# Population and community ecology of bacteria from the genus *Streptomyces*

By

Erik S. Wright

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2016

Date of final oral examination: 11/28/2016

The dissertation is approved by the following members of the Final Oral Committee:

Kalin H. Vetsigian, Assistant Professor, Bacteriology
David A. Baum, Professor, Botany
Anthony R. Ives, Professor, Entomology
Katherine D. McMahon, Professor, Bacteriology
Michael G. Thomas, Professor, Bacteriology

# ABSTRACT

Microbial communities underlie an enormous number of natural processes, yet we have only recently begun to build a predictive understanding of their ecological and evolutionary dynamics. Modeling complex communities requires knowledge of the patterns of ecological interactions and their frequency dependence within and between species. The dearth of available information about the pattern and type of microbial interactions has limited our comprehension of ecological systems and our ability to engineer microbial consortia. Here, I use a variety of approaches to uncover how interactions shape the community and population ecology of a genus of naturally antibiotic producing bacteria, *Streptomyces*. I find that the outcome of laboratory competitions between strains is often surprisingly sensitive to the initial relative abundance of the strains. Furthermore, I find that even individual cells of the same strain can have dramatically different numbers of descendants after growing together for a short amount of time. Both of these results are linked to inter- and intra-species interactions that may be mediated by antibiotics or other small molecules. Hence, two overarching conclusions are drawn: (i) the importance of frequency dependence in determining the effects of interactions and (ii) that experiments with relatively simple microbial communities can lead to unexpected results. These conclusions have several immediate implications for how we undertake predictive modeling of microbial communities. They indicate that interactions exhibit inherent nonlinearities that predispose the communities to multiple stable states and make developing predictive models more difficult. Also, these results suggest that it is worthwhile to continue developing an understanding of simple microbial communities that can serve as a foundation for models attempting to predict dynamics in communities as complex as those regularly encountered in natural systems.

To the *Streptomyces*:
without whose sacrifice this would not exist.

# ACKNOWLEDGEMENTS

By far, the greatest challenge that I have encountered in graduate school has been to write the acknowledgements section of my thesis. There are at least an order-of-magnitude more people to acknowledge than there are pages in this thesis. I debated several approaches: (i) spend hours recollecting my life, hoping not to miss anyone; (ii) write a program to infer the most important people from our interactions over email; (iii) come to terms with the fact that I cannot possibly accomplish this task with perfection. Each approach has advantages and disadvantages, but none would provide a comprehensive list. Alas, I have decided on the latter approach, and I will therefore be remiss in my duty to thank all of those who have contributed. I can only hope that I will someday be forgiven for omitting so many important people from these acknowledgements.

When reflecting upon my life, I feel very fortunate to have worked with my PhD advisor, Kalin Vetsigian. I interviewed with Kalin on one of his first days as a new professor at UW-Madison, on a day that also happened to be his birthday. Both of us knew within a matter of minutes that we would eventually work together. When we finished talking science he told me that our discussion had been the best birthday present he could have imagined. That’s just how much Kalin loves science; it is contagious. I became the first student (a.k.a. *kindred spirit*) in his lab, and we undeniably have made a great team. I have learned more from Kalin than I could ever list. One thing above all, though, deserves mention: Kalin treats members of his team as if they were members of his family. I will never forget just how much kindness and support he has given me, especially when my wife had our first child while I was in graduate school.

Kalin was one of many spectacular mentors I have had in graduate school. During my M.S. I was advised by Daniel Noguera, who single-handedly converted me into an academic. I was incredibly lucky that Dan took me under his wing, and he continues to mentor me to this day. I have the utmost respect for Dan, and I regret that my Master's thesis did not include a proper Acknowledgements section. It has been equally fortuitous to have been mentored by four post-doctoral scholars while I was in graduate school. The first, Safak Yilmaz remains a good friend despite having moved to a different city. The second, Sriram taught me a great deal of mathematics that I had long forgotten or never learned. The third, Ye Xu is an exemplary postdoc and sets the standard for all other postdocs. The fourth, Sailendharan Sudakaran is a master strategist, and wherever I end up next will be in no small part due to his help. Although not a postdoc, John Barkei had a huge impact on me while he was lab manager, during which time we went from an empty (new) lab to a full one. They each came into my life at the moment when I needed them most, and I can only hope that I will someday figure out how to manage without any of them being nearby.

I am exceedingly grateful for the wonderful people of the Center for High Throughput Computing (CHTC) at the UW-Madison. I have been immensely assisted by three of their incredible staff: Christina Koch, Lauren Michael, and Bill Taylor. Since they first introduced me to HTC, I have used it in almost every project, and it wasn't until preparing to leave campus that I *completely* realized just how blessed I had been to have access to this resource. Furthermore, the director of the CHTC, Miron Livny, advocated for my creation of a new class within the Computer Sciences Department. Because of him I was able to teach students from all over campus how to program, which was one of the best experiences I had in graduate school. I have been unbelievably fortunate to have been given this opportunity as a graduate student, and I am extremely thankful to the CHTC team for their backing.

I am no less appreciative of my students, who have put up with all of my eccentricities over the years while I have learned to become a better teacher. Admittedly, they all deserved more credits than they ever received in my class, but hopefully the empowerment of learning to program was worth it. Likewise, I have had the pleasure of closely mentoring 11 students in laboratory research while I was a graduate student. Each student was different and I learned something new every time. They are largely responsible making me into a better mentor, which will probably always be a work in progress. It has been a gift for me to watch each one grow and eventually leave for a new adventure. Also worth noting, one of the greatest parts of being at the UW-Madison was the ability to take courses on teaching and mentoring that have extensively sculpted my style.

Each Thanksgiving, when we share what we are thankful for in the past year, I have a slightly different list than I had the prior year. This year I am certainly thankful for my thesis committee members, who will come together four days after Thanksgiving to decide my fate. Each one of them has had a huge impact on my PhD. Trina McMahon served on both my M.S. and PhD committees, and she was both incredibly supportive and fun! Anthony Ives always asked the questions that would steer me in a new direction at our yearly committee meeting. Michael Thomas taught me a lot about antibiotics, in spite of chemistry being my Achilles heel, and gave me ample guidance as I decided on my next step in life. David Baum taught me evolution, and we ended up converting the term paper I wrote for his class into a manuscript. I hope that someday soon it is published. Every one of my committee members was an inspiration to me. Thank you all.

Finally, the love and support of my family has been invaluable throughout my life. My parents have taught me the importance of both ancestors and descendants, at least one of whom I hope discovers this thesis someday, as if it were a time capsule buried in a library somewhere. Moreover, they instilled in me a view of education as one of the highest ideals, which likely was passed-down to them from an ancestor of long-ago. My wife's parents and siblings are no less supportive, and I have always cherished their friendship. Of course, my wife is largely attributable for any success that may come my way. We have already taken many adventures together, and I am looking forward to those that lay ahead. Last but not least, I thank my barefoot boy (now almost 2 years old) for making me take a step back from work to enjoy the bigger picture.

## TABLE OF CONTENTS

## LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

It is currently, by many accounts, an amazing time period in which to live. The pace of change and our ability to affect the future are seemingly beyond that of almost any other time in history. This is particularly true in the broad field of biology, where we are just beginning to comprehend and manipulate our natural world. Having come from engineering, I was relatively unfamiliar with biology when I first began graduate school. Since day one I have found studying biology to be awe inspiring: as if I am an explorer at the forefront of a great expedition. Where it will take us I cannot know, but it is clear that an incredible journey lays ahead.

Of the several excursions I have so far taken into the realm of biology, three are included in this thesis, and I present them in the order in which they occurred. The first project involves understanding how interactions between different organisms shapes their ecological network. The second project delves into one of the obstacles uncovered during the first project, that is, how to overcome errors associated with high-throughput DNA sequencing. The third project reduces to only a single population of bacteria, and illustrates that even the simplest questions can have surprising answers.

In the next sections I attempt to take stock of the general lessons that I have learned from these three projects, their overarching themes, and how they are interconnected.

## Why use microbes to research ecology?

Ecologists, by definition, are concerned with the relationships among organisms and between organisms and their environment. Historically, the organisms of interest have been *macro*organisms such as ourselves, while *micro*organisms were as much an afterthought as they were invisible. In spite of this, microbes carry on providing a vast array of “ecosystem services” (for lack of a better term) and interactions, as they have for eons. In many ways, microbes are not just part of the ecosystem, they determined the ecosystem. Yet, this is not the only reason to study ecology from the microbial perspective. Along with relevance, an ecologist may seek a tractable experimental system, one that can readily be manipulated to quickly obtain new results. Ideally, such a system would allow the ecologist full flexibility in controlling its every aspect: its components (both biotic and abiotic), its structure, its size, its timeframe, and its scale.

Herein lies another major advantage to performing ecology research with microbes. Microbiologists can perturb many levels of scale: single nucleotides, genes, populations, communities, and ecosystems. While it is (*presumably*) cumbersome to manipulate the genome of an elephant and then grow a population of one billion elephants, with microbes this is comparatively trivial. In effect, we can match the relevant scale to the problem we hope to solve or iteratively switch between scales as needed. For example, at the beginning of the first project in this thesis, I began by assembling multi-species communities of different microbes in the laboratory. Although already much simpler than natural microbial communities, these synthetic communities were exceedingly complex, and stepping down to the level of two species communities resulted in enough simplicity to permit an interpretation of what was observed. It was only later when “scaling-up” again to multi-species communities that it was possible to explain these observations in light of information collected at smaller scales.

### Bottom-up versus top-down approaches to microbial ecology

A major difficulty of studying natural systems is their inherent complexity. Population abundances of elephants are connected to population abundances of grasses (*and elephant poachers*), which are in turn coupled to the weather, and so forth. We can control for many of these factors in laboratory experiments, with the downside that we must simplify some realities of complex systems in order to do so. The hope is that understanding a simplified system lends itself to understanding a real system. While the verdict is still out as to whether this is the case, it is (so far as I can tell) the only option to go beyond surveying approaches when studying microbial ecology. Here we are posed with a fundamental choice between two philosophically different strategies. In the *top-down* approach an existing microbial system is experimentally constrained and we study the community that emerges. Examples of this are transferring some seawater to a flask under a light source (Benincà et al. 2009), or inoculating a strip of cellulose paper with soil (Lewin et al. 2016). In such cases the complex community that once existed in nature will be reduced to a simplified community that can be studied in a controlled system.

The alternative is to begin with a panel of isolated organisms and combine them in such a manner as to directly elucidate aspects of the community assembly process. This *bottom-up* approach offers the most control over the resulting community's composition, but does not necessarily result in a diverse community. Most attempts at combining microorganisms in this manner, going back to the days of Gause's seminal experiments with *Paramecium* (see Gause 1935), have simply given rise to a single species community after one organism outcompetes all of the others. At the outset of undertaking the first project in this thesis, it was unclear what fraction of simple communities would have such a mundane fate. This is because, despite the immense popularity of surveying bacteria in natural environments, we still know very little about the fundamentals of how microbial communities arise in the first place. This is particularly true for microorganisms other than *E. coli*, and for laboratory ecosystems that are not established in well-mixed liquid environments.

## Constructing microbial communities

The rational design of synthetic microbial communities is a topic that has recently received considerable attention (Widder et al. 2016). To this end, the first goal has been to build a predictive understanding of how microbial communities will assemble or respond to perturbation. This is useful both for engineering microbial consortia to perform a given task (e.g., converting stover to ethanol) or for manipulating the composition of an existing microbial community (e.g., transitioning a gut microbiome to one without *Clostridium difficile*). Initially, experiments in constructing microbial communities from the bottom-up relied upon resource dependencies between community members to ensure coexistence (Wintermute and Silver 2010). This approach was later broadened to other interactions that would lead to coexistence in spatially structured environments (Kim et al. 2008; Majeed et al. 2010) or unstructured environments with predator-prey interactions (Hekstra and Leibler 2012; Friman et al. 2015). These types of lab communities avoid survival of the fittest by directly preventing one community member from single-handedly dominating all of the others.

Subsequent work in this field began to develop a phenomenological understanding of the community assembly process for microorganisms without prescribed interactions. The most common approach has been to systematically combine different microbes of interest and observe their dynamics (Rivett et al. 2016; Faith et al. 2014). An alternative approach, which was taken in the first project in this thesis, is to combine microbes at drastically different abundances, or (similarly) give some a chance to establish before adding others (Fukami et al. 2010). More recent work attempts to extend quantitative models for predicting the dynamics of mixed microbial communities (Trosvik et al. 2008; Friedman, Higgins, and Gore 2016). In such studies it is useful to quantify the resource-dependent and resource-independent interactions between organisms. For this reason, microbes with strong, measureable inter-population interactions are particularly useful for building predictive models.

### The marvelous *Streptomyces*

Microorganisms are incredible. No more so perhaps than members of the genus *Streptomyces*, which despite their name (meaning 'twisted' + 'fungus' in Greek) are bacteria. It is presently unclear why it is the case, but *Streptomyces* have given us almost half of the antibiotics that we use in the clinic. This has always struck me as puzzling, that is, why one genus out of thousands of described bacterial genera has developed such a special metabolic repertoire. Perhaps the answer lies in their habitat, as another group of microorganisms occupying a similar niche, the fungi, are also a major source of antibiotics. Both have a complex lifecycle involving sporulation, and both are saprophytes that live off of decaying organic matter in soil. *Streptomyces* are in fact so ubiquitous in soil that from a single spec almost too small to see with the naked eye, it is possible to isolate tens of different Streptomycetes. This is how I ended up amassing most of my collection of *Streptomyces*, all from a few neighboring specs of soil. Despite previously living together in a small space, these Streptomycetes had very strong interactions, as displayed by large zones of inhibition when grown together. The overarching question behind this thesis is therefore: *What are the ecological consequences of Streptomyces' complicated lifecycle and strong interactions?*

The first project in this thesis was conceptually simple: combine different *Streptomyces* and watch what happens. It may seem silly to study quasi-random assortments of bacteria in a laboratory system far removed from the realities of nature. However, if one takes the perspective that nature is everywhere then a laboratory ecosystem is just as valid as any other ecosystem. Furthermore, such assortments reflect the reality that microbes have high rates of dispersal. Thus, there is some non-zero probability that any microbe might be transferred to another's domain on a given day. Whether it can survive (or even invade) there is the question. Moreover, I tried to construct a laboratory microcosm that reflected the natural environment of soil bacteria: one having feast-and-famine cycles, spatial structure, and complex food sources. To accomplish this in high-throughput required engineering a device that would enable growing bacteria on the inside walls of test tubes, the development of which remains a point of pride for all those who were involved.

**Entering the era of high-throughput ecology**

Improvements in DNA sequencing are rapidly changing the practice of biology research. All of the projects in this thesis involve sequencing DNA using a relatively new technology being offered by a company named Illumina. At present, the most common approach is to sequence many samples simultaneously and then tell them apart based on unique "index" sequences that are incorporated into each sample. Using combinations of a few different index sequences allows hundreds of different samples to be multiplexed into the same sequencing run. Unfortunately, the de-multiplexing process is imperfect, and sequences are sometimes associated with the wrong sample. Surprisingly, despite the popularity of multiplexing, there is very little in the way of literature describing this pitfall. Such was the subject of the second project where I systematically studied and developed a solution to the de-multiplexing problem. The outcome of this project was a relief for two reasons: (i) I eventually learned that I was not the only one who had experienced the problem (*misery loves company*), and (ii) solving the problem turned out to be remarkably straightforward.

The third project in this thesis involved using multiplexed DNA sequencing to reveal a fundamental property of bacteria: the descendants distribution. That is, each bacterium in a population gives rise to a number of descendants after some period of time. How variable is this number? Does each bacterium have roughly the same number of offspring, or do some bacteria yield most of the offspring in a population? This is a deceptively simple question, and one that is impossible to answer without the ability to track sub-populations within a population. For this I made use of special *Streptomyces* (created by another researcher) that differ only by a short genetic "barcode" integrated into their chromosome. This project spanned many facets of biology: genetics, lab experiments, DNA sequencing, and computation. Moreover, the descendants distribution turned out to be *wild* (both literally and figuratively), which was more interesting than anyone had expected. This project illustrates how even the most basic questions can have surprising answers. Quite simply, it is science at its best.

# CHAPTER 2: INHIBITORY INTERACTIONS PROMOTE FREQUENT BISTABILITY AMONG COMPETING BACTERIA

## INTRODUCTION

Microbes undergo high rates of dispersal and intermixing in nature but are constrained in their ability to colonize new habitats by environmental filtering and local competition (Martiny et al. 2006). After undergoing the process of community assembly, established microbial communities are continually challenged by immigrants from other microbial communities that have the potential to upset the existing community balance. An understanding of the assembly and resilience of communities therefore requires knowledge of interactions among established community residents and outsiders that may arrive at low abundance though dispersal (Drake 1991; Ives and Carpenter 2007). Invasion experiments are a minimal way of capturing such competitive outcomes at the two extremes of relative abundance (Fig. 2.1). These pairwise relationships can be assembled into an invasibility network, which characterizes the effective interactions between microbes and their frequency-dependence (Chesson 2000).

This chapter was originally published in:

Wright, E. & Vetsigian, K. (2016). Inhibitory interactions promote frequent bistability among competing bacteria. Nature Communications, 7, 11274. http://doi.org/10.1038/ncomms11274.

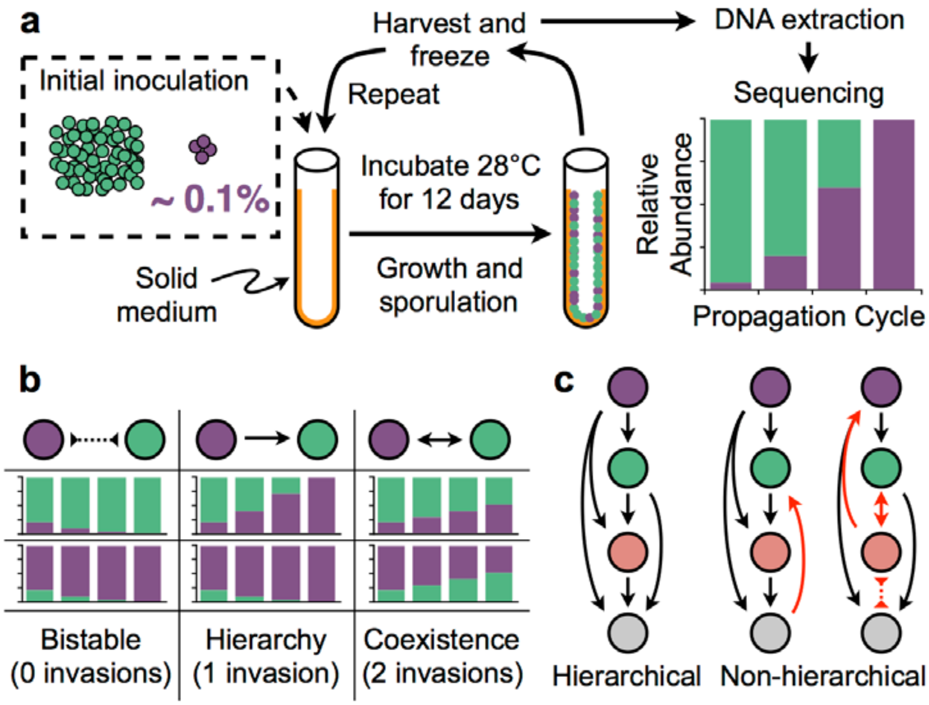


**Figure 2.1. Scheme for measuring pairwise invasions and potential outcomes. a**, Pairs of bacterial strains were added to test tubes at vastly different initial abundances and propagated for 3 cycles. Their relative abundance was quantified with high-throughput sequencing. **b**, Each strain from a pair was competed twice, as either the resident (high abundance) or invader (low abundance). The three potential outcomes are: bistability if neither resident was invaded, hierarchy if one resident was invaded and coexistence if both residents were invaded. **c**, The invasion network for a panel of strains may be either completely hierarchical, where strains can be ranked by relative fitness in a way that explains all pairwise outcomes, partly hierarchical with a small fraction of non-hierarchical features, or essentially non-hierarchical.

Even a basic knowledge of the statistical properties of invasibility networks for microbes from similar habitats can greatly enhance our understanding of the processes that structure natural communities (Davis, Thompson, and Grime 2005). The simplest expectation, based on the competitive exclusion principle (Kassen and Rainey 2004), is that most microbial strains can be ordered according to their competitive ability, which would result in a hierarchical network composed of asymmetrical invasions (Fig. 2.1) and dynamics dominated by the survival of the fittest for any particular environment (Foster and Bell 2012). On the other hand, networks exhibiting cyclic dominance (as in rock-paper-scissor games) can lead to diversity maintenance or alternating winners (Vetsigian, Jajoo, and Kishony 2011; Majeed et al. 2010; Kelsic et al. 2015).

Widespread negative frequency dependent selection would imply advantage for rare variants, which can promote diversity, whereas positive frequency-dependent selection can lead to alternate stable states (Petraitis 2013) and historical contingency (B. R. Levin 1988; Chao and Levin 1981). Despite their utility, the statistical properties of invasibility networks are generally unknown (Oliveira, Niehus, and Foster 2014; Coyte, Schluter, and Foster 2015; Allesina and Tang 2012).

To start filling this knowledge gap, we determined the invasion and antibiotic inhibition networks for a diverse panel of bacteria from the genus *Streptomyces* (Appendix A: Supplemental Fig. 2.1), most of which were isolated from neighboring grains in the same soil sample (see Methods). These bacteria are ubiquitous in soil and are prolific producers of antibiotics and other secondary metabolites when grown on a solid substrate (Vetsigian, Jajoo, and Kishony 2011). Understanding the ecological consequences of diverse secondary metabolites has been a major challenge (Ratcliff and Denison 2011; Andersson and Levin 1999), inspiring many theoretical and experimental works (Kinkel et al. 2013; Kelsic et al. 2015; Durrett and Levin 1997; Cordero et al. 2012). However, the role of secreted bioactive compounds in generating intransitive or frequency-dependent relationships has not been investigated systematically among large collections of microbes, particularly in unmixed environments, in which secreted molecules stay close to their producers.

We find that 'survival of the common' is nearly as widespread as 'survival of the fittest' among competing soil bacteria from the genus *Streptomyces*. The winner of a pairwise competition is often the species that starts at high initial abundance, making it impossible to completely rank the species based on their competitive ability. Instead of a single winner, the tournament between bacteria results in multiple winners that are in bistable relationships with each other. We also find that inhibitory interactions between species are an important factor shaping the network of invasions, and such inhibitory interactions promote survival of the common. These findings have several immediate implications for how we view the assembly, structuring, and diversity of microbial communities. They indicate that pairwise interactions lead to inherent nonlinearities that predispose communities towards multiple stable states. This may make microbial communities intrinsically sensitive to initial conditions during community assembly but, at the same time, could make them more resistant to change once they are established. Survival of the common may also promote mosaic spatial distributions with different populations dominating different patches or microbial hosts despite similar abiotic conditions.

## RESULTS

### Frequent bistable relationships between pairs of strains

To measure invasion, we inoculated a pair of strains at vastly different initial abundances inside a thin layer of solid (agar) defined medium and allowed them to grow and sporulate (Fig. 2.1a). Offspring spores were then collected from the surface of the agar and then used to inoculate another propagation cycle or determine relative abundances with high-throughput sequencing (see the Methods for details). After three propagation cycles, strains were said to invade if they had increased in abundance to at least 1% of the total community. Typically, invasions occurred rapidly, and the invader had almost completely displaced the resident within one or two propagation cycles (Appendix A: Supplemental Fig. 2).

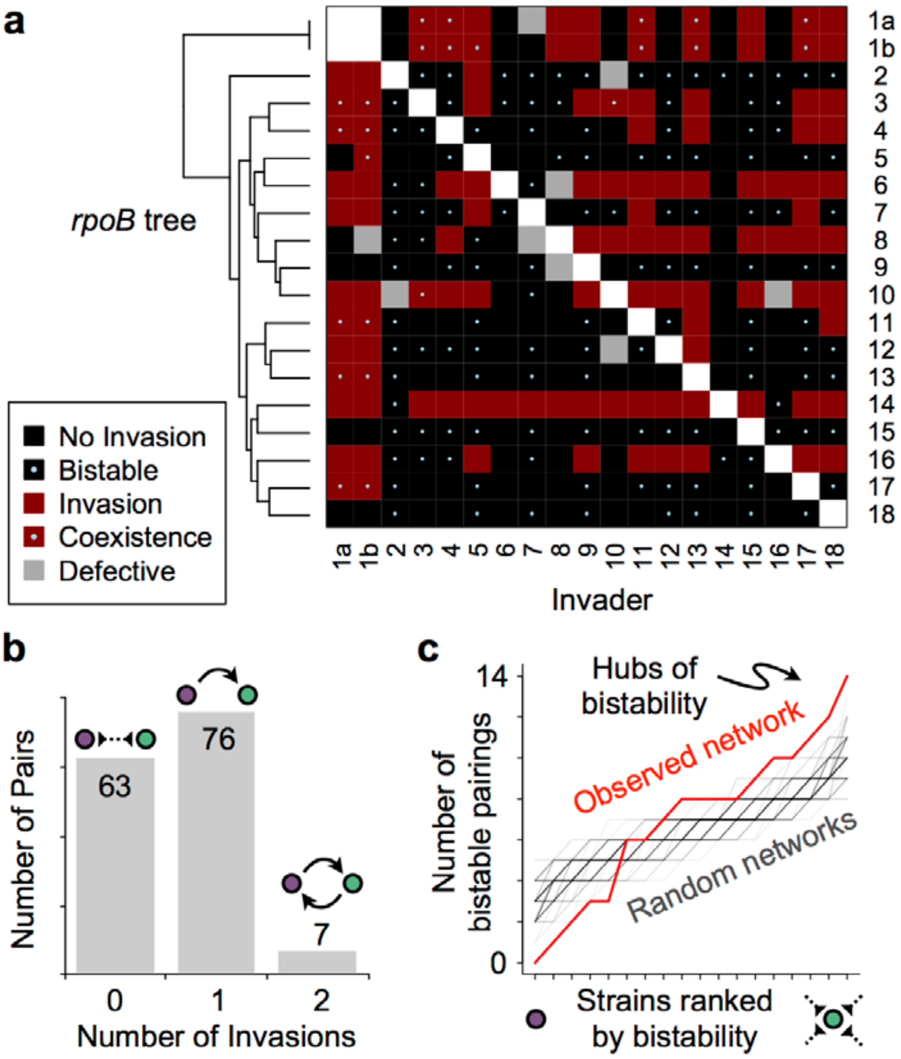


**Figure 2.2. Widespread bistability in pairwise invasions. a**, Pairwise invasion matrix for a panel of 18 diverse *Streptomyces* strains. Strains are sorted by phylogeny constructed from partial *rpoB* gene sequences. Strain #1 is present in two replicas (labeled 1a,b). b, Bistable pairings, in which two strains cannot invade each other, were a dominant feature of the invasion matrix. Coexistence was less frequent and mostly limited to strain #1, which was also the most phylogenetically distinct strain. **c**, A few strains were involved in many bistable pairings. These "hubs of bistability" were more frequent than in randomized matrices with the same number of each type of pairwise link (p = 1.7e-4).

