Spring 2022

On the Evolutionary Ecology of Microbial Metabolic Niche Construction

Jean Celestin Charles Vila
Yale University Graduate School of Arts and Sciences, jeanccvila@gmail.com

Abstract

On the Evolutionary Ecology of Microbial Metabolic Niche Construction

Jean Celestin Charles Vila

2022

All organisms construct the shared environment in which they live. This is especially notable in micro-organisms as they secrete and uptake a diverse range of metabolites depending on their genotype and environment. Over the past few decades, systems biologists have developed computational tools to predict the nutrient uptake and secretions of microbes across environments, using metabolic networks inferred from whole-genome sequencing. These tools provide an opportunity to quantify eco-evolutionary dynamics at the genomic level, by combining genome-scale metabolic models mapping genotype to phenotype, with consumer-resource models predicting population dynamics from phenotype. By leveraging these new computational approaches and combining them with experiments using microbial communities in synthetic environments, this dissertation will quantify the impact of metabolite production and consumption on the evolutionary and ecological dynamics of multi-species microbial communities.

In Chapter 1, I present a published paper in which I address how niche construction quantitatively determines evolutionary trajectories by deforming the fitness landscape of evolving populations. The chapter uses a combination of genome-scale metabolic modelling and experiments to systematically quantify the deformability of the E.coli metabolic fitness landscape. It shows that the effects of niche construction are quantitatively modest at short genomic scales but accumulate over longer evolutionary trajectories. These results suggest that fitness landscapes can predict evolution over short mutational distances, but that niche construction hampers predictability in the long term.

In Chapter 2 I present a published paper in which I ask whether communities assembling in the same metabolic environment show similar ecological interactions. This chapters leverages previously published 16s rRNA sequencing data from an experiment in which complex-microbial communities were allowed to self assemble in laboratory environments containing a single limiting resource. I benchmark a newly developed statistical tool, Dissimilarity-Overlap Analysis, and use it to determine whether interaction parameters are similar across communities assembled in the same metabolic environment. I find a

negative relationship between dissimilarity and overlap which is what we expect if interactions are strongly convergent. However, even in replicate, identical habitats, two different communities may contain the same set of taxa at different abundances in equilibrium. The formation of alternative states in community assembly is strongly associated with the presence of specific taxa suggesting that some taxa may differ in the niches they construct and occupy even across replicate abiotic conditions.

In Chapter 3 I present a published paper which asks how different components of the environment interact to collectively determine the taxonomic composition of microbial communities. This paper tests whether the composition of communities assembled in a pair of carbon sources could be predicted from those assembled in each single carbon source alone. This paper develops a null-additive model and show that it can explain a high variation of the relative abundance of families in communities assembled in pairs of carbon sources. Deviation from this additive model reveal a characteristic pattern with sugars 'dominating' organic acids. Using consumer-resource modelling, I show that nutrient dominance can be explained by experimentally validated asymmetries in the family level specialisation on different resource types. Quantifying the asymmetric effect of metabolites on community composition is a key step towards engineering microbial communities by modulating nutrient composition.

In Chapter 4, I present a draft manuscript in which I ask whether one can predict the composition of microbial communities assembling in different metabolic environments. I first use a combination of enrichment experiments, metabolomics and phenotypic assays to show that the predictability of community assembly depends on the phylogenetic distribution of quantitative metabolic traits selected for by different environments. This includes traits determining both the ability to exploit the supplied resource and the ability to grow on the constructed niches. I find that similarities in community composition across environments reflect correlations in conserved metabolic traits, which are predictable using metabolic models.Finally I show how one can use metabolic models to quantitatively predict the effect of novel environmental perturbations on microbial communities.

The work presented herein illustrates how genome-scale models can be combined with analytical models of population dynamics to develop quantitative and predictive eco-evolutionary theory. Whilst focusing on microbial communities, the concepts developed are applicable to other cellular populations as well as to macro-organism engaging in niche-constructing activities.By quantifying the effects of niche construction in an explicit manner, the work I have presented moves beyond semantic arguments and descriptive studies towards a predictive and mechanistic understanding of eco-evolutionary dynamics.

On the Evolutionary Ecology of Microbial Metabolic Niche Construction

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Jean Celestin Charles Vila

Dissertation Director: Alvaro Sanchez

May 2022

Acknowledgments

During my PhD I have been supported by a large community of colleagues, friends and family, without whom this research would not have been possible. Their support has often felt far in excess of what was warranted, and has provided me with the space to grow as a scholar, scientist and person. I hope including them in these acknowledgements can act as a small token of gratitude.

I would first and foremost like to thank my advisor Dr Alvaro Sanchez for opening the doors to join his laboratory as his first PhD student back in 2016. I feel uniquely fortunate to have arrived at Yale University just as his lab was getting established and I benefited enormously from the creative energy, excitement and intellectual freedom that came with joining his new research program. Alvaro gave me the freedom to explore different ideas, hopping between multiple projects and collaborating widely, whilst demonstrating a unique ability to translate the chaotic jumble of disorganized thoughts in my brain into coherent results. I am immensely grateful for the opportunities he has given me, the patience he has always shown, and the belief in my abilities even when I doubted myself. Most importantly, Alvaro has consistently shown me that being a rigorous, innovative and productive scientist is entirely compatible with being a kind, empathetic and generous person. Amongst the many lessons I have learnt from him, the one that stands out, is that the people producing the science are far more important than the science they produce.

I would also like to thank my committee members Dr Paul Turner, Dr Gunter Wagner, Dr David Post and Dr Maria Rebolleda-Gomez. Paul welcomed me into his lab-space when I arrived at Yale and I always looked forward to my meetings with him, not least for his enthusiastic encouragement and sharp critique of my ideas. I benefited immensely from many conversations with Gunter and David, which were insightful, though-provoking and which helped me place my research within a broader historical context. I was delighted to be able to ask Maria to join my committee as she is one of the most generous, open-minded and original thinkers I have ever met. During the time I spent in lab with her I have seen how Maria combined excellent research with unashamed authenticity, acting as role-model and helping me reconcile my own identities as a queer man and a scientist.

I am especially grateful to several post-docs who I have worked closely with at Yale and who each invested significant time and energy mentoring me. The most fun I have had on a research project was with Dr Djordje Bajic who combines unrelenting optimism and positivity with wonderfully anarchic and creative ways of thinking. Dr Alicia Sanchez-Gorostiaga trained me in experimental work and thanks to her endless patience and de-

tailed feedback achieved the impossible and taught me how to be an experimentalist. I am yet to meet anyone as thoughtful as Dr Sylvie Estrela who is far too modest for someone so talented, and repeatedly brought the most brilliant insights to our many joint projects. Dr Lisa Bono welcomed me the moment I arrived at Yale and has always given honest advice, friendship and support, especially during challenging times.

The Sanchez Lab at Yale has been a uniquely interactive, creative and supportive environment and I feel immensely privileged to have worked with its past and present members. Dr Nanxi Lu who set the highest standards of experimental precision whilst always supporting those of us who fell short. Dr Nora Pyenson who brings her wonderfully colourful sense of style, fun and creativity and always encouraged me to bring my authentic self to work. Dr Juan Diaz Colunga for our mutual joy in niche nerdy You-Tubers and for always going out of his way to help others. Dr Xin Sun whose positive attitude is infectious and always brings her enthusiasm as well as the occasional sweet treats. Abby Skwara who despite just joining the lab is likely to-surpass us all in a couple of years. Maddie Bender and Jackie Folmar, our undergraduates who dared to venture out to west campus and made me feel old in the process. Finally and not least my fellow graduate student, collaborator, friend and regular sounding board, Chang-Yu Chang who always managed to keep me motivated and excited about science.

I would also like to thank the graduate student and post-doc community of the Ecology and Evolutionary Biology Department at Yale. Thank you to Mike Blazanin, Nikunj Goel, Liam Taylor, Daniel Stadtmaur, Diego Ellis Soto, Lauren Melenthin, Petra Walker ,Brooke Bodensteiner , Dr Natasha Picciani, Dr Antonio Rodrigues, Muyang Lu, Anri Chomentowska, Dr Franz Simon, Dr Siyang Xia, Dr Arun Chavan, Mansa Srivastav, and Dr Jose Moreno Vilano. You have all made me feel at home here and I have fond memories of our times spent together, whether it be on a whiteboard in the seminar room, out in Connecticut bear country or on the dance floor during drag nights at Partners. A special thanks go to Jasmine Mah, Daemin Kim and Henry Camarillo, your company and friendship helped keep me sane during the long covid winter. Finally I want to send my deepest thanks to my former housemate and almost certainly a not-too distantly relative, Dr Alejandro Damien Serrano and his wife Rebecca Gordon. I think we drove each other a bit mad at times, but I could not have wished for better companions to brave the lockdown.

During my time at Yale, I benefited from the support of the distinguished faculty and staff of EEB. This includes the entire EEB business office and especially Deanna Brunson for being able to solve every problem as well as our custodial and support staff. Amongst the faculty I would like to acknowledge Dr Thomas Near, Dr Erika Edwards and Dr Casey

Dunn for regularly checking in on us, especially during the pandemic. I also want to acknowledge Dr Stephen Stearns, for sharing his wisdom, council and experience and for bringing me into Saybrook College as a Graduate Affiliate. Thank you for all that you have done for our field.

The Microbial Sciences institute has served as my second academic home at Yale over the past six years and I am indebted to the enormous number of people on whom the MSI and west campus depends, not least the shuttle drivers, cleaners, caterers and administrative staff, in particular Bianca Abreau and Lillian Smith, our senior administrative assistants at the MSI. I am also grateful for all the researchers at the MSI who have sought to build a positive community atmosphere across disparate fields. Dr Caroline , Catherine Shipps and Danny Dumitrescu, I shall miss our conversations at the Nespresso machine.

During my time at Yale I have been fortunate to collaborate with a large number of scientists from across the world. This includes visiting scholars to our lab: Dr Sotaro Takano and Dr Felipe Lino; our many friends from Boston University: Dr Robert Marsland III, Dr Daniel Segre, Dr Ilija Dukovski, Dr Mike Quintin; from Harvard University: Dr Yang-Yu Liu; from Stanford University: Dr KC Huang, Dr Andres Arande Diaz and Dr Lisa Willis, and from Michigan State University Dr Zachary Blount. Finally I want to thank Dr Pankaj Mehta and Dr Josh Goldford for welcoming me multiple times to Boston and for challenging me to think more deeply.

My academic career did not begin at Yale and completing my PhD was only possible thanks to the support of the many mentors I have had along the way. My undergraduate tutors and mentors at Oxford University who first fostered my fascination with evolutionary biology: Dr Ian Moore, Dr Nathalie Seddon, Dr Cedric Tan and Dr Jay Biernaskie. The researchers at the IGC in Portugal who introduced me to microbial experimental evolution: Dr Isabel Gordo, Dr Karina Xavier and Dr Ozan Ozkaya. The community of ecologists and evolutionary biologists at Silwood Park who helped me come into my own as a computational scientist, Dr Matthew Lloyd Jones, Dr Thomas Bell, Dr Thomas Scheurl, Dr Alberto Pascual-Garcia,Dr Matishalin Patel and Dr Samraat Pawaar. And not least my wonderful masters advisor Dr James Rosindell who fed my creativity and encouraged me to think outside the box whilst holding me accountable, gently pushing me to complete my first paper.

Outside of work I am indebted to the people of the city of New Haven ,the community of East-Rock and the friends who have shared this time with me. Garrett Nash, Mairead MacRae, Benjamin Morency and Kaltuma Mohamed,two years was far too short a time to spend with you all. Jon Beltz, Sophie Westacott and Max Shinn, thanks for completing

this marathon with me and helping me over the finish line. Finally to my wonderfully tolerant, quirky and genuine housemate Rahul Amin, it has been so fun much sharing my home with you for the past three years.

I am grateful to have been able to call on many friends from the UK when I felt nostalgic or in need of support. To the members of the NAME REDACTED Fan Club: Eilidh Macfarlane, Andy Mckay, Conor Mckenzie, Ab Brightman, James King and Brit Klugg, as well as associate members Nomi Farhi, Ali White and Robin McGhee. To my friends from Wadham College: Stephanie Faulkner, Katie Mcmahon, Maddy Butler, Kristina Kristova and Richard Steward who remind me to keep everything in moderation. To my college roommate Rowan Howell and his wife Esi Armah-Tetteh, I am so excited for the two of you, thank you for always being there. To my friends from the City of London School: Max Twivy, Jonny Lister, Tom Davidson, and Ted Elgar, who remind me how far we have all come. To Sebastian Neira Farriol and Ricardo Shankland for proving that new friendships can emerge when we are all isolated at home.

Finally, I would like to acknowledge my family who I love and find infuriating to equal measure and who have always cheered me on as I have sought to construct and occupy my own personal niche. I am especially grateful for the love of my sister Estelle Vila, my brother Dr Pierre Vila, my soon to be sister in law Dr Felicity De Vere, my step-mum Patricia Vila and my step-dad Dr Barry Wetherilt. I have also been supported by my cousins, aunts and uncles, Yohann Payrot, Cynthia Callegari, Jo-Marie Callegari, Katrien Antonis, Dr Sophie Antonis, Jose Payrot, Dominique Fremault, Patrick Antonis, Jon Whitemore and Claudine Fremault.

This dissertation is for my mum, Dr Anne Wetherilt, you have proof-read enough of my writing to be warranting a co-authorship, I will try to give you a rest during the post-doc, or at the very least, I owe you reciprocal services when you wrap up your current PhD thesis. This dissertation is for my dad Dr Jean-Luc Vila, you are probably the only one in my family who actually understands the maths in my papers, thank you for always believing in my abilities. This dissertation is for my grandparents Jacquie and Jean Vila who dedicated their lives to the education of others and encouraged us to value knowledge for knowledge's sake. The french poet Anatole France is quoted as saying "Les neuf dixiemes de l'education sont des encouragements"[nine tenth of education is encouragement]. Thank you to my family for always encouraging me.

Per citar Lluís Llach ”El dia en què quan se li pregunti a un infant ’Què vols ser de gran?’ respongui ’Vull ser una bona persona’, segurament la societat que somniem serà molt a prop”. Per ajudar-me a ser una bona persona, aquesta tesi està dedicada al meu avi Jean Pierre Etienne Vila i a la meva àvia Jacqueline Josette Raymonde Vila.

To qoute Lluís Llach ”The day when a child is asked ’What do you want to be when you grow up?’ answer ’I want to be a good person’, surely the society we dream of will be very close”. For helping me become good person, this dissertation is dedicated to my grandfather Jean Pierre Etienne Vila and my grandmother Jacqueline Josette Raymonde Vila. .

Contents

List of Figures

List of Tables

List of Supplementary Figures

List of Supplementary Tables

Chapter 1

On the deformability of an empirical fitness landscape by microbial evolution

Djordje Bajic[1]*., Jean C.C. Vila*[1]*., Zachary D. Blount and Alvaro Sanchez*

Bajic, D., Vila, J. C. C., Blount, Z. D., and Sanchez, A. (2018). On the deformability of an empirical fitness landscape by microbial evolution. Proceedings of the National Academy of Sciences, 115 (44), 11286-11291

[1]Djordje Bajic and Jean C.C. Vila contributed equally to this work

1.1 Abstract

A fitness landscape is a map between the genotype and its reproductive success in a given environment. The topography of fitness landscapes largely governs adaptive dynamics, constraining evolutionary trajectories and the predictability of evolution. Theory suggests that this topography can be deformed by mutations that produce substantial changes to the environment. Despite its importance, the deformability of fitness landscapes has not been systematically studied beyond abstract models, and little is known about its reach and consequences in empirical systems. Here we have systematically characterized the deformability of the genome-wide metabolic fitness landscape of the bacterium Escherichia coli. Deformability is quantified by the noncommutativity of epistatic interactions, which we experimentally demonstrate in mutant strains on the path to an evolutionary innovation. Our analysis shows that the deformation of fitness landscapes by metabolic mutations rarely affects evolutionary trajectories in the short range. However, mutations with large environmental effects produce long-range landscape deformations in distant regions of the genotype space that affect the fitness of later descendants. Our results therefore suggest that, even in situations in which mutations have strong environmental effects, fitness landscapes may retain their power to forecast evolution over small mutational distances despite the potential attenuation of that power over longer evolutionary trajectories. Our methods and results provide an avenue for integrating adaptive and eco-evolutionary dynamics with complex genetics and genomics.

1.2 Introduction

When a new genotype appears in a population its reproductive success is largely governed by the environment. Although the environment is often thought of as an external driver of natural selection, it can also be shaped by the evolving population itself, for instance through its metabolic activity or through interactions with the abiotic habitat or other species [Lewontin, 1983, John Odling-Smee et al., 2013, Laland et al., 2014]. These population-driven environmental changes can in turn modify the fitness effects of future mutations, closing in an eco-evolutionary feedback loop [Post and Palkovacs, 2009] . Eco-evolutionary feedbacks are well documented in natural [Hendry, 2016] and experimental [Jones et al., 2009] populations, and at all scales of biological organization: from the cellular scale [e.g., in the evolution of cancer [Basanta and Anderson, 2017]

and microbial populations [Sanchez and Gore, 2013]] to the organismal scale in animal [Matthews et al., 2016] and plant evolution [terHorst and Zee, 2016]. Given the growing evidence that evolutionary and ecological processes, including niche construction, occur on similar timescales, there is a critical need to understand the genomic bases of these eco-evolutionary feedbacks [Rudman et al., 2018].

The "map" between each genotype and its adaptive value in a given environment is known as the "fitness landscape" [Wright, 1932]. Because populations actively modify their environment, new mutations can, in principle, have environmental as well as fitness effects. Thus, evolving populations may dynamically reshape ("deform") the fitness landscapes on which they are adapting [Kauffman and Johnsen, 1991, Watson and Ebner, 2014]. Although they are often used only metaphorically to depict or visualize adaptation, fitness landscapes are a major determinant of evolution. In particular, the topography of a fitness landscape (i.e., the location of fitness peaks and valleys and their connectivity) plays a pivotal role, as it governs the accessibility of evolutionary trajectories [Weinreich et al., 2006, Poelwijk et al., 2007, Hartl, 2014], the role of population structure on evolution [Nahum et al., 2015], the degree of evolutionary convergence among populations [Van Cleve and Weissman, 2015], the expected role of drift, selection, and sex in the evolutionary process [Rozen et al., 2008, Moradigaravand and Engelstädter, 2012], the discovery of evolutionary innovations [Barve and Wagner, 2013], and the predictability of evolution [de Visser and Krug, 2014], a subject of growing importance for the management of pathogens and cancer treatment [Barber et al., 2015, Zhao et al., 2016, Luksza and Lässig, 2014, Nourmohammad et al., 2013, Lässig et al., 2017]. Given the fundamental role that fitness landscapes play in adaptation, if populations do indeed change the topography of their fitness landscapes as they evolve, it is imperative to understand precisely how. Do mutations that alter the environment generally also alter the fitness of all subsequent mutations or only a subset of them? If the latter, where are those deformations localized in the genotype space, and how strong are they? All these questions remain open, as the deformability (or "rubberness") of fitness landscapes has never been systematically studied in empirical systems at the genomic scale.

Substantial experimental evidence suggests that microbial fitness landscapes are likely to exhibit deformability [Paquin and Adams, 1983, Good et al., 2017, Le Gac and Doebeli, 2010, Rosenzweig et al., 1994, Friesen et al., 2004], making microbes an ideal system for addressing this issue. Microbial metabolism leads to large-scale environmental construction through the uptake and release of metabolites [Good et al., 2017, Rosenzweig et al., 1994]. Which nutrients are taken up, which byproducts are released, and in what amounts,

are all governed by the structure of the metabolic network and therefore by the genotype [Quandt et al., 2015, Paczia et al., 2012] As a result, new mutations that change the metabolic network can also change the patterns of metabolic uptake and secretion, altering the environment and potentially also altering the fitness of future mutations [Rosenzweig et al., 1994].

Microbial physiology and growth can be explicitly simulated using genome-scale metabolic models [Orth et al., 2011, Lewis et al., 2010, O'Brien et al., 2015] Due to their excellent predictive capabilities [Orth et al., 2011] and utility for easily and rapidly screening millions of genotypes, these models have been successfully used to systematically explore the genotype space [Matias Rodrigues and Wagner, 2009]. Recent advances in dynamic metabolic modeling make it possible to explicitly simulate the growth of microbial communities and their environmental feedbacks with evolution [Mahadevan et al., 2002, Harcombe et al., 2014], making genome-wide dynamic metabolic modeling of microbial genotypes a promising method to examine the deformability of fitness landscapes (Figure. 1.1A).

Here, we first use metabolic modeling to show that the environmental effect of new mutations can make genetic interactions (or "epistasis") noncommutative or dependent on the order in which mutations occur. We then use evolved strains from one of the populations in Lenski and coworkers' [Lenski et al., 1991, Blount et al., 2008] Escherichia coli Long-Term Evolution Experiment (LTEE) to experimentally demonstrate the presence of noncommutative epistasis and quantitatively validate the predictive capabilities of our model. We then scale up our study to include tens of millions of genotypes from the metabolic genotype space. By systematically screening the in silico metabolic fitness landscape of E. coli, we are able to offer a precise view of how deformability by eco-evolutionary feedbacks plays out over short and long mutational distances.

1.3 Results

1.3.1 Non-commutative Epistasis Characterizes Fitness Landscape Deformability

To investigate the effect of metabolic secretions on the fitness landscape, we used dynamic flux balance analysis (dFBA) to determine the distribution of fitness and environmental effects of new mutations in the local mutational neighborhood of a recently curated,

genome-scale metabolic model of E. coli [Orth et al., 2011]. Our screen included all possible single-addition and deletion mutants (Materials and Methods), whose growth was simulated on anaerobic glucose medium until saturation was reached. Of all nonessential mutations, 147 (3.3%) affected growth rate either positively or negatively (Figure. 1.1B). All these mutations also altered the chemical composition of the environment (Materials and Methods; also see Figure. 1.1C for a representative subset and Supplementary Figure. 1.1 for the full set), and the magnitude of the environmental and fitness effects were strongly correlated (Pearson's $\rho = 0.61, P < 10^{-6}$) (Supplementary Figure. 1.2).This suggests that the extracellular environment will change as new mutations fix in the population, which could in turn alter the fitness effects of new mutations, thus deforming the fitness landscape

We explored the extent to which this fitness landscape may be deformed by the effect of metabolic secretions using a dataset that consisted of $\sim 10^7$ single and double mutants, representing the entire second-order metabolic mutational neighborhood of E. coli. The fitness of each mutant (M) was determined in competition with its immediate ancestor (A) as $F_M^{(A)} = log([X'_M/X_M]/[X'_A/X_A])$ [Lenski et al., 1991, Travisano and Lenski, 1996],where X_A and X_M represent the initial densities of ancestor and mutant and X'_A and X'_M represent their final respective densities after 10 h of competition (Materials and Methods). All competitions were performed at an initial mutant frequency of 0.01. Using this measure, the fitness effects of two mutations are expected to combine additively when they act independently (Figure. 1.1D). As shown in Figure. 1.1E, when two mutations without an environmental effect interact with one another, epistasis (ϵ) will cause the fitness of the double mutant to deviate from additivity. This is the usual definition of epistasis in the literature, which is invariant as to the order in which mutations occur [Poelwijk et al., 2007]. In contrast, when at least one of the single mutants has an environmental effect, the double mutant experiences a different extracellular environment depending on which of the two single mutants was its immediate ancestor. For example, a double mutant could cross-feed on the byproducts of one of its possible single-mutant ancestors but not on the byproducts of the other (Figure. 1.1F). The result is a gene-by-environment-by-gene (G x E x G) interaction in which the magnitude of epistasis may depend on the order in which mutations occur. In other words, epistasis becomes noncommutative. The value of that noncommutative fitness shift (δ) characterizes the deformation of a two-step mutational trajectory (Figure. 1.1F).

Noncommutative epistasis and fitness intransitivity are closely related but not identical concepts (Supplementary Figure. 1.3) [de Visser and Lenski, 2002]. In its simplest,

qualitative formulation, "intransitivity" refers to situations in which the fitness of three mutants (A, B, C) in pairwise competition are nonhierarchical (i.e., A invades B, B invades C, and C invades A). A less stringent, quantitative definition of intransitivity has been applied when the relative fitness between a mutant and its ancestor cannot be predicted by the sum of cumulative fitness gains along a mutational trajectory [de Visser and Lenski, 2002]. This definition is close to but distinct from the concept of noncommutativity (Figure. 1). Noncommutativity quantifies the difference in cumulative fitness gains along two different mutational trajectories without regard for the fitness of the final point of the trajectory in competition with the original ancestor (Supplementary Figure. 1.3). Interestingly, noncommutativity and intransitivity are mathematically related to one another but must be estimated using independent experiments (Supplementary Figure. 1.3). Genotypes along an evolutionary trajectory usually compete with their immediate mutational ancestors rather than with their original ancestral strain [Paquin and Adams, 1983]. Therefore, noncommutativity is a suitable metric for characterizing fitness landscape deformability, while intransitivity can be more suitable for ecological questions, such as the possibility of coexistence of different genotypes [Kerr et al., 2002, Rainey and Travisano, 1998]

1.3.2 Deformability in the Path to an Evolutionary Innovation in E. coli

To experimentally validate and assess the potential relevance of noncommutative epistasis as a metric of fitness landscape deformability, we studied two mutations on the path to the evolutionary innovation of strong aerobic growth on citrate (Cit++) in the Ara-3 population of the LTEE [Blount et al., 2008]. The two principal mutations underlying this phenotype are known to have profound ecological consequences, suggesting that noncommutative epistasis may be present (Figure. 1.2A). The first mutation is a tandem amplification overlapping the citrate fermentation operon, cit, which occurred after 31,000 generations. This amplification caused aerobic expression of the CitT transporter, producing a weak citrate growth phenotype (Cit+) [Blount et al., 2012]. CitT is an antiporter that imports citrate, present in large amounts in the LTEE DM25 growth medium, while exporting intracellular C4-dicarboxylate TCA intermediates, e.g., succinate and malate [Quandt et al., 2015], thereby increasing their concentration in the extracellular environment. A subsequent mutation causes high-level, constitutive expression of DctA, a proton-driven dicarboxylic acid transporter. This mutation refines the Cit+ trait to Cit++ by allowing recovery

of the C4-dicarboxylates released into the medium by both the progenitor and the double mutant itself during growth on citrate (Figure. 1.2A) [Quandt et al., 2015]. We reasoned that these mutations together enable the exploitation of environments built by progenitor strains, producing a stronger increase in fitness than expected in the absence of environmental construction (Figure. 1.2B). In contrast, had the DctA mutation occurred before the CitT-activating duplication, it would have conferred no fitness benefit and would not have produced any changes in the environment relative to the ancestor (Figure. 1.2B).

We tested this prediction by performing competitive fitness assays with different combinations of a spontaneous Cit– mutant and dctA– knockout strains derived from ZDB89, a 35,000-generation Cit++ clone that possesses both the DctA-activating and CitT-activating mutations (Materials and Methods). Competitions were carried out with equal volumes of each combination of competitors, and relative fitness was determined using colony counts obtained after 0 and 24 h of growth [Lenski et al., 1991]. In parallel, we used our dFBA model to simulate these competitions, relying solely on known parameters from the experiments and on published parameters pertaining to the physiology of E. coli [(Materials and Methods) Harcombe et al. [2014], Gallet et al. [2017]]. Confirming our expectations, the dFBA model predicts strong noncommutative epistasis ($\delta = 1.50$) (Figure. 1.2C). This is confirmed by the experimental results ($\delta = 1.78 \mp 0.15$) (Fig. Figure 1.2D). The agreement between the empirically calibrated computational model and the experiments is not only qualitative but also is quantitative: With no fitting parameters, dFBA is predictive of the outcome of the experimental pairwise competitions, explaining 52% of the variance in colony counts from all experiments (n = 120) (Supplementary Figure. 1.4).

1.3.3 Short-Range Deformability in E. coli Is Weak and Rare

Although the above examples demonstrate the potential presence of fitness landscape deformability, its pervasiveness in empirical fitness landscapes remains unclear. To address this question, we screened the entire first- and second-order mutational neighborhood of E. coli using our computational model (Figure. 1.3A). In Figure. 1.3B we represent all pairs of mutations that exhibit deformability as nodes in a network that are connected if their noncommutative fitness shift (δ) is larger than 1% of the fitness effects (F_{MAX}; also see Supplementary Figure. 1.5). These represent only a small subset ($203/3343$, or 6.1%) of all epistatic interactions, which for the most part are not altered by the environmental effects of mutations.

Noncommutative interactions also tend to be unevenly distributed: Most mutations do

not deform the fitness of any other mutation, and only 15 (0.3%) of them deform the fitness of five or more other mutations (Figure. 1.3B and 1.3C). These few highly connected hubs on the network tend to be the mutations with the strongest environmental effects (Pearson's $\rho = 0.79, P < 10^{-6}$) (Supplementary Figure. 1.6). Noncommutative epistasis also tends to be small in magnitude (Figure. 1.3D); only 1.6% ($55/3343$) of epistatic pairs have a noncommutative epistatic shift larger than 10% of the total fitness increase ($\delta/F_{MAX} > 0.1$) (Figure. 1.3D). This reveals that the deformability of the local mutational neighborhood of the E. coli metabolic landscape is generally weak, rare, and highly anisotropic (i.e., nonhomogeneous), with deformations limited to localized directions in genotype space.

1.3.4 Long-Range Deformability of the E. coli Metabolic Fitness Landscape

The low deformability of the local mutational neighborhood could be explained by the strong genetic similarity between the mutants and the ancestral genotype: Genotypically close descendants will rarely be able to use metabolites that are discarded by their immediate ancestors. By the same logic, one may predict that over longer mutational distances metabolic differences might accumulate that enable the use of extracellular metabolites that are left as a "legacy" by previous mutations. Thus, we hypothesize that changes to the extracellular environment produced by a given mutation will primarily deform the fitness landscape at distant positions on the genotype space.

To test this hypothesis, we set out to introduce a mutation with a strong environmental effect and measure the deformation it causes at different distances in the genotype space. We chose the ACKr (acetate kinase) mutation (the deletion of the acetate kinase gene), which as shown in Figure. 1.1C modifies the environment by releasing large amounts of lactate at the expense of lower secretions of formate, acetate, and ethanol. To quantify the deformation introduced by this mutation, we compared the fitness of thousands of genotypes at increasing mutational distances from the ancestor in competition with either the ancestor E. coli model (A) or the ACKr mutant (M) (Figure. 1.4A). The deformation introduced by M at genotype G is thus quantified by the parameter $\Delta Fitness = |F_G^{(M)} - F_G^{(A)} - F_M^{(A)}|$ (Figure. 1.4B). Consistent with our hypothesis, and as shown in Figure. 1.4B and 1.4C, we found that the fitness landscape deformation $\Delta Fitness$ introduced by the ACKr mutation is negligible at short genotypic distances from it (e.g., 16 mutations or less), but it becomes stronger at longer distances. Fifteen other mutants in addition to

ACKr were also tested, with similar results (Supplementary Figure. 1.7.). Furthermore, by comparing the growth rate of thousands of genotypes in the environments constructed by A and M (noted by E_A and E_M, respectively), we found that increasingly distant genotypes become increasingly sensitive to the differences between the two environments (Figure. 1.4D and Supplementary Figure. 1.8). This provides an explanation for the observed pattern of fitness landscape deformation as a function of genotypic distance.

What are the genetic mechanisms underlying the long-range environmental effects of new mutations on growth rate? One possibility could be an increased probability of sampling mutations that produce a difference in growth rate between E_A and E_M. Alternatively, this effect could be caused by genetic interactions between two or more mutations that allow distant genotypes to use differently the resources secreted by A and M. To discriminate between these two possibilities, we compared the observed difference in growth rates (Figure 1.4D, gray line) with the difference expected if mutations do not interact (Figure 1.4D, red line) (Materials and Methods). As shown in Figure 1.4D, the null model that only incorporates increased sampling of mutations at growing mutational distances (while assuming no interactions) vastly underestimates the observed growth difference between E_A and E_M and thus is insufficient to explain our results. This suggests that, although both mechanisms are present, interactions between mutations dominate the deformation of the fitness landscape at large mutational distances (see also Supplementary Figure. 1.8).

To mechanistically illustrate the role of complex genetic interactions in long-range landscape deformation, in Figure 1.4E we show an adaptive trajectory in which a first mutation (lactate dehydrogenase; LDH) causes the release of lactate to the extracellular space. A complex metabolic innovation involving several reaction additions [ATP synthase (ATPS), pyruvate formate lyase (PFL), and ACKr] [Pál and Papp, 2017] is subsequently required to confer the ability to metabolize this lactate (Figure 1.4E; see also Supplementary Figure. 1.8). Notably, lactate becomes metabolized only by the final genotype, which contains all three required mutations.

1.4 Discussion

Darwin [Darwin, 1892] was perhaps the first to recognize that the environment experienced by an evolving population can also be shaped by the population itself. Long neglected, this concept was revived by Lewontin [Lewontin, 1978, 1983], and has gained added momentum in recent years as the important role played by eco-evolutionary feed-

backs in both ecology and evolution has become better appreciated [Laland et al., 2014, Post and Palkovacs, 2009, Rudman et al., 2018]. Due to technical limitations, experimental studies of eco-evolutionary feedbacks and the adaptive dynamics models that seek to explain them often lack explicit, genome-wide representations of the adaptive landscape, in particular with regard to complex traits and gene–gene interactions [Rudman et al., 2018]. The exact state of the environment, which is intrinsically complex and multidimensional, is also rarely measured experimentally or explicitly included in eco-evolutionary models. In return, genome-wide genotype fitness maps have largely ignored the effects of eco-evolutionary feedbacks, despite early abstract models of species coevolution, which introduced the idea of fitness landscape deformability (also referred to as "rubberness," refs. Kauffman and Johnsen [1991] and Solé and Sardanyés [2014]), and the many examples of their importance in coevolutionary arms races and other forms of coevolution [Morran et al., 2011]. This is particularly important in light of the argument, made by many authors, that the deformability of fitness landscapes (or its consequences, in the form of frequency-dependent selection) would erode their practical and conceptual utility [Schuster, 2012, Doebeli et al., 2017, Moran, 1964]

Our work empirically addresses this latter argument. Encouragingly, our results show that fitness landscapes may retain their local properties in the presence of mutations that significantly alter the environment. By systematically mapping an empirical fitness landscape, we have found that ignoring deformability and assuming a rigid landscape is a good approximation over short genotypic distances. This is because closely related genotypes are unlikely to differ from one another in their physiological response to the built environment. In contrast, over longer mutational distances, fitness landscapes are likely to be affected by environmental construction, an effect that is shaped by complex genetic interactions. This suggests an ecologically mediated mechanism by which historical contingency may shape downstream evolution even in clonal populations. In summary, our work suggests that, depending on the scale at which they are examined, fitness landscapes can either behave as a fixed externally determined topography on which adaptation proceeds, or become a dynamic property of the populations adapting on them [Doebeli et al., 2017, Moran, 1964].

One limitation of our study is the inability of our model to predict changes in the sign of the fitness effect of a new mutation. This is a common limitation of most FBA-based models (but see refs. Mori et al. [2016] and Beg et al. [2007]), as they do not consider the potential costs of adding a new biochemical reaction, or of maintaining a flux through it. These costs can arise in microbial cells either through the cost of increasing genome

size [Giovannoni et al., 2014] or through the cost of expressing the enzymes required for the new reaction. Given the absence of such costs in our FBA model, a deletion can never provide any advantage, and an addition will never be detrimental. Costs have already been incorporated in nondynamic FBA models [Mori et al., 2016, Beg et al., 2007], allowing the prediction of phenomena such as overflow metabolism. One can certainly imagine situations in which an addition that is detrimental due to its maintenance cost could become beneficial in the presence of the metabolic byproducts of its ancestor, leading to an ecologically mediated inversion of fitness effect (or "sign-δ"). Incorporating costs into a dynamic genome-scale modeling framework represents a promising future direction.

The idea that under frequency-dependent selection fitness landscapes change as populations move on them has been conceptually discussed and studied within the theoretical framework of adaptive dynamics [Kauffman and Johnsen, 1991, Watson and Ebner, 2014, Waxman and Gavrilets, 2005]. A solution to the problem of fitness landscape deformability was found in the formulation of the invasion fitness landscape, i.e., the map between the relative fitness S(x,y) of an invader with phenotype y against a resident phenotype x [Waxman and Gavrilets, 2005]. In principle our results and methods might allow one to map an empirical invasion fitness landscape, at least locally. However, one would need to identify a scalar phenotype that can be mapped to the invasion success against a resident genotype in the environment this resident constructs. Under what conditions this is possible is an open question that lies beyond the scope of this study, but it poses an interesting future challenge.

In line with this discussion, our results indicate that simulating cellular adaptive dynamics with an explicit and biologically realistic genome-wide representation of the genotype–phenotype map is within reach. Such an approach will shed light into the role played by dynamic niche construction in cellular evolution. We believe that it will also create multiple opportunities to incorporate genomics into the study of eco-evolutionary dynamics and thus reveal the genetic, biochemical, and environmental constraints that simultaneously govern the ecology and evolution of cellular populations.

1.5 Figures

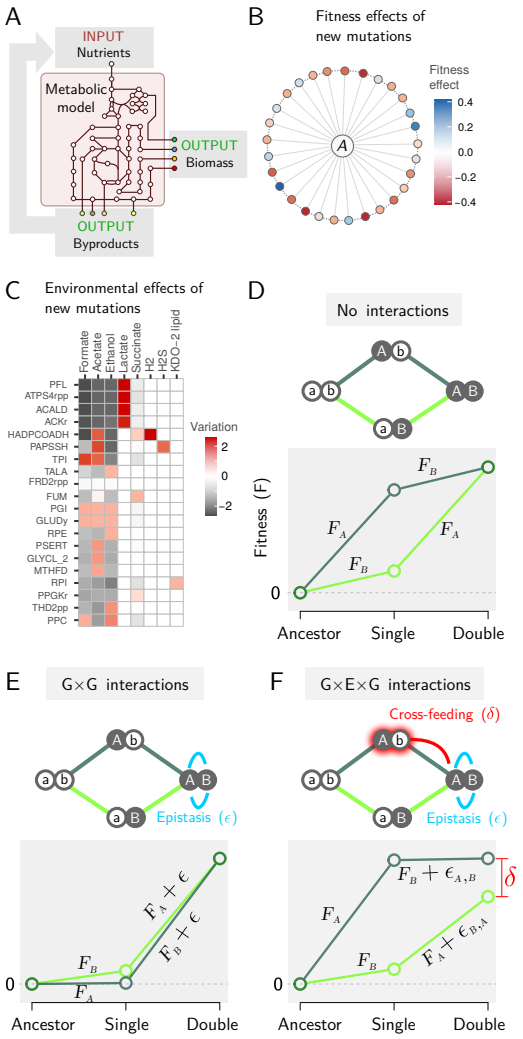


Figure 1.1: Measuring deformability in the E. coli metabolic fitness landscape.

Figure 1.1: (A) Schematic depiction of dFBA simulations. Given an input in the form of nutrients, metabolic fluxes through an explicit and empirically curated metabolic model are optimized to maximize the biomass growth yield. The optimal metabolic fluxes produce metabolic byproducts that are released to the external environment, becoming part of future inputs. (B) A subset of genotypes differing from our E. coli metabolic model by a single mutation (an added or deleted reaction), colored according to their effect on fitness in competition with the ancestor (A). (C) Environmental effects of a subset of mutants expressed as the variation in the profile of secreted metabolites compared with the ancestral E. coli genotype (computed as log-modulus transformed difference in the amount of a given secreted molecule; Materials and Methods). Mutant labels are given in Biochemical Genetic and Genomic (BiGG) database notation. (D) Two loci fitness landscapes in the absence of gene-gene interactions in which the fitness effect of each mutation is the same in all genetic backgrounds. The fitness of each genotype was calculated in direct competition with its immediate ancestor. Mutations A and B correspond to the addition of GLYCL _2 (glycine cleavage system) and AIRCr (phosphoribosylaminoimidazole carboxylase), respectively. (E) Two-loci fitness landscapes with gene–gene interactions giving rise to epistasis (ϵ). Mutations A and B were SO3R (sulfite reductase) and PAPSSH (phosphoadenylylsulfatase), respectively, simulated in a constant environment. (F) Two-loci fitness landscapes in which one of the mutants transforms the environment, leading to cross-feeding toward the double mutant. Mutations A and B correspond to the addition of PAPSSH and HADPCOADH (3-hydroxyadipyl-CoA dehydrogenase). In addition to regular epistasis, this led to a noncommutative epistatic shift ($\delta = \epsilon_{A,B} - \epsilon_{B,A}$).

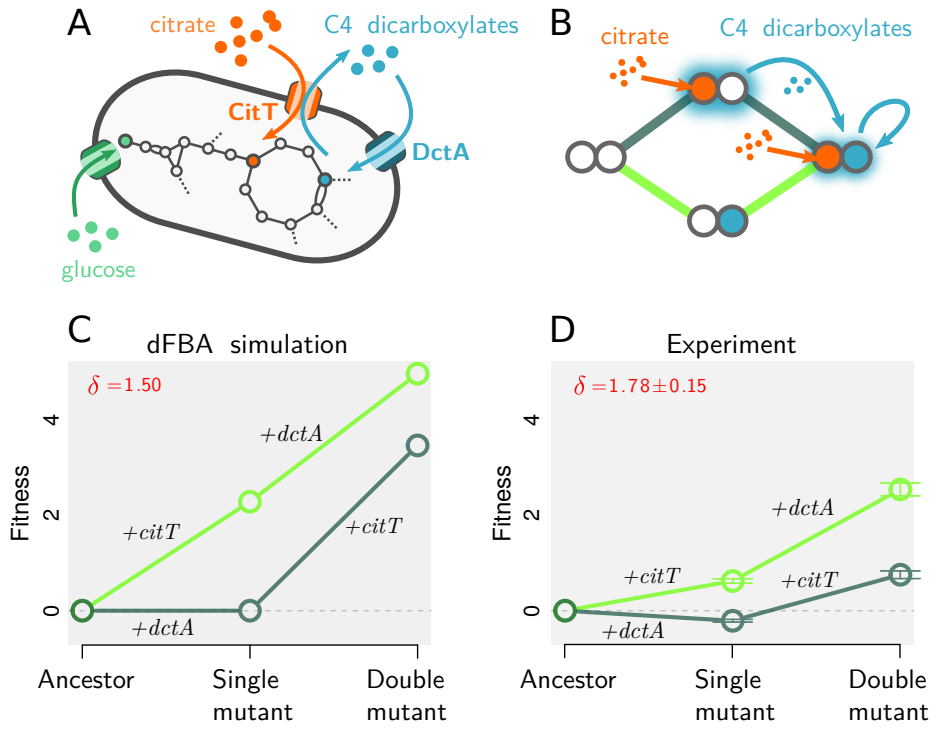


Figure 1.2: Noncommutative epistasis in the evolution of aerobic citrate use in E. coli. (A) Function of the two transporters involved in the innovation of strong aerobic growth on citrate (Cit++) in E. coli. CitT is an antiporter that exchanges extracellular citrate for internal C4-dicarboxylates (e.g., succinate, fumarate, and malate). DctA is a carboxylic acid transporter that imports C4-dicarboxylates from the extracellular space into the cytoplasm. (B) The two possible mutational trajectories leading to the Cit++ trait. If the mutation leading to expression of citT (+citT) occurs first, it will transform the environment leading to cross-feeding toward the double mutant. This should not occur if the dctA overexpression mutation (+dctA) occurs first. (C and D) Simulated (C) and experimentally measured (D) fitness landscapes in the DM25 medium used in the LTEE (Materials and Methods). Experimentally obtained values are reported as mean ± SEM (n = 10).

