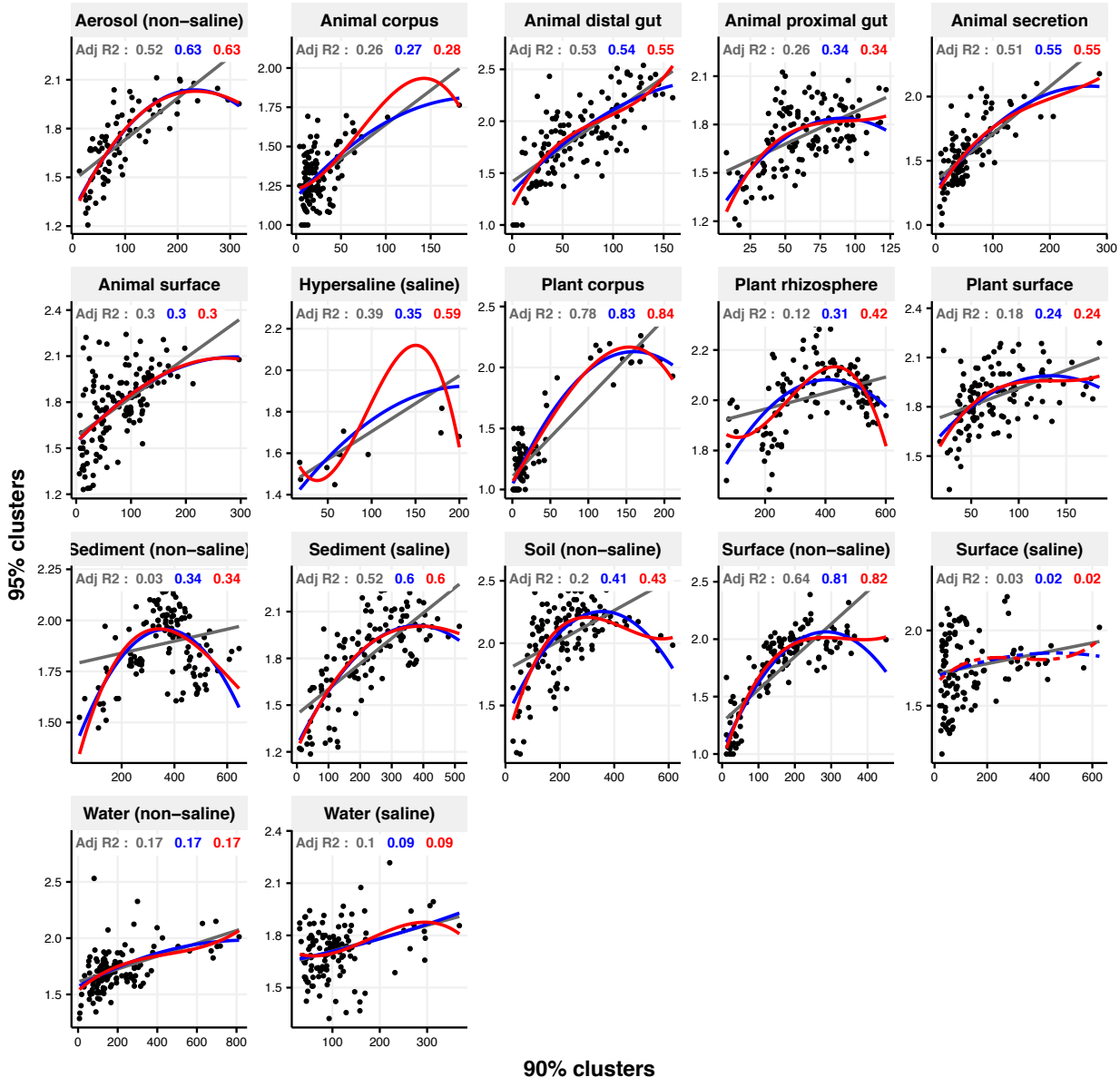

Figure S24. Focal clusters at 95% nucleotide identity.*Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. P-values are Bonferroni corrected for 17 tests. Significant, p<0.05 (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).*

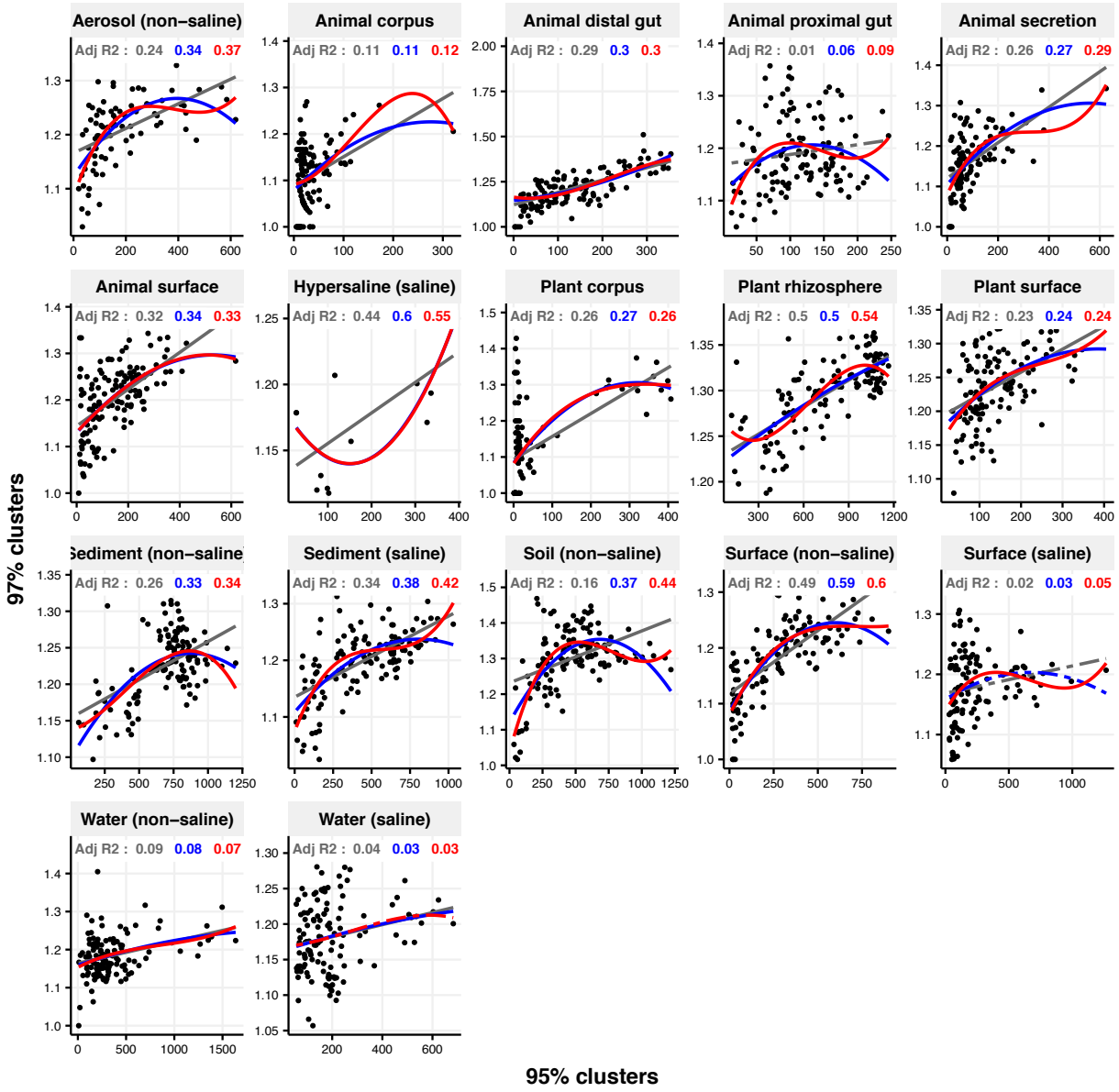

Figure S25. Focal clusters at 97% nucleotide identity. *Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. P-values are Bonferroni corrected for 17 tests. Significant, p<0.05 (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).*

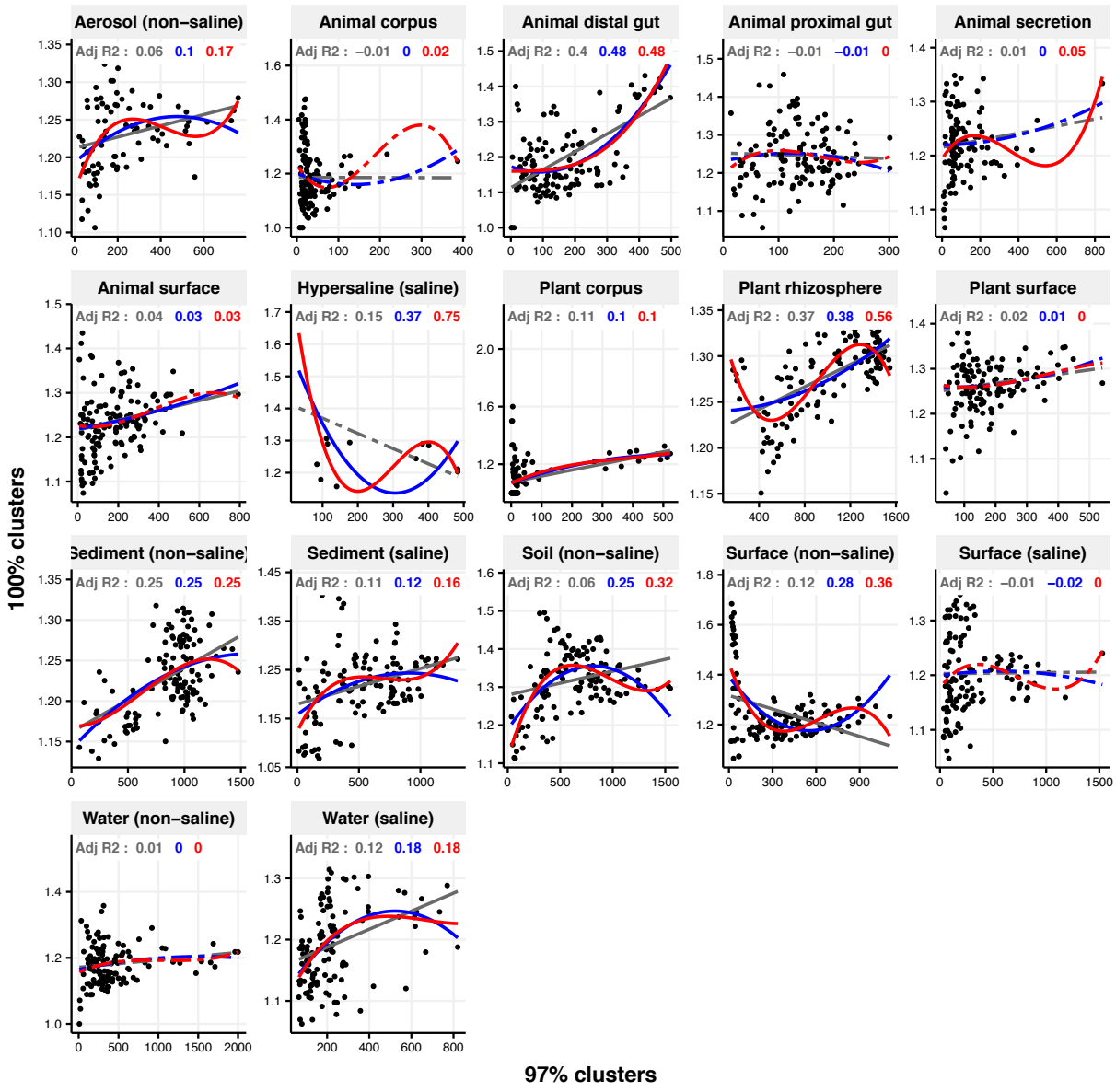

Figure S26. Focal clusters at 100% nucleotide identity. *Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. P-values are Bonferroni corrected for 17 tests. Significant, p<0.05 (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).*

Chapter 2. Community diversity is associated with intra-species genetic diversity and gene loss in the human gut microbiome

Naïma Madi[1], Daisy Chen[2,3,^], Richard Wolff[4,^], B. Jesse Shapiro[1,5,6,7,8*], and Nandita R. Garud[4,9,*]

1. Département de sciences biologiques, Université de Montréal, Canada;
2. Computational and Systems Biology, University of California, Los Angeles
3. Bioinformatics and Systems Biology Program, University of California, San Diego
4. Department of Ecology and Evolutionary Biology, University of California, Los Angeles
5. Department of Microbiology and Immunology, McGill University, Canada;
6. McGill Genome Centre, McGill University, Canada
7. Quebec Centre for Biodiversity Science, Canada
8. McGill Centre for Microbiome Research
9. Department of Human Genetics, University of California, Los Angeles

* Correspondence to jesse.shapiro@mcgill.ca and ngarud@ucla.edu. These authors contributed equally.

^ These authors contributed equally.

Abstract

The human gut microbiome contains a diversity of microbial species that varies in composition over time and across individuals. These species (and strains within species) can migrate across hosts and evolve by mutation and recombination within hosts. How the ecological process of community assembly interacts with intra-species diversity and evolutionary change is a longstanding question. Two contrasting hypotheses have been proposed based on ecological observations and theory: Diversity Begets Diversity (DBD), in which taxa tend to become more diverse in already diverse communities, and Ecological Controls (EC), in which higher community diversity impedes diversification within taxa. Previously, using 16S rRNA gene amplicon data across a range of environments, we showed a generally positive relationship between taxa diversity and community diversity at higher taxonomic levels, consistent with the predictions of DBD (Madi et al., 2020). However, this positive 'diversity slope' reaches a plateau at high levels of community diversity. Here we show that this general pattern holds at much finer genetic resolution, by analyzing intra-species strain and nucleotide variation in static and temporally sampled shotgun-sequenced fecal metagenomes from cohorts of healthy human hosts. We find that both intra-species polymorphism and strain number are positively correlated with community Shannon diversity. This trend is consistent with DBD, although we cannot exclude abiotic drivers of diversity. Shannon diversity is also predictive of increases in polymorphism over time scales up to ~4-6 months, after which the diversity slope flattens and then becomes negative—consistent with DBD eventually giving way to EC. Also, supporting a complex mixture of DBD and EC, the number of strains per focal species is positively associated with Shannon diversity but negatively associated with richness. Finally, we show that higher community diversity predicts gene loss in a focal species at a future time point. This observation is broadly consistent with the Black Queen Hypothesis, which posits that genes with functions provided by the community are less likely to be retained in a focal species' genome. Together, our results show that a mixture of DBD, EC, and Black Queen may operate simultaneously in the human gut microbiome, adding to a growing body of evidence that these eco-evolutionary processes are key drivers of biodiversity and ecosystem function.

Introduction

Our understanding of microbial evolution and diversification has been enriched by experimental studies of bacterial isolates in the laboratory, but it remains a challenge to study evolution in the context of more complex communities (Lenski 2017). Ongoing advances in culture-independent technologies have allowed us to study bacteria in the complex and dense communities in which they naturally occur (Garud and Pollard 2020). Within a community, individual players engage in many negative and positive ecological interactions. Negative interactions can originate from competition for resources and biomolecular warfare (Mitri and Richard Foster 2013; Hibbing et al. 2010), while positive interactions can stem from secreted metabolites that are used by other members of the community (cross-feeding) (Venturelli et al. 2018). These ecological interactions can create new niches and selective pressures, leading to eco-evolutionary feedbacks whose nature are yet to be fully understood.

Ecological interactions can yield positive or negative effects on the diversification of a focal species. Under the "Diversity Begets Diversity" (DBD) hypothesis, higher levels of community diversity increase the rate of speciation (or diversification, more generally) due to positive feedback mechanisms such as niche construction (Calcagno et al. 2017; Schluter and Pennell 2017). Competition for limited niche space could also drive DBD if species diversify into new niches to avoid competition (Meyer and Kassen 2007; Mitri and Richard Foster 2013; Schluter 2000; Schluter and Pennell 2017). By contrast, the "Ecological Controls" (EC) hypothesis posits that competition for a limited number of niches at high levels of community diversity results in a negative effect on further diversification. Metabolic models predict that DBD may initially spur diversification due to cross-feeding, but the diversification rate eventually slows and reaches a plateau as metabolic niches are filled (San Roman and Wagner 2021). These theoretical predictions are largely supported by our previous study involving 16S rRNA gene amplicon sequencing data from the Earth Microbiome Project, in which we observed a generally positive relationship (which we call the diversity slope; Figure 1) between community diversity and focal-taxon diversity at most taxonomic levels, reaching a plateau at the highest levels of diversity (Madi et al. 2020).

In this previous study, we found stronger support for DBD in the animal gut relative to more diverse microbiomes such as soils and sediments, which were closer to a plateau of diversity (Madi et al. 2020). While diversity slopes were generally positive at taxonomic levels as fine as amplicon sequence variants (akin to species or strains) within a genus, they were most positive at higher levels such as classes or phyla. A recent experiment on soil bacteria also found evidence of DBD at the family level, likely driven by niche construction and metabolic cross-feeding (Estrela et al. 2022). It therefore remains unclear if the predictions of DBD hold primarily at these higher taxonomic levels, involving the ecological process of community assembly, or if they also apply at the finer intra-species level. Within-host intra-species diversity can arise by co-colonization of a host by genetically distinct strains belonging to the same species or evolutionary diversification of a lineage via *de novo* mutation and gene gain/loss events within a host.

Such fine-scale strain-level variation has important functional and ecological consequences; among other things, strains are known to engage in interactions that cannot be predicted from their species identity alone (Goyal et al. 2022.). Although closely-related bacteria are expected to have broadly similar niche preferences, finer-scale niches may differ below the species level (Martiny et al. 2015). For example, the acquisition of a carbohydrate-active enzyme by *Bacteroides plebeius* allows it to exploit a new dietary niche in the guts of people consuming nori (seaweed) (Hehemann et al. 2010), and single nucleotide adaptations permit *Enterococcus gallinarum* translocation across the intestinal barrier resulting in inflammation (Yang et al. 2022). Despite their potential phenotypic effects, it is unknown if such fine-scale genetic changes are favored by higher community diversity (due for example to niche construction, as predicted by DBD) or suppressed (due to competition for limited niche space, as predicted by EC). Competition could also lead to DBD if focal species evolve new niche preferences to avoid extinction (Mitri and Richard Foster 2013; Schluter 2000) – an idea with some support in experimental microcosms (Meyer and Kassen 2007) but largely unexplored in natural communities.

Here, we investigate the relationship between intra-species genetic diversity and community diversity in the human gut microbiome, a well-studied system in which we previously found support for DBD at higher taxonomic levels. We use static and temporal shotgun metagenomic data from a large panel of healthy adult hosts from the Human Microbiome Project

(Lloyd-Price et al. 2017; Human Microbiome Project 2012) as well as from four healthy individuals sampled almost daily over the course of one year (Poyet et al. 2019). Using metagenomic data allows us to track change in single nucleotide variation, strain diversity, and gene gain or loss events within relatively abundant species in the microbiome, and study how these measures of intra-species diversity are associated with community diversity. Although such analyses of natural diversity cannot fully control for unmeasured confounding environmental factors, they are an important complement to controlled experimental and theoretical studies which lack real-world complexity.

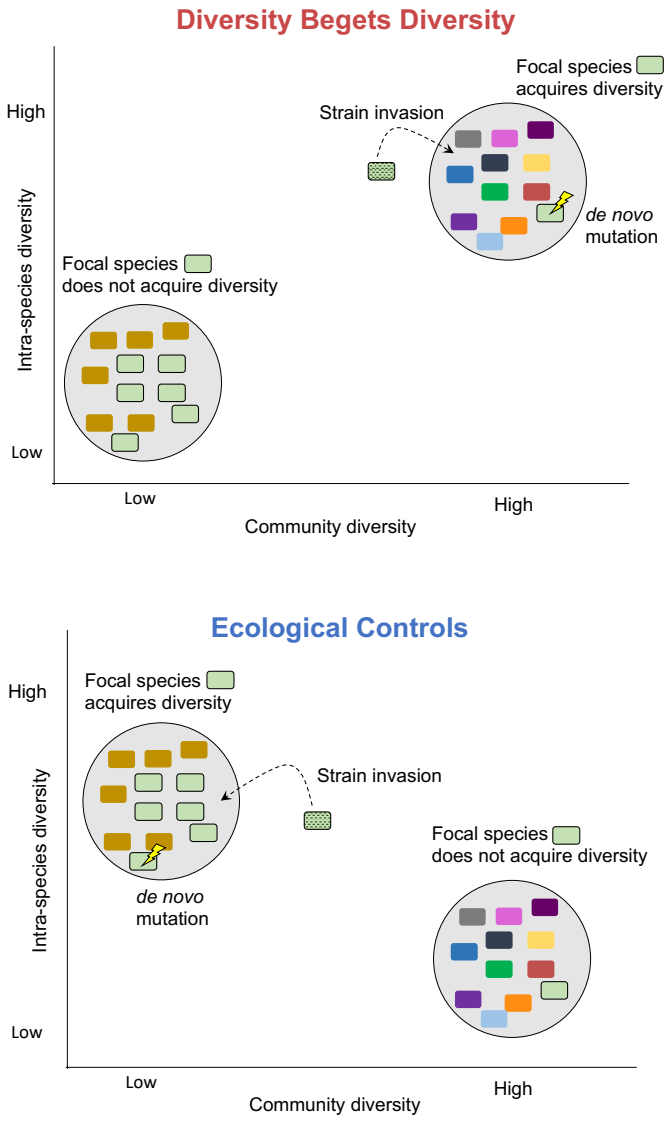

Figure 1. Diversity Begets Diversity (DBD) and Ecological Controls (EC) hypotheses illustrated. *Hypothetical*

microbial communities are illustrated as grey circles containing assemblages of microbial species, shown in different colors. 'Diversity begets diversity' means that the focal species is more likely to acquire diversity – through de novo mutation, invasion of a different strain of the same species, or a combination of both – in a community with high diversity. This is because new niches are created in a more diverse community. By contrast, 'Ecological Controls' means that the focal species is more likely to acquire diversity through strain invasion or mutation in a community with low diversity. This is because niches remain unfilled in a low-diversity community, while niche space is saturated in a high-diversity community, impeding further diversification.

Results

We investigated the relationship between community diversity and within-species genetic diversity in human gut microbiota using two shotgun metagenomic datasets. First, we analyzed data from a panel of 249 healthy hosts (Lloyd-Price et al. 2017; Human Microbiome Project 2012), in which stool samples were collected 1-3 times from each host at approximately 6-month intervals. Second, we analyzed data from four individuals sampled more densely over the course of ~18 months (Poyet et al. 2019). In both cases, we only consider intra-species diversity of relatively abundant species that are well sampled in these metagenomic datasets (Methods).

We examined several metrics of community diversity and intra-species diversity and calculated the slope of their relationship, defined as the diversity slope (Figure 1). We note that intra-species diversity can arise within hosts via *de novo* point mutation, gene gain or loss, or the coexistence of genetically distinct strains that diverged before colonizing the host. To quantify community diversity, we calculated Shannon diversity and richness at the species level. Shannon diversity is relatively insensitive to sampling effort (Madi et al. 2020; Walters and Martiny 2020) but richness can be underestimated in low sample sizes. We therefore computed richness on data rarefied to an equal number of reads per sample, yielding generally similar results to unrarefied data (described below). In all cases, we included the number of reads per sample (coverage) as a covariate in our models, as this could affect estimates of both community diversity and intra-species diversity. To quantify intra-species diversity, we used a reference genome-based approach to call single nucleotide variants (SNVs) and gene copy number variants (CNVs) within each focal species and computed polymorphism rates, measured as the fraction of synonymous nucleotide sites in a species' core genome with intermediate allele frequencies

(between 0.2 and 0.8) within a host (Methods). We also repeated the analysis on nonsynonymous sites, as these are subject to stronger selective constraints. As an additional metric of intra-species diversity, we inferred the number of strains within each species using StrainFinder applied to all polymorphic sites (including those outside the 0.2-0.8 frequency range) (Smillie et al. 2018).

Community diversity is positively associated with intra-species polymorphism in the human gut microbiome

As an exploratory visualization, we began by plotting the relationship between community diversity and intra-species polymorphism rate calculated at synonymous sites in cross-sectional HMP metagenomes for the nine most prevalent species (Figure 2A,B). The slope of this relationship (the diversity slope; Figure 1) provides an indicator of the evidence for DBD (positive slope) or EC (flat or negative slope). The relationship between polymorphism rate and community diversity was mostly positive in the top nine most prevalent species in HMP hosts (Figure 2A,B). These nine species are used as a simple illustration of the diversity slope, not as a formal hypothesis-testing framework.

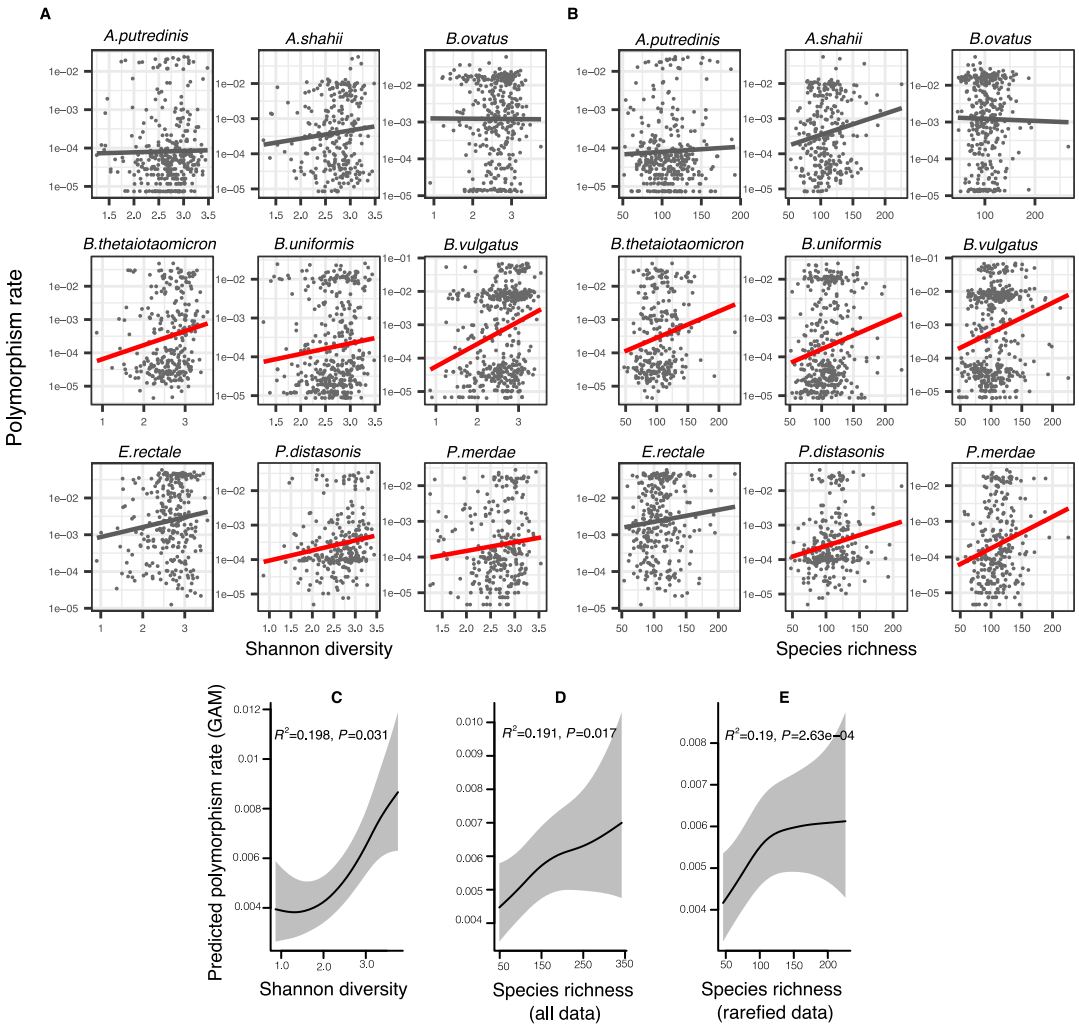

Figure 2. Positive association between community diversity and within-species polymorphism in cross-sectional Human Microbiome Project samples*. (A) Scatter plots showing the relationship between community Shannon diversity and within-species polymorphism rate (estimated at synonymous sites) in the nine most prevalent species in HMP. (B) Scatter plots showing the relationship between species richness and within-species polymorphism rate in the nine most prevalent species in HMP. These are simple correlations to show the relationships in the raw data. Significant correlations are shown with red trendlines (Spearman correlation, P<0.05); non-significant trendlines are in gray. Results of generalized additive models (GAMs) predicting polymorphism rate in a focal species as a function of (C) Shannon diversity, (D) species richness estimated on all sequence data, and (E) species richness estimated on rarefied sequence data. GAMs are based on data from 69 bacterial species across 249 HMP stool donors. Adjusted R^2 and Chi-square P-values corresponding to the predictor effect are displayed in each panel. Shaded areas show the 95% confidence interval of each model prediction. See Supplementary File 1a and supplementary file 2 section 1 for detailed model outputs.*

To generalize across species and to formally test the predictions of DBD, we fit generalized additive models (GAMs) to the HMP data. Using GAMs, we are able to model non-linear relationships and account for random variation in the strength of the diversity slope across

bacterial species, the uneven number of samples per host, and the non-independence of samples from the same host (Methods; see Supplementary File 1a and Supplementary File 2 section 1 for additional model details). These GAMs included 69 focal species with sufficient coverage to quantify within-species polymorphism (Methods); the results therefore apply to relatively abundant species in the human gut microbiome. GAMs showed an overall positive association between within-species polymorphism and Shannon diversity (Fig 2C, GAM, *P*=0.031, Chi-square test) as well as between within-species polymorphism and community richness after controlling for coverage as a covariate (Fig 2D, GAM, *P*=0.017, Chi-square test) or rarefying samples to an equal number of reads (Fig 2E, GAM, *P*=2.63e-04, Chi-square test). The random effect of species identity is highly significant in all models, indicating that each bacterial species has its own characteristic diversity slope (Supplementary File 1a). It appears that synonymous polymorphism reaches a plateau at high levels of community richness, which is particularly evident when using rarefied data (Fig 2E). Using the same GAMs applied to nonsynonymous polymorphism, we found no significant associations between diversity and within-species polymorphism rate (GAM, *P*>0.05, Chi-square test) (Supplementary File 1b, Supplementary File 2 section 4). This could be due to lower statistical power, since there are fewer nonsynonymous than synonymous sites, or could reflect a true difference in the diversity slope between these site categories.

These generally positive correlations between focal species polymorphism and species-level measures of community diversity also hold when community diversity is measured at higher taxonomic levels; specifically, synonymous polymorphism rate was significantly positively associated with Shannon diversity calculated at the genus and family levels (GAMs, *P*<0.05, Chi-square test) (Figure 2-figure supplement 1, Supplementary File 1c). However, synonymous polymorphism rate was not significantly associated with Shannon diversity calculated at the highest taxonomic levels (order, class and phylum, GAMs, *P*>0.05, Chi-square test). The positive correlation between polymorphism rate and richness held at all taxonomic levels (GAMs, *P*<0.05, Chi-square test) (Figure 2-figure supplement 1, Supplementary File 1c, Supplementary File 2 section 2 and 3). When estimated at nonsynonymous sites, polymorphism rate was not significantly correlated with Shannon diversity at any taxonomic level (GAMs, *P*>0.05, Chi-square test), but was positively correlated with richness at the highest levels (phyla, class and order,

P=3e-04, *P*=0.017 and *P*=6.11e-04 respectively, Chi-square test from GAMs) (Figure 2-figure supplement 2, Supplementary File 1d, Supplementary File 2 section 5 and 6). Even when not statistically significant, the diversity slopes were generally positive at all taxonomic levels for both synonymous and nonsynonymous polymorphism (Figure 2-figure supplements 1 and 2). Overall, these results are consistent with the predictions of DBD at most taxonomic levels. However, slightly different relationships are observed when considering different measures of community diversity (Shannon or richness) and different components of within-species diversity (nonsynonymous or synonymous).

Different measures of community diversity have contrasting associations with intra-species strain diversity

Within host polymorphism rates span several orders of magnitude (10^{-5}/bp to 10^{-2}/bp), largely due to the fact that strain content is variable across hosts. As previously argued (Garud et al. 2019), with conservatively high estimates for mutation rate (μ~10^{-9}) (Sung et al. 2012), generation times (~ 10 / day) (Poulsen et al. 1995), and time since colonization (<100 years), polymorphism rates of ~10^{-2}/bp or more are inconsistent with within-host diversification of a single colonizing lineage. Therefore, hosts with relatively high intra-host polymorphism rates are likely colonized by mixtures of multiple strains that diverged long before colonizing a host. Moreover, recent work suggests that the numbers and genetic composition of strains colonizing a host can vary from host to host (Garud et al. 2019; Olm, Brown, Brooks, Firek, et al. 2017; Russell and Cavanaugh 2017; Truong et al. 2017; Verster et al. 2017). The associations between polymorphism and community diversity (Figure 2) are likely driven by a combination of *de novo* mutation and co-colonization by multiple strains.

To separate these two sources of diversity and to explicitly account for the strain structure within hosts, we inferred the number of strains per focal species with StrainFinder (Smillie et al. 2018) (Methods) and used strain number as another quantifier of intra-species diversity. The relationship between community diversity and strain number varied depending on the focal species and the measure of community diversity. For example, the inferred number of *Bacteroides vulgatus* strains increased with community diversity, while *B. uniformis* strain count decreased or remained flat (Figure 3A, B). Expanding upon these examples, we used generalized

linear mixed models (GLMMs) to investigate the relationship between the number of strains per focal species and community diversity, while taking into account coverage per sample as a covariate and variation between species, hosts and samples as random effects (Methods). GLMMs are a special case of GAMs that can handle overdispersed, zero-truncated count data such as strain counts. The number of strains per focal species was positively correlated with community Shannon diversity (GLMM, *P*=3.58e-07, likelihood ratio test (LRT)) (Fig 3C, Supplementary File 1e, Supplementary File 2 section 7.1). This suggests that the positive correlation between polymorphism rate and Shannon diversity (Figure 2) is due at least in part to strain diversity.