We began by analyzing pairwise features of the invasion matrix. Invasions were highly repeatable, as we only observed a single difference between 32 replicate competitions performed with strain #1 (Fig. 2.2a). Overall, 31% of pairwise competitions resulted in an invasion (Fig. 2.2b). No strain was invaded by all other strains in the panel, although one strain (#14) was invaded by all but two others. Three strains were not invaded by any other strain, indicating that the strains cannot be ordered in a strict hierarchy. Six of seven cases of mutual invasion included strain #1 (Fig. 2.2b), which was also the most distantly related strain as it belongs to a separate genus (Fig. 2.2a). Mutual invasions are expected to lead to coexistence because neither strain can reach a low enough abundance that it is unable to recover. Accordingly, in all seven cases, the pairs of mutually invading strains were both found to be present at the end of three propagation cycles. In sharp contrast to the low number of mutual invasions, there were 63 mutually non-invading pairs of strains, where the most abundant strain was able to hold its ground against the less abundant (Fig. 2.2b). These bistable links centered on a small subset of strains that rarely invaded others and were rarely invaded by others, and therefore acted as "hubs of bistability" (Fig. 2.2c).

### Partly hierarchical invasion network with multiple winners

We next characterized triplet motifs in the invasion network relative to random networks with the same number of each type of pairwise link. We observed a strong enrichment for transitivity of hierarchy: given that strain A invades B and B invades C, A most likely also invades C (Fig. 2.3a). Surprisingly, we did not observe a single instance of the 'rock-paper-scissors' dynamic (C invading A). Similar to hierarchy, bistable links were also greatly enriched for transitivity (Appendix A: Supplemental Fig. 3).

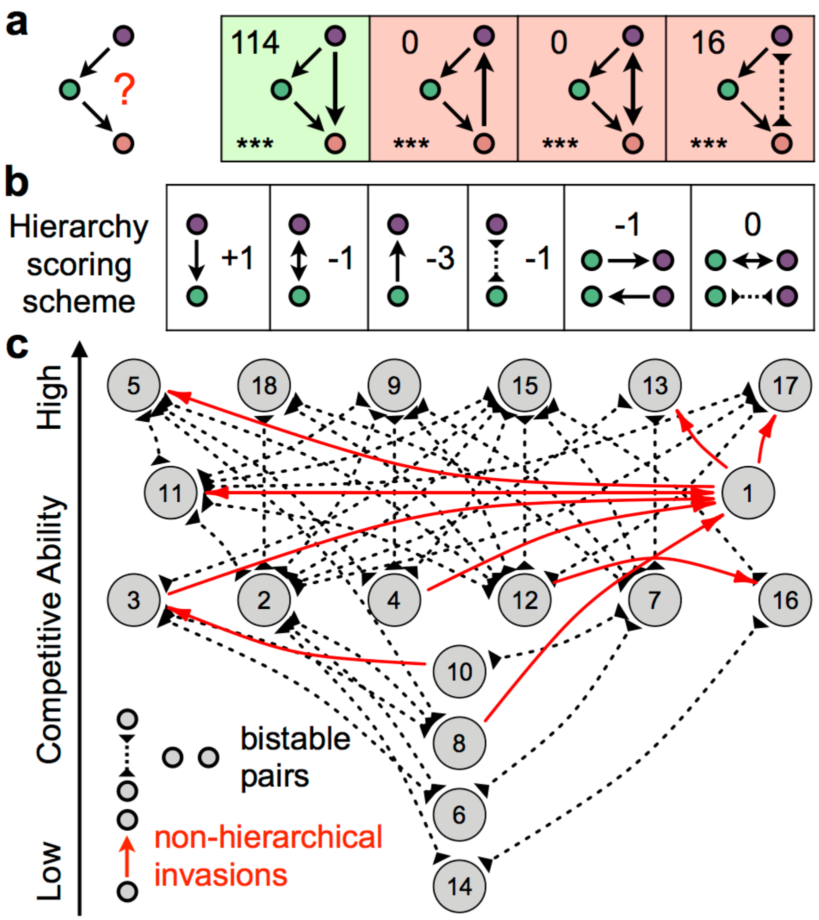


**Figure 2.3. The invasion network is partly hierarchical with multiple strains in the top level exhibiting bistable relationships with each other. a**, Enrichment (green) or depletion (pink) of different triplet motifs relative to randomized networks preserving the number of each pairwise motif (Appendix A: Supplemental Fig. 3). Number of occurrences for each motif are given in the upper-left corner. Transitive invasions (leftmost motif) were highly enriched (***, p < 1e-6), whereas the three intransitive motifs were highly depleted (***, p < 1e-6). **b**, The scoring scheme used in assigning hierarchy levels to strains rewards invasions pointing down the hierarchy and penalizes invasions directed against the hierarchy. **c**, The invasion network overlaid on the hierarchy assignments that maximize the score in b. Strains were placed into seven hierarchy levels with six strains at the top level exhibiting bistable relationships with each other. Invasions going down the hierarchy are not shown, while others are shown in red. Bistability is denoted by a missing link between strains at the same level or by a dashed line for strains at different levels.

The strong enrichment for transitivity of hierarchy motivated us to order species according to their competitive ability by determining the rank assignments that are most congruent with the observed invasibility network. To accomplish this, we developed a simple scoring scheme that rewarded invasions directed down the hierarchy, and penalized invasions going against the hierarchy (Fig. 2.3b). Under this scoring scheme, the optimal hierarchy placed the strains into seven levels (Fig. 2.3c). Only a few links were directed against the hierarchy, and most of them were due to mutual invasions with strain #1. Instead of a single fittest strain, six strains were tied for the top ranking, and, remarkably, all of them were in bistable relationships with each other. Thus, the hierarchical structure of the invasibility network revealed six mutually exclusive winners. Fascinatingly, there was considerable bistability between strains belonging to different hierarchical levels. In many of these cases an invader strain from the top of the hierarchy could form visible colonies or inhibition zones against a strain below it in the hierarchy, yet ultimately failed to invade (Appendix A: Supplemental Fig. 4). Furthermore, although strains exhibited widely different yields and growth rates in our experimental system, both measures were uncorrelated ($R^2$=0.006 and 0.174, respectively) with hierarchy level (Appendix A: Supplemental Fig. 5).

### Inhibitory interactions promote bistability

Given that bacteria from the genus *Streptomyces* are prolific antibiotic and siderophore (Griffin, West, and Buckling 2004; Darch et al. 2012; Keller and Surette 2006) producers, we hypothesized that inhibitory interactions played a major role in determining the hierarchy and generating bistability. To systematically examine the role of inhibition, we measured each strain's ability to prevent sporulation of other strains (Appendix A: Supplemental Fig. 6). The data revealed a strong tendency for inhibitions to point down the hierarchy (Fig. 2.4a), which is consistent with the notion that inhibition provides a competitive advantage (Ratcliff and Denison 2011). We found that strains were extremely unlikely to be invaded by strains they inhibit (p=1e-6, Fig. 2.4b), and there was a small increase in the probability to invade if inhibition was present (p=0.02, Appendix A: Supplemental Fig. 7a).

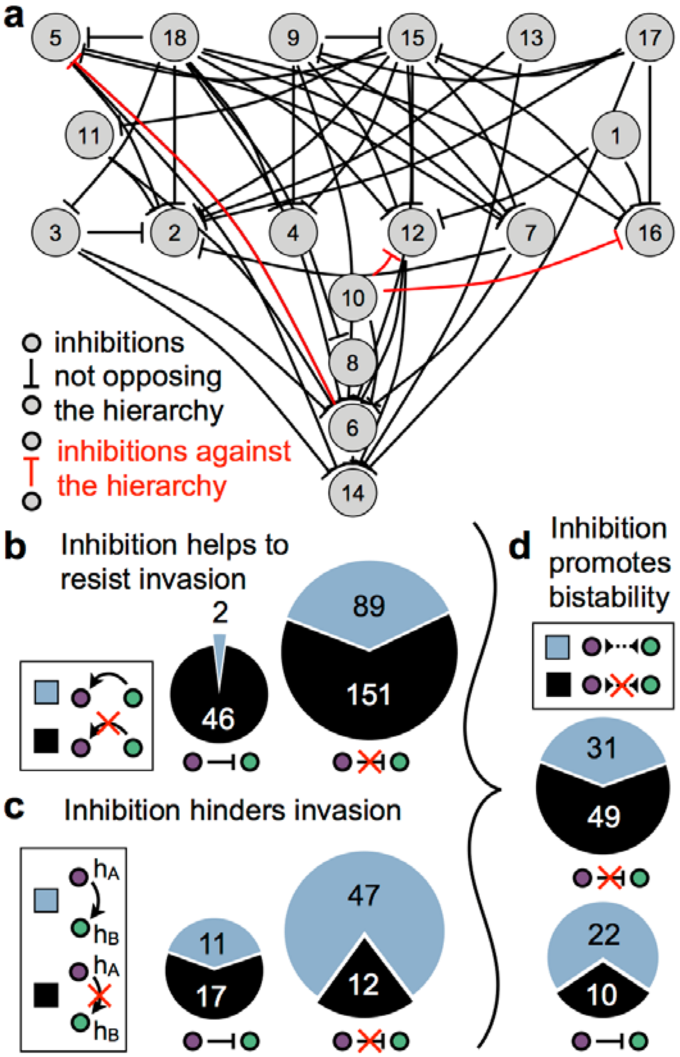


**Figure 2.4. Antibiotic inhibition helps bistability.** **a**, Inhibitions in the cross-streaking assay (Appendix A: Supplemental Fig. 6) overlaid on the invasion hierarchy. Almost all inhibitions were directed down the hierarchy (black inhibition arrows) and only a few were directed against the hierarchy (red inhibition arrows). **b**, The number of pairs with invasions (blue) and non-invasions (black) are shown for cases in which the invader is inhibited (left pie-chart) or not inhibited (right pie-chart). The vastly different fraction of invasions in the two pie-charts indicates that inhibition greatly assists residents in resisting invasion. **c**, The number of pairs with invasions (blue) and non-invasions (black) are shown for cases in which the resident is inhibited (left pie-chart) or not inhibited (right pie-chart). Only pairs in which the invader is at a higher hierarchy level ($3 \geq h_A - h_B > 0$) are considered to control for the tendency of inhibitions to point down the hierarchy. **d**, The number of bistable (blue) and non-bistable (black) pairs is shown for pairs with inhibition (bottom pie-chart) and without inhibition (top pie-chart). Strains #14 and #6 at the bottom of the hierarchy were not included, as they were invaded by almost all other strains. A significant enrichment for bistability is evident among pairs with an inhibitory interaction.

An alternative explanation for why inhibitions point down the hierarchy is that the hierarchy and the direction of inhibition are shaped by a common factor. For example, species better adapted to the growth medium may tend to outcompete other species while also having a head start in antibiotic production. To investigate this possibility, we reconstructed the hierarchy using only species pairs without inhibition. The new hierarchy was similar to the original one ($R^2$=0.916; Appendix A: Supplemental Table 1) and still exhibited a pronounced tendency for downward inhibitions (Appendix A: Supplemental Fig. 8), suggesting that a common factor was at least partially responsible for both the hierarchy and the direction of inhibition.

To control for this confounding effect, we recomputed the correlations between inhibition and invasion while focusing on pairs of strains with similar differences in hierarchy levels (see the Methods for details). Unexpectedly, this revealed that strains that inhibit other strains are significantly less likely to invade (Fig. 2.4c, p=1.3e-4). Hence, the small apparent increase in probability to invade as an inhibitor was entirely due to the tendency of inhibition to go down the hierarchy. This finding is consistent with the notion that investment in public goods might be counterproductive when the producers are at low abundance.

We concluded that although inhibitions likely help strains resist invasions at high-abundance, they also reduce their probability to invade when at low-abundance. The combination of these effects leads to a higher probability of bistability in pairs of strains in which there is inhibition (p=0.028). This enrichment was particularly pronounced after removing the two outlier strains at the bottom of the hierarchy, which are inhibited and outcompeted by almost everyone (Fig. 2.4d, p=0.002). This indicates that inhibition is one of the mechanisms promoting bistability.

## DISCUSSION

The finding of frequent bistability among *Streptomyces* isolated from the same environment is consistent with a recent theoretical work demonstrating that the counteraction of antibiotic production and degradation can lead to stable coexistence of many bacteria with different production and degradation capabilities (Kelsic et al. 2015). Extending this previous work, we proved that multi-species communities coexisting through this mechanism must contain bistable pairs (see the Methods for details). Thus, frequent pairwise bistability is expected theoretically among sets of three or more strains coexisting through this mechanism. Although it is unknown whether the strains used in this study would coexist in larger communities, this study demonstrates that, in addition to being prolific producers and degraders of antibiotics, *Streptomyces* exhibit the frequency-dependent relationships necessary for coexistence through an interplay between antibiotic production and degradation.

Although we found a link between inhibition and bistability, it is important to note that many bistable pairs had no measurable inhibition. These may be cases where inhibition was below our detection limit or was only partial and therefore did not lead to a visible zone of clearing. Bistability in these cases may also be due to other phenomena of intra-species cooperation that incur a cost of rarity (Allen, Gore, and Nowak 2013), including quorum sensing (Diggle et al. 2007) and the secretion of public goods such as extracellular enzymes. The existence of multiple life-stages, such as the germination – mycellium growth –sporulation lifecycle of *Streptomyces*, can also facilitate bistability (Moll and Brown 2008). Delineating the relative contribution of each of these potential causes of bistability in different microbial systems offers an interesting topic for future research.

Frequent bistability between bacterial strains has major implications for community assembly and structuring. First, it implies that the order of species arrival to a new environment could strongly influence the long-term community composition. In particular, pairwise bistability is expected to propagate to multistability in communities with more than two species. Furthermore, bistability may lead to an extreme sensitivity to initial abundances when many species arrive simultaneously, while also making established communities more resistant to invasions. These effects may manifest spatially through the generation of mosaic microbial distributions in which different communities are maintained in different spatial locations (Bayley et al. 2007) or in different individuals from a plant or animal host species. In particular, frequent bistability may help explain the recent finding that soil bacterial and fungal communities exhibit higher levels of dissimilarity across locations than expected from models assuming that environmental filtering and dispersal are the primary drivers of community assembly (Powell and Bennett 2015; Powell et al. 2015). Finally, on longer timescales, positive frequency-dependent selection could encourage the emergence of discrete microbial types as is typical for larger organisms (Bernstein et al. 1985), rather than a continuous spectrum of forms.

We expect the finding of widespread survival of the common to generalize to other microorganisms that produce strain-specific public goods, particularly when they are grown in unmixed environments and have strong resource overlaps (Vannette and Fukami 2013; Kinkel et al. 2013). Further insights into the survival of the common and its ramifications for ecosystems are likely to emerge from continued research on multi-species microcosm communities and gnotobiotic organisms.

## METHODS

### Isolation of Streptomycetes

The panel of 18 strains (Appendix A: Supplemental Table 1) used in this study included 13 *Streptomyces* isolates originating from the same soil sample. This sample was collected from the University of Wisconsin-West Madison Agricultural Research Station on 10 June 2014. The soil in the collection area was composed of Troxel silt loam at 1-3% slopes. A soil core was collected using a sterile 50 ml conical tube (VWR). A 0.25 gram piece of soil was extracted from a depth of 4 cm below the surface. This soil was separated into individual grains (~1 mg), each of which was used to inoculate a Petri dish containing Actinomycete Isolation Agar (AIA: 4 g/l Sodium Propionate, 10 g/l Soluble Starch, 0.4 g/l Sodium Caseinate, 2 g/l $KNO_3$, 2 g/l NaCl, 2 g/l $K_2HPO_4$, 0.05 g/l $MgSO_4$, 0.02 g/l $CaCO_3$, 0.01 g/l $FeSO_4$, 18 g/l Agar; pH adjusted to 7.5; to prevent fungal growth, cyclohexamide added after autoclaving to reach 50 mg/l).

Separated isolates on each Petri dish were automatically detected and pinned into individual wells of a 96 well plate containing Actinomycete Isolation Agar medium using an automatic colony picker (Hudson). To distinguish isolates, we sequenced a 935 base pair region of their DNA-directed RNA polymerase subunit β (*rpoB*) gene, which is commonly used as a species-level phylogenetic marker for *Streptomyces* (Doroghazi et al. 2014). We identified 28 distinct *Streptomyces* isolates, and selected 13 for invasion experiments based on having sufficient growth for sustained propagation on the defined medium used in the invasion experiment. These 13 isolates were named by their soil grain and microplate well. For example, sp. *S25E2* and sp. *S25H8* were both isolated from the same soil grain (S25). The remaining five strains used in this study were collected from various sources (Appendix A: Supplemental Table 1). Using the R (R Core Team 2016) package DECIPHER (E. S. Wright 2015), these 18 strains' *rpoB* sequences were aligned to those from related species in order to create the phylogenetic tree shown in Appendix A: Supplemental Fig. 1.

**Invasion experiments**

Before the start of the invasion experiment, the 18 strains were individually propagated for three growth cycles to reach equilibrium concentration on the defined medium (see below) used in the invasion experiments. A 50 µl aliquot of the *resident* strain at its equilibrium concentration was added to 80 µl of the diluted *invader* in the initial inoculum. The concentration in colony-forming units (CFUs) of each spore stock was determined using standard dilution plating. Each strain was diluted to achieve a target concentration of 100 CFUs per tube as *invader* in the first inoculation. The *invader* strains were again counted after dilution to determine their actual concentration in the tubes, and were generally close to the desired cell count (Appendix A: Supplemental Table 1). The *invader* concentration was typically less than 0.1% of the *resident* cells at the beginning of the first growth cycle.

Communities were grown in 18 mm x 150 mm glass tubes (Fisher Scientific) containing 4.5 mL of defined medium (10 g Starch, 0.4 g Proline, 0.4 g Asparagine, 2 g/l $KNO_3$, 2 g/l NaCl, 2 g/l $K_2HPO_4$, 0.05 g/l $MgSO_4$, 0.01 g/l $FeSO_4$, 25 g/l Agar; adjust pH to 7.0 with 5N NaOH). Inocula were injected into molten agar (~50°C), vortexed briefly to mix, and rolled horizontally (along their long-axis) at 1,800 r.p.m. for 45 seconds under high air flow to coat the inside of the tube with a thin (~0.8 mm) layer of solid agar. The inside of the tube was therefore hollow, which allowed oxygen to reach the cells. Tubes were stored upright at 28°C for 12 days before harvesting.

Tubes were harvested by first adding 4 ml of sterile 2 mm glass beads (Chemglass Life Sciences) and vortexing for 10 seconds to remove the hydrophobic spores from the agar surface. A 5 ml filter-sterilized solution of 0.1% Tween-80 and 20% glycerol was added to each tube and vortexed for 10 seconds. An aliquot of 1.7 ml of each community was collected and frozen at -80°C. Communities were frozen between growth cycles to ensure that they could be grown from a consistent state in future uses. In subsequent growths, 100 µl of each previously harvested community was used to inoculate its next propagation cycle (1/50th dilution per cycle). This process was repeated until the communities had been grown for three propagation cycles.

**DNA extraction and sequencing**

To efficiently extract DNA from harvested biomass, it was necessary to first germinate the spores. A 100 µl aliquot of each community was grown in a test tube with 2 ml of the defined medium without agar. The liquid cultures were incubated while shaking for 40 hours at 28°C. Next, the cultures were centrifuged at 1,000 relative centrifugal force (r.c.f.) for 10 minutes to pellet the cells. A 1.7 ml volume of supernatant was removed, the remaining volume was vortexed, and 200 µl of the concentrated mycelium was transferred to a 0.2 ml thin-wall tube (Corning). These tubes were sonicated at 100% amplitude for 60 seconds using a Model 505 Sonicator with Cup Horn (QSonica). After sonication, the samples were centrifuged, and the supernatant containing DNA was used as template for PCR amplification.

Primers were designed to optimally differentiate an 80 nucleotide region of the *rpoB* sequence of all 18 strains using DesignSignatures (E. S. Wright and Vetsigian 2016) (Appendix A: Supplemental Table 2). The targeted *rpoB* region differed by a minimum of 4 nucleotides between all strain pairs (median of 12 nucleotides different). Extracted DNA was first amplified with primers targeting sites that were universal to all species, diluted, and then re-amplified with barcoded primers (Appendix A: Supplemental Table 2). This two-step process can decrease amplification bias during PCR (Berry et al. 2011) and mitigate the amplification of PCR artifacts associated with long adapter primers. The two PCR reactions consisted of a 2 min denaturation step at 95°C, followed by 45 and 25 cycles, respectively, of 20 sec at 98°C, 15 sec at 67°C, and 15 sec at 80°C. Each PCR reaction was followed by a melt curve from 60°C to 95°C in 0.5°C increments every 10 sec to confirm the expected melt peak. The PCR reaction contained 2.5 µl of iQ Supermix (Bio-Rad), 0.4 µl of 5 µM forward and reverse primer, 0.5 µl of DNA template, and 1.6 µl reagent grade $H_2O$ per sample. The 177 base pair product of the first PCR reaction was diluted 1,000-fold for use as template in the second reaction with barcoded primers. Barcoded primers were staggered by inserting 0-3 additional nucleotides before the sequencing read to help with randomization for phasing (Wu et al. 2015). Groups of 24 PCR products (2.5 µl/sample) were pooled into 50 µL of 10 x TBE on ice. This mix of PCR products was separated by length in a 1% agarose gel. The band matching the desired length (~310 nucleotides) was excised from the gel, and purified with the Wizard SV-Gel and PCR Cleanup System (Promega). All samples were sequenced by the UW-Madison Biotechnology Center with an Illumina Hi-Seq in rapid mode.

### Determination of presence or absence of the invader

Barcoded primers contained Illumina adapters with i5 and i7 index sequences that were unique to each community. These index sequences allowed for de-multiplexing of the samples by exactly matching the pair of eight nucleotide index sequences to the unique combination belonging to each community. By using 25 different i5 and i7 index sequences, we were able to multiplex up to 625 samples in the same sequencing lane. This approach typically resulted in more than 10,000 reads of 101 nucleotides per community. The reads were exactly matched to the known *rpoB* sequences for the panel of 18 strains in order to count the relative abundance of each strain. Analysis of read counts was performed with the R (R Core Team 2016) package Biostrings (Pages et al.).

Based on the known species that could be present in each community, we identified a cross-indexing error during sequencing in which a wrong i7 index was associated with a read at a rate of approximately 0.1%. The error rate varied between i7 indices in proportion to the total number of reads with a given i7 index. This simple statistical model explained well the unexpected reads and enabled background subtraction. The background level of reads for each species and community was determined by summing the contributions from other index pairs sharing the same i5 index, and resulted in effective background levels of less than 1% per species in a community. This background model was further confirmed by the distribution of reads belonging to a strain that was not present in any community and had been amplified separately with a unique i5 index and i7 index pair using primers that were independently synthesized.

Strains were said to invade if they had increased in frequency and reached at least 1% of the total community after subtracting the background. Strains not appearing above the background level were considered below the detection limit and marked as non-invasions. In two of the communities the invaders appeared only slightly above the background read level and were marked as defective and not considered in the analyses. Other cases of defective communities were due to experimental failure.

We further assessed a subset of eight bistable strain pairs using quantitative PCR (Appendix A: Supplemental Table 3), which has a larger dynamic range than high-throughput sequencing. Primers were designed that were specific to each strain using the R (R Core Team 2016) package DECIPHER (E. S. Wright et al. 2014) (Appendix A: Supplemental Table 2). The resident and invader were targeted for amplification in two separate PCR reactions for each community, corresponding to 32 different reactions in total. The reaction conditions were identical to those described above, except that we used tenfold larger reactions (50 µl total). In 45 PCR cycles, we observed early amplification of all 16 residents (Appendix A: Supplemental Table 3), followed by delayed amplification with primers targeting the invader. Using melt curves, gel runs, and Sanger sequencing, we were able to confirm that all invader amplifications were standard PCR artifacts attributable to the high number of PCR cycles and large reaction size. Therefore, in all eight bistable pairs, we were unable to detect the presence of either invader after three growth cycles.

## Inhibition experiments

To quantify inhibition we used a standard cross-streaking test on a Petri dish (Appendix A: Supplemental Fig. 6). First, all strains were grown at high density and allowed to sporulate on defined medium (see above). We used a flat rectangular (8 mm x 81 mm) aluminum pinning tool to transfer spores from a *majority* strain to the center of a new plate with a thin layer of defined medium (5 ml per 88 mm diameter Petri dish). A cross-streak of a *minority* strain was plated perpendicular to the *majority* strain using a 60 mm microscope coverslip with a thickness of ~0.15 mm (VWR). Five *minority* strains, separated by 1 cm, were plated in parallel across the same *majority* strain on a single plate. The strains were allowed to grow for 12 days at 28°C, and imaged periodically with a flatbed scanner. We used an in-house R (R Core Team 2016) script to quantify the distance between the *majority* strain and where the *minority* strain had sporulated in the obtained images. The *majority* strain was defined as inhibiting the *minority* strain if it prevented sporulation within a distance of 1 mm or greater.

## Measurement of growth rate and yield

To measure growth rate, strains were grown on a thin layer of defined medium (see above) for 43 hours and imaged under a microscope. The surface area of three to seven separated colonies was determined with ImageJ (Abràmoff, Magalhães, and Ram 2004), averaged, and scaled to units of mm$^2$ for plotting in Appendix A: Supplemental Fig. 5. To measure yield, each strain was grown alone for three sequential growth cycles as in the invasion experiments described above. The final concentration in CFU/μl at the end of all three growth cycles was determined by standard tenfold dilution plating and counting.

**Hubs of bistability**

Bistability occurs when strain *i* does not invade *j* and *j* does not invade *i*, corresponding to non-invasions on opposite sides of the diagonal in the invasion matrix. The strains can be sorted by their number of bistable interactions, ranging from 0 (strain #1) to 14 (strain #2). We compared the number of bistable pairings per strain to that of a random invasion network with the same total number of each type of pairwise link. To determine the statistical significance of the observed hubs, we calculated the fraction of random networks for which the sum of bistabilities associated with the top three strains was greater than or equal to that of the three most bistable strains in the measured invasion network.