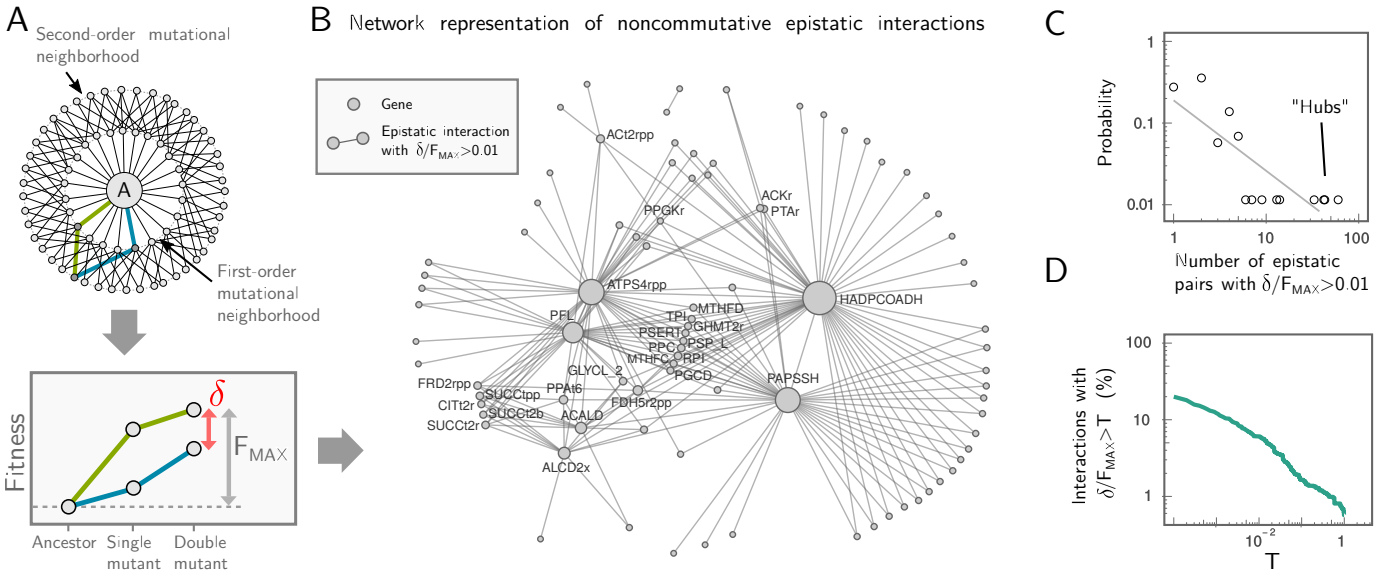


Figure 1.3: Short-range deformability in E. coli is rare, weak, and directional. (A) Systematic exploration of the second-order mutational neighborhood of E. coli. We exhaustively simulated every possible mutational trajectory starting from the ancestral (A) metabolism and ending in each double mutant. Noncommutative epistasis (δ) was measured for each pair of mutants and was normalized by F_{MAX}, i.e., the maximal cumulative fitness effect of the double mutant: $F_{MAX} = max[|F_i^{(A)} + F_{ij}^{(i)}|, |F_j^{(A)} + F_{ij}^{(j)}|]$, where, e.g., $F_x^{(y)}$ denotes the fitness of mutant x when invading its immediate ancestor y at low frequency (Materials and Methods). (B) Network representation of all noncommutative epistatic pairs. Nodes represent mutations, and two nodes are joined by an edge if $\delta/F_{MAX} > 0.01$ for that pair. Node labels (BiGG database notation) are shown for hubs (mutations with more than four noncommutative interactions). (C) Distribution of deformability for each gene in the network, measured as the number of other genes with which it has a noncommutative epistatic interaction. (D) Strength of all noncommutative epistatic interactions, i.e., the percentage of epistatic pairs with $\delta/F_{MAX} > T$ as a function of T.

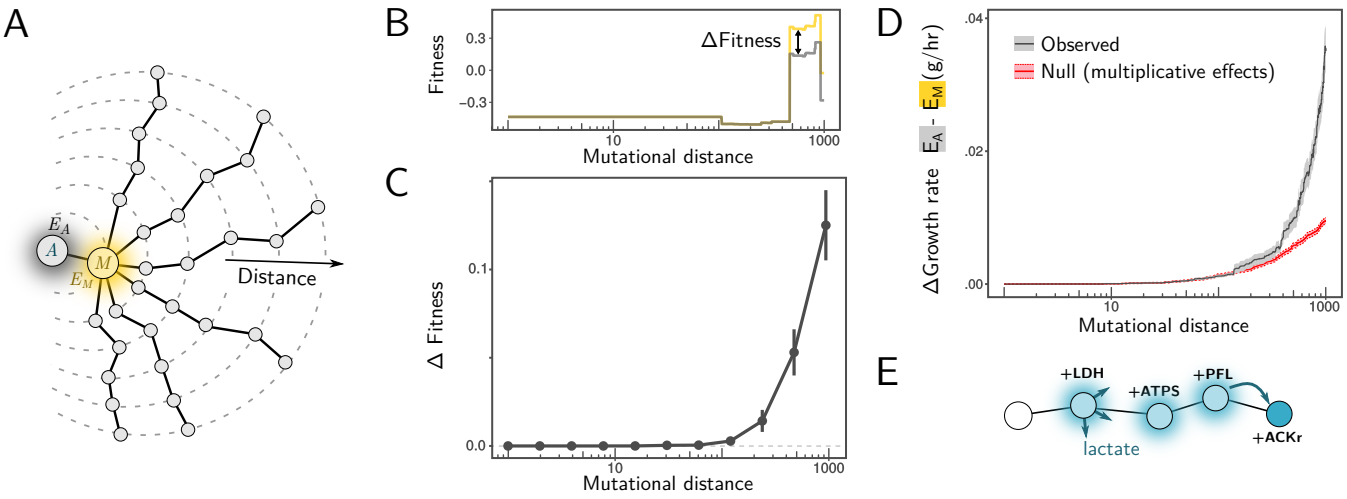


Figure 1.4: Long-range deformability of the E. coli metabolic fitness landscape. (A) We performed random walks (length = 1,000 mutations) in genotype space starting from an E. coli ancestor (A; gray) and first passing through a mutant (M; orange) with large environmental effect. (B and C) Fitness of mutants along these random walks was measured in competition with A (gray) in the environment it generates (EA), as well as in competition with M (orange) in the environment it generates (EM). In B we show the result for a single example of a random walk. Note that fitness in competition with M is shifted by the difference in fitness between M and A so all observed differences in fitness are due to deformation (for any genotype G, $\Delta Fitness = |F_G^{(M)} - F_G^{(A)} - F_M^{(A)}|$). (C) Average$\Delta$ Fitness at increasing mutational distances from A in over n = 100 random walks (error bars represent SEM; n = 100). (D) Average difference (absolute value) in growth rate between environments EM and EA (in grams of dry cell weight xh^{-1}) at varying genotype distances (gray line; shading represents SEM; n = 1,000). In red, we show the predicted difference in growth rates for a null model that assumes independent effects of mutations (Materials and Methods). (E) An example of an adaptive trajectory showing complex genetic interactions in a long-range deformation. The addition of LDH leads to the release of lactate as a by-product. Three additional mutations, ACKr, ATPS, and PFL, are required together for lactate to be used by a descendant genotype.

1.6 Supplementary Material

1.6.1 Reconstruction of a prokaryotic genotype space

All in silico explorations of genotype [King et al., 2016] space in this work took as a reference the E. coli model iJO1366 and consisted of both gene additions and deletions. Gene deletions were performed by constraining both upper and lower bounds of the reaction to zero. Gene additions were performed from a set of all known prokaryotic reactions. We used the BiGG database [King et al., 2016] to compile a dataset of all known reactions found across prokaryotic species. Conflicts in reaction directionality were resolved as follows i) if a reactions is found in the well benchmark E. coli iJO1366 model, use the properties given by this model, ii) if a reaction conflicts in directionality, only accept directions found across all models (e.g. if there is one model where a given reaction is irreversible, we set it as irreversible). We used this dataset to create a "universal" metabolic model that included all reactions found in E. coli iJO1366 as well as a set of all potential novel reactions. We removed reactions that would lead to erroneous energy-generating cycles using the GlobalFit algorithm [Fritzemeier et al., 2017]. The algorithm was constrained to conserve reactions present in the original E. coli model. Removing any futile cycles from this "universal" model ensures that there will not be any futile cycles in any subset. The resulting network contains 4999 metabolic reactions and 585 nutrient uptake or sink reactions, of which 2758 and 255 were not found in the original E. coli model.

1.6.2 In silico simulation of growth through metabolic modeling

Dynamic Flux Balance Analysis simulations were performed using the COMETS package ("Computation of Microbial Ecosystems in Time and Space", [Harcombe et al., 2013]) and the gurobi optimizer software. For computationally intensive simulations, we used the High Performance Facility at Yale University. For standard (non dynamic) FBA simulations, we used the COBRApy python package [Ebrahim et al., 2013a]. Both Dynamic and Standard FBA optimizations were done using the parsimonious algorithm, in which a first optimization is done to maximize biomass yield, and a second one fixes this yield and minimizes total fluxes throughout the network [Lewis et al., 2010]. Unless otherwise stated, the default V_{max} was set in dynamic FBA simulations to $-10mmol \times gr^{-1} \times hr^{-1}$ for all uptake reactions. Inorganic ions and gases where kept at high concentrations and where kept undepleted throughout the simulation (i.e lower bound : $-1000mmol \times gr^{-1} \times hr^{-1}$

, amount of metabolite: 1000 mmol). This was done to constrain our analysis to situations where growth is limited only by uptake of carbon sources. The unbounded nutrients are: ca2_e, cbl1_e, cl_e, co2_e, cobalt2_e, cu2_e, fe2_e, fe3_e, h_e, h2o_e, k_e, mg2_e, mn2_e, mobd_e, na1_e, nh4_e, ni2_e, pi_e, sel_e, slnt_e, so4_e, tungs_e, zn2_e. For the citrate simulation to avoid oxygen, nitrogen or proton limitation uptake was unconstrained by setting the V_{max} to $1000mmol \times gr^{-1} \times hr^{-1}$. Analysis of results was performed using GNU R language [R Core Team, 2021].

1.6.3 Fitness, environmental effects and deformability measurements

To measure fitness, we use here (in both experiments and simulations) the Malthusian fitness measure that allows for a quantitative comparison across environments [Wagner, 2010]. Fitness of mutant M relative to ancestor A is therefore given as $F_M^{(A)} = log([X'_M/X_M]/[X'_A/X_A])$ where X and X' represent initial and final densities. For a pair of mutations, deformability can be then measured as: $\delta_{ij} = F_{ij}^{(i)} + F_i^{(A)} - F_{ij}^{(j)} + F_j^{(A)}$ where $F_x^{(y)}$ represents the fitness of genotype x in competition with genotype y. To compute environmental effects of mutations, the difference in secretion profile of mutants (as shown in Figure. 1.1C and Supplementary Figure. 1.1) is computed for a given released molecule as $sign(D) \times (log(D) + 1)$ where D is the amount released by the mutant minus that secreted by Ancestral E. coli. This log-modulus transformation [John and Draper, 1980] is applied to help visualization of the generally small differences in released amount, which can be either positive or negative. To measure environmental effect of a mutation (as used in Supplementary Figure. 1.2 and Supplementary Figure. 1.5), we use the Euclidean distance in the profile of released metabolic byproducts between a mutant and the E. coli ancestor using standard Flux Balance Analysis [Ebrahim et al., 2013a].

1.6.4 Simulation of the fitness landscape of aerobic growth on citrate by E.coli

Starting with E.coli model iJO1366 we constructed metabolic models of the four mutants necessary to predict the fitness landscape involved in the evolution of aerobic citrate utilization in the Ara3 population of the LTEE. Unlike the LTEE ancestral strain REL606 (and E. coli generally), which possess the necessary genes for citrate utilization but do not express them in aerobic conditions, iJO1366 is able utilize both citrate and succinate if

these reactions are unbounded (as FBA optimizes precisely regulation). Thus, the ancestral phenotype was recreated by knocking out three reactions CITt7pp (citT), SUCCt2_2pp (dctA) and SUCCt2_3pp (dcuA or dcuB). The reactions encoded by the first two genes (citT and dctA) are known to be involved in the evolution of aerobic growth on citrate in the LTEE whereas dcuA and dcuB are involved in dicarboxylate uptake in anaerobic conditions and are inactive in aerobiosis [Six et al., 1994]. This triple knockout represents the pre-citrate E. coli ancestor strain. The addition of CITt7pp simulates the promoter capture and consequent aerobic expression of CitT. Similarly, the addition of SUCCt2_2pp is equivalent to the first mutation (aerobic expression of dctA). We used dynamic FBA to predict the fitness landscape of these two mutations, calibrating the simulations to reflect the the experimental conditions. This involved i) setting the in silico media to reflect DM25 minimal glucose media (0.139mM glucose, 1.7mM citrate). Aerobic condition was simulated by keeping oxygen (o2_e) undepleted. ii) using published parameters pertaining to the physiology of E. coli (3) and iii) estimating the initial biomasses of each mutant prior to competition. Initial biomass for citrate simulations was determined using initial plate counts from pairwise competitions experiments (see also Supplementary Figure. 1.4.). We assume that average cell dry mass is $3.9 \times 10^{-13} g$ which is the empirically measured cell dry mass of REL606 the ancestral strain used in the LTEE [Gallet et al., 2017]

1.6.5 E. coli Long-Term Evolution Experiment

Briefly, twelve populations of E. coli B were founded in 1988 from clone REL606. The populations were initially identical, save for half having a mutation that permitted growth on arabinose. (See below.) These have since been evolved in DM25 minimal glucose medium under conditions of daily, 100-fold serial transfer, and incubation at 37° with 120 rpm orbital shaking. Samples of each population are frozen every 500 generations 38 . DM25 is Davis-Mingioli broth supplemented with 25 mg/L glucose. (Per liter: 7g potassium phosphate dibasic trihydrate, 2g potassium phosphate monobasic anhydrous, 1g ammonium sulfate, 0.5g sodium citrate, 1mL 10% magnesium sulfate, and 1mL 0.2% thiamine.)

1.6.6 Isolation and Preparation of Test Strains

ZDB89 is a Cit++ clone isolated from the Ara-3 population sample frozen for generation 35,000 during the LTEE. Cit– revertants arise spontaneously from Cit+ and Cit++

clones due to recombination-mediated collapse of the tandem cit amplification to the ancestral genotype at that locus. We isolated a Cit– revertant, ZDB757, by first passaging ZDB89 in a glucose-only medium for five days. This passaging does not constitute a selection, but nonetheless enriches for Cit– revertants by eliminating the selective penalty for losing the ability to grow on citrate. Passage cultures were spread on LB plates, and Cit – mutants screened for by patching colonies to LB and Minimal Citrate (MC) plates to identify clones that no longer grew on citrate. The Cit– phenotype was confirmed by streaking on Christensen's Citrate Agar. Recombineering with the pKO3 suicide plasmid [Link et al., 1997] was used to delete the dctA gene from ZDB89 and ZDB757, producing the Cit+ dctA– and Cit– dctA– constructs, ZDB912 and ZDB904, respectively. To permit differentiation of competitors during fitness assays, we isolated Ara+ revertants of each of the aforementioned clones and constructs. Briefly, Ara– strains lack the ability to use arabinose, and form red colonies on Tetrazolium Arabinose (TA) plates, while Ara+ revertants are mutants with restored ability to grow on arabinose, and form white colonies on TA. The ancestral strain of the Ara-3 population and its descendants are Ara– . We isolated Ara+ revertants by plating clone or construct cultures on Minimal Arabinose (MA) plates. Revertants were competed against their Ara– parents to verify marker state neutrality. Clones, constructs, and revertants are listed in Supplementary Table 1. Derivation of constructs and revertants are shown in Supplementary Figure. 1.10 .

1.6.7 Experimental fitness Assays

Fitness was assayed in pairwise competitions. Competitors with opposite Ara marker states were inoculated from frozen stocks into 10 mL LB broth, and incubated overnight at 37° with 120 rpm orbital shaking to permit revival and elimination of traces of glycerol cryoprotectant. To precondition the competitors, each competitor revival culture was then diluted 100-fold in 0.85% saline, and 100 L of the diluted culture used to inoculate 9.9 mL DM25 with ten-fold replication. These culture were grown for 24 hours at 37° with 120 rpm orbital shaking, after which they were transferred via 100-fold dilution into 9.9 mL volumes of fresh DM25, and grown for another 24 hours under the same conditions. Ten competition cultures were prepared for each competitor pairing by inoculating each 9.9 mL DM25 with 50 L of each preconditioned competitor. A single replicate preconditioning culture of each competitor for each competition was inoculated so that each competition was inoculated from a single preconditioned culture of the competitors. Upon inoculation with the competitors, 100 L of a 100-fold dilution of each was spread on TA

to permit enumeration of the initial frequency of each competitor. 100 L of a 1000-fold dilution was also plated for each competition including at least one Cit+ or Cit++ competitor. Colonies were counted following 48 hours of plate incubation at 37°. Following 24 hours incubation under the same conditions used for preconditioning, 100 L of 10,000-fold dilutions of each competition were plated on TA to permit final enumeration of the competitors. 100,000-fold dilutions were also plated for competitions including at least one Cit+ or Cit++ competitor.

1.6.8 Exploration of deformability in the local mutational neighborhood of E. coli

To systematically analyze the local mutational neighborhood of E.coli we construct a set of metabolic models consisting of every viable single and double mutation, considering both additions and deletions from our universal reaction set and using as a reference the E. coli iJO1366 model (4389 and 9636050 genotypes, respectively). We removed from this analysis and all subsequent analysis essential genes, sink reactions, diffusion reactions, as well as those genes leading to artifacts (H2 or CO2 limitation). We used dynamic flux balance analysis to simulate competition assays of each mutant with its immediate ancestor. We chose to perform our in-silico competitions in anaerobic conditions because in under aerobic conditions FBA incorrectly predicts complete oxidation of glucose at saturating level. The simulations started with an initial glucose concentration of 0.0001mM and assayed co-culture growth during 10hr, a period during which glucose was never exhausted, i.e. growth remained exponential. All simulations were done with the mutant starting at low frequency ($1\%, 1.0 \times 10^{-10} gr.$ dry cell weight, for 9.9×10^{-9} gr. for the ancestor) in anaerobic glucose minimal media (unless otherwise stated, see detailed parameters in supplement and Supplementary Table 2).

1.6.9 Simulation of long-range fitness landscape deformation

In order to explore the long-range effects of landscape deformation, we started from a one-step mutation from ancestor E. coli model and performed random walks in genotype space by sampling 1024 mutations (without replacement) among both deletions and additions. In addition to previously mentioned artifacts we excluded from this analysis reactions that had led to CO2 or H2 limitation in a unique pair, even if they did not have this effect with other reactions . The pairs are: SHSL2 and SHSL2r ,DHORD_NAD and DHORDi, ENO

and HADPCOADH, LEUTA and LLEUDr, P5CRx and PRO1y, in BIGG database notation (1). To prevent irreversible loss of viability, the sampling procedure also ignored all reactions that were essential in a minimal model capable of growing on glucose in anaerobic conditions. The minimal model was built by sequentially removing reactions while possible, following [Pál et al., 2006]). At regular intervals along the random walk, fitness was measured as before in competition with the mutant and the ancestor (wild-type) using dynamic flux balance analysis (COMETS [Harcombe et al., 2014]). To determine the growth rates of genotype in ancestral vs mutant environments we repeated this procedure except at each step, growth rate was measured in the environment provided by the mutant and the ancestor using standard flux balance analysis (COBRApy (4)). These environments were simulated by setting uptake rates for each secreted metabolites to the excretion rate of the respective ancestor.

1.6.10 Computation of null models for growth in the absence of epistasis

We built a null model for the expected growth rates of mutants, under the assumption that the effect of each mutation on the growth rate is independent. We denote the growth rate of the ancestral genotype M by g_M , and that of each single mutant i by $g_i = w_i g_M$, where i can represent any of all N possible mutations (in our case N=4181). Here we introduced the parameter $w_i = g_i/g_M$ representing the relative effect of mutation i on the growth rate. If two mutations i and j do not interact with one another, their effects on growth rate are multiplicative: $w_{ij} = w_i w_j$. For a mutant that contains Q mutations relative to the ancestor M, we can thus calculate its growth rate relative to the ancestor as:

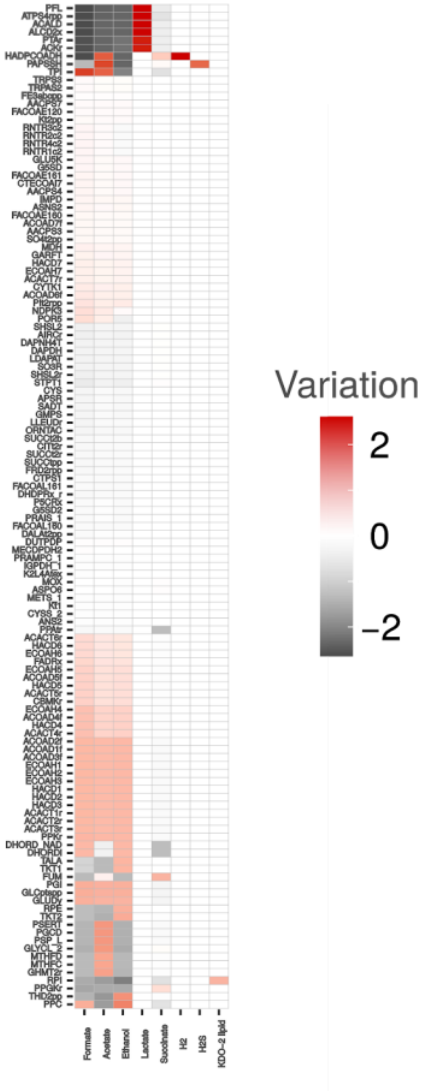
$$g_Q = \left(\prod_{i=1}^{P} w_i \right) g_M \tag{1.1}$$

1.6.11 Long range deformation in an adaptive trajectory

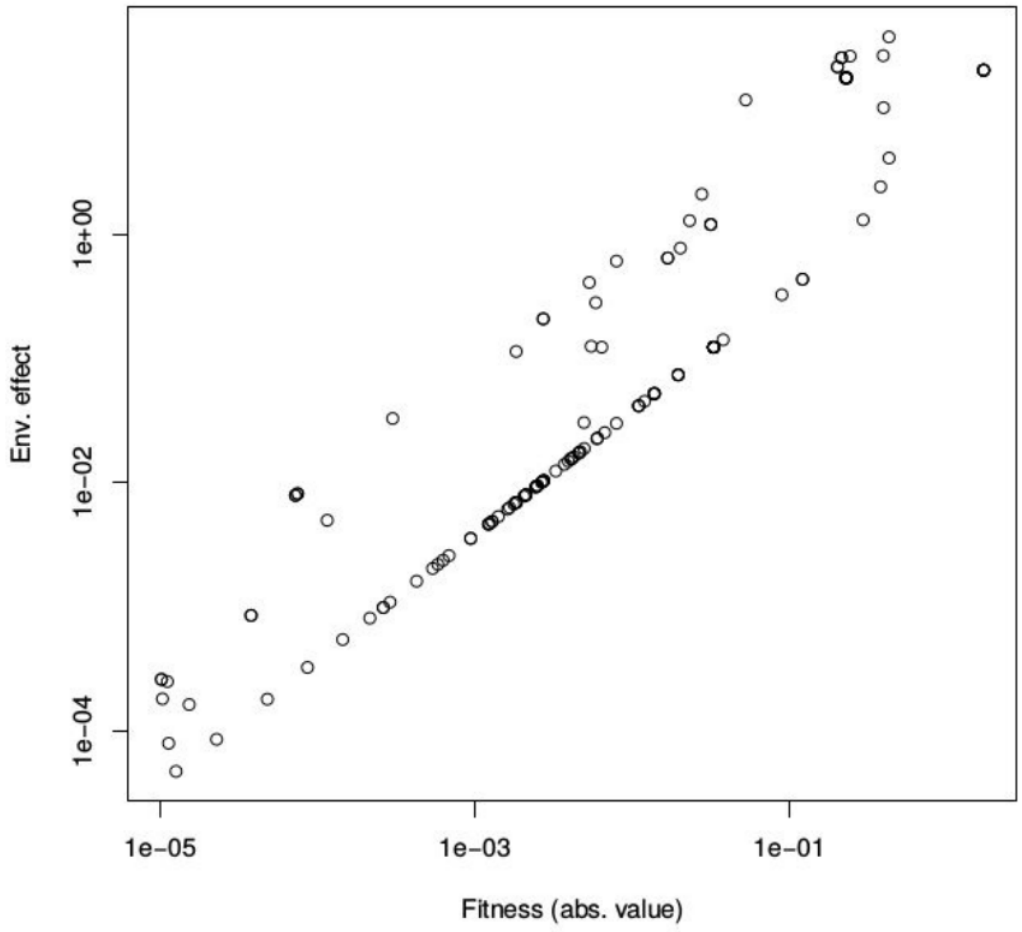
To provide a mechanistic example of an adaptive long-range deformation with epistasis, we sequentially removed reactions that i) had a detrimental effect, and ii) affected lactate secretion. We reached a 6 step mutant genotype that was unable to use lactate. The subsequent addition of 4 of these removed reactions to this mutant led first to the secretion of lactate (upon addition of LDH, lactate dehydrogenase), and then to the consumption of

this lactate only after 3 additional mutational steps (ACKr - acetate kinase, PFL - pyruvate-formate lyase, ATPs - cytosolic ATP synthase). For each mutation along the trajectory (LDH -ATPs-PFL-ACKr,) we measured fitness as before, by simulating the competition of each mutant with its immediate ancestor. The effect of lactate cross-feeding on fitness was assessed by repeating this analysis, albeit with lactate removed from the environment at each step (Supplementary Figure. 1.8).

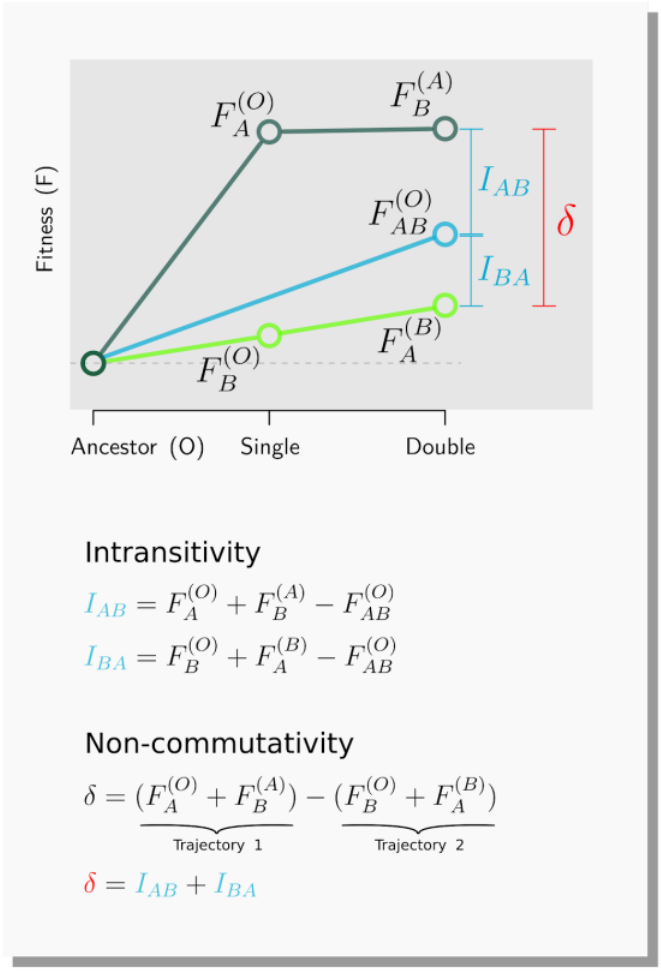
1.6.12 Supplementary Figures



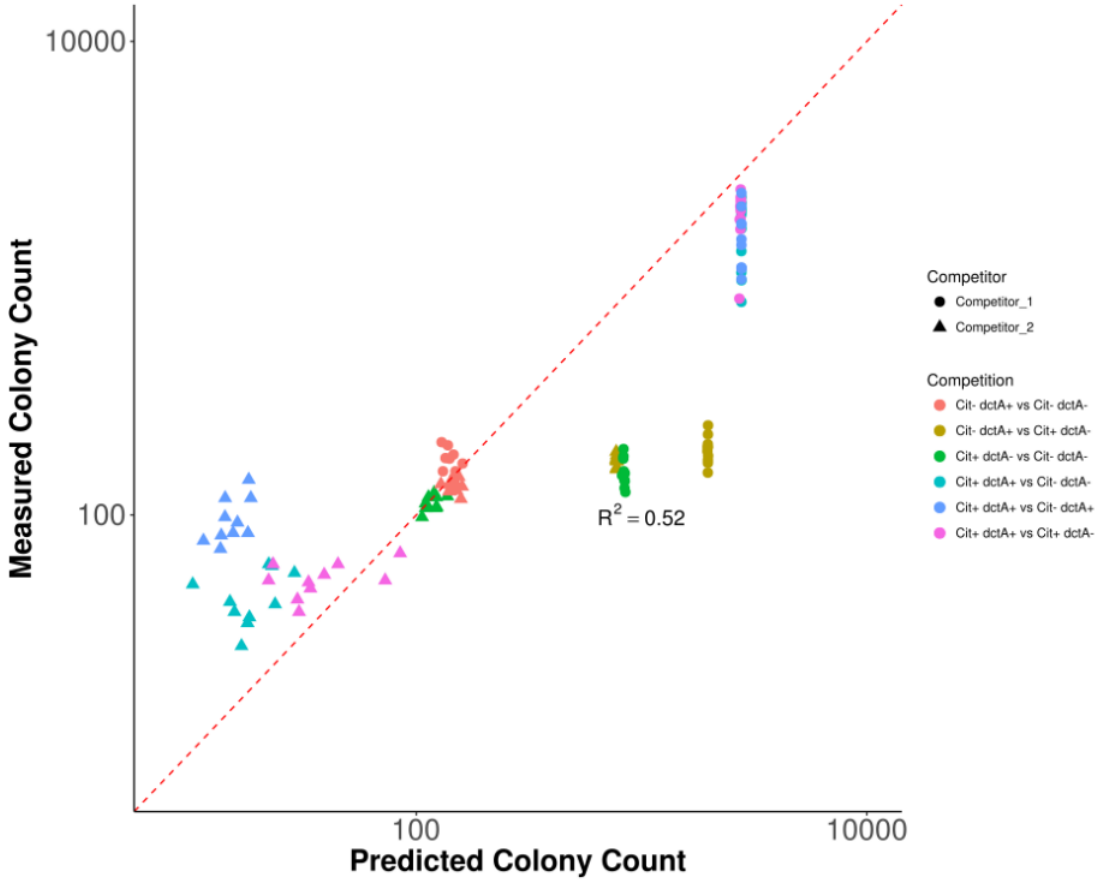
Supplementary Figure 1.1: Variation in the secretion profile of single mutants (see Figure 1.1C)., full set of mutants with environmental effect



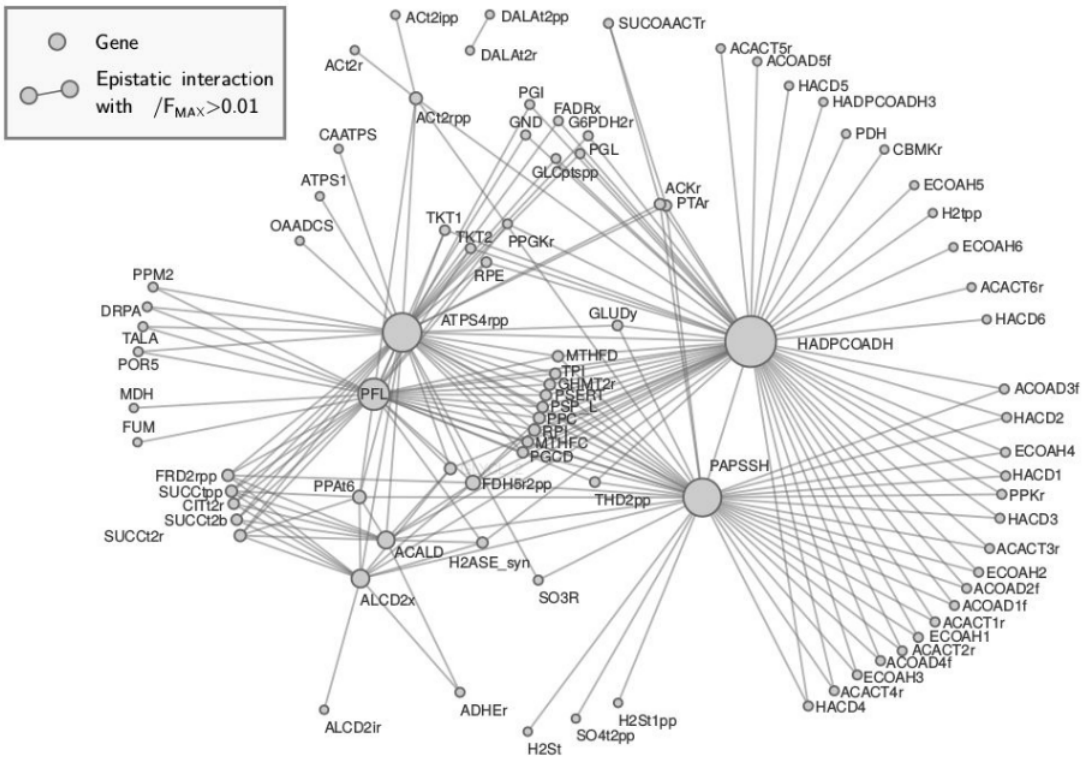
Supplementary Figure 1.2: Fitness and environmental effects are correlated. We plot the environmental effects of each single mutant (calculated as discussed in the methods) as a function of the fitness effect of that mutation. Both metrics are strongly correlated (Pearson's $/rho = 0.61, P < 10^{-6}$).



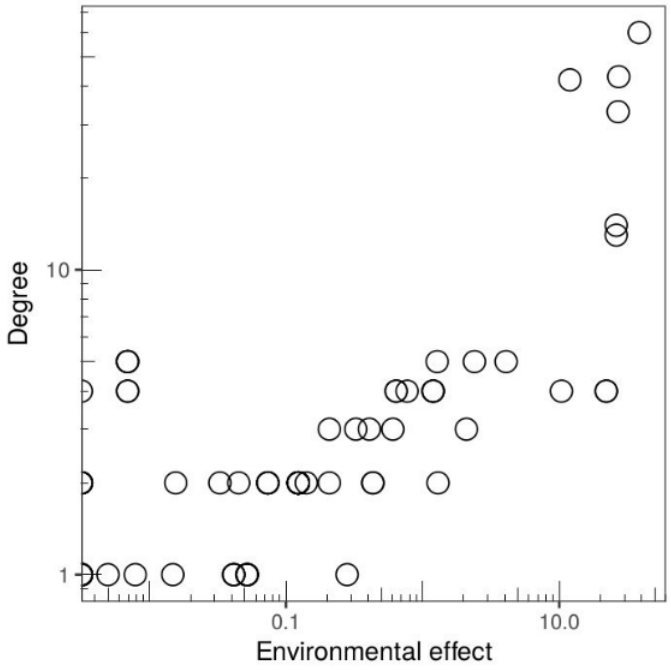
Supplementary Figure 1.3: Relation between non-commutativity ($/delta$) and intransitivity (I) in a hypothetical two mutation genotype space. We denote the fitness of mutant X in competition at low frequency with mutant Y by FX (Y) . As shown in the diagram,noncommutativity ($/delta$) is the sum of the intransitivities in both trajectories (I_{AB} and I_{BA}). However, to compute intransitivity, we need to perform a competition of the double mutant (AB) versus the original ancestor (O).



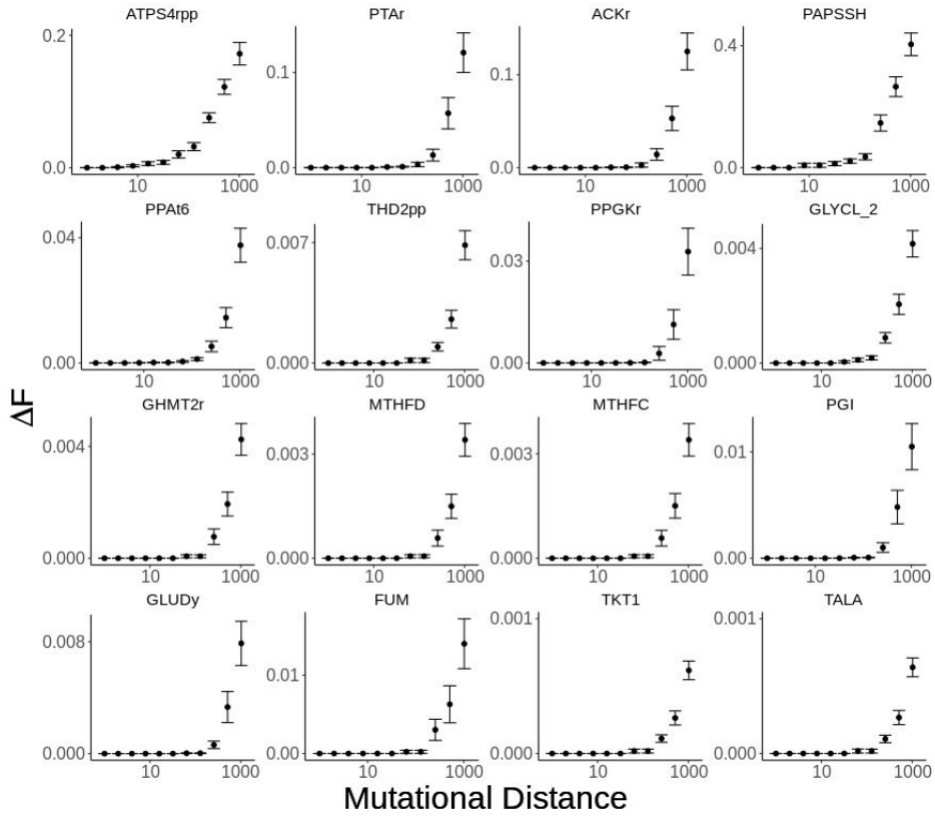
Supplementary Figure 1.4: Prediction of competition assay outcomes in the path to strong aerobic growth on citrate in E. coli was compared to the measured colony counts for each competition assay. All 120 competition assays were simulated using dynamic FBA (see Methods), and the experiments were performed as explained in the main text (Methods)



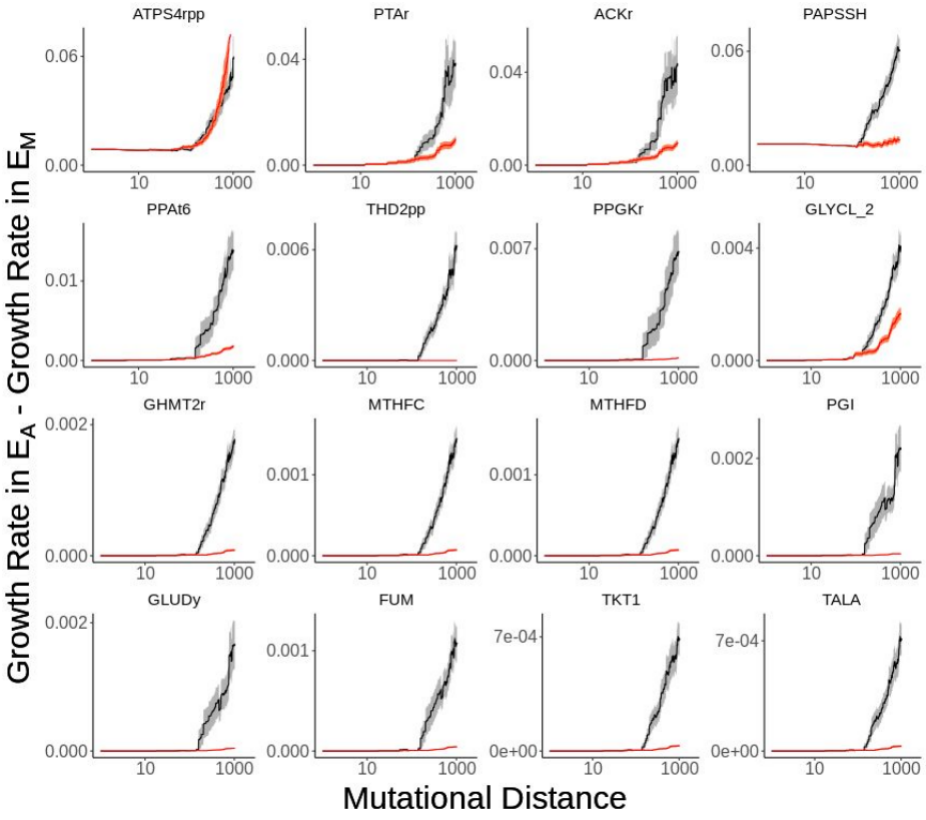
Supplementary Figure 1.5: Same network shown in Figure. 1.3B showing all reaction labels.



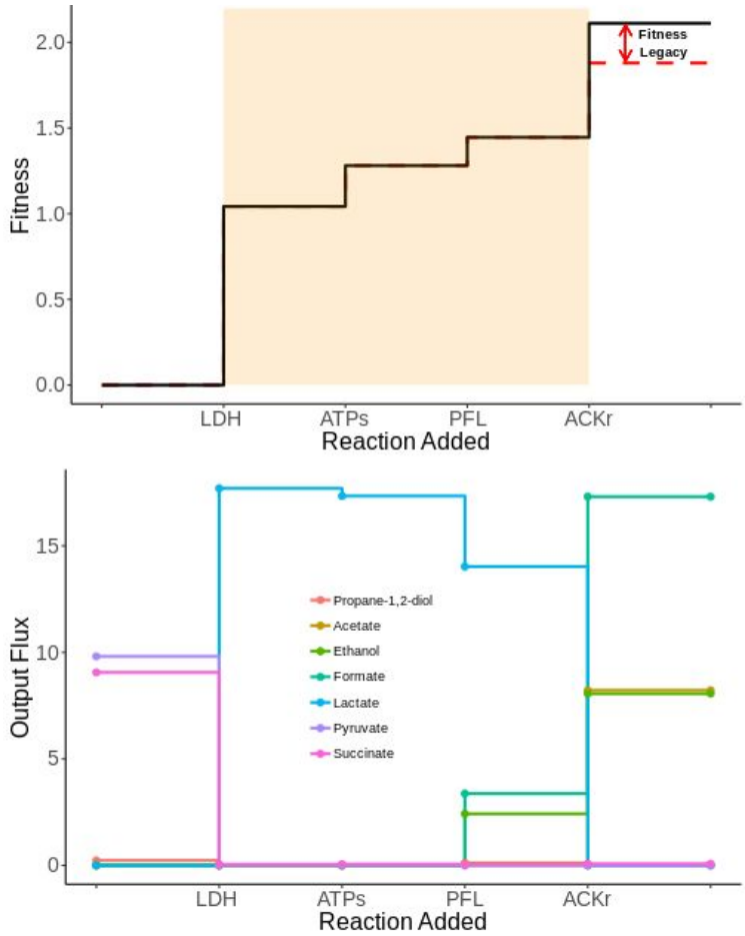
Supplementary Figure 1.6: Environmental effect and deformability are correlated in the local genotype space of E. coli. Here, the degree (i.e. number of interactions) in the network presented in Figure 1.3C (main text) is used as a proxy for deformability.