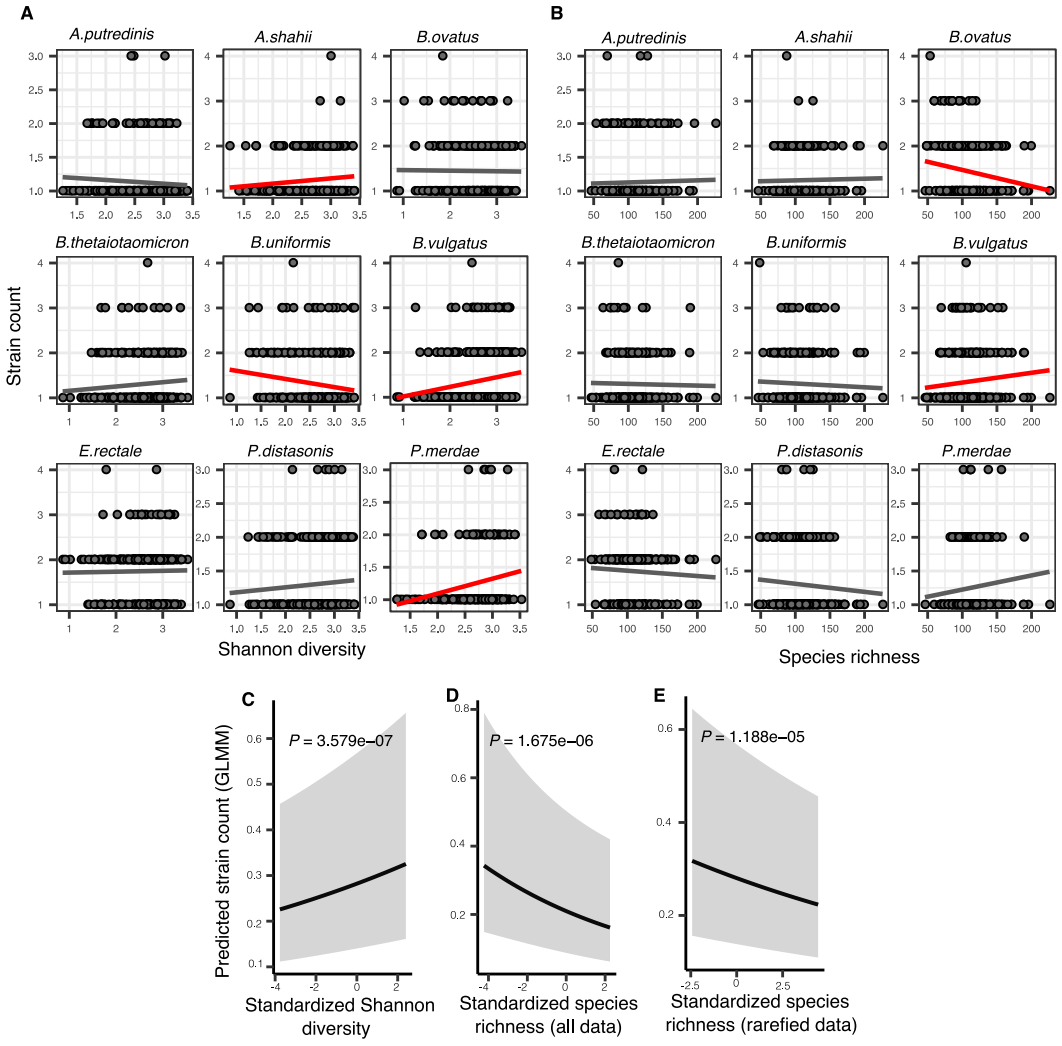

Figure 3. Associations between community diversity and strain number in cross-sectional Human Microbiome Project samples*. (A) Scatter plots showing the relationship between Shannon diversity and the inferred number of strains within each of the nine most prevalent species in HMP. (B) Scatter plots showing the relationship between species richness and the inferred number of strains within each of the nine most prevalent species in HMP. Significant*

linear correlations are shown with red trendlines (Pearson correlation, P<0.05); non-significant trend lines are in gray. Results of generalized linear mixed models (GLMMs) predicting strain count in a focal species as a function of (C) Shannon diversity, (D) species richness estimated on all data, and (E) species richness estimated on rarefied sequence data. Diversity estimates (x-axis) are standardized to zero mean and unit variance in the models. The Y-axis shows the mean number of strains per focal species predicted by the GLMM. GLMMs are based on data from 184 bacterial species across 249 HMP stool donors. P-values (likelihood ratio test) are displayed in each panel. Shaded areas show the 95% confidence interval of each model prediction. See Supplementary File 1e and Supplementary File 2 section 7 for detailed model outputs.

By contrast, species richness was negatively correlated with strain number (GLMM, *P*=1.67e-06, LRT) (Fig 3D, Supplementary File 1e, Supplementary File 2 section 7.2). The negative relationship with richness was unlikely to be confounded by sequencing depth, since the same result was obtained using rarefied data (Fig 3E, Supplementary File 1e, Supplementary File 2 section 7.3). The negative strain number-richness relationship also held at all other taxonomic ranks (GLMM, *P*<0.05, LRT), while the strain number-Shannon diversity relationship was generally positive (Fig 3-Figure supplement 1, Supplementary File 1f, Supplementary File 2 section 8-9). These effects also appear to be species-specific: for example, the number of *B. vulgatus* strains per host is positively correlated with both Shannon diversity and richness (consistent with DBD predictions) whereas *B. ovatus* has no relationship with Shannon diversity but a negative correlation with richness (consistent with EC; Fig 2A, B). Together, these results reveal that different components of community diversity can have contrasting effects on the diversity slope.

Community Shannon diversity is a predictor of intra-species polymorphism and gene loss in time series data

Our analyses thus far have considered only individual time points, which represent static snapshots of the dynamic processes of community assembly and evolution in the microbiome. To interrogate these phenomena over time, we analyzed 160 HMP subjects who were sampled 2-3 times ~6 months apart. Under a DBD model, we expect community diversity at an earlier time point to result in higher within-species polymorphism at a future time point. To test this expectation, we defined 'polymorphism change' as the difference between polymorphism rates at the two time points (Methods). We also investigated the effects of community diversity on gene loss and gain events within a focal species, as such changes in gene content are known to occur frequently within host gut microbiomes (Garud et al. 2019; Groussin et al. 2021; Yaffe and

Relman 2020; Zhao et al. 2019). Here a gene was considered absent if its copy number (c) was <0.05 and present if $0.6 \leq c \leq 1.2$. As in the cross-sectional analyses above, we also controlled for sequencing depth of the sample and excluded genes with aberrant coverage or presence in multiple species (Methods).

In HMP samples, polymorphism change showed no significant relationships with community diversity at the earlier time point, whether it was estimated with Shannon index or species richness (GAM, P>0.05) (Supplementary File 2 section 10.1). These results suggest that DBD is negligible or undetectable over ~6-month time lags in the human gut. By contrast, we found that gene loss in a focal species between two consecutive time points was positively correlated with community diversity at the earlier time point (Figure 4; GLMM, P=0.028, P=0.034 and P=0.049, LRT for Shannon, richness and rarefied richness respectively) (Supplementary File 1g, Supplementary file 2 section 10.3). Gene gains did not show any significant relationships with community diversity (GLMM, P>0.05). Selection for gene loss in more diverse communities is a prediction of the Black Queen Hypothesis (BQH), provided that higher community diversity results in more redundant gene functions that compensate for losses in a focal species (Morris, Lenski, and Zinser 2012). Most species in HMP samples lost fewer than ten genes over ~6 months – consistent with *de novo* deletion events of a few genes – but occasionally hundreds of genes were lost from a host, suggesting that strains with smaller genomes were selected in more diverse communities (Figure 4A, 4B).

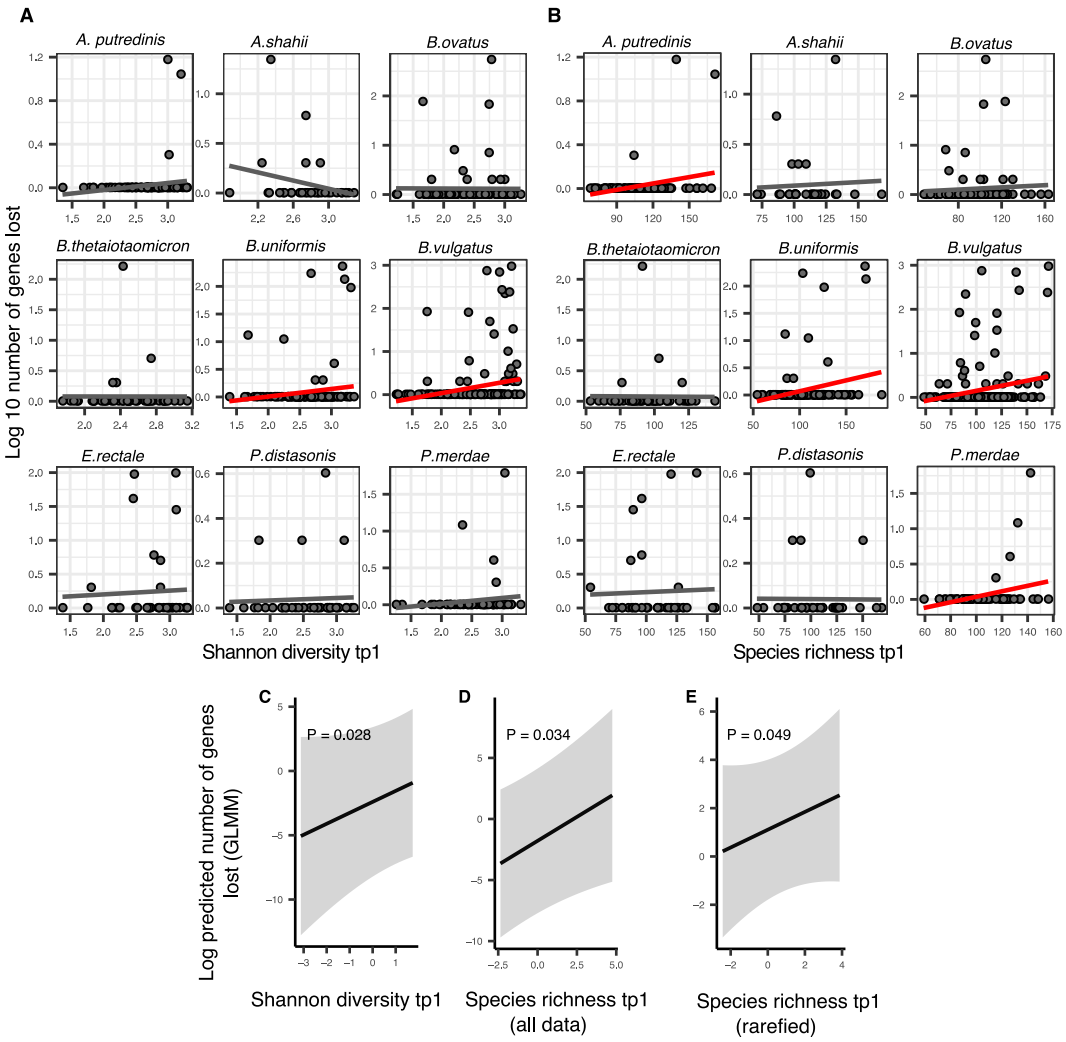

Figure 4. Positive association between community diversity and gene loss in Human Microbiome Project time series*. (A) Scatter plots showing the relationship between Shannon diversity at time point 1 (tp1) and gene loss between tp1 and tp2 within each of the nine most prevalent species in HMP. (B) Scatter plots showing the relationship between species richness at tp1 and gene loss between tp1 and tp2 within each of the nine most prevalent species in HMP. Significant linear correlations are shown with red trendlines (Pearson correlation, P<0.05); non-significant trend lines are in gray. The Y-axis is plotted on a log10 scale for clarity. Results of generalized linear mixed models (GLMMs) predicting gene loss in a focal species as a function of (C) Shannon diversity, (D) species richness estimated on all data, and (E) species richness estimated on rarefied sequence data. P-values (likelihood ratio test) are displayed in each panel. Shaded areas show the 95% confidence interval of each model prediction. The Y-axis is plotted on the link scale, which corresponds to log for negative binomial GLMMs with a count response. GLMMs are based on data from 54 bacterial species across 154 HMP stool donors sampled at more than one time point. See Supplementary file 1g and Supplementary File 2 section 10 for detailed model outputs.*

To study these dynamics at higher temporal resolution, we analyzed shotgun metagenomic data from four more frequently sampled healthy individuals from a previous study (Poyet et al. 2019). Stool from donor *am* was sequenced over 18 months with a median of one

day between samples; *an* over 12 months (median 2 days between samples); *ao* over 5 months (median 1 day between samples); and *ae* over 7 months (median 2 days between samples). In this data, we tracked both polymorphism change and gene gains and losses between two successive time points in 15 species with a minimal marker gene coverage of 10 in at least ten samples. These include seven species of *Bacteroides*, two *Eubacterium,* two *Faecalibacterium*, two *Ruminococcus,* as well as *Alistipes putredinis* and *Parabacteroides merdae*.

Using the Poyet dataset, we asked whether community diversity in the gut microbiome at one time point could predict polymorphism change at a future time point by fitting GAMs with the change in polymorphism rate as a function of the interaction between community diversity at the first time point and the number of days between the two time points. Shannon diversity at the earlier time point was correlated with increases in polymorphism (consistent with DBD) up to ~150 days (~4.5 months) into the future (Fig 5-Figure supplement 1), but this relationship became weaker and then inverted (consistent with EC) at longer time lags (Fig 5A, Supplementary File 1h, GAM, P=0.023, Chi-square test). The diversity slope is approximately flat for time lags between four and six months, which could explain why no significant relationship was found in HMP, where samples were collected every ~6 months. No relationship was observed between community richness and changes in polymorphism (Supplementary File 1h, GAM, P>0.05).

We next asked if community diversity at one time point could predict gene gains or losses at future time points by fitting GLMMs (analogous to the GAMs above, but more appropriate for gain/loss count data). Our method does not explicitly distinguish between gene gain/loss arising from recombination or deletion versus replacement of strains with different gene content. We found that community Shannon diversity predicted future gene loss in a focal species, and this effect became stronger with longer time lags (Fig 5B, Supplementary File 1i, GLMM, P=0.006, LRT for the effect of the interaction between the initial Shannon diversity and time lag on the number of genes lost). The model predicts that increasing Shannon diversity from its minimum to its maximum would result in the loss of 0.075 genes from a focal species after 250 days. In other words, about one of the 15 focal species considered would be expected to lose a gene in this time frame.

Higher Shannon diversity was also associated with fewer gene gains, and this relationship also became stronger over time (Fig 5C, Supplementary File 1i, GLMM, *P*=1.11e-09, LRT). We found a similar relationship between community species richness and gene gains, although the relationship was slightly positive at shorter time lags (Fig 5D, Supplementary File 1i, GLMM, *P*=3.41e-04, LRT). No significant relationship was observed between richness and gene loss (Supplementary File 1i, GLMM, *P*>0.05). Taken together with the HMP results (Fig 4), these longer time series reveal how the sign of the diversity slope can vary over time and how community diversity is generally predictive of reduced focal species gene content.

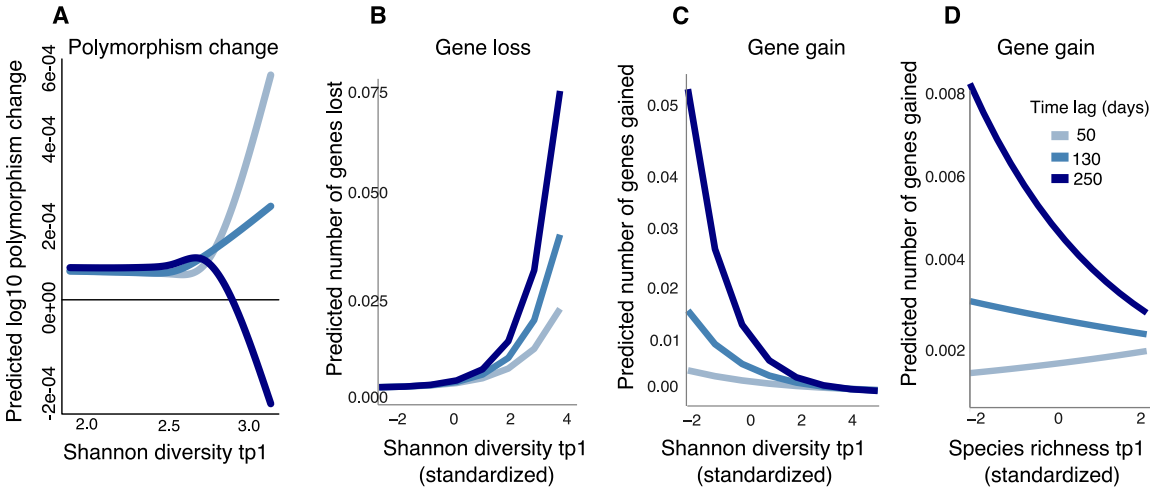

Figure 5. Community diversity is associated with increases in focal species polymorphism over short time lags and net gene loss in dense gut microbiome time series*. (A) Results of a GAM predicting polymorphism change in a focal species as a function of the interaction between Shannon diversity at the first time point and the time lag (days) between two time points in data from Poyet et al. The response (Y-axis) was log transformed in the Gaussian GAM. Results of GLMMs predicting (B) Number of genes lost and (C) Number of genes gained between two time points in a focal species as a function of the interaction between Shannon diversity at the first time point and the time lag between the two time points. (D) Results of the GLMM predicting the number of genes gained in a focal species as a function of the interaction between rarefied species richness at the first time point and the time lag between the two time points. The illustrated time lags correspond to the first quartile (50 days), the median (130 days), and the third quartile (250 days). See Supplementary Files 1h and i and Supplementary File 2 section 11 for detailed model outputs. These analyses are based on data from 15 bacterial species across 4 stool donors from Poyet et al. Only statistically significant relationships are plotted. Non-significant relationships are not shown: the GAM predicting polymorphism change as a function of rarefied richness (P>0.05) and the GLMM predicting the number of genes lost as a function of rarefied richness (P>0.05).*

Discussion

How eco-evolutionary feedbacks shape biological communities is an open question that to date has received substantial experimental and theoretical attention but is challenging to address in nature. In our previous study using 16S rRNA amplicon sequences from the Earth Microbiome Project, we found generally positive diversity slopes that eventually flattened at high levels of community diversity (Madi et al. 2020). This pattern is generally consistent with the predictions of DBD during the early stages of community assembly, but at later stages becomes more consistent with EC as niches become filled. Based on the time series metagenomic data analyzed here, the predictions of DBD also tend to hold over short time scales but fail over longer time scales of several months. Whether this leads to a terminal plateau of diversity, or whether ecological disturbances lead to cycles of DBD and EC, deserves further study.

In our previous study, the animal gut microbiome had one of the highest positive diversity slopes, making it an ideal candidate for investigating eco-evolutionary interactions at greater intra-species resolution using metagenomic data. In this follow-up study, we investigate the same phenomenon at a subspecies level, with results that are broadly consistent with the predictions of DBD giving way to EC over long time scales. We note that experiments supporting DBD have generally been conducted over short time scales ranging from two to 20 days (Estrela et al. 2022; Jousset et al. 2016b), consistent with the importance of DBD early in community assembly. We also identify several nuances and caveats to this general conclusion, which are discussed below in detail.

Another recent study also found evidence for eco-evolutionary feedbacks in the HMP, in the form of a positive relationship between evolutionary modifications or strain replacements in a focal species and community diversity (Good and Rosenfeld 2022). Using a model, they further showed that these eco-evolutionary dynamics could be explained by resource competition and did not require the cross-feeding interactions previously invoked (Estrela et al. 2022; San Roman and Wagner 2021; San Roman and Wagner 2018) to explain DBD at higher taxonomic levels. This could be because cross-feeding operates at the family- or genus- level and is less relevant at finer evolutionary scales.

There are several noteworthy caveats to our study. First, using metagenomic data from human microbiomes allowed us to study genetic diversity, but limited us to considering only relatively abundant species with genomes that were well-covered by short sequence reads. Deeper or more targeted sequencing may permit us to determine whether the same patterns hold for rarer members of the microbiome. However, it is notable that the majority of the dozens of species across the two datasets analyzed support DBD, suggesting that the phenomenon may generalize.

Second, we cannot establish causal relationships without controlled experiments. We are therefore careful to conclude that positive diversity slopes are consistent with the predictions of DBD, and negative slopes with EC, but unmeasured environmental drivers could be at play. For example, increased dietary diversity could simultaneously select for higher community diversity and also higher intra-species diversity. In our previous study, we found that positive diversity slopes persisted even after controlling for potential abiotic drivers such as pH and temperature (Madi et al. 2020), but a similar analysis was not possible here due to a lack of metadata. Neutral processes can account for several ecological patterns such as species-area relationships (Hubbell 2001), and must be rejected in favor of niche-centric models like DBD or EC. Using neutral models without DBD or EC, we found generally flat or negative diversity slopes due to sampling processes alone and that positive slopes were hard to explain with a neutral model (Madi et al. 2020). These models were intended mainly for 16S rRNA gene sequence data, but we expect the general conclusions to extend to metagenomic data. Nevertheless, further modeling and experimental work will be required to fully exclude a neutral explanation for the diversity slopes we report in the human gut microbiome.

Based on controlled experiments (Estrela et al. 2022) and modeling studies (San Roman and Wagner 2021), DBD is a plausible causal explanation for positive diversity slopes in the gut microbiome. Although they also note that causality is difficult to establish, Good and Rosenfeld (2022) suggest the importance of focal species evolution as a driver of changes in community structure, as shown in an experimental study of *Pseudomonas* in compost communities (Padfield et al. 2020). Clearly, further work is needed to establish the extent and relative rates of eco-

evolutionary feedback in both directions. How these feedbacks among bacteria are influenced by abiotic factors and by interactions with fungi, archaea, and phages also deserve further study.

Third, the diversity slope changes depending on which component of within-species diversity or community diversity is considered. Notably, the number of strains within a focal species is positively correlated with Shannon diversity, but inversely correlated with species richness, suggesting that the ability of strains to colonize a host may be associated with higher community evenness rather than total species count. Higher evenness might maximize the chance of inter-species interactions, whereas higher richness might be driven by rare species that are less likely to interact. Although Shannon diversity is considered to be more robust and informative than richness in estimating bacterial diversity (He et al. 2013; Reese and Dunn 2018), we observe the same contrasting results between Shannon diversity and richness when community diversity is calculated at higher taxonomic levels, suggesting that this pattern is not due to artifacts such as sequencing effort.

Our measures of intra-species diversity included both synonymous and nonsynonymous single nucleotide variants, inferred strain richness, and gene content. Synonymous nucleotide variation was consistently and positively associated with both community richness and Shannon diversity at all taxonomic levels (although not always with statistical significance). Nonsynonymous variation also tended to track positively with both measures of community diversity but was only statistically significantly associated with phylum and class richness. This suggests that evolutionarily older, less selectively constrained synonymous mutations and more recent nonsynonymous mutations that affect protein structure both track similarly with measures of community diversity. Nonetheless, a parsimonious explanation for possible differences between the two classes is that while they are affected similarly, we have more statistical power to identify correlations in the more numerous synonymous mutations. This merits further investigation.

Metagenomes from the same individual sampled over time allowed us to detect gene gain and loss events. In both HMP and Poyet et al. time series, community diversity was predictive of future gene loss in a focal species. This phenomenon is not explicitly predicted by either DBD or EC but it is compatible with aspects of the Black Queen Hypothesis, with some caveats. BQH

predicts that a focal species will be less likely to encode genes with functions provided by other members of the surrounding community if such functions are "leaky" and available as diffusible public goods (Morris and Lenski 2012). The BQH could also act as a driver of polymorphism within a species (Morris, Papoulis, and Lenski 2014). Gene loss may be adaptive, provided that there is a cost to encoding and expressing the relevant genes (Albalat and Cañestro 2016; Koskiniemi et al. 2012; Simonsen 2022). The tendency for reductive genome evolution in bacteria is well established (Albalat and Cañestro 2016; Koskiniemi et al. 2012; Puigbò et al. 2014). Genome reduction is a particular hallmark of endosymbiotic bacteria, which depend on their hosts for many metabolic gene products (McCutcheon and Moran 2012; Nikoh et al. 2011). It has been shown that uncultivated bacteria from the gut have undergone considerable genome reduction, which may be an adaptive process that results from reliance on public goods (Nayfach et al. 2019). In the gut microbiome, the BQH has been invoked to explain the distribution of genes involved in vitamin B metabolism (Sharma et al. 2019) and iron acquisition (Vatanen et al. 2019).

Our findings in human gut metagenomes are compatible with the BQH under the assumption that increasing community diversity also increases the availability of leaky gene products – which may not be the case if genomes in the gut microbiome are functionally redundant, as inferred in a recent study (Tian et al. 2020). This study found that species in the gut microbiome were highly redundant at the level of annotated metabolic pathways (KEGG orthologs) and that more functionally redundant microbiomes were more resistant to colonization by fecal transplants. Relatively low-redundancy microbiomes could therefore be more easily colonized but might also require migrants to encode more gene functions in order to persist. Importantly, functional redundancy may be high at the level of well-annotated metabolic functions, but low at the finer level of individual gene families, as demonstrated in marine microbiomes (Galand et al. 2018) but not yet studied explicitly in the gut. Here we report that genome reduction in the gut is higher in more diverse gut communities. This could be due to *de novo* gene loss, preferential establishment of migrant strains encoding fewer genes, or a combination of the two. The mechanisms underlying this correlation remain unclear and could be due to biotic interactions – including metabolic cross-feeding as posited by some models (Estrela et al. 2022; San Roman and Wagner 2021; San Roman and Wagner 2018) but not others

(Good and Rosenfeld 2022) – or due to unknown abiotic drivers of both community diversity and gene loss. Finally, we measured community diversity from the phylum to the species level, not below. We therefore did not investigate how the BQH could extend to maintain gene content variation within a species, as has been shown experimentally in *E. coli* (Morris, Papoulis, and Lenski 2014). This could be an avenue for future work.

In our previous analysis of lower-resolution 16S rRNA amplicon sequences, we reported a tendency for focal genera with larger genomes to have higher diversity slopes, perhaps because they experience stronger DBD (Madi et al. 2020). At face value, this tendency seems at odds with the BQH, which predicts genome reduction in more diverse communities. This apparent contradiction may be reconciled by considering eco-evolutionary dynamics on different time scales. A recent study used phylogenetic and metabolic reconstructions to show that gene gains often drive metabolic dependencies among bacteria (Goyal 2021), potentially explaining why genera with larger maximum genome size could experience stronger DBD. Our earlier study only had the genetic resolution to consider focal taxa down to the genus level, and by using the maximum genome size observed in a public database we did not capture the dynamic process of gene gain and loss within a species, as was possible in the current metagenomic study. It is therefore possible that on longer (ecological) time scales, larger genomes have more metabolic interactions and thus experience stronger DBD, while genome reduction in more diverse communities occurs on shorter (evolutionary) time scales.

In summary, we demonstrate how metagenomic data can be used to test the predictions of eco-evolutionary theory, including DBD, EC, and the BQH. It remains to be seen whether the distinct eco-evolutionary processes proposed by DBD and the BQH operate orthogonally or whether they interact. If BQH leads to gene losses that remain polymorphic rather than being lost entirely from the species (Morris, Papoulis, and Lenski 2014) – or invasions of strains with fewer genes that remain incomplete and do not replace the resident strain – this could be viewed as a form of diversification and perhaps a special case of DBD. Here we considered gene loss as a directional process; we did not attempt to distinguish between directional changes in gene copy number and the complete extinction of a gene, which is difficult to show using metagenomic

data. Future work could attempt to resolve this point and to potentially combine DBD and BQH into a unified theory.

Data and materials availability

The raw sequencing reads for the metagenomic samples used in this study were downloaded from the Human Microbiome Project Consortium 2012 and Lloyd-Price et al. (2017) (URL: https://aws.amazon.com/datasets/human-microbiome-project/); and Poyet et al. 2019 (NCBI accession number PRJNA544527). All computer code for this paper is available at https://github.com/Naima16/DBD_in_gut_microbiome.

Methods

We used MIDAS (Metagenomic Intra-Species Diversity Analysis System, version 1.2, downloaded on November 21, 2016) (Nayfach et al. 2016) to estimate within-species nucleotide and gene content of raw metagenomic whole genome shotgun sequencing data for HMP1-2 and Poyet et al. 2019 data. MIDAS relies on a reference database comprised of 31,007 bacterial genomes that are clustered into 5,952 species, covering roughly 50% of species found in human stool metagenomes from "urban" individuals. Described below are the parameters used to estimate species abundances, single nucleotide variants (SNVs), and gene copy number variants (CNVs) with MIDAS.

Estimation of species content

We estimated species abundances, SNVs and CNVs by mapping metagenomic shotgun reads to reference genomes. Since a component of this work relies on quantifying polymorphism and CNV changes over time, we constructed a "personal" reference database to avoid spurious inferences of allele frequency and CNV changes due to errors in mapping of reads to regions of the genome shared by multiple species (Garud et al. 2019). This per-host reference database was comprised of the union of all species present at one or more timepoints so as to be as inclusive

as possible to prevent reads from being "donated" to reference genome, while also being selective to prevent a reference genome from "stealing" reads from a species truly present.

To estimate the species relative abundances for each host x timepoint sample, we mapped reads to 15 universal single-copy marker genes that are a part of the MIDAS pipeline (Nayfach et al. 2016) (Wu, Jospin, and Eisen 2013) and belong to the 5,952 species in the MIDAS reference database. A species with an average marker gene coverage ≥ 3 was considered present for the purposes of building a per-host database for mapping reads to infer SNVs and CNVs below. The per-host database was constructed by including all species present at one or more timepoints with coverage ≥3. However, more stringent thresholds were imposed for calling SNVs and CNVs, as described below.

Estimation of copy number variation

To estimate gene copy number variation (CNV), we mapped reads to the pangenomes of species present in a host's personal database using Bowtie2 (Langmead and Salzberg 2012) with default MIDAS settings (local alignment, MAPID≥94.0%, READQ≥20, and ALN_COV≥0.75). Each gene's coverage was estimated by dividing the total number of reads mapped to a given gene by the gene length. These genes included the aforementioned 15 universal single-copy marker genes. A given gene's copy number (c) was estimated by taking the ratio of its coverage and the median coverage of the species' single-copy marker genes.

With these copy number values, we estimated the prevalence of genes in the between-host population, defined as the fraction of samples with copy number $c \leq 3$ and $c \geq 0.3$ (conditional on the mean single gene marker coverage being ≥ 5x). For each species, we computed "core genes", defined as genes in the MIDAS reference database that are present in at least 90% of samples within a given cohort. Within-host polymorphism rates were computed in core genes.

Orthologous genes present in multiple species can result in read "stealing" and read "donating" to species from which the reads did not originate. Thus, we excluded a set of genes belonging to a 'blacklist' composed of genes present in multiple species. This blacklist was constructed in (Garud et al. 2019) using USEARCH (Edgar 2010) to cluster all genes in human-associated reference genomes with a 95% nucleotide identity threshold. Since some genes may

be absent from the MIDAS database but may nevertheless be shared across species, we implemented another filter (as in Garud et al. 2019) in which genes with $c \geq 3$ in at least one sample in our cohort were excluded from analysis of polymorphism rate or gene changes over time.

Inferring single nucleotide variants (SNVs) within bacterial species

To call SNVs, we mapped reads to a single representative reference genome as per the default MIDAS software. Reads were mapped with Bowtie2, with default MIDAS mapping thresholds: global alignment, MAPID≥94.0%, READQ≥20, ALN_COV≥0.75, and MAPQ≥20. Species were excluded from further analysis if reads mapped to ≤ 40% of their genome. We additionally excluded samples from further analysis if they had low median read coverage ($\underline{D}$) at protein coding sites. Specifically, samples with $\underline{D} < 5$ across all protein coding sites with nonzero coverage were excluded. This MIDAS SNV output was then used for computing within-species polymorphism rates and inferring the number of strains present for each species in each sample (see below).

To compute polymorphism rates, additional bioinformatic filters were imposed to avoid read stealing and donating across different species. First, we did not call SNVs in blacklisted genes present in multiple species. Additionally, we excluded sites in a given sample if $D < 0.3\underline{D}$ or $D > 3\underline{D}$ as these sites harbor anomalously low or high coverage compared to the genome-wide average $\underline{D}$. Additional filters are described below.

Shannon diversity, species richness and polymorphism rate calculations

Shannon diversity and richness were computed within each sample by including any species with abundance greater than zero. Rarefied species richness estimates are based on HMP1-2 samples rarefied to 20 million reads and Poyet samples rarefied to 5 million reads. SNV and gene content variation within a focal species were ascertained only from the full dataset and not the rarefied dataset.

The polymorphism rate of a species in a sample was computed as the proportion of synonymous sites in core genes with intermediate allele frequencies ($0.2 \leq f \leq 0.8$), as was done

in Garud et al. 2019. Only species with a MIDAS marker gene coverage of 10 or more in 10 or more samples were included, yielding 69 species in 249 HMP stool donors and 15 species in four Poyet et al. 2019 donors. As explained in SI text 1 in Garud et al. 2019, this is quantitatively similar to the more traditional population genetic measure of heterozygosity, $H=E[2f(1-f)]$, in which intermediate frequency alleles contribute the most weight to heterozygosity. By computing polymorphism with the criteria $0.2 \leq f \leq 0.8$, we avoid inclusion of low frequency sequencing errors, which can otherwise greatly influence the mean heterozygosity. Polymorphism rates were computed separately for synonymous (4-fold degenerate) and nonsynonymous (1-fold degenerate) sites. The degeneracy of sites was determined based on MIDAS output.

Temporal changes in polymorphism rates and gene content

Polymorphism change was computed as the difference in polymorphism rates between time points within a host. Gene gains and losses between time points were computed in species with sufficient prevalence (at least 10 samples with marker gene coverage of at least 10, as in the polymorphism analysis above) by identifying genes with copy number $c \leq 0.05$ (indicating gene absence) in one sample and $0.6 \leq c \leq 1.2$ (with marker coverage ≥20x) in another (indicating single copy gene presence). These thresholds were used in Garud et al. 2019 when inferring gene changes in temporal data and reflect a range of copy numbers expected in either the absence of a gene or presence of a single copy of a gene given typical coverage values in growing cells (Korem et al. 2015). These copy number cutoffs were chosen to avoid spuriously analyzing genes linked to multiple species. In such cases, mapping artifacts in which reads can be arbitrarily assigned to multiple species cannot be disentangled. For example, a gene present in multiple species would likely have copy number significantly deviating from 1 (including values that lie in an ambiguous zone of 0.05 to 0.6, as well as >>1), reflecting the joint abundances of the multiple species. Thus, although we may miss many biologically interesting multi-copy genes (e.g. transporter genes in Bacteroides (Wexler and Goodman 2017), our thresholds avoid confounding our analysis with read stealing or donating among different species. Filters for coverage and blacklisted genes were applied as described above.

Strain number inference

We used StrainFinder (Smillie et al. 2018) to infer the number of strains present within each species in each HMP1-2 metagenomic sample. To do so, we used allele frequencies from MIDAS SNV output, generated as described above. For each species in each host, all multi-allelic sites with coverage of 20x or greater were passed as input to StrainFinder. Species/host pairs which had fewer than 100 sites with 20x coverage were removed from the analysis. StrainFinder was then run on each sample separately for strain number 1, 2, 3, and 4, and the optimal strain number was chosen based on the Bayesian Information Criterion (BIC). This range of strain number was chosen for biological reasons. A number of studies have demonstrated that at most a small handful of strains (between one and four) not sharing a common ancestor within the host are ever observed within a single gut microbiome at any one time (Garud et al. 2019; Truong et al. 2017; Verster et al. 2017; Yassour et al. 2018). Additionally, for the four densely longitudinally sampled hosts in Poyet et al. 2019, multiple analyses employing distinct sequencing strategies and strain phasing techniques have similarly concluded that a maximum of four strains were present at any one time within a host for the ~30 most prevalent species (Poyet et al. 2019; Wolff, Shoemaker, and Garud 2021; Zheng et al. 2022). Thus, four strains were chosen as the maximum to accommodate the range of observed possibilities.