**Analysis of triplet motifs in the invasion network**

There are 16 distinct triplet motifs possible when allowing for bistability (0 invasions), hierarchy (1 invasion) and coexistence (2 invasions). To assess the statistical significance of the different triplet motifs in the invasion network, we compared their frequency with those expected for random networks with the same number of pairwise links (0, 1, or 2 invasions) (Appendix A: Supplemental Fig. 3).

**Determination of the invasion network hierarchy**

A simple scoring model was constructed to assess a given ranking of strains based on the invasion matrix, as described in the Results. This model was based on a previous study of assigning hierarchy to directed networks (Gupte et al. 2011). The optimal ranking of strains was determined using the R (R Core Team 2016) package rgenoud (Mebane and Sekhon 2011) for integer genetic optimization. We repeated the optimization procedure 1,000 times from different initial rankings, and the highest scoring network was found in 70% of cases.

To determine the hierarchy without the effects of inhibition, we excluded pairs having inhibition in either direction when calculating the optimality score and then repeated the optimization procedure described above. In this case, as there were multiple networks with the same score, we averaged the hierarchy levels across all unique rankings (Appendix A: Supplemental Table 1).

**Calculation of correlations between inhibition and invasion**

To test whether residents that inhibit invaders are less likely to be invaded, we constructed a 2 x 2 contingency table by noting for each pair of species, A and B, whether A inhibits B or not, and whether B invades A or not. We computed the ratio of invasions to non-invasions for the cases with and without inhibition. We then took the ratio of the two ratios as a measure of relative enrichment for invasion in cases with inhibition. To calculate the statistical significance of the association between inhibition and invasion, we compared the observed enrichment to that expected for random inhibition networks. During randomization of the inhibition matrix, the number of each type of pairwise link (0, 1, or 2 inhibitions) was kept constant. Similarly, to test whether an invader inhibiting a resident is more likely to invade, we constructed a 2 x 2 contingency table by noting for each pair of species, A and B, whether A inhibits B or not, and whether A invades B or not.

**Controlling for downward pointing tendency of inhibitions**

To determine whether inhibitions play a role in invasions independently of (or in addition to) the downward bias, we repeated the above analysis focusing only on pairs for which $3 \geq h_A - h_B > 0$, where $h_A$and $h_B$ are the hierarchical levels of strains A and B in the invasion network. This controls for the fact that the ratio of pairs with $h_A \leq h_B$ and $h_A > h_B$ is very different for cases with and without inhibition, which might have led to spurious correlations between inhibition and invasion (that is, Simpson's paradox). Inhibitory interactions were randomly permuted only within the pairs considered, while preserving the number of inhibitions pointing up or down the hierarchy. Unfortunately, this approach could not be used to confirm with high statistical confidence that inhibition helps to resist invasion independently of the downward bias, because there are very few invasions or inhibitions against the hierarchy (Appendix A: Supplemental Fig. 7b). Nevertheless, invasions against the hierarchy were less frequent when the invaders were inhibited (p=0.07).

**Analysis of inhibition's role in bistability**

Bistability was observed in 51% of pairs having an inhibitory interaction and in 34% of pairs without an inhibitory interaction. To test whether bistable pairings were enriched in cases with inhibition, we randomized the inhibition matrix while maintaining the number of each type of pairwise link (0, 1, or 2 inhibitions). In 2.8% of random inhibition networks the enrichment for bistability in pairs with inhibition relative to pairs without inhibition was more than the observed value, corresponding to the p-value reported in the Results.

**Proving requirement of pairwise bistability for coexistence**

We will show that coexistence through the interplay between antibiotic production and degradation requires at least one bistable pair in the "Mixed Inhibition-Zone Model" introduced in Kelsic *et al.* (2015) for the case where antibiotic producers do not derive immediate benefit from inhibiting neighbors.

Coexistence requires the fastest growing species to be inhibited by another community member. If this were not the case, it would have the highest fitness for any combination of species abundances, and therefore will unconditionally outcompete all other species. Let species 1 be the species with the highest growth rate and species 2 be the species that inhibits it most strongly. We will show that if species 1 and 2 are a part of a coexisting community, then they are in a bistable relationship.

To calculate the invasibility relationships between species 1 and 2, we set the abundances of all other species to zero and obtain the following equations for the dynamics (in accordance with the notation used in Kelsic *et al.* (2015)):

$$f_1 = g_1 e^{-X_2 K_{P2}}$$

$$f_2 = g_2 e^{-X_1 K_{P1}}$$

$$X_i(t+1) = \frac{X_i(t) f_i(t)}{\sum_j X_j(t) f_j(t)},$$

where $\{X_i\}$ are the relative species abundances, $\{g_i\}$ are the growth rates, $\{f_i\}$ are the fitness values for given species abundances, $K_{P1} \geq 0$ and $K_{P2} > 0$ are the areas of inhibition caused by species 1 and 2.

To determine if species $i$ can invade species $j$ we set $X_i \to 0$ and $X_j \to 1$. The conditions for bistability (mutual non-invasion) of 1 and 2 are therefore:

$g_2 e^{-K_{P1}} < g_1$ and $g_1 e^{-K_{P2}} < g_2$.

The first condition is satisfied by construction because $g_2 < g_1$. Therefore species 1 and 2 are not bistable iff $g_1 e^{-K_{P2}} > g_2$.

The minimum fitness of species 1 over all possible abundances $\{X_i\}$ of the coexisting species is $\min f_1 = g_1 e^{-K_{P2}}$ because by construction species 2 is the species that inhibits species 1 the strongest. At the same time, the maximum fitness of species 2 over all possible abundances $\{X_i\}$ of the coexisting species is $\max f_2 = g_2$. Therefore, lack of bistability between species 1 and

2 implies $\min f_1 > \max f_2$, which means that species 1 is unconditionally outcompeting species 2 in contradiction to our assumption that the species are part of a coexisting community.

Therefore, every coexisting community has at least one bistable pair. The proof does not depend on the exact functional form of antibiotic inhibition.

## ACKNOWLEDGEMENTS

We are grateful to John Barkei and Amelia Remiarz for assisting with experiments related to this project. We thank Sri Ram, Hannah Haight, Sailendharan Sudakaran, Michael Sullivan, and Ye Xu for numerous helpful discussions. We also thank George Petry for helping to design and manufacture the device used for coating the insides of tubes with agar. We acknowledge support from the Simons Foundation, Targeted Grant in the Mathematical Modeling of Living Systems Award 342039, the National Science Foundation Grant DEB 1457518, and the National Institute of Food and Agriculture, US Department of Agriculture, Hatch project 1006261.

## AUTHOR CONTRIBUTIONS

Both authors contributed extensively to this work.

# CHAPTER 3: QUALITY FILTERING OF ILLUMINA INDEX READS MITIGATES SAMPLE CROSS-TALK

## BACKGROUND

In recent years Illumina sequencing has emerged as a mainstay for numerous biological applications. The Illumina platform uses sequencing by synthesis, whereby DNA containing adapters anneals to a flow cell and forms sequence clusters through bridge amplification before being sequenced. Due to the immense number of sequences that can be obtained, it is often useful to sequence DNA from multiple samples in a single run. This multiplexing process relies upon unique "index" sequences, termed i5 and i7, that are added to both sides of the DNA being sequenced. With only a few unique i5 and i7 sequences, hundreds of different i5 and i7 combinations can be created, enabling many samples to be simultaneously sequenced. De-multiplexing the samples after sequencing only requires finding the sequencing reads associated with each index pair that was added to the sequencing run.

As with other sequencing approaches, the Illumina method has been characterized for the frequency and type of errors that are generated (Cox, Peterson, and Biggs 2010). Substitutions, where one base is misread as another, are the most frequent class of error and occur more often toward the end of the sequence (Dohm et al. 2008; Schirmer et al. 2015). Insertions, deletions and motif-specific errors occur less frequently, but can still cause problems for certain applications (Nakamura et al. 2011; McMurdie et al. 2016).

Another type of error involves cross-talk among multiplexed samples and has received far less attention despite recent reports that error rates can be significant (E. Wright and Vetsigian 2016; Kircher, Sawyer, and Meyer 2011; Nelson et al. 2014). Such errors are particularly insidious

This chapter was originally published in:
Wright, E. S. & Vetsigian, K. H. (2016). Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics, 1–7. http://doi.org/10.1186/s12864-016-3217-x.

in applications that require the detection of variants that are rare in one sample but abundant in others, which includes biosphere surveys, investigations of ancient DNA, and the identification of cancerous cells (Nelson et al. 2014). Cross-talk errors can also be problematic if a large number of samples are multiplexed, such that each sample is a small fraction of the total number of reads. Since cross-talk can come from multiple sources, it has sometimes been attributed to experimental mistakes, cross-contamination during primer synthesis, multiple misread bases within index sequences, or sample carryover from previous sequencing runs on the same machine (D'Amore et al. 2016).

In view of the increasing importance of multiplexing on the Illumina platform, we systematically investigated cross-talk errors in order to rule out certain causes and determine whether there are any satisfactory solutions to the problem. To this end, we constructed 14 unique combinations of i5 indices, i7 indices, and read sequences (Fig. 3.1a), while carefully controlling for potential sources of cross-contamination such as primer synthesis. The sequences of all reads were well-separated in sequence space to minimize cross-talk due to misread bases. Surprisingly, we observed that cross-talk was due to three different types of misassignments (Fig 3.1b) that occurred at similar rates. Furthermore, we found that quality filtering of the index pairs was sufficient to all but eliminate misassignments between samples without sacrificing a substantial fraction of reads.

## RESULTS

Using standard de-multiplexing protocols, we observed a 0.09% rate of *sequence misassignments*, which have the correct i5 and i7 index pair but incorrect sequence, and a 0.16% rate of *index misassignments*, which have a correct sequence read but a single incorrect i5 or i7 index. These rates are consistent with prior studies that found misassignment rates between 0.06% and 0.21%

(Nelson et al. 2014; D'Amore et al. 2016). Furthermore, the rate of sequence misassignment was similar to that of i5 or i7 index misassignment (Fig. 3.1b), indicating that the sequence is being misassigned rather than both index sequences being independently misassigned. Both sequence and index misassignments will contribute to cross-talk between samples when each sample is separated from other samples by a single index, whereas only sequence misassignments are relevant when unique dual-indexing is used. Nevertheless, the existence of sequence misassignments indicates that even the use of two unique index sequences is insufficient to eliminate cross-talk.

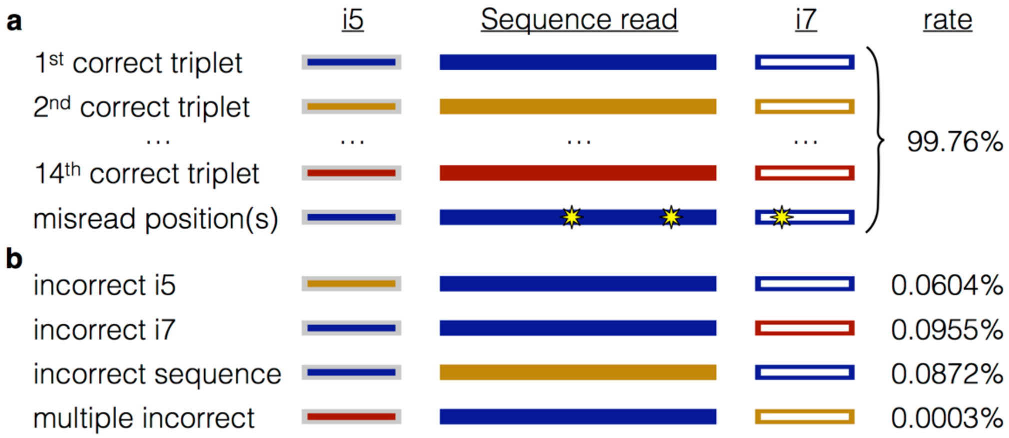
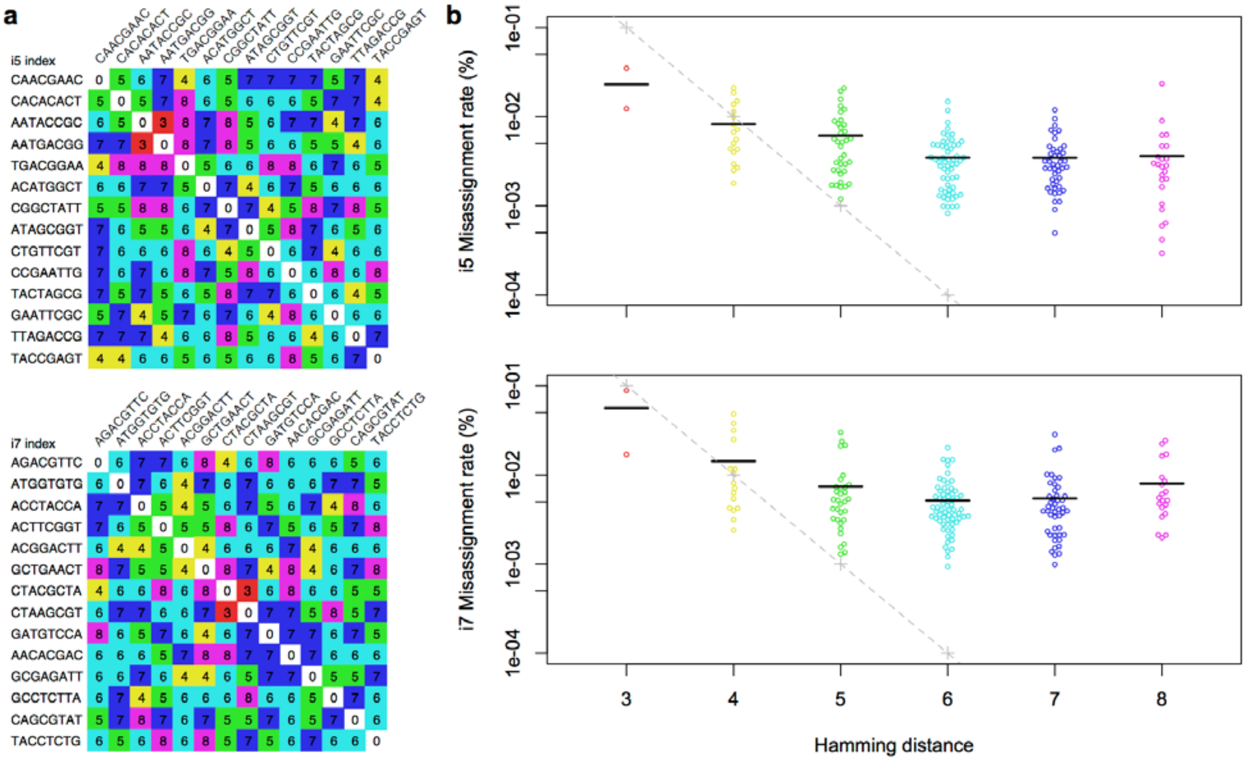


**Figure 3.1. Rates of different misassignment errors on the Illumina platform. a**, Unique index and read sequences that were well separated in sequence space (colored rectangles) were used to form 14 distinct samples and multiplexed in the same Illumina sequencing run. Misread bases (yellow stars) make up the most common error type, but are still attributable to their correct triplet. **b**, Misassigned reads appear as unexpected triplets, and can be categorized as either index misassignments (0.16% total) or sequence misassignments (0.09%).

Misassignments can in principle result from multiple misread bases within an index sequence. However, even at a high average error rate of 1% (Q20), the chance of at least three positions being misread is $10^{-6}$ assuming that errors are independent. The observed rate of index misassignment was far greater than expected, regardless of the number of differences between index sequences (Fig. 3.2). If two unique index sequences are used, the probability of both the i5 and i7 being misread as another index pair is expected to be around $10^{-12}$. Therefore, since we obtained approximately 10 million reads per sample, we would expect zero sequence misassignment due to misread bases. To further verify these assumptions, we de-multiplexed another index pair where neither the i5 or i7 index was included in the experiment. There were no reads attributed to this index pair, confirming that the per-base error rate of Illumina sequencing does not explain the observed rate of cross-talk.



| i5 index | CAACGAAC | CACACACT | AATACCGC | AATGACGG | TGACGGAA | ACATGGCT | CGGCTATT | ATAGCGGT | CTGTTCGT | CCGAATTG | TACTAGCG | GAATTCGC | TTAGACCG | TACCGAGT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CAACGAAC | 0 | 5 | 6 | 7 | 4 | 6 | 5 | 7 | 7 | 7 | 7 | 5 | 7 | 4 |
| CACACACT | 5 | 0 | 5 | 7 | 8 | 6 | 5 | 6 | 6 | 6 | 5 | 7 | 7 | 4 |
| AATACCGC | 6 | 5 | 0 | 3 | 8 | 7 | 8 | 5 | 6 | 7 | 7 | 4 | 7 | 6 |
| AATGACGG | 7 | 7 | 3 | 0 | 8 | 7 | 8 | 5 | 6 | 6 | 5 | 5 | 4 | 6 |
| TGACGGAA | 4 | 8 | 8 | 8 | 0 | 5 | 6 | 6 | 8 | 8 | 6 | 7 | 6 | 5 |
| ACATGGCT | 6 | 6 | 7 | 7 | 5 | 0 | 7 | 4 | 6 | 7 | 5 | 6 | 6 | 6 |
| CGGCTATT | 5 | 5 | 8 | 8 | 6 | 7 | 0 | 7 | 4 | 5 | 8 | 7 | 8 | 5 |
| ATAGCGGT | 7 | 6 | 5 | 5 | 6 | 4 | 7 | 0 | 5 | 8 | 7 | 6 | 5 | 6 |
| CTGTTCGT | 7 | 6 | 6 | 6 | 8 | 6 | 4 | 5 | 0 | 6 | 7 | 4 | 6 | 6 |
| CCGAATTG | 7 | 6 | 7 | 6 | 8 | 7 | 5 | 8 | 6 | 0 | 6 | 8 | 6 | 8 |
| TACTAGCG | 7 | 5 | 7 | 5 | 6 | 5 | 8 | 7 | 7 | 6 | 0 | 6 | 4 | 5 |
| GAATTCGC | 5 | 7 | 4 | 5 | 7 | 6 | 7 | 6 | 4 | 8 | 6 | 0 | 6 | 6 |
| TTAGACCG | 7 | 7 | 7 | 4 | 6 | 6 | 8 | 5 | 6 | 6 | 4 | 6 | 0 | 7 |
| TACCGAGT | 4 | 4 | 6 | 6 | 5 | 6 | 5 | 6 | 6 | 8 | 5 | 6 | 7 | 0 |

| i7 index | AGACGTTC | ATGGTGTG | ACCTACCA | ACTTCGGT | ACGGACTT | GCTGAACT | CTACGCTA | CTAAGCGT | GATGTCCA | AACACGAC | GCGAGATT | GCCTCTTA | CAGCGTAT | TACCTCTG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AGACGTTC | 0 | 6 | 7 | 7 | 6 | 8 | 4 | 6 | 8 | 6 | 6 | 6 | 5 | 6 |
| ATGGTGTG | 6 | 0 | 7 | 6 | 4 | 7 | 6 | 7 | 6 | 6 | 6 | 7 | 7 | 5 |
| ACCTACCA | 7 | 7 | 0 | 5 | 4 | 5 | 6 | 7 | 5 | 6 | 7 | 4 | 8 | 6 |
| ACTTCGGT | 7 | 6 | 5 | 0 | 5 | 5 | 8 | 6 | 7 | 5 | 6 | 5 | 7 | 8 |
| ACGGACTT | 6 | 4 | 4 | 5 | 0 | 4 | 6 | 6 | 6 | 7 | 4 | 6 | 6 | 6 |
| GCTGAACT | 8 | 7 | 5 | 5 | 4 | 0 | 8 | 7 | 4 | 8 | 4 | 6 | 7 | 8 |
| CTACGCTA | 4 | 6 | 6 | 8 | 6 | 8 | 0 | 3 | 6 | 8 | 6 | 6 | 5 | 5 |
| CTAAGCGT | 6 | 7 | 7 | 6 | 6 | 7 | 3 | 0 | 7 | 7 | 5 | 8 | 5 | 7 |
| GATGTCCA | 8 | 6 | 5 | 7 | 6 | 4 | 6 | 7 | 0 | 7 | 7 | 6 | 7 | 5 |
| AACACGAC | 6 | 6 | 6 | 5 | 7 | 8 | 8 | 7 | 7 | 0 | 7 | 6 | 6 | 6 |
| GCGAGATT | 6 | 6 | 7 | 6 | 4 | 4 | 6 | 5 | 7 | 7 | 0 | 5 | 5 | 7 |
| GCCTCTTA | 6 | 7 | 4 | 5 | 6 | 6 | 6 | 8 | 6 | 6 | 5 | 0 | 7 | 6 |
| CAGCGTAT | 5 | 7 | 8 | 7 | 6 | 7 | 5 | 5 | 7 | 6 | 5 | 7 | 0 | 6 |
| TACCTCTG | 6 | 5 | 6 | 8 | 6 | 8 | 5 | 7 | 5 | 6 | 7 | 6 | 6 | 0 |

**Figure 3.2. Misassignment rates were weakly correlated with the hamming distance between index sequences. a**, Matrices showing the hamming distance between i5 and i7 index sequences used in this study. **b**, The rate of triplets with an incorrect i5 (or i7) index as a function of the hamming distance to the correct i5 (or i7) index. Horizontal lines indicate the mean misassignment rate at each hamming distance. Note the log-scale y-axis. The theoretical misassignment rates based on independent substitutions are shown in gray for an exaggerated 10% substitution rate (Q10); lower substitution rates would simply shift the dashed-line to the left. The observed misassignment rate does not decrease exponentially as would be expected if misread errors are independent, indicating that misread bases are not the cause of misassignment errors.

Having ruled out misread bases as the cause of most misassignments, we next investigated whether incorrect reads were associated with low quality scores. Figure 3.3 shows that correct triplets (i5, i7, and sequence) tended to have high quality in both index read steps, whereas index misassignments tended to be low quality in the step for which they were misassigned. The average quality scores of i5 and i7 index reads appear to be largely independent, i.e. low quality in one does not imply low quality in the other. This may be due to the fact that the two index sequences are read separately after the cluster is inverted on the flow cell. In contrast, sequence misassignments tended to have poor quality i5 and i7 index reads, in addition to a low quality sequence read (Fig. 3.3). Moreover, quality scores were generally lower across the entire length of misassigned reads, rather than only being low quality in a specific region (Appendix B: Supplemental Fig. 1).

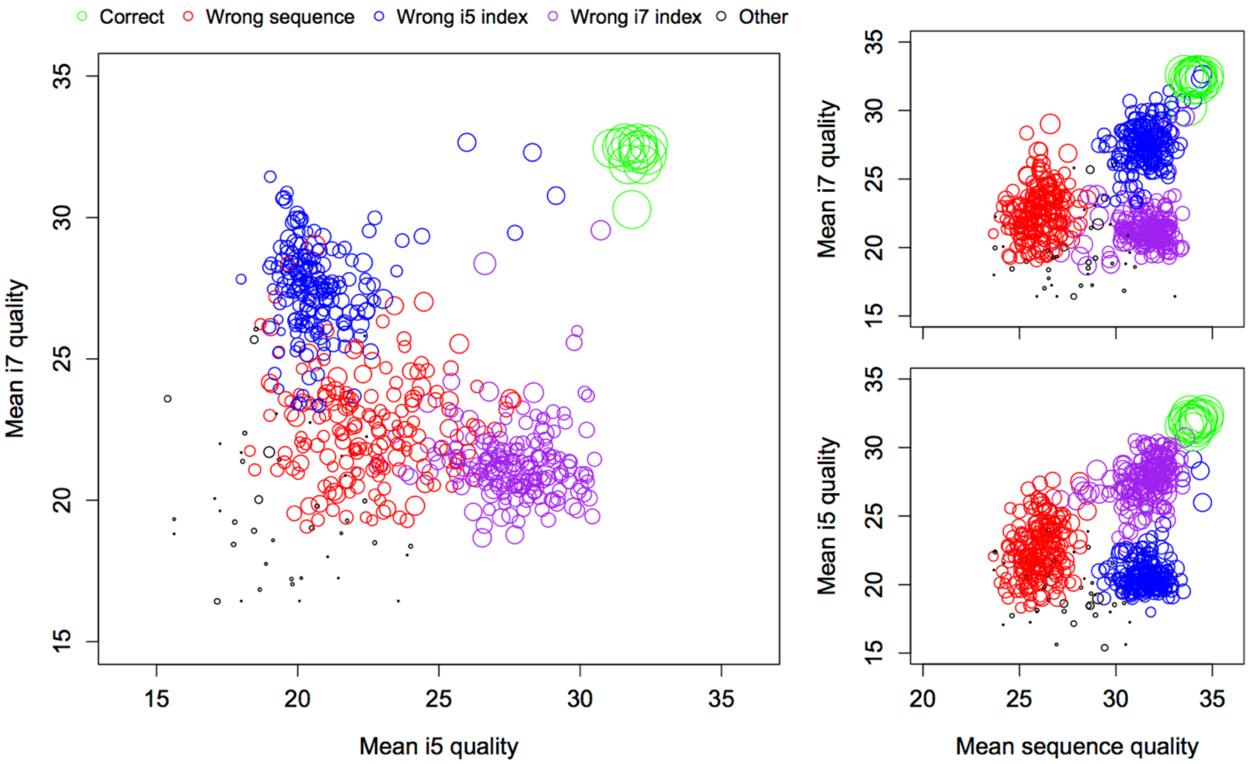


**Figure 3.3. Breakdown of average quality scores by error type.** Each point represents the reads obtain for one triplet (i5, i7, and sequence read), and is scaled to the log of the read count. Correct triplets (green) have high quality across all read steps, whereas sequence misassignments have low quality in all three read steps. In contrast, index misassignments tend to have low quality in the step for which they are misassigned.

The observed quality score pattern has several implications for filtering incorrect reads. First, filtering low quality sequence reads is expected to be insufficient to eliminate anything other than sequence misassignments. This implies that the i5 and i7 must be quality filtered to eliminate index misassignments. Second, filtering low quality i5 and i7 index reads may be sufficient to eliminate both sequence misassignments and index misassignments without needing to quality filter the sequence read. We tested these hypotheses by applying increasing stringencies of quality score filtering and observing the remaining cross-talk. Here we distinguished between three strategies: quality filtering only the sequences, only the index pairs, and filtering all read steps. As expected, keeping only high quality sequence reads nearly eliminated sequence misassignments but not index misassignments (Fig. 3.4), whereas filtering the index sequences largely prevented all types of misassignment. By filtering the index reads to an average quality score of ≥ 26 (0.25% probability of error per base) it was possible to reduce the overall rate of incorrect triplets from 0.24% to 0.03% while maintaining 88% of total reads. A combined strategy was only slightly more effective at eliminating both types of misassignment. Thus, quality filtering of index reads provides a simple way to minimize cross-talk while preserving the vast majority of reads.