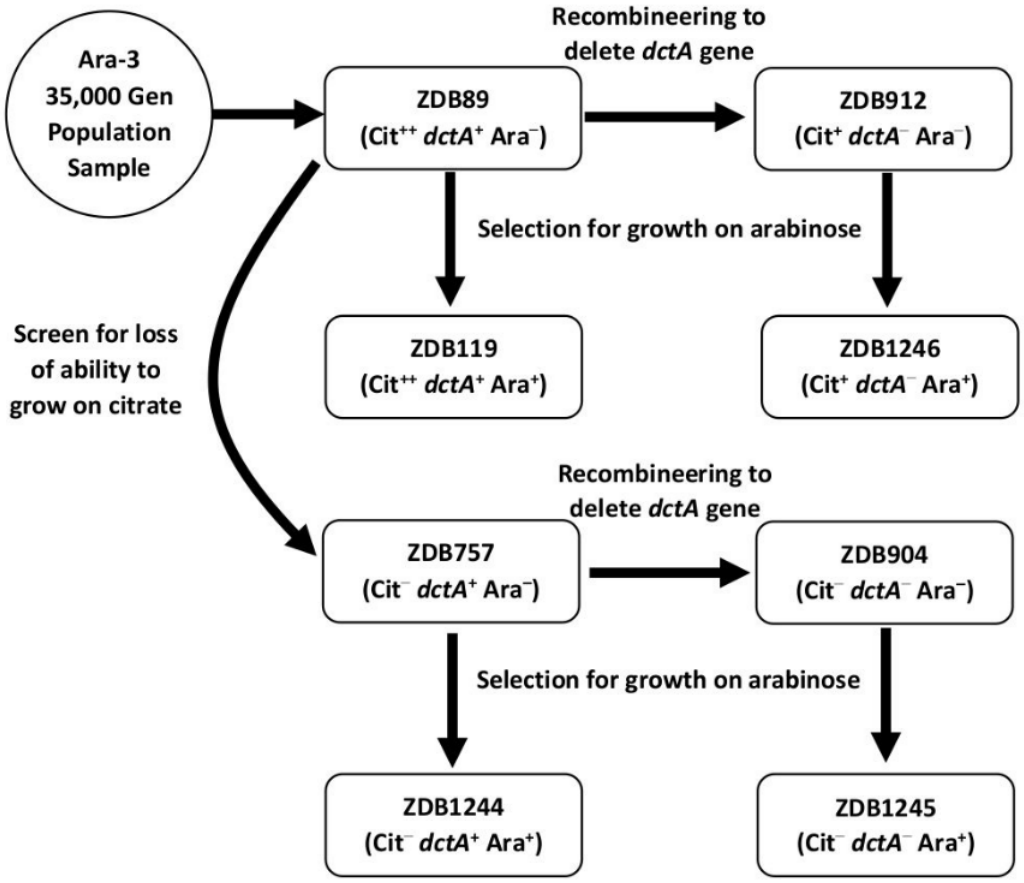
Supplementary Figure 1.7: Additional examples of long range deformation similar to Figure 1.4C (main text) using the 16 non-essential mutations M with largest environmental effect (shown as subpanel titles) other than ACKr (which is shown in Figure. 1.4C in the main text). We show average fitness differences (ΔF) in competition with A vs. in competition with M. ΔF always increases with mutational distance. Error bars represent SEM (N=100)



Supplementary Figure 1.8: Average difference (absolute value) in growth rate between environments E_M and E_A (in grams of dry cell weight xhr^{-1}) at varying genotype distances (gray line, shading represents SEM; N=100). Additional examples are shown, similar to Figure. 1.4D (main text), using the environments generated by the 16 non-essential mutations with largest environmental effect. Average difference in growth across environments alway increases with mutational distance. Gray shading is SEM (N=100).



Supplementary Figure 1.9: Breakdown of the example of long range effects in an adaptive trajectory given in Figure. 1.4. In the top panel we show the incremental fitness increase of each mutant as predicted by competition with it's immediate ancestor. The dotted red line shows the fitness predicted when excreted lactate is removed from the environment. Whilst lactate production only requires a single mutation, this environmental change does not affect the fitness of immediate descendants and instead leaves a 'legacy' (shaded region) that persists and requires multiple interacting mutations to be 'felt'. In the bottom panel we show the FBA predicted output flux of secondary metabolites when glucose is in excess (i.e uptake rate $= 10mmol \times gr^{-1} \times hr^{-1}$).



Supplementary Figure 1.10: Derivation of Ara-3 strains used in competition experiments

1.6.13 Supplementary Tables

Clone	Phenotype	*citT* Duplication	*dctA* Mutation	Ara marker
ZDB89	Cit++	+	+	–
ZDB119	Cit++	+	+	+
ZDB757	Cit–	–	+	–
ZDB904	Cit–	–	–	–
ZDB912	Cit+	+	–	–
ZDB1244	Cit–	–	+	+
ZDB1245	Cit–	–	–	+
ZDB1246	Cit+	+	–	+

Supplementary Table 1.1: Ara-3 strains used in competition experiments.

Parameters	Default Parameters (Fig 1,3 and 4)	Citrate Simulations (Fig2)
Default V_{max} (mmol × gr.$^{-1}$ × hr^{-1})	10	10
Default uptake Km (uM)	0.01	0.01
Death rate (%)	0	0.01
Time Step (hr)	0.1	0.1
Cycles	100	240
Space Width (cm)	0.02	0.02
Simulation layout	Single cell	Single cell

Supplementary Table 1.2: Parameters used in dFBA simulations.

1.7 Acknowledgements

We would like thank Gunter Wagner, Steve Stearns, David Post, and members of the A.S. laboratory for helpful discussions and feedback on the manuscript; Daniel Segrè and Ilija Dukovski for helpful advice during the early implementation of the software COMETS [Computation of Microbial Ecosystems in Time and Space (COMETS)]; Maia Rowles, Kiyana Weatherspoon, and Brooke Sommerfeld for assistance in the construction of Ara-3–derived strains; Richard Lenski for use of LTEE strains and materials; Jonathan Fritzemeier and Martin Lercher for assistance with model curation; and two anonymous reviewers for their critical comments, which significantly improved the manuscript. This work was partially funded by Young Investigator Award RGY0077/2016 from the Human Frontier Science Program (to A.S.) and John Templeton Foundation Foundational Questions in Evolutionary Biology Grant FQEB RFP-12- 13 (to Z.D.B.). The LTEE is supported in part by National Science Foundation Grant DEB-1019989.

Chapter 2

Dissimilarity–Overlap analysis of replicate enrichment communities

Jean C.C. Vila, Yang-Yu Liu and Alvaro Sanchez

Vila, J.C.C., Liu, YY. and Sanchez, A. Dissimilarity–Overlap analysis of replicate enrichment communities. ISME J 14, 2505–2513 (2020)

2.1 Abstract

The taxonomic composition of microbial communities can vary substantially across habitats and within the same habitat over time. Efforts to build quantitative and predictive models of microbial population dynamics are underway, but fundamental questions remain. How different are population dynamics in different environments? Do communities that share the same taxa also exhibit identical dynamics? In vitro communities can help establish baseline expectations that are critical towards resolving these questions in natural communities. Here, we applied a recently developed tool, Dissimilarity–Overlap Analysis (DOA), to a set of experimental in vitro communities that differed in nutrient composition. The Dissimilarity and Overlap of these communities are negatively correlated in replicate habitats, as one would expect if microbial population dynamics were on average strongly convergent (or "universal") across these replicate habitats. However, the existence of such a negative correlation does not necessarily imply that population dynamics are always universal in all communities. Even in replicate, identical habitats, two different communities may contain the same set of taxa at different abundances in equilibrium. The formation of alternative states in community assembly is strongly associated with the presence of specific taxa in the communities. Our results benchmark DOA, providing support for some of its core assumptions, and suggest that communities sharing the same taxa and external abiotic factors generally (but not necessarily) have a negative correlation between Dissimilarity and Overlap.

2.2 Introduction

Microorganisms grow and thrive in all habitats throughout the biosphere [Locey and Lennon, 2016, Hunter, 2016, Louca et al., 2016a, Blaser et al., 2016]. This includes the human body, where they form rich ecological communities made of large numbers of interacting species [Ley et al., 2006, Falony et al., 2016, Shafquat et al., 2014, Lloyd-Price et al., 2017]. The taxonomic composition of these communities can vary substantially between body sites, reflecting their different ecological, physical, and biochemical conditions [Human Microbiome Project Consortium, 2012]. Even for the same body site, community composition may vary widely between individuals, as well as within the same individual over time [David et al., 2014a,b]. In order to understand how microbiomes change longitudinally and over the lifespan of an organism, and to design effective strategies that enable

us to manipulate microbiomes towards desirable states, it is critical to develop predictive quantitative models of microbial population dynamics [Faust and Raes, 2012].

Models of dynamic ecosystems vary in their level of description, which is typically chosen to capture the specific phenomena under study. A detailed population dynamics model of microbial communities would have to include mechanistic microbial interactions (due to cross-feeding [Goldford et al., 2018, Machado et al., 2021], direct secretion of substances such as bacteriocins, antibiotics, or extracellular enzymes [Sanchez-Gorostiaga et al., 2019, Cornforth and Foster, 2013], or competition for the same nutrient), spatial structure of the particular habitat [Harcombe et al., 2014], and environment–microbiome or host–microbiome interactions. Building such detailed models can be daunting due to (i) a huge number of model parameters which need to be inferred from experimental data; and (ii) many environmental variables (such as the concentrations of bacteriocins and nutrients) which are hard to measure in real time.

To avoid those difficulties, an alternative modeling framework focuses on exploring the impact that any given microbial species has on the abundance of other microbial species [Bashan et al., 2016]. In this phenomenological modeling framework, one only needs to consider a simple population dynamics model written as a set of ordinary differential equations: $dX(t)/dt = f(X(t), \Theta)$. Here, f is a nonlinear function characterizing the population dynamics of the microbial community, $X(t) = (x_1(t), \ldots, x_i(t), \ldots, x_N(t))$ is an N-dimensional vector with $x_i(t)$ denoting the abundance of the i-th microbial species at time t, and Θ captures all the ecological parameters (such as intrinsic growth rates, intra- and inter-species interaction strengths, etc.). Note that those ecological parameters depend on environment- or host-independent factors, such as biochemical processes and microbial metabolic pathways; as well as environment- or host-specific ones, such as pH, temperature, nutrient intake, host genetic make-up, etc. Hence, environmental or host factors are not explicitly considered in this modeling framework but are absorbed in the ecological parameters [Bashan et al., 2016].

Generally, the ecological parameters estimated from a given habitat with certain characteristic environmental conditions do not necessarily map to other habitats with different environmental conditions. One can ask, however, whether those parameters (Θ) are strongly similar ("universal") for microbiomes that assemble in similar habitats. Addressing this fundamental question has important consequences for the applicability and predictive power of quantitative models of microbial community dynamics. If the interaction parameters were highly similar across habitats of a certain type, such as the guts of different human subjects, this will facilitate the development of generic microbiome-based

therapeutics. By contrast, if the ecological parameters and microbial dynamics are strongly host-specific, we must design truly personalized interventions, which need to consider not only the highly personalized microbial composition of each individual but also the unique dynamics of the underlying microbial ecosystem.

Directly addressing the above question would require one to infer all of the ecological parameters and fit the population dynamics $f(X(t), \Theta)$ from the microbiome data of each local community or host. Doing this for a large collection of communities is both logistically and computationally challenging. Recently, an indirect method called Dissimilarity–Overlap Analysis (DOA) was proposed [Bashan et al., 2016]. DOA relies on two mathematically independent measures between any two local communities: Overlap (O), which is defined as half of the sum of relative abundances of the shared species; and Dissimilarity (D), which is defined as the divergence between the renormalized abundance profiles of the shared species (Methods) [18]. DOA is based on the following two assumptions. First, the abundance profiles of the microbiome samples represent the steady states X^* of the microbial ecosystem and hence the fixed points of the underlying population dynamics that satisfy $f(X^*, \Theta) = 0$. Second, if any two local communities that have the same species collection also have the same abundance profile (steady state), i.e., $O = 1$ and $D = 0$, then the two communities should share universal microbial dynamics $f(X, \Theta)$ characterized by the same set of ecological parameters Θ. Mathematically, this means that if X^* satisfies both $f(X^*, \Theta^{(1)}) = 0$ and $f(X^*, \Theta^{(2)}) = 0$, given the large number of species and all the other levels of complexity in their interactions (encoded in the highly nonlinear function f), we conclude that generically $\Theta^{(1)} = \Theta^{(2)}$. In general, since D is mathematically not constrained by any value of $O > 0$, any constraints of D by O observed from real data deserve ecological interpretations. In particular, even if we do not have any steady state pair satisfying $O = 1$ and $D = 0$ (which is the typical case for host-associated microbial communities, such as the human gut microbiome, due to highly personalized microbial compositions), as long as steady state pairs with higher O tend to have lower D, i.e., there is a negative slope in the high-Overlap region of the Dissimilarity–Overlap Curve (DOC). This particular statistical constraint of D by O is consistent with the hypothesis of universal dynamics across all habitats in the sample, and it is a foundation of DOA [18]. It is also consistent with alternative hypotheses, such as communities assembling in environmental gradients, or situations when only a small fraction of the habitats have highly similar interaction parameters [Kalyuzhny and Shnerb, 2017]. The former is a particularly important scenario, and was recognized in the original study by Bashan et al. In many instances enough is known about the habitats to exclude from

the analysis factors that can lead to environmental heterogeneity [Bashan et al., 2016].

A negative slope in the high-Overlap region of the DOC has been found in the gut microbiome samples collected from different healthy individuals [Bashan et al., 2016]. Yet, the complete set of selective pressures experienced by microorganisms in the same habitat (e.g., the same body site of different individuals), and their variation across a host group cannot be known exactly. Hence, one cannot account for all the potential factors that may conceivably influence the microbial communities assembled in the same habitat, and so cannot provide an entirely conclusive answer regarding the universality of the underlying microbial dynamics. In other words, we cannot unambiguously attribute the negative slope of the DOC to universal dynamics and completely rule out the alternative explanation of environment or host factors. A more definitive demonstration would require a comparison between experimental communities assembled in well-controlled replicate habitats to those assembled in nonidentical habitats.

Benchmarking DOA against well-controlled in vitro communities, ideally assembled in multiple replicates of habitats that are either identical to each other, or different from each other in well-understood ways, would be necessary to understand the limitations and potential of DOA for its application to natural communities [Verbruggen et al., 2018]. To address this need, here we perform DOA for a large set of in vitro communities that meet these requirements: close to 300 independent enrichment communities assembled in multiple, replicate synthetic habitats on three different limiting nutrient conditions, and assembled to equilibrium under periodic serial dilution cycles [Goldford et al., 2018].

2.3 Results

2.3.1 Data Set

In a recent study [Goldford et al., 2018], we reported the assembly of a total of 276 enrichment communities in three different synthetic environments: M9 minimal medium with either glucose, citrate, or leucine as the only carbon source. These enrichment communities were assembled from twelve different environmental sources (including various soil samples and plant matter collected near Yale University in New Haven, CT). Seven or eight biological replicates of each inoculum were propagated in each of the three nutrient environments, under serial dilution with transfers every 48h with a dilution factor of $125\times$. A diagrammatic summary of the experiment is presented in Figure. 2.1A. As reported in

[13], communities were initially very diverse ($N = 110 = 1290$ unique Exact Sequence Variants, or ESVs). They typically converged to an approximately stable composition (containing $N = 2 - 22$ ESVs) after 50-60 generations, suggesting that communities were close to a steady state. Metabolic cross-feeding was found to be widespread and critical for the coexistence of multiple species on a single limiting resource [Goldford et al., 2018].

2.3.2 Communities assembled in identical environments exhibit a negative correlation between Dissimilarity and Overlap

We first addressed the question of whether communities assembled in identical environments do indeed give rise to a negatively sloped DOC at high Overlap. To test this prediction, we took all pairs of communities that had been assembled in the same nutrient-limited habitats, by pooling together every possible pair of glucose communities, as well as every pair of citrate (and of leucine) assembled communities. We then measured the Dissimilarity and Overlap for each pair. Applying the same type of statistical analysis used in the original study by Bashan et al. (Robust LOWESS regression; see "Methods" for details [Bashan et al., 2016]), we find that the DOC (but not the controls; Supplementary Figure. 2.1) does indeed exhibit a negative slope at high values of Overlap (Figure. 2.1B), and this is also confirmed by a standard linear regression ($D = D_0 + m * O; m = -0.56, p < 0.002$ by Bootstrapping; Methods) applied to the points with higher than median O (Figure. 2.1B, inset; Methods). The same is true when we analyze each of the three nutrient environments separately (Figure. 2.1C), and it also holds when we separate those communities assembled from either the same or different inoculum (Figure. 2.1D, E). In contrast, a statistically significant negative correlation between Dissimilarity and Overlap is not observed for community pairs that are assembled in different environments (e.g., one in citrate medium, one in glucose medium) (Supplementary Figure. 2).

An expectation of DOA is that communities that contain the same taxa in identical habitats should have them at highly similar equilibrium abundances, as the underlying population dynamics would be strongly similar. By contrast, communities assembled in different nutrient habitats are not expected to have similar species abundances even when they happen to share a large proportion of common species (high Overlap), as we do not necessarily expect their dynamical equations to be similar. Consistent with the first hypothesis, we find that, in identical nutrient conditions, the majority of our high-Overlap ($O > 0.98$) communities have low Dissimilarity (Figure. 2.1F, see Supplementary Figure. 2.1 for comparison to null) ($Mean = 0.188$, $Median = 0.132$, $IQR = 0.23$). To test

the second hypothesis, we considered "mixed" pairs of communities, where each community in the pair was assembled in a different environment. For instance, we find that glucose–citrate pairs (which exhibit no correlation between D and O (Supplemntary Figure 2.2 and 2.3)) have a similar Overlap distribution to glucose–glucose and citrate–citrate pairs (Figure . 2.1G). Yet, the distribution of Dissimilarities for these high-Overlap pairs ($O > 0.98$) is shifted up compared to glucose–glucose pairs ($t = -12.79$, $p < 0.002$ by Bootstrapping; Methods) and citrate–citrate pairs ($t = -11.965$, $p < 0.002$, by Bootstrapping; Methods) (Figure. 2.1H). This shift persists even if we only consider communities assembled from the same inoculum (Supplementary Figure. 2.4) and is robust to the Overlap threshold chosen (Supplementary Figure. 2.5)).

This last finding is consistent with the idea that population dynamics and equilibria are strongly convergent when the environments are identical, but not necessarily when the environments are different. This lends support to the null assumption that species interactions with the environment and with each other are different in different environments, but strongly convergent in identical environments. Our results support the prediction that identical environments will generate a negative statistical correlation between D and O, whereas different environments will not.

2.3.3 Specific taxa can be strongly associated with high Dissimilarity in replicate habitats

As can be visually appreciated in Figure. 2.1, when environments are identical and Overlap is high the Dissimilarity in our experimental communities is generally small. However there are numerous deviations from this rule, and we find multiple community pairs with high Overlap that still show high levels of Dissimilarity. Considering only communities with an Overlap > 0.98, we find that 12% of glucose pairs, 9% of citrate pairs, and 17% of leucine pairs have $Dissimilarity > \sqrt{log(2)/2}$ (which is half of the maximum possible Dissimilarity, calculated through the root Jensen–Shannon divergence) (Figure. 2.1F). Similar values are also obtained if we only consider communities assembled from the same inoculum, (14% of glucose pairs, 9% citrate pairs, 8% of leucine pairs) (Supplementary Figure. 2.4A). In sum, we find that communities can be dominated by the same set of ESVs in identical habitats, yet these ESVs, may exist at very different abundances.

We set out to investigate whether communities that deviate from the average trend captured by the DOC could be associated with the presence of specific taxa. To that end, we first selected pairs of glucose communities with high Overlap (higher than the median

of 0.98) that were assembled from the same initial pool of species (for a similar analysis in citrate or leucine, see Supplementary Table 2.1 and 2.2). We then tested whether any of the ten most commonly observed ESVs (corresponding to ESVs found in at least 18 of the 92 glucose communities), were statistically associated with high Dissimilarity. Four of these ESVs had higher Dissimilarity than expected by chance ($p < 0.05$, by Bootstrapping, Methods). Of these, an ESV of the genus Citrobacter had the largest effect (Table 2.1).

To further investigate this point, we split all pairs of communities by whether both contain this Citrobacter ESV (group I), only one contains it (group II), or none does (group III) (Figure. 2.2a). The mean Dissimilarity is higher for group I than group II (0.427 vs 0.175, $p < 0.003$, by Bootstrapping; Methods), and for group II than group III (0.175 vs 0.088, $p = 0.06$ by Bootstrapping; Methods) Figure. 2.2B. As shown in Supplementary Figure 2.6, our results are robust to our choice of a "High-Overlap" threshold of 0.98. If Citrobacter ESV was indeed associated with alternative dynamical states (either through multistability, or through their contribution to alternative dynamical equations when they are part of the community), we would also expect the DOC to flatten for glucose–glucose communities that contain it, relative to those that do not. That is indeed the case, as shown in Figure. 2.2C, 2.2D, see Supplementary Figure 2.7 for controls).

Similar results were found for citrate communities and leucine communities, where we found a Raoultella ESV and a Pseudomonas ESV associated with higher Dissimilarity respectively (Supplementary Tables 2.1 and 2.2, Supplementary Figures. 2.8 and 2.9). Our results thus reveal that the presence of a single ESV in a community may be strongly associated with alternative states in community assembly, even amongst communities that are assembled in the same environment from the same regional pool and contain highly overlapping sets of taxa.

2.4 Discussion

The first part of this paper tests two fundamental predictions of DOA. First, using publicly available data from a recent experiment we show that community assembly in identical environments does lead to a negative correlation between Dissimilarity and Overlap and a negatively sloped DOC. Second, we show that, as expected, communities assembled in identical environments that also contain highly overlapping sets of taxa have them at strongly convergent abundances. This is consistent with what one would expect if their dynamics were describable by the same equations. Likewise, we would not expect iden-

tical dynamical equations in different nutrient habitats. Consistent with this expectation, we found that communities assembled in different nutrient habitats do not have convergent species abundances, even when they happen to share a high number of taxa. This is reflected in the data by a shift in the Dissimilarity to higher values for those communities with very large Overlap.

Our results also indicate the existence of specific taxa that are associated with high Dissimilarity in replicate habitats. In glucose communities, we find an ESV of the genus Citrobacter, which not only predicts high Dissimilarity between community pairs when it is present in at least one of the communities, but it also flattens the DOC. Intriguingly, these results are in line with one of the main findings of the Bashan study, which was that microbiomes disrupted by C. difficile infection did not exhibit a negative correlation between Dissimilarity and Overlap, and Dissimilarity remained constant even as Overlap increased [Bashan et al., 2016]. The negative slope was recovered after a fecal microbiota transplantation, which restored a healthy microbiota and cured the disease. Although the reasons for this finding may be very different from the similar result found in our communities, our results indicate that a tight association between specific taxa and the flattening of the DOC may be seen even in the absence of an immune system or a complex host.

The publicly available data of Goldford et al. [Goldford et al., 2018] includes the measurement of the complete population dynamics for one glucose pair with high-Overlap and high Dissimilarity (highlighted in Figure. 2.2A). Both communities in the pair contain the Citrobacter ESV. For this pair, we find that the population dynamics are initially strongly convergent between the two communities, but bifurcate after 20 generations and subsequently diverge into alternative compositions (Figure. 2.2E, 2.2F), suggesting the potential presence of true multistability (i.e., multiple stable steady states are associated with the same set of species).

Of course, other possibilities exist. For example, the population dynamics may just not be identical even in replicate habitats due to the violation of the first assumption of DOA (that the communities are at equilibrium steady state). Communities may either have not reached equilibrium after 12 transfers, or they may be at a non-equilibrium steady state, undergoing oscillatory or cyclical dynamics. Neutral population dynamics and stochastic population dynamics can also lead to variation in community composition that may lead to increased Dissimilarity [Kalyuzhny and Shnerb, 2017, Hubbell, 2011]. These dynamics would be observed if communities are moving on a shallow attractor where selection is weak, or if non-accounted environmental fluctuations shift the position of the fixed points in the community. The high overlap of the two communities at the ESV level might also

reflect their differences at the strain level due to rapid evolution. Citrobacter may also have more sensitive interactions with the rarer members of the community. It is important to consider that environments do not just passively select for species and determine their interactions, but rather they are dynamically shaped by the taxa growing in them [Bajić et al., 2018, Callahan et al., 2014, Laland et al., 2015, Lewontin, 1983]. Therefore, it is possible that although the supplied nutrients are the same in two communities, the environment experienced by the members of our communities is actually different, through the different effects that species have on it [22].

One limitation of our study is that the communities investigated are species-poor, and many of our community pairs will only share a few species in common. This can be potentially problematic because it may affect the Dissimilarity measurement [Bashan et al., 2016]. Despite this caveat, the overall good agreement between our findings and the expectations of DOA suggests that species richness is not necessarily an impediment for the application of DOA to taxonomically poor natural communities. Further research would be needed to establish the precise conditions under which this would be true. More generally, others have argued that communities assembled along an environmental gradient may also give rise to negative correlations between Dissimilarity and Overlap [Kalyuzhny and Shnerb, 2017]. It would be important to test this prediction experimentally (for instance by establishing mixed nutrient habitats with varying concentrations of glucose and citrate in between the two "pure" habitats studied by Goldford et al.) but this falls beyond the scope of this paper.

It is thus important to remark that although sets of communities assembled in identical habitats present a negatively sloped DOC, the reverse statement is not necessarily true: the presence of a negatively sloped DOC does not necessarily mean that the habitats are identical to each other. In fact, grouping together all communities in the experiment, including those assembled in identical and different environments and projecting them all into the same Dissimilarity–Overlap plot, we find a negative correlation between D and O and a negatively sloped DOC (Fig. S10). The reason is that the strong effect of identical habitats overpowers the effect of nonidentical ones. In our case, we know the environmental factor that was critically different among these habitats (the single limiting nutrient), but this is not something that is trivial to identify in any given natural habitat, even if we remove any known factors of variation across habitats.

Notwithstanding these important caveats, our results confirm in a controlled set of experiments that microbial dynamics in replicate habitats are strongly convergent on average and lead to a negatively sloped DOC. In recent years, negatively sloped DOC has

been detected for the microbial communities assembled in some (but not all) human body sites [Bashan et al., 2016], as well as for mycorrhizal fungal communities [Verbruggen et al., 2018]. In order to correctly interpret these results, it is critical to benchmark the technique not only against simulations, but also against experimental communities whose assembly is well understood. Our results have provided a first empirical benchmark using well-controlled communities. We hope that these findings will contribute to grounding the expectations and intuitions behind DOA, and contribute to its application to microbial communities assembled in natural environments. More generally we hope that these results will encourage other researchers to benchmark novel statistical methods in microbial ecology using well-controlled in vitro communities.

2.5 Figures

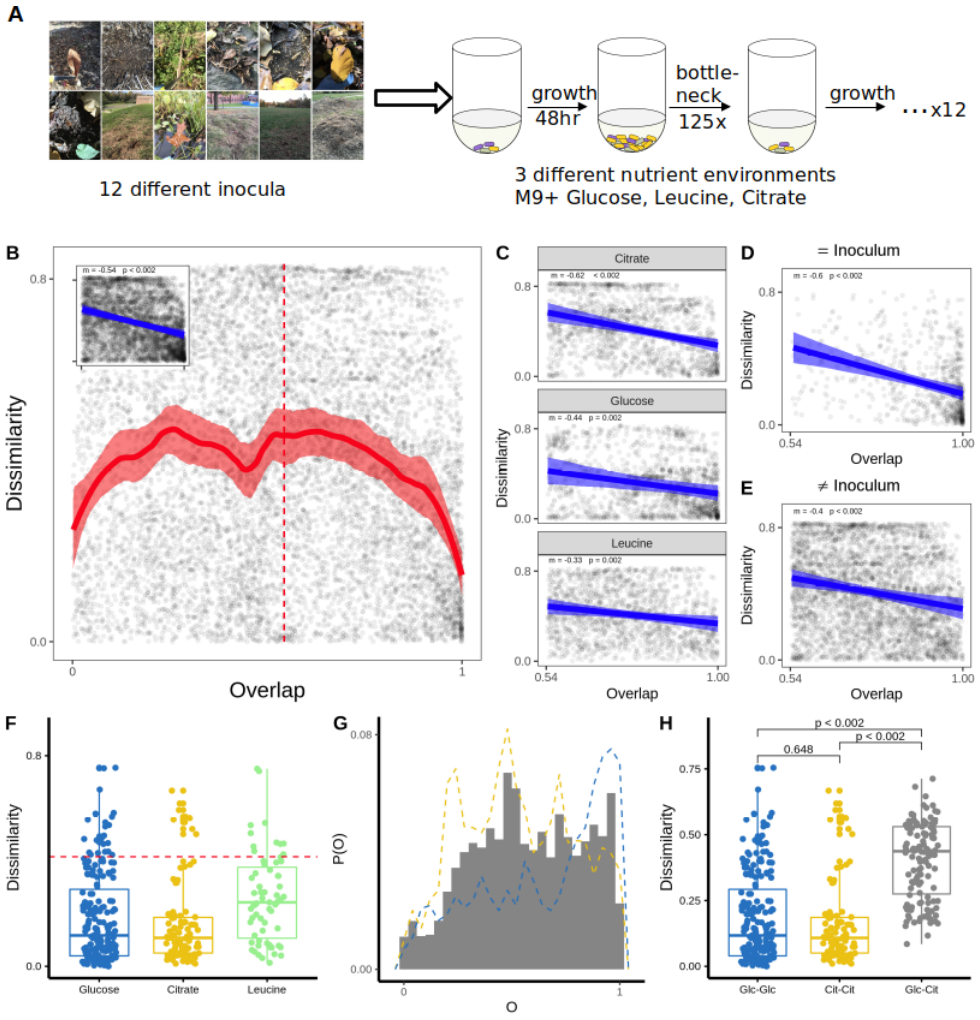


Figure 2.1: Communities assembled in the same environment show a negative correlation between Dissimilarity and Overlap

Figure 2.1: (A) Schematic description of the experiments in ref. [13]. (B) DOC of all microbial community pairs that have been assembled in the same environment (n=276 Samples). Shaded regions indicate the 95% confidence interval (Methods). The vertical dotted red line represents the median Overlap (0.543). The inset shows a linear regression for communities above the median Overlap. We repeat this regression over the same region, subsetting the data to consider (C) each nutrient environment separately; (D) subsets of pairs that have been assembled from the same inoculum (E); subsets of the pairs that have been assembled different inoculum. For each regression, we report m (slope of the linear regression) and a p value calculated as the fraction of bootstrap realization in which this slope is negative (see Methods). (F) Distributions for community pairs assembled in the same environment (both in glucose, both in citrate, or both in leucine) with high Overlap ($O > 0.98$). The dotted red line is at half the maximum possible dissimilarity $\sqrt{log(2)/2}$. (G) Histogram showing distributions of Overlaps for community pairs where one has been assembled on glucose and the other on citrate. The dotted lines give the frequency polygon for glucose–glucose community pairs and citrate–citrate community pairs (blue and yellow, respectively). We use the same binwidth (0.04) for both histograms and frequency polygons so the two are comparable. h Glucose–citrate communities with high-Overlap ($O > 0.98$) have significantly higher mean dissimilarity than glucose–glucose communities or citrate–citrate communities in the same Overlap range ($O > 0.98$). Displayed p values are computed by bootstrapping (see Methods)

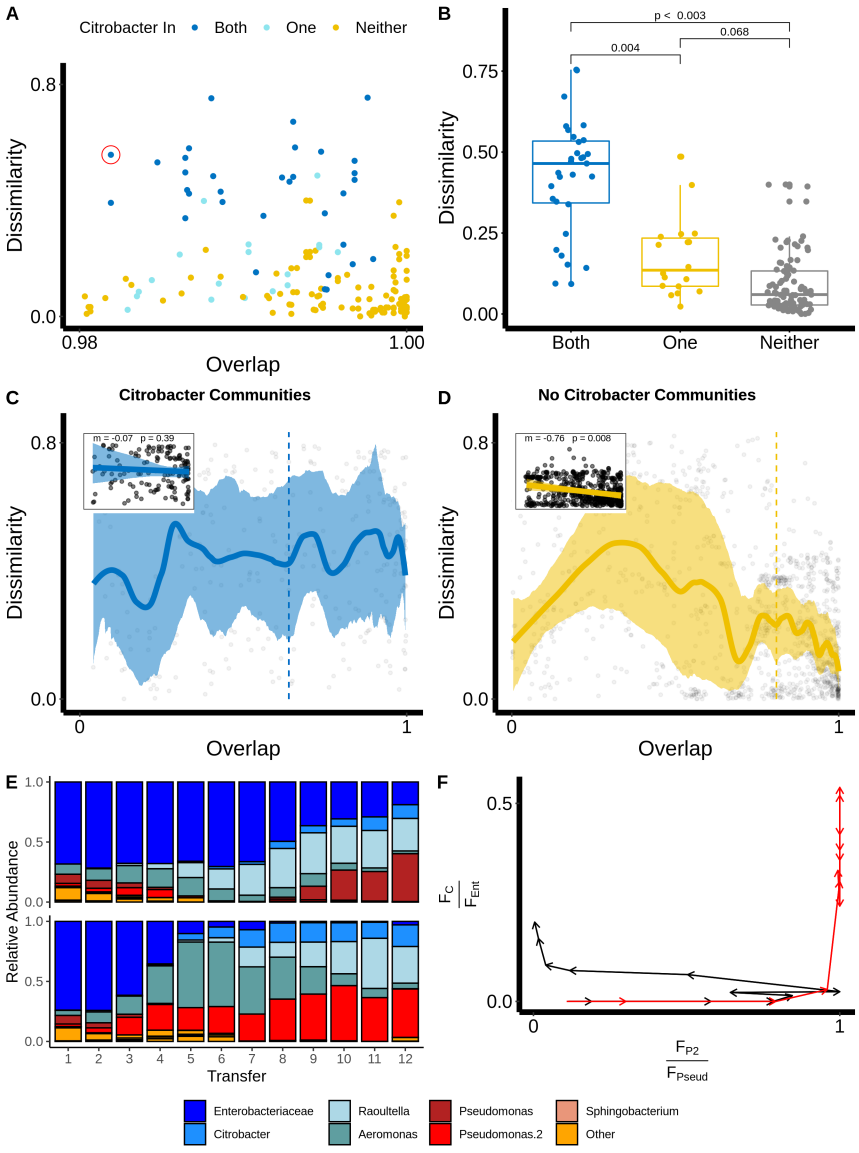


Figure 2.2: A Citrobacter ESV is associated with dynamical dissimilarity in communities assembled in replicate environments

Figure 2.2: (A) Dissimilarity and Overlap of microbial community pairs assembled from the same regional pool on M9+glucose with above-median Overlap ($O > 0.98$). (B) Dissimilarity of the same set of communities. For (A) and (B) we label communities by whether Citrobacter ESV is found in both communities (dark blue), only in one community (light blue) or in neither community (yellow). (C) DOC of all pairs of microbial communities assembled on glucose that contain Citrobacter ESV ($n = 25$). (D) DOC of all pairs of microbial communities assembled on glucose that do not contain Citrobacter ($n = 67$). (E) Population dynamics for one pair of glucose communities with high Overlap ($O = 0.98$) and high Dissimilarity ($D = 0.56$) (highlighted in (a) with the red circle). Structure of the two communities at the ESVl evel at every transfer. (F) Phase portraits illustrate the dynamics of the most abundant Enterobacteriaceae and Pseudomonadaceae ESV within those two communities. That black line corresponds to the top community in (E) and the red line corresponds to the bottom community in (E). FC and FP2 represent the fraction of the Citrobacter and Psuedomonas.2 ESVs in the population, whereas FEnt and FPseud represent the fractions of the Enterobacteriaceae and Psuedomonoadacea families. Dynamics are highly convergent until the third transfer, after which the communities diverge to alternative states.

2.6 Tables

ESV	N	Difference in Mean	T−statistic	P.value
Citrobacter	25	0.319974325	9.6322380	<0.002
Yersinia	19	0.289481747	7.3186745	0.00209205020920502
Enterobacteriaceae	38	0.248081124	6.6420427	0.00598802395209581
Pseudomonas	57	0.175009032	6.5060298	0.00399201596806387
Pseudomonas.2	41	0.131673553	3.7136131	0.093812375249501
Raoultella	45	0.103170657	3.2164266	0.0818363273453094
Pseudomonas.3	18	−0.002248539	−0.0671654	0.552238805970149
Pseudomonas.1	24	−0.031053732	−0.9691620	0.63872255489022
Pseudomonas.4	27	−0.044317786	−1.6107989	0.756487025948104
Klebsiella	65	−0.281428827	−7.4501771	0.98997995991984

Table 2.1: For each of the ten most commonly observed ESVs on Glucose we performed one-tailed t-tests to determine whether pairs of highly overlapping communities ($O > 0.98$) both containing that ESV had higher Dissimilarity than pairs where at least one community did not contain the ESV. In this table, we report the identity of the ESV, the number of communities in which that ESV is found (N) the difference in mean Dissimilarity, the T-statistic and the p value (obtained by bootstrapping, see Methods). See Supplementary Tables 2.1 and 2.2 for a similar analysis applied to leucine and citrate communities.

2.7 Material and Methods

2.7.1 Community Assembly Experiment

We analyzed publicly available data from a recent set of in vitro microcosms experiments [Goldford et al., 2018]. Briefly, diverse microbial communities were isolated from natural ecosystems and used as the inoculum for a batch culture containing M9 + one of three carbon sources (Glucose, Citrate, Leucine). Cultures were passaged every 48h with a dilution factor of $125\times$ and after each transfer, a sample was taken and stored for 16S community sequencing (Figure. 2.1A). The experiment was conducted for 12 transfers by which point communities appear to have reached a stable population equilibrium. In

total, 276 communities were allowed to self assemble (7–8 replicates per Inoculum and 12 Inoculum per carbon source). Community structure was determined at the end of the 12th growth period for all communities using 16s rRNA amplicon sequencing. A subset of communities was sequenced at each transfer allowing community structure to be tracked through time.

2.7.2 Calculating Dissimilarity and Overlap for community pairs

To account for differences in the sequencing depth of different communities we first normalize all communities so that each community is represented by the same number of sequences. Briefly, for each community we create a sample community of N reads sampled from the original read pool without replacement. Here N=4397 was used as this is the minimum number of reads for all communities analyzed. We then calculated the relative abundance of each ESV in each community. For each pair of communities, we follow Bashan et al. [Bashan et al., 2016] and calculate Overlap and Dissimilarity between relative abundance vectors x and y. The Overlap is given by

$$O\left(x,y\right)=\sum_{i\in S}\left(\frac{x_i+y_i}{2}\right) \tag{2.1}$$

where S is the set of ESVs found in both communities. For the calculation of Dissimilarity only the shared ESVs are considered, and the relative abundance of shared taxa is renormalized to add up to 1. The Dissimilarity between renormalized vectors X and Y can be calculated as root Jensen–Shannon divergence

$$(x,y)=\sqrt{\frac{1}{2}\left(\sum_i X_i log\frac{X_i}{M_i}+\sum_i Y_i\ log\frac{Y_i}{M_i}\right)}, \tag{2.2}$$

where $M_i=\frac{X_i+Y_i}{2}$

We calculated the Overlap and Dissimilarity for every pair of our 276 communities at transfer 12 (a total of 37,950 pairwise comparisons). From this dataset, any pairwise comparison in which the communities shared fewer than two taxa in common is excluded. These were removed as for these pairs $D(x,y)$ is always 0. In total this gave us a dataset consisting of 23592 pairwise comparisons.

2.7.3 Fitting DOCs

In Bashan et al. [Bashan et al., 2016] estimated using the Robust LOWESS a nonparametric scatterplot smoothing method. When analyzing all pairs of communities (Figure. 2.1B, Supplementary Figure 2.2A, Figure 2.2C,D), we implement the same method using the LOESS function from R default stats package with the following parameters span=0.2, family="symmetric", iterations=5). To compare slopes across different subsets of pairs (Insets Figure. 2.1B, Figure. 2.1C-E, Supplemntary Figure. 2B-D, Insets Figures. 2.2C, 2.2D) we use a simple OLS regression on data points with above-median Overlap as was also done by Bashan et al. [Bashan et al., 2016] and later by Kalyuzhny and Shnerb [Kalyuzhny and Shnerb, 2017].

2.7.4 Estimating confidence intervals and P value for DOC

We implemented the same bootstrapping algorithm used by Bashan et al. [Bashan et al., 2016]. We repeat this bootstrap algorithm 500 times and repeated all our analysis on every bootstrap realization. Confidence intervals in Figure. 2.1 and Figure. 2.2 represent 95% percentiles of the curves fitted to the bootstrapped data. The reported p values for the regression slopes (m) represent the fraction of bootstrap realization for which the OLS slope is positive (main text and Figure. 2.1, Figure. 2.2, Supplementary Figure. 2.3).

2.7.5 Bootstrapped Welch-tests

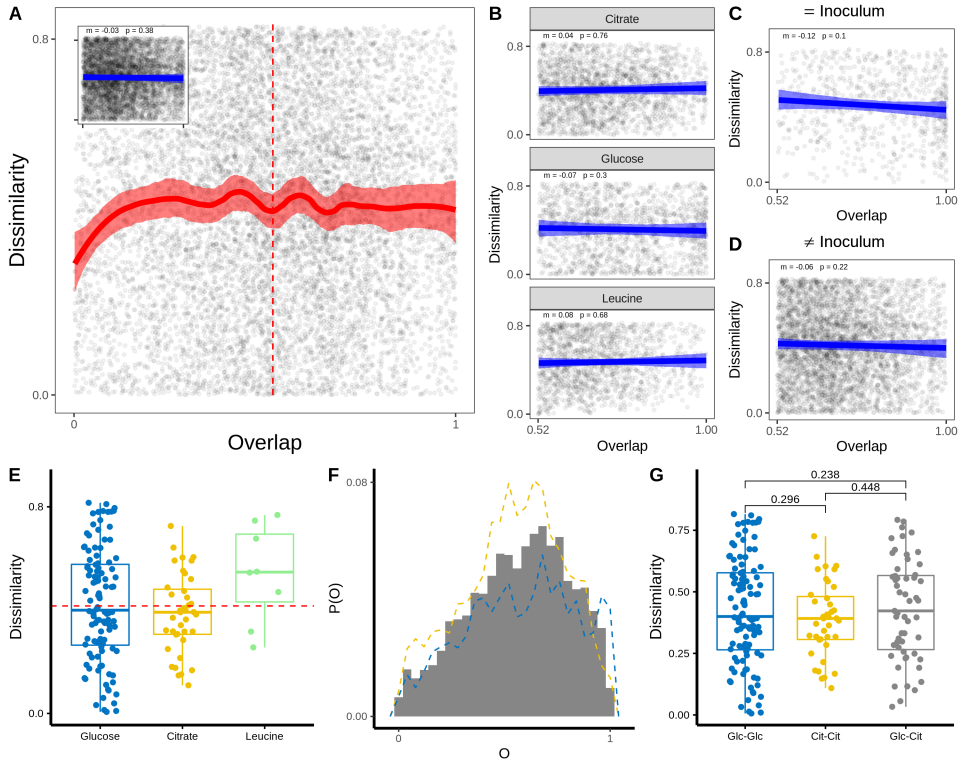
To account for the nonindependence of groups of sample pairs, all t-tests were performed on every bootstrap realization. The reported p values represent the fraction of bootstrap realizations in which the t-statistic has a different sign to the one calculated from the original dataset. When calculating this fraction we excluded bootstrap realizations in which some groups of sample pairs were unrepresented and so no t-statistic could be obtained.

2.7.6 Randomized data

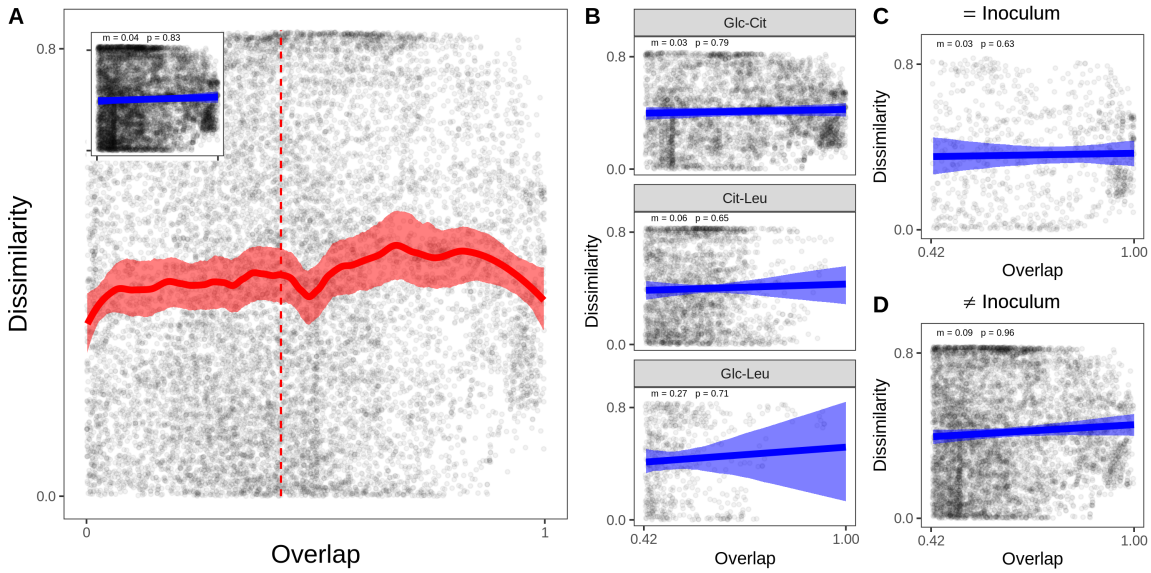
Each time a DOC is shown, we repeat the analysis on a randomized dataset in which species assemblage and abundance distribution are kept but the abundance of each taxon is randomized following Bashan et al. [Bashan et al., 2016]. The randomized results are shown in Supplementary Figure. 2.1, 2.3 and 2.7.