Statistical analyses

Model construction and evaluation

Using data from the Human Microbiome Project (HMP) and Poyet et al. 2019, we examined the relationship between within-species genetic diversity and the gut microbiome community diversity. Within-species genetic diversity was estimated with polymorphism rate and strain richness. Community diversity was estimated with the Shannon index, species richness estimated on the whole data, and species richness calculated on the data rarefied to an equal number of reads per sample (as described above). Generalized additive mixed models (GAMs) (mgcv function from the mgcv R package - RStudio version 1.2.5042) were used for most analyses, except when the response data were counts, such as the number of strains, gene gains or gene

losses. In these cases, we used generalized linear mixed models (GLMMs) (glmmTMB function from the glmmTMB R package - RStudio version 1.2.5042). GLMMs are currently more flexible than GAMs in the range of count models that it can fit (https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html). glmmTMB can deal with overdispersion in count data via two versions of negative binomial distributions (negative binomial1 and negative binomial2, respectively with linear and quadratic parameterization (Hardin and Hilbe 2018), and can handle zero-truncated count data (Shonkwiler 2016) with truncated Poisson and truncated negative binomial for both linear and quadratic parameterizations. In our case, strain count is an overdispersed positive variable, so a zero-truncated distribution was needed. We fit three different GLMMs with truncated-Poisson, truncated-negative binomial1 and truncated-negative binomial2, and then selected the best model based on the Akaike Information Criterium score (AIC) as described in (Brooks et al. 2017). The same methods were used to fit GLMMs for gene gains and losses.

To account for variation in sequencing depth, which can affect estimates of both community diversity and within-species genetic diversity, we added read count per sample (coverage) as a covariate to all generalized mixed models. Species name, subject identifier and sample identifier were added as random effects to account for variation between different species and subjects, and to account for non-independence between observations. The R syntax and statistics of all generalized models are reported in Supplementary File 2.

In GLMMs, the predictors were standardized to zero mean and unit variance before analyses. We first assessed random effects significance by comparing nested models where each random effect was dropped one at a time using the likelihood-ratio test (LRT, anova function from the R stats package) and only significant random effects were included in the final models. We then assessed the fixed effects' significance with likelihood-ratio tests implemented in the drop1 function in the R stats package. This function drops individual terms from the full model and reports the AIC and the LRT *P*-value. All the *P*-values reported for the GLMMs correspond to LRT and not to the Wald *P*-values reported by glmm.summary function from the R package glmmTMB, as was recommended in https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html. We again used LRTs to compare the full significant models to null models including all random effects but no fixed effects other than the intercept. The difference in Akaike

information criterion (ΔAIC) between full and null model and their associated *P*-values are reported in Supplementary File 1e,f,g. As an additional evaluation of the goodness of fits, we estimated the coefficient of determination (R^2) using the r2 function from the performance R package. Two values are reported: the marginal R^2, a measure of the variance explained only by fixed effects, and the conditional R^2, a measure of the variance explained by the entire model.

We evaluated model fits by inspecting the residuals using the DHARMa library in R (simulateResiduals and plot functions) for the GLMMs and by inspecting residual distributions and fitted-observed value plots using the gam.check function from the mgcv R package for the GAMs. Adjusted R^2 values (from gam.summary function from the mgcv R package) are reported as a goodness of fit for the GAMs. All model outputs (summary function from mgcv and glmmTMB R packages) are reported in the Supplementary File 2.

To study the relationship between focal species polymorphism and community diversity calculated at higher taxonomic ranks (from genus to phylum), we used GTDBK and the Genome Taxonomy Database (GTDB) (Chaumeil et al. 2020) to annotate MIDAS reference genomes. Richness at each level was estimated with the total number of distinct taxonomic units in the sample. The Shannon index was calculated based on the relative abundances table from MIDAS: at each taxonomic level, we used the sum of the abundances of all species belonging to that taxonomic level to calculate the Shannon index (using the diversity function from the R vegan library). We then fit two separate GAMs for each taxonomic rank (from genus to phylum) with either Shannon diversity or richness as the predictors of within-species polymorphism rate (with the coverage per sample as a covariate and species name, sample and subject identifiers as random effects). These GAMs were fitted with a beta error distribution with logit-link function, chosen because polymorphism rate is a continuous value strictly bounded by 1, and all the terms were smoothed terms (See Supplementary File 1c and Supplementary File 2 section 1-3 for additional model details).

We repeated the same methods for focal species synonymous and nonsynonymous polymorphism separately. See Supplementary File 1b and d and Supplementary File 2 section 4-6 for details of the models applied to nonsynonymous polymorphism.

Analysis of strain counts per focal species

To study the relationship between community diversity and the number of strains within a focal species in the HMP, we restricted the analysis to 184 focal species genomes with at least 100 nucleotide sites with 20x coverage in a sample. We fit separate GLMMs with strain count in a focal species as a function of community diversity estimated with Shannon diversity, species richness, or rarefied species richness. Since strain number is positive count data, we compared alternative zero-truncated count models based on the Akaike information criterion (AIC) score (AICtab function from bbmle R library) (Brooks et al. 2017). We fit the model with the truncated negative binomial distribution (truncated_nbinom2 or truncated_nbinom1 in glmmTMB; the second best fit) in order to resolve the overdispersion detected in the best fit (the truncated Poisson model). Overdispersion was tested using the check_overdispersion function from the performance R package as described here: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html.

As described above for focal species polymorphism, we tested the relationship between focal species strain count and community diversity at higher taxonomic levels from genus to phylum, fitting a separate GLMM with strain count in a focal species as a function of each metric of diversity (Shannon and richness) at higher taxonomic levels (from genus to phylum). All GLMM details are reported in Supplementary File 1f and Supplementary File 2 section 7-9.

Analysis of time series data

To test the predictions of DBD over time, we used HMP samples with multiple time points from the same person to look at the relationship between within-species polymorphism change, defined as the difference between polymorphism rate at two time points, and community diversity at the earlier time point. We fit GAMs with log transformed polymorphism change as a function of community diversity at the earlier time point, and added the coverage per sample at the earlier time point as a covariate as well as species name, sample and subject identifiers as random effects (Supplementary File 2 section 10.1).

In addition, we investigated the effect of community diversity at one time point on gene content variation (gains and losses considered separately) at the subsequent time point. Gene gains and losses were both overdispersed count data, so we selected the best negative binomial

model (between linear and quadratic parameterization) based on the AIC, and fit separate negative binomial GLMMs with gene gain as the response and each of the metrics of community diversity as the predictor, with the same covariates and random effects used in the previous models (Supplementary File 2 section 10.2). The same method was used to test how gene loss was related to community diversity (Supplementary File 1g, Supplementary File 2 section 10.3).

HMP longitudinal data consisted of hosts sampled at a time lag of ~6 months. To assess the relationship between within-species genetic diversity and community diversity at higher temporal resolution, we used the same methods to analyze longitudinal metagenomic data from four more frequently sampled healthy stool donors (hosts *am*, *an*, *ao* and *ae*) (Poyet et al. 2019). Stool from donor *am* was sequenced over 18 months with a median of one day between samples; *an* over 12 months (median 2 days between samples); *ao* over 5 months (median 1 day between samples); and *ae* over 7 months (median 2 days between samples). We looked at polymorphism change and gene gains and losses between two time points in the 15 species with a minimal marker gene coverage of ten in at least ten samples. Community diversity was estimated with Shannon diversity (unrarefied) and richness calculated on rarefied data to 5 million reads per sample.

We used the same methods as in HMP time series to study the relationship between community diversity at the initial time point and polymorphism change between the initial time point and all the future time points. We fit Gaussian generalized additive mixed models with log-transformed polymorphism change as the response and the interaction between community diversity at the first time point and the number of days between time points as the predictor. Covariates included coverage, species name, sample, and subject identifiers as random effects (Supplementary File 1h, Supplementary File 2 section 11.1 and 11.2). To study the relationship between gene variation (gains and losses separately) and diversity at the first time point, we fit negative binomial generalized linear mixed models with gene variation as a function of the interaction between diversity at the first time point and the number of days between the two time points, with the same covariate and random effects as used above for polymorphism change over time (Supplementary File 1i, Supplementary File 2 section 11.3-11.6).

Acknowledgements

We sincerely thank members of the Garud and Shapiro labs, and Pleuni Pennings, for their feedback during the development of this paper. NRG received support from the Paul Allen Frontiers Group, a University of California Hellman fellowship, a UCLA Faculty Career Development award, and the Research Corporation for Science Advancement. DWC received funding support from NIH R25 MH 109172. BJS was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Canada Research Chair. We also thank Djordje Bajić and two anonymous reviewers for their constructive suggestions that substantially improved the manuscript.

Supplementary figures

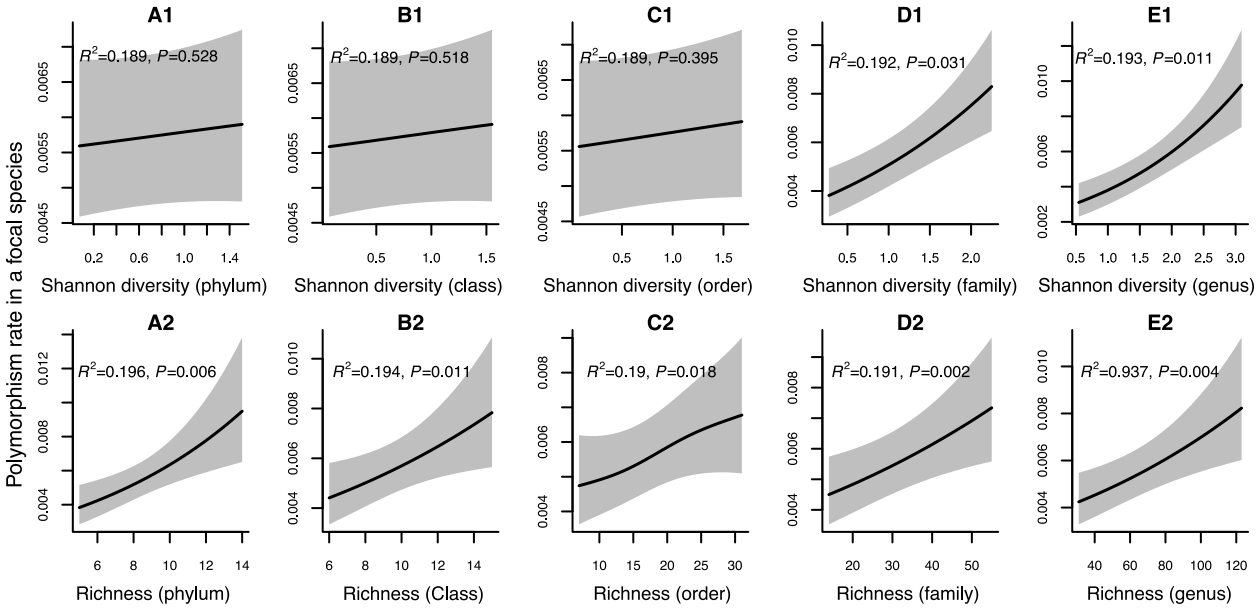

Figure S1. Results of generalized additive models predicting within-species polymorphism rate (at synonymous sites) as a function of community diversity at higher taxonomic levels (HMP data)*. (A1-E1) The predictor is Shannon diversity. (A2-E2) The predictor is richness. Adjusted r-squared (R^2) and Chi-squared P-values corresponding to the predictor are displayed in each panel (gam.summary function from mgcv R package). Shaded areas show the 95% confidence interval of each model prediction. See Supplementary File 1c and supplementary file 2 sections 2-3 for further details about model outputs.*

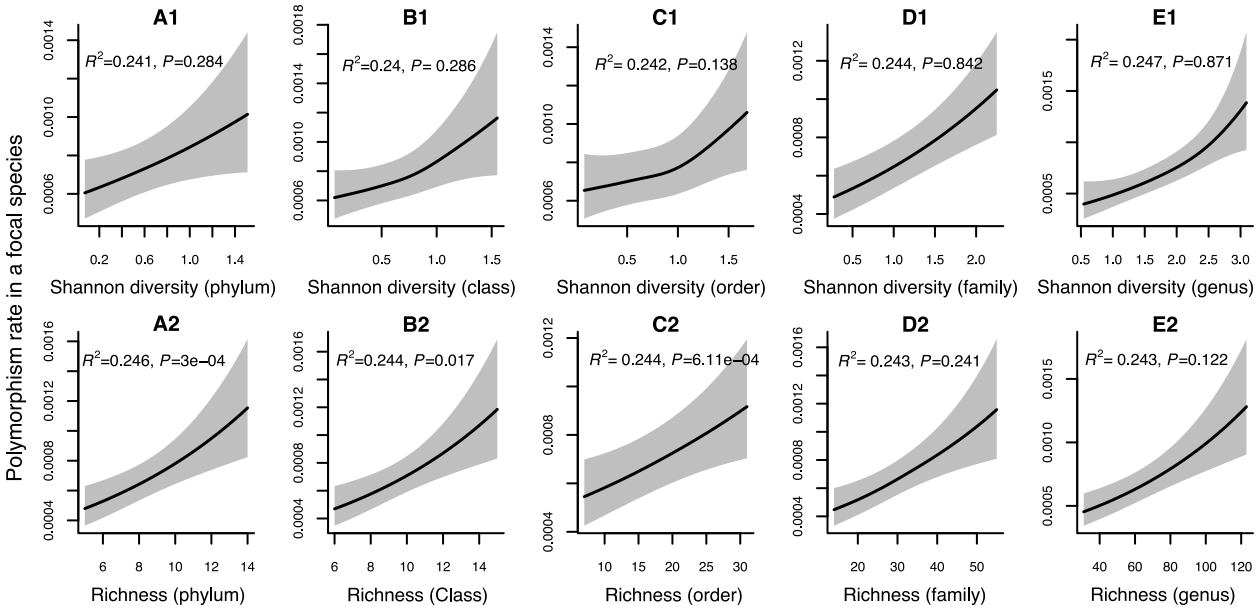

Figure S2. Results of generalized additive models predicting within-species polymorphism rate (at nonsynonymous sites) in a focal species as a function of community diversity at higher taxonomic levels (HMP data). *(A1-E1) The predictor is Shannon diversity. (A2-E2) The predictor is richness. Adjusted r-squared (R^2) and Chi-squared P-values corresponding to the predictor are displayed in each panel (gam.summary function from mgcv R package). Shaded areas show the 95% confidence interval of each model prediction. See Supplementary File 1d and supplementary file 2 sections 5-6 for further details about model outputs.*

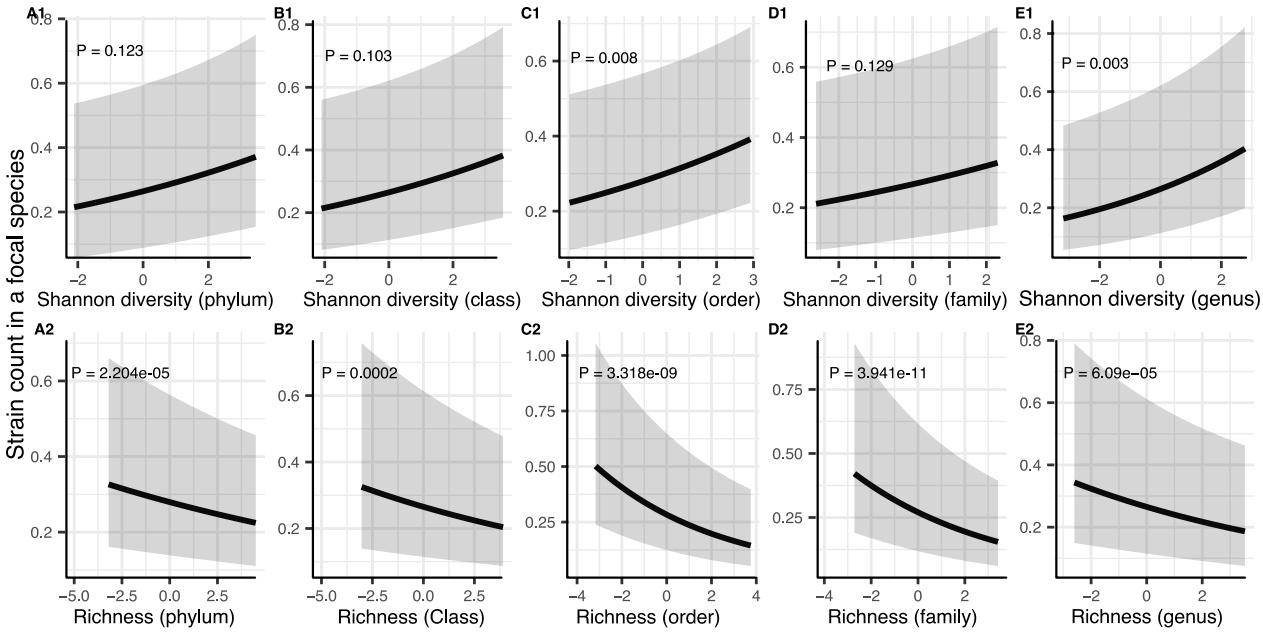

Figure S3. Results of generalized linear mixed models predicting strain count in a focal species as a function of community diversity at higher taxonomic levels (HMP data). *Strain number in a focal species is positively correlated with Shannon (A1-E1) whereas its correlation with richness remains negative (A2-E2) through all taxonomic levels. The Y-axis is the predicted mean number of strains within a focal species. P-values (drop1 function from R stats package, LRT). Shaded areas show the 95% confidence interval of each model prediction. See Supplementary File 1f and supplementary file 2 section 9 for model details.*

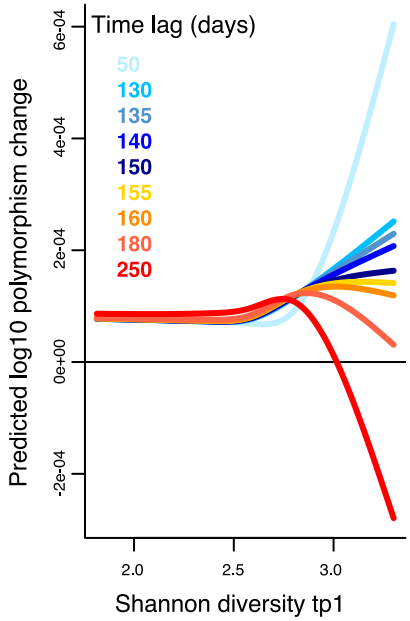

Figure S4. Results of a GAM predicting polymorphism change in a focal species as a function of the interaction between Shannon diversity at the first time point and the time lag (days) between two time points in the Poyet time series. *The response (Y-axis) was log transformed in the Gaussian GAM. Several different time lags are shown to illustrate the inversion of the relationship around a lag time of 150 days. See Supplementary File 1h and supplementary file 2 section 11 for further model details.*

Chapter3. Phage predation and antibiotic exposure are inversely associated with disease severity and shape pathogen genetic diversity in cholera patients

N. Madi[1,2], E. T. Cato[3], Md. Abu Sayeed[3], K. Islam[4], Md. I.UI. Khabir[4], Md. T. R. Bhuiyan[4], Y. Begum[4], A. Creasy-Marrazzo[3], M. Kamat[4], A. Cuénod[1], L. S. Bailey[5], K. B. Basso[5], F. Qadri[4], A. I. Khan[4*], B. J. Shapiro[1,2,6,7*], E. J. Nelson[3*]

[1] Department of Microbiology & Immunology, McGill University, Montréal, Canada
[2] Département de Sciences Biologiques, Université de Montréal, Montréal, Canada
[3] Departments of Pediatrics and Environmental and Global Health, University of Florida, Gainesville, FL, USA
[4] Infectious Diseases Division (IDD) & Nutrition and Clinical Services Division (NCSD), International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, Bangladesh
[5] Department of Chemistry, University of Florida, Gainesville, FL, USA
[6] McGill Genome Centre, McGill University, Montréal, Canada
[7] McGill Centre for Microbiome Research, McGill University, Montréal, Canada

[*] Corresponding authors

Email: eric.nelson@ufl.edu; jesse.shapiro@mcgill.ca; ashrafk@icddrb.org

Keywords: Cholera, diarrhoea, diarrhea, *Vibrio cholerae*, bacteriophage, phage, phage resistance, antibiotics, antimicrobial resistance, AMR, predator, prey, metagenomics

One sentence summary: A metagenomic survey of cholera patients in Bangladesh identifies phages and antibiotics as markers of improved clinical outcomes and as selective pressures favoring resistance mechanisms.

ABSTRACT

A century has passed since the discovery that virulent bacteriophages are associated with increased survival among cholera patients. Despite an increasingly detailed picture of the molecular mechanisms that influence phage-bacterial interactions, we lack an understanding of how these interactions impact disease severity. Here we report a year-long, nationwide study of diarrheal disease patients in Bangladesh. We quantified relative abundances of *Vibrio cholerae* (*Vc*) and associated phages using metagenomics while accounting for antibiotic exposure using quantitative mass spectrometry. Both phages and antibiotics suppressed *Vc* and were inversely associated with severe dehydration; these effects were dampened by resistance mechanisms. In the absence of phage resistance genes, phage (ICP1) selected for nonsynonymous point mutations that likely have evolutionary consequences. Our results point to a hierarchy of phages and antibiotics serving as key selective pressures and suggest the ratio of phage to pathogen can serve as a biomarker of disease severity in humans.

INTRODUCTION

The secretory diarrheal disease cholera is caused by the Gram-negative bacterium *Vibrio cholerae* (*Vc*) and can progress to life-threatening hypovolemic shock in less than 12 hours (Andrews et al. 2017). Cholera remains a major public health problem due to inadequate sanitation and restricted access to safe drinking water. Global estimates of cholera burden are 1.3-4.0 million cases and 21,000 to 143,000 deaths annually (Ali et al. 2015). As of January 2023, there were over 30 countries with active outbreaks necessitating the WHO to escalate the response to its highest level (Madi, Chen, et al. 2023). Rehydration is the primary life-saving maneuver for cholera patients. With effective rehydration, mortality rates fall from over 20% to less than 1%; antibiotics reduce stool volume and duration of diarrhea but are considered supportive and reserved for patients with more severe disease to reduce selection for antibiotic resistance (Islam et al. 1995; Dromigny et al. 2002; Weill et al. 2017; Nelson et al. 2009; Nelson et al. 2011). Nevertheless, antibiotic-resistant *V. cholerae* has emerged and spread globally (Weill et al. 2017; Das et al. 2020; Lassalle et al. 2022). Mechanisms of resistance are diverse and reside in the core genome and on mobile genetic elements, including the SXT integrative conjugative element (ICE) (Rivard, Colwell, and Burrus 2020). SXT is a ~100 kb ICE that can harbor resistance to sulfamethoxazole and trimethoprim, ciprofloxacin (qnr_{vc}), trimethoprim (*dfra31*) and streptomycin (*aph(6)*) (Waldor, Tschäpe, and Mekalanos 1996; Dalsgaard et al. 1999; Beaber, Hochhut, and Waldor 2002; Creasy-Marrazzo et al. 2022). Recent work has also shown that the ICE can encode diverse phage resistance mechanisms (LeGault et al. 2021).

V. cholerae is targeted by numerous virulent bacteriophages (phages) (Nelson et al. 2008; Boyd et al. 2021; Seed et al. 2011). With rising levels of antibiotic resistance, phages are a promising alternative or complementary therapy. The first clinical trials of phage therapy cocktails occurred during the Cholera Phage Inquiry from 1927 to 1936 (Summers 1993; D'Herelle and Malone 1927). In a *proto* randomized controlled trial, the Inquiry found the odds of mortality were reduced by 58% among those with phage therapy, with an absolute reduction in mortality of 10% (95% CI 0.27-0.64) (Pasricha, de Monte, and O'Flynn 1936) (reanalyses by E. J. Nelson). Despite these early findings, there is a lack of evidence in the modern era that links phage predation with disease severity during *natural* infection in humans. A collection of tangentially

related studies support the hypothesis that phages mitigate severity: Environmentally, phages in the aquatic environment are negatively correlated with cholera cases in Dhaka, Bangladesh over time, suggesting a role for phages in influencing epidemic dynamics (Faruque, Naser, et al. 2005). Clinically, a higher percentage of cholera patients shed phage towards the end of an outbreak (Faruque, Islam, et al. 2005). Computationally, models predict that phage can dampen outbreaks (Jensen et al. 2006b). Experimentally, animal studies found phage exposure was negatively associated with burden of colonization and disease severity (Nelson et al. 2008; Zahid et al. 2008; Jaiswal et al. 2013b; Yen, Cairns, and Camilli 2017). In this context, a key unanswered mechanistic question is how phage, antibiotics, and associated resistance factors interact to impact disease severity.

To address this question, we conducted a national prospective longitudinal study in the cholera endemic setting of Bangladesh. Stool samples were collected at hospital admission from diarrheal disease patients and screened for *V. cholerae* or *V. cholerae* phage. From positive samples, we sequenced shotgun metagenomes to quantify the relative abundances of *V. cholerae,* phages, and gut microbiome taxa, while accounting for antibiotic exposure determined by mass spectrometry. We demonstrate that severe dehydration is inversely associated with phage ICP1. We also find evidence for non-linear predator-prey interactions within infected hosts, but these interactions are suppressed at high levels of azithromycin. Known phage-resistance elements (ICEs) are associated with lower phage:*V. cholerae* ratios. In the absence of detectable ICEs, phages select for nonsynonymous point mutations in the *V. cholerae* genome, with potential evolutionary consequences. Together, our results support a hierarchy of selective pressures and resistance mechanisms evolving within individual cholera patients.

RESULTS & DISCUSSION

Study overview

A total of 2574 stool samples were collected from enrolled participants from March to December 2018; collection continued to April 2019 at one site (icddr,b). A total of 282 samples (10.9%) were culture-positive for *V. cholerae*. Among culture-negative samples, 10% of samples

were randomly selected and screened by PCR for *V. cholerae* or phages (ICP1,2,3) which generated an additional 107 positive samples (**Table S1**). Stool metagenomes were successfully sequenced from 88.4% (344/389) of these samples; 35% (122) were from the icddr,b. Detection rates for *V. cholerae,* ICP1, ICP2, and ICP3 were 55% (190), 18% (61), 1% (4) and 8% (28), respectively. Select antibiotics (**Table S2**) were quantified in stool using liquid chromatography-mass spectrometry.

Metagenomic correlates of disease severity and succession

In the context of enrollment at hospital admission, we expected to sample patients at different stages in their disease progression, beginning with high levels of *Vc* followed by *E. coli* and, in most cases, a return to a normal microbiome in which anaerobic bacteria dominate (David et al. 2015). We hypothesized that this ecological succession would be accompanied by a progression from severe to mild dehydration. Consistent with this hypothesis, we identified *Vc* as an indicator species of severe dehydration (Methods). *Vc was* more likely to be present and higher in relative abundance in patients with severe dehydration, while two *Bifidobacterium* species and *E. coli* were indicators of mild dehydration (**Figure 1A, Table S3**). In a time series of cholera patients, the phage ICP1 peaked on the first day of infection (David et al. 2015) yet intriguingly ICP1 was an indicator of mild dehydration in our analysis (**Figure 1A**). This contrasts with ICP3, which despite being less frequently detected in our study (28 samples with >0.1% ICP3 reads, compared to 61 samples with >0.1% ICP1), was an indicator of severe dehydration. This suggests that different phages can have contrasting disease associations. For subsequent analyses, we focus on the more prevalent ICP1 phage.

Although the distribution of ICP1 relative abundance is variable and less clearly associated with dehydration status than *Vc* (**Figure S1**), higher ratios of ICP1 to *Vc* were associated with mild dehydration (**Figure 1B**). This ratio is particularly informative, as it provides insight into how strongly ICP1 might suppress its bacterial host. These analyses suggest that ICP1 plays a role in modulating disease severity, or that it could be associated with later stages of diseases than previously thought. Our study design prevents us from confidently disentangling these possibilities; however, we recorded self-reported duration of diarrhea, providing an approximate

control for disease progression. We found that higher relative abundances of ICP1 were associated with mild dehydration at the early stages of disease (duration of diarrhea <72h) but not at late stages (**Figure S1B and D**). We therefore cannot exclude a model in which ICP1 suppresses *Vc* at early disease stages to reduce disease severity; nor can we exclude ICP1 as a non-causal biomarker of dehydration severity. In either case, ICP1 could provide a clinically useful indicator of disease severity.

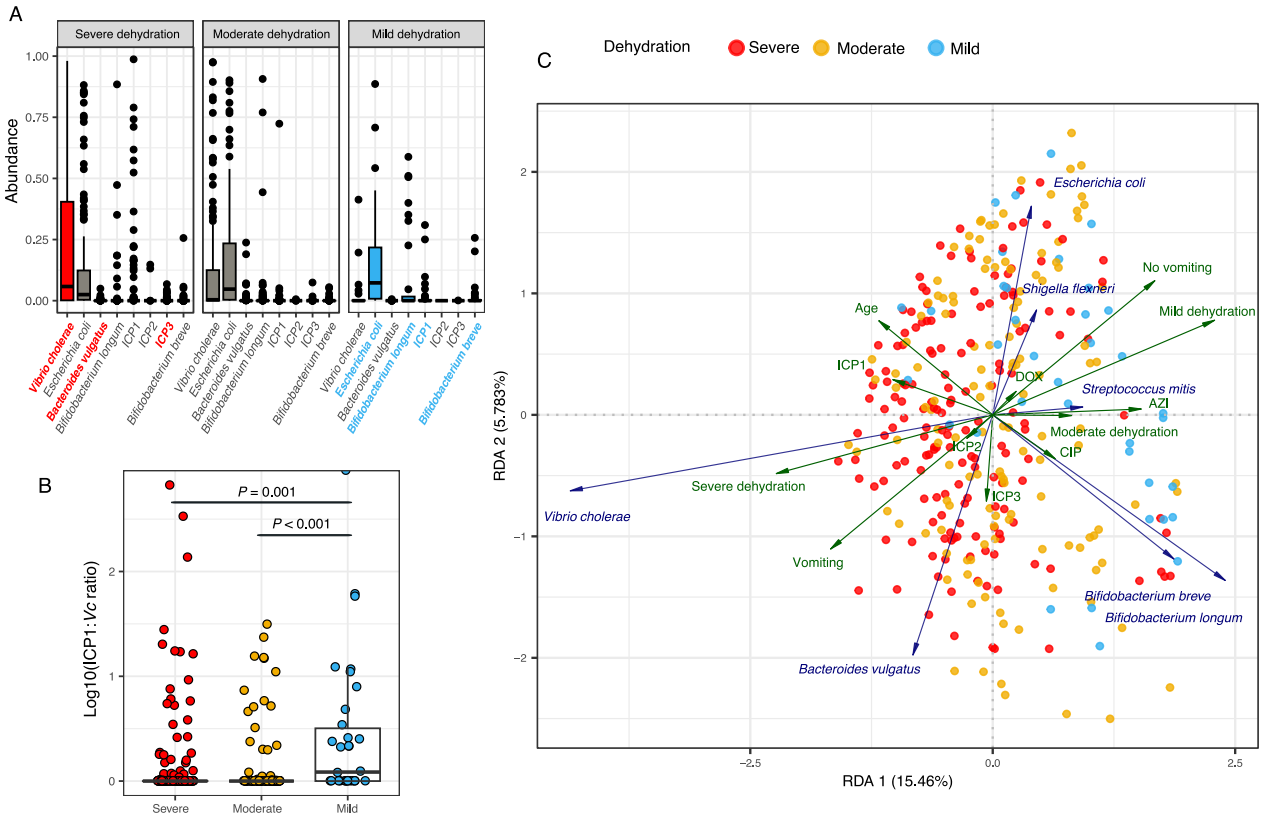

Figure 1. Dehydration severity is inversely associated with higher ICP1:Vc ratios in stool metagenomes. *(A) Relative abundances of the most prevalent species in patients with severe, moderate, or mild dehydration; these conventions equate to the World Health Organization (WHO) conventions of severe, some and no dehydration, respectively. Significant indicator species for severe or mild dehydration are shown in red or blue bold text, respectively. See Table S3 for full indicator species results. (B) The ICP1:V. cholerae ratio is higher in patients with mild dehydration. P-values are from a Kruskal-Wallis test with Dunn's post-hoc test, adjusted for multiple tests using the Benjamini-Hochberg (BH) method. Only significant P-values (<0.05) are shown. Only 323 samples with V. cholerae>0% of metagenomic reads were included, with 165 from severe, 128 from moderate, and 30 from mild cases. A pseudocount of one was added to the ratio before log transformation. In (A) and (B) the solid horizontal line is the median and the boxed area is the interquartile range. (C) Redundancy analysis (RDA) showing relationships among the seven most dominant bacterial species and explanatory variables: phages, patient metadata, and antibiotic concentration. Angles between variables (arrows) reflect their correlations; arrow length is proportional to effect size. Samples (points) are colored by dehydration severity. All displayed variables are significant (P<0.05, permutation test) except ICP2, ICP3, and*

doxycycline (Table S4). For the RDA: R^2=0.25 and adjusted R^2=0.184, permutation test P = 0.001). To improve readability, collection date and location are not shown (see Figure S3 for these details).

Antibiotic exposure is associated with mild disease and resistance genes

To visualize the complex relationships between disease severity, bacteria, and phage in the context of antibiotic exposure, we used redundancy analysis (RDA; **Figure 1C; Table S4**). For simplicity, the seven most dominant bacterial species identified by principal components analysis were included (**Figure S2**). As explanatory variables, we used the most important clinical data, chosen by forward selection beginning with phages and antibiotic concentrations (**Figure S3**). In accordance with the indicator species analysis (**Figure 1A, Table S3**), *Vc* was positively correlated with severe dehydration (**Figure 1C**). ICP1 was moderately associated with *Vc*, consistent with a phage's reliance on its host for replication, but less associated with severe dehydration. Both azithromycin (AZI) and ciprofloxacin (CIP) were negatively correlated with *V. cholerae* and severe dehydration, suggesting their role in suppressing cholera infection and disease.