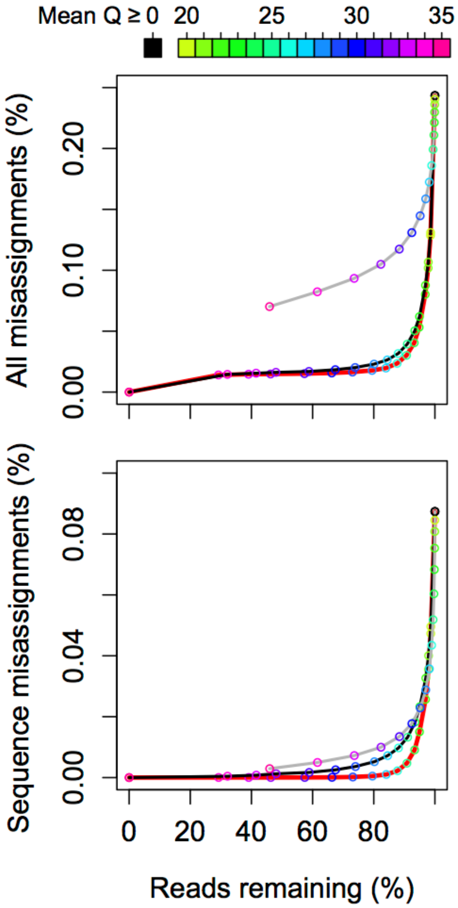


**Figure 3.4. Trade-off between removing misassigned and preserving correct reads during quality filtering.** (Top) Misassignments were not efficiently removed by quality filtering the sequence reads (gray line), whereas quality filtering the i5 and i7 index sequences was highly effective (black line). Quality filtering sequence reads in addition to index reads (red line) did not remove substantially more cross-talk. (Bottom) Quality filtering either sequence reads or index reads was effective at removing sequence misassignments.

## DISCUSSION

Misassignment errors could result from distinct cluster originators forming at an overlapping spot on the flow cell (Nelson et al. 2014). If this were the case, we might expect the quality score profiles of incorrect reads to oscillate between low quality in positions where the two sequence clusters differ (e.g., one A, one C) and high quality where they are identical (e.g., both A). However, we did not observe any such pattern in the quality score signals of incorrect triplets, perhaps because there is a poor correlation between the quality score and the actual probability of error (Schirmer et al. 2015) or because neighboring positions are taken into account when assigning quality scores. Nevertheless, we would expect overlapping clusters to lower the quality of all read steps due to competing signals, yet this was also not observed. Instead it appears that one cluster tends to overpower the other during each read step (i5, i7, or sequence), and the overpowering cluster in the pair can switch between read steps.

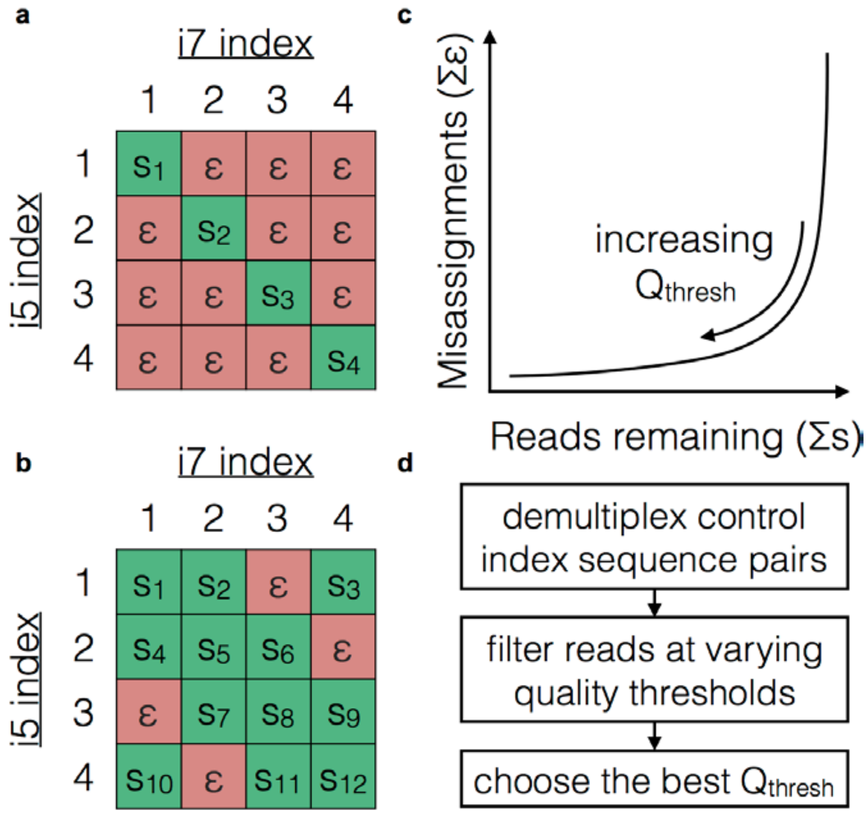


**Figure 3.5. Recommended procedure for removing background reads. a**, When using unique dual index sequences for every sample ($s_i$), each missing index pair offers a negative control that provides an estimate of the number of misassigned reads (ε). **b**, When almost all index combinations are being used, controls can be added by purposefully omitting samples for some combinations of index sequences. **c**, The quality score threshold ($Q_{thresh}$) can then be optimized by plotting the sum of misassignments versus the number of reads remaining. **d**, A value of $Q_{thresh}$ can be selected that minimizes misassignments while maximizing the number of reads that remain.

While a quality score threshold of 26 was sufficient to eliminate most misassignments in this study, this threshold may vary from run-to-run depending on the run's overall quality and other factors. For this reason, it may be useful to detect misassignments and then vary the quality score threshold to observe its effect on their removal (Fig. 3.5). Misassignments can be detected by de-multiplexing index combinations that should not be present in the sequencing run but for which the i5 and i7 index sequences exist separately in other samples. In the absence of misassignments the number of sequences attributable to missing index pairs should be zero. This provides a straightforward method for both verifying misassignments and confirming their removal. Also, this method does not depend upon knowing the sequence variants that belong to each sample.

## CONCLUSIONS

To our knowledge, this is the first systematic study of cross-talk on the Illumina platform that uses standard dual indexing as opposed to custom or single indexing schemes. Previous studies of cross-talk identified the advantages of dual indexing over single indexing and of quality filtering index sequences (Kircher, Sawyer, and Meyer 2011; Nelson et al. 2014). Here we extended these findings by showing that there are three independent modes of cross-talk: incorrect i5 index, i7 index, and sequence. The existence of sequence misassignments prevents dual indexing from completely eliminating cross-talk without quality filtering. It also means that if only a single (i7) index is used, filtering on sequence quality in addition to index quality is the best strategy. In agreement with previous work (Kircher, Sawyer, and Meyer 2011), we determined that no amount of quality filtering can completely eliminate cross-talk when samples are only separated by one of two index sequences. Thus, unique dual indexing is required when identification of extremely rare variants is critical. We also proposed a simple method for both quantifying cross-talk and choosing run-specific or application-specific thresholds for mitigating it by counting reads assigned to unexpected index pairs during quality filtering (Fig. 3.5).

Cross-talk between samples effectively limits the number of index pair combinations that can be reliably used. As the fraction of clusters sharing an i5 or i7 increases, the number of misassigned reads will concomitantly increase. Eventually, even at small rates of misassignment the incorrect reads would rise to an intolerable level if enough index combinations were used. This is supported by a previous study in which the rate of cross-talk was estimated to approach 1% when 625 index pair combinations were used (E. Wright and Vetsigian 2016). For this reason, we believe it is necessary to quality filter index reads in addition to the sequencing read when employing a multiplexing strategy. Furthermore, to mitigate the issue of spurious results due to cross-talk in the literature, we recommend that repositories such as the Sequence Read Archive (SRA) (Leinonen et al. 2010) enable and encourage the submission of quality scores for index sequences and unexpected (control) index pairs. This would allow retroactive filtering of published sequences, and would also provide a means for automatic accumulation of data on the magnitude of sample cross-talk as sequencing platforms evolve.

## METHODS

### Template DNA extraction and PCR amplification

A total of 13 strains (Appendix B: Supplemental Table 1) belonging to the genera *Amycolatopsis* or *Streptomyces* were grown at 28°C in 1 mL of 1/10th concentration ISP2 medium (10 g Malt extract, 4 g Yeast extract, and 4 g Dextrose per 1 L) for 9 days. The remaining protocol closely paralleled that of a previous study (E. Wright and Vetsigian 2016). Briefly, the cultures were centrifuged at 1000 rcf for 10 minutes to pellet the cells. A 700 µL volume of supernatant was removed, the remaining volume was vortexed, and 200 µL of the concentrated mycelium was transferred to a 0.2 mL thin-wall tube (Corning). These tubes were sonicated at 100% amplitude for 60 seconds using a Model 505 Sonicator with Cup Horn (QSonica) while the samples were completely enclosed. After sonication, the samples were centrifuged, and the supernatant containing DNA was used as template for PCR amplification.

Extracted DNA was amplified using indexed primers containing adapters (Appendix B: Supplemental Table 2). Samples were carefully arranged into a 96 well plate in alternating rows and columns to prevent any possibility of cross-contamination. Primers were designed to target either a stably integrated chromosomal barcode or the RNA polymerase subunit β (*rpoB*) gene. The PCR reaction consisted of a 2 min denaturation step at 95°C, followed by 40 cycles of 20 sec at 98°C, 15 sec at 67°C, and 15 sec at 80°C. The PCR reaction contained 10 µL of iQ Supermix (Bio-Rad), 0.8 µL of 10 µM forward primer, 0.8 µL of 10 µM reverse primer, 4 µL of DNA template, and 5.9 µL of reagent grade $H_2O$ per sample. Primers were synthesized by Integrated DNA Technologies using their TruGrade service that is intended to prevent cross-contamination during synthesis. Furthermore, primers were purchased across multiple orders that were staggered in time to further ensure that primer cross-contamination could not occur.

**DNA purification, sequencing, and analysis**

PCR products were purified separately with the Wizard SV-Gel and PCR Cleanup System (Promega). Samples were sequenced by the UW-Madison Biotechnology Center on an Illumina Hi-Seq 2500 in rapid mode. Sample concentrations were determined using an Agilent 2100 Bioanalyzer, and pooled immediately prior to sequencing in order to reach a target density of 8.5e5 to 1e6 clusters per $mm^2$. Spiking PhiX was unnecessary because the sequences' first 5 bases were well separated (hamming distance from 2 to 5), and we have not noticed a reduction in cross-talk from adding PhiX in prior runs. Single-end sequencing was performed for 51 cycles. After sequencing the cluster density was determined to be $9.9e5/mm^2$.

Samples were de-multiplexed using Illumina's bcl2fastq (v2.17) software and its associated defaults, that is, allowing 1 mismatch per index and only outputting reads that "pass filter". Illumina's pass filter algorithm screens out reads based on the signal intensities over the first 25 cycles of the sequencing read. The additional parameter "--create-fastq-for-index-reads" was specified to force the program to output fastq files for both index sequences (i5 and i7). Raw index and sequence reads are available from the sequence read archive (SRA) under accession number SRP083789. We also de-multiplexed another randomly selected index pair (i5: ACGTAAGG; i7: GGCCAATT) that was not used with any sample. This index pair had zero associated reads, confirming that the observed rates of sequence misassignment are larger than expected from misread bases alone.

Reads were assigned to the nearest expected sequence within an edit distance of four (including mismatches, insertions, and deletions) using the DECIPHER (v2.1.6) package (E. S. Wright 2016) in R (R Core Team 2016) (http://DECIPHER.codes). The sequences belonging to each sample were separated by an edit distance of at least 14, meaning that a small number of misread bases would not prevent correct matching. Barring insertions and deletions, which are uncommon on the Illumina platform, the 14 sequence variants were separated by between 21 and 43 substitutions. The probability of 17 (21 differences – 4 mismatches) or more substitutions within 51 bases is $10^{-23}$ at a high misread rate of 1% (Q20). Between 4.8 million and 9.6 million reads were mapped to each of the 14 sequences having a known index pair, with a mean of 7.5 million reads per expected triplet. A total of 99.9% of unexpected triplets differed from an expected triplet by a single read step, with the remainder differing by two read steps (e.g., incorrect i5 and i7).

Quality score filtering was applied with the TrimDNA function of DECIPHER (E. S. Wright 2016), which allows specification of a maximum average error rate. The quality score (Q) can be converted to a probability of error (p) using the formula $p = 10^{(Q/-10)}$. The sequence misassignment rate was calculated as the fraction of reads having the same i5 and i7 index pair that mapped to the wrong sequence, divided by the total number of mapped reads having that index pair. The index misassignment rate was calculated as the fraction of reads that mapped to a sequence with a known index pair, but differing by a single i5 or i7 index from the expected index pair.

## ACKNOWLEDGEMENTS

The authors thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility for performing the Illumina sequencing associated with this study. This work was supported by the Simons Foundation, Targeted Grant in the Mathematical Modeling of Living Systems Award 342039, the National Science Foundation Grant DEB 1457518, and the National Institute of Food and Agriculture, US Department of Agriculture, Hatch project 1006261.

## AUTHOR CONTRIBUTIONS

EW performed the experiments. EW and KV designed the study, analyzed the data and wrote the manuscript. Both authors read and approved the final manuscript.

# CHAPTER 4: JACKPOTS SKEW THE DISTRIBUTION OF DESCENDANTS ARISING FROM INDIVIDUAL BACTERIA

## INTRODUCTION

Since the dawn of population genetics, it has been clear that the distribution in the number of offspring per parent is central to developing a quantitative understanding of the evolution of genetic variants (Gillespie 1974). The offspring distribution provides a mapping between generations and directly determines the extent that genetic drift affects allele frequencies in a population (Der, Epstein, and Plotkin 2012). In idealized populations the offspring distribution is typically assumed to be Poisson distributed with a variance that is inversely proportional to the effective population size (Charlesworth 2009). However, for some animals there is high variance in reproductive success, with a minority of males fathering a large fraction of the progeny in each generation (Araki et al. 2007; Lallias et al. 2010). Such highly-skewed offspring distributions have fundamental implications for how we predict and interpret fluctuations in allele frequencies (Der, Epstein, and Plotkin 2012; Hedrick 2005; Hoban et al. 2013).

In contrast to animals, the offspring distribution is a largely unexplored concept for microorganisms. Unlike many sexually reproducing organisms, the offspring distribution is trivial for bacteria undergoing binary fission because each replicating bacterium always gives rise to two individuals. However, this concept can be generalized to the non-trivial descendants distribution by asking: How many progenitor cells does a random bacteria yield after a given amount of time $\boldsymbol{\tau}$, where the time $\boldsymbol{\tau}$ need not refer to a single generation (Fig. 4.1a). This is a variable quantity which is described by some probability distribution. In a system with seasonality, for example, one might look at this distribution after one season. Defined as such, the descendants distribution is a fundamental quantity of which little is known for bacteria.

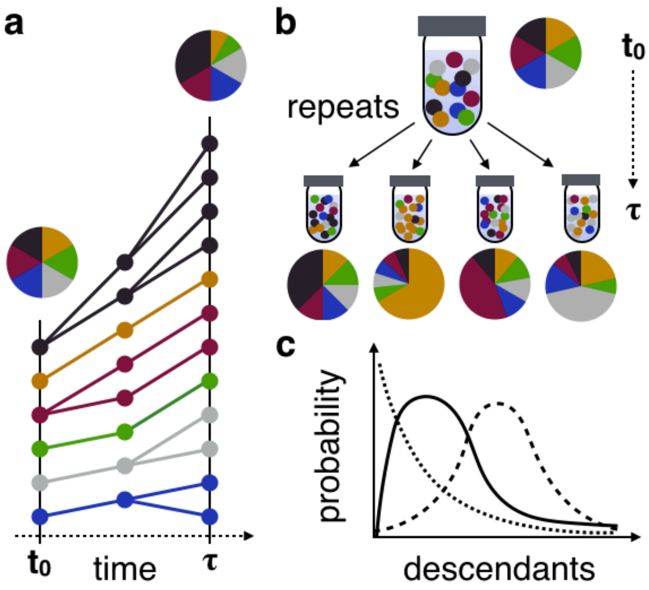


**Figure 4.1. Measurement of the distribution of descendants arising from a population. a**, Clonal cells, represented by colored circles, are grown for a period of time (**τ**) before their relative abundances are measured. **b**, The variability in the proportion of descendants between replicate populations of cells is used to determine the descendants distribution. **c**, The descendants distribution may take on a variety of shapes that have different rates of converging to zero. Heavy-tailed distributions would result in jackpots where individuals have much greater reproductive output than expected based on their initial frequency.

Due to bacteria's large population sizes (Lynch 2003), selection is generally considered to be the predominant evolutionary force determining fluctuations in allele frequencies (Price and Arkin 2015). However, some bacteria undergo dramatic fluctuations in population number that would strongly expose them to the effects of genetic drift (Batut et al. 2014). For example, entire populations can arise from a single cell during between-host bottlenecking (Kaltenpoth et al. 2009) or strong selective sweeps of a single lineage (Bendall et al. 2016). At the other extreme, the shape of the descendants distribution is largely unknown for a single genetically identical population grown in an unstructured environment. This leads to the question of whether it is feasible to directly measure the descendants distribution of microbial populations in a scalable fashion. Such a tool would allow exploration of the distribution in many different organisms and contexts, potentially allowing a determination of the extent of genetic drift and the mechanistic basis of observed stochasticity.

The descendants distribution is unknown for bacteria in part because it is challenging to measure for clonal individuals. Here we used a barcode tagging approach that enabled us to track descendants from hundreds of sub-populations differing only by a short DNA barcode inserted in their chromosome. Using the variability between replicate populations, we show that the descendants distribution is fat tailed, that is some bacteria represent a far greater proportion of the final population than their initial frequency. We propose hypotheses as to why some bacteria effectively "win the jackpot", and discuss implications of heavily skewed descendants distributions.

## RESULTS

### Measurement of the descendants distribution

Directly determining the descendants distribution would require tracking each individual cell and all of its offspring within a clonal population. Such a brute force strategy is exceedingly difficult, if not impossible. Therefore, we developed an alternative method to track sub-populations of cells and infer the shape of the descendants distribution based on changes in the relative abundance of sub-populations between replicates (Fig. 4.1b). This method required tagging bacterial lineages of an otherwise clonal population with a unique 30 base pair random sequence inserted at a fixed site on the chromosome. A similar technique has been used previously to tag yeast lineages (Levy et al. 2015). After growth in liquid medium, the relative abundance of each lineage is determined by PCR amplification of these lineage-specific "barcodes" followed by high-throughput sequencing (see Methods). The lower detection limit of a barcode is at least 10-fold lower than the least abundant cell in the initial population.

This method results in a frequency distribution of lineages at a time point, , that arose from sub-populations of bacteria at time point 0. Importantly, this approach requires that the technical variability due to the experimental procedure be far less than the biological variability. To investigate both of these components of variability, we compared the frequency distribution determined from technical (PCR) replicates to that originating from distinct biological replicates. We found that technical replicates had substantially higher correlation than biological replicates (Appendix C: Supplemental Fig. 1), confirming that most of the variability is biological in nature. This allows the shape of the descendants distribution to be inferred from the variability between biological replicates (Fig. 4.1c). To this end, we independently grew 5 different strains belonging to the genus *Streptomyces* in 8 separate replicate populations starting from 3 different initial concentrations (see Methods).

### The descendants distribution is skewed with a heavy tail

Two extremes of the barcode frequency distribution reveal characteristics of the descendants distribution (Fig. 4.2a). At one end, the distribution of barcodes present at high frequency is expected to be normally distributed because each barcode represents a large number of initial cells. Based on the central limit theorem, the means of large samples drawn from any distribution should be normally distributed, so long as the underlying distribution has a finite mean. We tested whether the relative frequencies of the 8 replicates belonging to the most abundant barcodes could be normally distributed using the Shapiro-Wilk test (threshold alpha = 0.02). Each of these barcodes is estimated to be shared by over 1,000 initial cells per replicate. For 4 out of 9 of these abundant barcodes the normal distribution was rejected (Fig. 4.2b). This result suggests that the underlying descendants distribution is heavy tailed, because convergence to normality would otherwise be expected for such large samples.

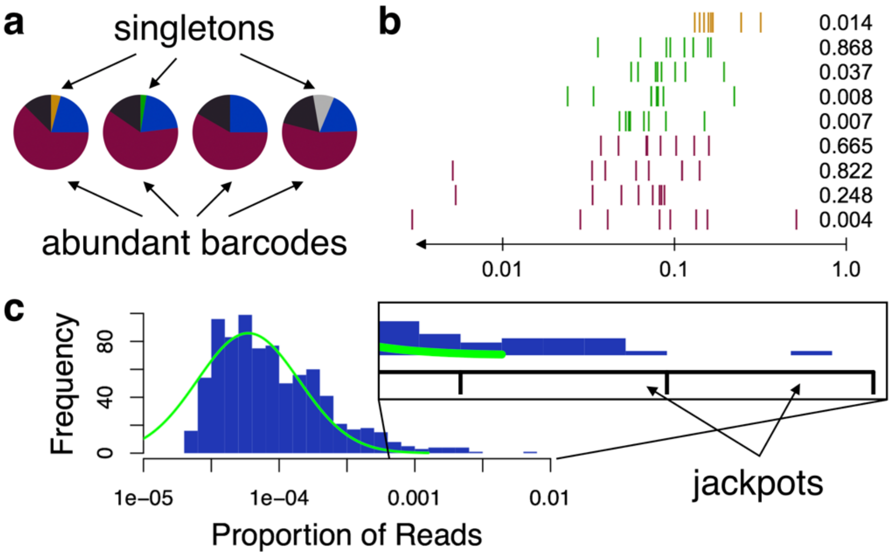


**Figure 4.2. Inferring the descendants distribution. a**, Barcodes at the two extreme of relative abundance reflect the shape of the descendants distribution. **b**, Abundant barcodes, those shared by more than 1000 cells in the initial population, are expected to converge to a normal distribution due to the central limit theorem. However, many of the most abundant barcodes were not normally distributed, based to their p-values (at right) in the Shapiro-Wilk test. Instead, the abundant barcodes originating from three different strains (colors) were widely scattered in terms of their final proportion of the population (x-axis). **c**, The singletons, those barcodes occurring in only 1 out of 8 replicates, approximate the shape of the descendants distribution since they likely started from single cells. For the strain with the most singletons, *Streptomyces G4A3*, we observed that their relative abundances at the end of the experiment were more heavy-tailed than a fitted log-normal distribution (green curve). The outlying "jackpots" represent cells that grew to a far higher abundance than average, in many case by more than 10-fold.

At the other extreme, as the initial frequency of a barcode approaches a single cell (Fig. 4.2a), the distribution of final barcode frequencies should approximate the descendants distribution. We made the assumption that strains appearing in only 1 out of 8 replicates of a given initial concentration were sufficiently rare to have originated from a single cell. While we would expect this assumption to be violated in about 10% of cases, the impact of starting from 2 cells should be on the order of 2-fold. The resulting distribution of barcode frequencies for these "singletons" appeared approximately log-normally distributed (Fig. 4.2c). Surprisingly, for many strains the distribution spanned over three orders of magnitude, meaning that some barcodes were over-represented by more than 1000-fold that of a typical barcode starting from an identical initial frequency.

### Sub-exponential and some heavy-tailed distributions can be rejected

The drawback of the two previous approaches for characterizing the underlying descendants distribution is that they only take into account the extremes of the dataset. Therefore, we devised an alternative procedure for using the entire dataset to more accurately estimate the descendants distribution. Here we simulated the entire process, starting from sampling an initial frequency distribution of barcodes, then "growing" each barcode according to an underlying distribution, and sampling the resulting frequency distribution for PCR amplification and sequencing. Since we know the number of initial templates per replicate based on quantitative PCR and the number of reads obtained in sequencing, the entire simulation only has two free parameters: (i) the shape of the distribution and (ii) the initial population size. The shape of each distribution comes from both the type of distribution and the parameter controlling its shape.

We used the average of 300 simulations to optimize both parameters for each distribution that we tested (see Methods). Briefly, the optimality criterion was based on the area between the cumulative distributions of the simulations and the observed data. For each shape value, we first optimized the initial (census) population size to yield the same number of unique barcodes that were observed in the real data. The optimal population sizes generally matched the approximate number of initial cells determined through plate counting (Appendix C: Supplemental Fig. 2), validating the simulation procedure. Based on the simulated data, distributions were rejected when their optimal shape value resulted in a distribution of barcodes that fell outside of the variability encompassed by the 300 replicate simulations (see Methods). The results of the simulations are summarized in Table 1.