2.8 Supplementary Material

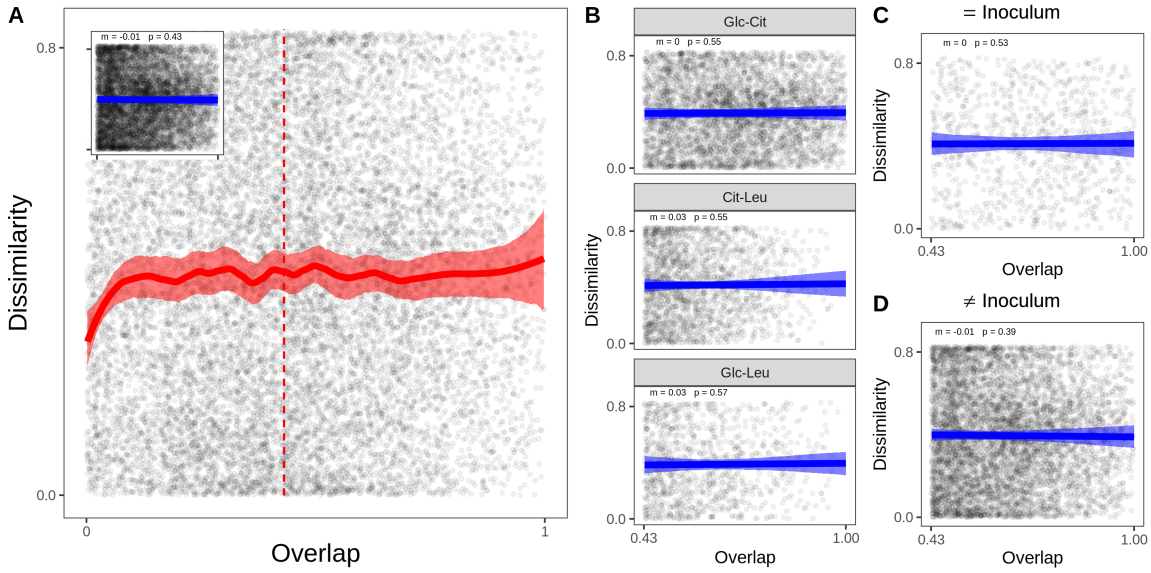
2.8.1 Supplementary Figures



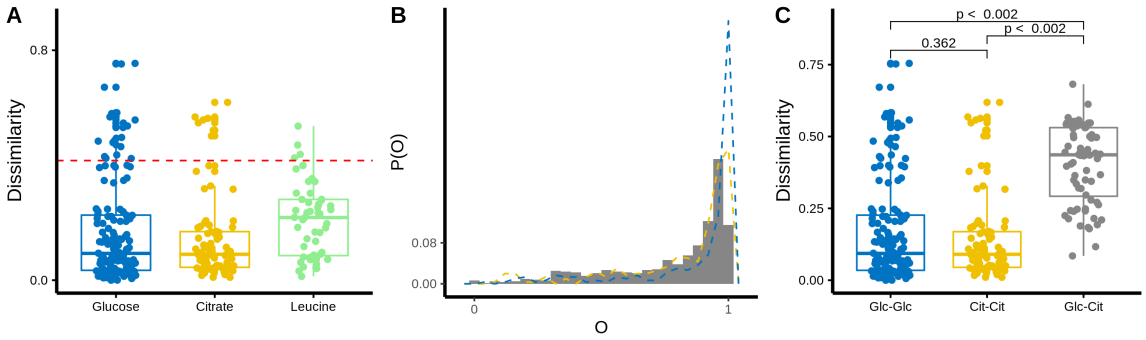
Supplementary Figure 2.1: A statistically significant negative correlation between Dissimilarity and Overlap is not observed for randomized samples (see Methods). Panel A is a randomized version of Fig. 2.1B. Panel B is a randomized version of Fig. 2.1C. Panel D is a randomized version of panel Fig. 2.1E. Panel E is a randomized version of Fig. 2.1F, Panel F is a randomized version of Fig. 2.1G. Panel G is a randomized version of Fig. 2.1H.



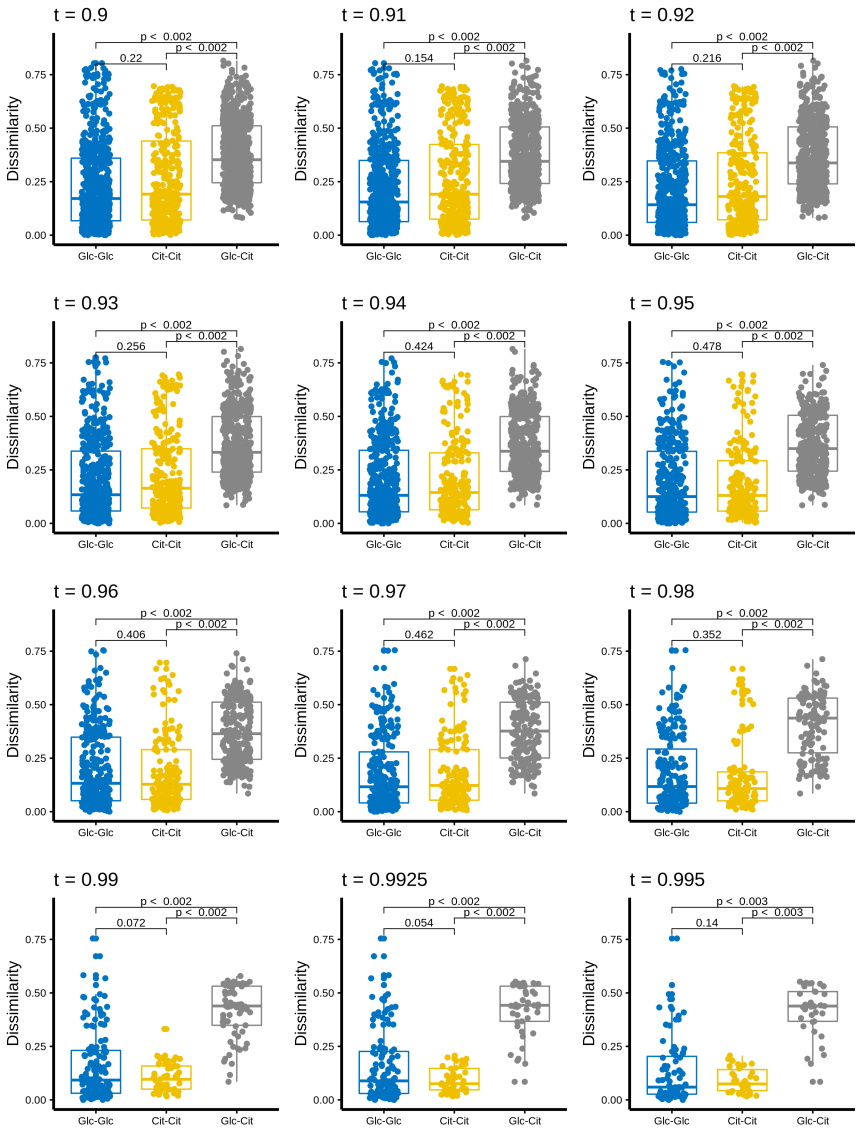
Supplementary Figure 2.2: A statistically significant negative correlation between Dissimilarity and Overlap is not observed for community pairs that are assembled in different environments (e.g. one in citrate medium, one in glucose medium). This holds true regardless of whether we consider: (A) all pairs in different environments; (B) Each contrasting environment separately. (C) Community pairs assembled from the same inoculum; or (D) Community pairs assembled from different inocula



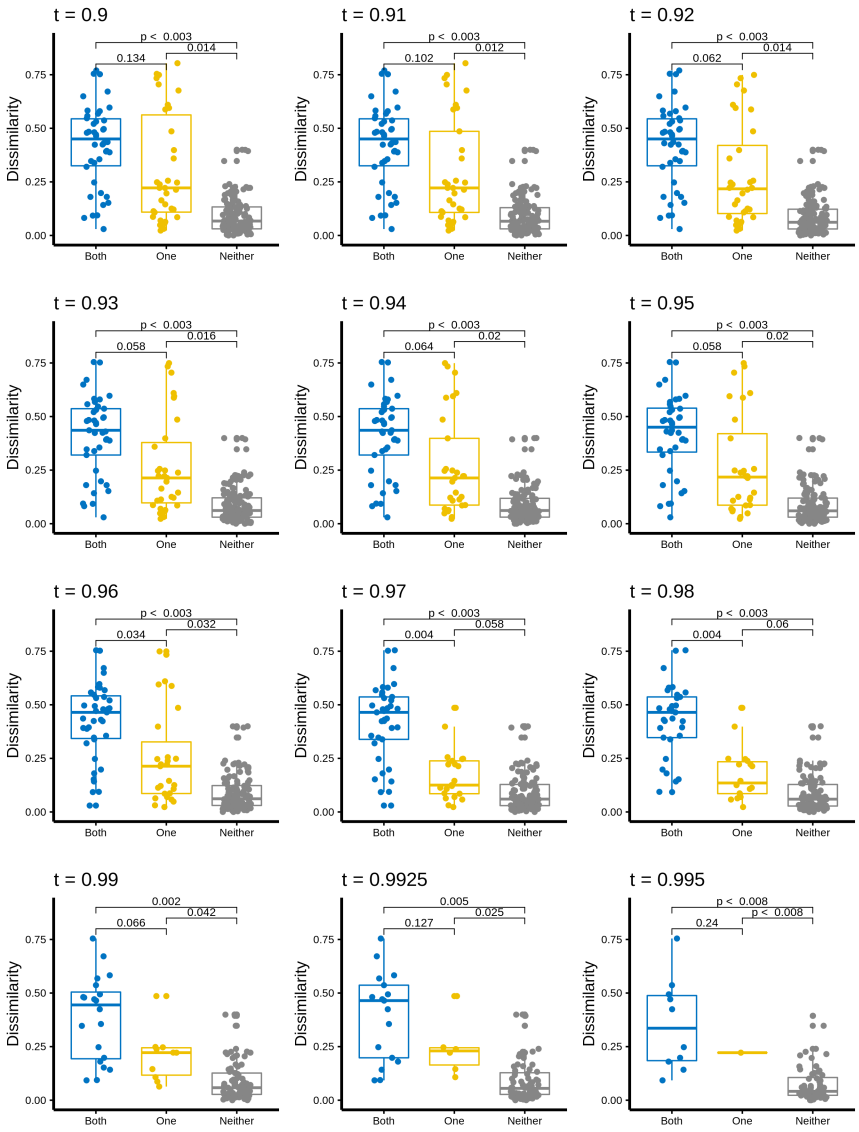
Supplementary Figure 2.3: Randomized Controls for Figure S2.2 also do not show a negative correlation. Panel A ,B C and D show the randomized controls for Figure S2.2 A,B,C and D respectivel



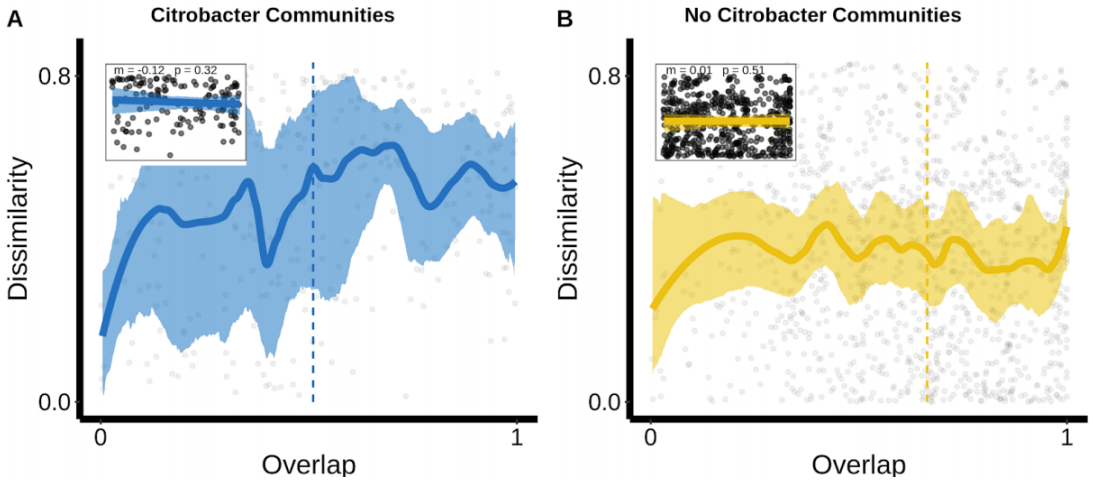
Supplementary Figure 2.4: Reproduction of Figure 2.1 F-H considering communities that come from the same inoculum. (A) Distributions for community pairs assembled in the same environment from the same inoculum (both in glucose, both in citrate, or both in leucine) with high overlap ($O > 0.98$). The dotted red-line is at half the maximum possible dissimilarity . (B) Histogram showing distributions of Overlaps for community pairs assembled from the same inoculum where one has been assembled on glucose and the other on citrate. The dotted lines give the frequency polygon for Glucose-Glucose Community pairs and Citrate-Citrate community pairs (Blue and yellow respectively). We use the same binwidth (.04) for both histograms and frequency polygons so the two are comparable. (C) Glucose-Citrate Communitie from the same Inoculum with high-Overlap ($O > 0.98$) have significantly higher mean Dissimilarity than Glucose-Glucose communities or Citrate-Citrate communities in the same Overlap range ($O > 0.98$) (P values computed using bootstrapping (see methods)).



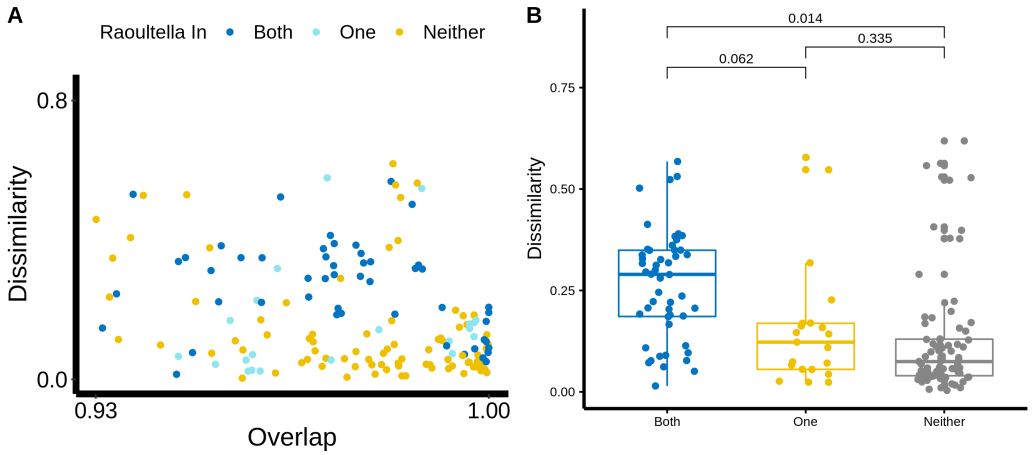
Supplementary Figure 2.5: The results plotted in Figure 2.1H are robust to the exact choice of threshold Overlap (t). In the main text, we show the result for $t = 0.98$, whereas here we give the results for a wide range of thresholds (t)



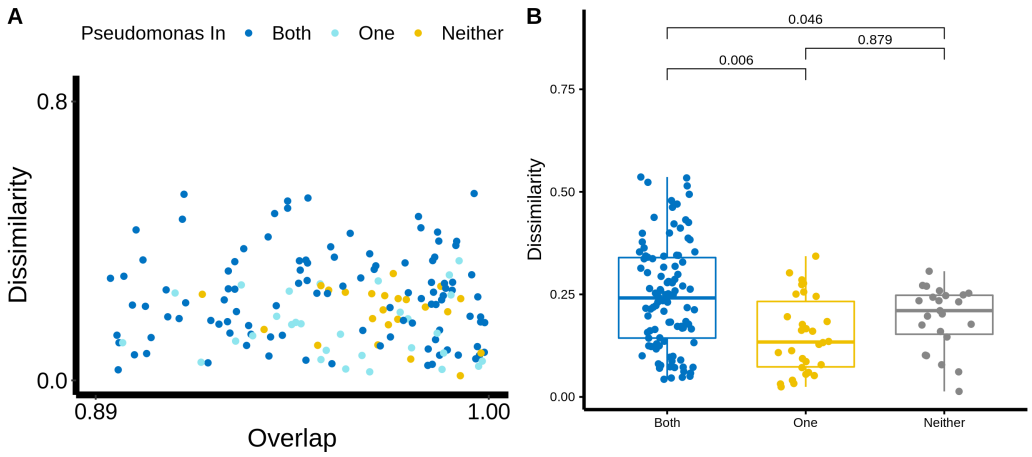
Supplementary Figure 2.6: The results plotted in Figure 2.2B are robust to the exact choice of threshold Overlap (t). In the figure, we show the result for the median overlap, whereas here we give the results for a wide range of thresholds (t)



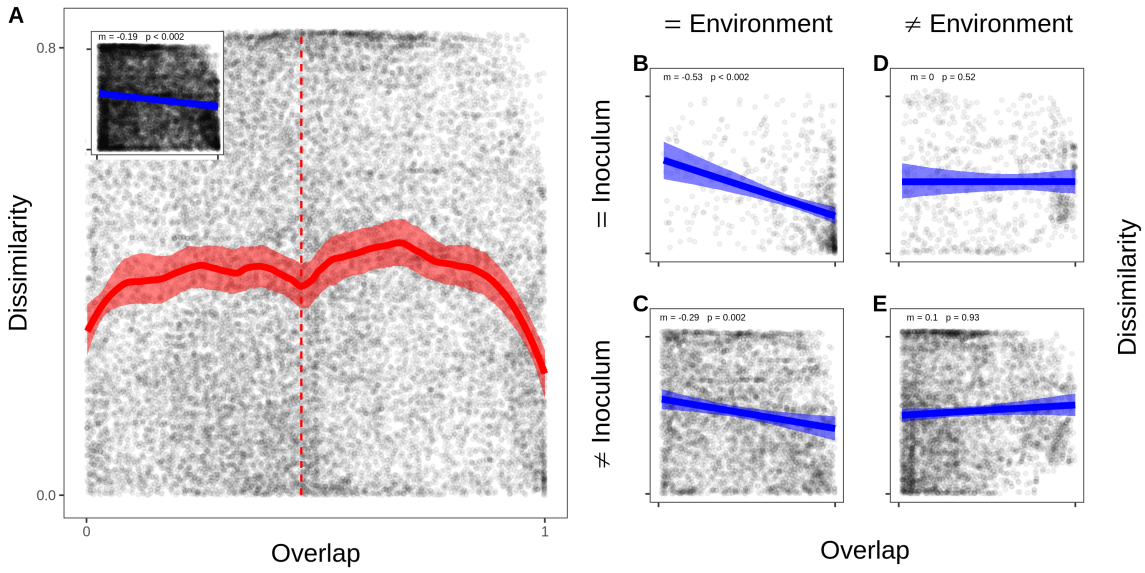
Supplementary Figure 2.7: Randomized Controls for Figure 2.C and 2.2D



Supplementary Figure 2.8: A Raoultella ESV is associated with dynamical dissimilarity in community assembled on Citrate from the same inoculum. Dissimilarity and Overlap of microbial community pairs assembled from the same regional pool on M9 + Citrate with above-median Overlap ($O > 0.93$) (B) Dissimilarity of the same set of communities. For (A) and (B) we label communities by whether Raoultella ESV is found in both communities (dark blue), only in one community (light blue) or in neither community (yellow).



Supplementary Figure 2.9: A Pseudomonas ESV is associated with dynamical dissimilarity in community assembled on Leucine from the same inoculum. Dissimilarity and Overlap of microbial community pairs assembled from the same regional pool on M9 + Leucine with above-median Overlap (O¿0.89) (B) Dissimilarity of the same set of communities. For (A) and (B) we label communities by whether Pseudomonas ESV is found in both communities (dark blue), only in one community (light blue) or in neither community (yellow)



Supplementary Figure 2.10: DOC of all microbial community pairs analyzed in this study. (n = 276 Samples). Shaded regions indicate the 95% confidence interval (Methods). We do observe a negative DOC but this is largely driven by community pairs assembled in the same environment. (Panel C and D).

2.8.2 Supplementary Tables

ESV	N	Difference in Mean	T−statistic	P.value
Raoultella	35	0.130070935	5.4394804	0.0119760479041916
Ochrobactrum	35	0.121718859	3.3184574	0.024
Stenotrophomonas	25	0.110209372	3.2080971	0.0200803212851406
Pseudomonas.1	52	0.048818453	1.9503271	0.203592814371257
Pseudomonas	76	0.036811427	1.3958867	0.239520958083832
Bordetella	33	0.062913891	1.3465327	0.255489021956088
Enterobacteriaceae	36	0.015863696	0.3767875	0.405189620758483
Pseudomonas.2	34	0.005168227	0.1446810	0.475049900199601
Klebsiella	64	−0.041667553	−1.1565688	0.728542914171657
Enterococcus	28	−0.037562471	−1.4167476	0.69061876247505

Supplementary Table 2.1: Repeat of analysis in Table 2.1 for Citrate communities

ESV	N	Difference in Mean	T−statistic	P.value
Pseudomonas	74	0.0775527033	4.409407808	0.00399201596806387
Pseudomonas.8	25	0.0664511098	2.471457821	0.0742971887550201
Delftia	72	0.0354502112	1.802358471	0.189620758483034
Pseudomonas.9	29	0.0491845272	1.476821279	0.248
Pseudomonas.1	55	0.0312316141	1.475657721	0.259481037924152
Stenotrophomonas	70	0.0156490342	0.791022079	0.331337325349301
Aeromonas	25	0.0001090002	0.004951194	0.474
Comamonas	30	−0.0089940703	−0.489915622	0.588822355289421
Bordetella	33	−0.0103111734	−0.556737490	0.610778443113772
Pseudomonas.3	28	−0.0333388745	−1.657847839	0.806387225548902

Supplementary Table 2.2: Repeat of analysis in Table 2.1 for Leucine communities

2.9 Acknowledgements

We thank members of AS and YYL Laboratories for helpful discussion about the work. We thank Josh Goldford for providing us with the experimental data analyzed in this paper.

Chapter 3

Nutrient Dominance Governs the Assembly of Microbial Communities in Mixed Nutrient Environments

Sylvie Estrela [1]*, Alicia-Sanchez Gorostiaga*[1]*, Jean C.C. Vila*[1] *and Alvaro Sanchez*

Estrela S, Sanchez-Gorostiaga A, Vila JCC, Sanchez A. 2021, Nutrient dominance governs the assembly of microbial communities in mixed nutrient environments. eLife 10, e65948

[1]Sylvie Estrela, Alicia-Sanchez Gorostiaga and Jean C.C. Vila contributed equally to this work

3.1 Abstract

A major open question in microbial community ecology is whether we can predict how the components of a diet collectively determine the taxonomic composition of microbial communities. Motivated by this challenge, we investigate whether communities assembled in pairs of nutrients can be predicted from those assembled in every single nutrient alone. We find that although the null, naturally additive model generally predicts well the family-level community composition, there exist systematic deviations from the additive predictions that reflect generic patterns of nutrient dominance at the family level. Pairs of more-similar nutrients (e.g. two sugars) are on average more additive than pairs of more dissimilar nutrients (one sugar–one organic acid). Furthermore, sugar–acid communities are generally more similar to the sugar than the acid community, which may be explained by family-level asymmetries in nutrient benefits. Overall, our results suggest that regularities in how nutrients interact may help predict community responses to dietary changes.

3.2 Introduction

Understanding how the components of a complex biological system combine to produce the system's properties and functions is a fundamental question in biology. Answering this question is central to solving many fundamental and applied problems, such as how multiple genes combine to give rise to complex traits [Phillips, 2008, Mackay, 2014], how multiple drugs affect the evolution of resistance in bacteria and cancer cells[Michel et al., 2008, Woods et al., 2006], how multiple environmental stressors affect bacterial physiology [Cruz-Loya et al., 2019], or how multiple species affect the function of a microbial consortium [Sanchez-Gorostiaga et al., 2019, Gould et al., 2018, Guo and Boedicker, 2016].

In microbial population biology, a major related open question is whether we can predict how the components of a diet collectively determine the taxonomic and functional composition of microbial communities. Faith and co-workers tackled this question using a defined gut microbial community and a host diet with varying combinations of four macronutrients [Faith et al., 2011]. This study found that community composition in combinatorial diets could be predicted from communities assembled in separate nutrients using an additive linear model. Given the presence of a host and its own possible interactions with the nutrients and resident species, it is not immediately clear whether such additivity

is directly mediated by interactions between the community members and the supplied nutrients or whether it is mediated by the host, for instance by producing additional nutrients, or through potential interactions between its immune system and the community members. More recently, Enke et al., 2019 found evidence that marine communities assembled in mixes of two different polysaccharides could be explained as a linear combination of the communities assembled in each polysaccharide in isolation.

Despite the important insights provided by both of these studies, we do not yet have a general quantitative understanding of how specific nutrients combine together to shape the composition of self-assembled communities [Pacheco et al., 2021]. Motivated by this challenge, here we use an enrichment community approach (i.e. where natural microbial communities are grown in a defined growth medium under well-controlled lab conditions) to systematically investigate whether the assembly of enrichment microbial communities in a collection of defined nutrient mixes could be predicted from the communities that assembled in each of the single nutrients in isolation.

3.3 Results

3.3.1 A null expectation for community assembly in mixed nutrient environments

To investigate whether communities assembled in pairs of nutrients can be predicted from those assembled in every single nutrient alone, we must first develop a quantitative null model that predicts community composition in a mixed nutrient environment in the case where each nutrient recruits species independently. Any deviation between the null model prediction and the observed (measured) composition reveals that nutrients are not acting independently, but rather 'interact' to shape community composition. This definition of an interaction as a deviation from a null model that assumes independent effects is commonplace in systems-level biology[Sanchez, 2019, Tekin et al., 2018].

In order to formulate the null expectation for independently acting nutrients, let us consider a simple environment consisting of two unconnected demes where two bacterial species, A and B, can grow together. The first deme contains a single growth limiting nutrient (nutrient 1), while the second deme contains a different single limiting nutrient (nutrient 2) (Figure 3.1A). In this scenario, each nutrient influences the abundance of species A and B independently: the microbes growing on nutrient one do not have ac-

cess to nutrient two and vice versa. Let's denote the abundance of species A in demes 1 and 2 by $n_{A,1}$ and $n_{A,2}$, and the abundance of species B as $n_{B,1}$ and $n_{B,2}$, respectively. If we now consider the two-deme environment as a whole, the abundance of species A is the sum of its abundance in each deme $n_{A,12} = n_{A,1} + n_{A,2}$ (likewise, for species B $n_{B,12} = n_{B,1} + n_{B,2}$). This example illustrates that in the scenario when two limiting nutrients act independently, each of them recruits species just as if the other nutrient were not there. In such case, the abundance of each species in a nutrient mix is the sum of what we would find in the single-nutrient habitats.

Under the null model, the relative abundance of species i in a mix of nutrients 1 and 2 can be written as $f_{i,12}(null) = w_1 f_{i,1} + w_2 f_{i,2}$ where $f_{i,1}$ and $f_{i,2}$ are the relative abundances of i in nutrients 1 and 2, respectively, and w_1 and w_2 are the relative number of cells in nutrients 1 and 2 (Materials and methods). Any quantitative difference between the null model prediction and the observed composition quantifies an 'interaction' between nutrients. Accounting for the presence of such interactions, the model can be re-written as $f_{i,12} = f_{i,12}(null) + \epsilon_{i,12}$ where $\epsilon_{i,12}$ represents the interaction between nutrients 1 and 2 (Figure 3.1B).

3.3.2 Experimental system

Equipped with this null model, we can now ask to what extent the nutrients recruit species independently in mixed environments. To address this question, we followed a similar enrichment community approach to the one we have used in previous work for studying the self-assembly of replicate microbial communities in a single carbon source [Goldford et al., 2018, Estrela et al., 2022] (Materials and methods, Figure 3.2A). Briefly, habitats were initially inoculated from two different soil inocula. Communities were then grown in synthetic (M9) minimal media supplemented with either a single carbon source or a mixture of two carbon sources, and serially passaged to fresh medium every 48 hr for a total of 10 transfers (dilution factor = $125\times$) (Figure 3.2A). The carbon source pairs consisted of a focal carbon source mixed at equal C-molar concentrations with one of eight additional carbon sources. We previously found that stable multi-species communities routinely assemble in a single carbon source (which is limiting under our conditions), and they converge at the family level in a manner that is largely governed by the carbon source supplied, while the genus or lower level composition is highly variable (Goldford et al., 2018). We chose glucose as the focal carbon source because we have previously carried out multiple assembly experiments in this nutrient [Goldford et al., 2018, Estrela et al., 2022].

As the additional carbon sources, we chose nutrients that are simple and metabolically diverse (sugar vs acid, that contain a different number of atoms of carbon, and that enter metabolism at different points), namely ribose, fructose, cellobiose, and glycerol (i.e. a pentose, a hexose, a disaccharide, and a sugar alcohol) and fumarate, benzoate, glutamine, and glycine (two organic acids and two amino acids). All carbon sources were also used in single carbon source cultures.

Communities assembled in single sugars contained 5–24 exact sequence variants (ESVs), mainly belonging to the Enterobacteriaceae family (mean relative abundance $\pm$ SD of $\approx 00.98 \pm 0.03$), a sugar specialist (Supplementary Figure 3.1). In contrast, communities assembled in organic acids exhibited a higher richness (12–36 ESVs), and unlike in sugars, Enterobacteriaceae were generally rare (mean $\pm$SD $\approx 0.06 \pm 0.06$). Instead, communities were dominated by respirative bacteria mainly belonging to the Pseudomonadaceae (mean $\pm$ SD $\approx 0.51 \pm 0.25$), Moraxellaceae (mean $\pm$ SD $\approx 0.18 \pm 0.21$), and Rhizobiaceae (mean $\pm$ SD $\approx 0.11 \pm 0.13$) families (Supplementary Figure 3.1). Because of the observed family-level convergence across carbon sources, which is consistent with previous studies[Goldford et al., 2018, Estrela et al., 2022, Diaz-Colunga et al., 2022], we focus our analysis below on family-level abundance.

3.3.3 The null model of independently acting nutrients explains a high fraction of the variation observed

To investigate the predictive power of the null (additive) model, we compare the predicted and observed relative abundances of each family for each carbon source pair across all experiments. Our results show that the null model predicts reasonably well the family-level abundances on average (Pearson's R = 0.95 and $p < 0.001$; RMSE = 0.073, N = 223) (Figure 3.2B,Supplementary Figure 3.2 and 3.3). To confirm that the strong predictive power of the null model is not an idiosyncrasy of using glucose as the focal carbon source in the pairs, we repeated the same experiment with succinate (an organic acid) as the focal carbon source. Although the correlation between observed and predicted abundance is lower than for glucose, the null additive model is still predictive (Pearson's R = 0.87 and $p<0.001$; RMSE = 0.094; N = 257) (Figure 3.2B).

This result seems to indicate that, at the family level, a simple model that assumes that nutrients act independently can predict community composition in a pair of nutrients (for an analysis of this point at the genus and ESV level, see Supplementary Figure 3.4). However, when we looked at this more closely and broke down our results by carbon source and

family, we found consistent and systematic deviations from the null model (Figure 3.2C). For example, across all succinate–sugar pairs, Enterobacteriaceae are significantly more abundant than predicted by the null model (one-tailed paired t-test, N = 8, $p<0.05$ based on 1000 permutations; see Materials and methods), while both Rhizobiaceae and Moraxellaceae are less abundant than predicted (one-tailed paired t-test, N = 8, $p<0.05$ based on 1000 permutations; see Materials and methods) (Figure 3.2C). The null 'interaction-free' model also predicts species abundance better in certain carbon source combinations (e.g. glucose + ribose) than in others (e.g. glucose + glutamine) (Figure 3.2C). The existence of systematic deviations from the null prediction reveals that some nutrient pairs do not recruit families independently, but instead 'interact' with each other to affect the abundance of specific families.

3.3.4 A simple dominance rule in mixed nutrient environments: sugars generally dominate organic acids

To map the regularities in nutrient interactions observed, we next sought to characterize the nature of these interactions for each carbon source pair and every family. One helpful way of visualizing nutrient interactions is to draw the pairwise abundance landscape for each species and carbon source pair (Figure 3.3A). For instance, a species could be either more abundant in a pair of nutrients than it is in any of them independently (synergy). Or it could be less abundant than it is in any of the two (antagonism). Dominance is a less extreme interaction that can be visualized by the pushing of a species abundance toward the value observed in one of the two nutrients and away from the average, thus overriding the effect of the second available nutrient (Figure 3.3A).

When the interaction is positive ($\epsilon > 0$), the dominant nutrient is the one where the family grew to a higher abundance. When the interaction is negative ($\epsilon < 0$), the dominant nutrient is the one where the species grew less well. Mathematically, dominance occurs when $|\epsilon| > 0$ and $min(f_{i,1}, f_{i,2}) \leq f_{i,12} \leq max(f_{i,1}, f_{i,2})$ while synergy and antagonism (forms of super-dominance) occur when $|\epsilon| > 0$ and $f_{i,12} > max(f_{i,1}, f_{i,2})$ and $f_{i,12} < min(f_{i,1}, f_{i,2})$, respectively (Materials and methods). Figure 3.3B shows representative examples of dominant carbon source interactions. For instance, Moraxellaceae and Rhizobiaceae grow strongly on succinate, but they are not found in fructose. When fructose is mixed with succinate, both families drop dramatically in abundance, despite their high fitness in succinate alone. Interestingly, however, the dominance of fructose over succinate is not observed for all families: those two nutrients do not interact on Pseu-

domonadaceae, whose abundance is well predicted by the null model. Using this framework, we then systematically quantified the prevalence of dominance, antagonism, and synergy between nutrients for each family (Supplementary Figure 3.6). While $66\%$of the nutrient pair combinations exhibited no significant interaction, dominance was by far the most common interaction amongst those that interacted (75%, Figure 3.3—Supplementary Figure 3.6A). It occurred predominantly in the sugar–acid pairs, and to a lesser extent in the acid–acid pairs, and only rarely in the sugar–sugar pairs (Figure 3.3—Supplementary Figure 3.6B). This result strongly suggests that nutrient interactions are not random but do have a specific structure that is conserved at the family level (Figure 3.3—Supplementary Figure 3.6C).

To systematically characterize and quantify nutrient dominance, we developed a dominance index (δ) (Materials and methods). For visualization purposes, the dominance index for the sugar–acid pairs (we will discuss the aci-acid pairs later) is written as $\delta_i = -|\epsilon_{12}|$ when the sugar dominates and as $\delta_i = |\epsilon_{12}|$ when the acid dominates. If $\epsilon_{12} = 0$, then $\delta_i = 0$. That is, in the absence of interaction between nutrients, there is no dominance. By plotting the dominance index for each pair of nutrients and each family, we observe a generic pattern of dominance of sugars over acids (Figure 3.3C). The families Moraxellaceae or Rhizobiaceae are recruited to the community by most organic acids in isolation, but they are not found in most sugar communities. When sugars and organic acids are mixed together, the sugar dominates and both families are at much lower abundances (by ≈ 6-fold in the case of Moraxellaceae and ≈ 114-fold in Rhizobiaceae) than expected by the null model, even though the organic acid where they thrived is present in the environment. Consistent with this pattern, we found that pairs of more-similar nutrients (a pair of sugars or a pair of organic acids) were significantly better predicted by the null model than mixed organic acid–sugar pairs (Figure 3.3D). No generic pattern of dominance was observed in the acid–acid mixtures (Supplementary Figure 3.7). When we examine interactions and dominance at the genus level, we find that sugars do not exhibit the same dominance for all genera within the same family (Supplementary Figure 3.9). This result is consistent with the convergence of community structure at the family level (despite substantial variation at lower levels of taxonomy), which we have reported for communities assembled in a single nutrient [Goldford et al., 2018, Estrela et al., 2022]. Together, these results indicate that interactions between nutrients are not universal, but rather they are conserved at the family-level.

3.3.5 An extension of the null consumer-resource model with an asymmetry in nutrient benefits recapitulates the dominance pattern observed

Our findings pose intriguing questions about the mechanisms behind the nutrient interaction patterns we have observed. For instance, is it reasonable to expect that the additive null model should have worked as well as it did, and better at the family than at the species level? Why are pairs of more-similar nutrients better explained by the null model than pairs of more dissimilar nutrients? What may explain why nutrients dominate over others at the family level? And why do sugars generally dominate organic acids for most families?

We have previously shown that many of the properties of our experimental enrichment communities reflect the generic emergent behavior of consumer-resource models [Goldford et al., 2018, Marsland et al., 2019], and subsequent work extended this finding to complex natural communities ([Marsland et al., 2020a]). We thus sought to ask whether our observations regarding the assembly of communities in pairs of resources are similarly reflecting a generic emergent property of consumer-resource models. To address this question, we followed the same procedure as we and others have done in previous work [Goldford et al., 2018, Marsland et al., 2020a,b, Serván and Allesina, 2021] and simulated the top-down assembly of microbial communities in pairs vs single nutrients using a recently developed Microbial Consumer Resource Model (MiCRM) [Goldford et al., 2018, Marsland et al., 2019, 2020b](see Materials and methods). The MiCRM differs from the classical MacArthur-Levins model (MacArthur, 1970) in that it includes metabolic cross-feeding in a manner that preserves thermodynamic balance. The model and the details of the simulations are described in the Materials and methods section. In brief, 200 species are seeded into each habitat at the start of a simulation. Each of these is represented by a different vector of resource uptake rates. These vectors are randomly sampled in a manner that captures the existence of two functional guilds, each of which specializes in a different group of resources (e.g. sugars vs organic acids) (Figure 3.4A). Members of the family specializing on sugars (i.e. the Enterobacteriaceae) have on average a higher uptake rate on each sugar whereas members of the family specializing on acids (i.e. the Pseudomonadaceae) have on average a higher uptake rate on each acid. The magnitude of specialization by each family on its preferred resource type is tuned by two parameters, qA and qS, which modulate the mean and variance of the probability distribution from which the uptake rates are sampled (see Materials and methods for more details). We

note that this specialization is quantitative rather than discrete, as all species are assumed to be able to consume all of the resources (a point that is in general consistent with our experimental findings [Supplementary Figure 3.10]). Communities are allowed to find a dynamical equilibrium, at which point we stop the simulation. In total, and in order to get to generic behavior, we generated 100 simulations each with a different random set of species (Materials and methods).

A generic property of these simulations is that the species-level community composition on mixtures of two limiting nutrients is reasonably well described by the additive null model (Pearson's R = 0.7 and $p<0.001$; RMSE = 0.097; N = 2440) (Figure 3.4B; Materials and methods), which is consistent with previous consumer-resource modelling work [Marsland et al., 2020a]. In addition, when we group species by the functional groups they belong to (i.e. family), the predictive ability of the additive null model improves (Pearson's R = 0.99 and $p<0.001$; RMSE = 0.03; N = 414) (Figure 3.4B), a point that is consistent with our experimental findings (Supplementary Figure 3.4). This family-level additivity holds when communities are randomly colonized by a different set of species (Supplementary Figure 3.11), suggesting that family-level additivity is robust to species-level taxonomic variability. The predictive accuracy of our null model is, however, influenced by the level of resource specialization. The less specialized (i.e. more generalist) the families are, the lower the predictive power of the null additive model (Supplementary Figure 3.12).

By contrast, the simulated communities do not exhibit any systematic dominance, neither at the species nor at the family level (Figure 3.4B). What feature of the MiCRM might be causing us to miss this experimental behavior? One assumption of the model, which we had made for the sake of simplicity and for consistency with previous work, is that all nutrients are equally valuable for the microbes that specialize on them. In other words, the benefits of growing in each type of nutrient are symmetric. Yet, this assumption is not really consistent with the empirical reality that glucose specialists, such as Enterobacteriaceae, grow more strongly in sugars than organic acid specialists do on organic acids. This is illustrated in Figure 3.4C, where we plot the growth advantage for seven Enterobacteriaceae isolates in sugar media vs the growth advantage of Pseudomonadaceae, Rhizobiaceae, and Moraxellaceae isolates in organic acids.

We postulated that including this asymmetry may unbalance the competition for resources and give rise to nutrient dominance at the family level, as the family that lies on the winning side of that asymmetry may leverage its enhanced competitive ability in the most valuable nutrient to displace the losing family from its lower-value nutrient niche. To

test this intuition, we relaxed the symmetry in resource value that was imposed by default in the model, and repeated our simulations for different levels of nutrient value asymmetry (the simulations still include facilitation via metabolite secretion, as we had done in all prior simulations) (Figure 3.4A, Supplementary Figure 3.13). As we show in Figure 3.4D, and consistent with our intuition, nutrient dominance at the family level may emerge as a generic property of microbial consumer-resource models when a nutrient is substantially more valuable than the other. Reassuringly, our experiments indicate that dominance is generally favorable to the taxa that benefits from growth asymmetry for example to Enterobacteriaceae in sugar–acid mixes, and unfavorable to families in the losing end of growth asymmetry (Pseudomonadaceae, Rhizobiaceae, and Moraxellaceae) in those same environments. This observation is consistent with the behavior of the model (Figure 3.4D).

Our consumer-resource model shows that dominance is a general outcome of consumer-resource interactions when there exists an asymmetry in nutrient benefit (an asymmetry that is indeed observed for the families in our communities), but other mechanisms of dominance may be at play too. For instance, one plausible mechanism that could lead to dominance is oxygen limitation, in particular if different carbon sources were to have different oxygen requirements [Hempfling and Mainzer, 1975, Skrinde and Bhagat, 1982]. To explore this idea of asymmetric oxygen demands, we used flux-balance analysis (FBA) to determine the oxygen demands of growth on each of the single-carbon sources. We found that, except for benzoate, all carbon sources have similar oxygen demands (Supplementary Figure 3.14). This does not rule out, however, the possibility that kinetics of growth and oxygen uptake may still contribute to oxygen depletion in a manner that may further stimulate dominance (in addition to the asymmetric resource benefits we report in Figure 3.4).

3.4 Discussion

Our analysis indicates that our empirical observations regarding the assembly of microbial communities in nutrient mixes are consistent with generic behavior of consumer-resource models. Based on this finding, we cautiously suggest that family-level asymmetries in nutrient uptake rates may be a possible mechanism for the general nutrient dominance patterns we have observed, and that a null additive model is in general a good first approximation for the assembly of microbial communities in simple nutrient mixtures (a pattern that is consistent with previous work [Faith et al., 2011, Enke et al., 2019]). It is

important to recognize, however, that other explanations and mechanisms of dominance may be at play too. Generally, these can be split into two main categories: asymmetries in how species respond to the provided nutrient and asymmetries that emerge as a result of the constructed environment. Below, we discuss several specific mechanisms that may contribute to each of these.