To test the hypothesis that antibiotic exposure selects for antibiotic resistance genes (ARGs), we assessed the relative abundances of 634 known ARGs conferring resistance to 34 antibiotic classes using deepARG (Arango-Argoty et al. 2018) applied to the metagenomic data. To identify associations between these ARGs and gut microbes, we ran a multiple factor analysis (MFA) with the 37 most dominant species from a PCA (20 highest coordinates on both axes, Methods) and all the ARGs. ARGs were primarily associated with three different taxonomic clusters dominated by (i) *V. cholerae*, (ii) *Escherichia* spp. and *Shigella* spp., and (iii) *Prevotella* spp. and *Bifidobacterium* spp. (**Figure S4**). The *V. cholerae* cluster was associated with 12 ARGs: *ompT, ompU*, *varG, dfrA1, dfrA16, aph3, aph6, mexI, floR, tet34, tet35*, and a chloramphenicol exporter. Several of these targets were previously linked to antibiotic resistance in *V. cholerae (Creasy-Marrazzo et al. 2022; Monir et al. 2023)*. To determine if these ARGs correlate with antibiotic exposure, we compared the relative abundance of ARGs (normalized to *V. cholerae* reads) at different levels of antibiotic exposure: above or below the minimum inhibitory concentration (MIC) set by experiments under aerobic or anaerobic conditions (**Table S5**) (Methods). CIP exposure was associated with higher relative abundance of *dfrA16* (Wilcoxon test,

BH corrected *P*=0.0178 and 0.0026 for anaerobic and aerobic MIC, respectively) and *aph6* (BH corrected *P*=0.084 for anaerobic and 0.0552 for aerobic MIC) (**Figure S5)**. No significant correlations were observed with azithromycin or doxycycline. Together, these results provide evidence that exposure to certain antibiotics selects for resistance genes in the human gut.

Predator-prey dynamics between phage and pathogen

Offset oscillations between *Vc* and its phages in the aquatic environment have been observed at the broader population level in Bangladesh, supporting the hypothesis that these predator-prey dynamics influence the progression of epidemics (Faruque, Naser, et al. 2005; Jensen et al. 2006b). This hypothesis has not been revisited in nearly two decades, and it remains unclear if the dynamics of aquatic phages are mirrored by phages within infected patients. To address these questions, we focused on a one-year time series from our most sampled site, icddr,b, in Dhaka, and tracked the mean relative abundances of *Vc* and ICP1 within patients over time. The abundance of *V. cholerae* was high from March through July 2018; ICP1 abundance was low in March 2018 at the beginning of the epidemic, increased in July, and spiked in August as *V. cholerae* declined (**Figure 2A, B**). ICP1 and *V. cholerae* were inversely correlated from September 2018 onward, with a spike in *V. cholerae* corresponding to a decline of ICP1 in September, and a peak in ICP1 associated with a decline of *V. cholerae* in November. *V. cholerae* and ICP1 were inversely correlated within patients over time (Spearman r=-0.25, P=0.017; **Figure 2C**), mirroring the trend previously observed between recorded cholera cases and environmental phage concentrations (Faruque, Naser, et al. 2005). Sampling sites outside Dhaka had sparser time series, but also generally supported alternating peaks of *Vc* and ICP1 (**Figure S6**). Our results suggest that phages may contribute to suppressing *Vc* not just in the environment, but also within patients. The degree of coupling between phage concentrations in the environment and within patients could therefore be relevant to epidemic dynamics and deserves further study.

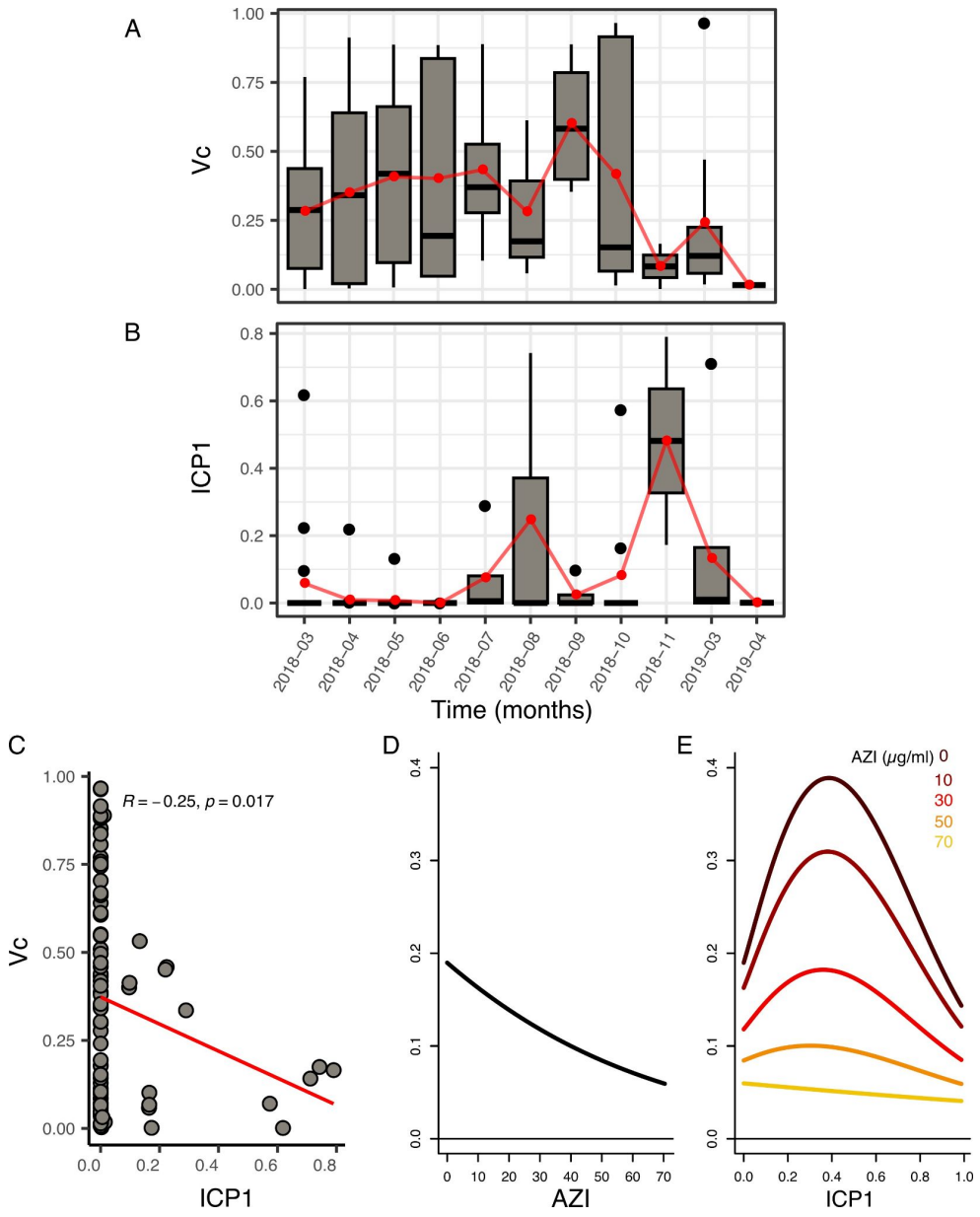

Figure 2. Offset temporal dynamics of V. cholerae and phage ICP1*. Relative abundances of (A) V. cholerae and (B) ICP1 over time (binned by month) in metagenomes sampled from the icddr,b (94 samples with Vc>0.5% or ICP1>0.1% of metagenomic reads). Red line shows the mean. The boxplots contain 16 samples from 2018-03, 25 samples from 2018-04,17 samples from 2018-05, 5 samples from 2018-06, 4 samples from 2018-07, 3 samples from 2018-08, 4 samples from 2018-09, 9 samples from 2018-10, 2 samples from 2018-11, 8 samples from 2019-03, and 1 sample from 2019-04. (C) Spearman correlation between V. cholerae and ICP1 in the same samples as in A and B. The solid horizontal line is the median and the boxed area is the interquartile range. Panels (D) and (E) show GAM results, fit to data from all 344 samples. (D) V. cholerae declines in relative abundance with higher concentration of azithromycin (AZI). (E) The relationship between ICP1 and V. cholerae is affected by azithromycin (AZI) concentration. The illustrated AZI concentrations show regular intervals between the minimum (0 ug/ml) and maximum (78 ug/ml) observed values.*

Antibiotic exposure suppresses predator-prey dynamics

Our data provide a unique opportunity to determine if and how antibiotics interact with phage to suppress *V. cholerae* within patients. To do so, we modeled the relationships between ICP1, *V. cholerae* and antibiotic exposure within each patient. We fit a generalized additive model (GAM) of *V. cholerae* relative abundance as a function of ICP1, antibiotic exposure, and their interaction, including dehydration status as a random effect. We fit GAMs with all antibiotics and their interaction with ICP1, as well as separate models with each antibiotic, alone or in combination, and compared them based on their AIC (**Table S6**). The most parsimonious model (with the lowest AIC) included the interaction between AZI and ICP1 as a predictor of *V. cholerae* abundance. *V. cholerae* was affected by AZI both directly (GAM, P=0.00236, Chi-square test) and in interaction with ICP1 (GAM, P=0.026, Chi-square test); the effect of ICP1 alone was not significant. Consistent with previous reports that AZI suppresses *V. cholerae (32)* and with our RDA results (**Figure 1C**), we found a negative relationship between *V. cholerae* and AZI (**Figure 2D**). This is also consistent with our observation that no annotated ARGs were associated with AZI (**Figure S7**), suggesting that *V. cholerae* in our samples is generally AZI-sensitive. The relationship between *V. cholerae* and ICP1 was quadratic: at low ICP1 abundance, the relationship was positive but became negative at higher ICP1 abundance (**Figure 2E**). This alternation between positive and negative correlations is consistent with predator-prey dynamics within infected patients (Carr et al. 2019b). However, at higher concentrations of AZI, the quadratic relationship was flattened, effectively suppressing the phage-bacteria interaction.

Integrative and conjugative elements (ICEs) are associated with phage suppression

SXTs are ~100kb integrative and conjugative elements (ICEs) that have been associated with resistance to both antibiotics and phages and were first discovered in *V. cholerae (Waldor, Tschäpe, and Mekalanos 1996)*. SXTs have conserved 'core' genes along with variable 'hotspots' encoding different genes; for example, hotspot 5 is a ~17kb region associated with phage resistance. At the time of our sampling, ICEVchInd5 and ICEVchind6 were the two most prevalent ICEs in Bangladesh (LeGault et al. 2021). These ICEs differ in their anti-phage systems: ICEVchInd5

(henceforth “ind5”) encodes a type 1 bacteriophage exclusion (BREX) system while ICEVchInd6 (“ind6”) encodes several other restriction-modification systems (LeGault et al. 2021).

We enumerated ICEs in metagenomes by mapping reads against reference sequences for ind5 (NCBI accession GQ463142.1) and ind6 (accession MK165649.1). An ICE was defined as present when 90% of its length was covered by at least one metagenomic read (**Figure S8A**). We found that 59% (144/244) of the samples contained ICEVchind5, 10.6% (26/244) contained ICEVchind6, and 22.1% (54/244) had no detectable ICE. The lack of ICE detections was not due to the lack of *V. cholerae* in a metagenome because ICE-negative samples did not contain fewer *V. cholerae* reads (**Figure S8B**).

Resistance mechanisms on ICEs have been shown to suppress phage *in vitro*, but their relevance within human infection remains unclear. We found that metagenomes without a detectable ICE (denoted as ICE-) were associated with higher phage to *V. cholerae* ratios (**Figure 3**). The effect was strongest for ICP1, which had the largest sample size (**Figure 3A**). This observation is consistent with ICE-encoded mechanisms suppressing phage within patients. Higher ICP1:*Vc* ratios, which tend to occur in ICE- patients, were also associated with mild dehydration (**Figure 3B**). The phage:*Vc* ratios varied according to the precise phage-ICE combination, suggesting that different ICE hotspots might confer resistance to different phages (**Figure 3, Figure S9**). Together, these results demonstrate a role for ICEs in suppressing phages during human infections, complementing and generally confirming the predictions of laboratory experiments (LeGault et al. 2021). That said, the suppression is not complete, and further experiments are needed to dissect the underlying causal relationships.

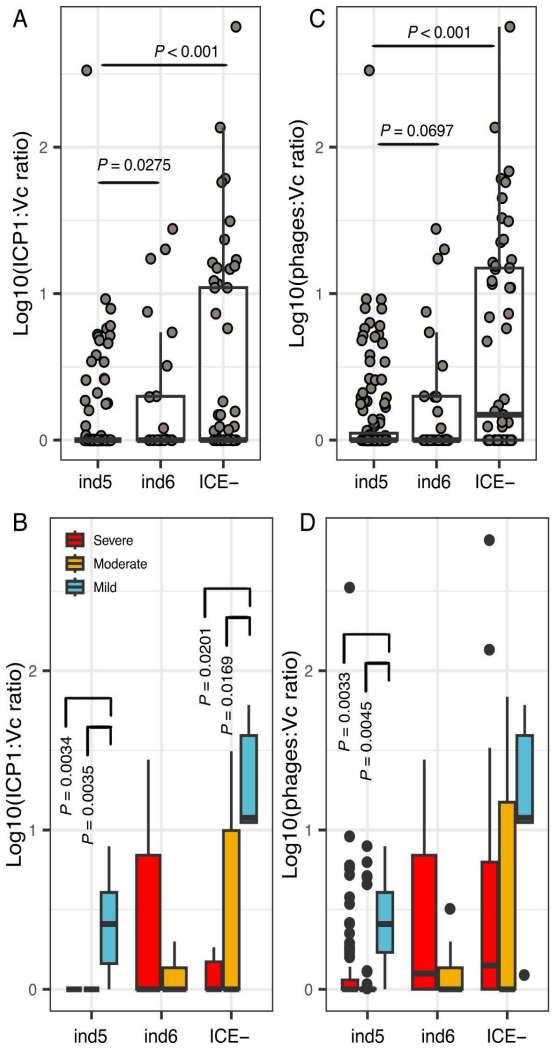

Figure 3. Integrative conjugative elements (ICEs) are associated with lower ICP1:V. cholerae ratios in patient metagenomes*. (A) Distribution of ICP1:Vc ratios across patients with different ICE profiles. (B) The same data as (A) binned into boxplots according to dehydration status. (C) Distribution of phage:Vc ratios, including the sum of all phages (ICP1, ICP2, ICP3). (D) The same data as (C) binned into boxplots according to dehydration status. P-values are from a Kruskal-Wallis test with Dunn's post-hoc test adjusted with the Benjamini-Hochberg (BH) method. Only P-values < 0.1 are shown. Only samples with sufficient Vc or ICP1 were included (224 samples with Vc>0.5% or phages >0.1% of metagenomic reads), of which 54 samples were ICE-, 26 were ind6+ and 144 were ind5+. For clarity, the Y-axes were log10 transformed after adding one to the ratios. The solid horizontal line is the median and the boxed area is the interquartile range.*

Hypermutation generates V. cholerae genetic diversity

In addition to variation in gene content in ICEs and other mobile elements, resistance to phages and antibiotics may be conferred by point mutations (single nucleotide variants; SNVs) that existed before or emerged *de novo* during infection. Although we cannot exclude mixed

infections by different *Vc* strains as a source of diversity, there was no evidence for more than one strain co-infecting a patient in our study population (**Figure S10**). We previously found a generally low level of *V. cholerae* genetic diversity within individual cholera patients(Levade et al. 2017), with the exception of hypermutation events characterized by DNA repair defects and hundreds of SNVs in the *Vc* genome, primarily transversion mutations (Levade et al. 2021). Here, with a much larger sample size, we can better quantify the frequency of hypermutator phenotypes, and test if within-host *V. cholerae* diversity is associated with phages or antibiotics – both of which could potentially select for resistance mutations. After quality filtering the metagenomic reads, we used InStrain (Olm et al. 2021) to quantify within-host *V. cholerae* diversity in 133 samples (Methods). To identify likely hypermutators, we independently counted samples with defects (nonsynonymous mutations) in any of 65 previously defined DNA repair genes (Jolivet-Gougeon et al. 2011) or a high rate of SNVs (25 or more) (Levade et al. 2021). We found that 35% of samples (47/133) had both a high SNV count and nonsynonymous mutations in DNA repair genes – making them likely to contain hypermutators within the infecting *V. cholerae* population. Higher SNV counts were significantly associated with DNA repair defects (Fisher's exact test, *P*=2.2e-16), consistent with these defects resulting in higher mutation rates within patients. The number of SNVs was not significantly confounded by *V. cholerae* genome coverage (**Figure S11A**). Consistent with our previous study (Levade et al. 2021), putative hypermutators had a distinct mutational profile, enriched in transversions (**Figure S11B**). This suggests that hypermutation is common and detectable in over one third of cholera patient stool metagenomes. We also observed that samples with high numbers of SNVs but no DNA repair defects had somewhat higher levels of phage ICP1 (**Figure S11D**). This suggested that phage could select for resistance mutations even in non-mutators. For subsequent analysis, we considered all SNVs together, regardless of whether they were generated by hypermutation or not.

Phages, not antibiotics, are associated with V. cholerae within-host diversity

We asked if *V. cholerae* within-host diversity could be predicted by phage or antibiotic concentrations – both likely strong selective pressures on bacterial populations. To do so, we fit GLMMs with phages and antibiotics as predictors of the number of high-frequency

nonsynonymous (NS) SNVs in the *V. cholerae* genome. We focused on higher-frequency SNVs (>10% within a sample) as more likely to be beneficial, and NS SNVs as those more likely to have fitness effects. We fit several models with different combinations of predictors: from a model with all antibiotics and their interaction with ICP1 to separate models with each antibiotic and its interaction with ICP1. We added *V. cholerae* abundance as a fixed effect to the model to control for any coverage bias in SNV calling (**Table S7**). The most parsimonious model included *V. cholerae* abundance and the interaction between *V. cholerae* and ICP1 as predictors of the number of high-frequency NS SNVs. Adding antibiotics or their interaction with ICP1 did not improve the model (**Table S7**), suggesting a limited role for antibiotics in selecting for point mutations within patients.

In the model, *V. cholerae* relative abundance and the interaction between *V. cholerae* and ICP1 both had significant effects (GLMM, Wald test, P=0.00246 and P=0.00494 respectively). The negative relationship between *V. cholerae* and the number of high-frequency NS SNVs (**Figure 4A**) is not easily explained by coverage artifacts, since the total number of SNVs is not associated with *V. cholerae* relative abundance (**Figure S11A**). The number of high-frequency NS SNVs rises with increasing ICP1 – but only when *V. cholerae* abundance is high (**Figure 4A**). As a control, we ran the same GLMM on NS SNVs without a minimum frequency cutoff and found no significant effects, suggesting that the interaction between ICP1 and *V. cholerae* on SNV count is specific to high-frequency mutations, rather than low-frequency mutations that are more likely neutral or deleterious. These data support a scenario in which ICP1 selects for NS SNVs when the *V. cholerae* population is large enough to respond efficiently to selection – for example, at the beginning of an infection or in the absence of antibiotics.

If phages select for beneficial mutations, we would expect these mutations to increase in frequency at higher phage abundance. We lack time-series data from individual patients, but the relative abundance of phage provides a proxy for the combined effect of time and strength of phage selection. To test this expectation, we fit a GAM with the average within-sample frequency of SNVs as a function of ICP1, antibiotics, and their interactions. We included the fixed effect of ICE presence/absence as another factor that could provide phage or antibiotic resistance, as well as mutation type to differentiate among non-synonymous (NS), synonymous (S), and intergenic (I) mutations. We fit GAMs with all antibiotics and their interaction with ICP1, as well as simpler

models with each antibiotic separately, and compared them based on AIC (**Supplementary Table S8**). The most parsimonious model included the interaction between ICP1, ICE and mutation type, but not antibiotics. ICP1 was a strong predictor of higher frequency NS SNVs in the absence of a detectable ICE (**Figure 4B**). Samples in this analysis were unambiguous in terms of their ICE presence/absence patterns (**Figure S12A, B**). To confirm this model prediction, we compared the distribution of the average frequency of NS SNVs between ICP1-positive and ICP1-negative samples. Consistent with the model, NS SNV frequency was significantly higher in ICP1-positive samples when the ICE was not detected (Wilcoxon test, P=0.0094) (**Figure 4C**). This effect was strongest for NS mutations, making them the likely targets of ICP1 selection. The slight (not significant, P>0.05) trends for synonymous and intergenic SNVs could be due to genetic linkage with nonsynonymous mutations. Together, these results suggest that, in the absence of detectable ICE-encoded phage resistance, ICP1 selects for non-synonymous point mutations that may confer alternative modes of resistance.

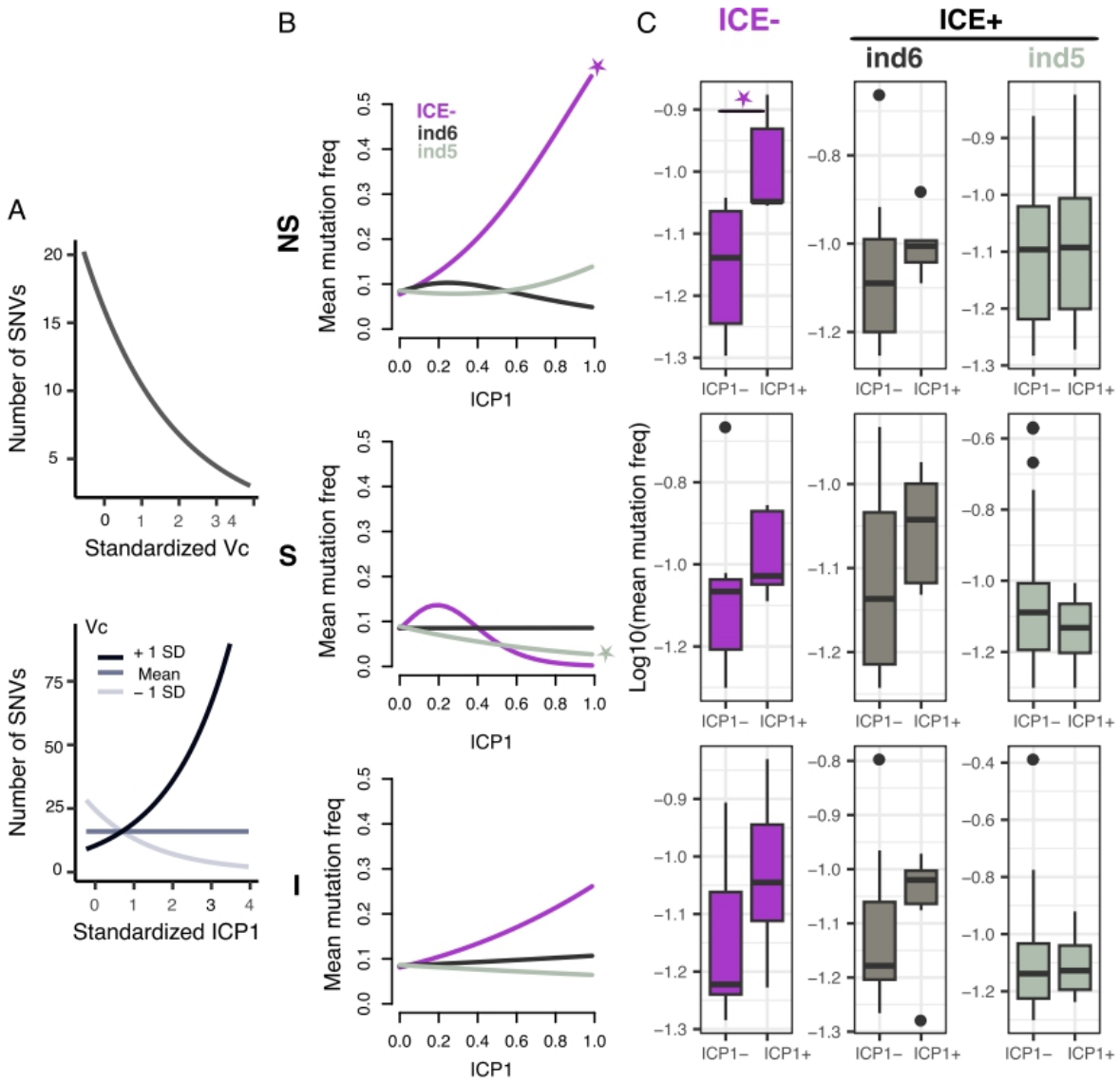

Figure 4. ICP1 selects for non-synonymous point mutations in the V. cholerae genome in the absence of ICE*. (A) Results of a GLMM modeling high frequency nonsynonymous SNV counts as a function of V. cholerae (Vc) and ICP1 standardized relative abundances. In the bottom panel, shades of gray indicate Vc relative abundance at the mean or +/- 1 standard deviation (SD) across samples. (B) GAM output with the mean mutation frequency as a function of the interaction between ICP1, ICE and mutation type (non-synonymous; NS, synonymous; S, or intergenic; I). Significant effects are shown with a star (P<0.05). The model was fit using 130 samples that passed the post-Instrain filter (Methods). (C) Boxplots of mutation frequency in the presence or absence of ICP1 and/or ICEs. The single significant comparison is indicated with a star (Wilcoxon test, P=0.0094). Boxplots include 130 samples, of which 32 are ICP1+ (ICP1>=0.1% of reads) and 98 are ICP- (ICP1<0.1% of reads). The solid horizontal line is the median and the boxed area is the interquartile range.*

To explore which point mutations might be specifically associated with phage resistance, we focused on the *V. cholerae* genes that accumulated repeated mutations under ICP1 pressure. We previously observed that the secreted hemolysin toxin gene *hlyA* was mutated more often

than expected by chance in cholera patients, but we were unable to identify an underlying selective pressure (Levade et al. 2021). In the larger sample reported here, *hlyA* was among the genes most frequently mutated in patients with relatively high *Vc* to ICP1 ratios (**Table S9**). This gene, along with others mutated at either high or low levels of phage (**Table S9; Table S10**) – including several with membrane or virulence-related functions – provide candidate phage resistance mechanisms that can be explored in future experiments.

We then investigated if within-ICP1 diversity was associated with *V. cholerae* anti-phage resistance genes. We focused on high-frequency SNVs (frequency>10%) in the ICP1 genome, which are more likely to be beneficial. Plotting the distribution of high frequency SNVs across different ICE profiles showed that NS SNVs (expected to have a fitness effect) are more common in ind5+ samples, suggesting that ind5 might exert stronger pressure on ICP1 than ind6 (**Figure S13**). The relationship between the frequency of NS mutations (only those higher than 10%) and ICP1:Vc ratio was significantly positive in ind5+ samples (spearman test, Benjamini-Hochberg corrected *P*=0.045) (**Figure S14**), suggesting that they are beneficial for ICP1 against ind5. We then identified genes that accumulated high frequency NS mutations in different ICE profiles. Most of the genes were annotated at NCBI as hypothetical proteins. Certain genes involved in nucleotide metabolism and an endonuclease were associated with ind5+ (**Table S11**) and ICE- samples (**Table S12**), and those encoding phage tail were associated with ind6+ samples (**Table S13**).

CONCLUSION

The tripartite interactions between pathogens, phage, and antibiotics have been studied in the laboratory, *in silico* with mathematical models, and to a lesser extent in the field, but how these factors interact during human infection remains an open question. Our objective was to characterize these interactions in the context of cholera. We analyzed more than 300 stool metagenomes from symptomatic patients enrolled at hospital admission across Bangladesh during an entire seasonal outbreak period. We discovered that higher levels of phage (ICP1) relative to *V. cholerae* were associated with mild dehydration, providing a biomarker of disease severity. The presence of phage alone did not strongly correlate with severity. However, the ratio

of phage to *V. cholerae* did strongly correlate with severity, suggesting that "effective predation" is a key metric to be used when considering phage as a biomarker.

As predicted by laboratory experiments showing that phage resistance elements on ICEs can protect against ICP1 (LeGault et al. 2021), we found that patient samples without detectable ICEs were associated with higher phage to *V. cholerae* ratios. While we have previously reported loss of ICE genes within cholera patients (Levade et al. 2017), lack of ICE detection in metagenomes could also be explained by colonization of a strain lacking the ICE or encoding an ICE variant with low similarity to those previously sequenced. In the absence of detectable ICE, phage (ICP1) was associated with increased rates of nonsynonymous point mutations in the *V. cholerae* genome. Many of these mutations likely arose by hypermutation, which generates deleterious mutations – but may also rapidly confer phage resistance, as observed in experiments with *Pseudomonas fluorescens (Pal et al. 2007)*. *V. cholerae* can evolve resistance to phage ICP2 within patients, yet the resistance mutations in a surface protein reduce its potential for onward transmission (Seed et al. 2014). Whether the mutations associated with ICP1 in our study mediate similar fitness tradeoffs for *V. cholerae* remains to be seen. Antibiotic exposure was not associated with increased rates of point mutations but was associated with less *V. cholerae* and less severe disease. Azithromycin appeared to be particularly effective at suppressing *V. cholerae* and was not associated with any known resistance genes in metagenomes. By contrast, ciprofloxacin exposure was associated with several known resistance genes and was – presumably for this reason – less effective at suppressing *V. cholerae*.

Our study has several limitations which provide opportunities for future research. First, samples were collected at a single time point (hospital admission) which allows us to establish statistical correlations, but in the absence of time-series or interventional experiments, we cannot infer causality. Notably, our results are consistent with ICP1 suppressing *V. cholerae* and reducing disease severity, but time-series studies of individual patients, or randomized controlled trials, will be needed to show causality. Second, shotgun metagenomic sequencing provides rich data on bacteria and phages in the gut microbiome, but is limited to the most abundant taxa, such as ICP1 and *V. cholerae*. Phages ICP2 and ICP3 were less prevalent in our study population, and less abundant within patients, making it challenging to infer their interactions with *V. cholerae* and

associations with disease severity. Short read sequencing also made it challenging to link ICE variants with *V. cholerae* genomes, so we relied on read mapping approaches. From a clinical perspective, quality of assessments of dehydration may have varied by provider and location which could produce bias and noise. Finally, our study lacked information about host genetics or immunity, which also contribute to disease outcomes (Harris et al. 2008; Nelson et al. 2009). Future studies combining rich patient metadata, time series, clinical interventions, long-read metagenomics, and isolate genome sequencing will complement and expand upon our work.

Despite these limitations, our study implicates both phages and antibiotics as determinants of cholera disease severity and paves the way for future enquiries into their interacting roles in disease progression and recovery. We propose a hierarchy of selective pressures acting on *V. cholerae* in the gut. In the absence of resistance genes, antibiotics are effective at suppressing cholera and reducing disease severity. In the absence of effective antibiotics, virulent phages suppress *V. cholerae* – particularly when the bacteria do not encode phage resistance in the ICE. Finally, in the absence of ICE-encoded resistance, phages may select for point mutations conferring phage resistance and potentially longer-term fitness consequences.

MATERIALS AND METHODS

Ethics Statement

The samples analyzed were collected within the confines of two previously published IRB approved clinical studies conducted in Bangladesh: (i) The mHealth Diarrhea Management (mHDM) cluster randomized controlled trial (IEDCR IRB/2017/10; icddr,b ERC/RRC PR-17036; University of Florida IRB 201601762) (Khan, Mack, et al. 2020). (ii) The National Cholera Surveillance (NCS) study (icddr,b ERC/RRC PR-15127) (Khan, Rashid, et al. 2020); See supplementary materials for further details.

Study Design

The study design was a prospective longitudinal study of patients presenting with diarrhoeal disease at five Bangladesh Ministry of Health and Family Welfare district hospitals (both mHDM and NCS sites) and two centralized NCS hospitals (BITID; icddr,b) from March 2018 to December 2018. See supplementary materials.

Sample collection. Stool samples were collected at hospital admission. Aliquots for transport and subsequent culture were stabbed into Cary-Blair transport media; aliquots for molecular analysis were preserved in RNAlater. See supplementary materials.

Microbiological and molecular analysis. Culture was performed via standard methods (Balows 2003); total nucleic acid (tNA) was extracted from the RNAlater preserved samples using standard methods. Cholera samples for subsequent metagenomic analysis were identified by screening all samples by culture for *V. cholerae*. Among culture negative samples, a random 10% of the remaining samples were screened for *V. cholerae* specific phage (ICP1, 2, 3) by PCR using total nucleic acid (tNA) extracts. From samples positive by culture or PCR, sequencing libraries were prepared using the NEB Ultra II shotgun kit and sequenced on illumina NovaSeq 6000 S4, pooling 96 samples per lane, yielding a mean of >30 million paired-end 150bp reads per sample.

Antibiotic detection by liquid chromatography mass spectrometry (LC-MS/MS)

Those cholera samples identified for metagenomic analysis were also analyzed by qualitative and quantitative LC-MS/MS for antibiotics (Creasy-Marrazzo et al. 2022; Alexandrova et al. 2019). While the target list for the qualitative analyses was broad, the list for the quantitative analyses was narrow: Ciprofloxacin, Doxycycline/Tetracycline, and Azithromycin. Standard curves were made for each quantitative target by preparing a dilution series of the three native and isotopic forms of the quantitative targets (Ciprofloxacin, Doxycycline, Azithromycin); for quantitative LC-MS/MS, clinical samples were spiked with the isotopes as internal references. See supplementary materials.

Metagenomic data analysis

We taxonomically classified short reads using Kraken2 (*44*) and Bracken v.2.5 (Lu et al. 2017). Reads were assembled using MEGAHIT v.1.2.9 (Li et al. 2015) and binned with DAS tool (Sieber et al. 2018). To characterize diversity within *V. cholerae*, we used StrainGE (van Dijk et al. 2022) and InStrain v.1.5.7 (Olm et al. 2021). To identify antibiotic resistance genes in metagenomes, we used deepARG v 1.0.2 (Arango-Argoty et al. 2018). See supplementary materials for details.

Statistical analyses

Statistics and visualizations were done in R studio version 1.2.5042. See supplementary materials for details.

Data availability

All sequencing data are deposited in the NCBI SRA under BioProject PRJNA976726. See supplementary materials for further information.

Code availability

Computer code for this paper will be available at https://github.com/Naima16/Cholera-phage-antibiotics.

Acknowledgements

We thank the patients for participating in this study and the clinical and laboratory teams that collected the samples. We are grateful to S. Flora and colleagues at the Institute of Epidemiology, Disease Control and Research (IEDCR), Ministry of Health and Family Welfare, Government of Bangladesh who collaborated on the original clinical studies in which the samples analyzed herein were collected. We are also grateful to R. Autrey and K. Berquist for their administrative expertise at the University of Florida. AIK and FQ were the principal investigators in Bangladesh and PIs of the ERC/RRC approvals at the icddr,b. This collective research infrastructure and support was invaluable to the success of this study. We thank members of the Nelson, Khan, Qadri and Shapiro labs for discussions that improved the manuscript.

Financial Support

This work was supported by the National Institutes of Health grants to EJN [R21TW010182] and KBB [S10 OD021758-01A1] and internal support from the Emerging Pathogens Institute at the University of Florida and the Departments of Pediatrics/ Children's Miracle Network (Florida). BJS and NM were supported by a Canadian Institutes for Health Research Project Grant. AC was supported by a Postdoctoral Mobility Fellowship of the Swiss National Science Foundation [P500PB_214356].

Disclaimer

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Potential conflicts of interest.

All authors: No reported conflicts.

Supplemental materials and methods

Ethics Statement

The samples analyzed were collected within the confines of two previously published IRB approved clinical studies conducted in Bangladesh: the mHealth Diarrhea Management (mHDM) cluster randomized controlled trial (IEDCR IRB/2017/10; icddr,b ERC/RRC PR-17036; University of Florida IRB 201601762) (Khan, Mack, et al. 2020) and a hospital-based National Cholera Surveillance (NCS) study (icddr,b ERC/RRC PR-15127) (Khan, Rashid, et al. 2020); the consent processes are described within these published studies.

Study Design

The study design was a national prospective longitudinal study of patients presenting with diarrhoeal disease at five Bangladesh Ministry of Health and Family Welfare district hospitals (both mHDM and NCS sites) and two centralized hospitals (NCS sites alone) (BITID, icddr,b) from March 2018 to December 2018. Recruitment was based on the census of diarrheal disease patients that sought care at the hospitals. For the mHDM study, inclusion criteria were patients two-months of age or older, and presentation with acute diarrhea defined as three or more episodes of loose stools in the 24 hours prior to admission and the duration of disease was less than seven days. For the NCS study, patients with diarrheal disease of all ages were included. For those under 2 months of age, diarrhea was defined as a change in stool habit from 'usual' (increased frequency with less formed stool). For those 2 months and older, diarrhea was defined as 3 or more loose or liquid stools within 24 hours or 3 loose/liquid stools or fewer causing dehydration in the last 24 hours. There was no exclusion based on enrollment in a second study (e.g., mHDM vs NCS).

Sample Collection

For the mHDM study, the intent was to collect four stool samples per study site per day. For the NCS study, the intent was to collect daily stool samples from two participants less than 5

years old and two participants 5 years of age and older; if the target number for an age group was not met, samples were collected from the other age group to achieve a total of 4 samples per day. For the samples collected at the five district hospitals and BITID, an aliquot was placed in transport media (Cary-Blair) and transferred to the icddr,b laboratory for culture, and a 0.5 ml aliquot was added to 1.3 ml of RNAlater to stabilize the sample for subsequent analysis; cold chain below 4°C was not consistently available at these sites. For samples collected at the icddr,b, samples were stabilized in RNAlater; culture was performed directly. Samples were stored at the centralized icddr,b laboratory at -80°C.

Antibiotic detection by liquid chromatography mass spectrometry (LC-MS/MS).

LC-MSMS methodology for both qualitative and quantitative approaches. The approach was based on prior studies (REF). Stool supernatants from the primary collection were obtained by centrifugation without filtration to minimize loss. Proteins were precipitated (1:7 ratio (v/v) of water:methanol). Supernatants were diluted with methanol and water (1:1 v/v) in 1% formic acid for liquid chromatography, and 5 □l of supernatant was injected for analysis. LC/MSMS was performed on a 2.1 x 150-mm Hypersil Gold aQ column (particle size, 3 µm) using a high-performance liquid chromatography system (Thermo UltiMate 3000 series) with an LTQ XL ion trap mass spectrometer (Thermo Fisher Scientific). Mobile phases were 1% formic acid in water (A) and 1% formic acid in methanol (B) and held at a constant 5%B for 2min before ramping to 95%B at 15 min where it was held for an additional minute before returning to starting conditions for a total run time of 25 min. Blanks were run in between every two samples, as well as before and after quality control samples and standards. µg

Eluent was ionized using electrospray ionization (ESI) in positive mode at a spray voltage of 5 kV, a nitrogen sheath gas flow rate of 8 L min^{-1}, and capillary temperature of 300°C. Two scan events were programmed to perform an initial scan from *m/z* 100 to 1000, which was followed by targeted collision induced dissociation based on a retention time and mass list. Retention time windows ranged from 1.7 minutes to 5.1 min, depending on the elution range of the standards at

high and low concentrations. Masses were targeted for the most abundant adduct or ion associated with each antibiotic (typically the $[M+H]^+$ ion) with a *m/z* 1 window.

LC-MSMS methodology specific to qualitative analysis. The target list for the qualitative analysis was broad and included both antibiotics and non-antibiotics commonly used in the clinical settings where the samples were obtained: Acetaminophen, Metronidazole, Zofran, Furazolidone, Nalidixic Acid, Sulfamethoxazole, Trimethoprim, Omeprazole, Ciprofloxacin, Cephalexin, Penicillin V (M+H), Amoxicillin, Penicillin V, Doxycycline or Tetracycline, Ceftriaxone, Erythromycin, and Azithromycin. Data analysis was performed using extracted ion chromatograms and MSMS matching with a control antibiotic MSMS library using Xcalibur 2.2 SP 1.48 (Thermo Fisher Scientific).

LC-MSMS methodology specific to quantitative analysis

The target list for the quantitative analysis was narrow: Ciprofloxacin, Ciprofloxacin d-8, Doxycycline or Tetracycline, Doxycycline d-5, and Azithromycin, Azithromycin d-5; the approach included scan filters for sulfamethoxazole/trimethoprim, metronidazole, and nalidixic acid which could be potential common confounding antibiotics. A standard curve was made for each quantitative target by preparing a dilution series of a mix of the three native forms of the quantitative targets (Ciprofloxacin, Doxycycline, Azithromycin). The dilutions were as follows: 0.5, 0.25, 0.125, 0.063, 0.05, 0.02, 0.01 μg/ml. All samples, including standards, were spiked with a 0.25 μM mix of isotope labeled antibiotics (Ciprofloxacin d-8, doxycycline d-5, Azithromycin d-5).

A calibration curve was generated by plotting the ratio of the area under the curve for analyte peaks in the ion chromatograms to the AUC of the isotope peaks against known concentrations of the seven standards. The linear line of best fit produced by this plot was generated and used to extrapolate the quantity of drug in mHDM samples. In quantitative analysis of the samples, the AUC of chromatogram peaks produced by the target compound was compared to that of the peak produced by the isotope spike in to generate a ratio. This ratio was then converted to a concentration (in units of μg/ml) using the calibration curve equation. Data analysis was performed manually by viewing the extracted ion chromatograms and MSMS from

each sample and matching with an in-house antibiotic library using 2.2 SP 1.48 (Thermo Fisher Scientific). The detection of a drug was confirmed when a peak in the chromatogram at the proper retention time window was identified and the most abundant adduct or ion associated with that drug (typically the $[M+H]^+$ ion) with a m/z 1 window.

Molecular analyses

Standard methods were used to extract total nucleic acid (tNA) from the samples stabilized in RNAlater. In brief, the samples were thawed at room temperature and centrifuged for 5 minutes at max speed and room temperature. All but 200 µl of the supernatant was removed; the supernatant was refrozen for subsequent mass spectrometry. The remaining pellet, with the 200 µl supernatant, was combined with 380 mg of glass beads and 1 ml InhibitEx buffer, and processed using the QIAamp Fast DNA Stool Mini Kit. Elutions were performed with 50 µl of Qiagen ATE buffer followed by a second elution with 50 µl ATE, quantification (Nanodrop) and storage at -80°C. For PCR, primers pairs were developed for the virulent bacteriophages ICP1 (gp58.2), ICP2 (gp24), and ICP3 (gp19). Primers were validated with (i) synthetic templates and (ii) biologic templates (*V. cholerae,* ICP1, ICP2, ICP3); primer sequences are provided in a Table S1.

Metagenomic analyses

Short read classification using kraken2/braken

We classified all short-read data with a Kraken2 database (Wood, Lu, and Langmead 2019) containing bacterial, archaeal, and viral domains, along with the human genome and a collection of known vectors (UniVec_Core) from NCBI, in July 2020. A Bracken database (Lu et al. 2017) was also built with a read length of 150 bp and the default k-mer length of 35. Kraken2 and Bracken version 2.5 were run with default parameters.

Quality filtering

Before calling SNVs within the *V. cholerae* population, metagenomes were decontaminated of human and PhiX reads by mapping the reads against the GRCh 38 assembly

of the human genome and the PhiX genomes, with bowtie2 version 2.3.5 (Langmead et al. 2009). Unmapped reads were used for subsequent analyses. Then, we used iu_filter_quality_minoche from the Illumina Utils package (v2.10) with default parameters (Minoche, Dohm, and Himmelbauer 2011) to trim adapters, remove duplicates and for quality filtering on the short reads.

SNV profiling

Short reads were assembled with MEGAHIT (Li et al. 2015) version 1.2.9 with the default parameters (k-mers length: k=21 et maxk=99 and min contig length of 1000. The contigs were grouped into bins with concoct (Alneberg et al. 2014) (v1.0.0) and MetaBAT 2 (Kang et al. 2015) (v2.12.1) with default parameters. Bins were then aggregated together using DAS Tool version 1.1.0 (Sieber et al. 2018). Then we used drep (Olm, Brown, Brooks, and Banfield 2017) version 2.0.0 with S_ani 0.98, to de-replicate the bins and identify unique metagenome-assembled genomes (MAGs) with a secondary clustering threshold of 98% and minimum completeness of 75% and maximum contamination of 25%. To prevent mismapping of reads from other species to the *V. cholerae* genome, we competitively mapped reads from all samples with sufficient *V. cholerae* or phages (*Vc*>0.5% of reads or phages>0.1% of metagenomic reads) against the concatenation of the 677 unique MAGs using bowtie2 v 2.3.5. To infer potential mixed infections, we used strainGST from the strainGE toolkit (van Dijk et al. 2022). To do so, we compiled a reference database of publicly available *V. cholerae* assemblies, including (i) all complete genomes from NCBI RefSeq (n=106, accessed 2023/05/12) and (ii) all assemblies published on NCBI Genbank between 2015/01/01 and 2023/05/12 and originating from samples collected between 2015/01/01 and 2019/12/31 (n=758). To characterize within-patient *V. cholerae* genetic diversity, we profiled the resulting bam files using inStrain v1.5.7 (Olm et al. 2021) with default parameters (minimum coverage of 5 to call a variant).

To reduce false positive SNVs, we applied a stringent post-InStrain filter: all positions with coverage <20 were removed. We removed the first and last 100 positions of every scaffold, as well as positions with coverage below 0.3*median and above 3*median (median coverage across all the positions in the sample). We also removed sites that did not pass the coverage filter in more than 2 samples. Finally, we removed 420 sites that were variable in a *V. cholerae* isolate

genome sequenced alongside the metagenomes. These sites were considered prone to sequencing error since they varied in an isolate genome that theoretically should contain no variation. After applying all these quality filters, we were left with 130 samples with SNV calls.

Hypermutator definition

Hypermutators were identified as samples with one or more nonsynonymous mutations in DNA repair genes (DNA repair defects) in *V. cholerae*, defined previously (Jolivet-Gougeon et al. 2011). We used prodigal v2.6.3 (Hyatt et al. 2010) with the default parameters to predict genes in the *Vc* MAG and annotated them with eggNOG-Mapper v2 with default parameters and eggNOG database v2 downloaded on April 2021 (Cantalapiedra et al. 2021).

Antibiotic resistance gene identification

Antibiotic resistance genes (ARGs) were predicted from short reads using deepARG (Arango-Argoty et al. 2018) v 1.0.2 with default parameters. Deeparg is a deep learning model that can predict ARGs from short-read metagenomic data; it uses DeepARG-DB, a merged database from 3 databases: Antibiotic Resistance Genes Database [ARDB], Comprehensive Antibiotic Resistance Database [CARD], and UniProt. It was downloaded in December 2021. We used the relative abundance of ARGs normalized by the 16S rRNA gene content in the sample.

SXT ICE identification

We used Bowtie2/2.3.5 with the 'very-sensitive' option to map short reads against the two most prevalent SXT ICEs in Bangladesh at the time of our sampling (LeGault et al. 2021): ICEVchind5 (GQ463142.1) and ICEVchind6 (MK165649.1) reference sequences downloaded from NCBI. The ICE was considered as present in a metagenome when 90% of its length was covered by at least one read. Using this criterion, 144 samples (59%) were ind5+, 26 (10.6%) ind6+, and 54 (22.1%) ICE-.

Statistical analyses

All statistics and visualizations were done in R studio version 1.2.5042.

To reduce dimensions of the species composition table, we ran a principal component analysis (PCA) on the Hellinger transformed abundance data (5719 species*344 samples) (decostand function and rda function in the vegan R library) and selected the 20 most dominant species on each axis (37 in total).

To understand relationships between taxa in the microbiome, phages, antibiotics and patient metadata, we used a Redundancy analysis (RDA) on bacterial species abundances (rda function from vegan R package). Only the 7 most dominant species identified with the PCA were included. As explanatory variables, we used the most contributing patient metadata, selected with forward selection method (ordiR2step function from the vegan R library). We began the forward selection with a model with phages and antibiotic concentrations (µg/ml). The RDA was run on log-chord transformed abundances, and a permutation test was used to assess the statistical significance of both the model and of each explanatory variable (anova.cca function from vegan package in R). The explained variation R^2 and adjusted R^2 were estimated with the RsquareAdj function in the vegan R package.

To identify species associated with each degree of dehydration, we used the indicator species analysis (Dufrene and Legendre 1997) as implemented in the multipatt R function from the indicspecies R library with (no group combination, duleg parameter set to TRUE and 9999 permutations) on the log-chord transformed species table (334 samples, 37 dominant species from the PCA) using the decostand and rda functions from the R vegan library). Reported P-values correspond to a permutation test with 999 iterations.

To study associations between ARGs and species, we ran a Multiple factor analysis (MFA) (Pagès 2002) using the mfa function from the FactoMineR R package, with the 37 most dominant species and all 634 resistance genes identified by deepARG. The visualization was done with the fviz_mfa_var function from the factoextra R library.

To model the relationships between phage, antibiotics and *V. cholerae* within each patient, we fit a generalized additive model (GAM) with *V. cholerae* relative abundance as a function of ICP1, antibiotics, and the interaction between them. We added the degree of dehydration as a random effect because it improved the model (smallest Akaike information

criteria (AIC) compared to the other models). We fit several GAMs with different combinations of predictors: from a model with all antibiotics and their interaction with ICP1 to separate models with each antibiotic and its interaction with ICP1, and compared them based on their AIC (AICtab function from the bblme R package). The most parsimonious model was retained (the one with the smallest AIC). The GAMs were fitted using a beta error distribution with log-link function because *Vc* relative abundance is a continuous value between 0 and 1. We evaluated the selected model fit by inspecting residual distributions and fitted-observed value plots using the gam.check function from the mgcv R package. All P-values reported for the GAMs correspond to the Chi-square test from gam.summary function from mgcv R package. We also reported the adjusted R^2 (from the same function) as an evaluation of the goodness of fit (**Table S6**).

We tested whether phages or antibiotics select for potentially adaptive mutations in *Vc* by fitting generalized linear mixed models (GLMMs) (glmmtmb function from glmmTMB R package) with phages and antibiotics as predictors of the number of high-frequency nonsynonymous SNVs in the *V. cholerae* genome. We added *Vc* abundance as a fixed effect to the model to control for any coverage effects. We focused on higher-frequency SNVs (>10% within a sample) as more likely to be beneficial, and nonsynonymous SNVs as those more likely to have fitness effects. We fit several models with different combinations of predictors: from a model with all antibiotics and their interaction with ICP1 to separate models with each antibiotic and its interaction with ICP1, and compared them based on their AIC (AICtab function from bblme R package). The most parsimonious model was retained (the one with the smallest AIC). Because the response is count data, we fit three different count GLMMs: Poisson, negative binomial1 and negative binomial2, implemented in the glmmTMB R package, and then selected the model with the lowest AIC as described (Brooks et al. 2017). The selected model was fitted with the nbinom2 error distribution. All continuous predictors were standardized to zero mean and unit variance before analyses to improve convergence. We evaluated the selected model fits by inspecting the residuals using the DHARMa library in R (simulateResiduals and plot functions). All the *P*-values reported for the GLMMs correspond to the Wald *P*-values reported by the glmm.summary function from the R package glmmTMB. As an evaluation of the goodness of fit, we compared the GLMMs and the corresponding null models (the same model but with no fixed effects other than the intercept)

with likelihood-ratio tests (anova function from the stats R package), and reported the corresponding P-values (**Table S7**). We also reported the marginal R^2 as a measure of the variance explained by fixed effects, estimated with the r2 function from the performance R package. We used GLMMs because they are currently more flexible than GAMs in the range of count models that they can fit (https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html). They can deal with overdispersion with two versions of negative binomial distributions: negative binomial 1 and negative binomial 2, respectively with linear and quadratic parametrization (Hardin and Hilbe 2018).

To further test the hypothesis that phages or antibiotics select for adaptive mutations in *V. cholerae*, we examined the relationship of the frequency of nonsynonymous SNVs in *V. cholerae* with phages and antibiotics. We fit a generalized additive model (GAM) (gam function in mgcv R library) with the average frequency of nonsynonymous SNVs as a function of ICP1, antibiotics, and their interactions. We also included the fixed effect of ICE presence/absence as another factor that could provide phage or antibiotic resistance as well as mutation type to differentiate among non-synonymous (NS), synonymous (S), and intergenic (I) mutations. We fit GAMs with all antibiotics and their interaction with ICP1, as well as several simpler models with each antibiotic separately, and compared them based on their AIC. The GAMs were fitted using a beta error distribution with a log-link function because the average minor allele frequency of nonsynonymous SNVs is a continuous value between 0 and 0.5. We evaluated the selected model fit by inspecting residual distributions and fitted-observed value plots using the gam.check function from the mgcv R package. All the P-values reported for the GAMs correspond to the Chi-square test from the gam.summary function of the mgcv R package. We also reported the adjusted R2 (from the same function) as an evaluation of the goodness of fit (**Table S8**).

SUPPLEMENTARY FIGURES

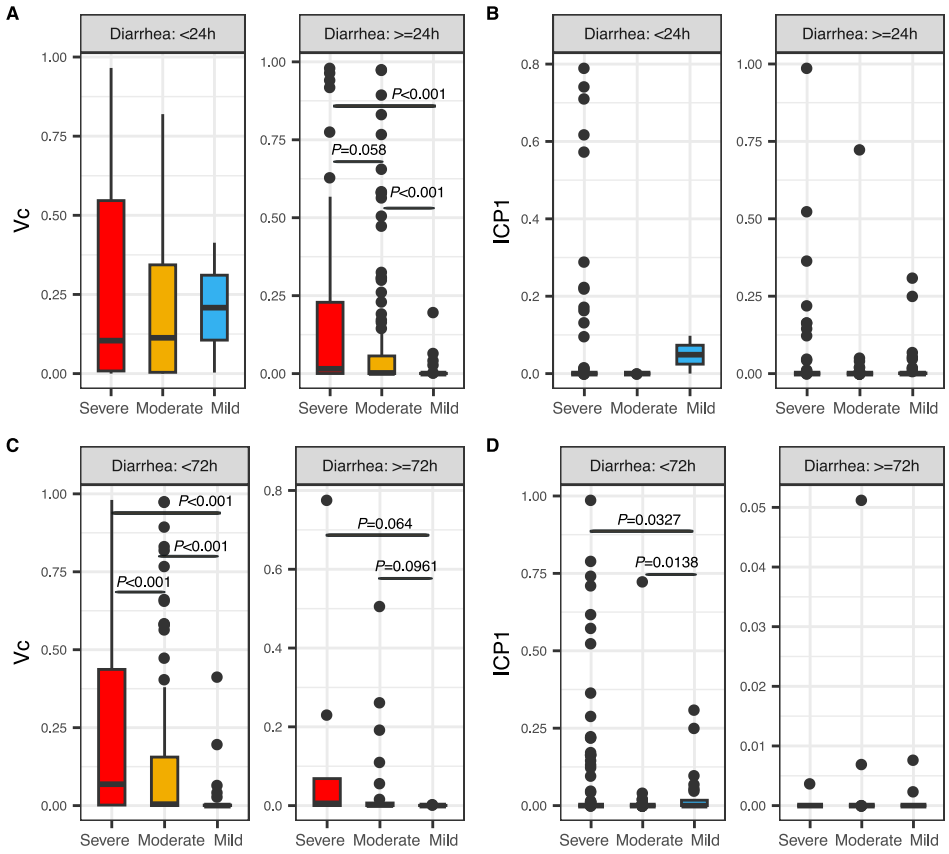

Figure S1. *ICP1:V. cholerae ratios among patients with different dehydration status binned by self-reported duration of diarrhea. P-values are from a Kruskal-Wallis test with Dunn's post-hoc test, adjusted for multiple tests using the Benjamini-Hochberg (BH) method. Only significant (P<0.05) and marginally significant P-values (<0.1) are shown. Only 323 samples with V. cholerae>0% of metagenomic reads were included, with 165 from severe, 128 from moderate, and 30 from mild cases. A pseudocount of one was added to the ratio before log transformation. The solid horizontal line is the median and the boxed area is the interquartile range.*

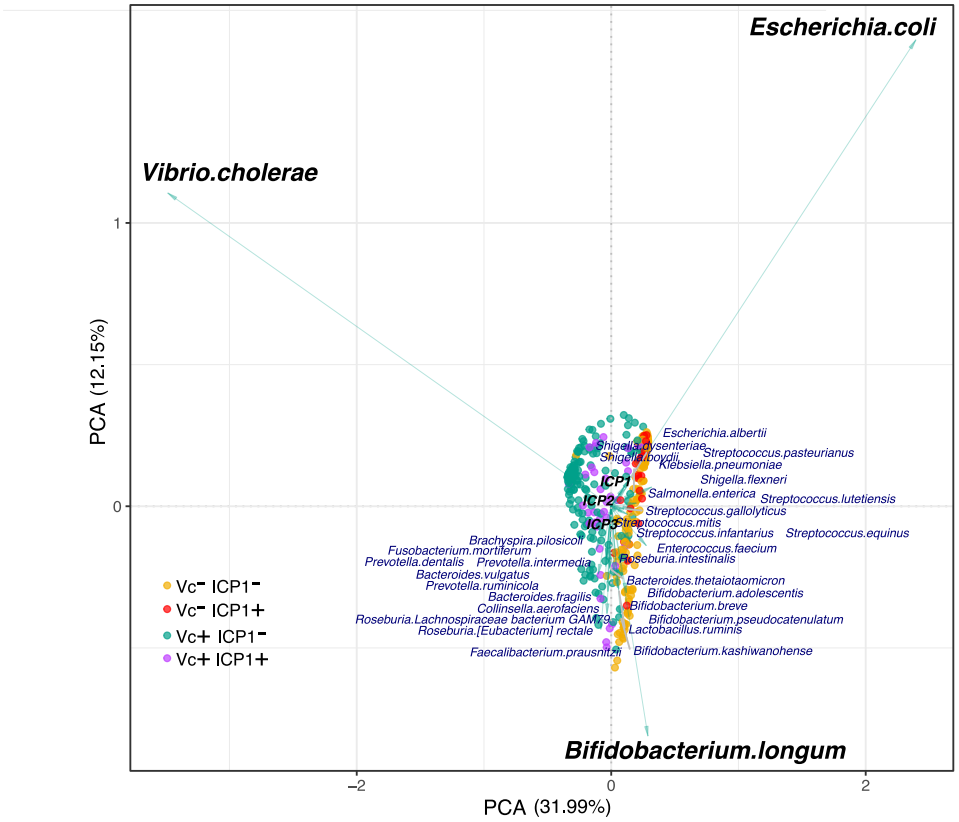

Figure S2. Principal component analysis (PCA) of the taxonomic composition of cholera patient stool samples*. Circles indicate samples; arrows indicate species. Samples are colored by their V. cholerae and ICP1 relative abundances (percentage of metagenomic reads). Yellow: patients with V. cholerae < 0.05% and ICP1 < 0.01%, red: V. cholerae < 0.05% and ICP1 >= 0.01%, green: V. cholerae >= 0.05% and ICP1 < 0.01%, purple: V. cholerae >= 0.05% and ICP1 >= 0.01%.*

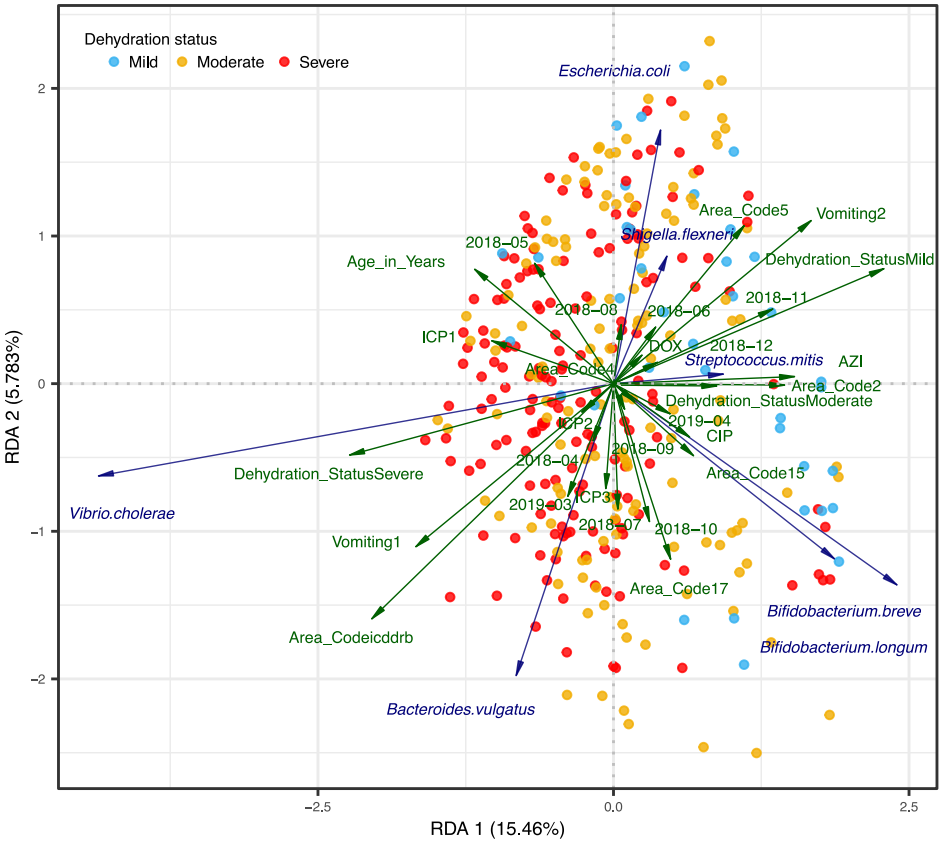

Figure S3. The two first axes from the RDA on prevalent bacterial species, patient metadata, and antibiotic concentrations*. All variables are shown (the collection date and area code were omitted in Figure 1C for clarity).*

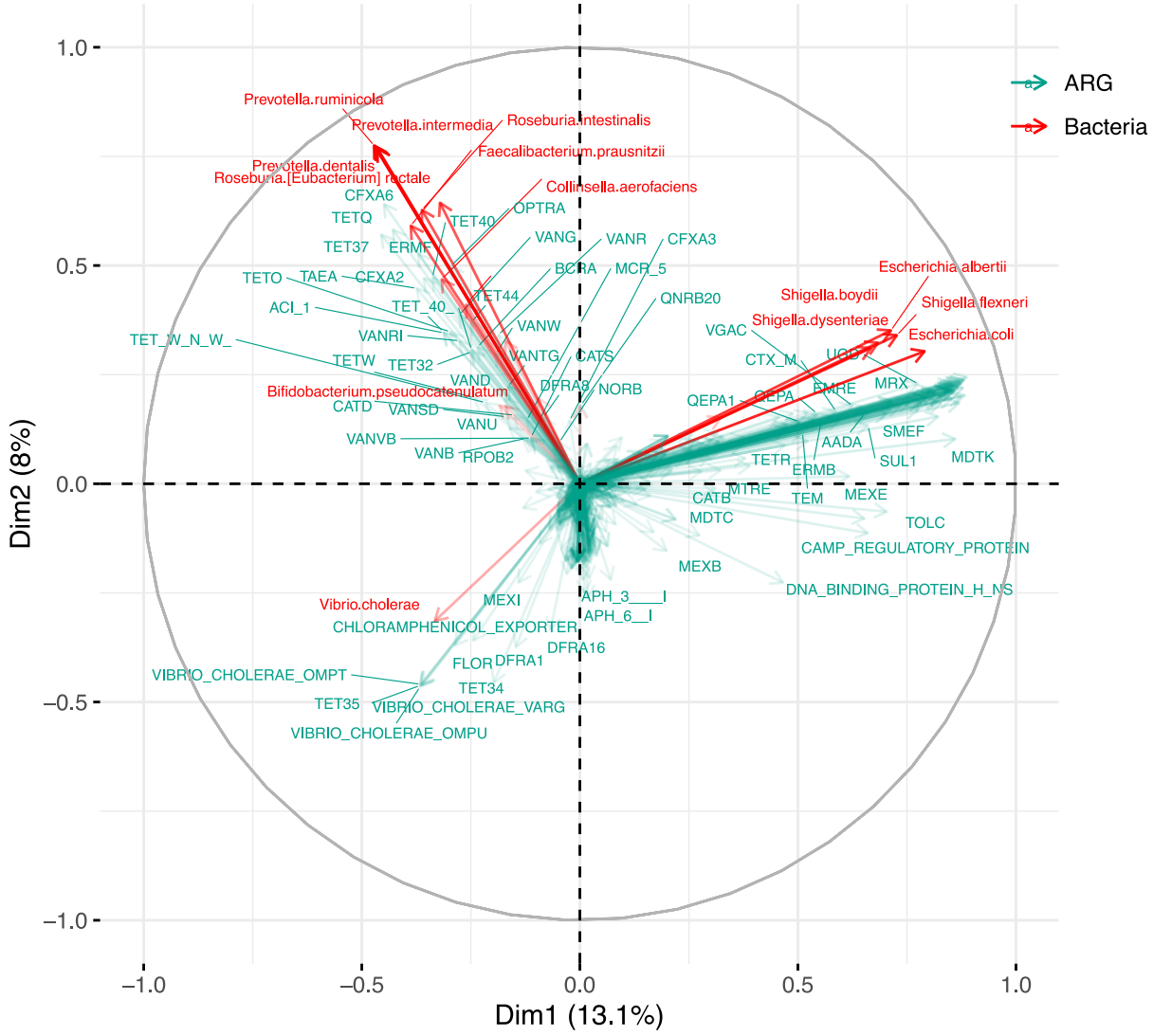

Figure S4. MFA showing correlations between bacterial species and antimicrobial resistance genes (ARGs) in patient gut microbiomes. *Only the greatest correlations on each axis are shown for clarity.*

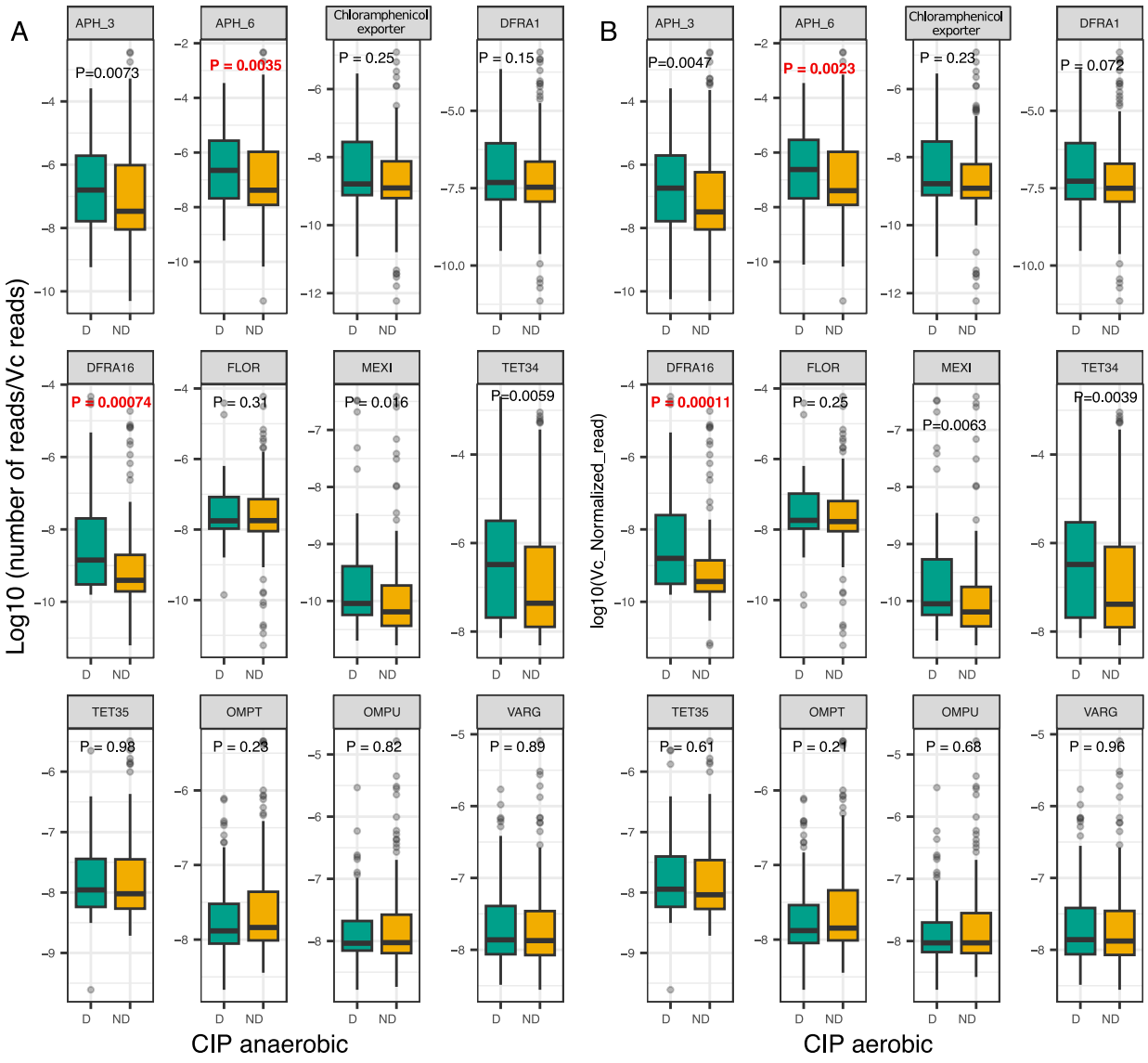

Figure S5. Distribution of the relative abundance of V. cholerae-associated ARGs in patients with different exposures to ciprofloxacin (CIP)*. CIP>=MIC (detected; D) and CIP<MIC (not detected; ND) under anaerobic conditions (A) or aerobic conditions (B). The Y-axis is the relative abundance of ARGs in metagenomes normalized by 16S rRNA gene reads and by V. cholerae reads. Only V. cholerae-positive samples are plotted (V. cholerae>0 reads). P-values are from a Wilcoxon test. In red: BH corrected P<0.05 after correction for 24 tests.*

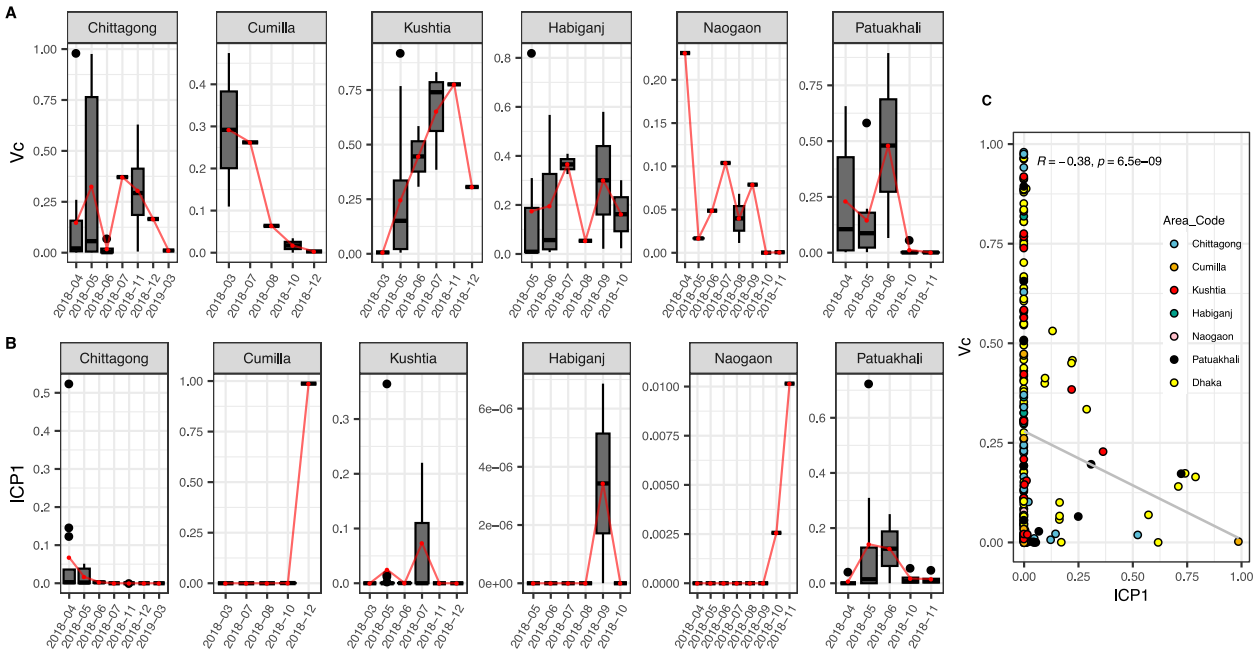

Figure S6. Relative abundances of (A) V. cholerae and (B) ICP1 over time (binned by month) in metagenomes sampled from six sampled regions of Bangladesh*. Red line shows the mean. (C) Spearman correlation between V. cholerae and ICP1 in all the data, including these six regions and the icddr,b (Figure 2 in the main text). The solid horizontal line is the median and the boxed area is the interquartile range.*

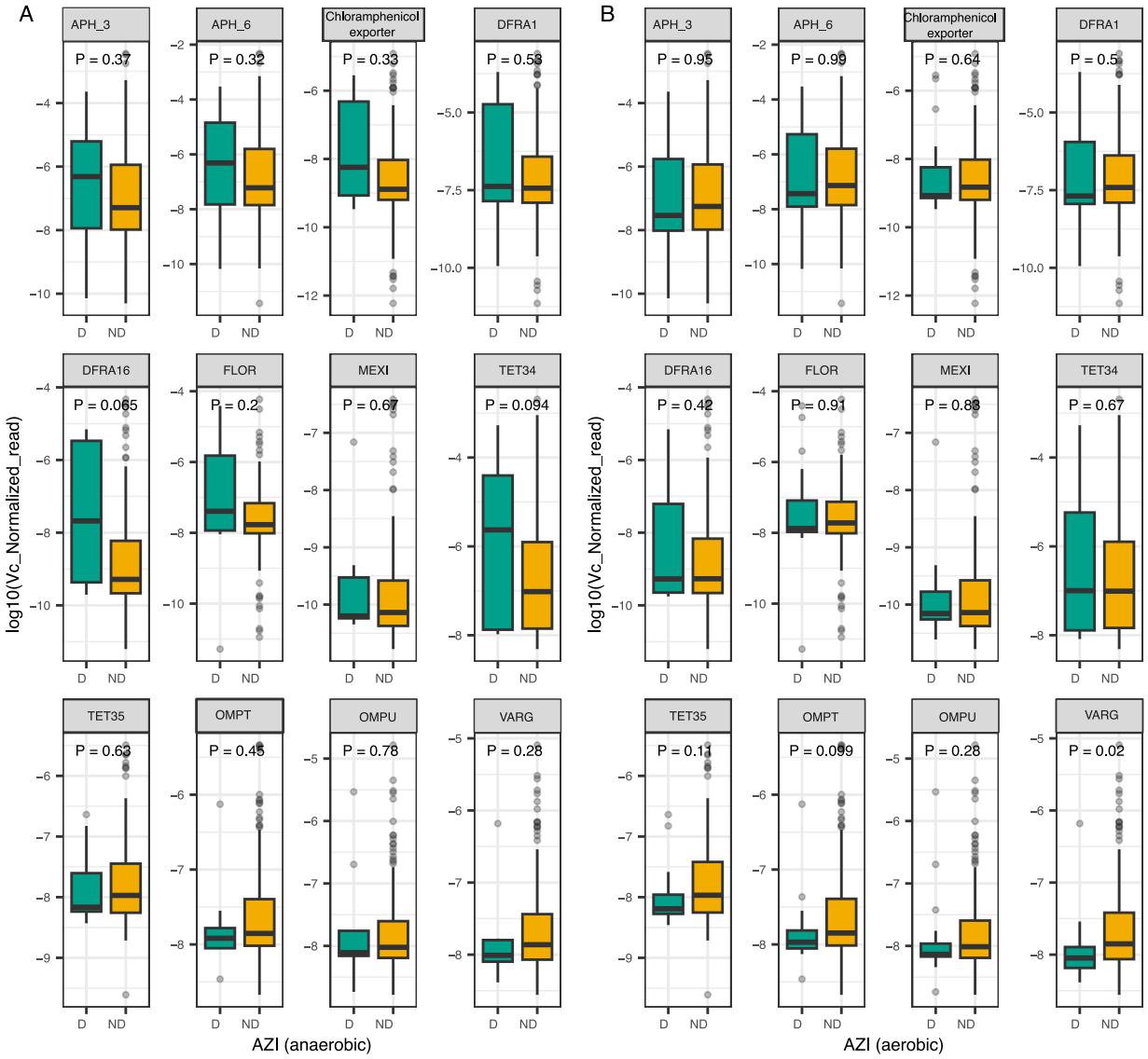

Figure S7. Distribution of the relative abundance of V. cholerae-associated ARGs in patients with different exposures to azithromycin (AZI)*. AZI>=MIC (detected; D) and AZI<MIC (not detected; ND) under anaerobic conditions (A) or aerobic conditions (B). The Y-axis is the relative abundance of ARGs in metagenomes normalized by 16S rRNA gene reads and by V. cholerae reads. Only V. cholerae-positive samples are plotted (V. cholerae>0 reads). P-values are from a Wilcoxon test. No comparisons (D vs. ND) are significant (P<0.05) after BH correction for 24 tests.*

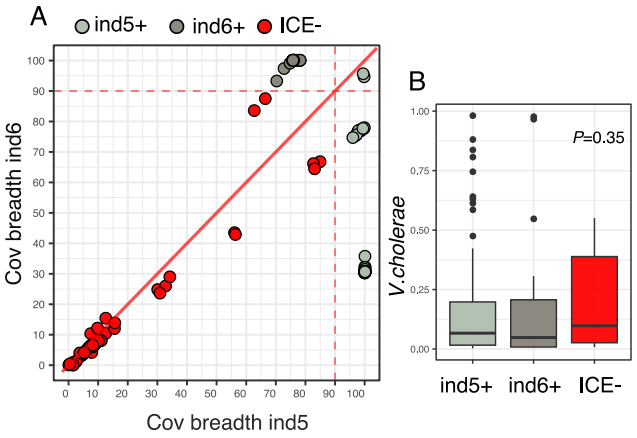

Figure S8*. A) Breadth of metagenomic read coverage of the ICEs ind5 and ind6. Samples were identified as ind5-positive or ind6-positive based on the mapping breadth. The ICE was considered as present when 90% of the reference ICE length was covered by at least one read. Only 2 samples had 100% ind5 and >90% ind6, which were considered ind5-positive. B) Boxplot showing that ICE-negative samples are not associated with lower relative abundance of V. cholerae. The P-value is from a Kruskal-Wallis test.*

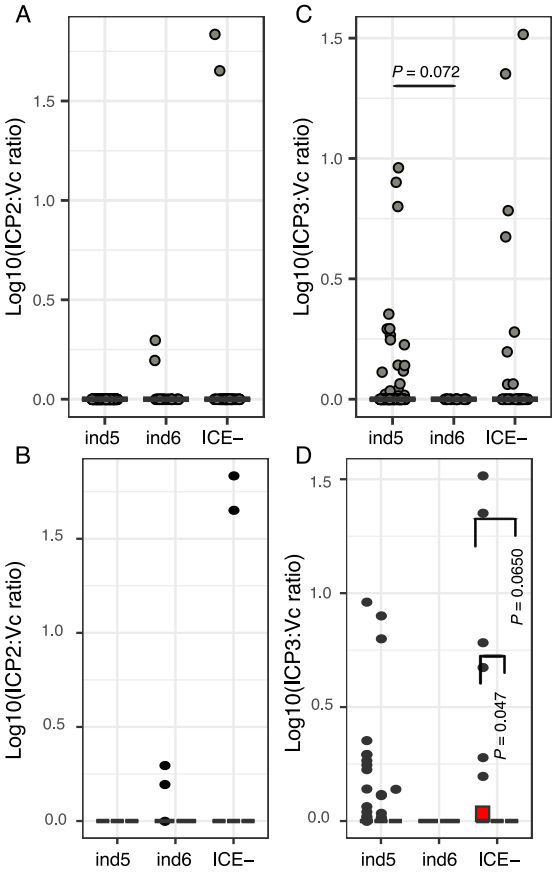

Figure S9. *(A) Distribution of ICP2:Vc ratios across patients with different ICE profiles. (B) The same data as (A) binned into boxplots according to dehydration status. (C) Distribution of ICP3:Vc ratios across patients with different ICE profiles, (D) The same data as (C) binned into boxplots according to dehydration status. P-values are from a Kruskal-Wallis test with Dunn's post-hoc test adjusted with the Benjamini-Hochberg (BH) method. Only P-values < 0.1 are*

shown. Only samples with sufficient Vc or ICP were included. For clarity, the Y-axes were log10 transformed after adding one to the ratios. The solid horizontal line is the median and the boxed area is the interquartile range.

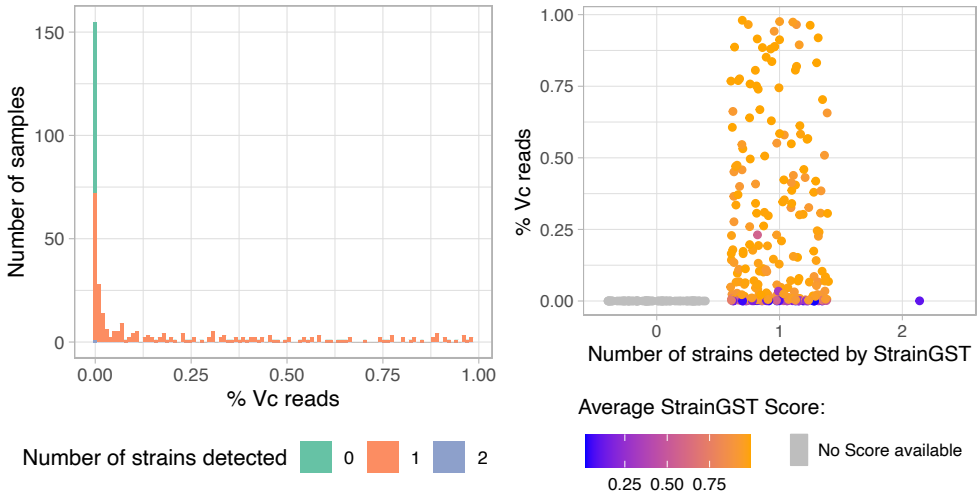

***Figure S10**. Mixed infections by more than on V. cholerae strain is unlikely in our patients. (A) The distribution of the number of distinct strains detected across samples. In 260/344 samples, strainGST identified only one reference strain. In 83/344 samples, no reference strain could be identified with confidence, likely due to low coverage of V. cholerae (these samples all had <1% Vc reads). (B) The relative abundance of Vc (% reads) in samples with zero, one, or two strains identified. In one sample, strainGST identified two reference strains, but with a low confidence score and at low Vc abundance.*

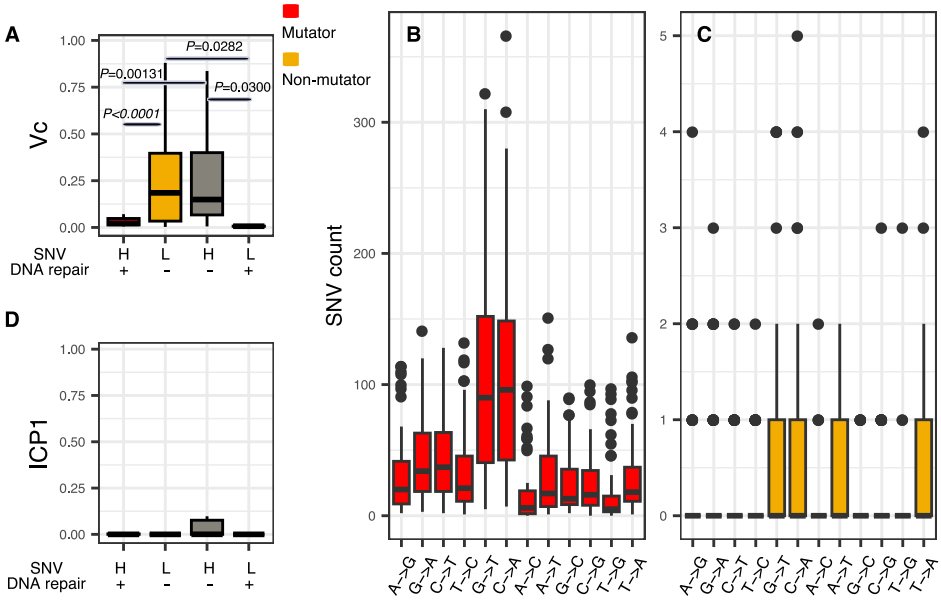

Figure S11. *A) Low V. cholerae relative abundance is associated with DNA repair mutations independently of the number of SNVs. Mutators are defined as having a high (H) number of SNVs (25 or more) in the V. cholerae genome, along with one or more nonsynonymous mutations in a DNA repair gene (+) resulting in a predicted defect in DNA repair. Non-mutators have neither a high number of SNVs nor a DNA repair defect. P-values are from a Kruskal-Wallis test with Dunn's post-hoc test adjusted with the Benjamini-Hochberg (BH) method. Only P-values < 0.1 are shown. (B and C) Transversion mutations, particularly G->T and C->A, are more common in mutators compared to non-mutators. D) Relative abundance of phage ICP1 in samples with different V. cholerae mutation profiles. Kruskall-wallis test (P= 0.07197). n=133, with 47 belonging to mutators, 70 to non-mutators, 14 to high SNV/no DNA repair defect, and 2 to low SNV/DNA repair defect groups.*

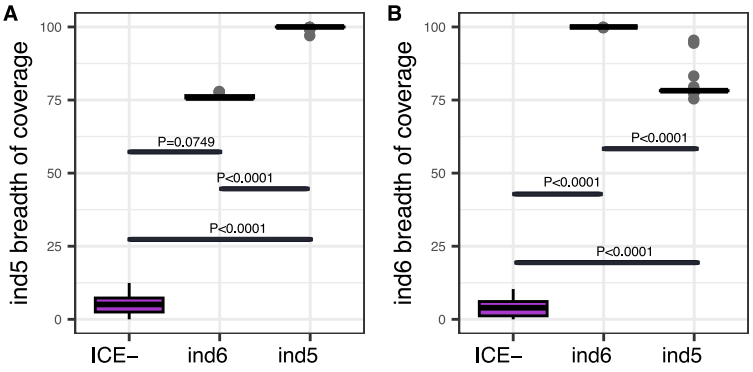

Figure S12. ICE- are readily distinguishable from ind5+ and ind6+ samples. *Breadth of ICE coverage across patients with different ICE profiles. (A) Breadth of ind5 coverage. (B) Breadth of ind6 coverage. Note that the few samples with ambiguous breadth of ICE coverage (in the 40-80% range) were not included in the InStrain analysis, and are not included here. P-values are from a Kruskal-Wallis test with Dunn's post-hoc test adjusted with the Benjamini-Hochberg (BH) method. n=131 with 24 belonging to ICE-, 18 to ind6+, and 89 to ind5+ groups.*

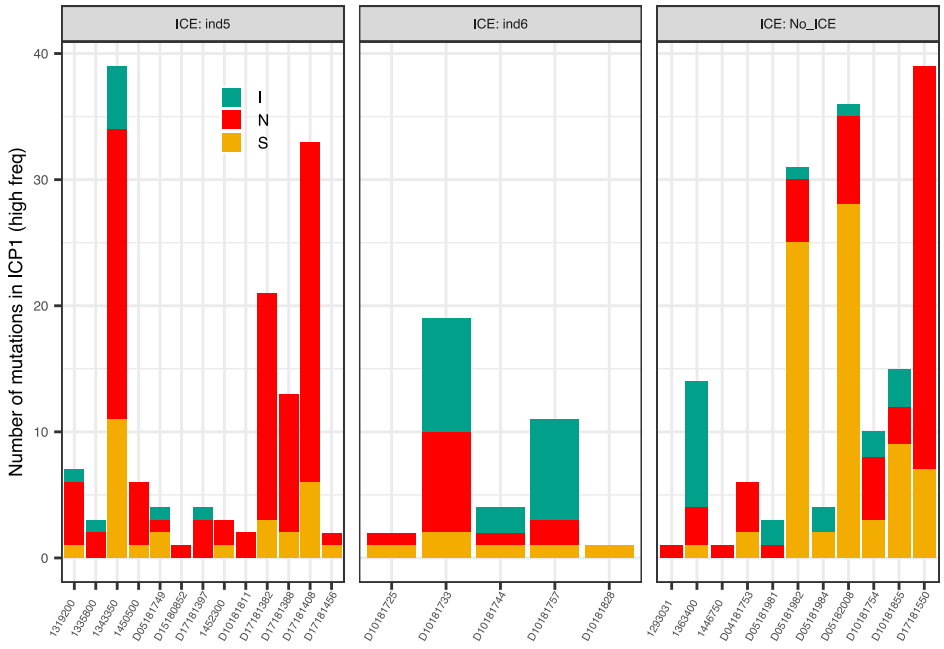

Figure S13. Genetic diversity in ICP1 varies among patients. There are more high frequency SNVs in ind5+ and ICE- than in ind6+ samples but NS mutations are more common in ind5+ samples.

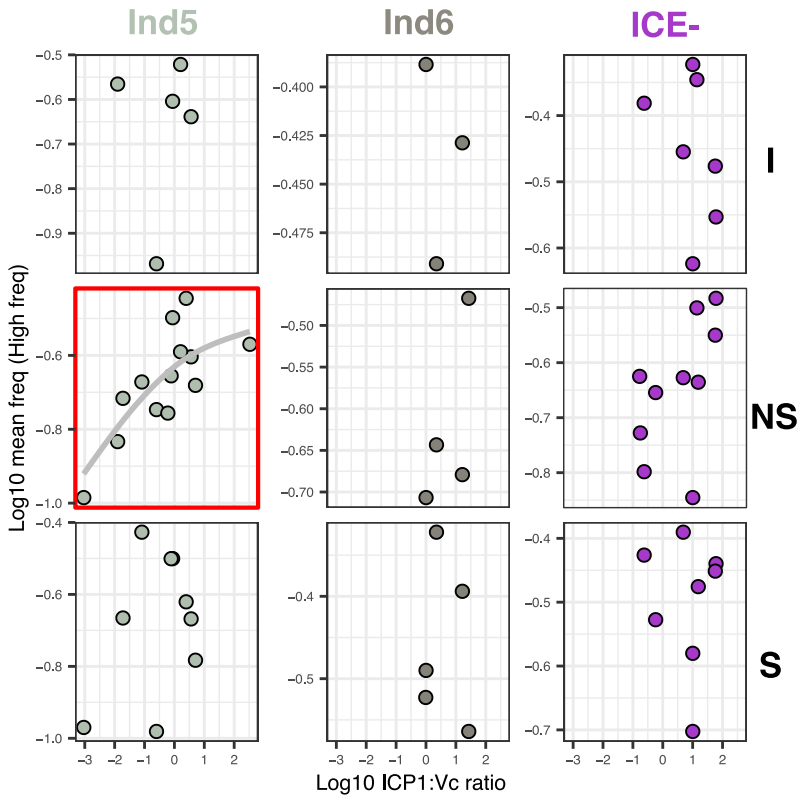

Figure S14. The frequency of NS mutations is correlated with ICP1:Vc ratio in ind5+ samples (spearman test, HB corrected P=0.045, in red).

SUPPLEMENTARY TABLES

Table S1. qPCR targets and primers

Target	Primer Name	Sequence 5'-3'	Reference
tcpA	*tcpA set2*_F	ACACGATAAGAAAACCGGTCA	(Alexandrova et al. 2019)
	*tcpA set2*_R	GCCTTGGTCATATTCTGCGA	
ICP1	GP58.2_F	CAAAGGCAGCAGGTAGGACA	This study
	GP58.2_R	CCCTTCAAGCCGTAGTTGGT	
ICP2	GP24_F	AGAAGTCGCAAACGGGGTAC	This study
	GP24_R	AACGTGGTTCTCGTGAGTGG	
ICP3	GP19_F	AGACCAACGCCGACTGTTAG	This Study
	GP19_R	CGATACCACGGAAAGCCTGT	
16S rDNA1	Maeda_1048_1067_F	GTGSTGCAYGGYTGTCGTCA	(Maeda et al. 2003)
	Maeda_1175_1194_R A	ACGTCRTCCMCACCTTCCTC	

Table S2. LC MS/MS targets and parameters.

Antibiotic	RT (min)	Extraction Ions
Azithromycin	11.4 – 13.1	750,591
Ciprofloxacin	9.5 – 11.2	332
Doxycycline	8.9 – 14.0	445,428
Nalidixic Acid	12.3 – 14.4	233
Azithromycin-d5	11.4 – 13.1	754,595
Ciprofloxacin-d8	9.5 – 11.2	339
Doxycycline-d5	8.9 – 14.0	450,433
Metronidazole	3.5-5.5	172

Table S3. Indicator species analysis. For each group, we report the indicator value between 0-1 ("stat"), with 1 being the perfect indicator species (occurs exclusively in one group). P-values are from a permutation test. Total number of species: 37. Selected number of species: 24.

Group	**stat**	***p***
Group 1: mild dehydration		
Bifidobacterium longum	0.372	**1e-05**
Bifidobacterium breve	0.304	**6e-05**
Escherichia coli	0.266	**0.00057**
Enterococcus aecium	0.214	0.00437
Bifidobacterium kashiwanohense	0.2	0.00818
Salmonella enterica	0.183	0.01262
ICP1	0.167	**0.02641**
Group 2: moderate dehydration		
Streptococcus pasteurianus	0.23	0.00272
Roseburia intestinalis	0.21	0.00596
Streptococcus gallolyticus	0.16	0.03152
Group 3: severe dehydration		
Vibrio cholerae	0.333	**6e-05**
Fusobacterium mortiferum	0.31	0.00011
Faecalibacterium prausnitzii	0.291	0.00020
Prevotella ruminicola	0.284	0.00028
Prevotella dentalis	0.281	0.00030
Brachyspira pilosicoli	0.249	0.00084
Collinsella aerofaciens	0.246	0.00149
Bacteroides thetaiotaomicron	0.234	0.00221
Bacteroides vulgatus	0.226	**0.00309**
Prevotella intermedia	0.209	0.00560
Bacteroides fragilis	0.201	0.00823
Roseburia[Eubacterium] rectale	0.194	0.01063
ICP3	0.189	**0.01050**
Roseburia lachnospiraceae bacterium GAM79	0.168	0.02409

Table S4. RDA variables and p-values.

Variable	***p*[a]**
CIP	0.073
DOX	0.701
AZI	0.001***
ICP1	0.012*
ICP2	0.336
ICP3	0.104
Area_Code	0.001***
Dehydration status	0.001***
Collection date	0.003**
Vomiting	0.002**
Age_in_Years	0.006**

[a] Permutation test (anova function from the vegan R package).

Table S5. Antibiotics grouping thresholds. *Minimal inhibitory concentrations (MICs) were established under aerobic or anaerobic conditions in (Creasy-Marrazzo et al. 2022). Concentrations are in units of µg/ml.*

	Aerobic		**Anaerobic**
Ciprofloxacin	0.016		0.063
Azithromycin	1		8
Doxycycline	0.13		0.13

Table S6. Generalized additive models included in the model selection (set 1). *Generalized additive models included in the model selection. GAMs were fit with* Vc *abundance as a function of ICP1, antibiotics and their interactions. The selection was based on ΔAIC. We report ΔAIC, R syntax for the formula, the predictors (fixed effects) and the corresponding P-values (Chi-square test), as well as the* P-*value corresponding to dehydration random effect (RE) and the adjusted r-squared of the model. nt : not tested.*

model	ΔAIC	Formula	ICP1	AZI	ICP1*AZI	Dehyd (RE)	R2
Gam1 (selected)	0	Vc ~ s(ICP1) + s(AZI) + te(ICP1, AZI) + s(Dehydration_Status, bs = "re")	0.614	0.002	0.026	3.66e-07	0.031
Gam2	1.3	Vc ~ s(ICP1) + s(AZI) + s(Dehydration_Status, bs = "re")	0.077	0.004	Not tested	2.76e-07	0.035
Gam3	3.9	Vc ~ s(ICP1) + s(AZI) + s(CIP) + s(DOX) + te(ICP1, AZI) + te(ICP1, CIP) + te(ICP1, DOX) +s(Dehydration_Status, bs = "re")	0.606	0.002	0.025	4.08e-07	0.027
Gam4	5.2	Vc ~ s(ICP1) + s(AZI) + s(CIP) + s(DOX) + s(Dehydration_Status, bs = "re")	0.078	0.003	Not tested	3.06e-07	0.032
Gam5	11.7	Vc ~ s(ICP1) + s(DOX) + te(ICP1, DOX) + s(Dehydration_Status, bs = "re")	0.059	nt	nt	3.65e-08	0.035
Gam6	11.9	Vc ~ s(ICP1) + s(CIP) + te(ICP1, CIP) + s(Dehydration_Status, bs = "re")	0.664	nt	4.35e-08	4.35e-08	0.022
Gam7	12.4	Vc ~ s(ICP1) + s(AZI) + te(ICP1, AZI, by = Dehydration_Status)					
Gam1(B)[a]	29.3	Vc ~ s(ICP1) + s(AZI) + te(ICP1, AZI)	0.914	0.003	0.015	nt	-0.017

[a] Without dehydration RE

Table S7. Generalized linear mixed models (set 2). *GLMMs with* Vc *SNV count as a function of phage and antibiotics and their interaction. We report ΔAIC, the R syntax for the formula, the predictors and the corresponding P-values (Wald test). The adjusted r-squared and* P*-value from the comparison of the models and null models (exactly the same model but with no fixed terms) (LRT, anova function from stats package in R). Antibiotics terms are not shown (not significant,* P*>0.05, GLMM, wald test). nt : not tested.*

Model	**dAIC**	**Formula**	**Vc**	**ICP1**	**ICP1*Vc**	**R2**	**P/null model**
Mod1 (Selected)	0.0	SNV_nb ~ Vc + Vc:ICP1	0.002	nt	0.004	0.377	7e-04
Mod2	0.6	SNV_nb ~ ICP1 + Vc + Vc:ICP1	0.007	0.266	0.012	0.408	1.19e-03
Mod3	1.6	SNV_nb ~ ICP1 + Vc + AZI + Vc:ICP1	0.008	0.267	0.015	0.409	0.002
Mod4	3.2	SNV_nb ~ ICP1 + Vc + CIP + AZI + Vc:ICP1	0.007	0.273	0.015	0.412	0.004
Mod5	5.0	SNV_nb ~ Vc	0.002	nt	nt	0.228	0.006
Mod6	5.2	SNV_nb ~ ICP1 + Vc + CIP + DOX + AZI + Vc:ICP1	0.00745	0.27359	0.01554	0.412	0.008
Mod6	5.9	SNV_nb ~ ICP1 + Vc + CIP + DOX + AZI + ICP1:CIP + ICP1:AZI + Vc:ICP1	0.0428	0.1634	0.0385	0.458	0.008
Mod7	7.0	SNV_nb ~ ICP1 + Vc + CIP + AZI + DOX + Vc:ICP1:CIP + Vc:ICP1:AZI + Vc:ICP1:DOX	0.015	0.732	nt	0.442	0.012
Mod8	7.1	SNV_nb ~ ICP1 + Vc + CIP + DOX + AZI + ICP1:AZI + Vc:ICP1	0.047	0.275	0.059	0.422	0.015
Mod9	12.1	SNV_nb ~ ICP1	nt	0.506	nt	0.018	0.527

Table S8. Generalized additive models (set 3). *GAMs used to select the most parsimonious model. The response is the average frequency of NS mutations in* V. cholerae *and the predictors are ICP1, antibiotics, ICE presence/absence and mutation type (NS: non-synonymous, S:synonymous and I:intergenic) as well as their interactions. We defined a variable (ICE.by.mut) as the combination between ICE and mutation type with 9 levels (mutation type*ICE) to represent the interaction between mutation type and ICE factors. Model selection was based on ΔAIC. We report ΔAIC, the R syntax for the formula, the predictors and the corresponding P-values (Chi-square test) and the adjusted r-squared of the model. nt: not tested.*

	ΔAIC	**formula**	***P* (ICP1*ICE*mut)**			***P* (antbx*mut)**	**R2**
			Ind5*NS	Ind6*NS	ICE-*NS		
Mod1	0.0	mean_freq ~ s(ICP1, by = ICE.by.mut)	0.1336	0.642	0.0469	nt	0.021
Mod2	5.8	mean_freq ~ s(ICP1, by = ICE.by.mut) + s(AZI, by = mutation_type) + s(CIP, by = mutation_type) + DOX * mutation_type	0.13242	0.58891	0.04677	> 0.05	0.060
Mod3	9.0	mean_freq ~ s(ICP1, by = ICE.by.mut) + s(AZI, by = mutation_type) + s(CIP, by = mutation_type) + DOX * mutation_type + te(ICP1, AZI, by = mutation_type) + te(ICP1, CIP, by = mutation_type) + DOX:ICP1:mutation_type	<2e-16	<2e-16	<2e-16	<2e-16 (AZI)	0.067
Mod4	10.9	mean_freq ~ s(ICP1, by = ICE.by.mut) + s(AZI, by = mutation_type) + s(CIP, by = mutation_type)	0.1223	0.6346	0.0496	> 0.05	0.002

Table S9. Top 10 genes with high frequency nonsynonymous mutations when V. cholerae>ICP1. *In bold, genes mutated only when V. cholerae>ICP1. 59 patients had high frequency non-synonymous SNVs.*

COG category (eggnog)	PFAM (eggnog) annotation	NCBI annotation	Mutation count	Patient count
O	PPC_Peptidase_M9_Peptidase_M9_N	TPA: collagenase HAS4622795.1	56	15
MQ	**ACD_ADPrib_exo_Tox_Anthrax_toxA_Hydrolase_4_MLD_Peptidase_C80_RtxA**	**TPA: MARTX multifunctional-autoprocessing repeats-in-toxin holotoxin RtxA HAS4620517.1**	**14**	**6**
PT	**GGDEF_Hemerythrin**	**GGDEF domain-containing protein WP_001190450.1**	**12**	**4**
T	**EAL_GGDEF**	**EAL domain-containing protein WP_000160066.1**	**11**	**2**
S	**Glyco_hydro_129**	**hypothetical protein VC_A0254 AAF96165.1**	**11**	**2**
S	**Betaprism_lec_Hemolysin_N_Leukocidin**	**Cytolysin vcc (ncbi) WP_001125271.1 (hlyA gene) [Levade et al 2021]**	**10**	**3**
C	FixO	cytochrome-c oxidase, cbb3-type subunit II WP_000097777.1	9	1
L	**Fapy_DNA_glyco_H2TH_zf FPG_IleRS**	**bifunctional DNA-formamidopyrimidine glycosylase/DNA-(apurinic or apyrimidinic site) lyase WP_001114647.1**	**9**	**3**
I	AcylCoA_dh_1_AcylCoA_dh_M_AcylCoA_dh_N_DUF1974	acyl-CoA dehydrogenase FadE WP_000404358.1	8	4
L	Phage_int_SAM_4	hypothetical protein WP_000222725.1	8	4

Table S10. **Genes with high frequency N mutations with a prevalence of 2 mutations or more when ICP1>*V. cholerae*.** In bold, genes mutated only when ICP1>*V. cholerae*. 10 patients had high frequency non-synonymous SNVs.

COG (eggnog)	PFAM (eggnog) annotation	NCBI annotation	Mutation count	Patient count
M	**Peptidase_S13**	D-alanyl-D-alanine carboxypeptidase/D-alanyl-D-alanine-endopeptidase (genbank : AAF93798.1)	**4**	**1**
V	ACR_tran	multidrug resistance protein, putative AAF93795.1	4	1
K	**MerR_1**	**MerR family transcriptional regulator WP_000226962.1**	**3**	**1**
S	DUF3302	**DUF3302 domain-containing protein** WP_000478180.1	2	1
O	PPC_Peptidase_M9_Peptidase_M9_N	TPA: collagenase HAS4622795.1	4	1
P	BPD_transp_1	ABC transporter permease subunit WP_000252168.1	4	1
IQ	PPbinding	Chain A, 3-oxoacyl-[acyl-carrier-protein] synthase 2 PDB: 4JRH_A	2	1
C	**Fer4_12_Radical_SAM**	**methyl-accepting chemotaxis protein WP_000383592.1**	**2**	**2**
Q	FtsX_MacB_PCD	ABC transporter permease WP_000645916.1	2	1
C	**CCG_Fer4_8**	**anaerobic glycerol-3-phosphate dehydrogenase subunit GlpC**	**2**	**1**

		WP_001014995.1		
GM	CoA_binding_3_Polysacc_synt_2	nucleoside-diphosphate sugar epimerase/dehydratase WP_000494952.1	2	1
F	PpxGppA	guanosine-5'-triphosphate,3'-diphosphate diphosphatase WP_000076046.1	2	1
S	**DUF3157**	**DUF3157 family protein WP_000733640.1**	**2**	**1**
C	CCG_FADoxidase_C_FAD_binding_4_Fer4_7_Fer4_8	FAD-binding and (Fe-S)-binding domain-containing protein WP_000188699.1	3	1
Q	HemolysinCabind_Peptidase_M10_C	retention module-containing protein WP_001191814.1	5	2
C	**Gp_dh_C_Gp_dh_N**	**glyceraldehyde 3-phosphate dehydrogenase GenBank: AAF96741.1**	**3**	**1**

***Table S11. Genes with high frequency N mutations with a prevalence of 2 mutations or more when ind5 was* detected, 41 in total.**

NCBI annotation	Mutation count	Patient count
anaerobic nucleoside diphosphate reductase NCBI	4	3
hypothetical protein NCBI	33	3
hypothetical protein NCBI	3	3
nicotinate phosphoribosyltransferase NCBI	3	3
hypothetical protein NCBI	4	3
hypothetical protein NCBI	5	2
hypothetical protein NCBI	4	2
putative homing endonuclease NCBI	2	2
hypothetical protein NCBI	2	2
hypothetical protein NCBI	2	2
ribonucleoside diphosphate reductase, beta chain NCBI	2	2
hypothetical protein NCBI	2	2
putative baseplate component NCBI	2	2
putative distal tail protein NCBI	2	2

Table S12. Genes with high frequency N mutations when no ICE was detected (all genes).

NCBI annotation	Mutation count	Patient count
hypothetical protein NCBI	5	5
hypothetical protein NCBI	9	5
NrdD-like anaerobic ribonucleotide reductase large subunit NCBI	4	2
hypothetical protein NCBI	2	2
hypothetical protein NCBI	2	2
hypothetical protein NCBI	2	2
hypothetical protein NCBI	6	2
hypothetical protein NCBI	1	1
anaerobic nucleoside diphosphate reductase NCBI	4	1

hypothetical protein NCBI	6	1
hypothetical protein NCBI	3	1
putative homing endonuclease NCBI	2	1
hypothetical protein NCBI	14	1
hypothetical protein NCBI	2	1

Table S13. Genes with high frequency N mutations when ind6 was detected (all genes).

NCBI annotation	Mutation count	Patient count
hypothetical protein NCBI	3	3
anaerobic nucleoside diphosphate reductase NCBI	1	1
hypothetical protein NCBI	1	1
putative tail fiber NCBI	1	1
tail sheath NCBI	2	1
hypothetical protein NCBI	1	1
hypothetical protein NCBI	3	1

SUPPLEMENTARY FILES

These files are available in the biorxiv version of the paper: https://www.biorxiv.org/content/10.1101/2023.06.14.544933v1

File S1. Patient metadata.

File S2. Full list of genes containing mutations in samples where *Vc* > ICP1 (% of reads).

File S3. Full list of genes containing mutations in samples where *Vc* < ICP1 (% of reads).

File S4. Antibiotic concentrations in stool samples.

Conclusion

How biotic interactions affect species diversity has received less attention than the abiotic drivers of diversity, and this is particularly true for bacteria. With the exception of a few experimental studies that tracked the diversification of a focal species in communities of varying complexity (Brockhurst et al. 2007; Calcagno et al. 2017; Jousset et al. 2016a; Gómez and Buckling 2013), this question has not been further studied. Moreover, most laboratory experiments are restricted to relatively short evolutionary time scales and include only a small number of bacterial taxa. Because of this major knowledge gap, it is unclear if the dynamics observed in these laboratory-scale systems can be expanded to natural bacterial communities, which undergo more complex ecological interactions over a longer period of time. Filling this gap is critical to challenge the universal character of drivers underpinning microbial diversity. The two first studies of my PhD thesis contributed to addressing this knowledge gap.

In **Chapter 1**, I used 16S rRNA gene amplicons sequences from the Earth Microbiome Project (EMP) data, to demonstrate a general positive relationship between community diversity and within-taxa diversity at taxonomic levels from phylum to genus in a broad range of environments. However, this positive trend plateaus at high community diversity as niches become filled. Furthermore, controlling for environmental variables in soil microbial communities, I demonstrated the increasing impact of the abiotic variables on focal-lineage diversity in more diverse microbiomes as the DBD slope decreased.

The empirical support for the DBD model in the EMP data brings new insights on eco-evolutionary processes driving biodiversity of natural microbial communities; however, these processes observed at higher taxonomic levels could not be generalized to finer intra-species level due to the limited resolution of the 16S rRNA gene data.

I was able to increase the genetic resolution in **Chapter 2**. Using higher resolution metagenomic data from the Human Microbiome Project (HMP), I looked at within-species diversity and demonstrated a positive correlation between gut microbiome diversity and within-species polymorphism and strain number. This study provided evidence that the DBD hypothesis holds at within-species resolution, consistent with another recent study which also supported a

DBD model in the human gut, inferred to be driven by resource competition (Good and Rosenfeld 2022).

Besides DBD, I found evidence of another eco-evolutionary model playing in the gut. Tracking gene gain and loss events in the HMP longitudinal data, revealed that genome reduction in the gut is more prevalent in more diverse gut communities, consistent with the Black Queen Hypothesis (BQH). This could be due to *de novo* gene loss or the preferential establishment of migrant strains encoding fewer genes.

Chapter 3 advances our knowledge of how interactions between *V. cholerae* and its virulent phage ICP1 affect the infection dynamics and *V. cholerae* genetic diversity during human infection. To my knowledge, these interactions have been studied in the laboratory, *in silico* with mathematical models, and to a lesser extent in the field but never during human infection. Furthermore, only a very few earlier studies based on small datasets, addressed within-patient *V. cholerae* diversity (Seed et al. 2014; Levade et al. 2021). My work shows that higher levels of ICP1 relative to *V. cholerae* were associated with mild dehydration and that the evolution and adaptation of *V. cholerae* within hosts is quite a common phenomenon in cholera patients. These findings have important implications for understanding *V. cholerae*-phage dynamics in natural infections and may potentially be useful in phage therapy research.

As predicted by laboratory experiments showing that phage resistance elements on ICEs can protect against ICP1 (LeGault et al. 2021), we found that patient samples where ICEs were not detected were associated with higher phage to *V. cholerae* ratios. In the absence of detectable ICE, ICP1 was associated with increased rates of nonsynonymous point mutations in the *V. cholerae* genome. Many of these mutations likely arose by hypermutation.

Antibiotic exposure was not associated with increased rates of point mutations but was associated with less *V. cholerae* and less severe disease. Azithromycin appeared to be particularly effective at suppressing *V. cholerae* and was not associated with any known resistance genes in metagenomes.

Chapter 3 revealed that both phages and antibiotics are determinants of cholera disease severity and paves the way for future inquiries into their interacting impact on disease progression and recovery. My work also suggests a hierarchy of selective pressures acting on *V.*

cholerae in the gut: antibiotics are effective at suppressing cholera in the absence of resistance genes, ICP1 suppress *V. cholerae* in the absence of ICEs and effective antibiotics; and finally, point mutations conferring phage resistance arose under phage selective pressure in the absence of ICEs.

Together, my PhD thesis has advanced current understanding of how eco-evolutionary feedbacks shape complex microbial communities. It has also shed light on how inter-species or phage-bacteria interactions affect evolution in nature.

Future directions

Further studies will be required to complement and expand upon my work. Some of the possible topics are listed below.

1. Clarify the nature of the underlying mechanisms of DBD and BQH and how these two models might co-occur and interact in natural communities. Higher resolution genomics, metabolomics, and complementary experiments are suggested techniques to overcome some of the limitations of 16S rRNA gene amplicon sequencing and shotgun metagenomics. The most recent publication from the Earth Microbiome Project includes metabolomic and metagenomic data (Shaffer et al. 2022) that could provide insights into this question, in addition to generalizing the dynamics seen in the human gut (**Chapter 2**) to a broader range of environments.
2. Investigate how DBD dynamics are influenced by abiotic factors and by interactions with other microbes such as fungi, archaea, and phages in the human gut and other environments.
3. Disentangle between *de novo* mutations within hosts and co-infection by different lineages of *V. cholerae*. Longitudinal data from the same patient, along with long read or whole genome sequencing, may be required.
4. Demonstrate causality in the dynamics between ICP1 and *V. cholerae* and their effect on disease severity. Time series or interventional experiments would be potential ways to achieve this goal.

5. Infer the interactions between *V. cholerae* and the less abundant virulent phages ICP2 and ICP3. Whole genome sequencing of isolates, single-cell sequencing, or long-read metagenomics may be needed to overcome shotgun metagenomic sequencing sampling only the most abundant taxa.

References

Abedon, S. T. 2012. 'Bacterial 'immunity' against bacteriophages', *Bacteriophage*, 2: 50-54.

Albalat, R., and C. Cañestro. 2016. 'Evolution by gene loss', *Nat Rev Genet. doi:10.1038/nrg.2016.39*.

Alexandrova, Ludmila, Farhana Haque, Patricia Rodriguez, Ashton C. Marrazzo, Jessica A. Grembi, Vasavi Ramachandran, Andrew J. Hryckowian, Christopher M. Adams, Md Shah A. Siddique, Ashraful I. Khan, Firdausi Qadri, Jason R. Andrews, Mahmudur Rahman, Alfred M. Spormann, Gary K. Schoolnik, Allis Chien, and Eric J. Nelson. 2019. 'Identification of widespread antibiotic exposure in cholera patients correlates with clinically relevant microbiota changes', *J. Infect. Dis.*

Ali, Mohammad, Allyson R. Nelson, Anna Lena Lopez, and David A. Sack. 2015. 'Updated global burden of cholera in endemic countries', *PLoS Negl. Trop. Dis.*, 9: e0003832.

Allen, J. M., L. J. Mailing, G. M. Niemiro, R. Moore, M. D. Cook, B. A. White, H. D. Holscher, and J. A. Woods. 2018. 'Exercise Alters Gut Microbiota Composition and Function in Lean and Obese Humans', *Med Sci Sports Exerc*, 50: 747-57.

Alonso, D., and A. J. McKane. 2004. 'Sampling Hubbell's neutral theory of biodiversity: Sampling neutral theory', *Ecology letters*, 7(10), 901–910.

Anderson, G., M. Seo, M. Berk, A. F. Carvalho, and M. Maes. 2016. 'Gut Permeability and Microbiota in Parkinson's Disease: Role of Depression, Tryptophan Catabolites, Oxidative and Nitrosative Stress and Melatonergic Pathways', *Curr Pharm Des*, 22: 6142-51.

Andrews, Jason R., Daniel T. Leung, Shahnawaz Ahmed, Mohammed Abdul Malek, Dilruba Ahmed, Yasmin Ara Begum, Firdausi Qadri, Tahmeed Ahmed, Abu Syed Golam Faruque, and Eric J. Nelson. 2017. 'Determinants of severe dehydration from diarrheal disease at hospital presentation: Evidence from 22 years of admissions in Bangladesh', *PLoS Negl. Trop. Dis.*, 11: e0005512.

Arango-Argoty, Gustavo, Emily Garner, Amy Pruden, Lenwood S. Heath, Peter Vikesland, and Liqing Zhang. 2018. 'DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data', *Microbiome*, 6: 23.

Arthur, J. C., E. Perez-Chanona, M. Muhlbauer, S. Tomkovich, J. M. Uronis, T. J. Fan, B. J. Campbell, T. Abujamel, B. Dogan, A. B. Rogers, J. M. Rhodes, A. Stintzi, K. W. Simpson, J. J. Hansen, T. O. Keku, A. A. Fodor, and C. Jobin. 2012. 'Intestinal inflammation targets cancer-inducing activity of the microbiota', *Science*, 338: 120-3.

Auguet, J. C., A. Barberan, and E. O. Casamayor. 2010. 'Global ecological patterns in uncultured Archaea', *ISME J*, 4: 182-90.

Averill, C., M. A. Anthony, P. Baldrian, F. Finkbeiner, J. van den Hoogen, T. Kiers, P. Kohout, E. Hirt, G. R. Smith, and T. W. Crowther. 2022. 'Defending Earth's terrestrial microbiome', *Nat Microbiol*, 7: 1717-25.

Baez, A., and J. Shiloach. 2014. 'Effect of elevated oxygen concentration on bacteria, yeasts, and cells propagated for production of biological compounds', *Microb Cell Fact*, 13: 181.

Bailey, S. F., J. R. Dettman, P. B. Rainey, and R. Kassen. 2013. 'Competition both drives and impedes diversification in a model adaptive radiation', *Proc Biol Sci*, 280: 20131253.

Balows, Albert. 2003. 'Manual of clinical microbiology 8th edition: P. R. Murray, E. J. Baron, J. H. Jorgenson, M. A. Pfaller, and R. H. Yolken, eds., ASM Press, 2003, 2113 pages, 2 vol, 2003 + subject & author indices, ISBN: 1-555810255-4, US$ 189.95', *Diagn. Microbiol. Infect. Dis.*, 47: 625.

Barberan, A., K. S. Ramirez, J. W. Leff, M. A. Bradford, D. H. Wall, and N. Fierer. 2014. 'Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria', *Ecol Lett*, 17: 794-802.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. 'Fitting Linear Mixed-Effects Models Using lme4. ', *Journal of Statistical Software*, 67(1), 1–48.

Beaber, John W., Bianca Hochhut, and Matthew K. Waldor. 2002. 'Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from Vibrio cholerae', *J. Bacteriol.*, 184: 4259-69.

Beckman, D. A., and C. M. Waters. 2023. 'Three dominant Vibrio cholerae lytic phage all require O1 antigen for infection', *bioRxiv*.

Bonder, M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, A. V. Vila, P. Deelen, T. Vatanen, M. Schirmer, S. P. Smeekens, D. V. Zhernakova, S. A. Jankipersadsing, M. Jaeger, M. Oosting, M. C. Cenit, A. A. Masclee, M. A. Swertz, Y. Li, V. Kumar, L. Joosten, H. Harmsen, R. K. Weersma, L. Franke, M. H. Hofker, R. J. Xavier, D. Jonkers, M. G. Netea, C. Wijmenga, J. Fu, and A. Zhernakova. 2016. 'The effect of host genetics on the gut microbiome', *Nat Genet*, 48: 1407-12.

Borges, A. L., A. R. Davidson, and J. Bondy-Denomy. 2017. 'The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs', *Annu Rev Virol*, 4: 37-59.

Boyd, Caroline M., Angus Angermeyer, Stephanie G. Hays, Zachary K. Barth, Kishen M. Patel, and Kimberley D. Seed. 2021. 'Bacteriophage ICP1: A Persistent Predator of Vibrio cholerae', *Annu Rev Virol*, 8: 285-304.

Boyd., Caroline M., Angus Angermeyer., Stephanie G. Hays., Zachary K. Barth., Kishen M. Patel., and Kimberley D. Seed. 2021. 'Bacteriophage ICP1: A Persistent Predator of Vibrio cholerae', *The Annual Review of Virology*, 8:285–304.

Brockhurst, M. A., A. Buckling, and P. B. Rainey. 2005. 'The effect of a bacteriophage on diversification of the opportunistic bacterial pathogen, Pseudomonas aeruginosa', *Proc Biol Sci*, 272: 1385-91.

Brockhurst, M. A., N. Colegrave, D. J. Hodgson, and A. Buckling. 2007. 'Niche occupation limits adaptive radiation in experimental microcosms', *PLoS One*, 2: e193.

Broniewski, J. M., S. Meaden, S. Paterson, A. Buckling, and E. R. Westra. 2020. 'The effect of phage genetic diversity on bacterial resistance evolution', *ISME J*, 14: 828-36.

Brooks, M.E., K. Kristensen, K.J. Benthem, A. Magnusson, C.W. Berg, A. Nielsen, H.J. Skaug, M. Mächler, and B.M. Bolker. 2017. 'Modeling zero-inflated count data with glmmTMB', *BioRxiv*.

Calcagno, V., P. Jarne, M. Loreau, N. Mouquet, and P. David. 2017. 'Diversity spurs diversification in ecological communities', *Nature Communications*, 8: 15810.

Cantalapiedra, C. P., A. Hernandez-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas. 2021. 'eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale', *Mol Biol Evol*, 38: 5825-29.

Carr, A., C. Diener, N. S. Baliga, and S. M. Gibbons. 2019a. 'Use and abuse of correlation analyses in microbial ecology', *ISME J*, 13: 2647-55.

Carr, Alex, Christian Diener, Nitin S. Baliga, and Sean M. Gibbons. 2019b. 'Use and abuse of correlation analyses in microbial ecology', *ISME J.*, 13: 2647-55.

Chapman, R., L. Jones, A. D'Angelo, A. Suliman, M. Anwar, and S. Bagby. 2023. 'Nanopore-Based Metagenomic Sequencing in Respiratory Tract Infection: A Developing Diagnostic Platform', *Lung*: 1-9.

Chaumeil, P.A., A.J. Mussig, P. Hugenholtz, and D.H. Parks. 2020. 'GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database', *Bioinformatics 36*.

Chen, L. X., L. N. Huang, C. Mendez-Garcia, J. L. Kuang, Z. S. Hua, J. Liu, and W. S. Shu. 2016. 'Microbial communities, processes and functions in acid mine drainage ecosystems', *Curr Opin Biotechnol*, 38: 150-8.

Coyte, K. Z., J. Schluter, and K. R. Foster. 2015. 'The ecology of the microbiome: Networks, competition, and stability', *Science*, 350: 663-6.

Creasy-Marrazzo, Ashton, Morteza M. Saber, Manasi Kamat, Laura S. Bailey, Lindsey Brinkley, Emilee Cato, Yasmin Begum, Md Mahbubur Rashid, Ashraful I. Khan, Firdausi Qadri, Kari B. Basso, B. Jesse Shapiro, and Eric J. Nelson. 2022. 'Genome-wide association studies reveal distinct genetic correlates and increased heritability of antimicrobial resistance in Vibrio cholerae under anaerobic conditions', *Microb. Genom.*, 8.

Culp, E. J., and A. L. Goodman. 2023. 'Cross-feeding in the gut microbiome: Ecology and mechanisms', *Cell Host Microbe*, 31: 485-99.

Czaran, T. L., R. F. Hoekstra, and L. Pagie. 2002. 'Chemical warfare between microbes promotes biodiversity', *Proc Natl Acad Sci U S A*, 99: 786-90.

D'Herelle, F., and R. H. Malone. 1927. 'A Preliminary Report of Work Carried out by the Cholera Bacteriophage Enquiry', *Ind. Med. Gaz.*, 62: 614-16.

D'Herelle, F ., and R. H Malone. 1927. 'A preliminary report of work carried out by the cholera bacteriophage enquiry.', *The indian medical gazette*.

Dalsgaard, A., A. Forslund, N. V. Tam, D. X. Vinh, and P. D. Cam. 1999. 'Cholera in Vietnam: changes in genotypes and emergence of class I integrons containing aminoglycoside resistance gene cassettes in vibrio cholerae O1 strains isolated from 1979 to 1996', *J. Clin. Microbiol.*, 37: 734-41.

Das, Bhabatosh, Jyoti Verma, Pawan Kumar, Amit Ghosh, and Thandavarayan Ramamurthy. 2020. 'Antibiotic resistance in Vibrio cholerae: Understanding the ecology of resistance genes and mechanisms', *Vaccine*, 38 Suppl 1: A83-A92.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton, and P. J. Turnbaugh. 2014. 'Diet rapidly and reproducibly alters the human gut microbiome', *Nature*, 505: 559-63.

David, Lawrence A., Ana Weil, Edward T. Ryan, Stephen B. Calderwood, Jason B. Harris, Fahima Chowdhury, Yasmin Begum, Firdausi Qadri, Regina C. LaRocque, and Peter J. Turnbaugh. 2015. 'Gut microbial succession follows acute secretory diarrhea in humans', *MBio*, 6: e00381-15.

De Luca, F., and Y. Shoenfeld. 2019. 'The microbiome in autoimmune diseases', *Clin Exp Immunol*, 195: 74-85.

de Wit, R., and T. Bouvier. 2006. '"Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? ', *Environmental Microbiology*, 8(4), 755–758.

Delgado-Baquerizo, M., A. M. Oliverio, T. E. Brewer, A. Benavent-Gonzalez, D. J. Eldridge, R. D. Bardgett, F. T. Maestre, B. K. Singh, and N. Fierer. 2018. 'A global atlas of the dominant bacteria found in soil', *Science*, 359: 320-25.

Dolgin., Elie. 2020. 'Fighting cancer with microbes', *Nature*, 577.

Doron, S., S. Melamed, G. Ofir, A. Leavitt, A. Lopatina, M. Keren, G. Amitai, and R. Sorek. 2018. 'Systematic discovery of antiphage defense systems in the microbial pangenome', *Science*, 359.

Dromigny, Jacques-Albert, Olivat Rakoto-Alson, Davidra Rajaonatahina, René Migliani, Justin Ranjalahy, and Philippe Mauclére. 2002. 'Emergence and rapid spread of tetracycline-resistant Vibrio cholerae strains, Madagascar', *Emerg. Infect. Dis.*, 8: 336-38.

Dufrene, M., and P Legendre. 1997. 'Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach', *Ecological Monographs*, 67(3), 345–366.

Dunlap, Paul V. 2001. 'Microbial Diversity', *University ofMarylandBiotechnologyInstitute,Baltimore,MD,USA*.

Edgar, R.C. 2010. 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics 26*.

Elston, D. A., R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin. 2001. 'Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks', *Parasitology*, 122(Pt 5), 563–569.

Elton, C. 1946. 'Competition and the Structure of Ecological Communities', *The Journal of Animal Ecology*, 15(1), 54–68. https://doi.org/10.2307/1625.

Emerson, B. C., and N. Kolm. 2005. 'Species diversity can drive speciation', *Nature*, 434(7036), 1015–1017.

Estrela, S., J. Diaz-Colunga, J.C.C. Vila, A. Sanchez-Gorostiaga, and A. Sanchez. 2022. 'Diversity begets diversity under microbial niche construction', *BioRxiv*.

Falkowski, P. G., T. Fenchel, and E. F. Delong. 2008. 'The microbial engines that drive Earth's biogeochemical cycles', *Science*, 320: 1034-9.

Fan, X., A. V. Alekseyenko, J. Wu, B. A. Peters, E. J. Jacobs, S. M. Gapstur, M. P. Purdue, C. C. Abnet, R. Stolzenberg-Solomon, G. Miller, J. Ravel, R. B. Hayes, and J. Ahn. 2018. 'Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study', *Gut*, 67: 120-27.

Faruque, Shah M., M. Johirul Islam, Qazi Shafi Ahmad, A. S. G. Faruque, David A. Sack, G. Balakrish Nair, and John J. Mekalanos. 2005. 'Self-limiting nature of seasonal cholera epidemics: Role of host-mediated amplification of phage', *Proc. Natl. Acad. Sci. U. S. A.*, 102: 6119-24.

Faruque, Shah M., Iftekhar Bin Naser, M. Johirul Islam, A. S. G. Faruque, A. N. Ghosh, G. Balakrish Nair, David A. Sack, and John J. Mekalanos. 2005. 'Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages', *Proc. Natl. Acad. Sci. U. S. A.*, 102: 1702-07.

Faruque, Shah M., Iftekhar Bin. Naser, M. Johirul. Islam , A. S. G. Faruque, A. N. Ghosh, G. Balakrish Nair, David A. Sack, and John J. Mekalanos. 2005a. 'Seasonal epidemics of

cholera inversely correlate with the prevalence of environmental cholera phages', *PNAS*, 102.

Frank, D. N., A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace. 2007. 'Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases', *Proc Natl Acad Sci U S A*, 104: 13780-5.

Galand, P.E., O. Pereira, C. Hochart, J.C. Auguet, and D. Debroas. 2018. 'A strong link between marine microbial community composition and function challenges the idea of functional redundancy', *ISME Journal 12*.

Garud, N. R., and K. S. Pollard. 2020. 'Population Genetics in the Human Microbiome', *Trends Genet*, 36: 53-67.

Garud, N.R., B.H. Good, O. Hallatschek, and K.S. Pollard. 2019. 'Evolutionary dynamics of bacteria in the gut microbiome within and across hosts', *PLoS Biol 17:e3000102*.

Gause, G.F. 2003. 'The Struggle for Existence', *Williams & Wilkins, Baltimore*, 1934.

Ghosh, T. S., S. S. Gupta, T. Bhattacharya, D. Yadav, A. Barik, A. Chowdhury, B. Das, S. S. Mande, and G. B. Nair. 2014. 'Gut microbiomes of Indian children of varying nutritional status', *PLoS One*, 9: e95547.

Gomez, P., and A. Buckling. 2013. 'Real-time microbial adaptive diversification in soil', *Ecol Lett*, 16: 650-5.

Gómez, P., and A. Buckling. 2013. 'Real-time microbial adaptive diversification in soil', *Ecology letters*, 16(5), 650–655.

Good, B.H., and L.B. Rosenfeld. 2022. 'Eco-evolutionary feedbacks in the human gut microbiome', *BioRxiv*.

Gotelli, N. J., and R. K. Colwell. 2001. 'Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters, 4(4), 379–391', *Ecology letters*, 4: 379-391.

Goyal, A. 2021. 'Horizontal Gene Transfer Drives the Evolution of Dependencies in Bacteria', *SSRN Electronic Journal*.

Goyal, A., L.S. Bittleston, G.E. Leventhal, L. Lu, and O.X. Cordero. 2022. . 'Interactions between strains govern the eco-evolutionary dynamics of microbial communities', *Elife 11*.

Groussin, M., M. Poyet, A. Sistiaga, S. M. Kearney, K. Moniz, M. Noel, J. Hooker, S. M. Gibbons, L. Segurel, A. Froment, R. S. Mohamed, A. Fezeu, V. A. Juimo, S. Lafosse, F. E. Tabe, C. Girard, D. Iqaluk, L. T. T. Nguyen, B. J. Shapiro, J. Lehtimaki, L. Ruokolainen, P. P. Kettunen, T. Vatanen, S. Sigwazi, A. Mabulla, M. Dominguez-Rodrigo, Y. A. Nartey, A. Agyei-Nkansah,

A. Duah, Y. A. Awuku, K. A. Valles, S. O. Asibey, M. Y. Afihene, L. R. Roberts, A. Plymoth, C. A. Onyekwere, R. E. Summons, R. J. Xavier, and E. J. Alm. 2021. 'Elevated rates of horizontal gene transfer in the industrialized human microbiome', *Cell*, 184: 2053-67 e18.

Guo, R., L. H. Chen, C. Xing, and T. Liu. 2019. 'Pain regulation by gut microbiota: molecular mechanisms and therapeutic potential', *Br J Anaesth*, 123: 637-54.

Hajishengallis, G. 2015. 'Periodontitis: from microbial immune subversion to systemic inflammation', *Nat Rev Immunol*, 15: 30-44.

Hardin, J.W., and J.M. Hilbe. 2018. 'Generalized Linear Models and Extensions', *4th ed. Stata Press*.

Harris, J. B., R. C. LaRocque, F. Qadri, E. T. Ryan, and S. B. Calderwood. 2012. 'Cholera', *Lancet*, 379: 2466-76.

Harris, Jason B., Regina C. LaRocque, Fahima Chowdhury, Ashraful I. Khan, Tanya Logvinenko, Abu S. G. Faruque, Edward T. Ryan, Firdausi Qadri, and Stephen B. Calderwood. 2008. 'Susceptibility to Vibrio cholerae Infection in a Cohort of Household Contacts of Patients with Cholera in Bangladesh', *PLoS Negl. Trop. Dis.*, 2: e221.

Harris, K., T. L. Parsons, U. Z. Ijaz, L. Lahti, I. Holmes, and C Quince. 2017. 'Linking Statistical and Ecological Theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process', *Proceedings of the IEEE*, 105(3), 516–529.

He, Y., B.J. Zhou, G.H. Deng, X.T. Jiang, H. Zhang, and H.W. Zhou. 2013. 'Comparison of microbial diversity determined with the same variable tag sequence extracted from two different PCR amplicons', *BMC Microbiol 13*.

Hehemann, J.H., G. Correc, T. Barbeyron, W. Helbert, M. Czjzek, and G. Michel. 2010. 'Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota', *Nature 464:908–912*.

Hibbing, M. E., C. Fuqua, M. R. Parsek, and S. B. Peterson. 2010. 'Bacterial competition: surviving and thriving in the microbial jungle', *Nat Rev Microbiol*, 8: 15-25.

Hubbell, S. P. 2001. 'The Unified Neutral Theory of Biodiversity and Biogeography', *Princeton University Press*.

Hug, L. A., B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. 2016. 'A new view of the tree of life', *Nat Microbiol*, 1: 16048.

Human Microbiome Project, Consortium. 2012. 'A framework for human microbiome research', *Nature*, 486: 215-21.

Hyatt, D., G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. 2010. 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11: 119.

Ibarra-Chavez, R., M. F. Hansen, R. Pinilla-Redondo, K. D. Seed, and U. Trivedi. 2021. 'Phage satellites and their emerging applications in biotechnology', *FEMS Microbiol Rev*, 45.

Islam, M. S., A. K. Siddique, A. Salam, K. Akram, R. N. Majumdar, K. Zaman, N. Fronczak, and S. Laston. 1995. 'Microbiological investigation of diarrhoea epidemics among Rwandan refugees in Zaire', *Trans. R. Soc. Trop. Med. Hyg.*, 89: 506.

Jaiswal, A., H. Koley, A. Ghosh, A. Palit, and B. Sarkar. 2013a. 'Efficacy of cocktail phage therapy in treating Vibrio cholerae infection in rabbit model', *Microbes Infect*, 15: 152-6.

Jaiswal, Abhishek, Hemanta Koley, Amit Ghosh, Anup Palit, and Banwarilal Sarkar. 2013b. 'Efficacy of cocktail phage therapy in treating Vibrio cholerae infection in rabbit model', *Microbes Infect.*, 15: 152-56.

Jarvinen, O. 1982. 'Species-To-Genus Ratios in Biogeography: A Historical Note. Journal of Biogeography', *Journal of Biogeography*, 9(4), 363–370.

Jensen, A. Mark., Shah. M. Faruque, John Mekalanos, J., and Bruce. R. Levin. 2006a. 'Modeling the role of bacteriophage in the control of cholera outbreaks', *PNAS*, vol. 103, no.12, 4652-4657.

Jensen, Mark A., Shah M. Faruque, John J. Mekalanos, and Bruce R. Levin. 2006b. 'Modeling the role of bacteriophage in the control of cholera outbreaks', *Proc. Natl. Acad. Sci. U. S. A.*, 103: 4652-57.

Johnson, C. M., and A. D. Grossman. 2015. 'Integrative and Conjugative Elements (ICEs): What They Do and How They Work', *Annu Rev Genet*, 49: 577-601.

Johnson, C. M., M. M. Harden, and A. D. Grossman. 2022. 'Interactions between mobile genetic elements: An anti-phage gene in an integrative and conjugative element protects host cells from predation by a temperate bacteriophage', *PLoS Genet*, 18: e1010065.

Johnson, P.C. 2014. 'Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models.', *Methods in Ecology and Evolution / British Ecological Society,*, 5(9), 944–946.

Jolivet-Gougeon, Anne, Bela Kovacs, Sandrine Le Gall-David, Hervé Le Bars, Latifa Bousarghin, Martine Bonnaure-Mallet, Bernard Lobel, François Guillé, Claude-James Soussy, and Peter Tenke. 2011. 'Bacterial hypermutation: clinical implications', *J. Med. Microbiol.*, 60: 563-73.

Jousset, A., N. Eisenhauer, M. Merker, N. Mouquet, and S. Scheu. 2016a. 'High functional diversity stimulates diversification in experimental microbial communities', *Sci Adv*, 2: e1600124.

Jousset, A., N. Eisenhauer, M. Merker, N. Mouquet, and S. Scheu. 2016b. 'High functional diversity stimulates diversification in experimental microbial communities.', *Science Advances*, 2(6), e1600124.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. 'MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities', *PeerJ*, 3: e1165.

Kastman, E. K., N. Kamelamela, J. W. Norville, C. M. Cosetta, R. J. Dutton, and B. E. Wolfe. 2016. 'Biotic Interactions Shape the Ecological Distributions of Staphylococcus Species. ', *mBio, 7(5)*.

Kennedy, A. C, and L. Z de Luna. 2005. 'Rhizhosphere', *In D. Hillel (Ed.), Encyclopedia of Soils in the Environment (pp. 399–406). Elsevier.*

Khan, Ashraful I., Jasmine A. Mack, M. Salimuzzaman, Mazharul I. Zion, Hasnat Sujon, Robyn L. Ball, Stace Maples, Md Mahbubur Rashid, Mohammod J. Chisti, Shafiqul A. Sarker, Debashish Biswas, Raduan Hossin, Kevin L. Bardosh, Yasmin A. Begum, Azimuddin Ahmed, Dane Pieri, Farhana Haque, Mahmudur Rahman, Adam C. Levine, Firdausi Qadri, Meerjady S. Flora, Matthew J. Gurka, and Eric J. Nelson. 2020. 'Electronic decision support and diarrhoeal disease guideline adherence (mHDM): a cluster randomised controlled trial', *Lancet Digit Health*, 2: e250-e58.

Khan, Ashraful Islam, Md Mahbubur Rashid, Md Taufiqul Islam, Mokibul Hassan Afrad, M. Salimuzzaman, Sonia Tara Hegde, Md Mazharul I. Zion, Zahid Hasan Khan, Tahmina Shirin, Zakir Hossain Habib, Iqbal Ansary Khan, Yasmin Ara Begum, Andrew S. Azman, Mahmudur Rahman, John David Clemens, Meerjady Sabrina Flora, and Firdausi Qadri. 2020. 'Epidemiology of Cholera in Bangladesh: Findings From Nationwide Hospital-based Surveillance, 2014-2018', *Clin. Infect. Dis.*, 71: 1635-42.

Kim, Y. S., T. Unno, B. Y. Kim, and M. S. Park. 2020. 'Sex Differences in Gut Microbiota', *World J Mens Health*, 38: 48-60.

Konstantinidis, K. T, and J. M Tiedje. 2005. 'Towards a genome-based taxonomy for prokaryotes. ', *Journal of Bacteriology, 187(18), 6258–6264.*

Koopman, M., S. El Aidy, and M. IDtrauma consortium. 2017. 'Depressed gut? The microbiota-diet-inflammation trialogue in depression', *Curr Opin Psychiatry*, 30: 369-77.

Korem, T., D. Zeevi, J. Suez, A. Weinberger, T. Avnit-Sagi, M. Pompan-Lotan, E. Matot, G. Jona, A. Harmelin, N. Cohen, A. Sirota-Madi, C. A. Thaiss, M. Pevsner-Fischer, R. Sorek, R. Xavier, E. Elinav, and E. Segal. 2015. 'Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples', *Science*, 349: 1101-06.

Koskiniemi, S., S. Sun, O.G. Berg, and D.I. Andersson. 2012. 'Selection-driven gene loss in bacteria', *PLoS Genet 8*.

Kuo, C H, and Ochman. 2009b. 'Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. ', *Biology Direct, 4, 35.*

Kuo, C.-H, and H Ochman. 2009a 'Deletional bias across the three domains of life. ', *Genome Biology and Evolution, 1, 145–152.*

Kurapova., A. I., G. M. Zenova., I. I. Sudnitsyn., A. K. Kizilova., N. A. Manucharova., Zh. Norovsuren., and D. G. Zvyagintsev. 2010. 'Thermotolerant and Thermophilic Actinomycetes from Soils of Mongolia Desert Steppe Zone', *Microbiology*, Vol. 81, No. 1, pp. 98–108.

Labrie, S. J., J. E. Samson, and S. Moineau. 2010. 'Bacteriophage resistance mechanisms', *Nat Rev Microbiol*, 8: 317-27.

Laland, K. N, F. J Odling-Smee, and M. W Feldman. 1999. 'Evolutionary consequences of niche construction and their implications for ecology.', *Proceedings of the National Academy of Sciences of the United States of America, 96(18), 10242–10247*.

Langmead, B., and S.L. Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nat Methods 9*.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biol*, 10: R25.

Lapierre, P, and J. P Gogarten. 2009. ' Estimating the size of the bacterial pan-genome. ', *Trends in Genetics: TIG, 25(3), 107–110.*

Lassalle, Florent, Salah Al-Shalali, Mukhtar Al-Hakimi, Elisabeth Njamkepo, Ismail Mahat Bashir, Matthew J. Dorman, Jean Rauzier, Grace A. Blackwell, Alyce Taylor-Brown, Mathew A. Beale, Ali Abdullah Al-Somainy, Anas Al-Mahbashi, Khaled Almoayed, Mohammed Aldawla, Abdulelah Al-Harazi, Marie-Laure Quilici, François-Xavier Weill, Ghulam Dhabaan, and Nicholas R. Thomson. 2022. 'Genomic epidemiology of the cholera outbreak in Yemen reveals the spread of a multi-drug resistance plasmid between diverse lineages of Vibrio cholerae', *bioRxiv*.

Lauber, C. L, M Hamady, Knight R, and N Fierer. 2009. 'Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment.', *Applied and Environmental Microbiology. 75, 5111-5120.*

LeGault, K. N., Z. K. Barth, P. DePaola, and K. D. Seed. 2022. 'A phage parasite deploys a nicking nuclease effector to inhibit viral host replication', *Nucleic Acids Res*, 50: 8401-17.

LeGault, K. N., S. G. Hays, A. Angermeyer, A. C. McKitterick, F. T. Johura, M. Sultana, T. Ahmed, M. Alam, and K. D. Seed. 2021. 'Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts', *Science*, 373.

Leidenfrost, R. M., D. C. Pother, U. Jackel, and R. Wunschiers. 2020. 'Benchmarking the MinION: Evaluating long reads for microbial profiling', *Sci Rep*, 10: 5125.

Lenski, R.E. 2017. 'Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations', *ISME Journal*.

Lerouge, I., and J. Vanderleyden. 2002. 'O-antigen structural variation: mechanisms and possible roles in animal/plant-microbe interactions', *FEMS Microbiol Rev*, 26: 17-47.

Levade, Inès, Ashraful I. Khan, Fahima Chowdhury, Stephen B. Calderwood, Edward T. Ryan, Jason B. Harris, Regina C. LaRocque, Taufiqur R. Bhuiyan, Firdausi Qadri, Ana A. Weil, and B. Jesse Shapiro. 2021. 'A Combination of Metagenomic and Cultivation Approaches Reveals Hypermutator Phenotypes within Vibrio cholerae-Infected Patients', *mSystems*, 6: e0088921.

Levade, Inès, Yves Terrat, Jean-Baptiste Leducq, Ana A. Weil, Leslie M. Mayo-Smith, Fahima Chowdhury, Ashraful I. Khan, Jacques Boncy, Josiane Buteau, Louise C. Ivers, Edward T. Ryan, Richelle C. Charles, Stephen B. Calderwood, Firdausi Qadri, Jason B. Harris, Regina C. LaRocque, and B. Jesse Shapiro. 2017. 'Vibrio cholerae genomic diversity within and between patients', *Microb Genom*, 3.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics*, 31: 1674-76.

Li, L., and Z. S. Ma. 2016. 'Testing the Neutral Theory of Biodiversity with Human Microbiome Datasets', *Scientific Reports, 6, 31448.*

Lindow, S. E, and M. T Brandl. 2003. 'Microbiology of the phyllosphere', *Applied and Environmental Microbiology, 69(4), 1875–1883.*

Lloyd-Price, J., A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A. B. Hall, A. Brady, H. H. Creasy, C. McCracken, M. G. Giglio, D. McDonald, E. A. Franzosa, R. Knight, O. White, and C. Huttenhower. 2017. 'Strains, functions and dynamics in the expanded Human Microbiome Project', *Nature*, 550: 61-66.

Loftus, M., S. A. Hassouneh, and S. Yooseph. 2021. 'Bacterial associations in the healthy human gut microbiome across populations', *Sci Rep*, 11: 2828.

Louca, S., F. Mazel, M. Doebeli, and L. W. Parfrey. 2019. 'A census-based estimate of Earth's bacterial and archaeal diversity', *PLoS Biology, 17(2), e3000106.*

Louca, S., and M. W. Pennell. 2020. 'Extant timetrees are consistent with a myriad of diversification histories', *Nature, 580(7804), 502–505.*

Louca, S., P. M. Shih, M. W. Pennell, W. W. Fischer, L. W. Parfrey, and M. Doebeli. 2018. 'Bacterial diversification through geological time', *Nature Ecology & Evolution, 2(9), 1458–1467.*

Lozupone, C. A, and R Knight. 2007. 'Global patterns in bacterial diversity', *Proceedings of the National Academy of Sciences of the United States of America, 104(27), 11436–11440*.

Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. 'Bracken: estimating species abundance in metagenomics data', *PeerJ Comput. Sci.*, 3: e104.

MacIntyre, D. L., S. T. Miyata, M. Kitaoka, and S. Pukatzki. 2010. 'The Vibrio cholerae type VI secretion system displays antimicrobial properties', *Proc Natl Acad Sci U S A*, 107: 19520-4.

Mackelprang, R., A. Burkert, M. Haw, T. Mahendrarajah, C. H. Conaway, T. A. Douglas, and M. P. Waldrop. 2017. 'Microbial survival strategies in ancient permafrost: insights from metagenomics', *ISME J*, 11: 2305-18.

Madi, N., E. T. Cato, M. A. Sayeed, K. Islam, M. I. U. Khabir, M. T. R. Bhuiyan, Y. Begum, M. M. Rashid, A. Creasy-Marrazzo, L. Brinkley, M. Kamat, A. Cuenod, L. S. Bailey, K. B. Basso, F. Qadri, A. I. Khan, B. J. Shapiro, and E. J. Nelson. 2023b. 'Phage predation and antibiotic exposure are inversely associated with disease severity and shape pathogen genetic diversity in cholera patients', *bioRxiv*.

Madi, N., D. Chen, R. Wolff, B. J. Shapiro, and N. R. Garud. 2023a. 'Community diversity is associated with intra-species genetic diversity and gene loss in the human gut microbiome', *Elife*, 12.

Madi, N., M. Vos, C. L. Murall, P. Legendre, and B. J. Shapiro. 2020. 'Does diversity beget diversity in microbiomes?', *Elife*, 9.

Madigan, Michael T. 2000. 'Extremophilic bacteria and microbial diversity', *Annals of the Missouri Botanical Garden*, Vol. 87, No. 1.

Marshall, C. R. . 2017. 'Five palaeobiological laws needed to understand the evolution of the living biota', *Nature Ecology & Evolution, 1(6), 165.*

Martiny, J.B.H., S.E. Jones, J.T. Lennon, and A.C. Martiny. 2015. 'Microbiomes in light of traits: A phylogenetic perspective', *Science (1979)*.

McCutcheon, J.P., and N.A. Moran. 2012. 'Extreme genome reduction in symbiotic bacteria', *Nat Rev Microbiol*.

McKitterick, A. C., and K. D. Seed. 2018. 'Anti-phage islands force their target phage to directly mediate island excision and spread', *Nature Communications*, 9: 2348.

Meyer, J. R, and R Kassen. 2007. 'The effects of competition and predation on diversification in a model adaptive radiation', *Nature, 446(7134), 432–435*.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems', *Genome Biol*, 12: R112.

Mitri, S., and k. Richard Foster. 2013. 'The genotypic view of social interactions in microbial communities', *Annu Rev Genet*.

Monir, Md Mamun, Mohammad Tarequl Islam, Razib Mazumder, Dinesh Mondal, Kazi Sumaita Nahar, Marzia Sultana, Masatomo Morita, Makoto Ohnishi, Anwar Huq, Haruo Watanabe, Firdausi Qadri, Mustafizur Rahman, Nicholas Thomson, Kimberley Seed, Rita R. Colwell, Tahmeed Ahmed, and Munirul Alam. 2023. 'Genomic attributes of Vibrio cholerae O1 responsible for 2022 massive cholera outbreak in Bangladesh', *Nat. Commun.*, 14: 1154.

Morris, J. J., and R. E Lenski. 2012. 'The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss', *mBio. 3, e00036-12*.

Morris, J.J., R.E. Lenski, and E.R. Zinser. 2012. 'The black queen hypothesis: Evolution of dependencies through adaptive gene loss', *mBio 3*.

Morris, J.J., S.E. Papoulis, and R.E. Lenski. 2014. 'Coexistence of evolving bacteria stabilized by a shared Black Queen function', *Evolution (N Y) 68*.

Nakagawa, S., and H. Schielzeth. 2013. 'A general and simple method for obtaining R 2 from generalized linear mixed-effects models', *Methods in Ecology and Evolution / British Ecological Society, 4(2), 133–142*.

Nayfach, S., B. Rodriguez-Mueller, Garud. N., and K.S. Pollard. 2016. 'An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography', *Genome Res 26:1612–1625*.

Nayfach, S., Z. J. Shi, R. Seshadri, K. S. Pollard, and N. C. Kyrpides. 2019. 'New insights from uncultivated genomes of the global human gut microbiome', *Nature*, 568: 505-10.

Needham, D. M., and J. A. Fuhrman. 2016. 'Pronounced daily succession of phytoplankton , archaea and bacteria following a spring bloom', *Nature Microbiology, 1, 16005*.

Nelson, E. J., J. B. Harris, J. G. Morris, Jr., S. B. Calderwood, and A. Camilli. 2009. 'Cholera transmission: the host, pathogen and bacteriophage dynamic', *Nat Rev Microbiol*, 7: 693-702.

Nelson, Eric J., Ashrafuzzaman Chowdhury, James Flynn, Stefan Schild, Lori Bourassa, Yue Shao, Regina C. LaRocque, Stephen B. Calderwood, Firdausi Qadri, and Andrew Camilli. 2008.

'Transmission of Vibrio cholerae Is Antagonized by Lytic Phage and Entry into the Aquatic Environment', *PLoS Pathog.*, 4: e1000187.

Nelson, Eric J., Danielle S. Nelson, Mohammed A. Salam, and David A. Sack. 2011. 'Antibiotics for both moderate and severe cholera', *N. Engl. J. Med.*, 364: 5-7.

Nikoh, N., T. Hosokawa, K. Oshima, M. Hattori, and T. Fukatsu. 2011. 'Reductive evolution of bacterial genome in insect gut environment', *Genome Biol Evol 3*.

Ofir, G., S. Melamed, H. Sberro, Z. Mukamel, S. Silverman, G. Yaakov, S. Doron, and R. Sorek. 2018. 'DISARM is a widespread bacterial defence system with broad anti-phage activities', *Nat Microbiol*, 3: 90-98.

Olm, M. R., C. T. Brown, B. Brooks, and J. F. Banfield. 2017. 'dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication', *ISME J*, 11: 2864-68.

Olm, Matthew R., Alexander Crits-Christoph, Keith Bouma-Gregson, Brian A. Firek, Michael J. Morowitz, and Jillian F. Banfield. 2021. 'inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains', *Nat. Biotechnol.*, 39: 727-36.

Oprea, M., E. Njamkepo, D. Cristea, A. Zhukova, C. G. Clark, A. N. Kravetz, E. Monakhova, A. S. Ciontea, R. Cojocaru, J. Rauzier, M. Damian, O. Gascuel, M. L. Quilici, and F. X. Weill. 2020. 'The seventh pandemic of cholera in Europe revisited by microbial genomics', *Nature Communications*, 11: 5347.

Padfield, D., A. Vujakovic, S. Paterson, R. Griffiths, A. Buckling, and E. Hesse. 2020. 'Evolution of diversity explains the impact of pre-adaptation of a focal species on the structure of a natural microbial community', *ISME J*, 14: 2877-89.

Pal, Csaba, María D. Maciá, Antonio Oliver, Ira Schachar, and Angus Buckling. 2007. 'Coevolution with viruses drives the evolution of bacterial mutation rates', *Nature*, 450: 1079-81.

Palmer, M. W., and T. A. Maurer. 1997. 'Does Diversity Beget Diversity? A Case Study of Crops and Weeds', *Journal of Vegetation Science, 8(2), 235–240*.

Parks, D. H, M Chuvochina, D. W Waite, C Rinke, A Skarshewski, P.-A Chaumeil, and P Hugenholtz. 2018. 'A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life', *Nature Biotechnology, 36(10), 996–1004*.

Pasricha, C. L., A. J. H. de Monte, and E. G. O'Flynn. 1936. 'Bacteriophage in the Treatment of Cholera', *Ind. Med. Gaz.*, 71: 61-68.

Pasricha, C.L., A.J. MDe Monte, and S.K Gupta. 1931. 'Seasonal variations of cholera bacteriophage in natural waters and in man, in Calcutta during the year 1930.', *The indian medical gazette*.

Penades, J. R., and G. E. Christie. 2015. 'The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites', *Annu Rev Virol*, 2: 181-201.

Pennekamp, F, M Pontarp, A Tabi, F Altermatt, R Alther, Y Choffat, E. A Fronhofer, P Ganesanandamoorthy, A Garnier, J. I Griffiths, S Greene, K Horgan, T. M Massie, E Mächler, G. M Palamara, M Seymour, and O. L Petchey. 2018. 'Biodiversity increases and decreases ecosystem stability', *Nature, 563(7729), 109–112.*

Piarroux, R., R. Barrais, B. Faucher, R. Haus, M. Piarroux, J. Gaudart, R. Magloire, and D. Raoult. 2011. 'Understanding the cholera epidemic, Haiti', *Emerg Infect Dis*, 17: 1161-8.

Poulsen, L.K., T.R. Licht, C. Rang, K.A. Krogfelt, and S. Molin. 1995. 'Physiological state of Escherichia coli BJ4 growing in the large intestines of streptomycin-treated mice', *J Bacteriol 177*.

Power, J. F., C. R. Carere, C. K. Lee, G. L. J. Wakerley, D. W. Evans, M. Button, D. White, M. D. Climo, A. M. Hinze, X. C. Morgan, I. R. McDonald, S. C. Cary, and M. B. Stott. 2018. 'Microbial biogeography of 925 geothermal springs in New Zealand', *Nat Commun*, 9: 2876.

Poyet, M., M. Groussin, S. M. Gibbons, J. Avila-Pacheco, X. Jiang, S. M. Kearney, A. R. Perrotta, B. Berdy, S. Zhao, T. D. Lieberman, P. K. Swanson, M. Smith, S. Roesemann, J. E. Alexander, S. A. Rich, J. Livny, H. Vlamakis, C. Clish, K. Bullock, A. Deik, J. Scott, K. A. Pierce, R. J. Xavier, and E. J. Alm. 2019. 'A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research', *Nat Med*, 25: 1442-52.

Price, T. D., D. M. Hooper, C. D. Buchanan, U. S. Johansson, D. T. Tietze, P. Alstrom, U. Olsson, M. Ghosh-Harihar, F. Ishtiaq, S. K. Gupta, J. Martens, B. Harr, P. Singh, and D. Mohan. 2014. 'Niche filling slows the diversification of Himalayan songbirds', *Nature*, 509: 222-5.

Puigbò, P., A.E. Lobkovsky, D.M. Kristensen, Y.I. Wolf, and E. V. Koonin. 2014. 'Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes', *BMC Med 12*.

Rabosky, D. L., J. Chang, P. O. Title, P. F. Cowman, L. Sallan, M. Friedman, K. Kaschner, C. Garilao, T. J. Near, M. Coll, and M. E. Alfaro. 2018. 'An inverse latitudinal gradient in speciation rate for marine fishes', *Nature*, 559: 392-95.

Rabosky, D. L., and A. H Hurlbert. 2015a. 'Species richness at continental scales is dominated by ecological limits', *The American Naturalist, 185(5), 572–583*.

Reese, A.T., and R.R. Dunn. 2018. 'Drivers of microbiome biodiversity: A review of general rules, feces, and ignorance', *mBio 9*.

Reyes-Robles, T., R. S. Dillard, L. S. Cairns, C. A. Silva-Valenzuela, M. Housman, A. Ali, E. R. Wright, and A. Camilli. 2018. 'Vibrio cholerae Outer Membrane Vesicles Inhibit Bacteriophage Infection', *J Bacteriol*, 200.

Rivard, Nicolas, Rita R. Colwell, and Vincent Burrus. 2020. 'Antibiotic Resistance in Vibrio cholerae: Mechanistic Insights from IncC Plasmid-Mediated Dissemination of a Novel Family of Genomic Islands Inserted at trmE', *mSphere*, 5.

Rousset, F., F. Depardieu, S. Miele, J. Dowding, A. L. Laval, E. Lieberman, D. Garry, E. P. C. Rocha, A. Bernheim, and D. Bikard. 2022. 'Phages and their satellites encode hotspots of antiviral systems', *Cell Host Microbe*, 30: 740-53 e5.

Russell, S.L., and C.M. Cavanaugh. 2017. 'Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes', *Mol Biol Evol 34:2747–2761*.

San Roman, M, and A. Wagner. 2018. 'An enormous potential for niche construction through bacterial cross-feeding in a homogeneous environment', *PLoS Comput Biol 14*.

San Roman, M., and A. Wagner. 2021. 'Diversity begets diversity during community assembly until ecological limits impose a diversity ceiling', *Mol Ecol 30*.

Sandro, Azaele., Suweis. Samir, Grilli. Jacopo, Volkov. Igor, R. Banavar. Jayanth, and Maritan. Amos. 2016. 'Statistical mechanics of ecological systems: Neutral theory and beyond', *Review of modern physics*, 88, 035003.

Sayed, A. M., M. H. A. Hassan, H. A. Alhadrami, H. M. Hassan, M. Goodfellow, and M. E. Rateb. 2020. 'Extreme environments: microbiology leading to specialized metabolites', *J Appl Microbiol*, 128: 630-57.

Schluter, D. 2000. 'The Ecology of Adaptive Radiation', *Oxford Series in Ecology and Evolution*.

Schluter, D., and M. W. Pennell. 2017. 'Speciation gradients and the distribution of biodiversity', *Nature*, 546: 48-55.

Schopf, J. W., K. Kitajima, M. J. Spicuzza, A. B. Kudryavtsev, and J. W. Valley. 2018. 'SIMS analyses of the oldest known assemblage of microfossils document their taxon-correlated carbon isotope compositions', *Proc Natl Acad Sci U S A*, 115: 53-58.

Seed, K. D., K. L. Bodi, A. M. Kropinski, H. W. Ackermann, S. B. Calderwood, F. Qadri, and A. Camilli. 2011a. 'Evidence of a dominant lineage of Vibrio cholerae-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh', *mBio*, 2: e00334-10.

Seed, K. S., M. Yen, B. J. Shapiro, I. J. Hilaire, R. C. Charles, J. E. Teng, L. C. Ivers, J. Boncy, J. B. Harris, and A. Camilli. 2014. 'Evolutionary consequences of intra-patient phage predation on microbial populations', *Elife*, 3: e03497.

Sender, R., S. Fuchs, and R. Milo. 2016. 'Revised Estimates for the Number of Human and Bacteria Cells in the Body', *PLoS Biology, 14(8), e1002533*.

Seth, E. C., and M. E. Taga. 2014. 'Nutrient cross-feeding in the microbial world', *Front Microbiol*, 5: 350.

Shaffer, J. P., L. F. Nothias, L. R. Thompson, J. G. Sanders, R. A. Salido, S. P. Couvillion, A. D. Brejnrod, F. Lejzerowicz, N. Haiminen, S. Huang, H. L. Lutz, Q. Zhu, C. Martino, J. T. Morton, S. Karthikeyan, M. Nothias-Esposito, K. Duhrkop, S. Bocker, H. W. Kim, A. A. Aksenov, W. Bittremieux, J. J. Minich, C. Marotz, M. M. Bryant, K. Sanders, T. Schwartz, G. Humphrey, Y. Vasquez-Baeza, A. Tripathi, L. Parida, A. P. Carrieri, K. L. Beck, P. Das, A. Gonzalez, D. McDonald, J. Ladau, S. M. Karst, M. Albertsen, G. Ackermann, J. DeReus, T. Thomas, D. Petras, A. Shade, J. Stegen, S. J. Song, T. O. Metz, A. D. Swafford, P. C. Dorrestein, J. K. Jansson, J. A. Gilbert, R. Knight, and Consortium Earth Microbiome Project. 2022. 'Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity', *Nat Microbiol*, 7: 2128-50.

Shah M. Faruque, M. Mostafizur Rahman, Asadulghani, K.M. Nasirul Islam, John J. Mekalanos. 1999. 'Lysogenic Conversion of Environmental Vibrio mimicus Strains by CTXphi', *Infection and immunity, Nov. 1999, p. 5723-5729.*

Sharma, V., D. A. Rodionov, S. A. Leyn, D. Tran, S. N. Iablokov, H. Ding, D. A. Peterson, A. L. Osterman, and S. N. Peterson. 2019. 'B-Vitamin Sharing Promotes Stability of Gut Microbial Communities', *Front Microbiol*, 10: 1485.

Shehata, T. E., and A. G. Marr. 1971. 'Effect of nutrient concentration on the growth of Escherichia coli', *J Bacteriol*, 107: 210-6.

Shetty, S. A., B. Kuipers, S. Atashgahi, S. Aalvink, H. Smidt, and W. M. de Vos. 2022. 'Inter-species Metabolic Interactions in an In-vitro Minimal Human Gut Microbiome of Core Bacteria', *NPJ Biofilms Microbiomes*, 8: 21.

Shetty, S. A., H. Smidt, and W. M. de Vos. 2019. 'Reconstructing functional networks in the human intestinal tract using synthetic microbiomes', *Curr Opin Biotechnol*, 58: 146-54.

Shonkwiler, J.S. 2016. 'Variance of the truncated negative binomial distribution', *J Econom 195*.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. 'Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy', *Nat Microbiol*, 3: 836-43.

Simonsen, A.K. 2022. 'Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria', *ISME Journal 16*.

Smillie, C. S., J. Sauk, D. Gevers, J. Friedman, J. Sung, I. Youngster, E. L. Hohmann, C. Staley, A. Khoruts, M. J. Sadowsky, J. R. Allegretti, M. B. Smith, R. J. Xavier, and E. J. Alm. 2018. 'Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation', *Cell Host Microbe*, 23: 229-40 e5.

Smith, W. P. J., B. R. Wucher, C. D. Nadell, and K. R. Foster. 2023. 'Bacterial defences: mechanisms, evolution and antimicrobial resistance', *Nat Rev Microbiol*.

Smith., Nick W., Paul R. Shorten, Eric Altermann, Nicole C. Roy, and Warren C. McNabb. 2019. 'The Classification and Evolution of Bacterial Cross-Feeding', *Front. Ecol. Evol., 14 May 2019*.

Sobolev, D., and M. F. Begonia. 2008. 'Effects of heavy metal contamination upon soil microbes: lead-induced changes in general and denitrifying microbial communities as evidenced by molecular markers', *Int J Environ Res Public Health*, 5: 450-6.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. 2006. 'Microbial diversity in the deep sea and the underexplored "rare biosphere." ', *Proceedings of the National Academy of Sciences of the United States of America, 103(32), 12115–12120.*

Sommer, F., and F. Backhed. 2013. 'The gut microbiota--masters of host development and physiology', *Nat Rev Microbiol*, 11: 227-38.

Sriswasdi, S., C.-C. Yang, and W. Iwasaki. 2017. 'Generalist species drive microbial dispersion and evolution', *Nature Communications, 8(1), 1162*.

Summers, W. C. 1993. 'Cholera and plague in India: the bacteriophage inquiry of 1927-1936', *J. Hist. Med. Allied Sci.*, 48: 275-301.

Sunagawa, S., L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, coordinators Tara Oceans, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, and P. Bork. 2015. 'Ocean plankton. Structure and function of the global ocean microbiome', *Science*, 348: 1261359.

Sung, W., M.S. Ackerman, S.F. Miller, T.G. Doak, and M. Lynch. 2012. 'Drift-barrier hypothesis and mutation-rate evolution', *Proc Natl Acad Sci U S A 109*.

Svoboda, Elizabeth. 2020. 'Autism and the gut', *MNature*, 577.

Tamar, E. S., and R. Kishony. 2022. 'Multistep diversification in spatiotemporal bacterial-phage coevolution', *Nature Communications*, 13.

Thompson, L. R., J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, J. A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vazquez-Baeza, A. Gonzalez, J. T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M. F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N. A. Bokulich, J. Lefler, C. J. Brislawn, G. Humphrey, S. M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J. A. Fuhrman, A. Clauset, R. L. Stevens, A. Shade, K. S. Pollard, K. D. Goodwin, J. K. Jansson, J. A. Gilbert, R. Knight, and Consortium Earth Microbiome Project. 2017. 'A communal catalogue reveals Earth's multiscale microbial diversity', *Nature*, 551: 457-63.

Tian, L., X. W. Wang, A. K. Wu, Y. Fan, J. Friedman, A. Dahlin, M. K. Waldor, G. M. Weinstock, S. T. Weiss, and Y. Y. Liu. 2020. 'Deciphering functional redundancy in the human microbiome', *Nat Commun*, 11: 6217.

Travisano., Michael, Judith A. Mongold., Albert F. Bennett., and Richard E. Lenski. 1995. 'Experimental Tests of the Roles of Adaptation, Chance, and History in Evolution', *Science*, 267.

Truong, D.T., A. Tett, E. Pasolli, C. Huttenhower, and N. Segata. 2017. 'Microbial strain-level population structure & genetic diversity from metagenomes', *Genome Res 27:626–638*.

Ulanova, D., and K. S. Goo. 2015. 'Diversity of actinomycetes isolated from subseafloor sediments after prolonged low-temperature storage', *Folia Microbiol (Praha)*, 60: 211-6.

Valverde, A., M. Tuffin, and D. A. Cowan. 2012. 'Biogeography of bacterial communities in hot springs: a focus on the actinobacteria', *Extremophiles*, 16: 669-79.

van Dijk, Lucas R., Bruce J. Walker, Timothy J. Straub, Colin J. Worby, Alexandra Grote, Henry L. th Schreiber, Christine Anyansi, Amy J. Pickering, Scott J. Hultgren, Abigail L. Manson, Thomas Abeel, and Ashlee M. Earl. 2022. 'StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities', *Genome Biol.*, 23: 74.

van Houte, S., A. Buckling, and E. R. Westra. 2016. 'Evolutionary Ecology of Prokaryotic Immune Mechanisms', *Microbiol Mol Biol Rev*, 80: 745-63.

Van Rossum, T., P. Ferretti, O. M. Maistrenko, and P. Bork. 2020. 'Diversity within species: interpreting strains in microbiomes', *Nat Rev Microbiol*, 18: 491-506.

Van Tienderen, P. H. 1991. 'Evolution of Generalists and Specialists in Spatially Heterogeneous Environments', *Evolution*, 45: 1317-31.

Vassallo, C. N., C. R. Doering, M. L. Littlehale, G. I. C. Teodoro, and M. T. Laub. 2022. 'A functional selection reveals previously undetected anti-phage defence systems in the E. coli pangenome', *Nat Microbiol*, 7: 1568-79.

Vatanen, T., D. R. Plichta, J. Somani, P. C. Munch, T. D. Arthur, A. B. Hall, S. Rudolf, E. J. Oakeley, X. Ke, R. A. Young, H. J. Haiser, R. Kolde, M. Yassour, K. Luopajarvi, H. Siljander, S. M. Virtanen, J. Ilonen, R. Uibo, V. Tillmann, S. Mokurov, N. Dorshakova, J. A. Porter, A. C. McHardy, H. Lahdesmaki, H. Vlamakis, C. Huttenhower, M. Knip, and R. J. Xavier. 2019. 'Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life', *Nat Microbiol*, 4: 470-79.

Venturelli, O. S., A. C. Carr, G. Fisher, R. H. Hsu, R. Lau, B. P. Bowen, S. Hromada, T. Northen, and A. P. Arkin. 2018. 'Deciphering microbial interactions in synthetic human gut microbiome communities', *Mol Syst Biol*, 14: e8157.

Verma, J., S. Bag, B. Saha, P. Kumar, T. S. Ghosh, M. Dayal, T. Senapati, S. Mehra, P. Dey, A. Desigamani, D. Kumar, P. Rana, B. Kumar, T. K. Maiti, N. C. Sharma, R. K. Bhadra, A. Mutreja, G. B. Nair, T. Ramamurthy, and B. Das. 2019. 'Genomic plasticity associated with antimicrobial resistance in Vibrio cholerae', *Proc Natl Acad Sci U S A*, 116: 6226-31.

Verster, A. J., B. D. Ross, M. C. Radey, Y. Bao, A. L. Goodman, J. D. Mougous, and E. Borenstein. 2017. 'The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition', *Cell Host Microbe*, 22: 411-19 e4.

Vos, M. 2011. 'A species concept for bacteria based on adaptive divergence', *Trends in Microbiology, 19(1), 1–7*.

Waldor, Matthew K., Helmut Tschäpe, and John J. Mekalanos. 1996. 'A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in Vibrio cholerae O139', *J. Bacteriol.*, 178: 4157-65.

Walters, K.E. , and J.B.H. Martiny. 2020. 'Alpha-, beta-, and gamma-diversity of bacteria varies across habitats', *PLoS One 15*.

Wang, L., S. Jiang, Z. Deng, P. C. Dedon, and S. Chen. 2019. 'DNA phosphorothioate modification-a new multi-functional epigenetic system in bacteria', *FEMS Microbiol Rev*, 43: 109-22.

Wang, Z., E. Klipfell, B. J. Bennett, R. Koeth, B. S. Levison, B. Dugar, A. E. Feldstein, E. B. Britt, X. Fu, Y. M. Chung, Y. Wu, P. Schauer, J. D. Smith, H. Allayee, W. H. Tang, J. A. DiDonato, A. J. Lusis, and S. L. Hazen. 2011. 'Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease', *Nature*, 472: 57-63.

Weill, François-Xavier, Daryl Domman, Elisabeth Njamkepo, Cheryl Tarr, Jean Rauzier, Nizar Fawal, Karen H. Keddy, Henrik Salje, Sandra Moore, Asish K. Mukhopadhyay, Raymond Bercion, Francisco J. Luquero, Antoinette Ngandjio, Mireille Dosso, Elena Monakhova, Benoit Garin, Christiane Bouchier, Carlo Pazzani, Ankur Mutreja, Roland Grunow, Fati

Sidikou, Laurence Bonte, Sébastien Breurec, Maria Damian, Berthe-Marie Njanpop-Lafourcade, Guillaume Sapriel, Anne-Laure Page, Monzer Hamze, Myriam Henkens, Goutam Chowdhury, Martin Mengel, Jean-Louis Koeck, Jean-Michel Fournier, Gordon Dougan, Patrick A. D. Grimont, Julian Parkhill, Kathryn E. Holt, Renaud Piarroux, Thandavarayan Ramamurthy, Marie-Laure Quilici, and Nicholas R. Thomson. 2017. 'Genomic history of the seventh pandemic of cholera in Africa', *Science*, 358: 785-89.

Wexler, A.G., and A.L. Goodman. 2017. 'An insider's perspective: Bacteroides as a window into the microbiome', *Nat Microbiol*.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. 'Prokaryotes: the unseen majority', *Proceedings of the National Academy of Sciences of the United States of America, 95(12), 6578–6583*.

Whittaker, R.H. 1972b. 'Evolution and measurement of species diversity', *Taxon*, Vol. 21, No. 2/3 (May, 1972), pp. 213-251.

WHO. 2022. 'World Health Organization. ', https://www.who.int/news-room/fact-sheets/detail/cholera.

Wolff, R., W.R. Shoemaker, and N.R. Garud. 2021. 'Ecological Stability Emerges at the Level of Strains in the Human Gut Microbiome', *BioRxiv*.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. 'Improved metagenomic analysis with Kraken 2', *Genome Biol.*, 20: 257.

Wozniak, R. A., D. E. Fouts, M. Spagnoletti, M. M. Colombo, D. Ceccarelli, G. Garriss, C. Dery, V. Burrus, and M. K. Waldor. 2009. 'Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs', *PLoS Genet*, 5: e1000786.

Wu, D., G. Jospin, and J.A. Eisen. 2013. 'Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups', *PLoS One 8*.

Xiang Ng, Q., Mlzq De Deyn, W. Loke, and W. S. Yeo. 2020. 'Yemen's Cholera Epidemic Is a One Health Issue', *J Prev Med Public Health*, 53: 289-92.

Yaffe, E., and D.A. Relman. 2020. 'Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation', *Nat Microbiol 5*.

Yang, Y., M. Nguyen, V. Khetrapal, N.D. Sonnert, A.L. Martin, H. Chen, M.A. Kriegel, and N.W. Palm. 2022. 'Within-host evolution of a gut pathobiont facilitates liver translocation ', *Nature*.

Yassour, M., E. Jason, L. J. Hogstrom, T. D. Arthur, S. Tripathi, H. Siljander, J. Selvenius, S. Oikarinen, H. Hyoty, S. M. Virtanen, J. Ilonen, P. Ferretti, E. Pasolli, A. Tett, F. Asnicar, N.

Segata, H. Vlamakis, E. S. Lander, C. Huttenhower, M. Knip, and R. J. Xavier. 2018. 'Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life', *Cell Host Microbe*, 24: 146-54 e4.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon. 2012. 'Human gut microbiome viewed across age and geography', *Nature*, 486: 222-7.

Yen, M., L. S. Cairns, and A. Camilli. 2017. 'A cocktail of three virulent bacteriophages prevents Vibrio cholerae infection in animal models', *Nat Commun*, 8: 14187.

Zahid, M. S. H., S. M. N. Udden, A. S. G. Faruque, S. B. Calderwood, J. J. Mekalanos, and S. M. Faruque. 2008. 'Effect of Phage on the Infectivity of Vibrio cholerae and Emergence of Genetic Variants', *Infect. Immun.*, 76: 5266-73.

Zhao, S., T. D. Lieberman, M. Poyet, K. M. Kauffman, S. M. Gibbons, M. Groussin, R. J. Xavier, and E. J. Alm. 2019. 'Adaptive Evolution within Gut Microbiomes of Healthy People', *Cell Host Microbe*, 25: 656-67 e8.

Zheng, W., S. Zhao, Y. Yin, H. Zhang, D.M. Needham, E.D. Evans, C.L. Dai, P.J. Lu, E.J. Alm, and D.A. Weitz. 2022. ' High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome', *Science (1979) 376*.