**Table 4.1. Best-fit parameters for different distribution types at each initial concentration.**

| | | Initial strain concentration | | | | | |
|---|---|---|---|---|---|---|---|
| | | Full concentration | | One tenth | | One hundredth | |
| Distribution | Strain | p-value | shape | p-value | shape | p-value | shape |
| Exponential | *S. coelicolor* | 0.00 | 3.00 | 0.00 | 5.00 | 0.03 | 3.00 |
| | *S. albus J0174* | 0.00 | 5.00 | 0.00 | 5.00 | 0.00 | 5.00 |
| | *S. G4A3* | 0.00 | 7.00 | 0.00 | 6.00 | 0.00 | 2.00 |
| | *S. S26F9* | 0.00 | 2.00 | 0.00 | 2.00 | 0.00 | 3.00 |
| | *S. venezuelae* | 0.00 | 3.00 | 0.00 | 6.00 | - | - |
| Weibull | *S. coelicolor* | 0.00 | 0.15 | 0.14 | 0.35 | 0.05 | 0.45 |
| | *S. albus J0174* | 0.00 | 0.30 | 0.07 | 0.10 | 0.02 | 0.25 |
| | *S. G4A3* | 0.00 | 0.35 | 0.00 | 0.50 | 0.36 | 0.45 |
| | *S. S26F9* | 0.04 | 0.10 | 0.39 | 0.15 | 0.02 | 0.25 |
| | *S. venezuelae* | 0.15 | 0.30 | 0.48 | 0.15 | - | - |
| Lognormal | *S. coelicolor* | 0.39 | 3.10 | 0.46 | 1.75 | 0.23 | 1.40 |
| | *S. albus J0174* | 0.00 | 1.75 | 0.24 | 3.75 | 0.00 | 2.70 |
| | *S. G4A3* | 0.03 | 1.60 | 0.03 | 1.55 | 0.40 | 2.10 |
| | *S. S26F9* | 0.44 | 4.05 | 0.29 | 2.85 | 0.02 | 2.35 |
| | *S. venezuelae* | 0.14 | 1.85 | 0.37 | 3.50 | - | - |
| Pareto | *S. coelicolor* | 0.47 | 0.96 | 0.00 | 1.00 | 0.26 | 0.94 |
| | *S. albus J0174* | 0.08 | 0.75 | 0.34 | 1.09 | 0.43 | 1.04 |
| | *S. G4A3* | 0.14 | 0.74 | 0.00 | 0.81 | 0.02 | 1.11 |
| | *S. S26F9* | 0.44 | 1.09 | 0.15 | 1.02 | 0.01 | 1.11 |
| | *S. venezuelae* | 0.40 | 0.89 | 0.00 | 0.86 | - | - |

The exponential distribution, which is not heavy-tailed, was clearly a poor fit to the real data across all 5 strains and initial concentrations (Fig. 4.3c). Heavy-tailed distributions were a better fit to the data, although most could be rejected, especially for the strains with the greatest number of observed barcodes (data points). Both the log normal and Pareto distributions fit the data reasonably well, but the Pareto distribution was clearly a better fit to the most abundant barcodes (Fig. 4.3b). Strikingly, the optimal exponent (shape parameter $\boldsymbol{\alpha}$) of the Pareto distribution was near one for most strains. Distributions of this type are classified as *wild* because they have infinite variance. While it is extremely unlikely that a single cell could produce 100%

of all descendants in a short amount of time, this result highlights the heavy-tailed nature of the observed descendants distribution.

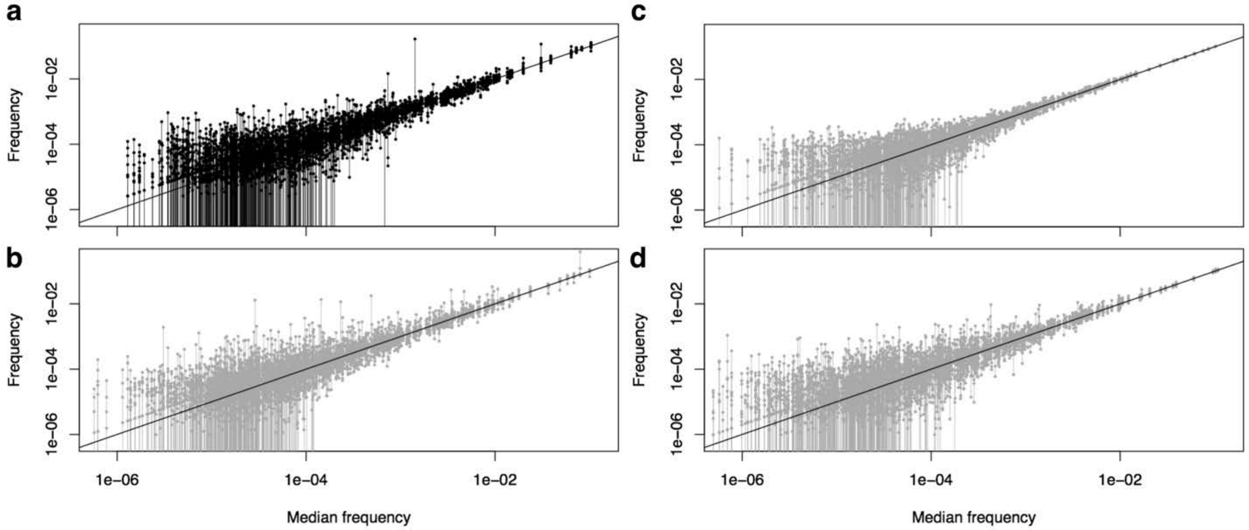


**Figure 4.3. Fitting the complete experimental result to simulation.** We performed simulations (gray) of the entire experiment with different descendants distributions to test their ability to recapitulate the observed data (black). **a**, The variation in relative barcode frequencies between 8 replicate populations is shown for the strain *S. albus* at full initial concentration, containing 1524 unique barcodes. The vertical lines connect the observed relative frequencies of each of the 8 biological replicates (points) corresponding to a given barcode. Note that the line may extend to zero in cases where a barcode was not observed in one or more of the 8 replicates. **b**, The full dataset was best fit by a simulation (gray) under a Pareto-shaped descendants distribution. The Pareto distribution was the only shape that could not be statistically rejected and is the most heavy-tailed of the distributions tested. **c**, The exponential distribution, which is not heavy-tailed, was a poor fit to the data as it quickly converged to low variance for the most abundant barcodes. **d**, The log normal distribution displayed less variation than the Pareto distribution among the most abundant barcodes.

## DISCUSSION

In this study we created and applied a procedure for determining the descendants distribution arising from a population of nearly-clonal bacteria. Surprisingly, the descendants distribution was closely approximated by a power-law distribution with a heavy-tail (Fig. 4.3), resulting in a wide range of relative abundances after only a short time (Fig. 4.2). This raises the question: what causes the extreme variability in abundances? To answer this question, it is helpful to separate sources of variability into two components: genetic and phenotypic. High genetic variability would require that some barcode lineages contained mutants of substantially higher fitness than the rest of the population. In contrast, high phenotypic variability would necessitate that some individuals have substantially higher fitness than their genetically identical siblings, where fitness is defined as the number of descendants.

One genetic basis for variation is that some barcodes have pre-existing mutations that impart a higher growth rate, resulting in an exponential divergence in relative abundance over time. However, the method we used to fit the data to the simulation (Fig. 4.3) is robust to per-barcode selection coefficients because it relies on intra-barcode variability. We confirmed this by incorporating selection coefficients into our simulation (see Methods) and observed no effect on the fitted parameters. Another way to uncover differences in inter-barcode selection coefficients is to look for correlations between the final relative frequencies of rare barcodes. We tested this by plotting the relative frequency of barcodes that were only present in 2 of 8 replicates (Appendix C: Supplemental Fig. 3). The correlation between replicate barcodes was extremely low (Pearson's r = 0.08), indicating that inter-barcode selection coefficients are not a major source of the observed variability between replicates.

These results do not rule out the possibility that there were rare individuals within a barcode lineage with new or recently acquired beneficial mutations. Such mutants would likely have had to arise after the start of the experiment in order to only be present in a minority of replicates. Given the high number of positively-skewed replicates, it is implausible that this many mutants of large effect size could occur so rapidly. Furthermore, we estimate that most cells only doubled about 10-15 times over the course of the experiment. Even a large growth rate advantage of 10% would be expected to result in at most a 3-fold variability in final abundances. Nonetheless, it is well known that mutation is a major determinant of fitness variation in populations, and we cannot rule out the fact that some of the variance in the descendants distribution was attributable to genetic differences.

Setting aside genetic variation, there are several sources of phenotypic variability that could result in large differences in the number of descendants. Previously it has been shown that germination times for a clonal population of *Streptomyces* are approximately normally distributed (Xu and Vetsigian). When followed by exponential growth this would be expected to result in a log normal distribution in the number of descendants per individual. This initial variation could be further amplified by small differences in growth rate, for example if larger colonies grow faster because they are more effective at obtaining nutrients. Any autocorrelation in growth rates between generations would result in multiplicative accumulation of variation over time, which is expected to lead to a heavy-tailed descendants distribution in accordance with the multiplicative central limit theorem.

The heavy-tailed nature of the descendants distribution is anticipated to have several effects on bacterial evolution. First, extreme stochastic variability will act to purge genetic diversity over time, causing the force of genetic drift to have a greater influence over allele frequencies than would otherwise be expected. Second, the effective population size is inversely proportional to the variance in the offspring distribution (see Methods). Hence, greater variation in the number of descendants per individual is expected to lower the effective population size (Hedrick 2005), thereby raising the lower-bound at which weak selective pressure can effectively act. While the implications of low effective population size have been considered for sexual species, they have largely been discounted for bacteria because of their large census population sizes. In the future, it would be interesting to study the descendants distribution for non-sporulating bacteria to see whether the conclusions drawn in this study can be generalized to other bacteria.

## METHODS

### Construction of barcoded strains

Oligonucleotides 5'-GATCCACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' and 5'-*S20*-*N30*-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTG/3Phos/ were purchased from Integrated DNA Technologies (IDT). The latter oligonucleotide is different for each strain library and contains a unique 20-nucleotide strain barcode (*S20*), a stretch of 30 random nucleotides that form the set of lineage barcodes (*N30*), and a 3'-phosphate modification. To permit robust identification of a strain in the presence of sequencing errors, the *S20* sequences were designed using Edittag (Faircloth and Glenn 2012). The 34-nucleotide complementary region of the two oligonucleotides were annealed, made double stranded using Klenow Polymerase (Promega), and then modified using T4 Polynucleotide Kinase (New England BioLabs), which removes the 3'-phosphate and adds 5'-phosphates. Subsequently, this DNA insert was ligated into plasmid pSRKV004 cut with BamHI and EcoRV (New England BioLabs). The plasmid pSRKV004 is a derivative of the integrating plasmid pSET152 (Hopwood et al. 2000) in which the orientation of EcoRV and BamHI sites in the multiple cloning site is reversed.

To reduce the background of pSRKV004 without inserts after ligation, the ligation mixture was digested with EcoRV and NotI (New England BioLabs) and then transformed into *E. coli* 10G ELITE cells (Lucigen) via electroporation. Transformants were selected on Luria broth (LB) plates with 50 μg/ml Apramycin and the pool of transformants underwent plasmid preparation (miniprep) using a commercial kit (Promega). The miniprep was again digested with EcoRV and NotI and the resulting library was introduced into the conjugation helper strain ET12567-pUZ8002 (Hopwood et al. 2000) via chemical transformation. Transformants were selected on LB + 15 µg/ml Chloramphenicol + 50 µg/ml Kanamycin and 50 µg/ml Apramycin plates, pooled, and grown in liquid LB containing 15 µg/ml Chloramphenicol, 50 µg/ml Kanamycin and 50 µg/ml Apramycin for 2-3 hours in a 37°C shaker.

This *E. coli* culture was used for conjugation into the desired *Streptomyces* strain using the standard protocol (Hopwood et al. 2000). Briefly, the transformed conjugation helper strain was mixed with *Streptomyces* spores, the bacterial mix was grown on mannitol-salt (MS) agar for 16 hours and then overlaid with Apramycin (100 µg/ml) and Nalidixic acid (50 µg/ml). Strains successfully undergoing conjugation integrate the plasmid at a phage attachment site in their genomic DNA (Sun et al. 1999). Barcoded libraries were prepared by scraping spores from exconjugants and selecting against *E. coli* carryover by propagating the spores on *Streptomyces* Isolation Medium (D’Costa et al. 2006) supplemented with 50 µg/ml Nalidixic acid and 100 µg/ml Apramycin for two growth cycles.

**Strains and growth conditions**

Five barcoded *Streptomyces* strains were chosen based on having more than 100 distinct barcodes per strain. These five strains were *S. coelicolor*, *S. albus J1074*, *S. G4A3* (Vetsigian, Jajoo, and Kishony 2011), *S. S26F9* (E. Wright and Vetsigian 2016), and *S. venezuelae*. Full concentration spore stocks were diluted 10-fold and 100-fold to generate three concentrations, and aliquoted into 8 replicates per concentration, each containing a single strain (120 total populations). Each replicate (30 µl) was used to inoculate 1 ml of 1/10th ISP2 liquid (10 g Malt extract, 4 g Yeast extract, and 4 g Dextrose per 1 L) in a sterile 1.5 ml polystyrene tube (Evergreen Scientific). A small hole was made in the cap of each tube to allow air flow. Tubes were incubated for 7.5 days at 28°C while shaking at 200 rpm.

## DNA extraction and sequencing

After growth, strains were centrifuged at 2000 rpm for 10 minutes to pellet the cells. A 750 µl volume of supernatant was removed, leaving about 150 µl remaining. Note that some of the original volume was lost to evaporation during growth. The remaining volume containing mycelium was sonicated at 100% amplitude for 3 minutes using a Model 505 Sonicator with Cup Horn (QSonica) while the samples were completely enclosed. After sonication, the samples were centrifuged, and the supernatant containing DNA was used as template for PCR amplification.

PCR primers (Appendix C: Supplemental Table 1) were designed with unique 8-nucleotide i5 and i7 index sequences and Illumina adapters. The random barcode (*N30*) sequence occurs at the start of the sequencing read to assist with cluster detection on the Illumina platform. Since strains could be distinguished by their unique sequence specific barcode (*S20*), we amplified each replicate using a unique dual-index combination, but used the same set of combinations for all 5 strains. Hence, the *S20* region effectively acted as a third index sequence that allowed the 5 strains sharing dual-index primers to be correctly de-multiplexed. This permitted all 24 samples per strain to be multiplexed without needing to have some samples only separated by a single i5 or i7 index. All strains were amplified separately before pooling, requiring a total of 120 PCR reactions (5 strains with 24 replicates each). In addition, we performed two more technical replicates of one sample belonging to each strain.

Extracted DNA was amplified using a qPCR reaction consisting of a 2 min denaturation step at 95°C, followed by 40 cycles of 20 sec at 98°C, 15 sec at 67°C, and 15 sec at 80°C. Each well contained 10 µL of iQ Supermix (Bio-Rad), 1.6 µL of 10 µM left primer, 1.6 µL of 10 µM right primer, 4 µL of DNA template, and 2.8 µL of reagent grade $H_2O$ per sample. A standard curve of pure template DNA was used to estimate the initial DNA copy number per sample. The resulting amplicons were pooled by sample and purified using the Wizard SV-Gel and PCR Cleanup System (Promega). Samples were sequenced by the UW-Madison Biotechnology Center on an Illumina Hi-Seq 2500 in rapid mode. Sequences were deposited into the Short Read Archive (SRA) repository under accession number PRJNA353868.

### DNA Sequence analysis

Using the R (R Core Team 2016) package DECIPHER (E. S. Wright 2016), DNA sequencing reads were filtered at a maximum average error of 0.1% (Q30) to lessen the degree of cross-talk between dual-indexed samples (E. S. Wright and Vetsigian, 2016). Sequences were assigned to the appropriate strain by exact matching the *S20*, and the nearest barcode by clustering *N30* sequences within an edit distance of 5. To completely eliminate any remaining cross-talk, we subtracted 0.01% + 5 reads from the count of every barcode by sample. The remaining reads were normalized by dividing by the total number of reads per sample. The final result of this process was a matrix of read counts for each unique barcode across every sample by strain.

### Complete simulations using different descendants distributions

We performed comprehensive simulations in order to test the fit of different distributions to the 8 replicates per strain at a given concentration. The simulation begins by averaging the relative barcode frequency distributions across all 8 replicates to generate a background barcode frequency distribution. This distribution is reasonably well approximated by an exponential distribution, but is truncated because very rare barcodes are not observed. We supplemented these rare barcodes by extrapolating the exponential distribution and adding back “virtual” barcodes at less than the 10th percentile of relative frequency. Since these barcodes are extremely rare, they collectively have minimal effect on the relative frequencies of the other barcodes.

The simulation begins by Poisson sampling a given number of initial barcodes from this background distribution. These barcodes then give rise to a number of final barcodes in accordance with the given distribution's shape. The barcode frequencies are then normalized to sum to 1, meaning that the simulation result is invariant to each distribution's scale (i.e., mean). The corresponding shape parameter for each distribution is the *rate* λ (exponential), *shape k* (Weibull), *scale* σ (log normal), and *exponent* α (Pareto). Next, the simulation subsamples the distribution in accordance with the observed number of sequencing reads and the predicted number of initial templates in PCR. To better reflect the real data, these two steps are performed based on the estimated number of reads and initial templates in each replicate.

Hence, there are only two free parameters in the simulation, one specifying the initial (census) population size, and a second controlling the shape of the distribution. For a given shape parameter, the initial population size was optimized to yield the same number of observed data points (barcodes) as in the real data. We found this method to generally approximate the initial population size estimated by plate counting the initial inoculum (Appendix C: Supplemental Fig. 2). We performed a sweep across a range of shape values to find the optimum based on the result of 300 simulations.

To define an optimality criterion, we split the simulation results into successive bins by relative frequency, with 10 bins that were evenly spaced in log-space per order of magnitude. We then compared the area between the cumulative frequency distributions of the real data within a bin and that of the combined result of the 300 simulations. The shape parameter with the least total separation between empirical cumulative distribution functions was considered optimal. We tested whether a distribution could be rejected by comparing the optimality score of the real data to that of the 300 simulations tested against one another through leave-one-out. The reported p-value represents the fraction of simulations with at least as extreme of a score as the real data.

**Effective population size**

In the Wright-Fisher (S. Wright 1931) idealized population model, the variance effective population size for haploid organisms is:

$$N_e = \frac{p_0(1 - p_0)}{\mathrm{var}(p_1)}$$

Where $p_0$ is the initial proportion of a genotype, and $p_1$ is the proportion after one generation. We are interested in the distribution of $p_1$ for single individuals, for which we can substitute:

$$p_0 = 1/N$$

For large *N*, we can approximate:

$$N_e = \frac{1/N\,(1 - 1/N)}{\mathrm{var}(p_1)} \approx \frac{1/N}{\mathrm{var}(p_1)} = \frac{N}{\mathrm{var}(C)}$$

Here we can consider $p_1$ as the proportion of children (C) arising from the original genotype ($p_1$=*C/N*). Therefore, the effective population size is inversely related to the variance of the offspring distribution (C). While this model is highly idealized, it serves as a basis for the notion that $N_e$ will be much less than *N* when the descendants distribution is highly skewed.

## ACKNOWLEDGEMENTS

We thank Sriram for preparing the barcoded strains used in this work, Ye Xu for help with microscopy experiments, the UW Biotechnology Center DNA Sequencing Facility for performing the Illumina sequencing associated with this study, and the UW-Madison Center for High Throughput Computing (CHTC) for providing compute resources. We are grateful for feedback from David Baum, Anthony Ives, and Laurence Loewe during preparation of the manuscript. This work was supported by the Simons Foundation, Targeted Grant in the Mathematical Modeling of

Living Systems Award 342039, the National Science Foundation Grant DEB 1457518, and the National Institute of Food and Agriculture, US Department of Agriculture, Hatch project 1006261.

## AUTHOR CONTRIBUTIONS

EW performed the experiments and simulations. EW and KV designed the study, analyzed the data and wrote the manuscript.

# CHAPTER 5: RECAPITULATION

## SUMMARY OF CHAPTER 2

It is largely unknown how the process of microbial community assembly is affected by the order of species arrival, initial species abundances and interactions between species. A minimal way of capturing competitive abilities in a frequency-dependent manner is with an invasibility network specifying whether a species at low abundance can increase in frequency in an environment dominated by another species. Here, using a panel of prolific small molecule producers and a habitat with feast-and-famine cycles, we show that the most abundant strain can often exclude other strains – resulting in bistability between pairs of strains. Instead of a single winner, the empirically determined invasibility network is ruled by multiple strains that cannot invade each other, and does not contain loops of cyclic dominance. Antibiotic inhibition contributes to bistability by helping producers resist invasions while at high abundance and by reducing producers' ability to invade when at low abundance.

## SUMMARY OF CHAPTER 3

Multiplexing multiple samples during Illumina sequencing is a common practice and is rapidly growing in importance as the throughput of the platform increases. Misassignments during de-multiplexing, where sequences are associated with the wrong sample, are an overlooked error mode on the Illumina sequencing platform. This results in a low rate of cross-talk among multiplexed samples and can cause detrimental effects in studies requiring the detection of rare variants or when multiplexing a large number of samples. We observed rates of cross-talk averaging 0.24% when multiplexing 14 different samples with unique i5 and i7 index sequences. This cross-talk rate corresponded to 254,632 misassigned reads on a single lane of the Illumina

HiSeq 2500. Notably, all types of misassignment occur at similar rates: incorrect i5, incorrect i7, and incorrect sequence reads. We demonstrate that misassignments can be nearly eliminated by quality filtering of index reads while preserving about 90% of the original sequences. Cross-talk among multiplexed samples is a significant error mode on the Illumina platform, especially if samples are only separated by a single unique index. Quality filtering of index sequences offers an effective solution to minimizing cross-talk among samples. Furthermore, we propose a straightforward method for verifying the extent of cross-talk between samples and optimizing quality score thresholds that does not require additional control samples and can even be performed *post hoc* on previous runs.

## SUMMARY OF CHAPTER 4

Variance in reproductive success is a major determinant of the degree of genetic drift in a population. While it is well known that many animals exhibit high variance in their number of progeny, far less is known about the corresponding distribution in microorganisms. Here we study the distribution of descendants that may arise from a single bacterium after a few generations of growth. We find that the descendants distribution is heavy-tailed, meaning that a few cells effectively "win the jackpot" to become a large proportion of the population. We attribute this skew to the amplification of small growth differences that begin with variation in lag time before exponential growth. The product of these differences results in a heavy-tailed distribution of descendants that is best fit by a power-law (Pareto) distribution with an exponent near 1. This result implies that stochastic effects have a major influence over allele dynamics, even in growth conditions that are far more homogeneous than the natural environment.

## CONCLUDING REMARKS

One of the perils of being a graduate student is that you see your project(s) everywhere you look. This was certainly the case for me with the first project in this thesis. One of the main conclusions was that bistability, the existence of two stable states, is widespread among *Streptomyces* in our experimental system. I began seeing bistability everywhere: gut communities (*Clostridium difficile* or not), antibiotic treatment outcomes (cured or not), the *lac* operon (expressed or not), light switches (on or not), and so on. Bistability, it seemed, was more ubiquitous than I had thought! Nevertheless, one place I have yet to find abundant bistability is in good scientists. One cannot simply be a biologist or a computer scientist, a theoretician or an empiricist, a reductionist or a holist (i.e., systems biologist), a field researcher or a lab rat. The conflict between these states is a false one, or, at least, the states themselves are unstable. Each is complementary to the other. Each is enlightened by the other. Each is incomplete without the other.

Hence, as I move into the next phase of my career, I reflect upon the origins of good science. In doing so I aspire to continue asking fundamental questions, and seek their answers through a combination of approaches. In other words, I do not wish to be part of a wet lab or a dry lab, but a *soggy* lab. One that is steeped in all the idiosyncrasies of biology. One where scientists are comfortable straddling the traditional walls between domains and excited about the continuous shades of gray that pervade nature. One where a problem's intricacies are viewed almost as a testament to the problem's authenticity. I believe this philosophy underlies any successes that may be contained in these pages, and I hope that I will find many others along this journey who are embracing the many facets of the thing that we all call *science*.

## REFERENCES

Abràmoff, Michael D, Paulo J Magalhães, and Sunanda J Ram. 2004. “Image Processing with ImageJ.” *Biophotonics International* 11 (7). Laurin Publishing: 36–42.

Allen, B, J Gore, and M A Nowak. 2013. “Spatial Dilemmas of Diffusible Public Goods.” *eLife* 2 (0): e01169–69. doi:10.7554/eLife.01169.010.

Allesina, Stefano, and Si Tang. 2012. “Stability Criteria for Complex Ecosystems.” *Nature* 483 (7388). Nature Publishing Group: 205–8. doi:10.1038/nature10832.

Andersson, D I, and B R Levin. 1999. “The Biological Cost of Antibiotic Resistance.” *Current Opinion in Microbiology* 2 (5): 489–93.

Araki, Hitoshi, Robin S Waples, William R Ardren, Becky Cooper, and Michael S Blouin. 2007. “Effective Population Size of Steelhead Trout: Influence of Variance in Reproductive Success, Hatchery Programs, and Genetic Compensation Between Life-History Forms..” *Molecular Ecology* 16 (5): 953–66. doi:10.1111/j.1365-294X.2006.03206.x.

Barbe, V, M Bouzon, S Mangenot, B Badet, J Poulain, B Segurens, D Vallenet, P Marliere, and J Weissenbach. 2011. “Complete Genome Sequence of *Streptomyces Cattleya NRRL 8057*, A Producer of Antibiotics and Fluorometabolites.” *Journal of Bacteriology* 193 (18): 5055–56. doi:10.1128/JB.05583-11.

Batut, Bérénice, Carole Knibbe, Gabriel Marais, and Vincent Daubin. 2014. “Reductive Genome Evolution at Both Ends of the Bacterial Population Size Spectrum.” *Nature Reviews Microbiology* 12 (12): 841–50. doi:10.1038/nrmicro3331.

Bayley, S E, I F Creed, G Z Sass, and A S Wong. 2007. “Frequent Regime Shifts in Trophic States in Shallow Lakes on the Boreal Plain: Alternative‘ Unstable’ States?.” *Limnology and Oceanography*.

Bendall, Matthew L, Sarah LR Stevens, Leong-Keat Chan, Stephanie Malfatti, Patrick Schwientek, Julien Tremblay, Wendy Schackwitz, et al. 2016. “Genome-Wide Selective Sweeps and Gene-Specific Sweeps in Natural Bacterial Populations,” January. Nature Publishing Group, 1–13. doi:10.1038/ismej.2015.241.

Benincà, Elisa, Klaus D Jöhnk, Reinhard Heerkloss, and Jef Huisman. 2009. “Coupled Predator-Prey Oscillations in a Chaotic Food Web.” *Ecology Letters* 12 (12): 1367–78. doi:10.1111/j.1461-0248.2009.01391.x.

Bentley, S D, K F Chater, A-M Cerdeno-Tarraga, G L Challis, N R Thomson, K D James, D E Harris, M A Quail, H Kieser, and D Harper. 2002. “Complete Genome Sequence of the Model Actinomycete *Streptomyces Coelicolor A3 (2)*.” *Nature* 417 (6885). Nature Publishing Group: 141–47.

Bernstein, H, H C Byerly, F A Hopf, and R E Michod. 1985. “Sex and the Emergence of Species.” *Journal of Theoretical Biology* 117 (4): 665–90.

Berry, D, K Ben Mahfoudh, M Wagner, and A Loy. 2011. “Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification.” *Applied and Environmental Microbiology* 77 (21): 7846–49. doi:10.1128/AEM.05220-11.

Chao, L, and B R Levin. 1981. “Structured Habitats and the Evolution of Anticompetitor Toxins in Bacteria.” *Proceedings of the National Academy of Sciences* 78 (10): 6324–28.

Charlesworth, Brian. 2009. “Fundamental Concepts in Genetics: Effective Population Size and Patterns of Molecular Evolution and Variation.” *Nature Reviews. Genetics* 10 (3): 195–205. doi:10.1038/nrg2526.

Chesson, Peter. 2000. “Mechanisms of Maintenance of Species Diversity.” *Annual Review of Ecology and Systematics*. JSTOR, 343–66.