Our null model (and consumer-resource simulations) assumes, by definition, that the growth of a single species on a mixture of nutrients (in terms of growth rate and yield) will be the aggregate sum of the growth on each nutrient alone. Multiple mechanisms, however, could lead to violations of this assumption. Firstly, a species might not consume both nutrients simultaneously but may instead consume them sequentially, or diauxically, resulting in fluctuations in the effective resource specialization of each species [Monod, 1942, Lendenmann et al., 2000, Erickson et al., 2017, Pacciani-Mori et al., 2020]. Secondly, even if a species is co-utilizing both nutrients, the biomass yields may not be additive, due to synergistic effects of using different nutrients for different cellular functions (such as energy versus biomass or for synthesis of different biomass precursors) [Lendenmann et al., 1996, Pacheco et al., 2019, Wang et al., 2019]. Thirdly, a molecule that can be used as a nutrient by one species may have an inhibitory effect on another species, for example benzoate is known to have antimicrobial activities against some bacteria (which may explain why benzoate dominates sugars for some families in Figure 3C; [Stanojevic et al., 2009]). The growth dynamics on mixtures of carbon sources have been extensively characterized in simple sugars for a few model organisms (such as Escherichia coli, Bacillus subtilis, and Pseudomonas aeruginosa), but we still lack a systematic understanding of mixed-substrate growth across taxa and environment [Harder and Dijkhuizen, 1982, Görke and Stülke, 2008, Bajic and Sanchez, 2020] . Systematically mapping mixed-resource utilization strategies represents an exciting direction for future work and would allow us to better predict the effects of environmental complexity on the emergent properties of complex microbial communities.

Importantly, even if species respond to the supplied pair of nutrients in an additive manner, niche construction (and thus the interactions between species) may not be additive. For example, species may secrete secondary metabolites or antimicrobial agents on nutrient mixtures, which may interact with each other [Sánchez et al., 2010, Mendonca et al., 2020, Fujiwara et al., 2020]. Moreover, cellular growth can change other physicochemical properties of the environment aside from carbon source availability, such as by changing the pH, the accessibility of non-carbon source nutrients leading to co-limitation, or oxygen availability [Harpole et al., 2011, Cremer et al., 2017, Sánchez-Clemente et al.,

2020].

The wealth of independent mechanisms that may contribute to nutrient dominance illustrates the potential importance of this phenomenon. Quantitatively elucidating the specific mechanisms that may explain the individual patterns of nutrient interactions (or lack thereof) for each family and in each pair of nutrients would require us to measure the amounts of all nutrients secreted by every species in each environment over time (i.e. in each nutrient and in each pair) and then characterize the growth curves of all species in those nutrients. Although such monumental effort is beyond the scope of this paper, we hope that our findings and methodology will be a stepping stone towards elucidating how microbial communities assemble in complex nutrient mixtures and that they will stimulate further theoretical and empirical work. We propose that top-down community assembly in combinatorially reconstructed nutrient environments can be a helpful approach not only to understand the origins of microbial biodiversity, but also to learn how to manipulate existing microbiomes by rationally modulating nutrient availability.

3.5 Figures

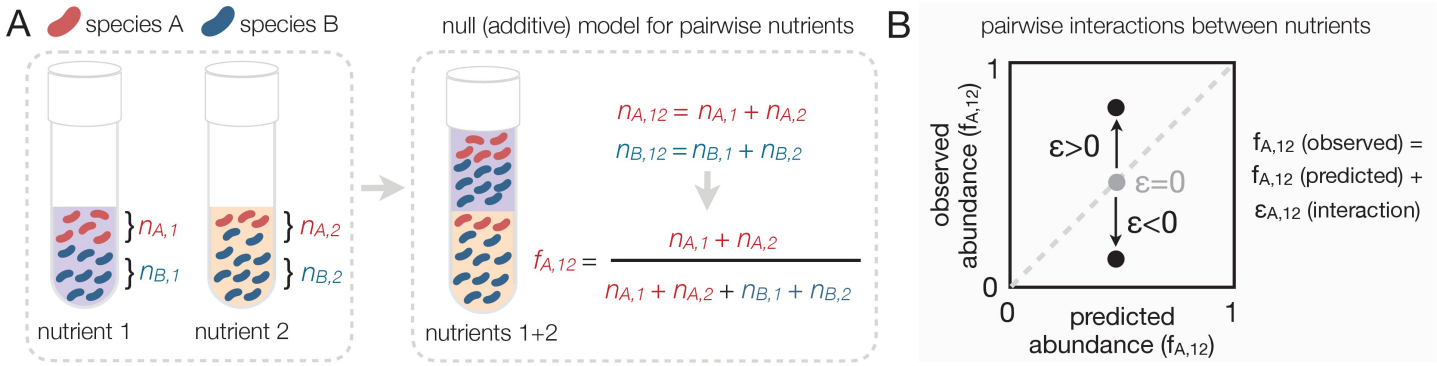


Figure 3.1: Predicting community composition in mixed nutrient environments:(A) Community composition in a single nutrient (nutrient 1 or nutrient 2) vs a mixture of nutrients (nutrient 1 nutrient 2). Assuming that nutrients act independently, the null model predicts that the abundance of each species in the mixture is the sum of its abundance in the single nutrients (i.e. additive). (B) Plotting the experimentally measured (observed) relative abundance in the mixed carbon sources against its predicted (from null model) relative abundance reveals the presence or absence of interactions. Any deviation from the identity line (predicted = observed) is the interaction effect (ϵ). When $\epsilon = 0$, there is no interaction between nutrients. When ϵ is non-zero, community composition is affected by nutrient interactions. If $\epsilon > 0$, the null model underestimates the abundance. If $\epsilon < 0$, the model overestimates the abundance.

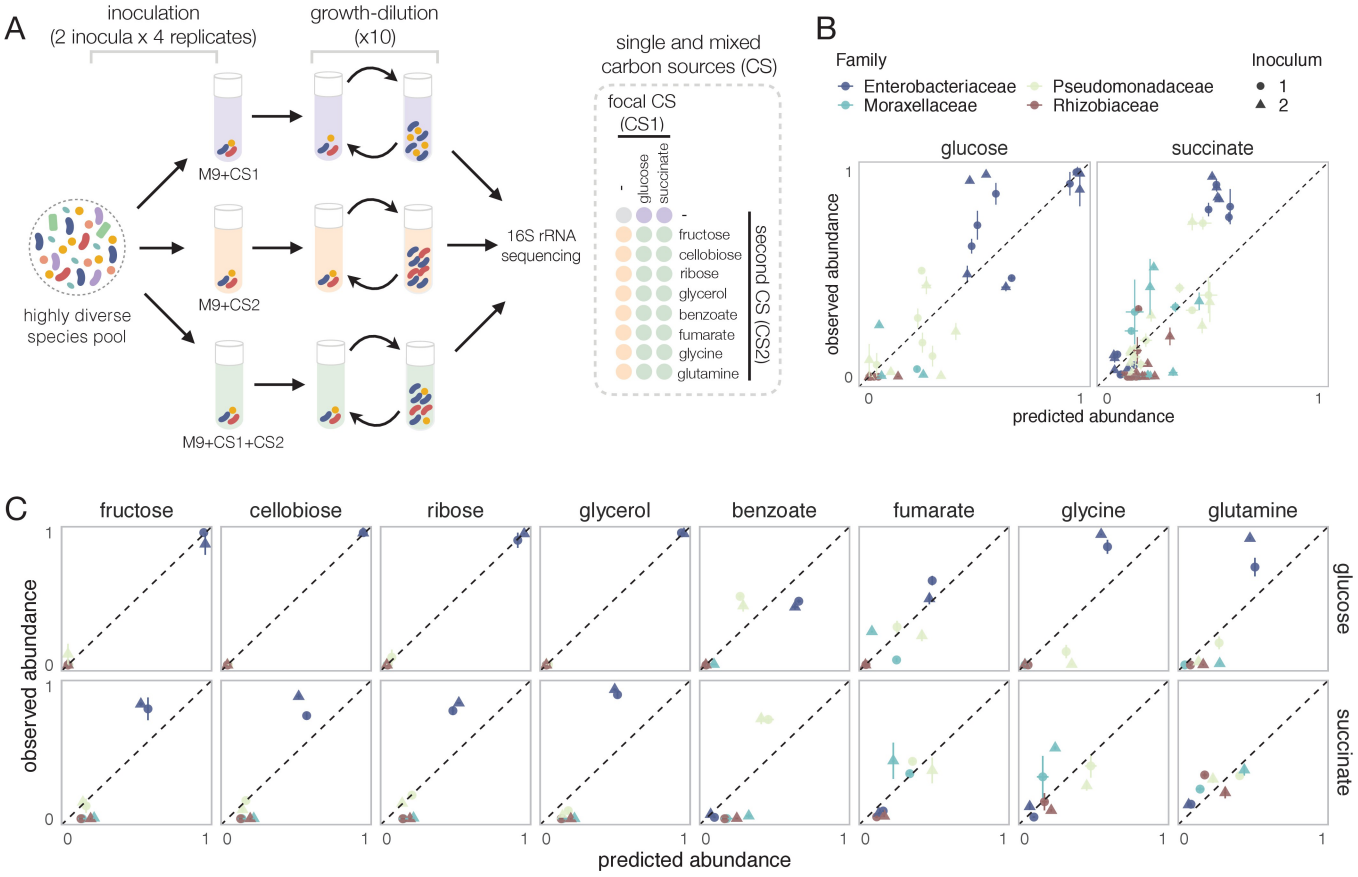
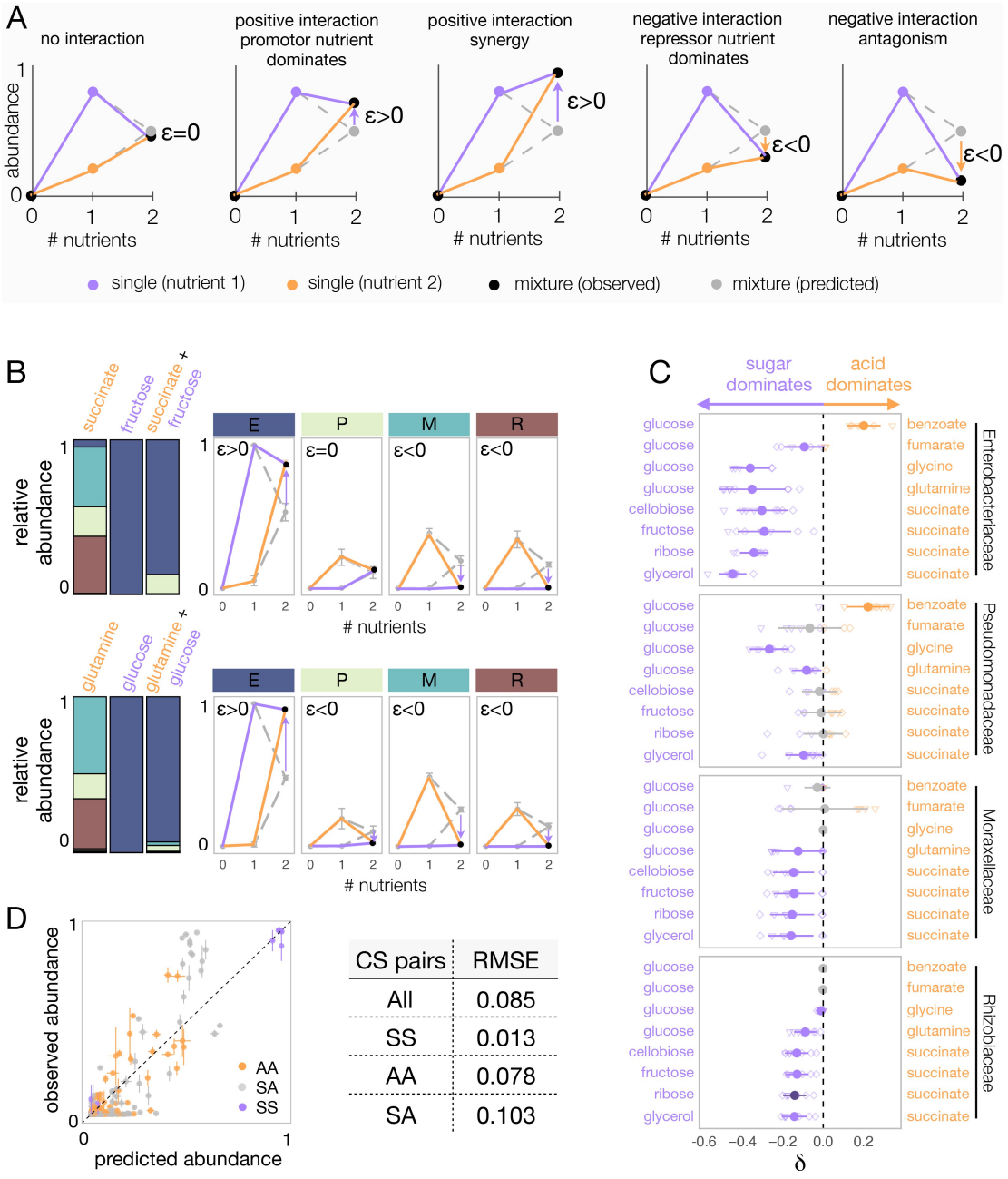


Figure 3.2: Systematic deviations from the null prediction reveals that some nutrients interact to shape community assembly : (A) Schematic of experimental design. Two different soil samples were inoculated in minimal M9 medium supplemented with either a single carbon source (CS1 or CS2) or a mixture of two carbon sources (CS1 + CS2) (three to four replicates each). Communities were propagated into fresh media every 48 hr for 10 transfers and then sequenced to assess community composition. Carbon source mixtures consisted of a focal carbon source (CS1; glucose or succinate) mixed with a second carbon source (CS2). (B, C) For each pair of carbon sources, we show the experimentally observed and predicted (by the null additive model) relative abundance of each family in the mixture. Any deviation from the identity line (predicted = observed) reveals an interaction effect. Only the four most abundant families are shown. Error bars represent mean $\pm$ SE.



CS pairs	RMSE
All	0.085
SS	0.013
AA	0.078
SA	0.103

Figure 3.3: Sugars generally dominate over organic acids

Figure 3.3: (A) Detecting interactions and hierarchies of dominance between nutrients on microbial community composition. Drawing the single and pairwise abundance landscapes for each species allows us to visualise interactions between nutrients. Multiple types of interactions are possible, including dominance, synergy, and antagonism. Interactions occur when ϵ is significantly different from 0 (Materials and methods). Synergy (antagonism) occurs when the abundance in the mixture is greater (lower) than the abundances in any of the single nutrients independently (Materials and methods). Dominance occurs when the abundance in the mixture is closer or similar to the abundance in one of the singles. The landscape also allows us to identify which carbon source has a dominating effect within the pair. When $\epsilon > 0$, the growth-promoting nutrient dominates and has an overriding effect in the community composition. In contrast, when $\epsilon < 0$, the growth-repressing nutrient dominates. (B) Two examples of nutrient interactions (succinate + fructose and glucose + glutamine) exhibiting sugar dominance. Barplots show a representative replicate from one of the inocula (Figure 2—figure supplements 1–2). For instance, the landscape for succinate-fructose shows that fructose overrides the effect of succinate by promoting Enterobacteriaceae (E), and repressing Moraxellaceae (M) and Rhizobiaceae (R) (purple arrows), whereas no interaction effect is observed for Pseudomonadaceae (P). Error bars represent mean $\pm$ SD of the four replicates. (C) Dominance index for the eight sugar–acid pairs and the four dominant families. Filled circles show the mean $\pm$ SD of the two inocula $\times$ four replicates for each pair of nutrients, and open symbols show all eight independent replicates (different shapes for different inocula), except for glycine pairs where $N = 6$. Purple indicates that the sugar dominates while orange indicates that the acid dominates. Lighter purple and orange indicate dominance, while darker purple and orange indicate super-dominance (synergy or antagonism). An interaction occurs when the abundance is greater ($\epsilon > 0$) or lower ($\epsilon < 0$) in the carbon source mixture than predicted by the null model (one-tailed paired t-test, $p < 0.05$, $N = 8$, based on 1000 permutations; Materials and methods). In gray are shown cases where there is no interaction, or when dominance is undefined because the two inocula exhibit opposite dominant nutrient (in which case δ is shown as both $-\delta$ and $+\delta$). (D) Predicted vs observed family-level abundance. For each pair of carbon sources (CS), shown is the experimentally observed and predicted (by the null model) relative abundance of each family in the mixed carbon sources. Any deviation from the identity line (predicted = observed) is the interaction effect. The colors show whether the carbon source pairs are sugar–sugar (SS), acid–acid (AA), or sugar–acid (SA). Error bars represent mean $\pm$ SE. Table shows RMSE for each carbon source pair type.

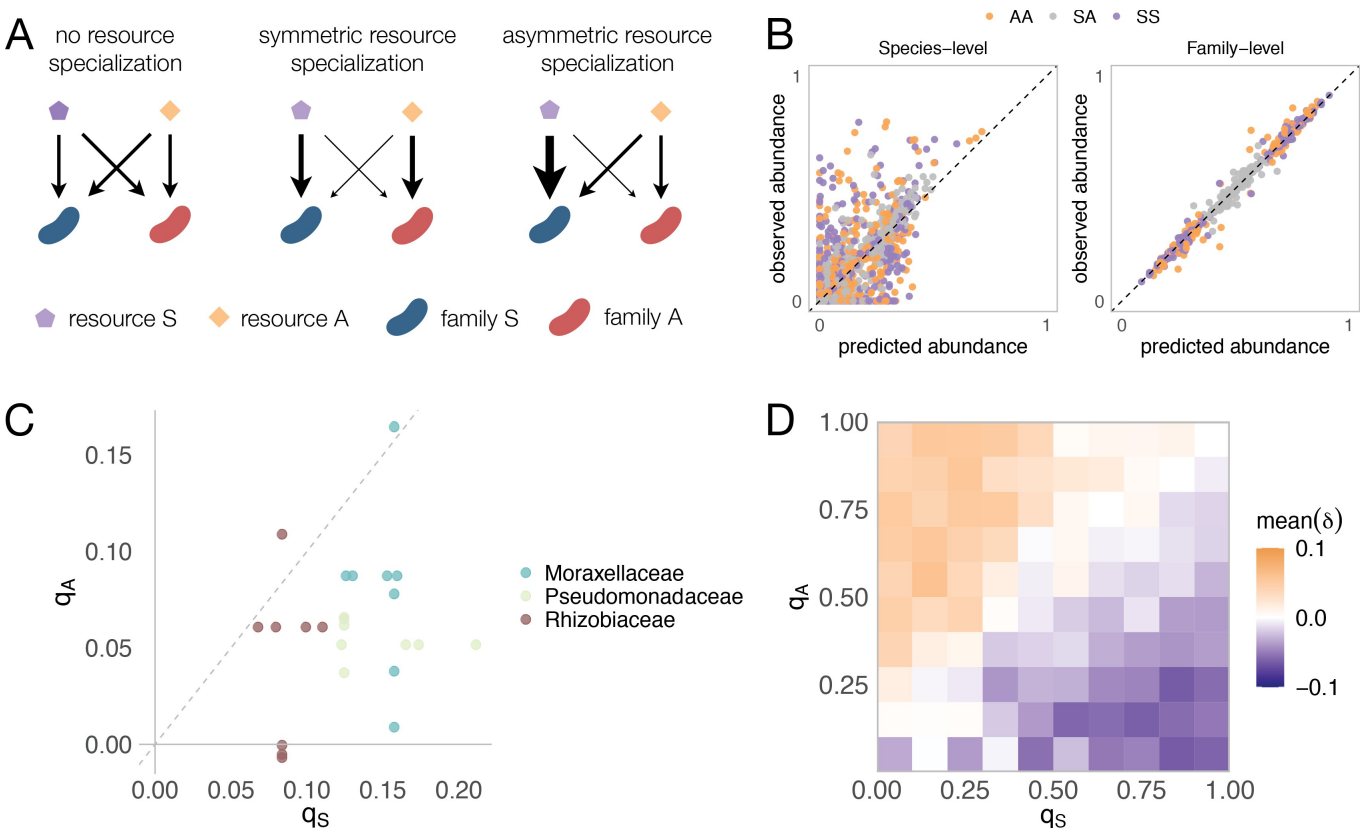


Figure 3.4: Family-level asymmetry in nutrient benefits can lead to dominance

Figure 3.4: (A) Schematic illustrating different scenarios of nutrient preference. There are two families (FS and FA) and two resource classes (RS and RA). Without resource specialization, FS and FA have equal access to RS and RA. With symmetric specialization, each family prefers its own resource class with the same strength. With asymmetric specialization, one family (FS) has better access to its own resource class (RS) relative to that of the other family (FA) on its own resource class (RA). (B) A mechanistically explicit consumer-resource model that incorporates resource competition, resource specialization and nonspecific cross-feeding (Materials and methods) recovers the predicted additivity pattern at both the species (left) and family (right) level of taxonomic organization. The observed relative abundance of each species or family in 300 communities grown on a different pair of nutrients (100 AA, 100 SS, and 100 SA) is plotted against the abundance predicted from the same communities grown on each of the relevant single nutrients (S, A). Each family specializes equally on its preferred nutrient ($q_S = q_A = 0.9$) as in previous work [Marsland et al., 2020b]. In Figure 4—figure supplement 4, we illustrate representative consumption matrices for different choices of q_A and q_S. (C) 22 strains were isolated from the assembled communities and their growth rates on minimal M9 media supplemented with one the 10 carbon sources were measured. qS represents the growth rate advantage of Enterobacteriaceae on sugars relative to the other dominant family (colored), while q_A represents the growth rate advantage of the other family on the acids relative to Enterobacteriaceae (Materials and methods). When q_S is positive, Enterobacteriaceae grow faster on the sugar than the other family, while when q_S is negative, Enterobacteriaceae grow more slowly on the sugar than the other family. When q_A is positive, the other family grows faster on the acid than Enterobacteriaceae, while when q_A is negative, the other family grows more slowly on the acid than Enterobacteriaceae. Each dot corresponds to a sugar-acid pair for a Enterobacteriaceae-other family pair (n = 24). The growth rate advantage of Enterobacteriaceae on sugars is significantly greater than the growth rate advantage of the other families on acids (i.e. $q_S > q_A$, mean of differences = 0.069, paired t-test, n = 24, p-value¡0.0001). (D) Here we repeat the same simulation as shown in (B), this time using different combinations of $q_A and q_S$ (0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95). Heatmap shows the mean dominance level (δ) for different combinations of q_A and q_S. When $\delta < 0$, the sugar dominates (purple); when $\delta > 0$, the acid dominates (orange).

3.6 Material and Methods

3.6.1 Null model for relative abundance

Let us consider a simple scenario of two co-cultures of species A and B growing together in two separate demes, each containing a single nutrient (labeled 1 and 2). The fractions of A and B in nutrient/deme one are $f_{A,1} = n_{A,1}/(n_{A,1} + n_{B,1})$ and $f_{B,1} = n_{B,1}/(n_{A,1} + n_{B,1})$, respectively, and similarly, the fractions of A and B in nutrient/deme two are $f_{A,2} = n_{A,2}/(n_{A,2} + n_{B,2})$ and $f_{B,2} = n_{B,2}/(n_{A,2} + n_{B,2})$ (where n is the total number of cells of species A or B). If we consider the two-deme system as a whole (i.e. if we pool together the amount of species in each nutrient/deme), the fractions of A and B in the mixture are given by: $f_{A,12} = (n_{A,1} + n_{A,2})/(n_{A,1} + n_{B,1} + n_{A,2} + n_{B,2})$ and $f_{B,12} = (n_{B,1} + n_{B,2})/(n_{A,1} + n_{B,1} + n_{A,2} + n_{B,2})$.

We can define $n_{t,1} = n_{A,1} + n_{B,1}$ and $n_{t,2} = n_{A,2} + n_{B,2}$ as the total number of cells in the nutrient demes 1 and 2, respectively. We can thus write $f_{A,12} = (n_{A,1} + n_{A,2})/(n_{t,1} + n_{t,2})$. Defining $w_1 = n_{t,1}/(n_{t,1} + n_{t,2})$ and $w_2 = n_{t,2}/(n_{t,1} + n_{t,2})$, it is straightforward to show that: $f_{A,12} = w_1 f_{A,1} + w_2 f_{A,2}$. By the same reasoning, we find that $f_{B,12} = w_1 f_{B,1} + w_2 f_{B,2}$.

3.6.2 Sample collection

Soil samples were collected from two different natural sites in West Haven (CT, USA), with sterilized equipment, and placed into sterile bottles. Once in the lab, 5 g of each soil sample were then transferred to 250 mL flasks and soaked into 50 mL of sterile $1\times$ phosphate buffer saline supplemented with 200 μ g/mL cycloheximide (Sigma, C7698) to inhibit eukaryotic growth. The soil suspension was well mixed and allowed to sit for 48 hr at room temperature. After 48 hr, samples of the supernatant solution containing the 'source' soil microbiome were used as inocula for the experiment or stored at $-80°C$ with 40% glycerol.

3.6.3 Preparation of media plates

Carbon source (CS) stock solutions (Supplementary file 1a) were prepared at 0.7 C-mol/L ($10\times$) and sterilized through 0.22 μM filters (Millipore). Carbon sources were aliquoted into 96 deep-well plates (VWR) as single CS or mixed in pairs at 1:1 (vol:vol) and stored at

-20C. The carbon sources were adjusted to equal C-molar concentrations because carbon is the main limiting factor. To keep the total amount of carbon constant across all treatments, pairs contained half the amount of each carbon source compared to their respective single CS. Synthetic minimal growth media was prepared from concentrated stocks of M9 salts, MgSO4, CaCl2, and 0.07 C-mol/L (final concentration) of single or pairs of CS. The final pH of all growth media is shown in Supplementary table 3.1.

3.6.4 Community assembly experiment

Starting inocula were obtained directly from the 'source' soil microbiome solution by inoculating 40 μL into 500 μL culture media prepared as indicated above. For each sample and carbon source, 4μL of the culture medium was dispensed into fresh media plates containing the different single or pairs of CS in quadruplicate. Bacterial cultures were allowed to grow for 48 hr at 30C in static broth in 96 deep-well plates (VWR). After 48 hr, each culture was homogenized by pipetting up and down 10 times before transferring 4 μL into 500 μL of fresh media, and cells were allowed to grow again. Cultures were passaged 10 times (˜70 generations). OD620 was measured after 48 hr growth. Samples were frozen at $-80°C$ with 40% glycerol.

3.6.5 DNA extraction, library preparation, and sequencing

Samples were centrifuged for 40 min at 3500 rpm, and the pellet was stored at $-80°C$ until DNA extraction. DNA extraction was performed with the DNeasy 96 Blood and Tissue kit for animal tissues (QIAGEN), as described in the kit protocol, including the pre-treatment step for Gram-positive bacteria. DNA concentration was quantified using the Quan-iTPicoGreen dsDNA Assay kit (Molecular Probes, Inc), and the samples were normalized to $5ng/\mu L$ before sequencing. The 16S rRNA gene amplicon library preparation and sequencing were performed by Microbiome Insights, Vancouver, Canada (https://microbiomeinsights.com/). For the library preparation, PCR was done with dual-barcoded primers [Kozich et al., 2013], targeting the 16S V4 region, and the PCR were cleaned up and normalized using the high-throughput SequalPrep 96-well Plate Kit. Samples were sequenced on the Illumina MiSeq using the 300 bp paired-end kit v3.chemistry.

3.6.6 Taxonomy assignment

The taxonomy assignment was performed as described in previous work (Estrela et al., 2020). Following sequencing, the raw sequencing reads were processed, including demultiplexing and removing the barcodes, indexes, and primers, using QIIME (version1.9, [Caporaso et al., 2010]), generating fastq files with the forward and reverse reads. DADA2 (version 1.6.0) was then used to infer ESVs [Callahan et al., 2016]. Briefly, the forward and reverse reads were trimmed at position 240 and 160, respectively, and then merged with a minimum overlap of 100 bp. All other parameters were set to the DADA2 default values. Chimeras were removed using the 'consensus' method in DADA2. The taxonomy of each ESV was then assigned using the naïve Bayesian classifier method [Wang et al., 2007] and the Silva reference database version [Quast et al., 2013] as described in DADA2. The analysis was performed on samples rarefied to 10,779 reads.

3.6.7 Quantification of total abundances, interactions, and dominance

We used OD620 after the 48 hr growth cycle as a proxy for total population size (community biomass) (Supplementary Figure 3.5). The predicted relative abundance of species i in a mix of nutrients 1 and 2 was then calculated as $f_{i,12}(null) = w_1 f_{i,1} + w_2 f_{i,2}$ where $f_{i,1}$ and $f_{i,2}$ are the relative abundances of i in nutrients 1 and 2, respectively, and $w_1 = (OD620_1/(OD620_1 + OD620_2))$ and $w2 = (OD620_2/(OD620_1 + OD620_2))$. In Figures 3.2 and 3.3D,Supplementary Figures 3.3 and 3.4, $f_{i,12}(null)$ is calculated as the mean of the two single carbon source-replicate pairwise combinations (N = 16). Pearson's R was calculated using the R function 'cor.test' from the 'stats' package, and the RMSE was calculated using the 'rmse' function from the 'Metrics' package.

To determine whether an interaction between nutrients exists (i.e. $\epsilon \neq f_{i,12} - f_{i,12}(null)$), we assess whether the abundance observed in the carbon source mixture is significantly greater or lower than the abundance predicted by the null additive model (i.e. $\epsilon > 0$ or $\epsilon < 0$, respectively) (one-tailed paired t-test). More specifically, considering the two inocula and four replicates per carbon source, the family-level analysis was done as follow. For each carbon source pair and inoculum, four predicted pairs are formed by randomly pairing one replicate of each carbon source. Four unique observed vs predicted pairs are then randomly formed from the 64 possible combinations (i.e. from the four single nutrient 1 $\times$ four single nutrient 2 $\times$ four mixed nutrient 12). Up to this point, the pairs are formed for each inoculum separately, in other words, there is no cross-inocula pairing. Once all N=8

pairs are formed (i.e. N = 4 pairs per inoculum), they are pooled to perform the one-tailed paired t-test. The N = 8 pairs are then randomly permuted 1000 times, determining the t-statistic for each permutation. We establish a 95% confidence threshold for the t-statistic. The effect observed is statistically significant (i.e. an interaction exists) if a significant difference is found in more than 95% of the permuted pairs. At the genus level, the analysis was performed in a similar way, except that the two inocula were kept separately. This is because, compared to families, the likelihood that genera that are sampled in one of the inocula are sampled in the other inoculum is much lower.

Once an interaction has been identified (i.e. $|\epsilon| > 0$), we can determine the type of interaction formed (Figure 3.3A). Synergy and antagonism (which are forms of super-dominance) occur when $f_{i,12} > max(f_{i,1}, f_{i,2})$ and $f_{i,12} < min(f_{i,1}, f_{i,2})$, respectively, while dominance occurs when $min(f_{i,1}, f_{i,2}) <= f_{i,12} <= max(f_{i,1}, f_{i,2})$ (Welch two sample t-test, $p < 0.05$). When $\epsilon > 0$, the nutrient with greater abundance dominates; when $\epsilon < 0$, the nutrient with lower abundance dominates. For visualization purposes, we developed a dominance index (ϵ). The dominance index for the sugar–acid pairs is written as $\delta_i = -|\epsilon_{12}|$ when the sugar dominates and as $\delta_i = |\epsilon_{12}|$ when the acid dominates. The dominance index for the sugar–sugar and acid–acid pairs is written as $\delta_i = -|\epsilon_{12}|$ when the focal carbon source (glucose or succinate) dominates and as $\delta_i = |\epsilon_{12}|$ when the additional carbon source dominates.

3.6.8 Isolation of Strains

Several communities (transfer 10) from different inocula and carbon sources were plated on chromogenic agar (HiCrome Universal differential Medium, Sigma) and grown for 48 hr at $30°C$. Single colonies exhibiting distinct morphologies and/or colours were picked, streaked a second time on fresh chromogenic agar plates for purity, and grown for 48 hr at $30°C$. A single colony was then picked from each plate and grown into Tryptic Soy Broth (TSB) for 48 hr at $30°C$. The single-strain cultures were stored with 40% glycerol at $-80°C$. The isolated strains were sent for full-length 16S rRNA Sanger sequencing (Genewiz), and their taxonomy was assigned using the online RDP naïve Bayesian rRNA classifier version 2.11.

3.6.9 Growth rate estimation

Twenty-two isolated strains belonging to the four dominant families, namely Enterobacteriaceae (7), Pseudomonadaceae (5), Moraxellaceae (6), and Rhizobiaceae (4) (Supplementary table 3.2), were streaked from frozen stock on chromogenic agar plates and grown for 48 hr at $30°C$. For each strain, a single colony was pre-cultured in 500 μL TSB in a deepwell plate for 24 hr at $30°C$. Each strain was then acclimated into the 10 single carbon sources (glucose, fructose, cellobiose, ribose, glycerol, succinate, fumarate, benzoate, glutamine, and glycine). For this, 2 μL of the grown pre-culture was inoculated into 500 μL of fresh minimal media with each carbon source at a concentration of 0.07 C-mol/l and grown for 48 hr at $30°C$. The growth curve assay was then performed in a 384-well plate by inoculating 1 μL of the grown isolate culture on 100 μL of fresh media of the same carbon source as for the acclimation step (three to four replicates each). OD620 was read every 10 min for 40 hr at $30°C$. The average growth rate of each strain in each carbon source was calculated as $r_{avg} = log_2(N_f/N_i)/(t_f - t_i)$ where N_f is the OD at 16 hr (ie. t_f) and N_i is the OD at 0.5 hr (i.e. t_i).

3.6.10 Growth rate asymmetry calculation

The growth rate asymmetry on sugars (q_S) is calculated as $q_S = r_{avg}(E, S) - r_{avg}(O, S)$ where $r_{avg}(E, S)$ is the mean average growth rate of Enterobacteriaceae on the sugar S, and $r_{avg}(O, S)$ is the mean average growth rate of one of the other dominant families (i.e. Pseudomonadaceae, Moraxellaceae, or Rhizobiaceae) on S. The growth rate asymmetry on organic acids (q_A) is calculated as $q_A = r_{avg}(E, S) - r_{avg}(O, S)$ where $r_{avg}(O, A)$ is the mean average growth rate of one of the other dominant families (i.e. Pseudomonadaceae, Moraxellaceae, or Rhizobiaceae) on the organic acid A, and $r_{avg}(E, A)$ is the mean average growth rate of Enterobacteriaceae on A.

3.6.11 Microbial consumer-resource model

Microbial community assembly is modeled using the Microbial Consumer Resource Model (MiCRM), with simulations implemented using *Community Simulator*, a freely available Python package [Marsland et al., 2020b]. This model has been outlined extensively in previous work and has been shown to qualitatively reproduce ecological patterns across both natural [Goldford et al., 2018] and laboratory [Marsland et al., 2020a] microbiomes. Here we describe the exact equations simulated and parameters used in this paper. A more

general description of this model is given elsewhere [Marsland et al., 2019, 2020b]. Our MiCRM simulations model the abundance N_i of n species and the abundance R_α of M resources in a well-mixed chemostat-like ecosystem with continuous resource flow. We focus on continuous resource flow for simplicity and because previous work has shown that the major qualitative features of the MiCRM are unaffected by periodic resource supply (as was the case in our experiments) [Marsland et al., 2020a]. Species interact by uptake and release of resources into their environment. The dynamics of the system are governed by the following set of ordinary differential equations:

$$\frac{dN_i}{dt} = N_i \sum_\alpha (1-l) R_\alpha c_{i\alpha} - m \tag{3.1}$$

$$\frac{dR_\alpha}{dt} = \frac{R_\alpha^0 - R_\alpha}{\tau} - \sum_j N_j R_\alpha C_{j\alpha} + \sum_{j,\beta} N_j D_{\alpha,\beta} R_\beta c_{j\beta} l \tag{3.2}$$

Here $c_{i\alpha}$ is the uptake rate of resource α by species i, m is the minimal energy requirement for maintenance of species i, τ is the timescale for supply of external resources, R_α^0 is the abundance of resource α supplied (i.e. the abundance in the media), l is the fraction of resource secreted as by-product, and $D_{\alpha,\beta}$ is the fraction of resource α secreted as by-product of β. In line with previous work, the following parameters are kept constant for all simulations $\tau = 1$, $m = 1$, and $l = 0.5$ [Goldford et al., 2018, Marsland et al., 2020a].

In the MiCRM, by-product production is encoded in the metabolic matrix D where each element $D_{\alpha,\beta}$ specifies the fraction of resource α secreted as by-product β. As in previous work, each column β in $D_{\alpha,\beta}$ is sampled from a Dirichlet distribution with concentration parameter $D_{\alpha,\beta} = 1/(sM)$ where $s = 0.3$ is a parameter that tunes the sparsity of the underlying metabolic network. The Dirichlet distribution ensures that each column sums to one so that the total secretion flux does not exceed the input flux. For simplicity we used a fixed concentration parameter and so are not assuming any underlying metabolic structure. The MiCRM also assumes that all species have the same D matrix, that is when growing on the same resource each species releases the same metabolic by-products.

In our simulations, species differ solely in the uptake rate for different resources $c_{i,\alpha}$ where i is the species and α is the resource. Taxonomic specialization is introduced in the form of two families F_A and F_S that each have a preference for one of two resource classes A and S, respectively. Each $c_{i\alpha}$ is sampled from a gamma distribution (to ensure positivity) whose mean $< c_{i\alpha} >$ and variance $var(c_{i\alpha})$) depends on the family of i and the resource class of α. This means that all species are capable of metabolizing all resources.

Specifically:

$$\langle c_{i\alpha} \rangle = \begin{cases} \frac{\mu_c}{M}(1+q_A) \text{ if } i \in F_A \text{ and } \alpha \in A \\ \frac{\mu_c}{M}(1-q_A) \text{ if } i \in F_S \text{ and } \alpha \in A \\ \frac{\mu_c}{M}(1+q_S) \text{ if } i \in F_S \text{ and } \alpha \in S \\ \frac{\mu_c}{M}(1-q_S) \text{ if } i \in F_A \text{ and } \alpha \in S \end{cases}$$

and

$$var\left(c_{i\alpha}\right) = \begin{cases} \frac{\sigma_c^2}{M}(1+q_A) \text{ if } i \in F_A \text{ and } \alpha \in A \\ \frac{\sigma_c^2}{M}(1-q_A) \text{ if } i \in F_S \text{ and } \alpha \in A \\ \frac{\sigma_c^2}{M}(1+q_S) \text{ if } i \in F_S \text{ and } \alpha \in S \\ \frac{\sigma_c^2}{M}(1-q_S) \text{ if } i \in F_A \text{ and } \alpha \in S \end{cases}$$

$\mu_c = 10$ determines the overall mean uptake rate and $\sigma_c^2 = 3$ determines overall variance in uptake rate (these parameters are the default value in the Community Simulator package). Parameters q_A and q_S tune the relative advantage each specialist family has on its preferred resource. When $q_A = 1$, only F_A consumes resources in A whereas when $q_A = 0$ both families have equal access to resources in A. Conversely, when $q_S = 1$, only F_S consumes resources in S whereas when $q_S = 0$ both families have equal access to resources in S.

For each simulation we consider 200 species (100 per family). Each community in one simulation is seeded with all 200 species. This means that there is no stochasticity in colonization (though see Supplementary Figure 3.11 where this assumption is relaxed). We choose 200 species as this is within the range of the number of ESVs in a typical inoculum for our experiments (110-1290 ESVs, reported in Goldford et al., 2018). The initial abundances are all set to 1 for simplicity. In line with our experiments, either one or two resources are supplied in the media and the rest are generated as metabolic by-products. For simulations with a single supplied resource, $R_0^\alpha = 1000$ if α is the supplied

resource and 0 otherwise. For simulations with two supplied resources, $R_0^\alpha = 500$ for each supplied resource and 0 otherwise. This ensures that the total amount of resources is kept constant as in our experiments. In total, we consider 20 resources in each simulation (with 10 resources in each resource class [A or S]) as this gives us communities with 7 ± 2 species (mean $\pm$ SD) at equilibrium, which is comparable to the diversity of our experimental communities.

In line with our experiments, each simulation consisted of three types of mixed-resource environments (one with two supplied resources in class R_A, one with two supplied resources in class R_S and one with one resource in class R_A and one resource in class R_S). We also included all four single resource environments needed to predict the mixes (i.e. two with the resources in class R_A and two with the resources in class R_S). Therefore, each simulation consisted of seven communities each in a different environment and all seeded with the same initial set of 200 species. The equilibrium for all seven communities was found using the SteadyState function in *Community Simulator* [Marsland et al., 2020b]. Failed runs where the SteadyState function returned an error were removed from our analysis. In addition, for each simulation we tested that the SteadyState algorithm had truly converged to an equilibrium using the same approach as in Marsland et al., 2020a and removed all non-convergent runs (defined as a run for which $|dln(N_i/dt)| > 10^{-5}$). Including these runs would not have qualitatively changed our results.

In the raw numerical output of the run, all species have non-zero abundances due to limits in numerical precision. A species was considered extinct if its abundance was less than 10^-6, which was set by looking at a histogram of the raw output of our simulations. Once the extinct species were removed, we predicted the relative abundance of each species i in the mixture of nutrients using the same approach that had been used for the experimental data. To obtain a statistically robust sample size, we repeated this procedure for 100 replicate simulations, resampling all randomly generated parameters in each simulation (i.e. resampling all $c_{i\alpha}$ and $D_{\alpha,\beta}$ as described above).

3.6.12 Flux balance analysis

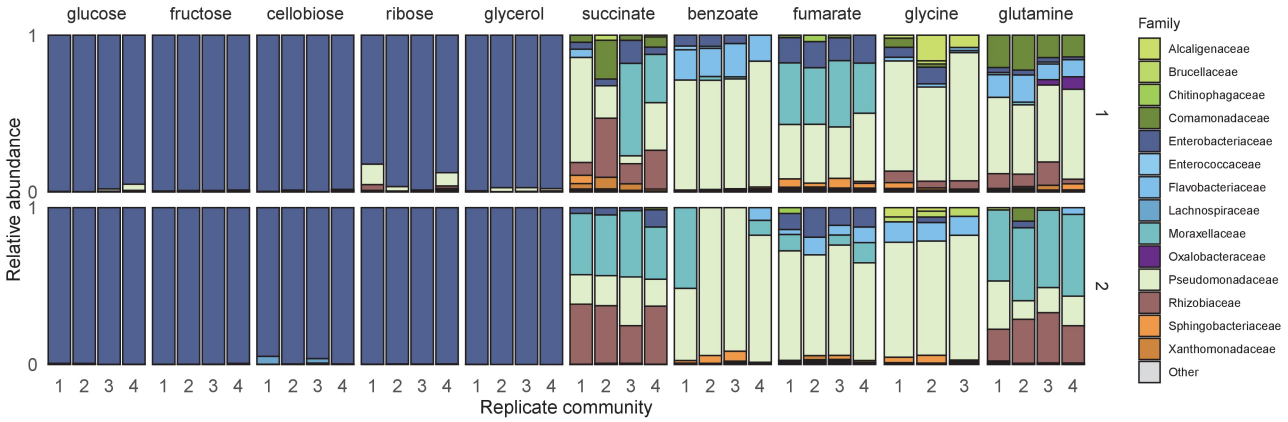
We use FBA to estimate whether the different carbon sources were likely to result in large differences in oxygen demand. FBA is a widely used constraint-based modelling approach that allows us to predict metabolic fluxes through a stoichiometric metabolic network (assuming optimal growth and that cells are in a steady state) [Orth et al., 2010]. For this analysis, we used a modified version of iJO1366 (see below), a high-quality genome scale-

metabolic network of E. coli [Orth et al., 2011]. FBA simulations were performed using the COBRApy Package [Ebrahim et al., 2013b]. We simulated the growth of E. coli on minimal synthetic media in aerobic conditions containing one of the 10 carbon sources used in our experiments (Figure 3.2A). These simulations were used to estimate the number of O2 molecules that would be consumed per carbon atom when growing on each of the 10 carbon sources (Supplementary Figure 3.14). Except for benzoate, we found that all of the carbon sources exhibited similar predicted oxygen demands (0.25–0.34 O2/C). This does not rule out the possibility that the kinetics of growth and O2 uptake may contribute to increased O2 depletion in one carbon source compared to another, nor that the different taxa selected for by the different carbon sources might display differences in O2 uptake. Nonetheless, they do suggest that differences in oxygen stoichiometry are unlikely to be the main mechanism for dominance across all carbon sources.

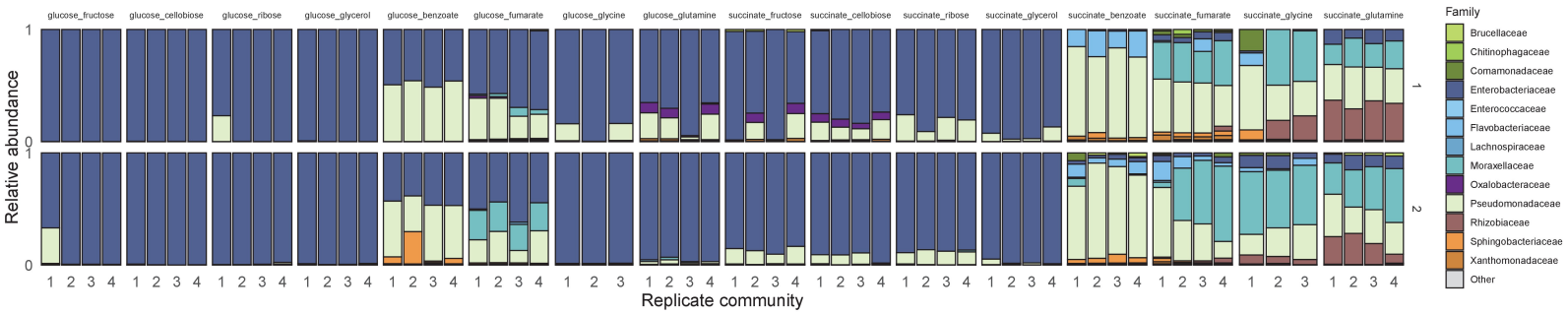
For these simulations, all inorganic compounds were assumed to be in excess and their exchange fluxes were unbounded by setting to an arbitrarily large negative value (-1000 mmol/gDWh). These compounds are as follows: ca2_e, cbl1_e, cl_e, co2_e, cobalt2_e, cu2_e, fe2_e, fe3_e, h_e, h2o_e, k_e, mg2_e, mn2_e, mobd_e, na1_e, nh4_e, ni2_e, pi_e, sel_e, slnt_e, so4_e, tungs_e, zn2_e, and o2_e. To estimate the optimal oxygen consumption per mole of carbon consumed, the exchange flux for each of the 10 carbon sources was set to -1 cmol/gDWh. We set the lower bound on ATPM maintenance to 0 as we wanted to estimate the O2/C when resources were in excess and so the effects of growth independent maintenance would be negligible. Similar results can be obtained using the default ATPM lower bound in the published model and setting a higher lower bound on the carbon uptake flux (such as the -60 cmol/gDWh typically used for E. coli on glucose) [Harcombe et al., 2014]. The biomass reaction (BIOMASS_Ec_iJO1366_core_53p95M) was used as the objective function.

3.7 Supplementary Material

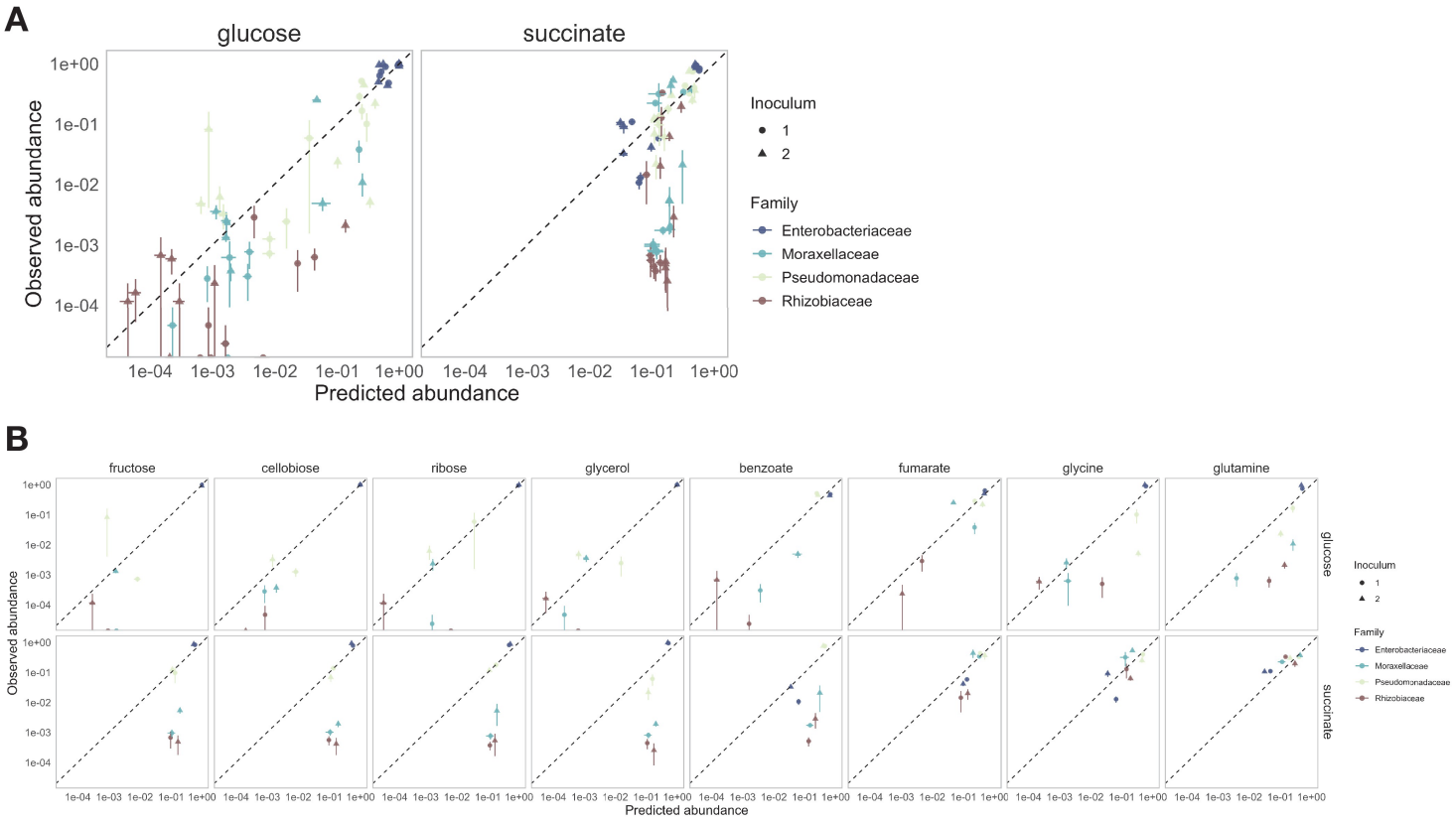
3.7.1 Supplementary Figures



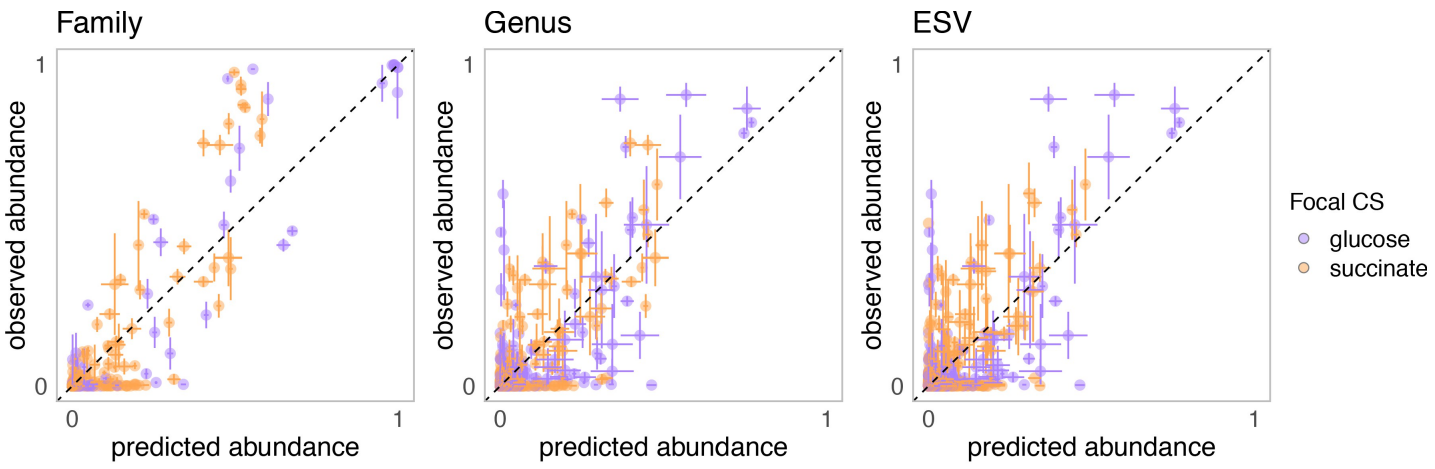
Supplementary Figure 3.1: Community assembly in a single carbon source: Two soil samples were inoculated in minimal M9 medium supplemented with a single carbon source (three or four replicates each) and propagated into fresh media every 48 hr for 10 transfers (Materials and methods). Shown is the family-level taxonomic composition at Transfer 10 for inoculum 1 (top) and inoculum 2 (bottom). Families with a relative abundance lower than 0.01 are shown as 'Other'.



Supplementary Figure 3.2: Community assembly in a mixture of two carbon sources: Two soil samples were inoculated in minimal M9 medium supplemented with two carbon sources (glucose or succinate + another carbon source), and propagated into fresh media every 48 hr for 10 transfers (Materials and methods). There are three/four replicates per carbon source pair. Shown is the family-level taxonomic composition at Transfer 10 for the two inocula. Families with a relative abundance lower than 0.01 are shown as 'Other'.

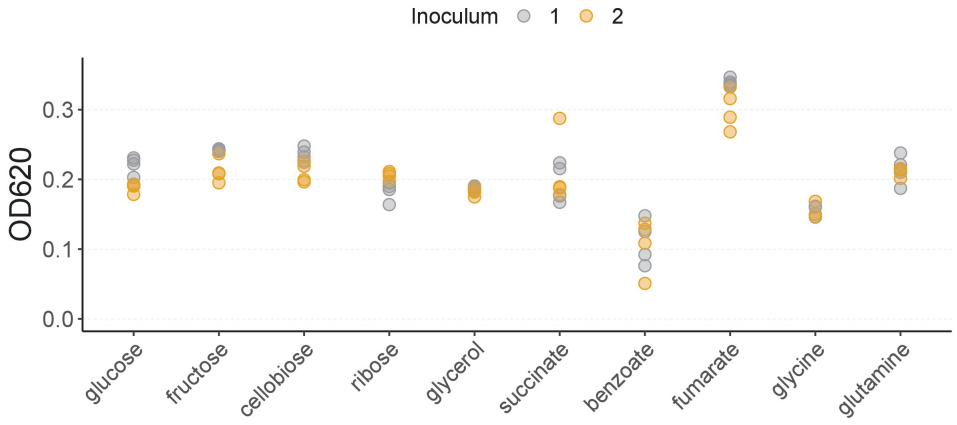


Supplementary Figure 3.3: Systematic deviations from the null (additive) prediction reveal interactions between nutrients: Shown is the same data as in Figure 3.2B (A) and Figure 3.2C (B) but displayed on a log–log scale so the datapoints at lower relative abundance are easier to visualize.

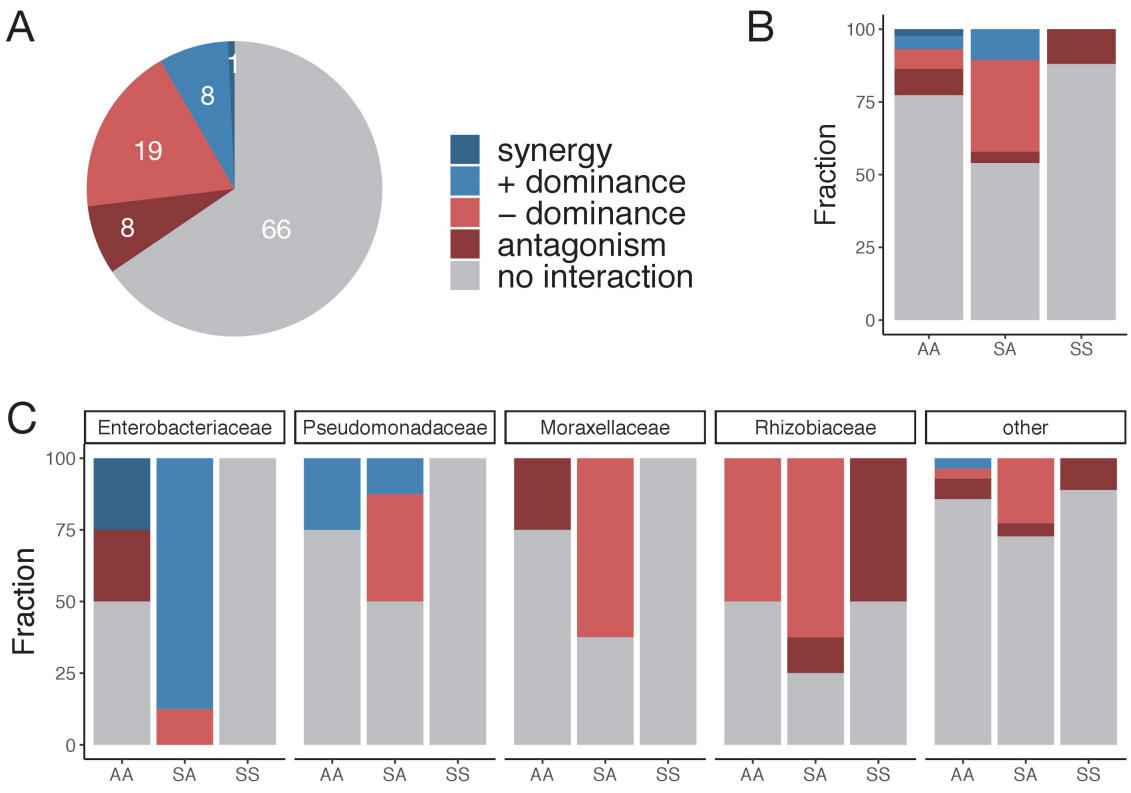


Level	Focal CS	Pearson's R	RMSE
Family	glucose	0.948	0.073
Family	succinate	0.871	0.094
Genus	glucose	0.782	0.082
Genus	succinate	0.765	0.066
ESV	glucose	0.773	0.069
ESV	succinate	0.674	0.056

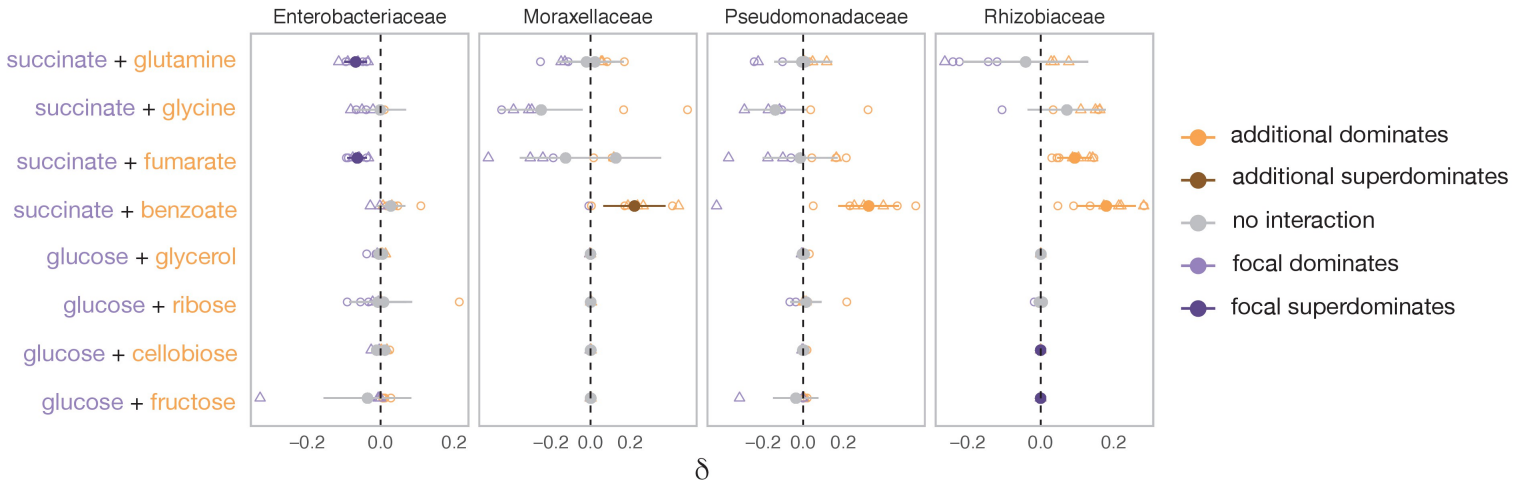
Supplementary Figure 3.4: Comparison of the observed relative abundance and abundance predicted by the null model. Shown is the observed vs predicted abundance for different taxonomic levels and focal carbon source (CS) (mean ± SE). Table shows the Pearson's R and RMSE for each family-focal carbon source combination.



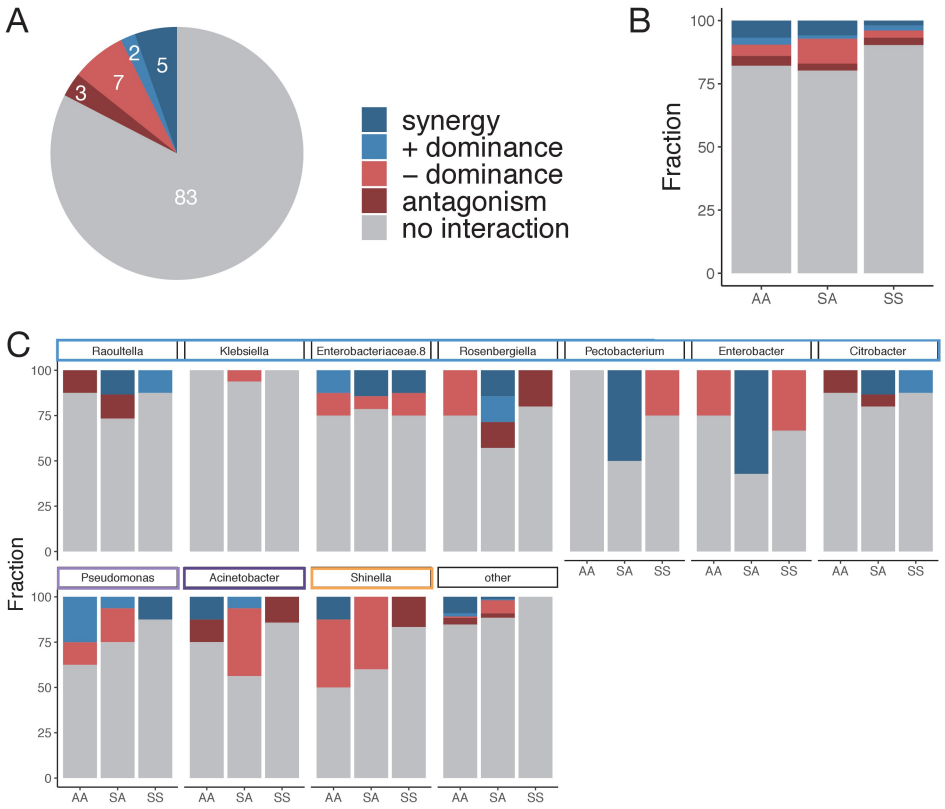
Supplementary Figure 3.5: Community yield in each single carbon source:Total community biomass (OD620) at the end of the 48 hr incubation period at Transfer 10. There are four biological replicates per carbon source per inoculum, except for glycine with three replicates.



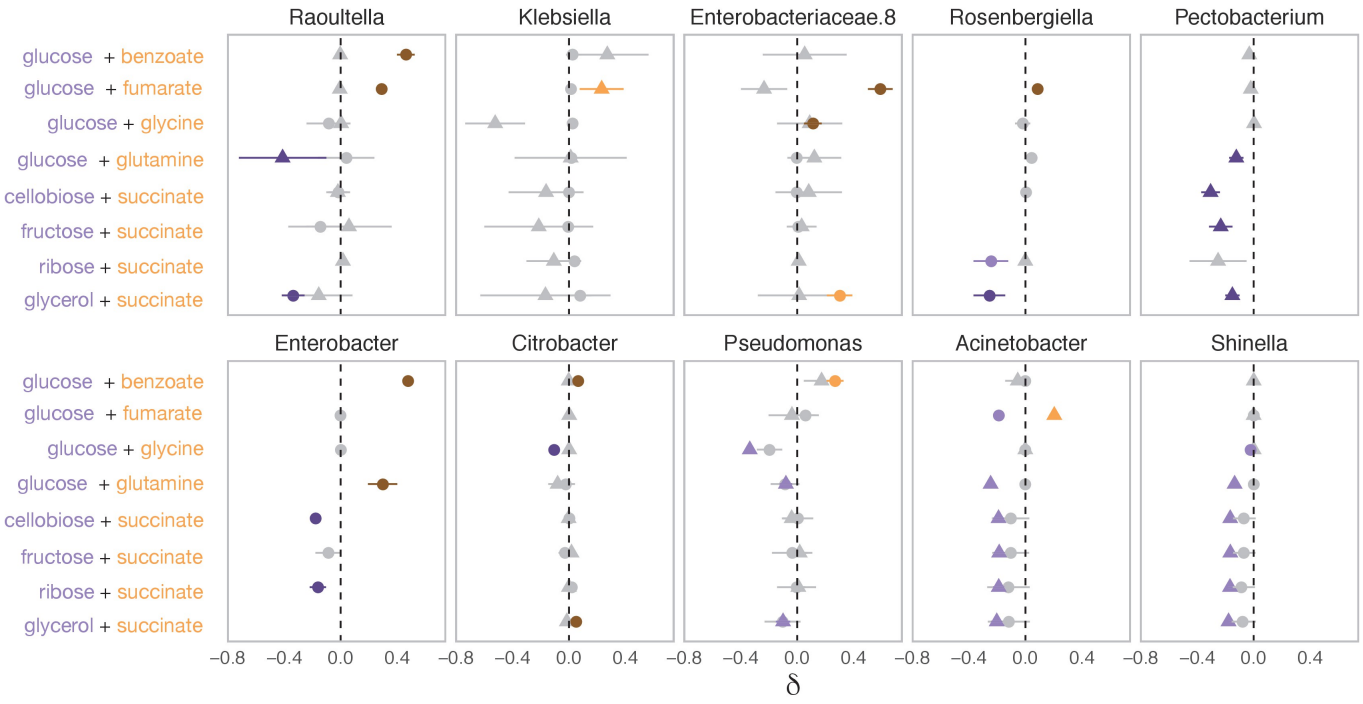
Supplementary Figure 3.6: Dominance is the most common type of nutrient interaction, especially in the sugar–acid mixtures: (A) Interaction type for each pair of carbon source and family. An interaction between nutrients occurs when the abundance in the mixture is significantly greater or lower than predicted by the null additive model (one-tailed paired t-test, p¡0.05 based on 1000 permutations; only cases where N $\geq$ 6 unique pairs are considered) (Materials and methods). Multiple types of nutrient interaction are possible: dominance, synergy, and antagonism. Synergy (antagonism) occurs when the abundance in the mixture is greater (lower) than the abundances in any of the single nutrients independently (Welch two sample t-test, $p < 0.05$) (Materials and methods). Dominance occurs when the abundance in the mixture is closer or similar to the abundance in one of the singles. (B) Interaction type by carbon source pair type. AA: mixture of two acids; SS: mixture of two sugars; and SA: mixture of a sugar and an acid. (C) Interaction type shown for the four most abundant families and 'other' families grouped together.



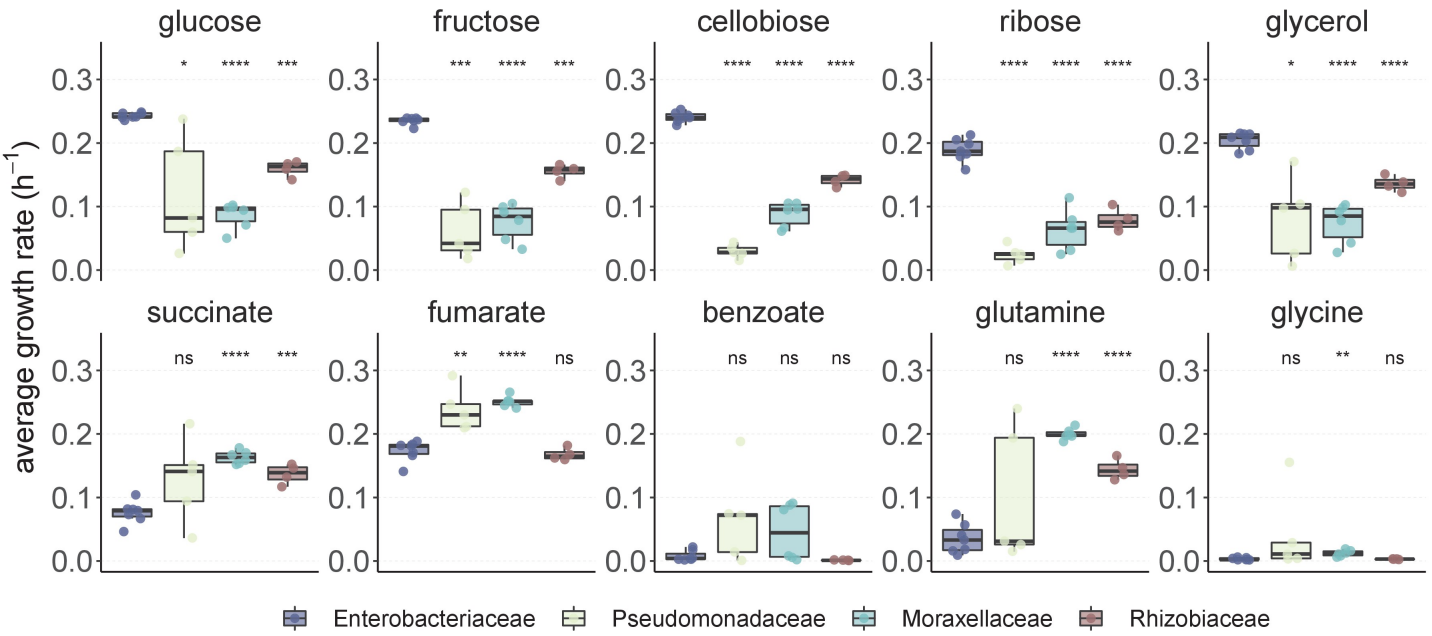
Supplementary Figure 3.7: Family-level dominance for mixtures of acid–acid and sugar–sugar: For each carbon source pair, the filled circles show the mean ± SD of N = 8 unique replicates (two inocula x four replicates each), and the open symbols show all eight replicates individually (except for glycine pairs where N = 6). The different shapes correspond to different inocula. When $\delta < 0$, the focal carbon source (succinate or glucose) dominates. When $\delta > 0$, the additional carbon source dominates (Materials and methods). Orange or purple corresponds to cases where nutrients interact, in which case there is dominance (or super-dominance). An interaction occurs when the abundance observed in the mixture is significantly greater ($\epsilon > 0$) or lower ($\epsilon < 0$) than predicted by the null additive model (one-tailed paired t-test, $p < 0.05$, based on 1000 permutations; see Materials and methods). Gray corresponds to cases where nutrients do not interact or dominance is undefined because one carbon source dominates in one of the inocula and the paired carbon source dominates in the other inocula (in which case δ is shown as both $-\delta$ and $+\delta$). Lighter orange or purple indicates dominance while darker orange or purple indicates super-dominance (synergy or antagonism) (Materials and methods). Only the four most dominant families are shown.



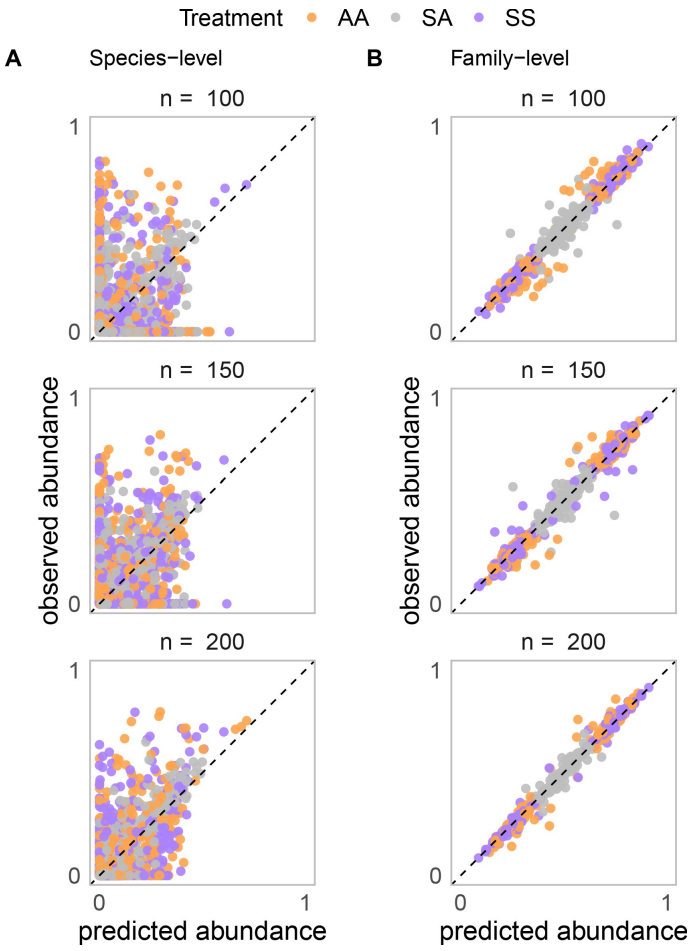
Supplementary Figure 3.8: Patterns of nutrient interaction at the genus level: (A) Multiple types of nutrient interactions are possible, including dominance, synergy, and antagonism (Figure 3A). An interaction occurs when ϵ is significantly greater or lower than 0 (one-sided paired t-test, $p < 0.05$ based on 1000 permutations, Materials and methods). Inocula are considered separately, and only cases where $N \geq 3$ unique pairs are considered. Synergy (antagonism) occurs when the abundance in the mixture is greater (lower) than the abundances in any of the singles separately. Dominance occurs when the abundance in the mixture is closer or similar to one of the single abundances but not above or below any of the single abundances independently (Materials and methods). (B) Interaction type by carbon source pair type. AA: mixture of two acids; SS: mixture of two sugars; and SA: mixture of a sugar and an acid. (C) Interaction type is broken down by the 10 most abundant genera spanning the Enterobacteriaceae family (blue), Pseudomonadaceae (light purple), Moraxellaceae (dark purple), and Rhizobiaceae (orange) families. Note that Enterobacteriaceae.8 is a non-identified genus belonging to the Enterobacteriaceae family. The other genera are grouped together and shown as 'other'.



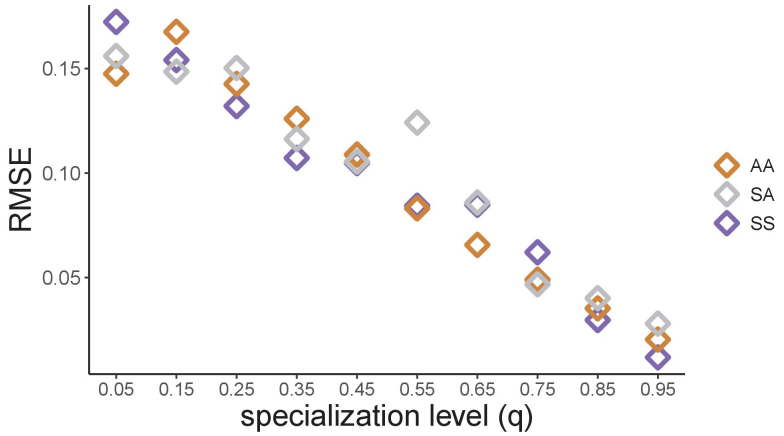
Supplementary Figure 3.9: The systematic dominance of sugars observed at the family level does not apply to the genus level: To determine the genus-level dominance, the two inocula are considered separately (different shapes) as the genera that are sampled in one inocula may not be sampled in the other inocula. Purple indicates that the sugar dominates while orange indicates that the acid dominates. Lighter purple and orange indicate dominance, while darker purple and orange indicate super-dominance (synergy or antagonism) (Materials and methods). No interaction is shown in gray. An interaction occurs when ϵ is significantly greater or lower than 0 (one-sided paired t-test, $p < 0.05$ based on 1000 permutations, Materials and methods). Shown are the 10 most abundant genera (mean $\pm$ SD). Note that Enterobacteriaceae.8 is a non-identified genus of the Enterobacteriaceae family.



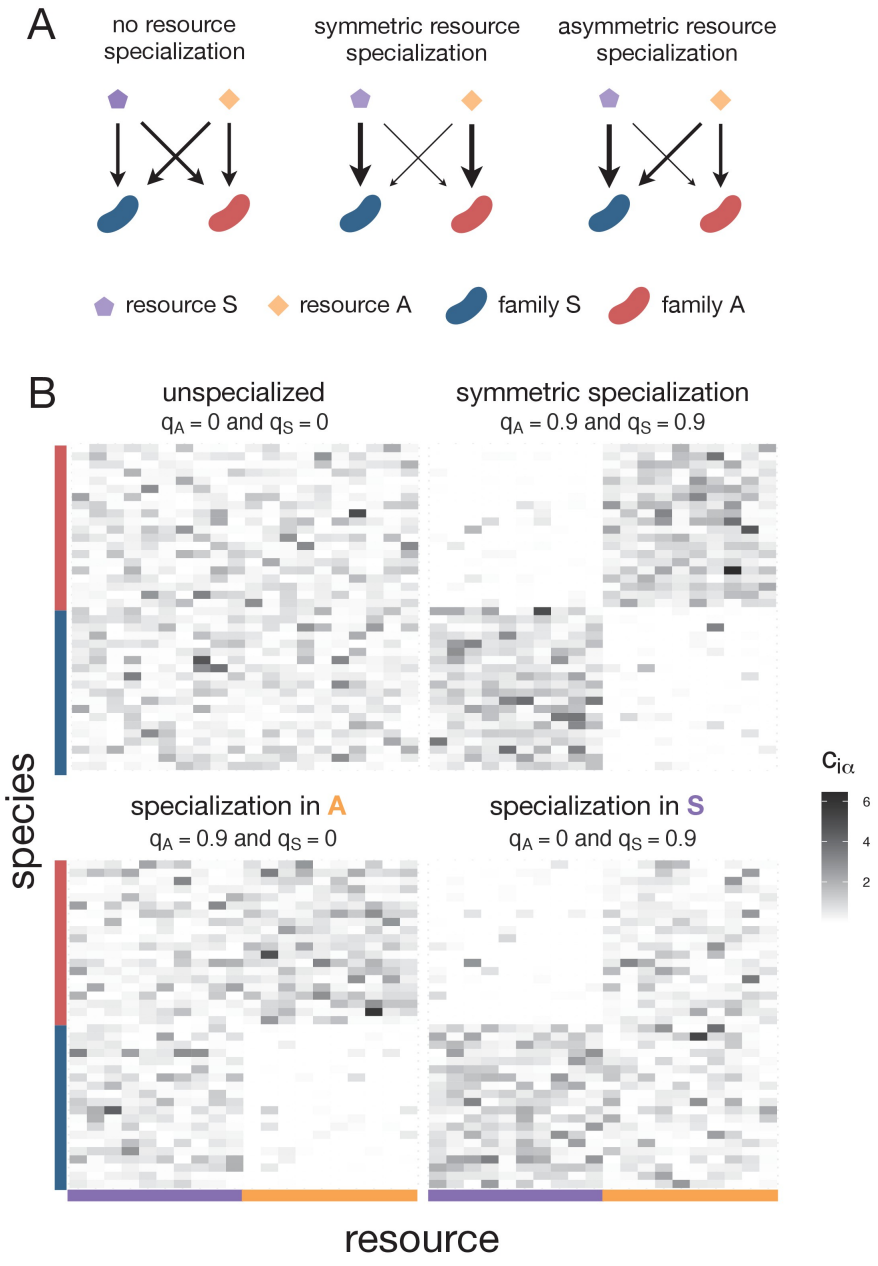
Supplementary Figure 3.10: Enterobacteriaceae generally have a strong growth advantage in sugars :Twenty-two strains belonging to the four dominant families, namely Enterobacteriaceae (7), Pseudomonadaceae (5), Moraxellaceae (6), and Rhizobiaceae (4) were isolated from the self-assembled communities and their growth rate on the 10 carbon sources was measured (Materials and methods, Supplementary file 3.1b). The average growth rate is measured as the mean cell divisions from 0.5 hr to 16 hr of growth (three or four replicates each) (Materials and methods). Thus, this approach takes into account both lag and growth rate, two growth traits that are important in determining the competitive ability of a strain. We use the first 16 hr of growth rather than a longer time window to better assess growth rate on the supplied nutrient and avoid potential artifacts from growth on secretions. Significance level (p-value) is measured by comparing the average growth rate between Enterobacteriaceae (reference) and each other family (paired t-test, $****p < 0.0001$; $***p < 0.001$; $**p < 0.01$; $*p < 0.1$).



Supplementary Figure 3.11: Stochastic colonization has no qualitative effect on the pattern of additivity found using a Microbial Consumer Resource Model: Relative abundance of each species (A) or species grouped by family (B) in simulated communities grown in a mixture of nutrients plotted against the predicted relative abundance from simulated communities grown in single nutrients assuming that nutrients act independently (Materials and methods). Communities are colonized with n species, randomly sampled from a regional pool of 200 species, while keeping the number of families constant. When n = 200, all species are sampled. Decreasing n reduces the initial species variability of the community and also introduces stochastic colonization through the random sampling of the regional species pool. For each n, the result of 100 simulations for communities grown in three carbon source pairs is shown (1 SS pair, 1 AA pair, and 1 SA pair).

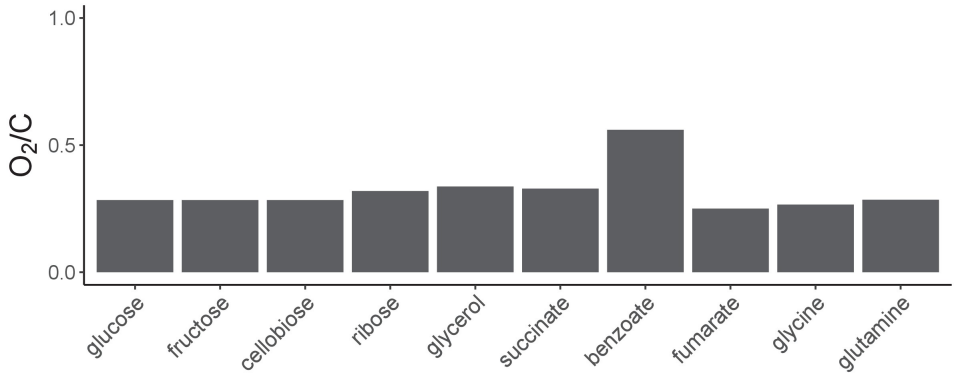


Supplementary Figure 3.12: The predictive accuracy of the null model decreases with lower levels of resource specialization. In Figure 3.4B (right-hand panel), we performed consumer-resource model simulations and plotted the observed and predicted relative abundance of each family in 300 communities grown on a different pair of nutrients (100 AA, 100 SS, and 100 SA). In those simulations, each family is specialized on its preferred nutrient ($q_S = q_A = 0.9$). Here, we repeat these exact simulations for different degrees of resource specialization ($q_S = q_A \in [0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95]$). When $q_S = q_A = 0.05$, the two families are largely unspecialized whereas when $q_S = q_A = 0.95$ both families are largely specialized (Figure Supplementary Figure. 3.13). The predictive accuracy of the null model is quantified using the RMSE calculated across n = 100 communities.



Supplementary Figure 3.13: Consumption matrices for different patterns of nutrient preference between families used in the consumer-resource model simulations

Supplementary Figure 3.13: (A) The schematics illustrate different scenarios of nutrient preference for two families (FS and FA) and two resource classes (RS and RA). Without resource specialization, FS and FA have equal access to RS and RA. With symmetric specialization, each family prefers its own resource class with the same strength. With asymmetric specialization, one family (FS) has better access to its own resource class (RS) relative to that of the other family (FA) on its own resource class (RA). (B) Consumption matrices for two families (FA and FS coloured in orange and purple respectively) in two resource classes (RS and RA). Each row corresponds to a different species (for visualization purposes we show 30 species per family) and each column corresponds to a different nutrient within a resource class (10 nutrients per resource class). The value $c_{i\alpha}$ corresponds to the uptake rate of species i in nutrient α. Four nutrient preference patterns are illustrated. Without family-level nutrient preference (specialization), species from the two families have equal access to resources in A ($q_A = 0$) and resources in S ($q_S = 0$). When each family has a strong and quantitatively similar preference for its own resource class, there is symmetric specialization ($q_A = q_S > 0$). When family FA has a strong preference for its own resource class A but both families have equal access to resources in S, then $q_A > 0$ and $q_S = 0$. When family FS has a strong preference for its own resource class S but both families have equal access to resources in A, then $q_S > 0$ and $q_A = 0$.



Supplementary Figure 3.14: Oxygen demands are similar across the different carbon sources: We carried out flux-balance analysis using a genome-scale metabolic model of E. coli to determine if different carbon sources are likely to exhibit large differences in oxygen demand (Materials and methods). On the y-axis, we plot the oxygen exchange flux/carbon flux for each of the 10 carbon sources used in this study.

3.7.2 Supplementary Tables

Carbon source	Supplier	Reference	pH (in M9)	Final concentration (mM)
D-Glucose	VWR	0188-500	6.83	11.67
D-Cellobiose	Sigma	22150-10G	6.84	5.83
D-Fructose	Acros Organics	161355000	6.79	11.67
D-Ribose	Acros Organics	AC132361000	6.81	13.99
Glycerol (80%, w/v)	Teknova	G8797	6.81	23.33
Sodium Succinate hexahydrate	Alfa Aesar	419A3	6.84	17.50
Sodium hydrogen fumarate	Alfa Aesar	B24683	6.11	17.50
Sodium benzoate	Alfa Aesar	A15946	6.80	10.0
L-Glutamine 200mM (29.23 mg/mL)	Sigma	G7513-100ML	6.80	14.0
Glycine	Sigma	G7126-100G	6.82	35.0

Supplementary Table 3.1: Carbon sources used in this study

Family	Genus	Transfer_CarbonSource_Inoculum_Replicate
Enterobacteriaceae	Raoultella	T10_glucose_I1_R2
Enterobacteriaceae	Citrobacter	T10_glucose-cellobiose_I1_R1
Enterobacteriaceae	Klebsiella	T10_glucose-cellobiose_I1_R1
Enterobacteriaceae	Citrobacter	T10_succinate_I2_R1
Enterobacteriaceae	Enterobacter	T10_succinate_I2_R1
Enterobacteriaceae	Klebsiella	T10_succinate_I2_R4
Enterobacteriaceae	Raoultella	T10_glutamine_I2_R2
Moraxellaceae	Acinetobacter	T10_succinate_I2_R1
Moraxellaceae	Acinetobacter	T10_succinate_I2_R1
Moraxellaceae	Acinetobacter	T10_succinate_I2_R4
Moraxellaceae	Acinetobacter	T10_succinate_I2_R4
Moraxellaceae	Acinetobacter	T10_glutamine_I2_R2
Moraxellaceae	Acinetobacter	T10_glutamine_I2_R2
Pseudomonadaceae	Pseudomonas	T10_glutamine_I2_R3
Pseudomonadaceae	Pseudomonas	T10_ribose_I1_R1
Pseudomonadaceae	Pseudomonas	T10_benzoate_I1_R3
Pseudomonadaceae	Pseudomonas	T10_fumarate_I2_R2
Pseudomonadaceae	Pseudomonas	T10_benzoate_I2_R3
Rhizobiaceae	Rhizobium	T10_succinate_I2_R1
Rhizobiaceae	Rhizobium	T10_succinate_I2_R1
Rhizobiaceae	Rhizobium	T10_succinate_I2_R4
Rhizobiaceae	Rhizobium	T10_glutamine_I2_R2

Supplementary Table 3.2: Taxonomy of strains used in the growth rate assay and community they were isolated from.

3.8 Acknowledgements

We want to thank members of the Sanchez lab for helpful discussions. This work was supported by the National Institutes of Health through grant 1R35 GM133467-01, and by a Packard Foundation Fellowship to AS.

Chapter 4

Predicting Microbial Community Assembly across Environments

4.1 Abstract

Predicting the effects of the metabolic environment on the taxonomic composition of microbial communities is a major goal for microbiome research. Here we combine an enrichment community approach with consumer resource modelling to show that the compositions of microbial communities assembled in different metabolic environments is predicted by the taxonomic distributions of the metabolic traits under selection. We hypothesize that environments selecting for correlated traits will select for quantitatively similar communities, a hypothesis we confirm experimentally. Correlations in metabolic traits across environments can be predicted using genome-scale metabolic models, allowing us to predict microbial community assembly in novel environments. Our results reveal that despite the combinatorial complexity of diverse communities, the effects of selection can be predicted if we can identify conserved metabolic structures determining trait values across environments.

4.2 Introduction

A major challenge in ecology and evolution is to quantitatively predict shifts in the genetic and phenotypic composition of diverse communities in response to environmental change. This extends across scales of ecological complexity; from predicting allele fre-

quency changes within a single species to predicting changes in the functional and taxonomic composition of multi-species communities [Lässig et al., 2017, McGill et al., 2006]. In microbes there is widespread interest in predicting: the abundance of resistance alleles when populations are exposed to antibiotics [Martínez et al., 2007, Suzuki et al., 2014, Pinheiro et al., 2021] ; the prevalence of viral escape mutants in response to neutralizing antibodies [Hie et al., 2021]; and the effects of diet on gut microbiome composition [Wu et al., 2011, Faith et al., 2011, David et al., 2014a].

A large body of microbiome research has shown that the composition of natural microbial communities is strongly influenced by the metabolites present in the environment [Faith et al., 2011, David et al., 2014b,a]. Recent studies; both in natural habitats [Burke et al., 2011, Human Microbiome Project Consortium, 2012, Louca et al., 2016a,b] and in well-controlled laboratory ecosystems [Goldford et al., 2018, Bittleston et al., 2020, Estrela et al., 2022, de Jesús Astacioa et al., 2021], have shown that microbial communities subject to identical metabolic environments typically converge to taxonomic attractors at coarse phylogenetic levels despite variability in species composition. Nonetheless, quantitative rules that can predict the taxonomic compositions of communities in different metabolic environments remain elusive [Costello et al., 2012, Koskella et al., 2017]. Attempts to quantitatively predict the response of communities to well defined environmental change has only been possible in communities that involve a small number of taxa (≤ 3), and where the phenotypic properties of each community member have been extensively characterized [Harcombe et al., 2014]. Extending these predictions to even moderately diverse communities is intractable due to combinatorial explosion in the number of possible interactions, both between pairs and groups of cells.

Intriguingly similar combinatorial challenges have been encountered when trying to model other complex self assembly processes in biology, notably protein folding [Jumper et al., 2021]. Predicting the stable 3D structure of a protein from chemical interactions between amino acid residues alone has proven impossible because of the vast conformational space that needs to be explored. Despite this, predictions of protein structure have been conducted using comparative approaches that leverage the shared evolutionary history of proteins[Shindyalov et al., 1994, Marks et al., 2011]. Using homology modeling, one can predict the three dimensional structure of an unknown protein, by identifying conserved structural motifs from homologous sequences [Roy et al., 2010]. Inspired by this approach, we set out to explore whether the response of complex communities to different environments could be predicted using a comparative approach leveraging the shared evolutionary history of microbial metabolism.

4.3 Results

4.3.1 Community composition in a given environments depends on the taxonomic distributions of quantitative metabolic traits

In order to address this question one would need to study the assembly of replicate diverse multi-species communities in many different environment whose biochemical compositions was known and well-defined. To that end, we have turned to our experimental enrichment community system, where diverse microbiomes taken from natural habitats are repeatedly passaged into fresh minimal media containing a single limiting carbon source (Figure 4.1A). Following our previously established protocol (see methods) communities from 11 different starting inoculum were serially passaged and allowed to self-assembled in 7 different carbon sources) each entering central metabolism using different metabolic pathways (Figure 4.1A)(D-Glucose, Pyruvate, Citrate, Acetate, L-Glutamine, L-Phenylalanine and L-Leucine). In line with our previous work communities assembling in different environments formed diverse multi-species communities (containing 6-49 ESVs) with coarse-grain family-level compositions that where highly reproducible in each environment (Figure 4.1B). Family-level community dissimilarity (Bray-Curtis) for communities assembling in the same carbon source is significantly lower (median: 0.23) than for communities assembling in different carbon sources (median, 0.45, one-tailed Kolmogorov-Smirnov, $p < 1e\text{-}06$, Supplementary Figure 4.1) and differences in the supplied carbon source accounts for a significant fraction of the variation in the the abundance of the most cosmopolitan families (84% for the Enterobacteriaceae and 47% for the Pseudomonadaceae (see Supplementary Figure 4.2 for other families)).

In previous work we have examined communities assembling in glucose limited habitats and shown that that the family level convergence reflects an emergent metabolic-self organisation arising from the phylogenetic conservation of the metabolic traits under selection [Estrela et al., 2022]. Specifically members of the Enterobacteriacia are selected for their ability to uptake and generate quantitatively similar levels of biomass from glucose whilst secreting quantitatively similar levels of the metabolic by-product acetate. Conversely Pseudomonadaceae are selected for the ability to uptake and generate quantitatively similar levels of biomass from acetate. These family level differences in growth on glucose stem from differences in homologous metabolic pathways conserved at a coarse-phylogenetic level: Enterobacteriacia (and the closely related Aeromonadaceae) metabo-

lize glucose using the Embden-Meyerhof-Parnas (EMP) pathway which generates a higher ATP yield than the Entner-Doudoroff Pathway (ED) on which Pseudomonadaceae (and the closely related Moraxellaceae) rely [Stettner and Segrè, 2013].

Does the family-level conservation of quantitative metabolic traits extend to the non-glucose carbon sources metabolized using different metabolic pathways? Other carbon sources used in our community assembly experiment are known be metabolized using alternative metabolic pathways which show varied taxonomic distributions. For example acetate can be metabolized using either the ACS pathway or the more inefficient PtA-AckA pathway [Enjalbert et al., 2017]. Moreover in the inferred metagenomes of communities assembled in different carbon sources we observed highers levels of enzymes involved in metabolizing the corresponding resource (Supplementary Figure 4.3). Given that the core metabolic pathways associated with carbohydrate, organic acid and amino acid metabolism are often conserved [Peregrín-Alvarez et al., 2009] we hypothesized that the metabolic traits under selection would generically show family-level conservation leading family-level convergence during in community assembly .

To test this hypothesis we studied a library of Gammaproteobacteria strains that we had isolated with unique full-length 16s sequences (N = 54) (see methods). These strains covered six of the eight most common families found in our communities including Enterobacteriacia (n=34), Pseudomonadaceae (n=16), Aeromonadacae(n=1), Moraxellaceae(n=1), Comamonadaceae(n=1), and Alcaligenaceae(n=1). The reconstructed phylogenetic tree for these isolates is shown in Figure 4.1C) (see methods). We quantified the growth rate of our isolates in all 7 supplied carbon sources (methods) (Figure 4.1C) and mapped the empirical growth rates onto the phylogeny (**7 CS growth rates**). We observed significant family level conservation of growth rates across environments, which we quantified using a phylogenetic imputation approach (methods). This quantified how well one could predict the growth rate of an isolate in a given carbon source from it's relative position on the 16s rRNA phylogenetic tree (Supplementary Figure 4.4). The growth rates on all carbon source showed some degree of predictability though this varied by carbon source (i.e Pearson's $R = 0.89$ for Glucose vs $R = 0.4$ for Acetate). ESVs belonging to families displaying a faster growth rate on the supplied carbon sources were consistently found out at higher abundances in the self-assembled communities (Figure 4.1D).Moreover on all 7 carbon sources the most abundant family was the one containing the fastest growing strains (Supplementary Figure 4.5) This confirms our hypothesis that differences in community composition across environments predictably arise from differences in metabolic traits conserved at the family level.

4.3.2 Convergent community assembly depends on the conservation of by-product production and nutrient uptake capabilities

The precise abundance of a family in a community at equilibrium depends not just on its growth rate on the supplied carbon source but also on cross-feeding interactions which we have previously shown to play a major role in structuring communities assembling in minimal Glucose [Estrela et al., 2022, Goldford et al., 2018]. If the family level conservation of growth on different carbon sources (shown in Figure 4.1C) stems from the evolutionary conservation of the underlying metabolic pathways, than we should expect member of the same family to secrete similar by-products resulting in conserved cross-feeding interactions Giri et al. [2021], Oña et al. [2021].

To determine whether this was indeed the case for our isolates we systematically quantified metabolic by-product production and nutrient uptake of all 54 isolates when growing on minimal glucose. In previous work we had measured the amount of the metabolic by-products: acetate, succinate and d-lactate (methods) secreted after 16 hrs of growth for a subset of our 54 strains Estrela et al. [2022]. Here we filled any gaps in these measurements and extended them to include the amounts of secreted gluconate, and 2-ketogluconate (see methods) as these metabolites are known to be major by-products of glucose metabolism for the Pseudomonadaceae [del Castillo et al., 2007] (Supplementary Figure 4.6A-B). We find that members of the same family tend to have highly similar by-product profiles (Figure 4.2A). To explore whether similar patterns of by-product secretion would be observed for other carbon sources we used liquid-chromatography mass spectrometry (LC-MS) and analyzed the byproducts of growth of a pair of enterobacteriaceae strains and a pair of Pseudomonadaceae strains (Methods) on 5 different carbon sources over a 48hr period. On all 5 carbon sources members of the same family produced more similar metabolic-by-products than members of different families (Supplementary Figure 4.7).

Does the predictability of family level composition depend on the conservation of metabolic niche construction? To explore this question we turned to a microbial consumer resource model (MiCRM) (section 3.6.11) which we have previously shown can reproduce quantitative ecological patterns in self assembled laboratory communities Marsland et al. [2020b], Goldford et al. [2018]. We paramaterized the model using the empirically measured trait values for the 54 isolates grown on minimal glucose i.e using by product quantification and growth rate quantification on the secreted by-products (Figure 4.2A-B, Supplementary Figure 4.6). Using this model we simulated 100 communities on minimal glucose each composed of a randomly chosen subset of the 54 isolates (Figure 4.2C). The

relative abundance of the families in the simulated communities is reproducible across replicates and is quantitatively similar to that observed experimental glucose communities shown in Figure 4.1. When we repeated these simulations randomizing either metabolic-by-production across families or nutrient uptake and yield across families communities no longer converge to similar family-level compositions. These simulations thus confirm our intuition that convergent community assembly in our experimental system depends on the conservation of both by-product production and of nutrient uptake capabilities (Figure 4.2C).

4.3.3 Substrates that use overlapping metabolic pathways select for similar communities

We have shown that predictability of community assembly in a given environments depends on the taxonomic conservation of quantitative metabolic traits. Related taxa when grown in the same carbon source tend to have similar growth rates and produce similar by-products because they use the same conserved metabolic pathways (Figure 4.2). By the same logic we reasoned that the same taxa grown in different carbon sources would show more similar uptake rates,yields and by-products if the carbon source are metabolized using the same metabolic pathway. To test this hypothesis we expanded the list of environments in which grew our isolates to include 12 additional carbon sources and quantified the growth of all 54 isolates in these carbon sources (Figure 4.3A). Across all isolates some pairs of carbon sources rates showed significant positive correlations in growth rate and some pairs carbon source displayed significant negative correlations in growth rate (Figure 4.3B). Hierarchical clustering on these correlations identified two clear substrate classes, 'Glycolytic' substrate which include all the sugars and enter metabolism through Glycolysis and 'Gluconeogenic' resources which include most organic acids and which enter metabolism directly through the TCA Cycle (Figure 4.3C) Buffing et al. [2018]. To determine whether these resource classes also lead to similar by-products we performed untargeted metabolomics on the spent media of an Enterobacter and a Pseudomonas strain after 24hrs of grow in the different minimal media (4.3D). We find that the spent media on different carbon sources clustered by whether the carbon source was Glycolytic or Gluconeogenic.

Given the similarity in metabolic trait across these two substrate classes we hypothesized that communities assembling in Gluconeogenic substrate should be more similar to one another than communities assembling in Glycolytic Substrates. To test this hypothesis

we took a single community and allowed it assemble in a library of carbon sources , 21 of which were Glycolytic and 20 of which were Gluconeogenic (Figure 4.3F). Glycolytic Carbon Sources selected for communities dominated by Enterobacteriaceae whereas Gluconeogenic Carbon Sources selected for communites dominated by Pseudomonadaceae. To test whether the similarities in community composition across pairs of 'metabolically similar' carbon sources was quantitative as well as qualitative, we turned to flux balance analysis and for each carbon source calculated the optimal metabolic fluxes through a universal gram-negative bacterial model. Carbon sources that used overlapping metabolic pathways (quantified by the Euclidean distance in Metabolic Flux) selected for similar family level community compositions (Figure 4.3F). Using this quantification of metabolic similarity we turned to a machine learning approach and trained a Lasso regularized linear Model to predict the family level abundance in communities assembled on a novel carbon source from the fba flux vector alone (Method). We obtained a high cross-validation accuracy and a significant correlation between the predicted and observed relative abundance of the most abundant families (Figure 4.3G).

4.4 Discussion

Previously we have shown that microbial communities will converge to similar family level compositions when assembling on minimal glucose reflecting the taxonomic partitioning of metabolic function [Estrela et al., 2022]. By extending this work to non-glucose environments here we demonstrate that family level community composition is generically predicted by the taxonomic distribution of metabolic traits (Figure 4.1). At first glance the fact that community assembly shows this predictability appears to be in conflict with a large body of work that has found that metabolic traits are often conserved at shallow phylogenetic depths Martiny et al. [2013, 2015]. For example, the ability to grow on a given substrate can be highly variable even at a strain level [Sabarly et al., 2011]. Our works helps to resolve this conflict as we show that convergence during community assembly depends not simply on the ability to grow, but on the quantitative traits selected for by both the supplied and constructed environment (Figure 4.2).

Moving beyond binary classifications of metabolic function and explicitly measuring quantitative metabolic traits led to the identification of conserved correlations that held across environments. Specifically substrates metabolized using overlapping metabolic pathways displayed similar growth rates and resulted in similar by-products, irrespective

of taxonomy (Figure 4.3A-D). Due to these similarities, carbon sources metabolized using overlapping pathways selected for quantitatively similar community compositions (Figure 4.3E-G). By taking of advantage of genome-scale metabolic modelling we can quantify the overlap in metabolic pathway utilization between pairs of substrates. One can leverage this quantification to predict the compositions of a novel set of communities assembling in a novel set of carbon sources (Figure 4.3G).

Just as the 3D structure of a proteins can be predicted by identifying conserved structural motifs, we propose that the taxonomic structure of complex communities can be predicted by identifying conserved functional motifs. Here we have shown that these types of predictions are possible in the context of self-assembled enrichment communities subject to different metabolic environment. Future work will be needed to explore whether other types of selective pressure can similarly be predicted in other types of communities.

4.5 Figures

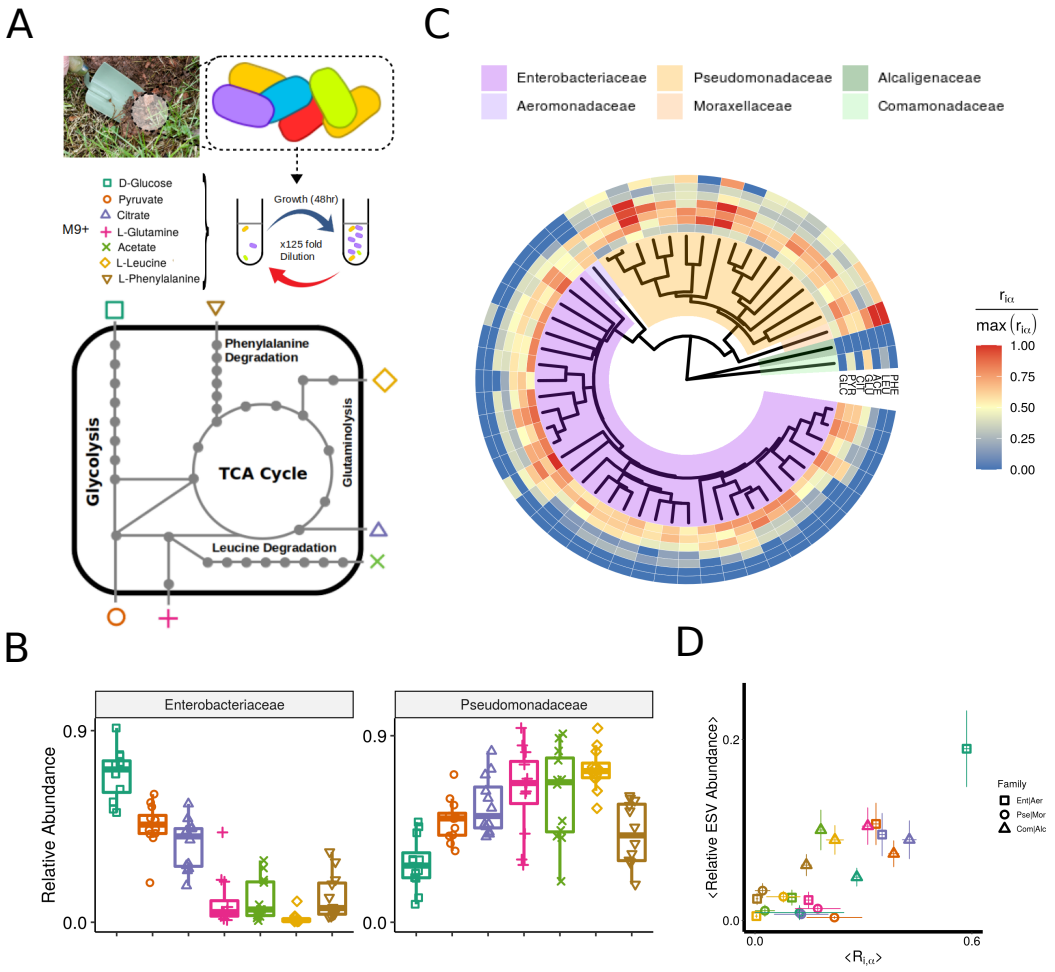


Figure 4.1: Community composition in different environments depends on the taxonomic distribution of quantitative metabolic traits: (A) Diagram of experimental scheme. Communities from 12 different inoculum were allowed to self-assemble in M9 minimal media containing one of 7 different carbon sources. These 7 carbon sources enter central metabolism through different metabolic pathways (B) Communities assembling on different carbon sources selected for distinct yet highly reproducible family level community compositions (n=84). Boxplot shows the relative Abundance of the two dominant families (Enterobacteriaceae and Pseudomonadaceae) in all the communities assembled on different carbon sources (See supplementary Figure 4.1 for other families (C) Phylogenetic tree and normalized growth rates on the 7 different carbon sources for n=54 isolates. Growth rates showed different phylogenetic distributions on the different carbon sources (D) For each carbon source the average growth rate of isolates belonging to each family is correlated with the Abundance of ESVs of that family in the self-assembled communities (Pearsons R =0.88, $P < 0.01$).

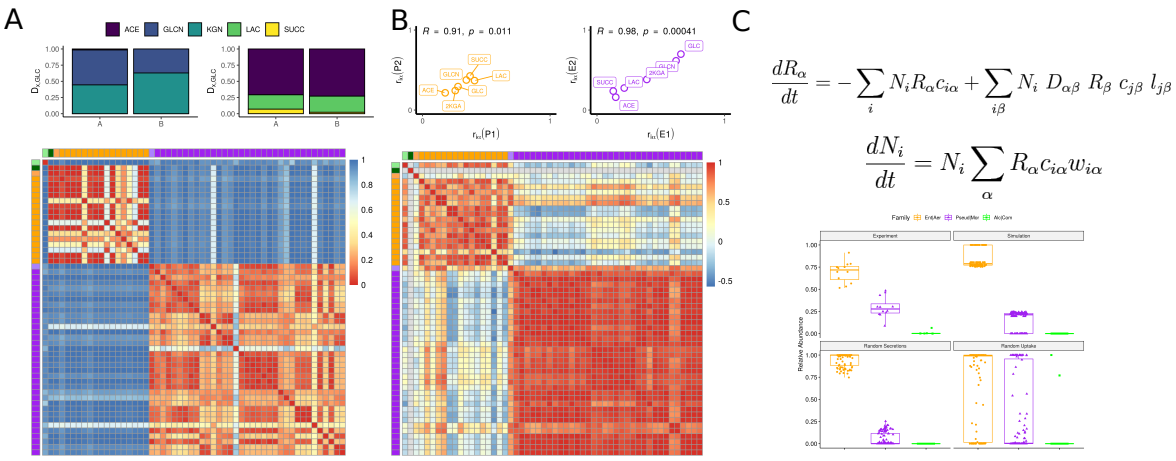


Figure 4.2: Predictability of community assembly depends on the conservation of by-product production and nutrient uptake capabilities: (A) Similarities in metabolic by-product profile ($D_{X,Glc}$) for all isolates after 16hr of growth on glucose. Barplots show metabolic by-product profile for a represantive pair of Pseudomonadaceae (left barplot) and Enterobacteriacea (right barplot). Heatmap shows euclidean distance in secretion profile for every pair of strains. Strains are grouped by phylogenetic similarity (with colours on the edge of the plot corresponding to the different families as in Figure 4.1. (B) Correlated growth on metabolic by-product ($R_{i,\alpha}$) of all isolates after 16hr. Scatter plots show metabolic by-product profile for a Representative pair of Pseudomonadaceae (left plot) and Enterobacteriacea (right plot). Heatmap shows pearsons correlation coefficent for every pair of strains. (C) Using these measured trait values we paramaterize a consumer resource model and simulate 100 random communities assembling on minimal glucose. Simulated communities (top Right) are similar in composition to the experimental communities (top left). This no longer holds true when we either byproduct production (bottom left) or the nutrient uptake capabilities are no longer conserved at the family level (bottom right)

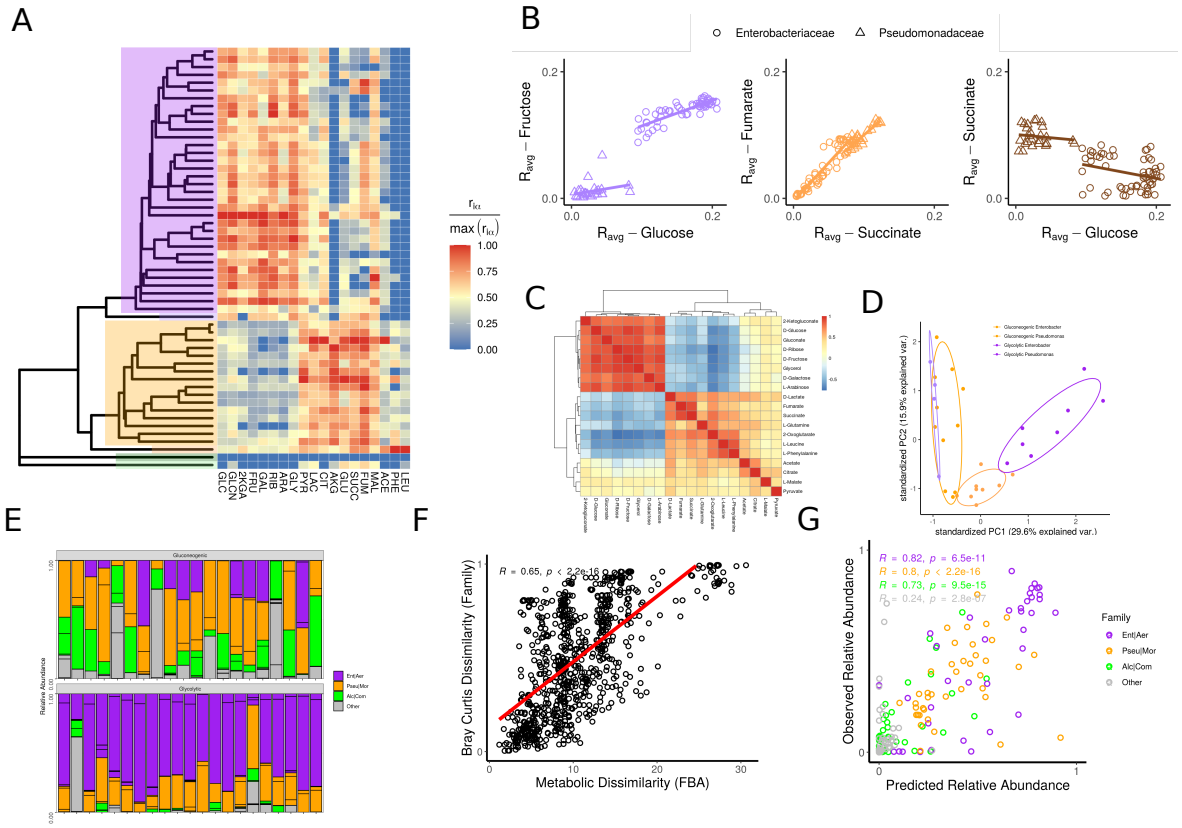


Figure 4.3: Substrates metabolized using overlapping metabolic pathways display similar metabolic traits resulting in similar communities: (A) We measured the growth rate of 54 isolates on 19 different carbon sources (B) Scatter plot shows three representative pairs of substrate. Isolates display significant positive correlations in growth rates (Pearsons $R > 0$, $P < 0.01$) across some pairs of substrate and significant negative correlations in growth rate (Pearsons $R < 0$, $P < 0.01$) across other pairs of bustrates. (C) Hierarchical clustering of correlation coeficients between the growth rates on a pair of substrate. The two clusters that emerge directly correspond to Glycolytic and Gluconeogenic carbon sources (D) For a single Enterobacter and single Pseudomonas strain we performed untarged metabolomics on the spent media after 24hr of growth for a library of different carbon sources. Carbon sources belonging to the same resource class lead to more similar by-products. (E) Communities were assembled in a library of 41 different carbon sources. Each bar corresponds to a different carbon source (grouped by whether the carbon source is glycolytic or Gluconegeonic (F) Metabolically similar carbon sources (quantified using FBA) select for similar communities at the family level (quantified using bray-curtis dissimilarity) (G) Predicted vs observed relative abundance for communities assembling in a novel carbon source (see section 4.6.9).

4.6 Methods

4.6.1 Community assembly experiments

In the first community assembly experiment (data shown in Figure 4.1) communities taken from 12 different soil inoculum where serially transferred in M9 minimal media supplemented with one of seven possible carbon source (D-Glucose, Pyruvate, Citrate, L-Glutamine, Acetate, L-Leucine and L-Phenylalanine). All carbon sources were added at equal c-molar concentrations (0.07 C-mol/L). Communities where initialized by inoculating $4\mu l$ of the source community into $500\mu l$ (as in Goldford et al 2018) of fresh minimal media. The communities where grown in 96 deep well plates(VWR) at 30C without shaking for 48hr. At the end of the 48hr growth cycle $4\mu l$ was transferred to $500\mu l$ of fresh media. This was repeated for a total of 12 transfers (84 generations).

In the second community assembly experiment (data shown in figure 4.3) 2 communities each taken from a different soil inoculum where first grown in Tryptic Soy Broth (TSB). Each enriched community was used as inoculum for M9 minimal media supplemented with equal c-molar concentrations (0.07 C-mol/L) of one of 41 different possible carbon sources .These carbon sources were: Benzoate, Methanol, Ethanol, 1-Propanol, Butanol, Glycolate, Galactitol, Propionate, Acetate, Formate, L-tartrate, Hexanoate, Ribitol, Myo-Inositol, D-Mannitol, Pyruvate, Melibiose, D-Glucose, D-Fructose, D-Galactose, L-Lactate, D-Sorbitol, Salicin, Cellobiose, D-Arabinose, L-Arabinose, Lactose, 2-Oxoglutarate, Trehalose, Sucrose, Glycerol, Raffinose, L-Rhamnose, D-Ribose, Maltose, Citrate, Fumarate, Succinate, L-Malate, Butyrate and Glyoxylate. The communities where grown under the same conditions as in the first community assembly

After the final transfer culture samples were stored at-80C after mixing with 40% Glycerol. Sequencing and Taxonomic Assignment was conducted as described in 3.6.5 and 3.6.6. For the first community assembly experiments samples were rarefied to a maximum sequencing depth of 19969 read. For second community assembly experiment samples were rarefied to a maximum sequencing depth of 17582 reads

4.6.2 Phylogenetic Tree Reconstruction

We started with a library of 100 Gammaproteobacteria isolates for which we had sequenced the full-length 16S rRNA gene [Estrela et al., 2022]. Low quality bases were removed from the start and end of the forward and reverse reads using the biopython

implementation of the motts motified trimming algorithm [Cock et al., 2009]. After merging the forward and reverse reads we eliminated every potentially duplicate sequence by aligning every pair of sequences using the NeedleCommandline function in biopython and counting the mismatches between all pairs of sequences. We conservatively only considered mismatches as a pair of non-identical bases which had phred quality score > 10 and that were less than 20 bp from the end of our sequences. Removing any potentially duplicate sequences(i.e any sequence with more than 1 mismatch) gave us 54 full-length 16s rRNA sequences. Multiple sequence alignemnt was then conducted using clustalw version 2.0 [Larkin et al., 2007]. We inferred a Maximum Likelihood Phylogenetic tree using IQTREE2 version 2.1.2 with 1000 boostrap replicates (iqtree -s alignment.fa -bb 1000) [Minh et al., 2013, Kalyaanamoorthy et al., 2017, Minh et al., 2020]. All subsequent analysis was performed on the consensus tree.

4.6.3 Growth trait quantification

Isolates were streaked from glycerol onto chromogenic agar and grown at 30C for 24hr (HiCrome Universal differential Medium from Sigma). Single colonies of each isolate were used to inoculate 500uL LB(Lennox) in a 96 deep-well plate. These pre-cultures were incubated at 30C shaking at 200rpm for 24hr. After 24hr 100uL of each sample was collected to measure Optical density (OD) 620 (which had reached between 0.2 and 1.5). Pre-Cultures were then diluted 1:100000 in M9 supplemented with one of 19 possible carbon sources (D-Glucose, D-Fructose,D-Galactose, D-Ribose, L-Arabinose, Glycerol, Gluconate ,2-Ketogluconate,Pyruvate, D-Lactate ,Citrate , 2-Oxoglutarate,Succinate,Fumarate,L-Malate,L-Glutamine,L-Leucien,Acetate,L-Phenylalanine).

The final volume for the growth assays was $100uL$ in 96 well plates. OD620 measurements were performed at 30 minute intervals over a 48hr period with an Epoch 2 microplate spectrophotometer (BioTek). Between readings culture plates were stored with lids on in a Microplate Stacker (Bioplate) at 30C without shaking.

From the growth curve of each isolate (i) on each carbon source (α) we estimated the yield ($w_{i\alpha} = N_f$), average growth rate ($r_{i\alpha} = log(N_f/N_0)/(t_f)$ and nutrient uptake rate ($c_{i\alpha} = r_{i\alpha}/w_{i\alpha}$). In these equations t_f is the time at which the supplied resource has been fully consumed, N_f is the OD620 at this time point and N_0 is the OD at the start of the growth curve (calculated using the OD of the inoculating pre-cultures).

Different strains on different carbon sources will consume the supplied resource on different timescales (i.e will have different t_f). Because it would be practically infeasible

to measure the supplied resource concentration of all isolates we assumed that the supplied resource had been fully consumed when growth curves reached either a diauxic shift or approached carrying capacity. To obtain a robust estimate of t_f across all growth curves we deployed a heuristic approach. We first eliminated any noise arising from the lag phase by only considered time-points after the OD had reached half-carrying capacity t_{mid} (estimated using the growth-curver package [Sprouffske and Wagner, 2016]). Only considering timepoints after t_{mid} we took as our t_f the minimum of the following three time-point i) first time at which a negative first derivative is observed ii) time at which the minimum first derivative is observed and iii) first time at which a drop in average growth rate was observed. Visual inspection of our growth curves confirmed that this combinations of criteria, avoided unreliable growth rates estimates due to noisy stationary phase dynamics (i.e due to biofilm formation). On glucose this heuristic gave us a mean (t_f) for the Enterobacteriaceae of $13.7hr$ and for the Pseudomonadaceae of $28hr$ which is in line with estimates obtained from glucose concentration curves [Goldford et al., 2018].

4.6.4 By-product measurements on glucose

Isolates were preconditions to growth on glucose minimal media ($500\mu L$) for 48hr. $4\mu L$ of the preconditioned cultures were inoculated into 500uL fresh glucose media. $100\mu l$ samples were collected after 16h and their OD620 was measured. The remaining sample was centrifuged at 3000rpm for 25 min to separate cells from supernantant. Supernatants were transferred to a 96 well plate $0.2\mu m$ AcroPrep filter plate on top of a 96 well NUNC plate fitted with the metal collar adaptor and centrifuged at 3000 rpm for 10 min. The supernatant was immediately frozen at -80C until processing.

Additional glucose,acetate,lactate and succinate measurements were conducted as described previously [Estrela et al., 2022].Gluconate concentrations were measured using a D-Gluconate assay kit (ab204703). 2-Ketogluconate concentrations were measured as described in Molina et al. [2019]. Briefly 50ul of o-phenylenediamine dihydrochloride was combined with 100ul of diluted filtered culture supernatant and heated at 100C for 30 minutes. The absorbance of the reaction mixture was measured at 330mm and 2-ketogluconate concentration were estimated by comparing the diluted supernatant to a standard curve.

4.6.5 Phylogenetic Imputation of Quantitative Metabolic Traits

In order to quantify the extent to which strain-level metabolic traits could be predicted from the relative position on the 16s rRNA phylogenetic tree we combined phylogenetic imputation using a brownian motion evolutionary model with leave-one-out cross-validation. For each isolate all measurements of its traits were removed and the value was imputed using the phylopars.predict function in the Rphylopars package[Bruggeman et al., 2009]. We repeated this predictions for all isolates and trait measurements independently. The degree of phylogenetic predictability is quantified as the correlation between the observed and predicted trait values across the complete cross-validation dataset.

4.6.6 Targeted LCMS of E.coli, Enterobacter, Pseudomonas and P.putida supernatant across carbon sources

E. coli MG1655, *P.putida KT2440*, an Enterobacter and a Pseudomonas isolate from the glucose communities in Goldford et al. [2018]) were revived on LB Agar. For each strain we picked two replicate colonies and inoculated them into 50ml falcon tubes containing 5ml of LB(Lennox). Falcon tubes were incubated at 30C (shaking) for 16hrs. After this all 8 populations were brought into balanced exponential phase by diluting (1:5) 3 times into fresh LB. The first three dilutions were performed at 1hr intervals after which cultures were allowed to grow for an 1hr and 30min. At this point cells were centrifuged and washed three times in M9 minimal media containing no carbon source to remove any left-over LB before being resuspended in M9 Minimal media. Cells were normalized to a pre-innoculation OD of 0.1.

All 8 samples were used to innoculate 3 replicates and grown at 30C in 500ul of M9 media containing one of 5 possible carbon sources (D-Glucose, D-Fructose, Glycerol, Pyruvate and L-Malate). Each replicate was used for a different timepoint (16hr,28hr and 48hr). At each timepoint OD readings were taken and the supernatant was extracted as described in subsection 4.6.4. Metabolite quantification in samples was conducted using liquid-chromatography mass spectrometry as described in Estrela et al. [2022]

4.6.7 Untargeted LCMS of Enterobacter and Pseudomonas supernatant across carbon sources

An Enterobacter and a Pseudomonas isolate from Estrela et al. [2022] were revived on chromogenic agar. Three replicates of each isolate were preconditions for 48hr on minimal media containing one of the following 15 carbon sources: D-Glucose ,D-Fructose ,D-Galactose ,D-Ribose ,L-Arabinose ,Glycerol ,Gluconate , 2-Ketoglucoante ,Pyruvate , D-Lactate ,Citrate ,Fumarate ,L-Malate ,Acetate, L-Glutamine,L-Leucine or L-Phenylalanine. After preconditioning 4ul of culture was transferred to fresh media which was grown for 24hr before spent media extraction. Growth conditions and spend media extraction were carried out as described in subsection 4.6.6. 50ul samples were submitted for untargeted metabolomics analysis via LCMS.

4.6.8 Genome-Scale Metabolic Modelling using Flux Balance Analysis

For every carbon source in the experiment shown in Figure 4.3 we used Flux Balance Analysis to obtain a vector of fluxes through a typical microbial metabolic network. For this analysis we used a previously published universal gram-negative bacterial model [Machado et al., 2018] that was capable of growing on every carbon source studied in this paper. For these simulations we followed the same procedure as described in subsection 3.6.12. All carbon sources were supplied to the model at equimolar concentrations by setting the exchange flux to -1 cmol/gDWh. Each carbon source could thus be associated with a unique vector of metabolic fluxes. All simulations were conducted using the cobrapy package [Ebrahim et al., 2013b].

4.6.9 LASSO prediction of community composition in a novel carbon source

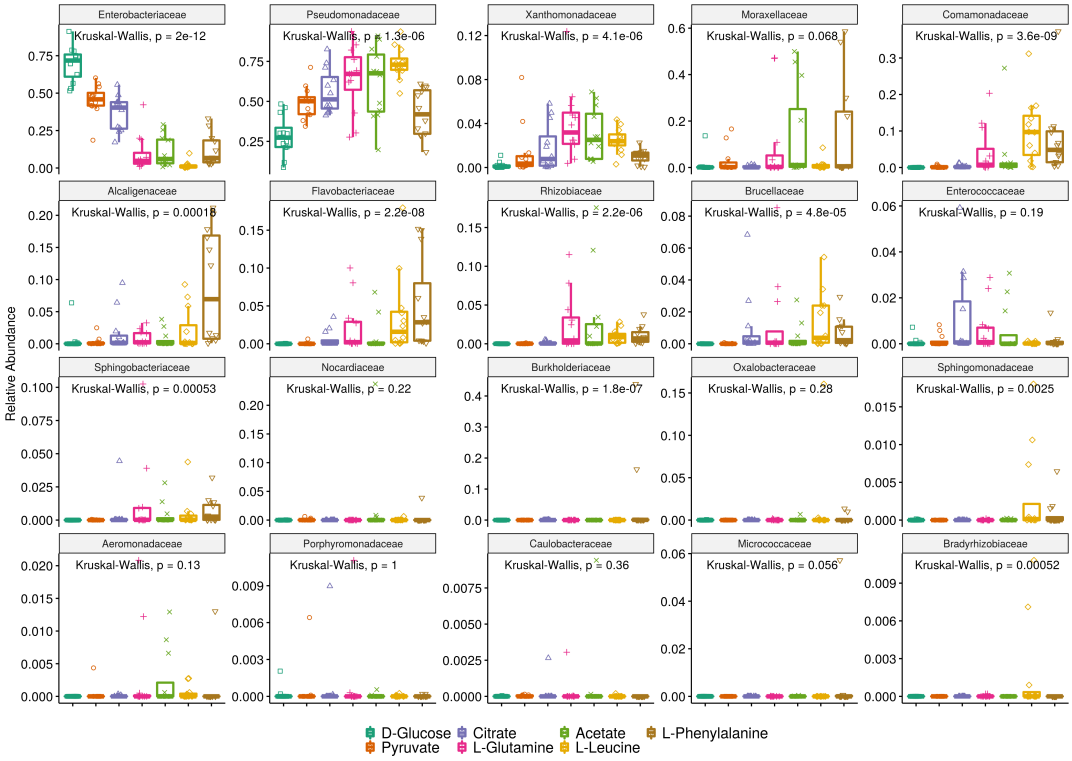
We used leave-one-out cross-validation to determine whether we could predict the composition of communities assembling in a novel carbon source. Linear regressions were trained on the 40 of the Carbon Source in the Community Assembly experiment shown in Figure 4.3 and used to predict the composition of communities assembled on the 41st carbon source. We repeated this for all 41 carbon source in that experiment. The FBA flux vectors for each carbon sources (subsection 4.6.8) were used as the independent variable

for the model, while the dependent variable was the family-level community composition. Linear regressions were regularized using a LASSO regularizer.

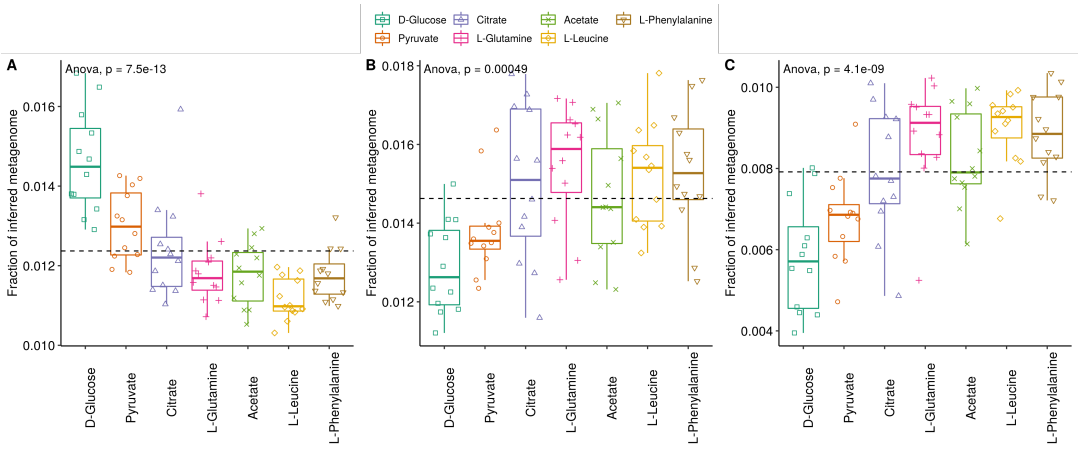
4.7 Supplementary Material

4.7.1 Supplementary Figures

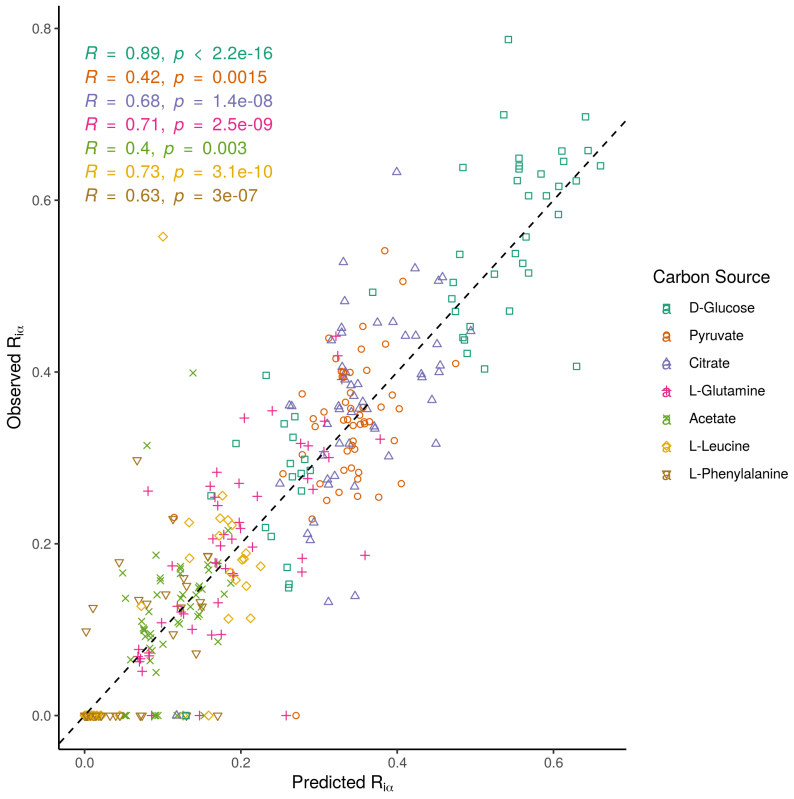
Supplementary Figure 4.1: Communities assembled in the same environment have more similar family-levels community composition than communities assembled in different environments (one-tailed Kolmogorov-Smirnov, $p < 1e\text{-}06$). Heatmap show Family-level Bray–Curtis dissimilarity for every pair of 84 communities assembled as part of the experiment shown in figure 4.1A.



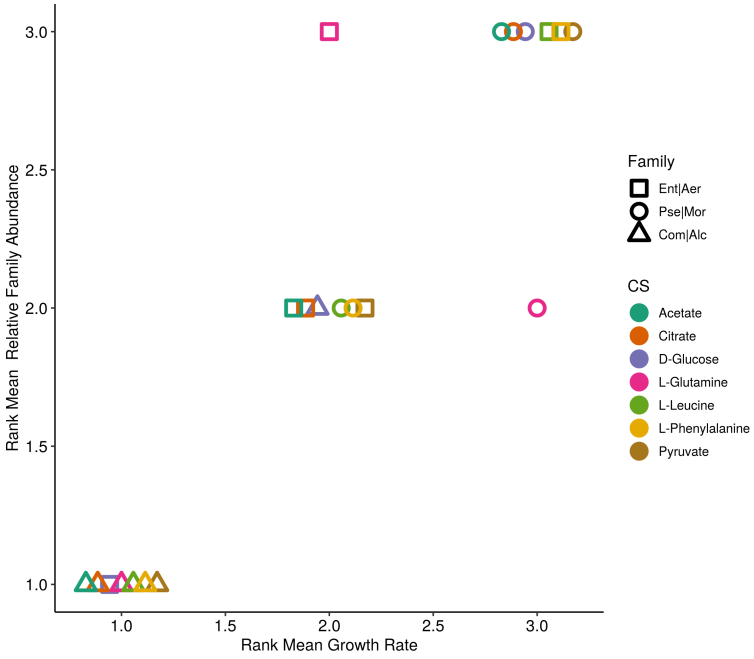
Supplementary Figure 4.2: Family level community composition is convergent across environments. Boxplots show the relative abundance for all families found in at least 10 of the 84 communities assembled as part of the experiment shown in figure 4.1A



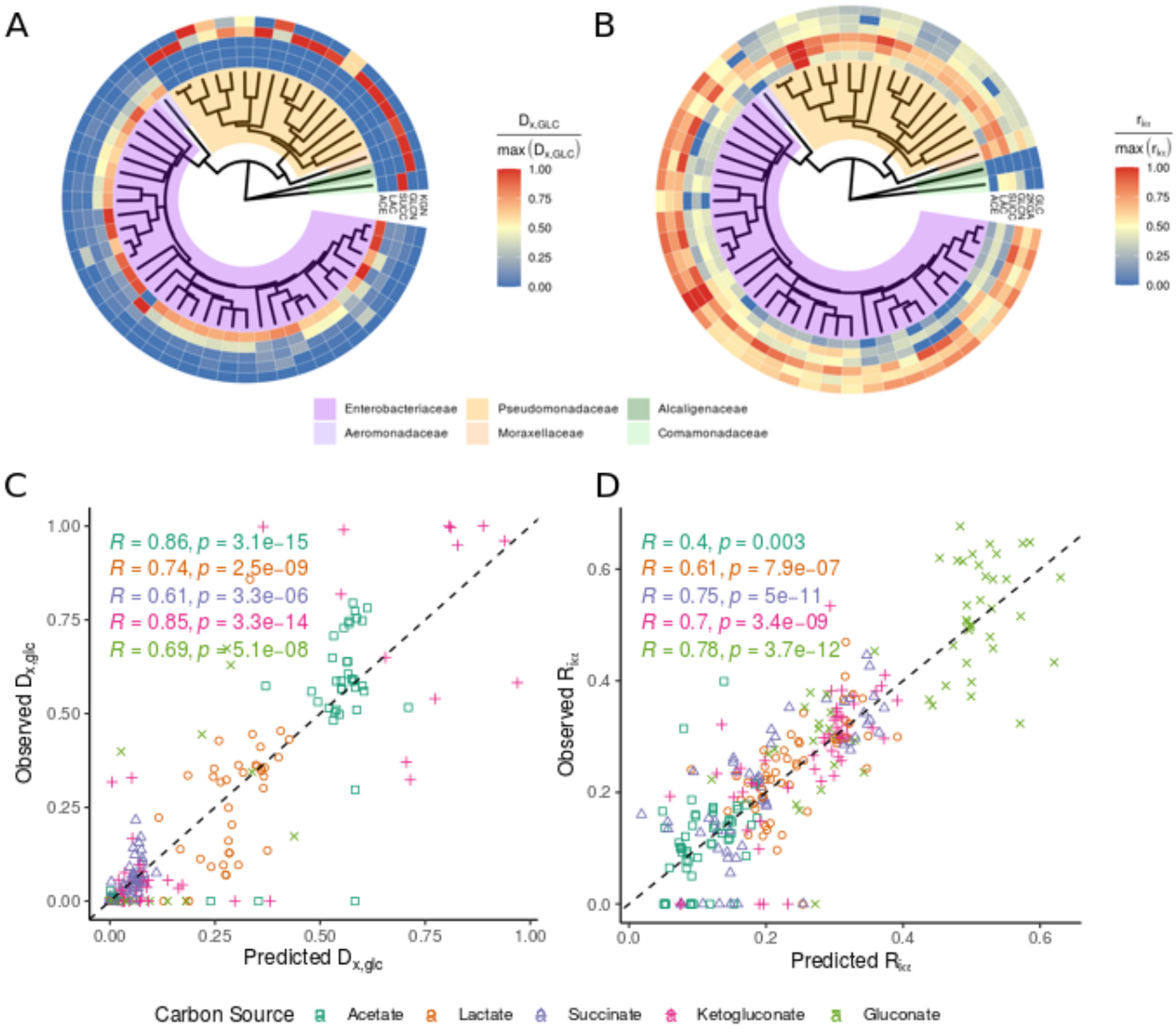
Supplementary Figure 4.3: Metagenomes were inferred using Picrust2. Each point corresponds to a community assembled in one of 7 carbon sources. The y axis shows the fractions of genes in the inferred metagenome that are involved in Glycolysis (A), The TCA Cycle (B) or Amino Acid Degradation (C). Communities assembled in different carbon sources showed significant differences in the relative abundance of genes involved these pathways (one-way anova).A post hoc Tukey test confirmed that communities assembled on D-Glucose and Pyruvate had significantly (P¡0.01) higher fraction of genes involved in Glycolysis, Communities assembled on Citrate,L-Glutamine and Acetate had higher levels of genes involved in the TCA Cycle and communities assembled on L-Leucine and L-Phenylanine had higher levels of genes involved in Amino acid degradation



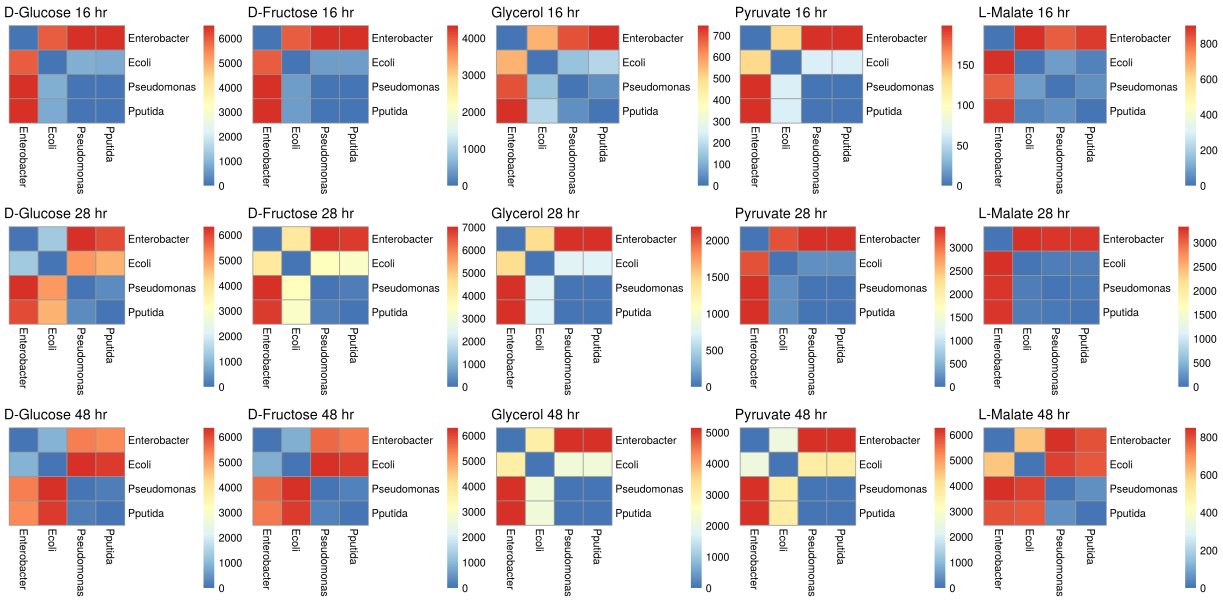
Supplementary Figure 4.4: Predicted versus observed Growth Rate for n= 54 Gammaproteobacteria strains on each of the 7 carbon sources shown in Figure 4.1. Predictions where made by combining leave-one cross-validation and phylogenetic imputation as outlined in Section 4.6.5. Points along the identity line (dotted black line where predicted = observed),correspond to perfect predictions. On all 7 carbon sources we find a significant positive correlation between predicted and observed growth rate (Pearson's $R > 0$ and $p < 0.001$).



Supplementary Figure 4.5: On all seven carbon sources the family containing fastest growing strains in monoculture (X axis = 1) had the highest average abundance in the self-assembled communities (Y axis = 1). On 6 of the seven carbon the second fastest growing 'family' (X axis = 2) was the second most abundant in the community (Y axis = 2).



Supplementary Figure 4.6: Phylogenetic conservation of cross feeding traits. (A) Experimentally measured Metabolite secretion profile ($D_{X,glc}$) for 54 isolates grown for 16 hours on M9 minimal glucose (B) Experimentally measured growth rates on secreted metabolite by-products of glucose metabolism. As in Supplementary Figure 4.7 we used phylogenetic imputation to quantify the predicted versus observed by-product secretions (C) and growth Rate (D) . For all measured substrate we find a significant positive correlation both between predicted and observed secretions and betweem growth rate (Pearson's $R > 0$ and $p < 0.001$).



Supplementary Figure 4.7: Members of the same family produce more similar by-products when grown on the same carbon source. We used Targeted LCMS to quantify the Concentration of metabolites secreted by E. coli, Enterobacter P.putida and Pseudomonas after 16,28 and 48 hours of growth. Heatmap shows the euclidean distance in metabolic by-products production for each carbon source at each timepoint . Blue squares show more similar secretions

4.8 Acknowledgements

We want to thank members of the Sanchez, Segre lab and Mehta Lab for helpful discussions. This work was supported by the National Institutes of Health through grant 1R35 GM133467-01, and by a Packard Foundation Fellowship to AS.

Bibliography

Richard C Lewontin. The organism as the subject and object of evolution. *Scientia*, 77 (18):65, 1983. 2, 9, 45

F John Odling-Smee, Kevin N Laland, and Marcus W Feldman. *Niche Construction*. Princeton University Press, February 2013. 2

Kevin Laland, Tobias Uller, Marc Feldman, Kim Sterelny, Gerd B Müller, Armin Moczek, Eva Jablonka, John Odling-Smee, Gregory A Wray, Hopi E Hoekstra, Douglas J Futuyma, Richard E Lenski, Trudy F C Mackay, Dolph Schluter, and Joan E Strassmann. Does evolutionary theory need a rethink? *Nature*, 514(7521):161–164, October 2014. 2, 10

David M Post and Eric P Palkovacs. Eco-evolutionary feedbacks in community and ecosystem ecology: interactions between the ecological theatre and the evolutionary play. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 364(1523):1629–1640, June 2009. 2, 10

Andrew P Hendry. *Eco-evolutionary Dynamics*. Princeton University Press, November 2016. 2

Laura E Jones, Lutz Becks, Stephen P Ellner, Nelson G Hairston, Jr, Takehito Yoshida, and Gregor F Fussmann. Rapid contemporary evolution and clonal food web dynamics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 364(1523):1579–1591, June 2009. 2

David Basanta and Alexander R A Anderson. Homeostasis back and forth: An ecoevolutionary perspective of cancer. *Cold Spring Harb. Perspect. Med.*, 7(9), September 2017. 2

Alvaro Sanchez and Jeff Gore. feedback between population and evolutionary dynamics determines the fate of social microbial populations. *PLoS Biol.*, 11(4):e1001547, April 2013. 3

Blake Matthews, Thierry Aebischer, Karen E Sullam, Bänz Lundsgaard-Hansen, and Ole Seehausen. Experimental evidence of an eco-evolutionary feedback during adaptive divergence. *Curr. Biol.*, 26(4):483–489, February 2016. 3

Casey P terHorst and Peter C Zee. Eco-evolutionary dynamics in plant–soil feedbacks. *Funct. Ecol.*, 30(7):1062–1072, July 2016. 3

Seth M Rudman, Matthew A Barbour, Katalin Csilléry, Phillip Gienapp, Frederic Guillaume, Nelson G Hairston, Jr, Andrew P Hendry, Jesse R Lasky, Marina Rafajlović, Katja Räsänen, Paul S Schmidt, Ole Seehausen, Nina O Therkildsen, Martin M Turcotte, and Jonathan M Levine. What genomic data can reveal about eco-evolutionary dynamics. *Nat Ecol Evol*, 2(1):9–15, January 2018. 3, 10

S Wright. The roles of mutation, inbreeding, cross-breeding and selection in evolution, p 356–366. In *Proceedings of the Sixth Annual Congress of Genetics*, volume 1, 1932. 3

S A Kauffman and S Johnsen. Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.*, 149(4):467–505, April 1991. 3, 10, 11

Richard A Watson and Marc Ebner. Eco-Evolutionary dynamics on deformable fitness landscapes. In Hendrik Richter and Andries Engelbrecht, editors, *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 339–368. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. 3, 11

Daniel M Weinreich, Nigel F Delaney, Mark A Depristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312 (5770):111–114, April 2006. 3

Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, January 2007. 3, 5

Daniel L Hartl. What can we learn from fitness landscapes? *Curr. Opin. Microbiol.*, 21: 51–57, October 2014. 3

Joshua R Nahum, Peter Godfrey-Smith, Brittany N Harding, Joseph H Marcus, Jared Carlson-Stevermer, and Benjamin Kerr. A tortoise–hare pattern seen in adapting structured and unstructured populations suggests a rugged fitness landscape in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, 112(24):7530–7535, June 2015. 3

Jeremy Van Cleve and Daniel B Weissman. Measuring ruggedness in fitness landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 112(24):7345–7346, June 2015. 3

Daniel E Rozen, Michelle G J L Habets, Andreas Handel, and J Arjan G M de Visser. Heterogeneous adaptive trajectories of small populations on complex fitness landscapes. *PLoS One*, 3(3):e1715, March 2008. 3

Danesh Moradigaravand and Jan Engelstädter. The effect of bacterial recombination on adaptation on fitness landscapes with limited peak accessibility. *PLoS Comput. Biol.*, 8 (10):e1002735, October 2012. 3

Aditya Barve and Andreas Wagner. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, 500(7461):203–206, August 2013. 3

J Arjan G M de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.*, 15(7):480–490, July 2014. 3

Louise J Barber, Matthew N Davies, and Marco Gerlinger. Dissecting cancer evolution at the macro-heterogeneity and micro-heterogeneity scale. *Curr. Opin. Genet. Dev.*, 30: 1–6, February 2015. 3

Boyang Zhao, Michael T Hemann, and Douglas A Lauffenburger. Modeling tumor clonal evolution for drug combinations design. *Trends Cancer Res.*, 2(3):144–158, March 2016. 3

Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507 (7490):57–61, March 2014. 3

Armita Nourmohammad, Torsten Held, and Michael Lässig. Universality and predictability in molecular quantitative genetics. *Curr. Opin. Genet. Dev.*, 23(6):684–693, December 2013. 3

Michael Lässig, Ville Mustonen, and Aleksandra M Walczak. Predicting evolution. *Nat Ecol Evol*, 1(3):77, February 2017. 3, 109

C E Paquin and J Adams. Relative fitness can decrease in evolving asexual populations of s. cerevisiae. *Nature*, 306(5941):368–370, 1983. 3, 6

Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50, November 2017. 3

Mickael Le Gac and Michael Doebeli. Epistasis and frequency dependence influence the fitness of an adaptive mutation in a diversifying lineage. *Mol. Ecol.*, 19(12):2430–2438, June 2010. 3

R F Rosenzweig, R R Sharp, D S Treves, and J Adams. Microbial evolution in a simple unstructured environment: genetic differentiation in escherichia coli. *Genetics*, 137(4): 903–917, August 1994. 3, 4

Maren L Friesen, Gerda Saxer, Michael Travisano, and Michael Doebeli. Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in escherichia coli. *Evolution*, 58(2):245–260, February 2004. 3

Erik M Quandt, Jimmy Gollihar, Zachary D Blount, Andrew D Ellington, George Georgiou, and Jeffrey E Barrick. Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the lenski evolution experiment. *Elife*, 4, October 2015. 4, 6, 7

Nicole Paczia, Anke Nilgen, Tobias Lehmann, Jochem Gätgens, Wolfgang Wiechert, and Stephan Noack. Extensive exometabolome analysis reveals extended overflow metabolism in various microorganisms. *Microb. Cell Fact.*, 11:122, September 2012. 4

Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard Ø Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism–2011. *Mol. Syst. Biol.*, 7(1):535, October 2011. 4, 5, 90

Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, Karl K Weitz, Roland Eils, Rainer König, Richard D Smith, and Bernhard Ø Palsson. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.*, 6(1):390, July 2010. 4, 17

Edward J O'Brien, Jonathan M Monk, and Bernhard O Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971–987, May 2015. 4

João F Matias Rodrigues and Andreas Wagner. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.*, 5(12):e1000613, December 2009. 4

Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle, 3rd. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophys. J.*, 83(3):1331–1340, September 2002. 4

William R Harcombe, William J Riehl, Ilija Dukovski, Brian R Granger, Alex Betts, Alex H Lang, Gracia Bonilla, Amrita Kar, Nicholas Leiby, Pankaj Mehta, Christopher J Marx, and Daniel Segrè. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.*, 7(4):1104–1115, May 2014. 4, 7, 22, 38, 90, 109

Richard E Lenski, Michael R Rose, Suzanne C Simpson, and Scott C Tadler. Long-Term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138(6):1315–1341, December 1991. 4, 5, 7

Zachary D Blount, Christina Z Borland, and Richard E Lenski. Historical contingency and the evolution of a key innovation in an experimental population of escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, 105(23):7899–7906, June 2008. 4, 6

M Travisano and R E Lenski. Long-term experimental evolution in escherichia coli. IV. targets of selection and the specificity of adaptation. *Genetics*, 143(1):15–26, May 1996. 5

J Arjan G M de Visser and Richard E Lenski. Long-term experimental evolution in escherichia coli. XI. rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evol. Biol.*, 2:19, October 2002. 5, 6

Benjamin Kerr, Margaret A Riley, Marcus W Feldman, and Brendan J M Bohannan. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature*, 418 (6894):171–174, July 2002. 6

P B Rainey and M Travisano. Adaptive radiation in a heterogeneous environment. *Nature*, 394(6688):69–72, July 1998. 6

Zachary D Blount, Jeffrey E Barrick, Carla J Davidson, and Richard E Lenski. Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature*, 489 (7417):513–518, September 2012. 6

Romain Gallet, Cyrille Violle, Nathalie Fromin, Roula Jabbour-Zahab, Brian J Enquist, and Thomas Lenormand. The evolution of bacterial cell size: the internal diffusion-constraint hypothesis. *ISME J.*, 11(7):1559–1568, July 2017. 7, 19

Csaba Pál and Balázs Papp. Evolution of complex adaptations in molecular systems. *Nat Ecol Evol*, 1(8):1084–1092, August 2017. 9

Charles Darwin. *The Formation of Vegetable Mould, Through the Action of Worms: With Observations on Their Habits*. J. Murray, 1892. 9

R C Lewontin. Adaptation. *Sci. Am.*, 239(3):212–8, 220, 222 passim, September 1978. 9

Ricard V Solé and Josep Sardanyés. Red queen coevolution on fitness landscapes. In Hendrik Richter and Andries Engelbrecht, editors, *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 301–338. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. 10

Levi T Morran, Olivia G Schmidt, Ian A Gelarden, Raymond C Parrish, II, and Curtis M Lively. Running with the red queen: Host-Parasite coevolution selects for biparental sex. *Science*, July 2011. 10

Peter Schuster. A revival of the landscape paradigm: Large scale data harvesting provides access to fitness landscapes. *Complexity*, 17(5):6–10, May 2012. 10

Michael Doebeli, Yaroslav Ispolatov, and Burt Simon. Point of view: Towards a mechanistic foundation of evolutionary theory. *Elife*, 6:e23804, 2017. 10

P A Moran. ON THE NONEXISTENCE OF ADAPTIVE TOPOGRAPHIES. *Ann. Hum. Genet.*, 27:383–393, June 1964. 10

Matteo Mori, Terence Hwa, Olivier C Martin, Andrea De Martino, and Enzo Marinari. Constrained allocation flux balance analysis. *PLoS Comput. Biol.*, 12(6):e1004913, June 2016. 10, 11

Q K Beg, A Vazquez, J Ernst, M A de Menezes, Z Bar-Joseph, A-L Barabási, and Z N Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by escherichia coli and constrains its metabolic activity. *Proc. Natl. Acad. Sci. U. S. A.*, 104(31):12663–12668, July 2007. 10, 11

Stephen J Giovannoni, J Cameron Thrash, and Ben Temperton. Implications of streamlining theory for microbial ecology. *ISME J.*, 8(8):1553–1565, August 2014. 11

D Waxman and S Gavrilets. 20 questions on adaptive dynamics. *J. Evol. Biol.*, 18(5): 1139–1154, September 2005. 11

Zachary A King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. BiGG models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, 44(D1):D515–22, January 2016. 17

Claus Jonathan Fritzemeier, Daniel Hartleb, Balázs Szappanos, Balázs Papp, and Martin J Lercher. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS Comput. Biol.*, 13(4):e1005494, April 2017. 17

William R Harcombe, Nigel F Delaney, Nicholas Leiby, Niels Klitgord, and Christopher J Marx. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. *PLoS Comput. Biol.*, 9(6):e1003091, June 2013. 17

Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. COBRApy: COnstraints-Based reconstruction and analysis for python. *BMC Syst. Biol.*, 7:74, August 2013a. 17, 18

R Core Team. R: A language and environment for statistical computing, 2021. 18

Günter P Wagner. The measurement theory of fitness. *Evolution*, 64(5):1358–1376, May 2010. 18

J A John and N R Draper. An alternative family of transformations. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 29(2):190–197, 1980. 18

S Six, S C Andrews, G Unden, and J R Guest. Escherichia coli possesses two homologous anaerobic c4-dicarboxylate membrane transporters (DcuA and DcuB) distinct from the aerobic dicarboxylate transport system (dct). *J. Bacteriol.*, 176(21):6470–6478, November 1994. 19

A J Link, D Phillips, and G M Church. Methods for generating precise deletions and insertions in the genome of wild-type escherichia coli: application to open reading frame characterization. *J. Bacteriol.*, 179(20):6228–6237, October 1997. 20

Csaba Pál, Balázs Papp, Martin J Lercher, Péter Csermely, Stephen G Oliver, and Laurence D Hurst. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440(7084):667–670, March 2006. 22

Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.*, 113(21):5970–5975, May 2016. 37

Philip Hunter. Plant microbiomes and sustainable agriculture: Deciphering the plant microbiome and its role in nutrient supply and plant immunity has great potential to reduce the use of fertilizers and biocides in agriculture. *EMBO Rep.*, 17(12):1696–1699, December 2016. 37

Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, September 2016a. 37, 109

Martin J Blaser, Zoe G Cardon, Mildred K Cho, Jeffrey L Dangl, Timothy J Donohue, Jessica L Green, Rob Knight, Mary E Maxon, Trent R Northen, Katherine S Pollard, and Eoin L Brodie. Toward a predictive understanding of earth's microbiomes to address 21st century challenges. *MBio*, 7(3), May 2016. 37

Ruth E Ley, Daniel A Peterson, and Jeffrey I Gordon. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–848, February 2006. 37

Gwen Falony, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, Marc Jan Bonder, Mireia Valles-Colomer, Doris Vandeputte, Raul Y Tito, Samuel Chaffron, Leen Rymenans, Chloë Verspecht, Lise De Sutter, Gipsi Lima-Mendez, Kevin D'hoe, Karl Jonckheere, Daniel Homola, Roberto Garcia, Ettje F Tigchelaar, Linda Eeckhaudt, Jingyuan Fu, Liesbet Henckaerts, Alexandra

Zhernakova, Cisca Wijmenga, and Jeroen Raes. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–564, April 2016. 37

Afrah Shafquat, Regina Joice, Sheri L Simmons, and Curtis Huttenhower. Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends Microbiol.*, 22(5):261–266, May 2014. 37

Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A Brantley Hall, Arthur Brady, Heather H Creasy, Carrie McCracken, Michelle G Giglio, Daniel McDonald, Eric A Franzosa, Rob Knight, Owen White, and Curtis Huttenhower. Erratum: Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 551(7679):256, November 2017. 37

Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012. 37, 109

Lawrence A David, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biol.*, 15(7):R89, 2014a. 37, 109

Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, Sudha B Biddinger, Rachel J Dutton, and Peter J Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, January 2014b. 37, 109

Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.*, 10(8):538–550, July 2012. 38

Joshua E Goldford, Nanxi Lu, Djordje Bajić, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segrè, Pankaj Mehta, and Alvaro Sanchez. Emergent simplicity in microbial community assembly. *Science*, 361(6401):469–474, August 2018. 38, 40, 41, 44, 51, 67, 68, 70, 71, 86, 87, 109, 112, 122, 123

Daniel Machado, Oleksandr M Maistrenko, Sergej Andrejev, Yongkyu Kim, Peer Bork, Kaustubh R Patil, and Kiran R Patil. Polarization of microbial communities between competitive and cooperative metabolism. *Nat Ecol Evol*, 5(2):195–203, February 2021. 38

Alicia Sanchez-Gorostiaga, Djordje Bajić, Melisa L Osborne, Juan F Poyatos, and Alvaro Sanchez. High-order interactions distort the functional landscape of microbial consortia. *PLoS Biol.*, 17(12):e3000550, December 2019. 38, 65

Daniel M Cornforth and Kevin R Foster. Competition sensing: the social side of bacterial stress responses. *Nat. Rev. Microbiol.*, 11(4):285–293, April 2013. 38

Amir Bashan, Travis E Gibson, Jonathan Friedman, Vincent J Carey, Scott T Weiss, Elizabeth L Hohmann, and Yang-Yu Liu. Universality of human microbial dynamics. *Nature*, 534(7606):259–262, June 2016. 38, 39, 40, 41, 44, 45, 46, 52, 53

Michael Kalyuzhny and Nadav M Shnerb. Dissimilarity-overlap analysis of community dynamics: Opportunities and pitfalls. *Methods Ecol. Evol.*, 8(12):1764–1773, December 2017. 39, 44, 45, 53

Erik Verbruggen, Merlin Sheldrake, Luke D Bainard, Baodong Chen, Tobias Ceulemans, Johan De Gruyter, and Maarten Van Geel. Mycorrhizal fungi show regular community compositions in natural ecosystems. *ISME J.*, 12(2):380–385, February 2018. 40, 46

Stephen P Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press, June 2011. 44

Djordje Bajić, Jean C C Vila, Zachary D Blount, and Alvaro Sánchez. On the deformability of an empirical fitness landscape by microbial evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 115(44):11286–11291, October 2018. 45

Benjamin J Callahan, Tadashi Fukami, and Daniel S Fisher. Rapid evolution of adaptive niche construction in experimental microbial populations. *Evolution*, 68(11):3307–3316, November 2014. 45

Kevin N Laland, Tobias Uller, Marcus W Feldman, Kim Sterelny, Gerd B Müller, Armin Moczek, Eva Jablonka, and John Odling-Smee. The extended evolutionary synthesis: its structure, assumptions and predictions. *Proc. Biol. Sci.*, 282(1813):20151019, August 2015. 45

Patrick C Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, 9(11):855–867, November 2008. 65

Trudy F C Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, 15(1):22–33, January 2014. 65

Jean-Baptiste Michel, Pamela J Yeh, Remy Chait, Robert C Moellering, Jr, and Roy Kishony. Drug interactions modulate the potential for evolution of resistance. *Proc. Natl. Acad. Sci. U. S. A.*, 105(39):14918–14923, September 2008. 65

Robert Woods, Dominique Schneider, Cynthia L Winkworth, Margaret A Riley, and Richard E Lenski. Tests of parallel molecular evolution in a long-term experiment with escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, 103(24):9107–9112, June 2006. 65

Mauricio Cruz-Loya, Tina Manzhu Kang, Natalie Ann Lozano, Rina Watanabe, Elif Tekin, Robert Damoiseaux, Van M Savage, and Pamela J Yeh. Stressor interaction networks suggest antibiotic resistance co-opted from stress responses to temperature. *ISME J.*, 13 (1):12–23, January 2019. 65

Alison L Gould, Vivian Zhang, Lisa Lamberti, Eric W Jones, Benjamin Obadia, Nikolaos Korasidis, Alex Gavryushkin, Jean M Carlson, Niko Beerenwinkel, and William B Ludington. Microbiome interactions shape host fitness. *Proc. Natl. Acad. Sci. U. S. A.*, 115 (51):E11951–E11960, December 2018. 65

Xiaokan Guo and James Q Boedicker. The contribution of High-Order metabolic interactions to the global activity of a Four-Species microbial community. *PLoS Comput. Biol.*, 12(9):e1005079, September 2016. 65

Jeremiah J Faith, Nathan P McNulty, Federico E Rey, and Jeffrey I Gordon. Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science*, 333(6038): 101–104, July 2011. 65, 73, 109

Alan R Pacheco, Melisa L Osborne, and Daniel Segrè. Non-additive microbial community responses to environmental complexity. *Nat. Commun.*, 12(1):2365, April 2021. 66

A Sanchez. Defining higher-order interactions in synthetic ecology: lessons from physics and quantitative genetics. cell syst 9: 519–520, 2019. 66

Elif Tekin, Pamela J Yeh, and Van M Savage. General form for interaction measures and framework for deriving Higher-Order emergent effects. *Frontiers in Ecology and Evolution*, 6, 2018. 66

Sylvie Estrela, Jean C C Vila, Nanxi Lu, Djordje Bajić, Maria Rebolleda-Gómez, Chang-Yu Chang, Joshua E Goldford, Alicia Sanchez-Gorostiaga, and Álvaro Sánchez. Functional attractors in microbial community assembly. *Cell Syst*, 13(1):29–42.e7, January 2022. 67, 68, 70, 109, 110, 112, 114, 120, 122, 123, 124

Juan Diaz-Colunga, Nanxi Lu, Alicia Sanchez-Gorostiaga, Chang-Yu Chang, Helen S Cai, Joshua E Goldford, Mikhail Tikhonov, and Álvaro Sánchez. Top-down and bottom-up cohesiveness in microbial community coalescence. *Proc. Natl. Acad. Sci. U. S. A.*, 119 (6), February 2022. 68

Robert Marsland, 3rd, Wenping Cui, Joshua Goldford, Alvaro Sanchez, Kirill Korolev, and Pankaj Mehta. Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS Comput. Biol.*, 15(2):e1006793, February 2019. 71, 87

Robert Marsland, 3rd, Wenping Cui, and Pankaj Mehta. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Sci. Rep.*, 10 (1):3308, February 2020a. 71, 72, 86, 87

Robert Marsland, Wenping Cui, Joshua Goldford, and Pankaj Mehta. The community simulator: A python package for microbial ecology. *PLoS One*, 15(3):e0230430, March 2020b. 71, 81, 86, 87, 89, 112

Carlos A Serván and Stefano Allesina. Tractable models of ecological assembly. *Ecol. Lett.*, 24(5):1029–1037, May 2021. 71

W P Hempfling and S E Mainzer. Effects of varying the carbon source limiting growth on yield and maintenance characteristics of escherichia coli in continuous culture. *J. Bacteriol.*, 123(3):1076–1087, September 1975. 73

James R Skrinde and Surinder K Bhagat. Industrial wastes as carbon sources in biological denitrification. *J. Water Pollut. Control Fed.*, 54(4):370–377, 1982. 73

Tim N Enke, Manoshi S Datta, Julia Schwartzman, Nathan Cermak, Désirée Schmitz, Julien Barrere, Alberto Pascual-García, and Otto X Cordero. Modular assembly of Polysaccharide-Degrading marine microbial communities. *Curr. Biol.*, 29(9):1528–1535.e6, May 2019. 73

Jacques Monod. Recherches sur la croissance des cultures bacteriennes. 1942. 74

Urs Lendenmann, Heinrich Senn, Mario Snozzi, and Thomas Egli. Dynamics of mixed substrate growth of escherichia coli in batch culture: the transition between simultaneous and sequential utilisation of carbon substrates. *Acta Universitatis Carolinae Environmentalica*, 14:21–30, 2000. 74

David W Erickson, Severin J Schink, Vadim Patsalo, James R Williamson, Ulrich Gerland, and Terence Hwa. A global resource allocation strategy governs growth transition kinetics of escherichia coli. *Nature*, 551(7678):119–123, November 2017. 74

Leonardo Pacciani-Mori, Andrea Giometto, Samir Suweis, and Amos Maritan. Dynamic metabolic adaptation can promote species coexistence in competitive microbial communities. *PLoS Comput. Biol.*, 16(5):e1007896, May 2020. 74

U Lendenmann, M Snozzi, and T Egli. Kinetics of the simultaneous utilization of sugar mixtures by escherichia coli in continuous culture. *Appl. Environ. Microbiol.*, 62(5): 1493–1499, May 1996. 74

Alan R Pacheco, Mauricio Moel, and Daniel Segrè. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat. Commun.*, 10(1):103, January 2019. 74

Xin Wang, Kang Xia, Xiaojing Yang, and Chao Tang. Growth strategy of microbes on mixed carbon sources. *Nat. Commun.*, 10(1):1279, March 2019. 74

D Stanojevic, L Comic, O Stefanovic, and Sl Solujic-Sukdolak. Antimicrobial effects of sodium benzoate, sodium nitrite and potassium sorbate and their synergistic action in vitro. *Bulgarian Journal of Agricultural Science*, 15(4):307–311, 2009. 74

W Harder and L Dijkhuizen. Strategies of mixed substrate utilization in microorganisms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 297(1088):459–480, June 1982. 74

Boris Görke and Jörg Stülke. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat. Rev. Microbiol.*, 6(8):613–624, August 2008. 74

Djordje Bajic and Alvaro Sanchez. The ecology and evolution of microbial metabolic strategies. *Curr. Opin. Biotechnol.*, 62:123–128, April 2020. 74

Sergio Sánchez, Adán Chávez, Angela Forero, Yolanda García-Huante, Alba Romero, Mauricio Sánchez, Diana Rocha, Brenda Sánchez, Mariana Avalos, Silvia Guzmán-Trampe, Romina Rodríguez-Sanoja, Elizabeth Langley, and Beatriz Ruiz. Carbon source regulation of antibiotic production. *J. Antibiot.*, 63(8):442–459, August 2010. 74

Caroll M Mendonca, Sho Yoshitake, Hua Wei, Anne Werner, Samantha S Sasnow, Theodore W Thannhauser, and Ludmilla Aristilde. Hierarchical routing in carbon metabolism favors iron-scavenging strategy in iron-deficient soil pseudomonas species. *Proc. Natl. Acad. Sci. U. S. A.*, 117(51):32358–32369, December 2020. 74

Ryosuke Fujiwara, Shuhei Noda, Tsutomu Tanaka, and Akihiko Kondo. Metabolic engineering of escherichia coli for shikimate pathway derivative production from glucose-xylose co-substrate. *Nat. Commun.*, 11(1):279, January 2020. 74

W Stanley Harpole, Jacqueline T Ngai, Elsa E Cleland, Eric W Seabloom, Elizabeth T Borer, Matthew E S Bracken, James J Elser, Daniel S Gruner, Helmut Hillebrand, Jonathan B Shurin, and Jennifer E Smith. Nutrient co-limitation of primary producer communities. *Ecol. Lett.*, 14(9):852–862, September 2011. 74

Jonas Cremer, Markus Arnoldini, and Terence Hwa. Effect of water flow and chemical environment on microbiota growth and composition in the human colon. *Proc. Natl. Acad. Sci. U. S. A.*, 114(25):6438–6443, June 2017. 74

Rubén Sánchez-Clemente, M Isabel Guijo, Juan Nogales, and Rafael Blasco. Carbon source influence on extracellular ph changes along bacterial Cell-Growth. *Genes*, 11 (11), October 2020. 74

James J Kozich, Sarah L Westcott, Nielson T Baxter, Sarah K Highlander, and Patrick D Schloss. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq illumina sequencing platform. *Appl. Environ. Microbiol.*, 79(17):5112–5120, September 2013. 83

J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann,

Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336, April 2010. 84

B J Callahan, P J McMurdie, M J Rosen, A W Han, A J A Johnson, and S P Holmes. DADA2: High resolution sample inference from illumina amplicon data. nat meth 13: 581–583, 2016. 84

Q Wang, G M Garrity, J M Tiedje, and J R Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. appl environ microbiol73: 5261–5267. *View Article*, 2007. 84

C Quast, E Pruesse, P Yilmaz, J Gerken, T Schweer, P Yarza, J Peplies, and F O Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. nucleic acids res41 (d1): D590–D596, 2013. 84

Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nat. Biotechnol.*, 28(3):245–248, March 2010. 89

A Ebrahim, J A Lerman, B O Palsson, and D R Hyduke. COBRApy: COnstraints-Based reconstruction and analysis for python. BMC syst biol. *COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol*, 7, 2013b. 90, 124

Brian J McGill, Brian J Enquist, Evan Weiher, and Mark Westoby. Rebuilding community ecology from functional traits. *Trends Ecol. Evol.*, 21(4):178–185, April 2006. 109

José L Martínez, Fernando Baquero, and Dan I Andersson. Predicting antibiotic resistance. *Nat. Rev. Microbiol.*, 5(12):958–965, December 2007. 109

Shingo Suzuki, Takaaki Horinouchi, and Chikara Furusawa. Prediction of antibiotic resistance by gene expression profiles. *Nat. Commun.*, 5:5792, December 2014. 109

Fernanda Pinheiro, Omar Warsi, Dan I Andersson, and Michael Lässig. Metabolic fitness landscapes predict the evolution of antibiotic resistance. *Nat Ecol Evol*, 5(5):677–687, May 2021. 109

Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021. 109

Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D Bushman, and James D Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, October 2011. 109

Catherine Burke, Peter Steinberg, Doug Rusch, Staffan Kjelleberg, and Torsten Thomas. Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. U. S. A.*, 108(34):14288–14293, August 2011. 109

Stilianos Louca, Saulo M S Jacques, Aliny P F Pires, Juliana S Leal, Diane S Srivastava, Laura Wegener Parfrey, Vinicius F Farjalla, and Michael Doebeli. High taxonomic variability despite stable functional structure across microbial communities. *Nat Ecol Evol*, 1(1):15, December 2016b. 109

Leonora S Bittleston, Matti Gralka, Gabriel E Leventhal, Itzhak Mizrahi, and Otto X Cordero. Context-dependent dynamics lead to the assembly of functionally distinct microbial communities. *Nat. Commun.*, 11(1):1440, March 2020. 109

Luis Miguel de Jesús Astacioa, Kaumudi H Prabhakara, Zeqian Li, Harry Mickalide, and Seppe Kuehn. Closed microbial communities self-organize to persistently cycle carbon. *Proc. Natl. Acad. Sci. U. S. A.*, 118(45), November 2021. 109

Elizabeth K Costello, Keaton Stagaman, Les Dethlefsen, Brendan J M Bohannan, and David A Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, June 2012. 109

Britt Koskella, Lindsay J Hall, and C Jessica E Metcalf. The microbiome beyond the horizon of ecological and evolutionary theory. *Nat Ecol Evol*, 1(11):1606–1615, November 2017. 109

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol

Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589, August 2021. 109

I N Shindyalov, N A Kolchanov, and C Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.*, 7(3):349–358, March 1994. 109

Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6(12):e28766, December 2011. 109

Ambrish Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, 5(4):725–738, April 2010. 109

Arion I Stettner and Daniel Segrè. The cost of efficiency in energy metabolism. *Proc. Natl. Acad. Sci. U. S. A.*, 110(24):9629–9630, June 2013. 111

Brice Enjalbert, Pierre Millard, Mickael Dinclaux, Jean-Charles Portais, and Fabien Létisse. Acetate fluxes in escherichia coli are determined by the thermodynamic control of the Pta-AckA pathway. *Sci. Rep.*, 7:42135, February 2017. 111

José M Peregrín-Alvarez, Chris Sanford, and John Parkinson. The conservation and evolutionary modularity of metabolism. *Genome Biol.*, 10(6):R63, June 2009. 111

Samir Giri, Leonardo Oña, Silvio Waschina, Shraddha Shitut, Ghada Yousif, Christoph Kaleta, and Christian Kost. Metabolic dissimilarity determines the establishment of cross-feeding interactions in bacteria. *Curr. Biol.*, 31(24):5547–5557.e6, December 2021. 112

Leonardo Oña, Samir Giri, Neele Avermann, Maximilian Kreienbaum, Kai M Thormann, and Christian Kost. Obligate cross-feeding expands the metabolic niche of bacteria. *Nat Ecol Evol*, 5(9):1224–1232, September 2021. 112

Teresa del Castillo, Juan L Ramos, José J Rodríguez-Herva, Tobias Fuhrer, Uwe Sauer, and Estrella Duque. Convergent peripheral pathways catalyze initial glucose catabolism in pseudomonas putida: genomic and flux analysis. *J. Bacteriol.*, 189(14):5142–5152, July 2007. 112

Marieke F Buffing, Hannes Link, Dimitris Christodoulou, and Uwe Sauer. Capacity for instantaneous catabolism of preferred and non-preferred carbon sources in escherichia coli and bacillus subtilis. *Sci. Rep.*, 8(1):11760, August 2018. 113

Adam C Martiny, Kathleen Treseder, and Gordon Pusch. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.*, 7(4):830–838, April 2013. 114

Jennifer B H Martiny, Stuart E Jones, Jay T Lennon, and Adam C Martiny. Microbiomes in light of traits: A phylogenetic perspective. *Science*, 350(6261):aac9323, November 2015. 114

V Sabarly, O Bouvet, J Glodt, O Clermont, D Skurnik, L Diancourt, D de Vienne, E Denamur, and C Dillmann. The decoupling between genetic structure and metabolic phenotypes in escherichia coli leads to continuous phenotypic diversity. *J. Evol. Biol.*, 24 (7):1559–1571, July 2011. 114

Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009. 121

M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, November 2007. 121

Bui Quang Minh, Minh Anh Thi Nguyen, and Arndt von Haeseler. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.*, 30(5):1188–1195, May 2013. 121

Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler, and Lars S Jermiin. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, 14(6):587–589, June 2017. 121

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37 (5):1530–1534, May 2020. 121

Kathleen Sprouffske and Andreas Wagner. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics*, 17:172, April 2016. 122

Lázaro Molina, Ruggero La Rosa, Juan Nogales, and Fernando Rojo. Pseudomonas putida KT2440 metabolism undergoes sequential modifications during exponential growth in a complete medium as compounds are gradually consumed. *Environ. Microbiol.*, 21(7): 2375–2390, July 2019. 122

Jorn Bruggeman, Jaap Heringa, and Bernd W Brandt. PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Res.*, 37(Web Server issue):W179–84, July 2009. 123

Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.*, 46(15):7542–7553, September 2018. 124