Cordero, O X, H Wildschutte, B Kirkup, S Proehl, L Ngo, F Hussain, F Le Roux, T Mincer, and M F Polz. 2012. “Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance.” *Science* 337 (6099): 1228–31. doi:10.1126/science.1219385.

Cox, Murray P, Daniel A Peterson, and Patrick J Biggs. 2010. “SolexaQA: at-a-Glance Quality Assessment of Illumina Second-Generation Sequencing Data.” *BMC Bioinformatics* 11 (1): 485. doi:10.1186/1471-2105-11-485.

Coyte, Katharine Z, Jonas Schluter, and Kevin R Foster. 2015. “The Ecology of the Microbiome: Networks, Competition, and Stability.” *Science* 350 (6261): 663–66. doi:10.1126/science.aad2602.

D'Amore, Rosalinda, Umer Zeeshan Ijaz, Melanie Schirmer, John G Kenny, Richard Gregory, Alistair C Darby, Migun Shakya, Mircea Podar, Christopher Quince, and Neil Hall. 2016. “A Comprehensive Benchmarking Study of Protocols and Sequencing Platforms for 16S rRNA Community Profiling.” *BMC Genomics* 17 (January): 55. doi:10.1186/s12864-015-2194-9.

Darch, Sophie E, Stuart A West, Klaus Winzer, and Stephen P Diggle. 2012. “Density-Dependent Fitness Benefits in Quorum-Sensing Bacterial Populations.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (21): 8259–63. doi:10.1073/pnas.1118131109.

Davis, M A, K Thompson, and J Philip Grime. 2005. “Invasibility: the Local Mechanism Driving Community Assembly and Species Diversity.” *Ecography*.

Der, R, C Epstein, and J B Plotkin. 2012. “Dynamics of Neutral and Selected Alleles When the Offspring Distribution Is Skewed.” *Genetics* 191 (4): 1331–44. doi:10.1534/genetics.112.140038.

Diggle, Stephen P, Ashleigh S Griffin, Genevieve S Campbell, and Stuart A West. 2007. “Cooperation and Conflict in Quorum-Sensing Bacterial Populations.” *Nature* 450 (7168): 411–14. doi:10.1038/nature06279.

Dohm, J C, C Lottaz, T Borodina, and H Himmelbauer. 2008. “Substantial Biases in Ultra-Short Read Data Sets From High-Throughput DNA Sequencing.” *Nucleic Acids Research* 36 (16): e105–5. doi:10.1093/nar/gkn425.

Doroghazi, James R, Jessica C Albright, Anthony W Goering, Kou-San Ju, Robert R Haines, Konstantin A Tchalukov, David P Labeda, Neil L Kelleher, and William W Metcalf. 2014. “A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics.” *Nature Chemical Biology* 10 (11): 963–68. doi:10.1038/nchembio.1659.

Drake, J A. 1991. “Community-Assembly Mechanics and the Structure of an Experimental Species Ensemble.” *American Naturalist*.

Durrett, R, and S Levin. 1997. “Allelopathy in Spatially Distributed Populations.” *Journal of Theoretical Biology* 185 (2): 165–71.

D’Costa, Vanessa M, Katherine M McGrann, Donald W Hughes, and Gerard D Wright. 2006. “Sampling the Antibiotic Resistome.” *Science* 311 (5759): 374–77. doi:10.1126/science.1120800.

Faircloth, Brant C, and Travis C Glenn. 2012. “Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels.” Edited by Shin-Han Shiu. *PloS One* 7 (8): e42543. doi:10.1371/journal.pone.0042543.s022.

Faith, Jeremiah J, Philip P Ahern, Vanessa K Ridaura, Jiye Cheng, and Jeffrey I Gordon. 2014. “Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice.” *Science Translational Medicine* 6 (220): 220ra11. doi:10.1126/scitranslmed.3008051.

Foster, Kevin R, and Thomas Bell. 2012. “Competition, Not Cooperation, Dominates Interactions Among Culturable Microbial Species.” *Curbio* 22 (19). Elsevier Ltd: 1845–50. doi:10.1016/j.cub.2012.08.005.

Friedman, Jonathan, Logan M Higgins, and Jeff Gore. 2016. “Community Structure Follows Simple Assembly Rules in Microbial Microcosms.” doi:10.1101/067926.

Friman, Ville-Petri, Alessandra Dupont, David Bass, David J Murrell, and Thomas Bell. 2015. “Relative Importance of Evolutionary Dynamics Depends on the Composition of Microbial Predator–Prey Community,” December. Nature Publishing Group, 1–11. doi:10.1038/ismej.2015.217.

Fukami, Tadashi, Ian A Dickie, J Paula Wilkie, Barbara C Paulus, Duckchul Park, Andrea Roberts, Peter K Buchanan, and Robert B Allen. 2010. “Assembly History Dictates Ecosystem Functioning: Evidence From Wood Decomposer Communities.” *Ecology Letters* 13 (6): 675–84. doi:10.1111/j.1461-0248.2010.01465.x.

Gause, G F. 1935. “Experimental Demonstration of Volterra's Periodic Oscillation in the Numbers of Animals.” *J. Exp. Biol* 12 (1): 44–48.

Gillespie, J H. 1974. “Natural Selection for Within-Generation Variance in Offspring Number.” *Genetics*.

Griffin, Ashleigh S, Stuart A West, and Angus Buckling. 2004. “Cooperation and Competition in Pathogenic Bacteria.” *Nature* 430 (7003): 1024–27. doi:10.1038/nature02744.

Gupte, Mangesh, Pravin Shankar, Jing Li, S Muthukrishnan, and Liviu Iftode. 2011. “Finding Hierarchy in Directed Online Social Networks.” ACM, 557–66.

Hedrick, Philip. 2005. “Large Variance in Reproductive Success and the $N_E$/N Ratio.” *Evolution* 59 (7): 1596–99.

Hekstra, Doeke R, and Stanislas Leibler. 2012. “Contingency and Statistical Laws in Replicate Microbial Closed Ecosystems.” *Cell* 149 (5). Elsevier Inc.: 1164–73. doi:10.1016/j.cell.2012.03.040.

Hoban, Sean M, Massimo Mezzavilla, Oscar E Gaggiotti, Andrea Benazzo, Cock van Oosterhout, and Giorgio Bertorelle. 2013. “High Variance in Reproductive Success Generates a False Signature of a Genetic Bottleneck in Populations of Constant Size: a Simulation Study.” *BMC Bioinformatics* 14 (October): 309. doi:10.1186/1471-2105-14-309.

Hopwood, D A, T Kieser, M J Bibb, M J Buttner, and K F Chater. 2000. *Practical Streptomyces Genetics*. The John Innes Foundation.

Ives, A R, and S R Carpenter. 2007. “Stability and Diversity of Ecosystems.” *Science* 317 (5834): 58–62. doi:10.1126/science.1133258.

Kaltenpoth, Martin, Wolfgang Goettler, Sabrina Koehler, and Erhard Strohm. 2009. “Life Cycle and Population Dynamics of a Protective Insect Symbiont Reveal Severe Bottlenecks During Vertical Transmission.” *Evolutionary Ecology* 24 (2): 463–77. doi:10.1007/s10682-009-9319-z.

Kassen, Rees, and Paul B Rainey. 2004. “The Ecology and Genetics of Microbial Diversity.” *Annual Review of Microbiology* 58 (1): 207–31. doi:10.1146/annurev.micro.58.030603.123654.

Keller, Laurent, and Michael G Surette. 2006. “Communication in Bacteria: an Ecological and

Evolutionary Perspective." *Nature Reviews Microbiology* 4 (4): 249–58. doi:10.1038/nrmicro1383.

Kelsic, Eric D, Jeffrey Zhao, Kalin Vetsigian, and Roy Kishony. 2015. "Counteraction of Antibiotic Production and Degradation Stabilizes Microbial Communities." *Nature* 521 (7553): 516–19. doi:10.1038/nature14485.

Kim, Hyun Jung, James Q Boedicker, Jang Wook Choi, and Rustem F Ismagilov. 2008. "Defined Spatial Structure Stabilizes a Synthetic Multispecies Bacterial Community." *Proceedings of the National Academy of Sciences of the United States of America* 105 (47): 18188–93. doi:10.1073/pnas.0807935105.

Kinkel, Linda L, Daniel C Schlatter, Kun Xiao, and Anita D Baines. 2013. "Sympatric Inhibition and Niche Differentiation Suggest Alternative Coevolutionary Trajectories Among Streptomycetes." *The ISME Journal*, October. Nature Publishing Group, 1–8. doi:10.1038/ismej.2013.175.

Kircher, M, S Sawyer, and M Meyer. 2011. "Double Indexing Overcomes Inaccuracies in Multiplex Sequencing on the Illumina Platform." *Nucleic Acids Research* 40 (1): e3–e3. doi:10.1093/nar/gkr771.

Lallias, D, N Taris, P Boudry, F Bonhomme, and S Lapègue. 2010. "Variance in the Reproductive Success of Flat Oyster Ostrea Edulis L. Assessed by Parentage Analyses in Natural and Experimental Conditions." *Genetical Research* 92 (3): 175–87. doi:10.1017/S0016672310000248.

Leinonen, R, H Sugawara, M Shumway, on behalf of the International Nucleotide Sequence Database Collaboration. 2010. "The Sequence Read Archive." *Nucleic Acids Research* 39 (Database): D19–D21. doi:10.1093/nar/gkq1019.

Levin, B R. 1988. "Frequency-Dependent Selection in Bacterial Populations." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 319 (1196): 459–72.

Levy, Sasha F, Jamie R Blundell, Sandeep Venkataram, Dmitri A Petrov, Daniel S Fisher, and Gavin Sherlock. 2015. "Quantitative Evolutionary Dynamics Using High-Resolution Lineage Tracking." *Nature*, February. doi:10.1038/nature14279.

Lewin, Gina R, Amanda L Johnson, Rolando D Moreira Soto, Kailene Perry, Adam J Book, Heidi A Horn, Adrián A Pinto-Tomás, and Cameron R Currie. 2016. "Cellulose-Enriched Microbial Communities From Leaf-Cutter Ant (Atta Colombica) Refuse Dumps Vary in Taxonomic Composition and Degradation Ability." Edited by Marie-Joelle Virolle. *PloS One* 11 (3): e0151840. doi:10.1371/journal.pone.0151840.s010.

Lynch, M. 2003. "The Origins of Genome Complexity." *Science* 302 (5649): 1401–4. doi:10.1126/science.1089370.

Majeed, Hadeel, Osnat Gillor, Benjamin Kerr, and Margaret A Riley. 2010. "Competitive Interactions in *Escherichia Coli* Populations: the Role of Bacteriocins" 5 (1). Nature Publishing Group: 71–81. doi:10.1038/ismej.2010.90.

Martiny, Jennifer B Hughes, Brendan J M Bohannan, James H Brown, Robert K Colwell, Jed A Fuhrman, Jessica L Green, M Claire Horner-Devine, et al. 2006. "Microbial Biogeography: Putting Microorganisms on the Map." *Nature Reviews Microbiology* 4 (2): 102–12. doi:10.1038/nrmicro1341.

McMurdie, Paul J, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, Susan P Holmes, and Benjamin J Callahan. 2016. "DADA2: High-Resolution Sample Inference From Illumina Amplicon Data." *Nature Methods*, May. Nature Publishing Group, 1–7.

doi:10.1038/nmeth.3869.

Mebane, W R, Jr, and J S Sekhon. 2011. “Genetic Optimization Using Derivatives: the Rgenoud Package for R.” *Journal of Statistical Software*.

Moll, Jason D, and Joel S Brown. 2008. “Competition and Coexistence with Multiple Life-History Stages.” *The American Naturalist* 171 (6): 839–43. doi:10.1086/587517.

Nakamura, Kensuke, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, et al. 2011. “Sequence-Specific Error Profile of Illumina Sequencers.” *Nucleic Acids Research* 39 (13): e90. doi:10.1093/nar/gkr344.

Nelson, Michael C, Hilary G Morrison, Jacquelynn Benjamino, Sharon L Grim, and Joerg Graf. 2014. “Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys.” Edited by Markus M Heimesaat. *PloS One* 9 (4): e94249. doi:10.1371/journal.pone.0094249.s006.

Oliveira, Nuno M, Rene Niehus, and Kevin R Foster. 2014. “Evolutionary Limits to Cooperation in Microbial Communities.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (50): 17941–46. doi:10.1073/pnas.1412673111.

Pages, H, P Aboyoun, R Gentleman, and S DebRoy. “Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms.”

Petraitis, Peter. 2013. *Multiple Stable States in Natural Ecosystems*. OUP Oxford.

Powell, Jeff R, and Alison E Bennett. 2015. “Unpredictable Assembly of Arbuscular Mycorrhizal Fungal Communities.” *Pedobiologia - International Journal of Soil Biology*, December. Elsevier GmbH., 1–5. doi:10.1016/j.pedobi.2015.12.001.

Powell, Jeff R, Senani Karunaratne, Colin D Campbell, Huaiying Yao, Lucinda Robinson, and Brajesh K Singh. 2015. “Deterministic Processes Vary During Community Assembly for Ecologically Dissimilar Taxa.” *Nature Communications* 6. Nature Publishing Group: 1–10. doi:10.1038/ncomms9444.

Price, Morgan N, and Adam P Arkin. 2015. “Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes.” *mBio* 6 (6): e01302–15. doi:10.1128/mBio.01302-15.

R Core Team. 2016. “R: a Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Ratcliff, W C, and R F Denison. 2011. “Alternative Actions for Antibiotics.” *Science* 332 (6029): 547–48. doi:10.1126/science.1205970.

Rivett, Damian W, Thomas Scheuerl, Christopher T Culbert, Shorok B Mombrikotb, Emma Johnstone, Timothy G Barraclough, and Thomas Bell. 2016. “Resource-Dependent Attenuation of Species Interactions During Bacterial Succession,” February. Nature Publishing Group, 1–10. doi:10.1038/ismej.2016.11.

Schirmer, M, U Z Ijaz, R D'Amore, N Hall, W T Sloan, and C Quince. 2015. “Insight Into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform.” *Nucleic Acids Research* 43 (6): e37–e37. doi:10.1093/nar/gku1341.

Seyedsayamdost, Mohammad R, Matthew F Traxler, Shao-Liang Zheng, Roberto Kolter, and Jon Clardy. 2011. “Structure and Biosynthesis of Amychelin, an Unusual Mixed-Ligand Siderophore From Amycolatopsis Sp.AA4.” *Journal of the American Chemical Society* 133 (30): 11434–37. doi:10.1021/ja203577e.

Sun, J, G H Kelemen, J M Fernández-Abalos, and M J Bibb. 1999. “Green Fluorescent Protein as a Reporter for Spatial and Temporal Gene Expression in *Streptomyces Coelicolor A3(2)*.” *Microbiology (Reading, England)* 145 ( Pt 9) (September): 2221–27. doi:10.1099/00221287-

145-9-2221.

Trosvik, Pål, Knut Rudi, Tormod Næs, Achim Kohler, Kung-Sik Chan, Kjetill S Jakobsen, and Nils C Stenseth. 2008. "Characterizing Mixed Microbial Population Dynamics Using Time-Series Analysis." *The ISME Journal* 2 (7): 707–15. doi:10.1038/ismej.2008.36.

Vannette, Rachel L, and Tadashi Fukami. 2013. "Historical Contingency in Species Interactions: Towards Niche-Based Predictions." Edited by Tim Wootton. *Ecology Letters* 17 (1): 115–24. doi:10.1111/ele.12204.

Vetsigian, Kalin, Rishi Jajoo, and Roy Kishony. 2011. "Structure and Evolution of *Streptomyces* Interaction Networks in Soil and in Silico." Edited by Jonathan A Eisen. *PLoS Biology* 9 (10): e1001184. doi:10.1371/journal.pbio.1001184.g005.

Widder, Stefanie, Rosalind J Allen, Thomas Pfeiffer, Thomas P Curtis, Carsten Wiuf, William T Sloan, Otto X Cordero, et al. 2016. "Challenges in Microbial Ecology: Building Predictive Understanding of Community Function and Dynamics," March. Nature Publishing Group, 1–12. doi:10.1038/ismej.2016.45.

Wintermute, E H, and P A Silver. 2010. "Dynamics in the Mixed Microbial Concourse." *Genes & Development* 24 (23): 2603–14. doi:10.1101/gad.1985210.

Wright, E S, and K H Vetsigian. 2016. "DesignSignatures: a Tool for Designing Primers That Yields Amplicons with Distinct Signatures." *Bioinformatics* 32 (10): 1565–67.

Wright, Erik S. 2015. "DECIPHER: Harnessing Local Sequence Context to Improve Protein Multiple Sequence Alignment." *BMC Bioinformatics* 16: 322. doi:10.1186/s12859-015-0749-z.

Wright, Erik S. 2016. "Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R." *The R Journal* 8 (1): 352–59.

Wright, Erik S, and Kalin H Vetsigian. 2016. "Quality Filtering of Illumina Index Reads Mitigates Sample Cross-Talk." *BMC Genomics*, 1-7.

Wright, Erik S, L Safak Yilmaz, Sri Ram, Jeremy M Gasser, Gregory W Harrington, and Daniel R Noguera. 2014. "Exploiting Extension Bias in Polymerase Chain Reaction to Improve Primer Specificity in Ensembles of Nearly Identical DNA Templates." *Environmental Microbiology* 16 (5). Wiley Online Library: 1354–65.

Wright, Erik, and Kalin Vetsigian. 2016. "Inhibitory Interactions Promote Frequent Bistability Among Competing Bacteria." *Nature Communications* 7: 11274. doi:10.1038/ncomms11274.

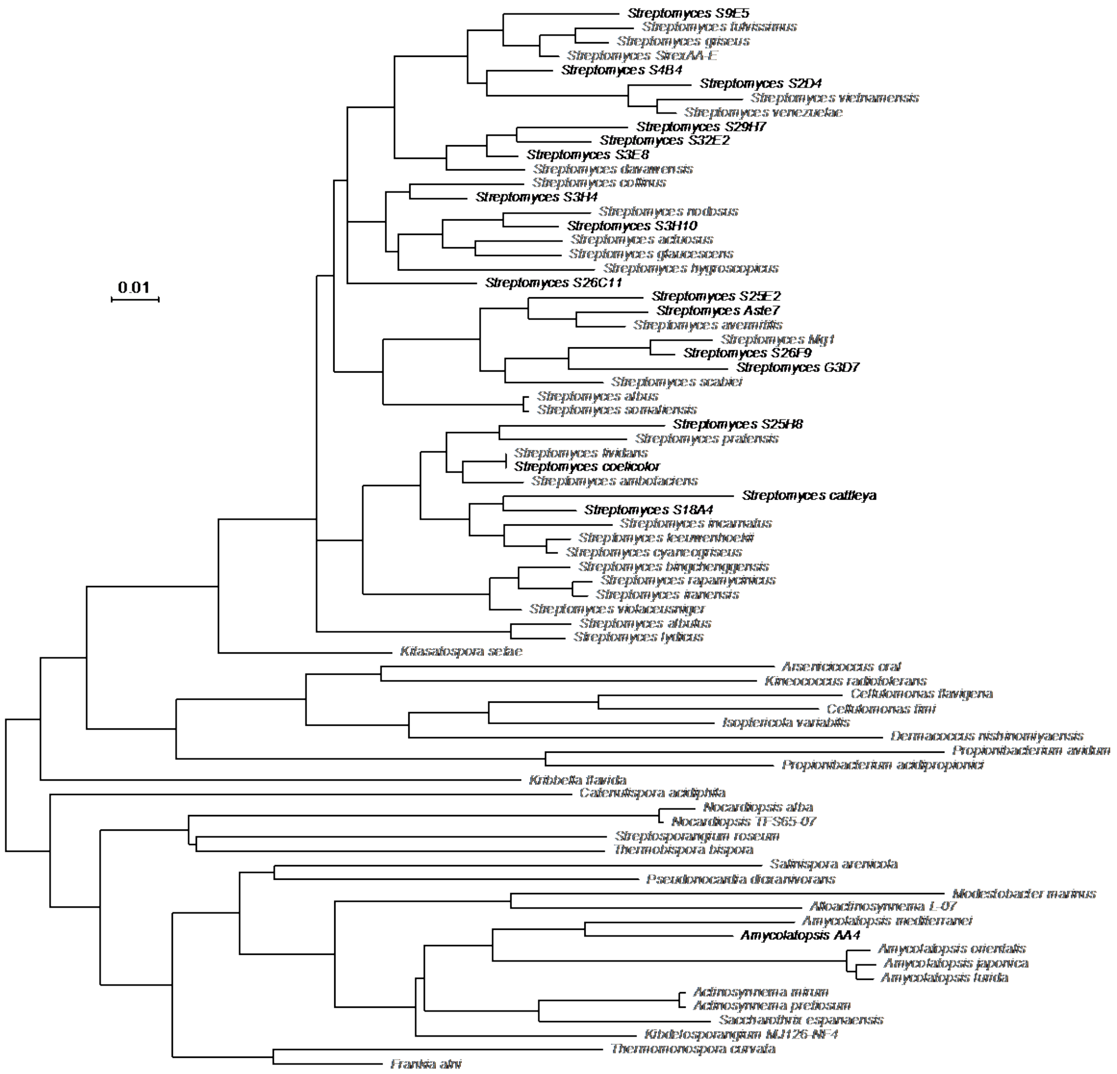
Wright, S. 1931. "Evolution in Mendelian Populations." *Genetics* 16 (2): 97–159.

Wu, Liyou, Chongqing Wen, Yujia Qin, Huaqun Yin, Qichao Tu, Joy D Van Nostrand, Tong Yuan, Menting Yuan, Ye Deng, and Jizhong Zhou. 2015. "Phasing Amplicon Sequencing on Illumina Miseq for Robust Environmental Microbial Community Analysis." *BMC Microbiology*, June. BMC Microbiology, 1–12. doi:10.1186/s12866-015-0450-4.

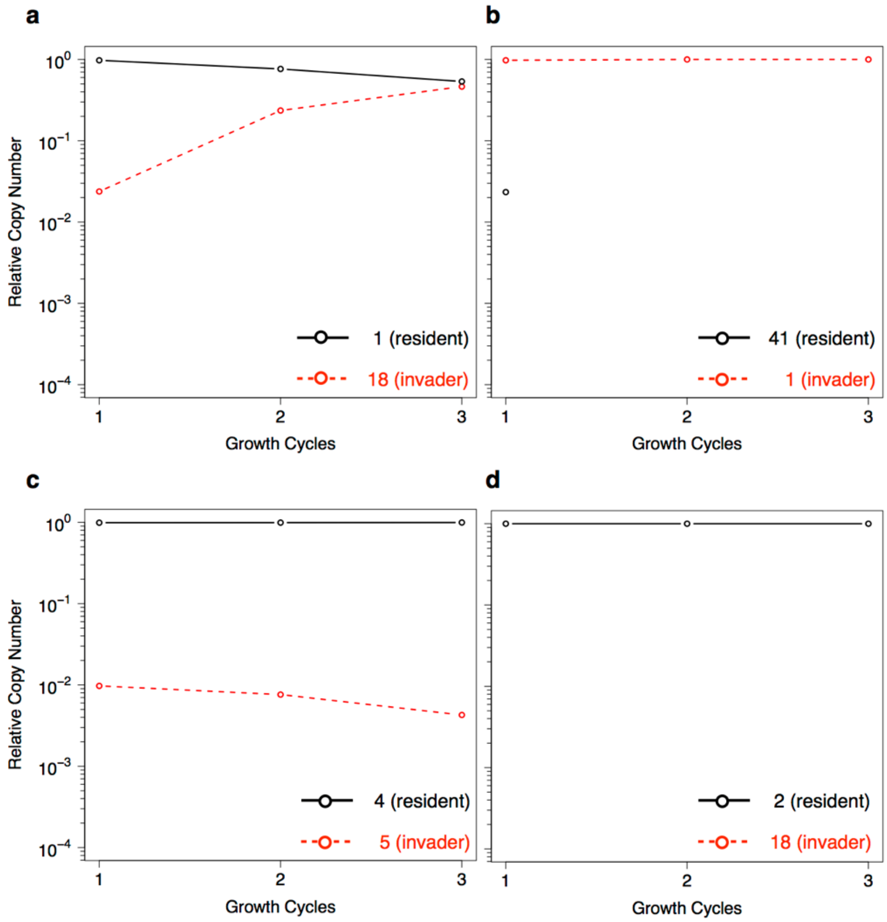
Xu, Ye, and Kalin H Vetsigian. "Diverse Promotional and Inhibitory Interactions Among Germinating *Streptomyces* Spores." *The ISME Journal*, inReview.

# APPENDIX

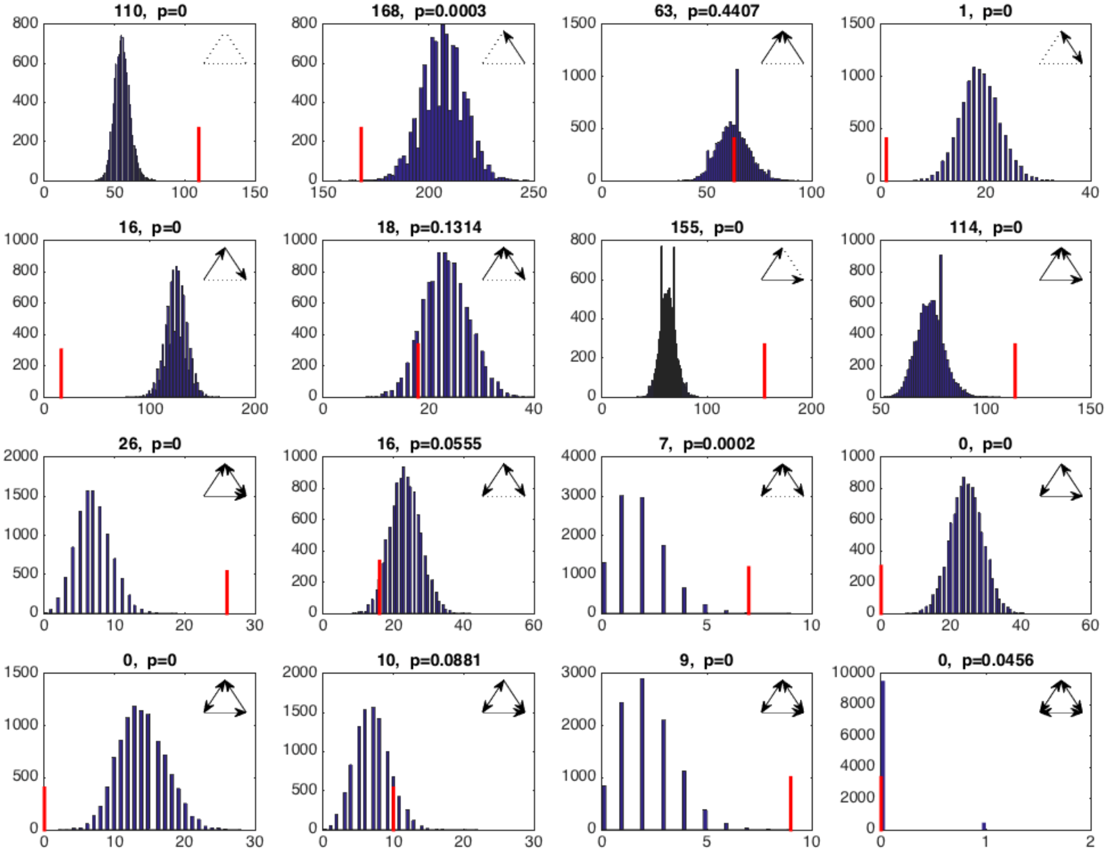
## APPENDIX A: SUPPLEMENTAL MATERIAL FOR CHAPTER 2



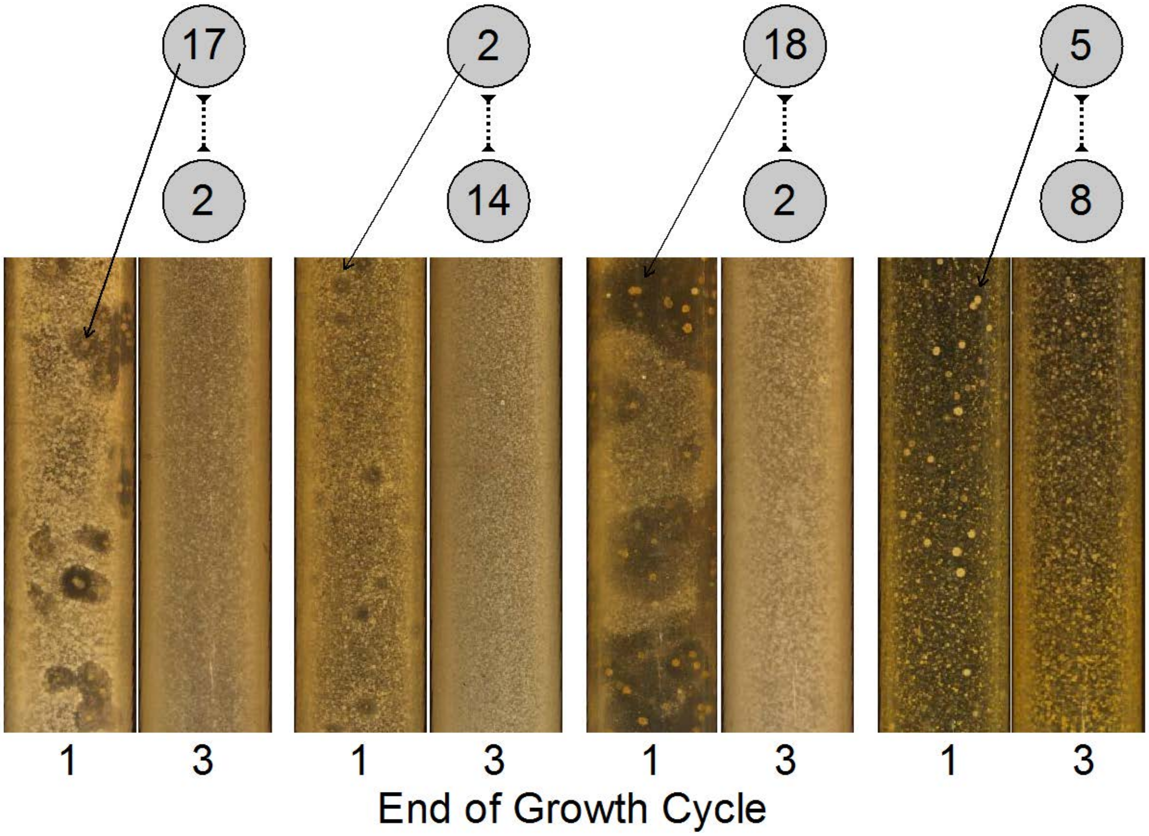
**Supplemental Figure 1.** Maximum likelihood tree based on 643 nucleotides of the *rpoB* gene belonging to strains used in this study (black labels) and other related strains (gray labels). Strains used in this study cover much of the breadth of known *Streptomyces*. Scale bar shows the expected number of substitutions per site.



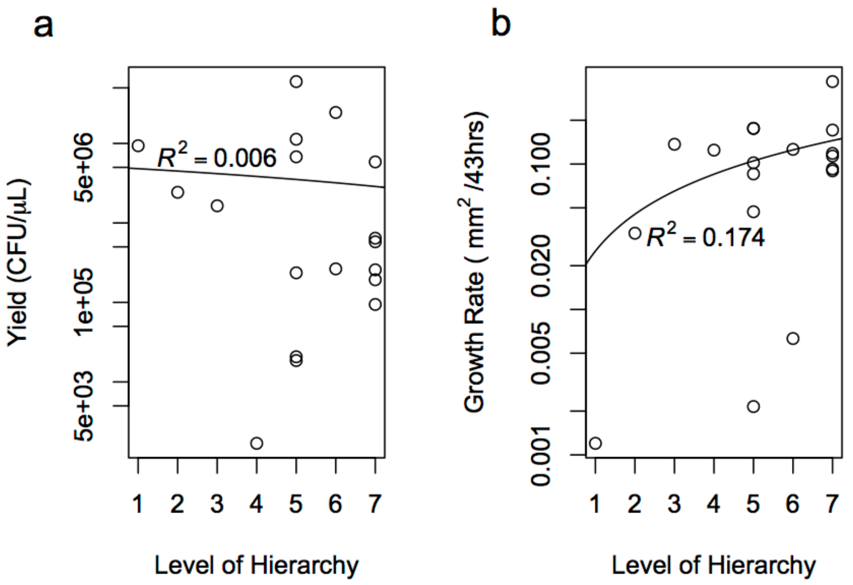
**Supplemental Figure 2.** Examples of community dynamics during (**a**) a slow invasion, (**b**) a fast invasion, or (**c**, **d**) no invasion. Relative copy number is the fraction of *rpoB* sequencing reads that matched each of the two species at a given growth cycle after background subtraction (see Methods section in the main text). In some cases of non-invasion (**d**) the invader never rose above the lower detection limit, although the presence of the invader could often be visually confirmed in the tube during the first growth cycle (Appendix A: Supplemental Fig. 4). In a subset of 16 of these cases we further confirmed the absence of the invader after the third growth cycle using quantitative PCR, which has a superior lower detection limit (Appendix A: Supplemental Table 3).



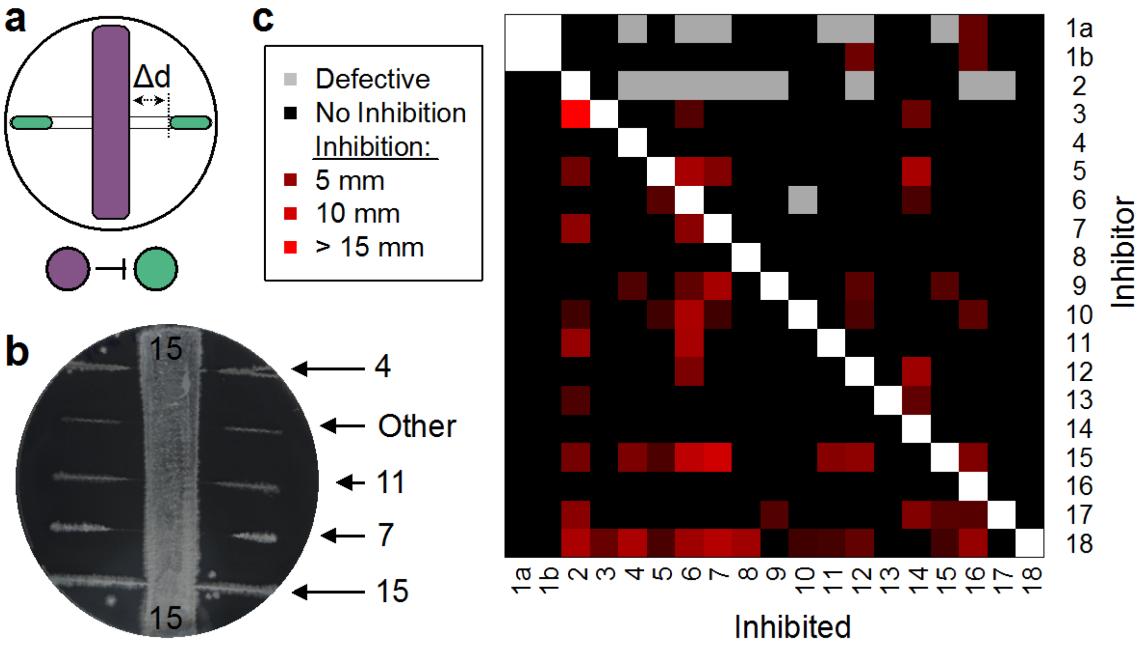
**Supplemental Figure 3.** Distribution of triplet motifs in randomized invasion networks (blue histograms) relative to the observed number of each motif (red lines). Enrichment for transitivity of hierarchy is evident from histograms (2, 1), (2, 4), (3, 4), and (4, 1) in (row, column) format. Enrichment for transitivity of bistability is evident from histograms (1, 1), (1, 2), and (1, 4). Absence of the 'rock-paper-scissors' dynamic is shown in (3, 4).



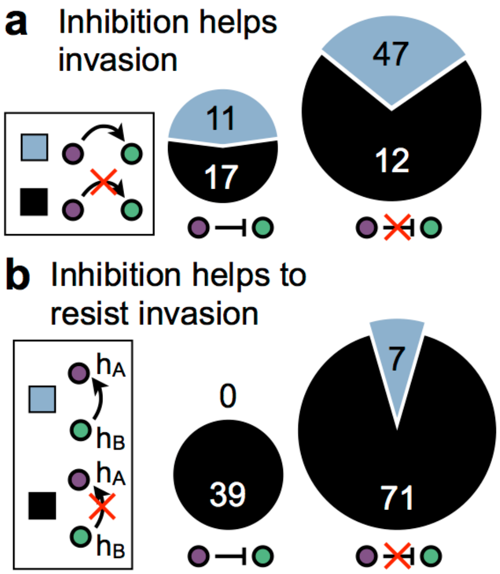
**Supplemental Figure 4.** Example images of cases where the higher ranked strain in a bistable pairing is visible in the tube after the first growth cycle, but disappears by the third growth cycle.



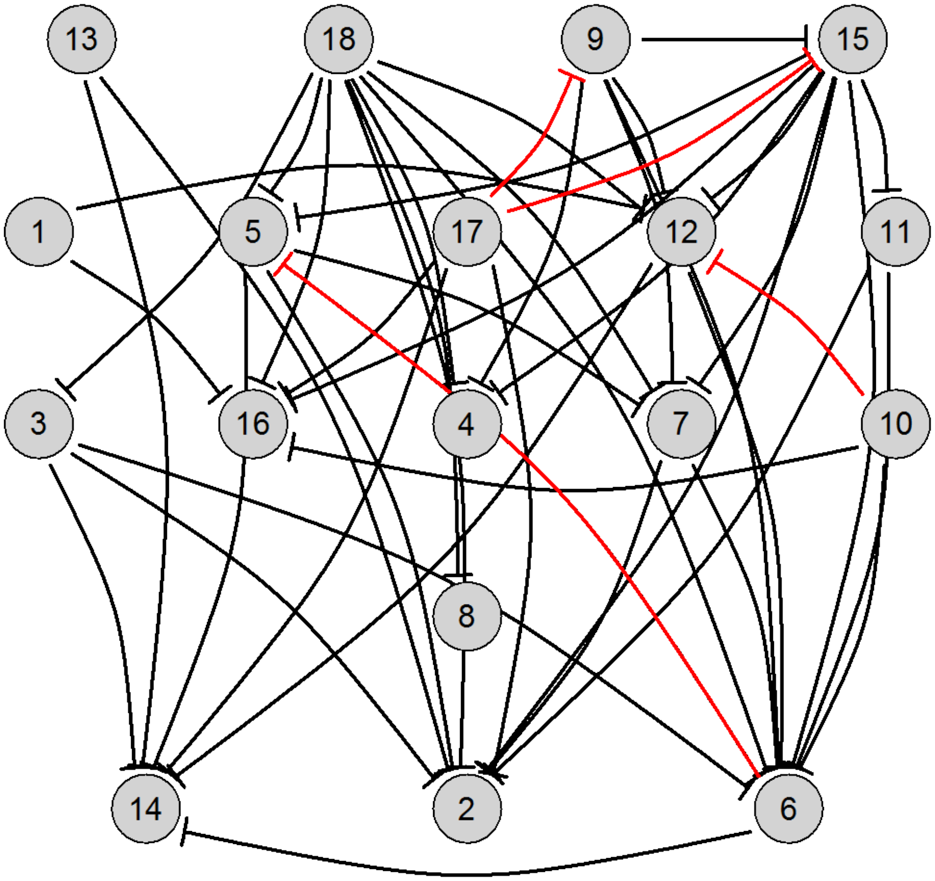
**Supplemental Figure 5.** Both yield (**a**) and growth rate (**b**) are largely uncorrelated with hierarchy level. Yield was measured for each strain grown by itself after three growth cycles. Colony size was measured under a microscope for separate colonies after 43 hours of growth (see Methods). Note the log-scaled y-axes, which cause the best-fit trend-lines to appear curved.



**Supplemental Figure 6.** Measurement of the inhibition matrix. **a**, Inhibition in the cross-streaking assay was measured as the distance an abundant strain (the inhibitor) was able to prevent sporulation of a less abundant strain intersecting it on a petri dish. **b**, Example experimental results for strain 15 as the inhibitor and four other strains being inhibited. **c**, The matrix of pairwise inhibitions included several strains that were inhibited by most others, and several strains that were not inhibited by any others. The diagonal is white because a strain can grow adjacent to itself (no self-inhibition).



**Supplemental Figure 7.** Correlations between invasions and inhibitions complementing Fig. 2.4bc in the main text. **a**, Inhibition appears to increase the likelihood of invasion when using data from all pairs (p = 0.02). **b**, Inhibition helps to resist invasion (p = 0.07) using a subset in which the resident is at a higher hierarchy level ($3 \geq h_A - h_B > 0$).



**Supplemental Figure 8.** One of 11 top scoring alternative hierarchies created by not scoring invasions between pairs that have an inhibitory interaction in either direction. Few inhibitions are directed against the invasion hierarchy, indicating that the downward bias of inhibitions is not entirely due to a role of inhibitions in shaping the invasion hierarchy.

**Supplemental Table 1.** Strains of *Streptomyces* bacteria used in this study.

| Strain Number | Strain Name | Source | Accession | Invader (CFU) [α] | Resident (CFU) [β] | Full Rank [γ] | Partial Rank [γ] |
|---|---|---|---|---|---|---|---|
| 1a | *Amycolatopsis AA4* | (Seyedsayamdost et al. 2011) | PRJNA33599 | 132 | 3.21E+06 | 6 | 4.91 |
| 1b | | | | 139 | 1.23E+07 | | |
| 2 | *Streptomyces cattleya* | (Barbe et al. 2011) | FQ859185 | 141 | 5.64E+06 | 5 | 3.18 |
| 3 | *Streptomyces Aste7* | Owen Park, Madison, WI | KT364455 | 79 | 1.18E+05 | 5 | 3.91 |
| 4 | *Strepomyces S25E2* | See Methods | KT364442 | 131 | 1.04E+04 | 5 | 3.91 |
| 5 | *Strepomyces S3H4* | See Methods | KT364431 | 182 | 2.93E+06 | 7 | 4.91 |
| 6 | *Strepomyces S3H10* | See Methods | KT364432 | 54 | 1.21E+06 | 2 | 1.45 |
| 7 | *Strepomyces S26C11* | See Methods | KT364445 | 14 | 3.00E+07 | 5 | 3.91 |
| 8 | *Strepomyces S32E2* | See Methods | KT364451 | 157 | 8.21E+05 | 3 | 2.36 |
| 9 | *Strepomyces S3E8* | See Methods | KT364429 | 191 | 4.71E+04 | 7 | 5.91 |
| 10 | *Strepomyces S29H7* | See Methods | KT364448 | 235 | 8.44E+02 | 4 | 3.36 |
| 11 | *Strepomyces S25H8* | See Methods | KT364444 | 81 | 1.32E+05 | 6 | 4.91 |
| 12 | *Strepomyces S18A4* | See Methods | KT364439 | 151 | 3.39E+06 | 5 | 4.91 |
| 13 | *Streptomyces coelicolor (M145)* | (Bentley et al. 2002) | AL645882 | 122 | 1.29E+05 | 7 | 5.91 |
| 14 | *Streptomyces G3D7* | (Vetsigian, Jajoo, and Kishony 2011) | KT364454 | 67 | 4.68E+06 | 1 | 1.18 |
| 15 | *Strepomyces S26F9* | See Methods | KT364446 | 92 | 9.64E+04 | 7 | 5.91 |
| 16 | *Strepomyces S9E5* | See Methods | KT364435 | 297 | 9.29E+03 | 5 | 3.91 |
| 17 | *Strepomyces S4B4* | See Methods | KT364433 | 198 | 3.21E+05 | 7 | 4.91 |
| 18 | *Strepomyces S2D4* | See Methods | KT364427 | 169 | 2.89E+05 | 7 | 5.91 |

[α] Initial concentration as the *invader* strain in Colony Forming Units (CFU) per tube.

[β] Initial concentration as the *resident* strain in Colony Forming Units (CFU) per tube.

[γ] Hierarchical ranking of strains using the full invasion matrix (Full Rank) or only the pairs of strains without inhibition (Partial Rank). The Partial Rank column gives the average of the 11 possible rankings with equivalent scores (see Appendix A: Supplemental Fig. S7 for one example).

**Supplemental Table 2.** PCR primers used in this study. For sequencing, the rpoB_amp_F and rpoB_amp_R are the primers used in the first amplification step, and the longer primers are used in the second amplification step. Barcoded primers are named by index (BC) number and, for the reverse primers, a phase offset (+0 to +3 nucleotides) (Wu et al. 2015).

| Primer Name | Primer Sequence (5' to 3') |
|---|---|
| rpoB_amp_F | AAGGTCGGCCGCTACAAGGT |
| rpoB_amp_R | GATGTCGTCGGTCTCGAC |
| rpoB_amp_F1_BC1 | CAAGCAGAAGACGGCATACGAGATTAAGAGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC3 | CAAGCAGAAGACGGCATACGAGATGCGAATTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC4 | CAAGCAGAAGACGGCATACGAGATACTGAGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC5 | CAAGCAGAAGACGGCATACGAGATTTAGGCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC6 | CAAGCAGAAGACGGCATACGAGATCTCCGATTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC7 | CAAGCAGAAGACGGCATACGAGATGATGCGAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC8 | CAAGCAGAAGACGGCATACGAGATCATGGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC9 | CAAGCAGAAGACGGCATACGAGATCGTGATCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC10 | CAAGCAGAAGACGGCATACGAGATTGCATTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC11 | CAAGCAGAAGACGGCATACGAGATACGTTCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC12 | CAAGCAGAAGACGGCATACGAGATTTCACAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC13 | CAAGCAGAAGACGGCATACGAGATAATTGGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC14 | CAAGCAGAAGACGGCATACGAGATCCTTACGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC15 | CAAGCAGAAGACGGCATACGAGATCAGAGGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC16 | CAAGCAGAAGACGGCATACGAGATACTCGGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC17 | CAAGCAGAAGACGGCATACGAGATGAGTTCCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC18 | CAAGCAGAAGACGGCATACGAGATTCACGTTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC19 | CAAGCAGAAGACGGCATACGAGATTATGACCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC2 | CAAGCAGAAGACGGCATACGAGATATACGCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC20 | CAAGCAGAAGACGGCATACGAGATGGCATCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC21 | CAAGCAGAAGACGGCATACGAGATTCTGGACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC22 | CAAGCAGAAGACGGCATACGAGATTGAGCAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |

| Primer Name | Primer Sequence (5' to 3') |
|---|---|
| rpoB_amp_F1_BC23 | CAAGCAGAAGACGGCATACGAGATTGACCTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC24 | CAAGCAGAAGACGGCATACGAGATCATCTGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_F1_BC25 | CAAGCAGAAGACGGCATACGAGATCGGTCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| rpoB_amp_R1+0_BC1 | AATGATACGGCGACCACCGAGATCTACACGCCTCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+1_BC2 | AATGATACGGCGACCACCGAGATCTACACCAGCGTATACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+2_BC3 | AATGATACGGCGACCACCGAGATCTACACGAATTCGCACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+3_BC4 | AATGATACGGCGACCACCGAGATCTACACAGCTCAGTACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+0_BC5 | AATGATACGGCGACCACCGAGATCTACACGTGCCTAAACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+1_BC6 | AATGATACGGCGACCACCGAGATCTACACAATCGGAGACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+2_BC7 | AATGATACGGCGACCACCGAGATCTACACTTCGCATCACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+3_BC8 | AATGATACGGCGACCACCGAGATCTACACATGCCATGACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+0_BC9 | AATGATACGGCGACCACCGAGATCTACACTGATCACGACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+1_BC10 | AATGATACGGCGACCACCGAGATCTACACGGAATGCAACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+2_BC11 | AATGATACGGCGACCACCGAGATCTACACGAGAACGTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+3_BC12 | AATGATACGGCGACCACCGAGATCTACACCCTGTGAAACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+0_BC13 | AATGATACGGCGACCACCGAGATCTACACGGCCAATTACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+1_BC14 | AATGATACGGCGACCACCGAGATCTACACACGTAAGGACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+2_BC15 | AATGATACGGCGACCACCGAGATCTACACTACCTCTGACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+3_BC16 | AATGATACGGCGACCACCGAGATCTACACTACCGAGTACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+0_BC17 | AATGATACGGCGACCACCGAGATCTACACTGGAACTCACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+1_BC18 | AATGATACGGCGACCACCGAGATCTACACGAACGTGAACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+2_BC19 | AATGATACGGCGACCACCGAGATCTACACCGGTCATAACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+3_BC20 | AATGATACGGCGACCACCGAGATCTACACATGATGCCACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+0_BC21 | AATGATACGGCGACCACCGAGATCTACACTGTCCAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+1_BC22 | AATGATACGGCGACCACCGAGATCTACACCTTGCTCAACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+2_BC23 | AATGATACGGCGACCACCGAGATCTACACTCAGGTCAACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+3_BC24 | AATGATACGGCGACCACCGAGATCTACACTCCAGATGACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| rpoB_amp_R1+0_BC25 | AATGATACGGCGACCACCGAGATCTACACTTAGACCGACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |

| **Primer Name** | **Primer Sequence (5' to 3')** |
|---|---|
| rpoB_strain2_F1 | GGCCGGTCTGGACGTG |
| rpoB_strain2_R1 | GCGGTCAGGTAGTGCACGT |
| rpoB_strain9_F1 | GGTCAGGTCACCGACGAC |
| rpoB_strain9_R1 | GTGAAGCGCATGTCGTCATTG |
| rpoB_strain11_F1 | GGGCCGCGAGATCATCG |
| rpoB_strain11_R1 | CGACGCCACCACGGG |
| rpoB_strain15_F1 | TCATCGACGGCGTCGTCA |
| rpoB_strain15_R1 | CATGTCCTCGGACAGGGC |
| rpoB_strain17_F1 | CCGTCTCTCGGCGCTC |
| rpoB_strain17_R1 | GCTCGTCGTTCAGGGTCG |
| rpoB_strain18_F1 | CGGCGAGAACGGCAACGAG |
| rpoB_strain18_R1 | CCGGATGTTGATCAGGGTCTGA |

**Supplemental Table 3.** Results of quantitative PCR amplification of a subset of 8 bistable pairs after the third growth cycle.

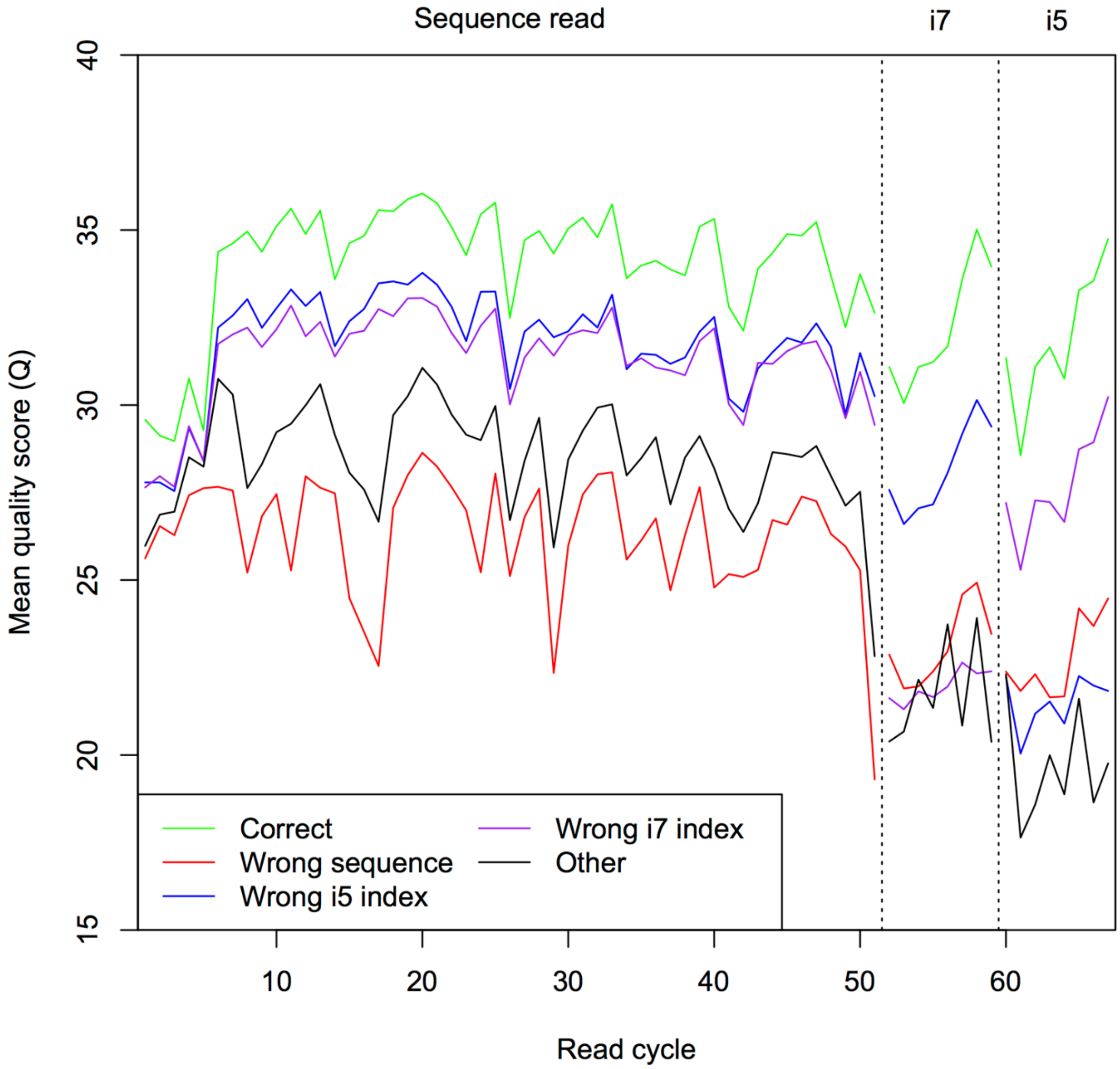
| Bistable pairs | | Threshold cycle ($C_t$) | | | | | |
|---|---|---|---|---|---|---|---|
| Resident | Invader | Resident $C_t$ | Invader $C_t$ | $\Delta C_t$ | Melt curve [α] | Gel run [β] | Sanger Sequencing [γ] |
| 9 | 17 | 17.48 | 36.49 | 19.01 | Different | Matched | Artifact |
| 17 | 9 | 15.14 | 29.48 | 14.34 | Different | Different | N/A |
| 15 | 18 | 15.90 | 34.59 | 18.69 | Different | Different | N/A |
| 18 | 15 | 14.09 | 34.92 | 20.83 | Different | Different | N/A |
| 15 | 9 | 17.03 | 31.35 | 14.32 | Different | Absent | N/A |
| 9 | 15 | 17.01 | 37.23 | 20.22 | Different | Matched | Artifact |
| 2 | 17 | 18.65 | 39.38 | 20.73 | Different | Different | N/A |
| 17 | 2 | 16.30 | 22.04 | 5.74 | Different | Different | N/A |
| 2 | 11 | 24.08 | 40.14 | 16.06 | Different | Different | N/A |
| 11 | 2 | 17.93 | 26.75 | 8.82 | Different | Different | N/A |
| 15 | 17 | 16.09 | 44.29 | 28.20 | Different | Absent | N/A |
| 17 | 15 | 17.72 | 36.88 | 19.16 | Different | Absent | N/A |
| 17 | 18 | 16.73 | 34.15 | 17.42 | Different | Absent | N/A |
| 18 | 17 | 14.57 | 36.63 | 22.06 | Different | Absent | N/A |
| 2 | 18 | 16.31 | 37.60 | 21.29 | Different | Absent | N/A |
| 18 | 2 | 15.21 | 26.97 | 11.76 | Different | Absent | N/A |

[α] Whether the shape of the melt curve matched or was different than would be expected if the invader's target DNA had amplified.

[β] Whether the amplicon length matched or was different than would be expected if the invader's target DNA had amplified. "Absent" indicates that the gel run band was too short to appear on the gel (< ~70 base pairs).

[γ] Whether the results of Sanger sequencing matched the invader's target DNA, appeared to be a PCR artifact (Artifact), or was not sequenced (N/A).

## APPENDIX B: SUPPLEMENTAL MATERIAL FOR CHAPTER 3



**Supplemental Figure 1.** Mean quality score per base for each read step (sequence, i7 index, or i5 index). The average quality was consistently lower across the entire length of the misassigned reads relative to correctly assigned reads. Furthermore, particular positions exhibited consistently lower scores across all read types, as well as across the 14 sequence variants.

**Supplemental Table 1**

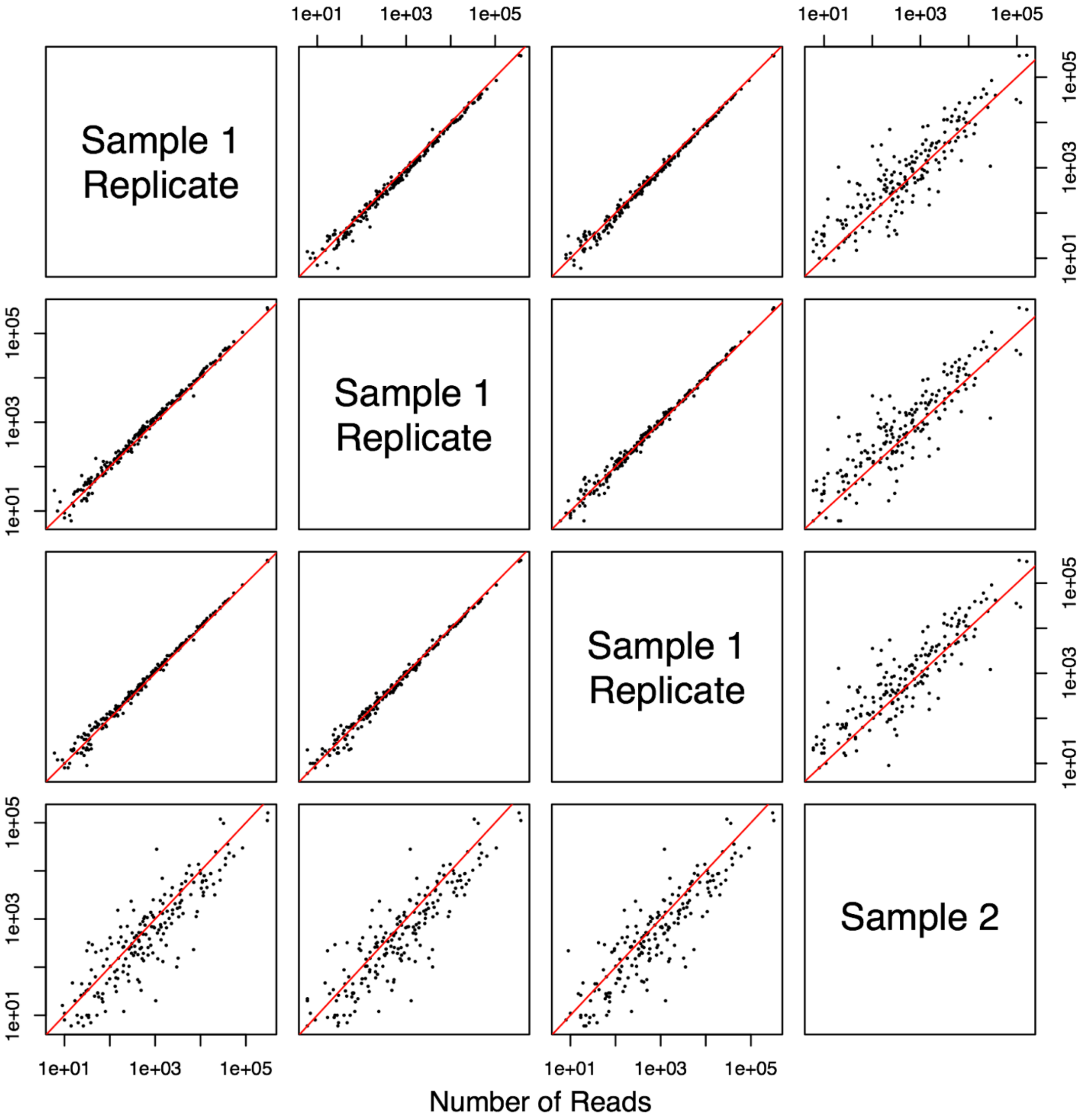
| Target | Strain | Primer 1 | Primer 2 | i5 index | i7 index |
|---|---|---|---|---|---|
| barcode | *Streptomyces lividans* | Left54 | Right28 | CAACGAAC | AGACGTTC |
| barcode | *Streptomyces S3H10* | Left55 | Right80 | CACACACT | ATGGTGTG |
| barcode | *Streptomyces coelicolor* | Left56 | Right66 | AATACCGC | ACCTACCA |
| barcode | *Streptomyces venezuelae* | Left41 | Right67 | AATGACGG | ACTTCGGT |
| barcode | *Streptomyces S26F9* | Left42 | Right29 | TGACGGAA | ACGGACTT |
| barcode | *Streptomyces albus J1074* | Left43 | Right81 | ACATGGCT | GCTGAACT |
| barcode | *Streptomyces G4A3* | Left78 | Right68 | CGGCTATT | CTACGCTA |
| barcode | *Streptomyces S25E2* | Left79 | Right30 | ATAGCGGT | CTAAGCGT |
| barcode | *Streptomyces S4B4* | Left57 | Right31 | CTGTTCGT | GATGTCCA |
| barcode | *Streptomyces S2D4* | Left44 | Right69 | CCGAATTG | AACACGAC |
| barcode | *Streptomyces S18A4* | Left45 | Right70 | TACTAGCG | GCGAGATT |
| *rpoB* | *Streptomyces S4B4* | Reverse3 | Forward1 | GAATTCGC | GCCTCTTA |
| *rpoB* | *Streptomyces cattelya* | Reverse25 | Forward2 | TTAGACCG | CAGCGTAT |
| *rpoB* | *Amycolatopsis AA4* | Reverse16 | Forward15 | TACCGAGT | TACCTCTG |

**Supplemental Table 2**

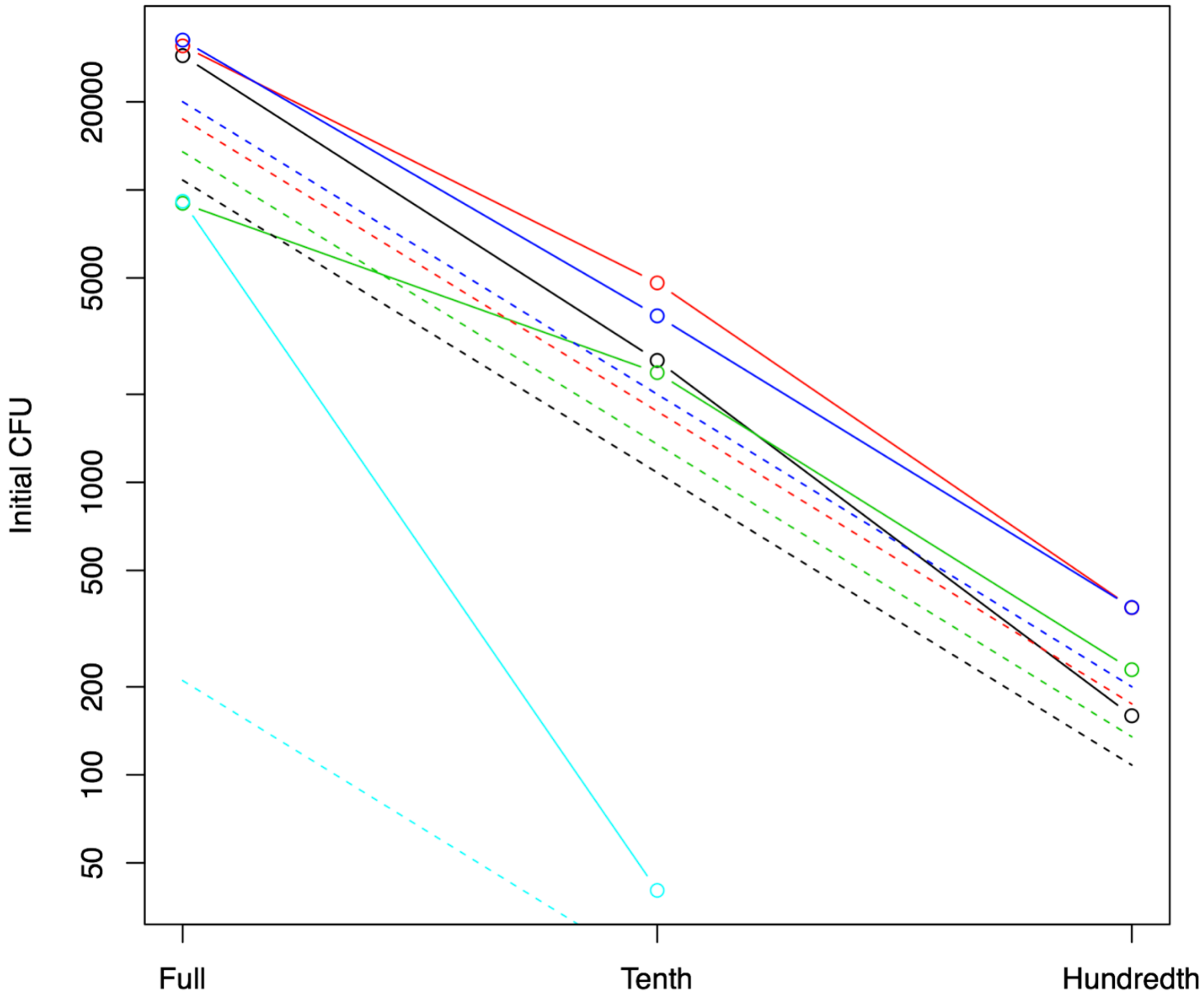
| Primer Name | Primer Sequence (5' to 3') |
|---|---|
| Forward1 | CAAGCAGAAGACGGCATACGAGATTAAGAGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| Forward15 | CAAGCAGAAGACGGCATACGAGATCAGAGGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| Forward2 | CAAGCAGAAGACGGCATACGAGATATACGCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGGTCGGCCGCTACAAGGT |
| Left41 | AATGATACGGCGACCACCGAGATCTACACAATGACGGACACTCTTTCCCTACACGACG |
| Left42 | AATGATACGGCGACCACCGAGATCTACACTGACGGAAACACTCTTTCCCTACACGACG |
| Left43 | AATGATACGGCGACCACCGAGATCTACACACATGGCTACACTCTTTCCCTACACGACG |
| Left44 | AATGATACGGCGACCACCGAGATCTACACCCGAATTGACACTCTTTCCCTACACGACG |
| Left45 | AATGATACGGCGACCACCGAGATCTACACTACTAGCGACACTCTTTCCCTACACGACG |
| Left54 | AATGATACGGCGACCACCGAGATCTACACCAACGAACACACTCTTTCCCTACACGACG |
| Left55 | AATGATACGGCGACCACCGAGATCTACACCACACACTACACTCTTTCCCTACACGACG |
| Left56 | AATGATACGGCGACCACCGAGATCTACACAATACCGCACACTCTTTCCCTACACGACG |
| Left57 | AATGATACGGCGACCACCGAGATCTACACCTGTTCGTACACTCTTTCCCTACACGACG |
| Left78 | AATGATACGGCGACCACCGAGATCTACACCGGCTATTACACTCTTTCCCTACACGACG |
| Left79 | AATGATACGGCGACCACCGAGATCTACACATAGCGGTACACTCTTTCCCTACACGACG |
| Reverse16 | AATGATACGGCGACCACCGAGATCTACACTACCGAGTACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGATGTCGTCGGTCTCGAC |
| Reverse25 | AATGATACGGCGACCACCGAGATCTACACTTAGACCGACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGTCGTCGGTCTCGAC |
| Reverse3 | AATGATACGGCGACCACCGAGATCTACACGAATTCGCACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATGTCGTCGGTCTCGAC |
| Right28 | CAAGCAGAAGACGGCATACGAGATGAACGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right29 | CAAGCAGAAGACGGCATACGAGATAAGTCCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right30 | CAAGCAGAAGACGGCATACGAGATACGCTTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right31 | CAAGCAGAAGACGGCATACGAGATTGGACATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right66 | CAAGCAGAAGACGGCATACGAGATTGGTAGGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right67 | CAAGCAGAAGACGGCATACGAGATACCGAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right68 | CAAGCAGAAGACGGCATACGAGATTAGCGTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |

| Primer Name | Primer Sequence (5' to 3') |
|---|---|
| Right69 | CAAGCAGAAGACGGCATACGAGATGTCGTGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right70 | CAAGCAGAAGACGGCATACGAGATAATCTCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right80 | CAAGCAGAAGACGGCATACGAGATCACACCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right81 | CAAGCAGAAGACGGCATACGAGATAGTTCAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |

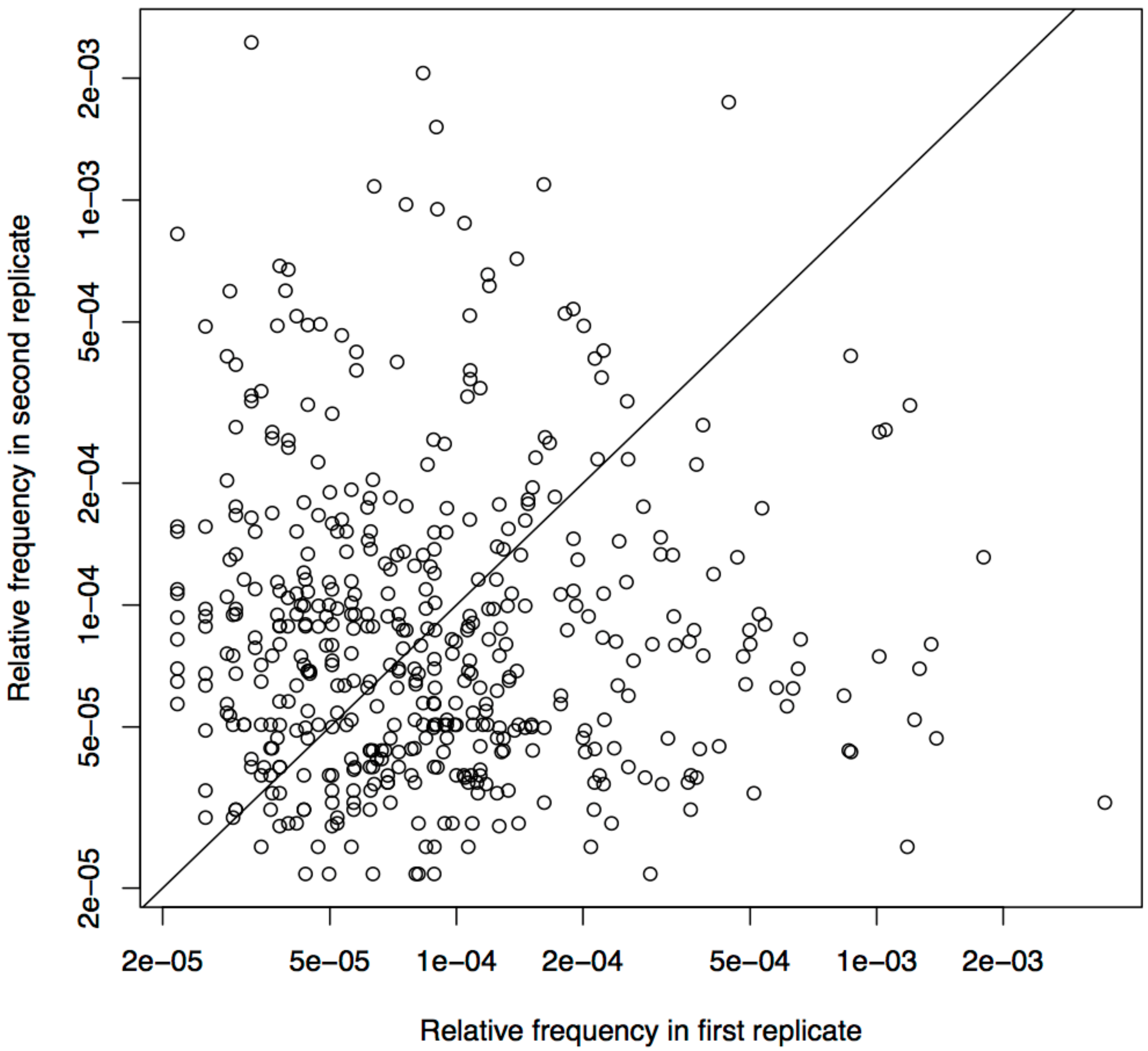
## APPENDIX C: SUPPLEMENTAL MATERIAL FOR CHAPTER 4



**Supplemental Figure 1. Most of the variability between replicates is biological in nature.** Three technical (separate PCR and sequencing) replicates of the same biological sample are plotted against each other and a different biological sample of *S. coelicolor*. Each point corresponds to a unique barcode that was present in both samples. Correlation between technical replicates was much higher than that for biological replicates. The line of identity is colored red.



**Supplemental Figure 2. Fitted values of the initial census population size (*N*) roughly corresponded to expectation.** For each of the 5 species, values of *N* determined by fitting the number of barcodes during simulations (with an underlying Pareto descendants distribution) are shown as points connected by solid lines. Expected values based on dilution plating of each strain are shown by the correspondingly colored dashed lines. Note that both axes are log-scaled.



**Supplemental Figure 3. The frequency of rare barcodes was largely uncorrelated between biological replicates.** The relative frequencies of barcodes appearing in only 2 of 8 replicates are shown for strain *S. S4G3*. The lack of correlation between replicate barcodes indicates that inter-barcode selection had a negligible influence over the variability between replicates. Note the log-scaled axes and the line of identity.

**Supplemental Table 1.** PCR primers used in this study.

| **Primer Name** | **Primer Sequence (5' to 3')** |
|---|---|
| Left41 | AATGATACGGCGACCACCGAGATCTACACAATGACGGACACTCTTTCCCTACACGACG |
| Left42 | AATGATACGGCGACCACCGAGATCTACACTGACGGAAACACTCTTTCCCTACACGACG |
| Left43 | AATGATACGGCGACCACCGAGATCTACACACATGGCTACACTCTTTCCCTACACGACG |
| Left44 | AATGATACGGCGACCACCGAGATCTACACCCGAATTGACACTCTTTCCCTACACGACG |
| Left45 | AATGATACGGCGACCACCGAGATCTACACTACTAGCGACACTCTTTCCCTACACGACG |
| Left46 | AATGATACGGCGACCACCGAGATCTACACAGGCATCTACACTCTTTCCCTACACGACG |
| Left47 | AATGATACGGCGACCACCGAGATCTACACAAGGTAGCACACTCTTTCCCTACACGACG |
| Left48 | AATGATACGGCGACCACCGAGATCTACACTCCATCGTACACTCTTTCCCTACACGACG |
| Left49 | AATGATACGGCGACCACCGAGATCTACACCATTCCGTACACTCTTTCCCTACACGACG |
| Left50 | AATGATACGGCGACCACCGAGATCTACACGTGAGACAACACTCTTTCCCTACACGACG |
| Left51 | AATGATACGGCGACCACCGAGATCTACACCGCTTAAGACACTCTTTCCCTACACGACG |
| Left52 | AATGATACGGCGACCACCGAGATCTACACCTCGAGTAACACTCTTTCCCTACACGACG |
| Left54 | AATGATACGGCGACCACCGAGATCTACACCAACGAACACACTCTTTCCCTACACGACG |
| Left55 | AATGATACGGCGACCACCGAGATCTACACCACACACTACACTCTTTCCCTACACGACG |
| Left56 | AATGATACGGCGACCACCGAGATCTACACAATACCGCACACTCTTTCCCTACACGACG |
| Left57 | AATGATACGGCGACCACCGAGATCTACACCTGTTCGTACACTCTTTCCCTACACGACG |
| Left58 | AATGATACGGCGACCACCGAGATCTACACGTCGTTGTACACTCTTTCCCTACACGACG |
| Left59 | AATGATACGGCGACCACCGAGATCTACACCACAGGAAACACTCTTTCCCTACACGACG |
| Left60 | AATGATACGGCGACCACCGAGATCTACACTAAGCCAGACACTCTTTCCCTACACGACG |
| Left61 | AATGATACGGCGACCACCGAGATCTACACGCATAGGTACACTCTTTCCCTACACGACG |
| Left62 | AATGATACGGCGACCACCGAGATCTACACATGCGTAGACACTCTTTCCCTACACGACG |
| Left63 | AATGATACGGCGACCACCGAGATCTACACCTTGTTGCACACTCTTTCCCTACACGACG |
| Left64 | AATGATACGGCGACCACCGAGATCTACACAAGTGGTGACACTCTTTCCCTACACGACG |
| Left65 | AATGATACGGCGACCACCGAGATCTACACTTGCTACGACACTCTTTCCCTACACGACG |
| Left78 | AATGATACGGCGACCACCGAGATCTACACCGGCTATTACACTCTTTCCCTACACGACG |

| Primer Name | Primer Sequence (5' to 3') |
|---|---|
| Left79 | AATGATACGGCGACCACCGAGATCTACACATAGCGGTACACTCTTTCCCTACACGACG |
| Right28 | CAAGCAGAAGACGGCATACGAGATGAACGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right29 | CAAGCAGAAGACGGCATACGAGATAAGTCCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right30 | CAAGCAGAAGACGGCATACGAGATACGCTTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right31 | CAAGCAGAAGACGGCATACGAGATTGGACATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right32 | CAAGCAGAAGACGGCATACGAGATCCAAGTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right33 | CAAGCAGAAGACGGCATACGAGATGGTAACGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right34 | CAAGCAGAAGACGGCATACGAGATTCGGTAACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right35 | CAAGCAGAAGACGGCATACGAGATGTCGCAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right36 | CAAGCAGAAGACGGCATACGAGATTCCATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right37 | CAAGCAGAAGACGGCATACGAGATGCCTTGAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right38 | CAAGCAGAAGACGGCATACGAGATGATCATGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right39 | CAAGCAGAAGACGGCATACGAGATATTGCCAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right66 | CAAGCAGAAGACGGCATACGAGATTGGTAGGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right67 | CAAGCAGAAGACGGCATACGAGATACCGAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right68 | CAAGCAGAAGACGGCATACGAGATTAGCGTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right69 | CAAGCAGAAGACGGCATACGAGATGTCGTGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right70 | CAAGCAGAAGACGGCATACGAGATAATCTCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right71 | CAAGCAGAAGACGGCATACGAGATACCTTGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right72 | CAAGCAGAAGACGGCATACGAGATAACCGATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right73 | CAAGCAGAAGACGGCATACGAGATCATAACGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right74 | CAAGCAGAAGACGGCATACGAGATCAACCACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right75 | CAAGCAGAAGACGGCATACGAGATCGATTGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right76 | CAAGCAGAAGACGGCATACGAGATGTTCTGGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right77 | CAAGCAGAAGACGGCATACGAGATTGAAGGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right80 | CAAGCAGAAGACGGCATACGAGATCACACCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |
| Right81 | CAAGCAGAAGACGGCATACGAGATAGTTCAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCGATTAAGTTGGGTAACG |