Phage-Host Interaction in Nature

Thesis by

Arbel D. Tadmor

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California
2011
(Defended May 17, 2011)

Acknowledgments

To many people I owe my gratitude for making my stay at Caltech both fascinating as well as incredibly fun. First and foremost I would like to thank my advisor, Prof. Rob Phillips, who taught me how to be the most critical scientist I can be; Prof. Jared Leadbetter, from whom I learned about microbial diversity and the art of phylogenetics and who graciously agreed to head my committee, and my two other committee members, Prof. David Baltimore and Prof. Victoria J. Orphan, for kindly agreeing to be on my committee and whose questions and critique were critical to the success of my research. I wish to also thank Dr. Eric G. Matson and Dr. Elizabeth A. Ottesen for kindly answering my many questions, and Eric and Dr. Adam Z. Rosenthal for going termite collecting with me on several occasions. I would like to also thank Dr. Blake W. Axelrod, Prof. David Bensimon, Dr. Bertrand Ducos, Prof. Michael L. Roukes and Lijun Xu for fruitful and fascinating collaborations, and Prof. Grant J. Jensen for allowing me access to his imaging facility working closely with Dr. Alasdair McDowall and Dr. Bill Tivol. I am also indebted to the many people who kindly agreed to review our work along the way, including Prof. Sherwood Casjens, Prof. Daniel S. Fisher, Prof. Roger W. Hendrix, Prof. Ron Milo, Prof. Stephen R. Quake, Prof. Edward M. Rubin and Prof. Nathan D. Wolfe. I would also like to thank past and present members of the Phillips group and the Leadbetter group, and especially Dr. Heun Jin Lee, Dr. Martin Lindén, Damien Soghoian, Dr. David Wu and Dr. David Van Valen for the many stimulating conversations and valuable feedback I received from them over the years. Finally I wish to thank Dr. Paul Grayson for introducing me to the world of phages.

Abstract

Though viruses may be the most abundant biological entities on the planet, very little is known about phage-host interaction in the wild due to the absence of proper experimental tools. In the present work we report of a method to pair environmental phages with their bacterial hosts at the single-cell level without having to culture either host or virus. The method utilizes microfluidic digital PCR in conjunction with a metagenome data mining tool that was developed to find a viral marker gene in an unknown environment. We implemented this technique on the microbial community residing in the hindgut of termites. Consequently, we discovered genus-wide infection patterns displaying remarkable intra-genus selectivity, with viral alleles displaying limited lateral gene transfer and/or host switching despite host proximity. To try and explain phage-host interactions from a theoretical perspective, we formulated a simple biophysical model describing the interaction of bacteria and viruses in aqueous environments. We predict that the radius r of a bacterium is the most critical parameter determining its fixed point concentration, which scales as r^{-4}. Given the hypothesis that there is no selection pressure on bacterial radii, our model predicts that the size spectrum of marine bacteria follows a power law with slope -1, close to the observed average spectrum. Moreover, given the total concentration of bacteria in the ocean, our model enables us to estimate the total number of bacterial "species" per volume of water providing a lower and upper bound on the total number of species in the oceans. To elucidate the concept of a "species", we consider a bacterial-viral co-speciation model, which is consistent with the observed narrow host range of phages. Our model hints that the bacterial-viral "arms race" may be a critical component in the process of co-speciation. We suggest further experiments to test both models. Finally, we consider a recent high resolution measurement of the force as a function of time generated by stress fibers within a single fibroblast cell and suggest a stochastic model that is capable of accounting for the observed data.

Table of Contents

Chapter 1 Introduction

Chapter 2 Probing Individual Environmental Bacteria for Viruses Using Microfluidic Digital PCR

Chapter 3 MetaCAT—Metagenome Cluster Analysis Tool

Chapter 4 The Biophysics of Prokaryotic and Viral Diversity in Aqueous Environments

Chapter 5 An evolutionary model of phage-host interaction

Chapter 6 A Kinetic Model for Stress Fiber Contraction and Relaxation

List of Figures

List of Tables

Chapter 1

Introduction

1.0 Preface

In the following introduction we begin with a brief overview of some basic facts known about phages and their interaction with bacterial hosts. Our purpose is not to exhaustively review the topic, but to introduce certain concepts that will be useful for the remaining chapters, especially Chapters 2, 4, and 5. We then highlight for each chapter the most interesting or promising findings. The remaining chapters of the thesis are organized as follows: Chapter 2 describes an experiment to co-localize phages with their hosts directly from the environment using single cell microfluidic technology. Chapter 3 describes a bioinformatic tool for metagenome analysis that was used in Chapter 2 to identify the must abundant viral genes in the metagenome of a higher Costa Rican termite. Chapter 4 analyzes the problem of phage-host interaction from a theoretical perspective. We first consider a biophysical model describing phage-host interaction of a single isolated phage-host system. We then make the leap to a distribution of phage-host systems in the environment, allowing us to calculate, for example, bounds on the total diversity in the ocean water column. Then, in Chapter 5, we consider the beginnings of an evolutionary model for phage-host co-speciation that we believe has much potential. The key feature of this model is that it is consistent with a “world” where phages have evolved to have a narrow host range. Presently, this model suggests how bacterial “species” and viral “species” are related (thus defining both terms), and hints that the arms race that bacteria and viruses are locked in is perhaps the engine driving bacterial and viral co-speciation, and thus perhaps bacterial evolution itself (a hypothesis). Finally in Chapter 6 we present an analysis of experimental data collected by Dr. Blake W. Axelrod, a research engineer in the Roukes lab, who measured with the highest resolution to date the force as a function of time of a stress fiber in a single fibroblast cell as this stress fiber is artificially disrupted and then allowed to naturally reassemble. Blake observed quantized steps in the force exhibiting exponential like temporal profiles that we can explain by a simple stochastic model, where each sarcomere perfectly obeys a law of exponentially distributed time delays.

1.1 Some facts phages in nature

1.1.1 Abundance and activity

Viruses may very well be the most abundant biological entities on the planet. In offshore surface waters viral concentrations are typically in the range of 10^5-10^6 ml^{-1}, whereas in coastal environments, viral concentrations can reach 10^6-10^7 ml^{-1} [1]. High viral concentrations were also found in sea ice (10^7-10^8 ml^{-1} [1]), marine sediments (10^7-10^{10} g^{-1}) [1,2], in soil (~10^8 g^{-1}) [1] as well as in the rumen gut (10^8-10^{10} ml^{-1}) [3,4]. Viral concentrations are typically correlated with bacterial concentrations. A variable often used by environmental virologists to gauge this correlation is the virus to bacterium ratio (VBR). The VBR for marine systems is consistently measured to be on the order of 10 [1,5,6,7], making viruses the most abundant life-forms in the oceans. The VBR can also reach as high as ~70 in sea ice [1] or could be as low as 0.04 in soil [1]. A VBR of ~6 is also observed when zooming in on a particular phage-host system. For example, *Synechococcus* cells from the Gulf of Mexico have been shown to be infected by an average of ~6 viral-like particles per bacterium [8].

Virions are also extremely active in the environment. It has been estimated that ~20% of all marine microbial cells, which constitute over 90% of the viable biomass in the Earth's oceans, are turned over daily by viral predation [6]. In the deep-sea, viruses are thought to be responsible for at least 80% of prokaryote mortality (calculated by taking the ratio of viral production and prokaryotic burst size) [9]. The same is true for low-oxygen lake waters (in which grazers do not thrive) where viruses are thought to be responsible for 50-100% of bacterial mortality [5]. Such high viral-induced mortality suggests that many if not most bacteria die from viral infection. For example, in environments where viral lysis accounts for 50-100% of the bacterial mortality, either every bacterial cell or every second cell will be lysed by a virus in order to maintain a steady state population of bacteria.

In terms of their life expectancy in the wild, marine viruses can survive only about 1 to 10 days without having to "feed". Viral decay rates of ~0.1 to ~2 day^{-1} have been measured for inshore and offshore regions, respectively [8], with comparable decay rates in deep-sea sediments [2].

1.1.2 Lytic or lysogenic?

Viruses are also very abundant in the form of lysogenic viruses, with an estimated ~60% of sequenced bacterial genomes encoding at least one integrated viral element [10,11]. However, one might expect that with viruses being so abundant in nature and having such a major contribution to bacterial mortality, that the observed viral-like particles are lytic viruses. If these viruses were lysogenic, they would probably need to be continuously induced in order to reach the observed levels of abundance, obviating the need to integrate or to encode a genetic switch. Indeed, growth experiments with native bacterial communities in freshly filtered sea water indicated that under typical natural conditions, induction of lysogens was rare, with the vast majority of the observed viruses probably the result of successive lytic infection [5,12]. Furthermore, attempts to induce lysogens with bright continuous sunlight or pulsed sunlight did not result in increased viral concentrations [5,12] suggesting that lysogens are not easily inducible. Therefore, it appears that lysogenic induction may be occurring at low levels either continuously or sporadically [5] (possibly occasionally on larger scales [1]), with the vast majority of viruses in the sea probably the result of lytic infection [5].

1.2 Phage-host interaction

1.2.1 Predator-prey dynamics

Phages have effect on bacteria in many different levels and vice versa. Our intention here is not to give a comprehensive review of all mechanism of interaction between phages and bacteria, but to highlight a few important concepts used in later chapters. The most basic level that viruses affect bacteria is through concentration control. In a classic case of a predator-prey dynamical system (one predator-one prey), the fixed point concentration of the prey is determined by the predator. Therefore the fixed point concentration of the prey does not depend on its growth rate. As long as there is positive growth of the prey, its final concentration will be the same. Therefore, if a bacterial species has a very low growth rate, the concentration of viruses infecting it will be low. Conversely, if the bacterium grows very fast, the concentration of the viruses infecting it will be very high.

1.2.2 Population control versus species control

It is generally accepted that bacterial host mortality is primarily due to either grazing by protists or lysis by viruses [5,13,14,15]. The fundamental difference between these two predators is that protists, to a first-order approximation, are omnivores, i.e., they are not host specific [15,16], while phages display a species- or strain-level host range [1,17,18]. Protist therefore either control the *total* bacterial concentration (sum of all species), or — if they are themselves prey — do not exert control over bacteria [15] and simply reduce the bacterial production rate (with bacterial concentration being determined by competition for nutrients [15]). Viruses on the other hand exert control at the species level. Therefore, through predator-prey dynamics, viruses directly control the genetic diversity of bacteria in the environment.

1.2.3 Kill the winner hypothesis

In nature, every environment contains many species of bacteria. Given the narrow host range of phages, to a first-order approximation, we can think of this environment as comprised of a collection of non-interacting phage-hosts systems[1]. Given the individual predator-prey dynamics, based on our explanation above, we expect the concentration of each bacterial species to be controlled separately and be independent of the growth rate of the bacteria. By having viruses control the population in this way, fast growing cells will not be "allowed" to take over the population. If a bacterium's growth rate increases, the concentration of the viruses infecting it will also increase, thus keeping the (fixed point) concentration of that bacterium in check. Thus the equilibrium diversity in these networks is maintained by mechanisms that are selectively ''killing the winner.'' (i.e., a superior competitor) [16,19,20].

1.2.4 The bacterial-viral "arms race"

Recently it has been discovered that bacteria have a primitive immune system in the form of CRISPRs — clustered regularly interspaced short palindromic repeats — arrays found in nearly half of all sequenced bacterial genomes [21]. Short (26–72bps [21]) "spacer" sequences derived

[1] In Chapter 4 we show that the situation of more than one virus controlling the same bacterial species leads to extinction events.

from viral genes, present between the CRISPRs, are transcribed and interfere with viral gene expression in a mechanism thought to be similar to RNA silencing [21,22]. Bacteria continuously acquire CRISPR spacer sequences from viruses to evade these viruses. To evade new acquired spacers, the viruses rapidly evolve their genes though mutation, homologous recombination and deletion [23]. Conversely, CRISPR repeats and their associated proteins undergo evolution to escape a shut-down mechanism for the CRISPR system encoded by the phage [21]. Thus, bacteria and viruses are locked in an arms race [21]. This arms race may have long term evolutionary consequences on the bacterial population. From inspection of the history of spacers stored on the bacterial genomes of many individuals in a population it has been observed that all individuals can have essentially the same older spacers, with the new diverse set of spacers at the tip of the array, where new spacers are added [23]. One explanation for this observation could be a recent strong selection event caused by an unusually virulent virus to which potentially only one cell in the population was immune [23].

1.3 A coarse-grained view of phage-host interaction

1.3.1 The biophysics of a single phage-host system

Two perspectives: biophysical versus dynamical

From a dynamical point of view phage-host systems can be analyzed as a classic predator-prey problem. This type of problem has been studied extensively and is considered a textbook problem. From a biophysical perspective, the problem of phage-host interaction is that of viral transport. While this problem seems difficult to address in environments like soil or sediment, in aqueous environments the problem of viral transport can be reduced to solving the diffusion equation. This intuition did not escaped biophysicists who worked with viruses in the early days, and the first solution to this problem appeared in the book of Stent "Molecular Biology of

Bacterial Viruses" in 1963. Although both perspectives are known and are made use of, we have not seen in the literature an attempt to merge these two perspectives in one package, obtaining expressions for the concentrations of bacteria and viruses in terms of basic biophysical parameters such as temperature, viscosity, radii, and so on. We have also not seen any model attempting to exploit the empirical correlation between burst size and the volume of the bacterium and the inverse correlation between burst size and the volume of the virion (with empirical correlations measured up to 1 μm) [1,24,25]. Such correlations can have great implications on the scaling laws of these systems, and may be critical when attempting to draw conclusions on an entire community.

Combining the two perspectives leads to new insight

In Sections 4.1–4.3 we construct a new biophysical model describing the interaction of a single isolated phage-host system and obtain interesting scaling laws for the steady-state concentration of bacteria and viruses. We find that the most critical parameter determining the fixed point concentration of a phage-host pair in the environment is the radius of the bacterium (Fig. 1.1). We found that the fixed point (i.e., steady state) concentration of bacteria scales as r^{-4} with r being the radius of the bacterium. Since in nature, the radii of bacteria vary by over three orders of magnitude, our model predicts that the concentration of bacteria can change by over 13 orders of magnitude! Furthermore, our model predicts that large bacteria will be exceedingly rare, with the largest known bacterium (*Thiomargarita namibiensis* having a diameter of 750 μm) predicted to have one cell in ~10^3 liters of water. On the other hand, we predict that the concentration of the viruses infecting large bacteria will be high enough so that, using molecular techniques, these viruses should be detectable in one ml of water.

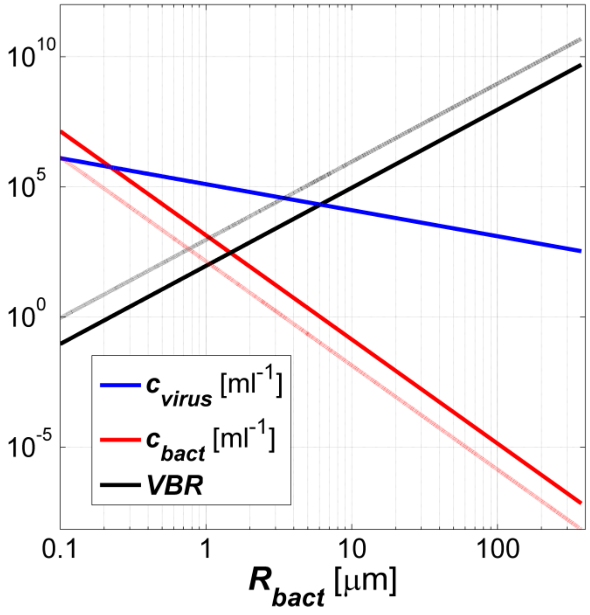


Figure 1.1. Scaling of the virus concentration, the bacterium concentration and the VBR with the radius of the bacterium for a single phage-host system. This figure shows that the radius of a bacterium is a critical parameter determining the fixed point concentration of the bacterium.

1.3.2 The biophysics of many phage-host systems

How many is many?

Thus far we have dealt with an artificial problem of a single phage-host system. In nature there are many such systems and in Section 4.4 we deal with the question of how to make the transition from a single phage-host system, to many such systems in the environment. How many is many? In the Venter expedition to the Sargasso Sea, every sample containing several hundreds of liters of ocean water was found to have at least 300 bacterial species (using a cutoff equivalent to a small subunit rRNA cutoff of 3%) [26]. Therefore, we expect that a natural environment will contain at least hundreds (probably thousands) of phage-host systems. Since the host range of phages is narrow, these phage-host systems can be treated, to a first-order approximation, as independent. Our realization from Section 1.3.1 that bacterial radii span such a wide range of values, and that the fixed point concentration of bacteria is extremely sensitive to this parameter, suggested to us that one cannot simply replace this parameter with an average value. One actually needs to calculate this average using the probability density function of bacterial radii for the given environment.

A simple evolutionary scenario

The difficulty in making the transition between a single phage-host system and many phage-host systems in the environment, is figuring out what is the *a priori* probability density of radii in a given environment, i.e., the probability per radius that a given environment *a priori* would contain a bacterial species with this radius. This function (which we denoted by $f_R(r)$) has evolutionary significance and can be interpreted as the density of bacterial species, perhaps analogous to the density of states in statistical mechanics, and reflects the evolutionary history of bacteria in the given environment. If the radii of all bacterial species that have adapted to survive in the given environment were known, one could calculate this function. Since we cannot calculate this function from first principles, we considered the simplest evolutionary scenario where this function is a constant, which means there is no selection pressure on bacterial radii, i.e., all radii are *a priori* equally probable. Given this assumption we were able to obtain expressions for basic quantities, such as, the total concentration of bacteria in a given environment, the total concentration of viruses in a given environment, the VBR, and the total bacterial biomass in a given environment. These results are especially interesting given that they are calculated from basic physical parameters such as temperature, viscosity, radii, and so on.

The size spectra of bacteria in the ocean

One additional quantity that we can calculate given $f_R(r)$ is the distribution of radii in a given environment, and from this function we can easily calculate the probability that a bacterium of random volume *V*, is greater than or equal to a given volume, *v,* or $\text{Prob}(V \geq v)$. This function is called the size spectra of radii and has been of interest to marine biologists for decades, with measurements dating back to the work of Sheldon in 1972 [27]. In 2001 the Chisholm lab from MIT measured the size spectra of microbes in the western north Atlantic Ocean. They found that the size spectra obeyed a power law with a slope between -1 and -1.4. The ensemble average of all environments was well described by a power law of slope -1.2. When expanding their dataset to include microzooplankton the slope was corrected to a value close to -1. Our calculation, given our simple evolutionary scenario, predicted a power law with a slope of -1, hopefully indicating that we are on the right track.

Species richness

In section 1.2.2 we mentioned that the total concentration of bacteria is determined either by the protists or by the availability of nutrients [15]. Thus, given the total concentration of bacteria in an environment, one can in fact calculate the number of predicted species. By considering two extreme models of spatial distribution of diversity — complete homogeneity and maximal heterogeneity — we were able to calculate bounds on the total diversity in the ocean water column. In Section 4.4 we also explore how the number of species scales with basic parameters and found that, quite intuitively, warm, nutrient-rich environments where viruses have a long lifetime will sustain the greatest diversity of species. Finally we compared estimates of diversity with observations from metagenome studies.

What is a species?

When applying our model to data we realized that there is something strange about our biophysical model. No where did we define what a “species” is! How different do two genomes (bacterial or viral) need to be in order to be considered different “species”? It is this question that we tried to address in Chapter 5.

1.4 The evolutionary perspective

1.4.1 A model for co-speciation of viruses and bacteria

In order to answer what a bacterial or viral species are, one needs to go to a higher theory that takes into account the genetic aspect of these entities, and not just parameterize them with a radius, decay rate, and so on. We therefore sought to formulate an evolutionary model that could hopefully supply us with a definition of what is a species. This model needed to respect a few basic rules so that it would be equivalent to our biophysical model. These rules were basically that: (1) each bacterial species was associated with a single viral species and vice versa (i.e., there is no cross interaction between phage-host systems) and (2) each species (bacterial or viral) was unique and distinguishable from all other species. We then asked ourselves the following question: if we start from a state of a single bacterial species interacting with a single viral species, how would this state evolve so that after some time we obtained a state comprised of two bacterial species and two viral species, where the new species were independent of the old species. Such an evolutionary model would create a “world” with single viral species paired with single bacterial species, and vice versa, and where each bacteria-viral species pair was independent of all other pairs. We found that in order for these strains to evolve we needed to (1) define the concept of a “strain”, which is like a species, only there is no restriction on whom this strain can or cannot interact with, and (2) assume that whenever a new bacterial strain emerges, a corresponding viral strain co-emerges so that the symmetry between bacteria and viruses is conserved. By considering how a species evolved through generation of strains, into a new species, a qualitative picture of what a species is, within the context of this model, emerged (Fig. 1.2; see also Fig. 5.1 that illustrates the process of co-speciation). How this complex structure was obtained is discussed in detail in Chapter 5. When this structure is viewed in a genetic coarse-grained way we recover the simple picture of our biophysical model: species interacting uniquely with species. Therefore we argue that this model can be used to interpret the results of the biophysical model. Such a situation is often encountered in physics, where one theory is the limiting case of another (such as nuclear physics' versus particle physics' description of a proton). Such limits are related to scale transformations in renormalization group theory, possibly suggesting a deeper connection between the two models.

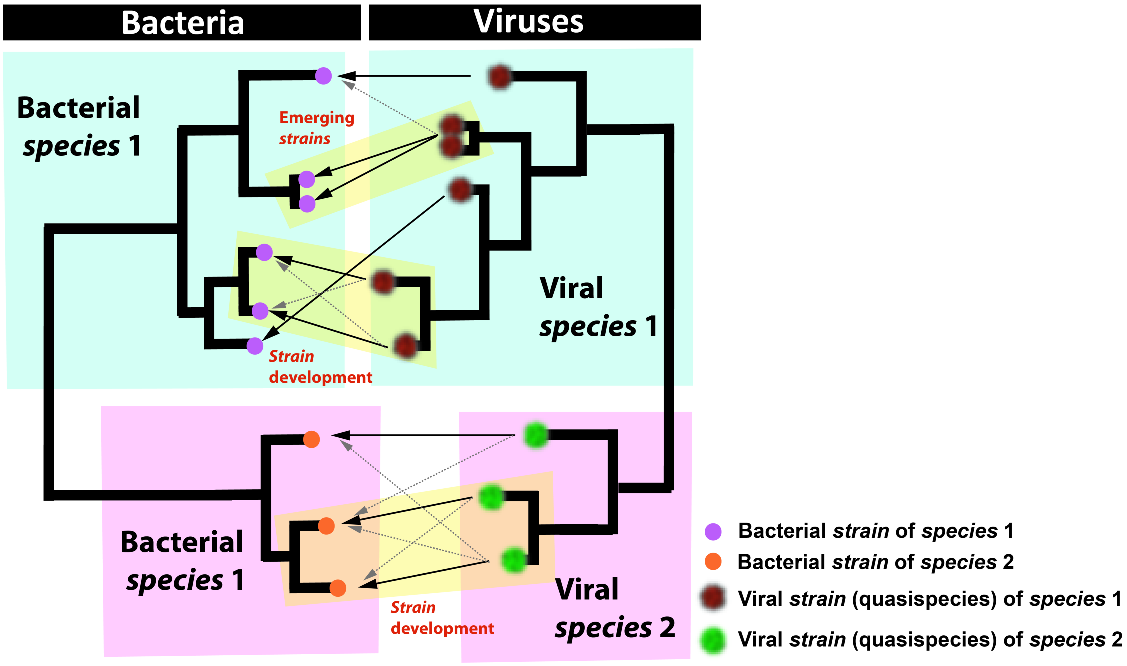

Bacteria
Viruses
Bacterial species 1
Emerging strains
Viral species 1
Strain development
Bacterial species 1
Strain development
Viral species 2
● Bacterial strain of species 1
● Bacterial strain of species 2
● Viral strain (quasispecies) of species 1
● Viral strain (quasispecies) of species 2

Figure 1.2 Schematic depiction of bacterial and viral species and strains. The relation between bacterial and viral species and strains according to a postulated evolutionary model considered in Chapter 5. Each bacterial species interacts with a single viral species. Bacterial strains on the other hand (that are simply emerging bacterial species) are initially part of a mesh of interactions with other strains. The interaction of the bacterial strain with the co-evolving viral strain is critical in order for both strains to evolve away from this state into a state of mutual independence (emerging as new species).

1.4.2 Is positive feedback driving co-speciation?

Perhaps the most interesting finding of this model was that in order for a new bacterial species and new viral species to co-emerge, the emerging bacterial and viral strains may be driving each other's evolution, through a positive feedback evolution mechanism. This positive feedback causes the strains to evolve as fast as possible from their initial state in order to sever the bonds with their parental strains and become independent (Fig. 1.3). This positive feedback evolution mechanism is the arms race between bacteria and viruses. The logic behind our suggestion is the following:

1. Phages for some reason have converged to an evolutionary solution where they have a narrow host range. This is not the most beneficial solution for a parasitic element, as a wide host range, such as that of a grazer, would be much more effective. Therefore, there appears to be some evolutionary advantage to this solution.
2. In the process of the arms race, viruses cause selective sweeps in the bacterial population. Such bottlenecks are known to accelerate evolution as traits in small populations can be fixed quickly. Thus, the phage is driving bacterial evolution, distancing the new bacterial strain as fast as possible from the bacterial strain from which is was born (Fig. 5.1). This evolution is necessary for the emerging bacterial species since this will lead the emerging viral species to lose its affinity to the parental bacterial strain and gain control over it. As it gains control, the concentration of the emerging bacterial species increases (Fig. 1.4). The concentration of the emerging bacterial species is maximal when the new viral species has total control over it. Thus, this process allows the emerging bacterial species in the end to “take up” its own concentration.
3. As the new viral species is emerging, it is controlling two populations, the parental bacterial species and the new bacterial strain (Fig. 5.1). In order for this phage to form a unique association with the new bacterial species (i.e., control only it) it must evolve away from its current state as far as possible until it can no longer infect the parental bacterial strain. This process appears to be achieved through the bacterial-viral arms race, since the new bacterial strain is forcing the virus to keep muting in order to track the new bacterial strain and the virus is causing the new bacterial strain to evolve.
4. Combining 2+3 we conclude that perhaps through positive feedback the bacterial and viral species are moving at the greatest possible pace from an initial state of one species to a final

state of two species (Fig. 5.1). Thus, viruses are the tool of evolution to generate species, and the narrow host range of phages is necessary to achieve this goal.

Chaotic evolution?

Since a positive feedback mechanism amplifies noise exponentially (like the shrill of a microphone in front of a speaker), the process of bacterial speciation may be simply a process of "amplifying noise". This may open the door to quantitative analysis via chaos theory. For example, it would be interesting to see if phylogenic trees of bacteria spanning many orders (strain, species, genus, family, order, etc.) display any features of self similarity, the hallmark of fractals generated by chaos theory. Further quantities that may be tractable are the rate speciation and the number of strains per species.

An experimental system to test the predictions of this theory would be Lenski-type evolution experiments with E. coli + a lytic phage. Specific experiments are suggested in section 5.4. In section 4.5 we suggest a series of experiments to test our biophysical model.

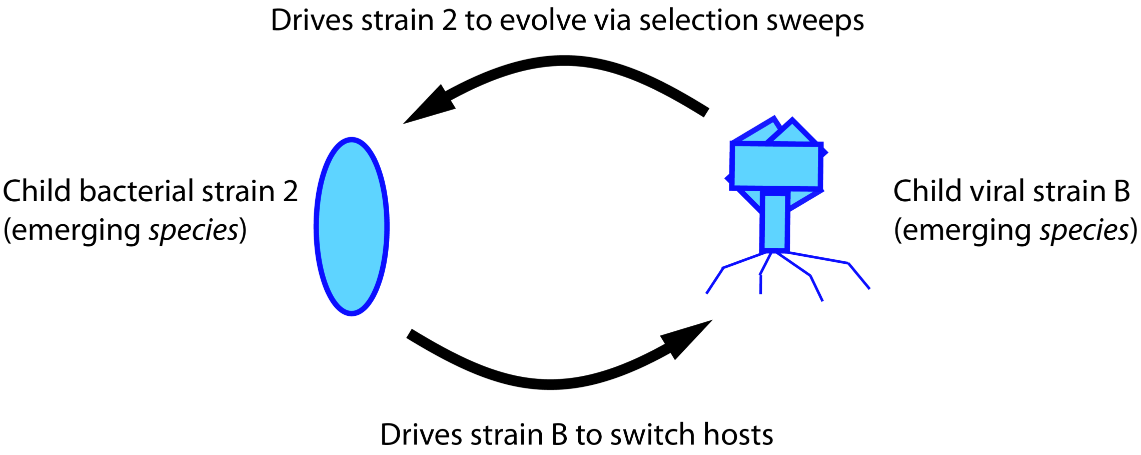

Drives strain 2 to evolve via selection sweeps
Child bacterial strain 2 (emerging *species*)
Child viral strain B (emerging *species*)
Drives strain B to switch hosts

Figure 1.3 Positive feedback evolution model for emerging bacterial and viral "species". The arms race between bacteria and viruses may be a critical step in the formation of a new bacterial and viral "species". This process is critical in order to allow viral strain B to relinquish its control of its parental bacterial strain (strain 1) while at the same time gaining control over the new bacterial strain (strain 2). Therefore this "arms race" may allow the two emerging "species" to form a one-to-one association, leading to the result that vial species have a narrow ("species") host range. This process may also be critical for the bacterium, where by selective sweeps, the controlling child viral strain drives the bacterial strain to evolve away from its original parental strain. This positive feedback model may be initially amplifying "noise". Thus the process of co-speciation is perhaps equivalent to "amplification of noise" and thus may be a chaotic effect. Covering the genome space at such an exponential rate may be required in order to converge to a solution on a practical timescale, especially given the fact that bacteria are much less efficient and exploring this space than diploid organisms. Thus, the arms race may be an equivalent solution of bacteria to sexual reproduction (possibly a good enough solution for a smaller genome size).

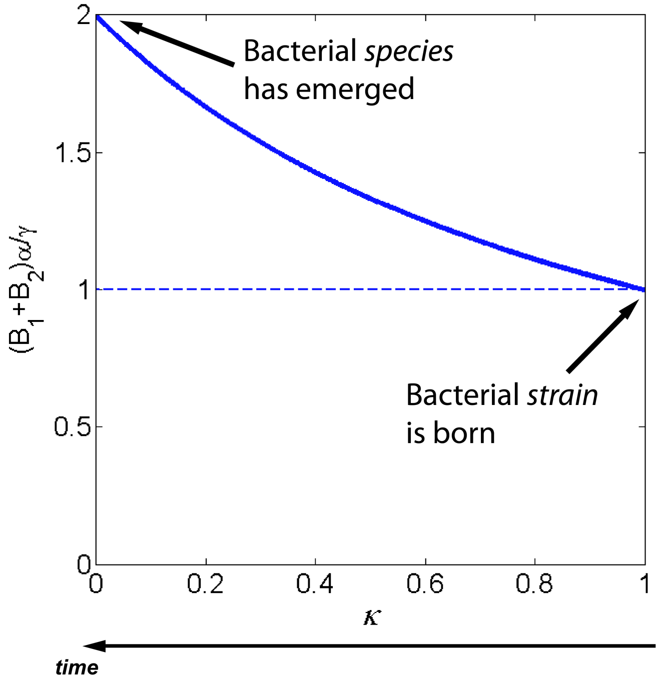


Figure 1.4. Total concentration "taken up" by the evolving bacterial strain and its parental strain. Initially, the total concentration of the parental bacterial species (B_1) and the just-emerging bacterial species (B_2), in normalized units, is 1 and is determined by the controlling viral species. As the new bacterial strain is emerging, it is driving the evolution of the emerging viral strain, causing its affinity to the parental bacterial strain to drop (i.e., κ, which is a measure of the affinity of the emerging viral species to the parental bacterial species, decreases). This causes the emerging viral species to gain more control over the emerging bacterial species, and so its concentration increases. When co-speciation is completed, the new viral species has total control of the new bacterial species and has lost its affinity to the parental bacterial species (i.e., it has a species host range). This allows the total concentration to double (i.e., B_1=B_2=1). See also Fig. 5.4.

1.5 The experimental frontier

1.5.1 Phage-host co-localization methodology

Thus far, phage-host interaction in the wild could only be investigated for certain systems such as cyanophages [28,29,30,31]. The challenge lays in the fact that traditional techniques in microbiology necessitate that hosts be culturable in order to isolate their phage. Yet when >99% of bacteria cannot be cultured [32] other methods need to be sought. In Chapter 2 we describe a method using digital microfluidic PCR array to pair phages with their bacterial host without having any prior assumptions regarding the host. The experimental scheme for a new environment is shown in Fig. 1.5.

The first stage involves obtaining a metagenome for the environment of interest. Once gene objects have been assembled and translated, one can run a bioinformatic tool called MetaCAT (Chapter 3) that was written for this purpose. MetaCAT (metagenome cluster analysis tool) is used to find the most abundant viral genes in a given metagenome by clustering together similar genes in the metagenome that are expected to be related (Fig. 1.6). This tool is used to find candidate viral marker genes. The idea behind using this tool for viral genes is that viral genes tend to have many mutations, and therefore are not collapsed by the assembler. Therefore we expect that the abundance of the genes in the metagenome reflects their abundance in the sample. Thus abundant viral genes found by MetaCAT would correspond to abundant and thus dominant genes in the sample. Furthermore, the more alleles one has for primer design, the more general the primers will be, and the more diversity they will recover.

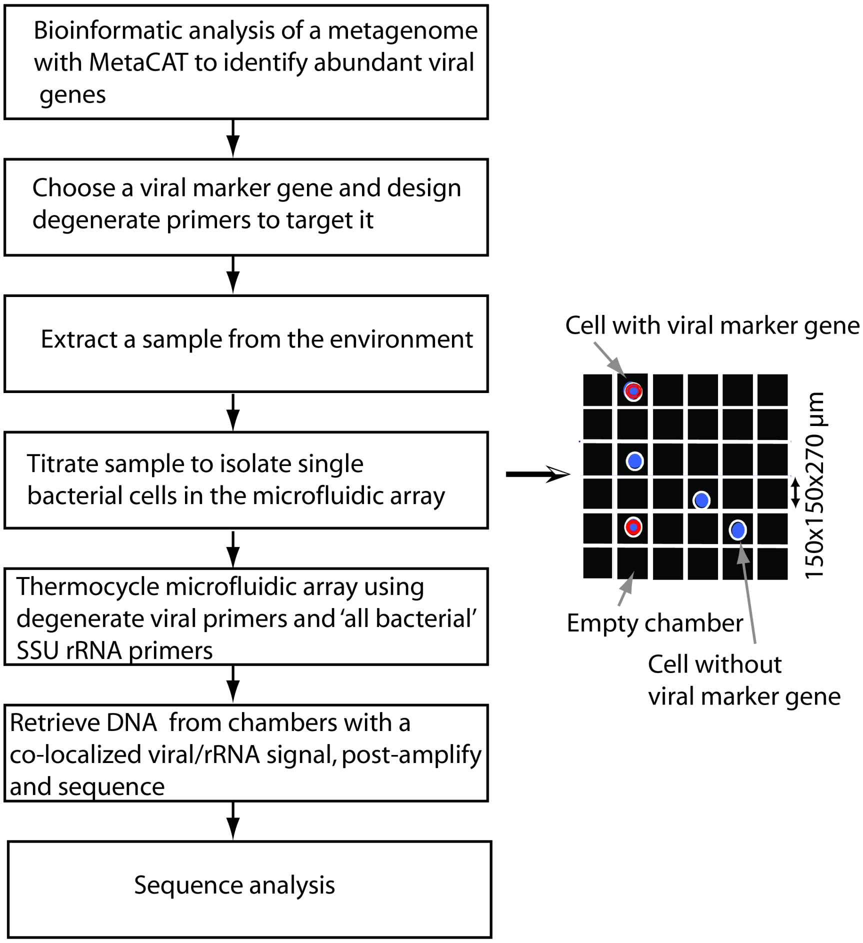

Bioinformatic analysis of a metagenome with MetaCAT to identify abundant viral genes
Choose a viral marker gene and design degenerate primers to target it
Extract a sample from the environment
Titrate sample to isolate single bacterial cells in the microfluidic array
Thermocycle microfluidic array using degenerate viral primers and 'all bacterial' SSU rRNA primers
Retrieve DNA from chambers with a co-localized viral/rRNA signal, post-amplify and sequence
Sequence analysis
Cell with viral marker gene
150x150x270 µm
Empty chamber
Cell without viral marker gene

Figure 1.5 Workflow using the microfluidic digital PCR array for host-virus co-localization in a novel environmental sample

After degenerate primers have been designed, one can load an environmental sample onto a digital PCR microfluidic array panel, which distributes the sample evenly among 765 6 nl chambers. Samples are titered such that a small fraction of chambers contain a single cell that is probed for a universal small subunit rRNA gene and a viral marker gene. In Fig. 1.7 we show a typical digital PCR panel after PCR cycling. Each chamber that contains both colours (red for a viral marker gene and green for the small subunit rRNA gene) is a potential co-localization signal. Chambers displaying co-localization are retrieved and sequenced allowing later phylogenetic analysis. The great challenge with this experiment was that the phage gene displayed many mutations from cell to cell of the same host species. Therefore we needed to devise a statistical criterion and sampling strategy to separate repeated co-localizations due to chance from genuine co-localization.

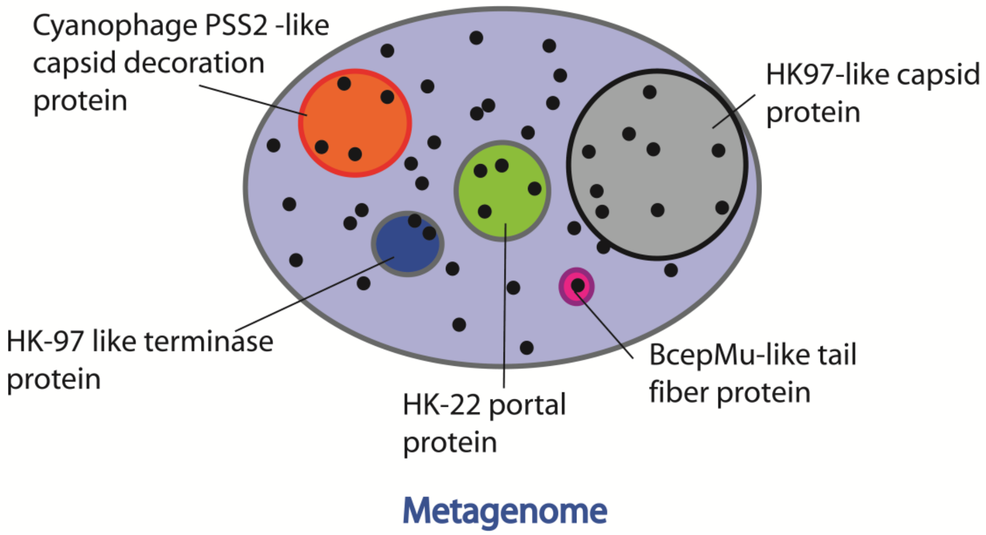

Cyanophage PSS2 -like capsid decoration protein
HK97-like capsid protein
HK-97 like terminase protein
HK-22 portal protein
BcepMu-like tail fiber protein
Metagenome

Figure 1.6 Ideal clustering of gene objects in a metagenome. Each dot represents a gene object in a metagenome, with the entire metagenome depicted by the blue oval. Similar genes are grouped into clusters (circles of different colors) and each cluster is represented by a single gene from a known reference database. In this schematic description, the distance between dots is interpreted in an abstract manner and does not correspond to a rigorous metric.

1.5.2 The case of the termite hindgut

The co-localization experiment described in Chapter 2 was performed for samples from the gut of a termite. Our analysis of the termite hindgut began by analyzing the metagenome of a higher termite collected from Costa Rica. This analysis detected several highly abundant unique viral genes. We then BLASTed these genes against the genomes of two spirochetes that were isolated from a lower termite collected from northern California. We found that two genes had very close homologs: a portal protein and a terminase protein. These two genes were part of larger prophage-like elements (two elements in each genome). We then proceeded to uncover the entire

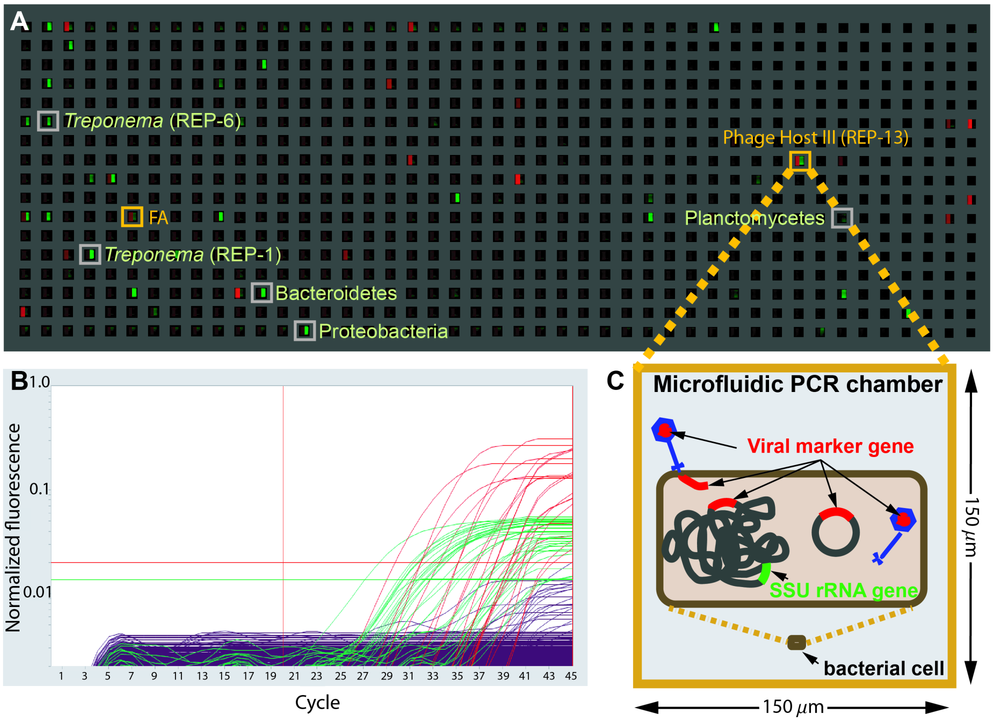


Figure 1.7. End-point fluorescence measured in a panel of a microfluidic digital PCR array. **A.** The measured end-point fluorescence from the rRNA channel (right half of each chamber) and the terminase channel (left half of each chamber) in a microfluidic array panel. **B.** Normalized amplification curves of all chambers (red/viral, green/rRNA). **C.** Specific physical associations between a bacterial cell and the viral marker gene resulting in co-localization include for example: an attached or assembling virion, injected DNA, an integrated prophage or a plasmid containing the viral marker gene.

prophage-like element in each genome (Fig. 1.8). To show that these prophage-like elements were also abundant in the metagenome we BLASTed each gene from in the prophage-like element against the metagenome. The result, shown in Fig. 1.8, indicated that these prophage-like elements were indeed abundant in both termites. We chose the terminase gene to be our viral marker gene and designed degenerate primers to amplify a large portion of this gene (~820 bp). To test the hypothesis that the prophage-like element that we found is ubiquitous in termites, we tested these primers against nine termite species belonging to seven families collected from five different geographical locations. Fig. 1.9 shows that indeed we obtained positive hits for all these termites confirming that this prophage-like element is ubiquitous to termites (at least of north and central America). We also obtained a positive hit for a wood feeding roach, raising the possibility that this prophage-like element has infected a common ancestor of termites and wood feeding roaches and has been transformed since. In Chapter 2 we describe the results of our co-localization experiment gut samples extracted from *Reticulitermes hesperus* using the same viral primers.

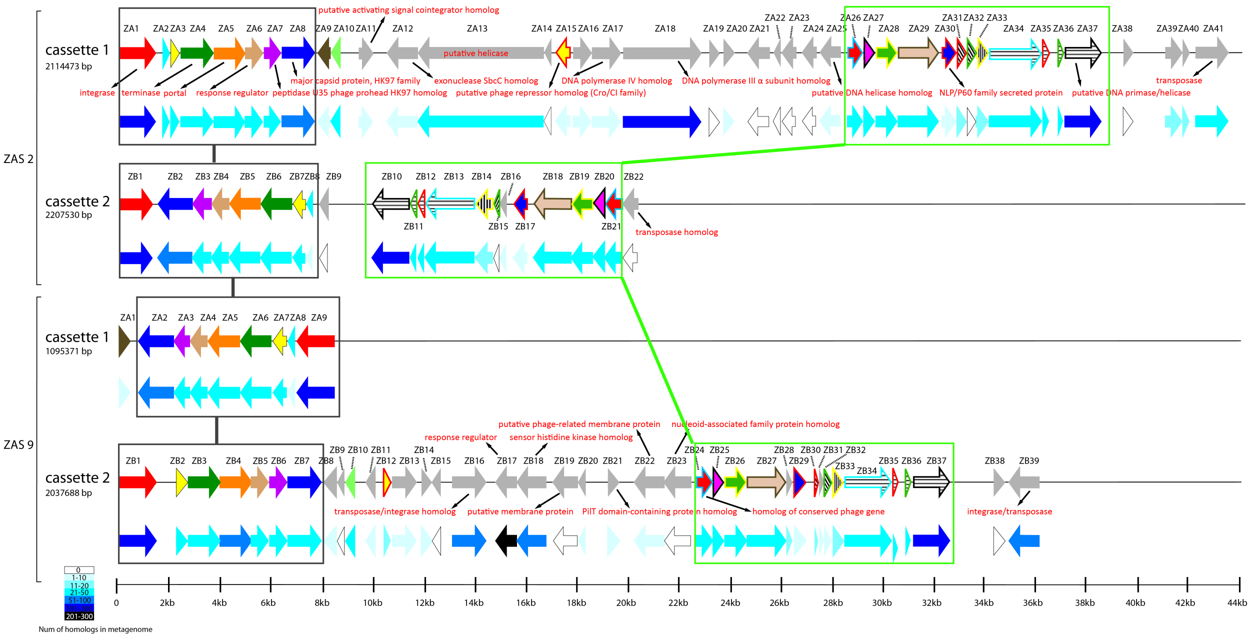

ZAS 2
cassette 1
2114473 bp
ZA1 ZA2 ZA3 ZA4 ZA5 ZA6 ZA7 ZA8 ZA9 ZA10 ZA11 ZA12 ZA13 ZA14 ZA15 ZA16 ZA17 ZA18 ZA19 ZA20 ZA21 ZA22 ZA23 ZA24 ZA25 ZA26 ZA27 ZA28 ZA29 ZA30 ZA31 ZA32 ZA33 ZA34 ZA35 ZA36 ZA37 ZA38 ZA39 ZA40 ZA41
putative activating signal cointegrator homolog
putative helicase
integrase
terminase
portal
response regulator
peptidase U35 phage prohead HK97 homolog
major capsid protein, HK97 family
exonuclease SbcC homolog
putative phage repressor homolog (Cro/CI family)
DNA polymerase IV homolog
DNA polymerase III α subunit homolog
putative DNA helicase homolog
NLP/P60 family secreted protein
putative DNA primase/helicase
transposase
cassette 2
2207530 bp
ZB1 ZB2 ZB3 ZB4 ZB5 ZB6 ZB7 ZB8 ZB9 ZB10 ZB11 ZB12 ZB13 ZB14 ZB15 ZB16 ZB17 ZB18 ZB19 ZB20 ZB21 ZB22
transposase homolog
ZAS 9
cassette 1
1095371 bp
ZA1 ZA2 ZA3 ZA4 ZA5 ZA6 ZA7 ZA8 ZA9
cassette 2
2037688 bp
ZB1 ZB2 ZB3 ZB4 ZB5 ZB6 ZB7 ZB8 ZB9 ZB10 ZB11 ZB12 ZB13 ZB14 ZB15 ZB16 ZB17 ZB18 ZB19 ZB20 ZB21 ZB22 ZB23 ZB24 ZB25 ZB26 ZB27 ZB28 ZB29 ZB30 ZB31 ZB32 ZB33 ZB34 ZB35 ZB36 ZB37 ZB38 ZB39
response regulator
putative phage-related membrane protein
sensor histidine kinase homolog
nucleoid-associated family protein homolog
transposase/integrase homolog
putative membrane protein
PilT domain-containing protein homolog
homolog of conserved phage gene
integrase/transposase
0
1-10
11-20
21-50
51-100
101-200
201-300
Num of homologs in metagenome
0 2kb 4kb 6kb 8kb 10kb 12kb 14kb 16kb 18kb 20kb 22kb 24kb 26kb 28kb 30kb 32kb 34kb 36kb 38kb 40kb 42kb 44kb

Figure 1.8 Map of viral cassettes in ZAS2 and ZAS9 highlighting gene frequency in the higher termite metagenome. Blue arrows represent abundance in the metagenome.
]

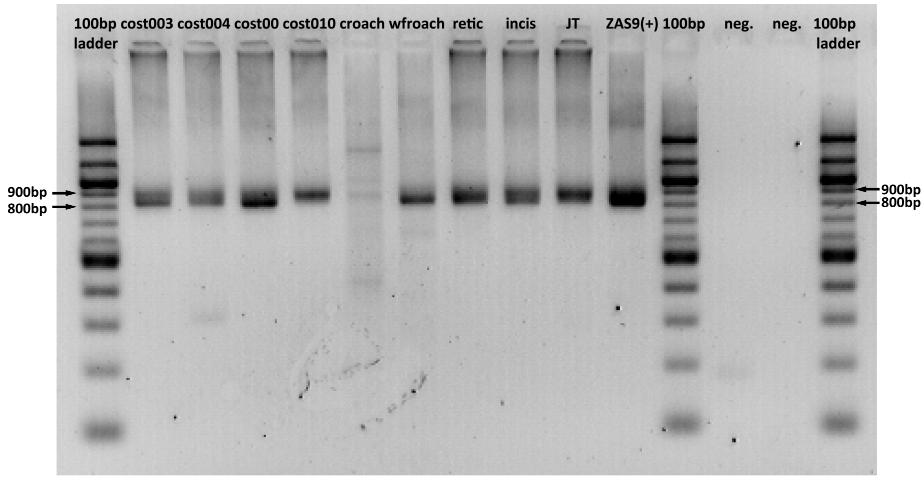


Figure 1.9. Agarose gel electrophoresis analysis of terminase PCR product amplified from termite and related insect species. PCR product using degenerate primers ter.7F and ter.5eR targeting the large terminase subunit gene. Specimens included were: *Nasutitermes sp.* (cost003), *Rhynchotermes sp.* (cost004), *Microcerotermes sp.* (cost008), *Amitermes sp.* (cost010), *Periplaneta americana* (croach), *Cryptocercus punctulatus* (wfroach), *Reticulitermes hesperus* (retic), *Incisitermes minor* (incis), Gnathiamitermes sp. JT5 (JT). ZAS9 was used as a positive control. Also shown are two negative PCR controls.

1.6 Stress fibers in single fibroblast cells

Dr. Blake W. Axelrod, a research engineer in the Roukes lab, built a microfluidic NEMS device allowing him to measure the force as a function of time of a stress fiber in a single fibroblast cell, performing the highest resolution measurement to date. In this experiment, a single fibroblast cell (Fig. 1.10A insert) contacts a NEMS force sensor. When the cell is placed in a recovery medium it exerts a force on the force sensor (Fig. 1.10A, blue region) corresponding to the force generated by an assembled stress fiber (about 20nN). Once a substance called cytochalasin D is flowed in, the stress fiber undergoes disassembly and consequently the force declines (Fig. 1.10A, red region). The process of disassembly is a reversible one, since when flowing the recovery medium back again, the stress fiber assembles again and the force is regenerated. When examined more closely, each assembly/disassembly profile appears to be comprised of steps that appear to, on average, increase in duration (Fig. 1.10B). These steps were remarkably uniform in the force amplitude and it was postulated that they are the result of individual sarcomeres failing or contracting.

At the time the data was presented it was not clear what the origin of the temporal dynamics is, what the mechanism leading to exponential-like "charging" and "discharging" curves is, and why the steps are increasing with time. In Chapter 6 we present our analysis of Blake's dataset. We proposed a simple stochastic model for stress fiber assembly and disassembly, whereby individual sarcomeres assemble or dissemble (a) abruptly, (b) irreversibly, (c) independently, (d) with the time to the event of assembly or disassembly exponentially distributed with a fixed time constant (Fig. 1.11). With this model it is simple to explain why, for example, steps increase in time. According to this model, in the case of stress fiber disassembly for example, the time from the perturbation (t=0) until a sarcomere fails is exponentially distributed. If there are N sarcomeres, then at t=0 there are N independent sarcomeres that can fail. Thus one needs to wait a short period of time to observe a step. As more steps fail, one needs to wait longer until one sees a step because there are less remaining sarcomeres. In Chapter 6 we show that the inverse duration times of each step increase linearly with time. In Fig. 1.12 we show the inverse duration times versus the linear prediction of the model (where parameters were estimated from the data based on the model). The data appears to be behaving qualitatively as predicted. In Chapter 6 we present more rigorous tests to check our model. We show the (a) sarcomeres appear to be failing or assembling statistically as exponential variables. (b) When data was rescaled using model parameters estimated from the data, all profiles collapsed to the predicted ensemble average.

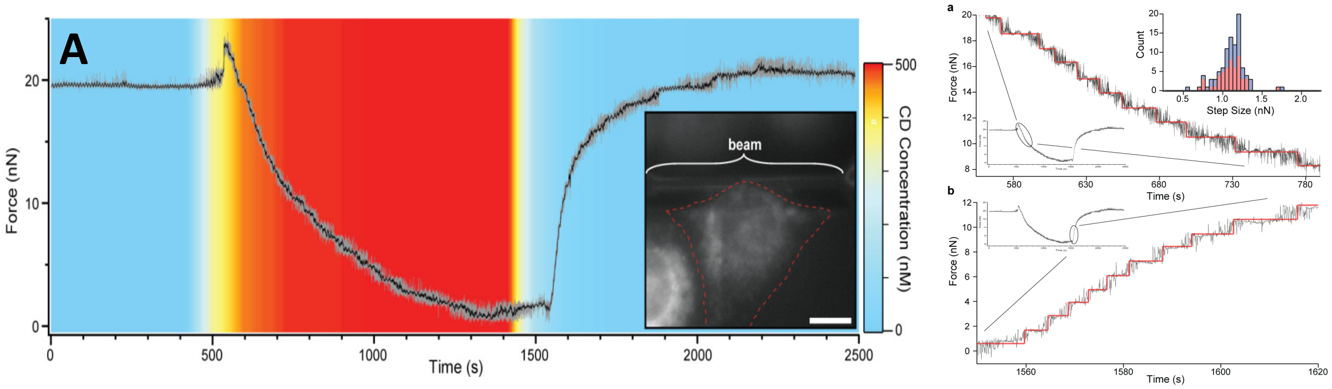


Figure 1.10 Typical force-time response to Cytochalasin D perturbation. A. Typical measured force response to the force disruptor Cytochalasin D (red region) and recovery medium (blue region). Cytochalasin D belongs to a class of substances called Cytochalasins that are fast acting and reversible disruptors of contractile force. When Cytochalasin D is flowed in, the force decays due to stress fiber disassembly. When recovery medium is flowed in, the stress fiber reassembles and the force increases with time. Inset shows fluorescent image of the cell attached to the beam taken immediately before data acquisition, scale bar is 10 µm. **B.** Steps during CD-induced force collapse (upper) or force recovery (lower). The average step size is 1.08 nN ± 0.18 nN (*n*=96). Figure and caption courtesy of Blake Axelrod (Roukes lab, Caltech).

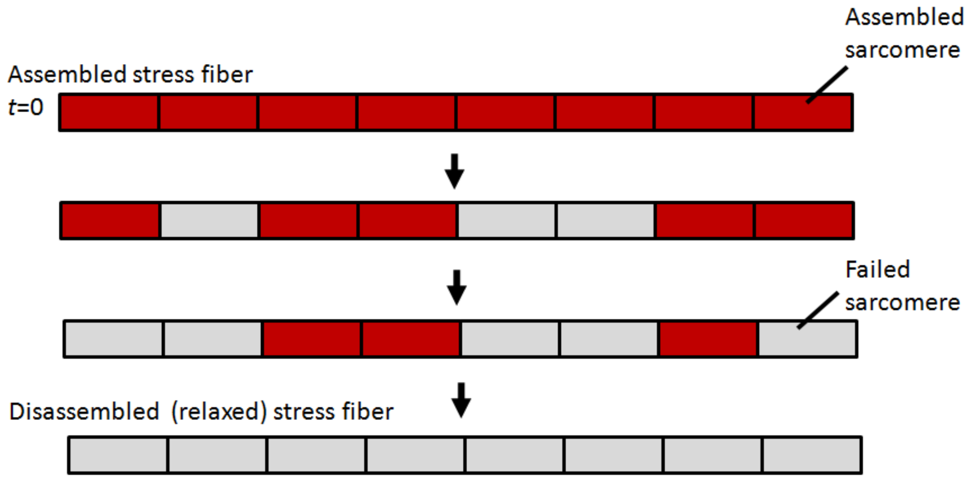


Figure 1.11. Schematic model for stress fiber relaxation. (1) Each sarcomere assembles or disassembles abruptly and irreversibly. (2) Sarcomeres assemble or disassemble independently of each other. (3) The time until a sarcomere assembles or disassembles is exponentially distributed (reflecting a certain constant probability rate for this event to occur). (4) The time constant for assembly is the same for all sarcomeres. Similarly, the time constant for disassembly is the same for all sarcomeres.

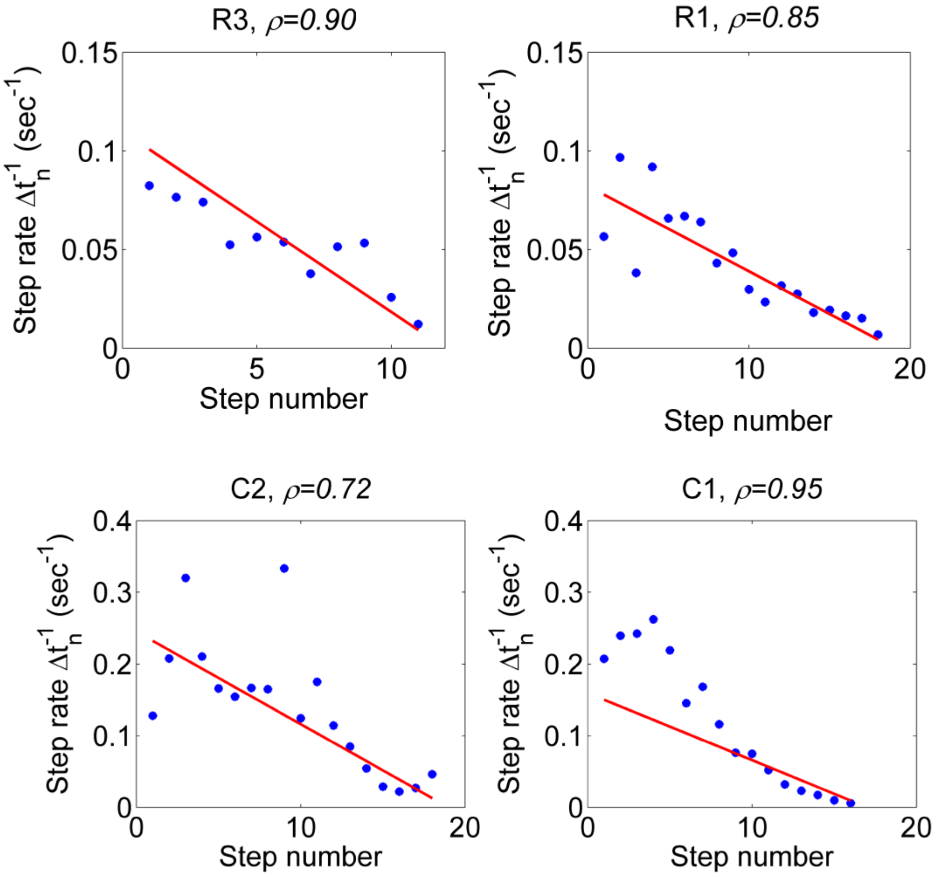


Figure 1.12. Stochastic model prediction of force step durations versus experimental data. The stochastic model prediction for the inverse of the force step durations (red curve) versus experimental data points (blue dots). ρ is the Pearson correlation coefficient measuring the strength of the correlation. The red line is not a fit, as these lines were predicted based on parameters that were estimated from the data according to our stochastic model. Note that we anticipate a high level of noise since the standard deviation equals the predicted rates.

1.7 References

1. Weinbauer M (2004) Ecology of prokaryotic viruses. FEMS Microbiology Reviews 28: 127-181.
2. Danovaro R, Corinaldesi C, Luna GM, Magagnini M, Manini E, et al. (2009) Prokaryote diversity and viral production in deep-sea sediments and seamounts. Deep Sea Research Part II: Topical Studies in Oceanography 56: 738-747.
3. Klieve AV, Swain RA (1993) Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. Applied and Environmental Microbiology 59: 2299.
4. Kamra D (2005) Rumen microbial ecosystem. Current Science 89: 124-135.

5. Fuhrman J (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541-548.
6. Suttle C (2007) Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5: 801-812.
7. Wommack K, Colwell R (2000) Virioplankton: viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews 64: 69.
8. Suttle CA, Chan AM (1994) Dynamics and distribution of cyanophages and their effect on marine Synechococcus spp. Applied and Environmental Microbiology 60: 3167.
9. Danovaro R, Dell'Anno A, Corinaldesi C, Magagnini M, Noble R, et al. (2008) Major viral impact on the functioning of benthic deep-sea ecosystems. Nature 454: 1084-1087.
10. Edwards R, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504-510.
11. Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? Molecular Microbiology 49: 277-300.
12. Wilcox R, Fuhrman J (1994) Bacterial viruses in coastal seawater: lytic rather than lysogenic production. Marine Ecology-Progress Series 114: 35-35.
13. Suttle C (2005) Viruses in the sea. Nature 437: 356-361.
14. Paul J, Kellogg C (2000) Ecology of bacteriophages in nature. Viral Ecology: 211–246.
15. Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. Nature Reviews Microbiology 3: 537-546.
16. Thingstad T, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquatic Microbial Ecology 13: 19-27.
17. Suttle C (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. Viral Ecology: Academic Press. pp. 247–296.
18. Kutter E, Sulakvelidze A (2005) Bacteriophages: biology and applications: CRC Press.
19. Thingstad T (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography: 1320-1328.
20. Weinbauer MG, Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? Environmental Microbiology 6: 1-11.
21. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology 6: 181-186.
22. Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS One 4: e4169.
23. Banfield J, Young M (2009) Variety--the Splice of Life--in Microbial Communities. Science 326: 1198.
24. Weinbauer M, Peduzzi P (1994) Frequency, size and distribution of bacteriophages in different marine bacterial morphotypes. Marine Ecology Progress Series 108: 11-20.
25. Weinbauer M, Hoefle M (1998) Size-specific mortality of lake bacterioplankton by natural virus communities. Aquatic Microbial Ecology 15: 103-113.
26. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66.
27. Sheldon R, Prakash A, Sutcliffe Jr W (1972) The size distribution of particles in the ocean. Limnology and Oceanography 17: 327-340.

28. Sullivan M, Huang K, Ignacio-Espinoza J, Berlin A, Kelly L, et al. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol 12: 3035-3056.
29. Lindell D, Jaffe J, Coleman M, Futschik M, Axmann I, et al. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449: 83-86.
30. Angly F, Felts B, Breitbart M, Salamon P, Edwards R, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368.
31. Williamson S, Rusch D, Yooseph S, Halpern A, Heidelberg K, et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3: 1456.
32. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: reviews0003.

Chapter 2

Probing Individual Environmental Bacteria for Viruses Using Microfluidic Digital PCR

2.1 Abstract

Viruses may very well be the most abundant biological entities on the planet. Yet neither metagenomic studies nor classical phage isolation techniques have shed much light on the identity of the hosts of most viruses. We used a microfluidic digital PCR approach to physically link single bacterial cells harvested from a natural environment with a viral marker gene. When we implemented this technique on the microbial community residing in the termite hindgut, we found genus-wide infection patterns displaying remarkable intra-genus selectivity. Viral marker allelic diversity revealed restricted mixing of alleles between hosts indicating limited lateral gene transfer of these alleles despite host proximity. Our approach does not require culturing hosts or viruses and provides a method for examining virus-bacterium interactions in many environments.

2.2 Introduction

Despite the pervasiveness of bacteriophages in nature and their postulated impact on diverse ecosystems (*1*), we have a poor grasp of the biology of these viruses and their host specificity in the wild. Though significant progress has been made with certain host-virus systems such as cyanophages (*2-5*), this is the exception rather than the rule. Conventional plaque assays used to isolate environmental viruses are not applicable to >99% of microbes in nature since the vast preponderance of the microbial diversity on Earth has yet to be cultured *in vitro* (*6*). Given the magnitude of the problem, the development of high-throughput, massively-parallel sequencing approaches that do not rely on cultivation to identify specific virus-host relationships are required. While metagenomics has revolutionized our understanding of viral diversity on Earth (*7-9*), that approach has as yet done little to shed light on the nature of specific viral-host interactions, except in restricted cases (*10*).

2.3 Proposed method for phage-host co-localization

Recent advances in microfluidic technology have enabled the isolation and analysis of single cells from nature (*11-13*). Here we present an alternative to the classical phage enrichment technique where we propose to use an uncultured virus to capture its hosts from the environment using a microfluidic PCR approach called digital multiplex PCR (*12, 14*). To this end, microbial cells were harvested directly from the environment, diluted and loaded onto a digital PCR array panel containing 765 PCR chambers operating at single-molecule sensitivity. Samples were diluted such that the majority of chambers were ideally either empty or contained a single bacterium (Fig. 2.1), achieving a Poisson distribution (*15*). Because there is no universally conserved gene in viruses (*7, 16*), degenerate primers (*17*) were designed to target a subgroup of diverse phage-like elements (*18*). Concurrently, the small subunit ribosomal RNA (SSU rRNA) gene encoded by each bacterial cell was amplified using universal “all bacterial” primers (see Fig. 2.4 for experimental design). Possible genuine host-virus associations detectable by this assay are depicted in Fig. 2.1C. Free phages may also co-localize with hosts, however these events are not expected to lead to statistically significant co-localizations due to the random nature of these associations (*19*).

2.4 Hunting for phages in the termite hindgut

The system we chose to investigate was the termite hindgut. This microliter-in-scale environment contains ~10^7 prokaryotic cells per μl (*20*) with over 250 different species of bacteria (*21*), making it ideally suited to explore many potential, diverse phage-host interactions. To find a viral marker gene relevant to such an environment, the more abundant candidate viral marker genes present in the sequenced metagenome from a hindgut of a higher termite from Costa Rica collected in 2005 (*22*) were examined (Table 2.2; search algorithm described in the Materials and methods section). We then checked if any of these viral genes had homologous counterparts in the sequenced genomes of two spirochetes isolated in 1997 from a laboratory colony of a genetically and geographically distant termite originally collected in 1986 from Northern California (*23-24*). We identified two such genes encoding a large terminase subunit protein (homologous to the T4 associated pfam03237 Terminase_6) and a portal protein (homologous to pfam04860 Phage_portal) exhibiting about 70–78% amino acid identity to their closest homologs in the higher termite gut metagenome (Table 2.3). This finding is surprising given that typically, across biology, portal proteins and terminase proteins from different phages exhibit little overall sequence similarity (*25-28*). Further analysis revealed that the spirochete viral genes were part of a larger prophage-like element, with the majority of recognizable genes most closely related to *Siphoviridae* phage genes (*19*). The association of these genes with prophage-like elements is consistent with the fact that both the Terminase_6 pfam and the Phage_portal pfam describe proteins in known lysogenic and lytic phages.

As a viral marker gene for this prophage-like element we chose the large terminase subunit gene. This gene is a component of the DNA packaging and cleaving mechanism present in numerous double-stranded DNA phages (*26*) and is considered to be a signature of phages (*29*). We consequently designed degenerate primers based on the collection of fifty metagenome and treponeme-isolate alleles of this gene. The ~820bp amplicon spanned by these primers covered about two thirds of this gene and approximately 77% of the predicted N-terminal domain containing the conserved ATPase center (*26, 30*), the “engine” of this DNA packaging motor (*31*) (see alignments in Figs. 2.5 and 2.6). Testing these primers against the RefSeq viral database (*32*) did not yield any hits (Fig. 2.5). Indeed, the closest homolog of this gene in the RefSeq viral database displayed only 25% amino acid identity (Table 2.3). Thus, while this terminase gene was clearly associated with the Terminase_6 pfam, the termite related alleles appear to be part of a novel assemblage of terminase genes in this environment and not closely related to previously sequenced phages (Fig. 2.5).

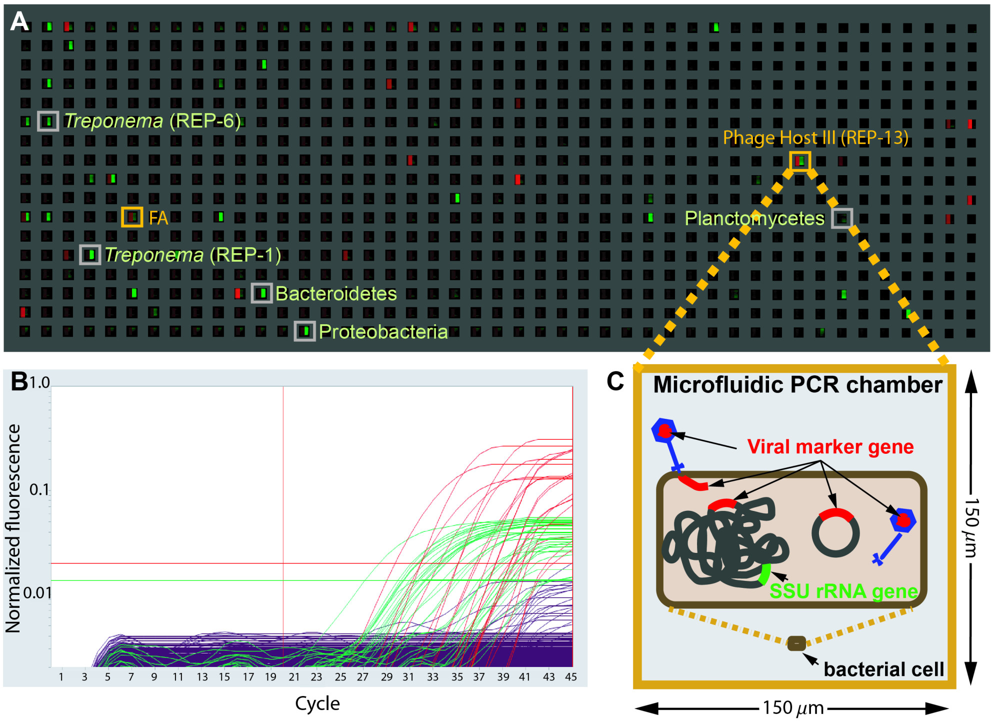


Figure 2.1. End-point fluorescence measured in a panel of a microfluidic digital PCR array. A. The measured end-point fluorescence from the rRNA channel (right half of each chamber) and the terminase channel (left half of each chamber) in a microfluidic array panel. Each panel in the array (one of twelve) consists of 765 150 x 150 x 270µm^3 (6 nL) reaction chambers. Retrieved co-localizations are outlined in orange and positive rRNA chambers randomly selected for retrieval are outlined in gray. FA indicates false alarm (a probable terminase primer-dimer). **B.** Normalized amplification curves of all chambers in (A) after linear derivative baseline correction (red/viral, green/rRNA). **C.** Specific physical associations between a bacterial cell and the viral marker gene resulting in co-localization include for example: an attached or assembling virion, injected DNA, an integrated prophage or a plasmid containing the viral marker gene.

Given that terminase genes of different phages often exhibit less sequence similarity (see above), the fact that we found such closely related terminase genes from such distantly related termites collected from well separated geographical locations (California and Costa Rica) and from specimens collected almost two decades apart led us to speculate that this family of viral genes and prophage-like elements might be ubiquitous in termites. Indeed, to date we have identified close homologs of the large terminase subunit gene in the gut communities of nine termite species belonging to seven families collected from five different geographical locations. We therefore wished to identify the bacterial hosts associated with this viral marker gene. To this end, we made collections of representatives of a third previously unexamined termite family (Rhinotermitidae; *Reticulitermes hesperus*, from a third geographical location in Southern California) over a span of six months (Table 2.4). We then performed seven independent experiments, where in each case the hindgut contents of three worker termites were pooled, diluted, and loaded onto a digital PCR array, screening in total ~3000 individual hindgut particles (i.e., individual cells or possibly clumps of cells positive for the SSU rRNA gene).

2.5 Identification of novel uncultured bacterial hosts

Of the 41 retrieved co-localizations, 28 were associated with just four phylotypes designated "Phage Hosts I, II, III and IV" (see Fig. 2.2, Table 2.1 and the phylogenetic analysis in Fig. 2.7 and Tables 2.5 and 2.6). Statistically, the reproducible co-amplifications were significant and cannot be explained by random co-localization of two unassociated genes (Table 2.1). Furthermore, these associations were independently reproduced in specimens from different colonies collected six months apart (Fig. 2.2), indicative that relationships between specific host bacteria and viral markers were being revealed.

All four of the phylotypes were members of the spirochetal genus *Treponema* and exhibited significant diversity within this genus (Table 2.5). No reproducible or statistically robust associations involving other bacteria were observed. The terminase alleles that associated with these cells shared ≥69.8% identity (average 81.9 ± 8.3% standard deviation, SD)(*33*) and were divergent from other currently known terminases

(Fig. 2.5), suggesting that the primer set amplifies elements exclusively found associated with termite gut treponemes. Analysis of the retrieved terminase gene sequences reveal that they are under substantial negative selection pressure with $\omega=\beta/\alpha=0.079$, where ω is the relative rate of non-synonymous, *β*, and synonymous, *α*, substitutions (*18*)(see Table 2.7 for additional estimates for individual hosts). Furthermore, none of the terminase sequences in Fig. 2.2 appeared to encode either errant stop codons or obvious frame shift mutations, and functional motifs appeared to be conserved (Fig. 2.5). Together, the sequence data suggest that these genes have been active in recent evolutionary history and are not degenerating pseudogenes (*19*).

Table 2.1 | Statistics of repeatedly co-localized SSU rRNA genes

Host	No. of repeated co-localizations[*] (*n*=41)	Occurrence in reference library[†] (*n*=118)	P value (one tailed, *n*=41)‡
Host I	13	5	5.4×10^{-18}
Host II	8	2	7.6×10^{-13}
Host III	4	1	5.7×10^{-7}
Host IV	3	1	3.8×10^{-5}

[*] Based on the DOTUR analysis described in Table 2.5
[†] Based on the DOTUR analysis described in Table 2.6. Reference library frequencies are roughly 1/3 of the co-localization frequencies indicating that sampling was unbiased.
‡The statistical test to determine the P value is explained in the supporting text.

Since the viral marker gene was present in hosts spanning a swath of species of termite gut treponemes, we were interested to see if this viral marker exhibited any selectivity within this genus. The relative frequency of free-living *Treponema* phylotypes was determined by randomly sampling chambers positive for the rRNA gene (*18*) (Fig. 2.3, Fig. 2.7). We found that Hosts I through IV were relatively infrequent, comprising 1.3% to 6.4% of the sampled *Treponema* cells (Table 2.1) and collectively about 9.8% of the sampled bacterial cells (correcting for reagent contaminants). Interestingly, the three most abundant

Treponema phylotypes in the survey constituting ~30, 10 and 9% of the free-swimming spirochetal cells (REPs 1, 2 and 3 in Fig. 2.3; see also Fig. 2.7 and Table 2.6) were never co-retrieved with the viral marker gene, to the extent that this target was spanned by our degenerate primers. Given that the degenerate core region (*17*) of each primer targets residues that were strictly conserved in gut microbes of highly divergent termite specimens (Fig. 2.5), and that these primers successfully amplified this gene from the guts of many different termite species (see above), it appears that these strains are most likely either insensitive to this virus or that only a small percentage are infected (*19*). Therefore we conclude that ~50% of the free-swimming spirochetal cells in the gut were likely not infected with an element encoding the targeted viral marker gene, whereas ~12% were hosts potentially infected (Fig. 2.3).

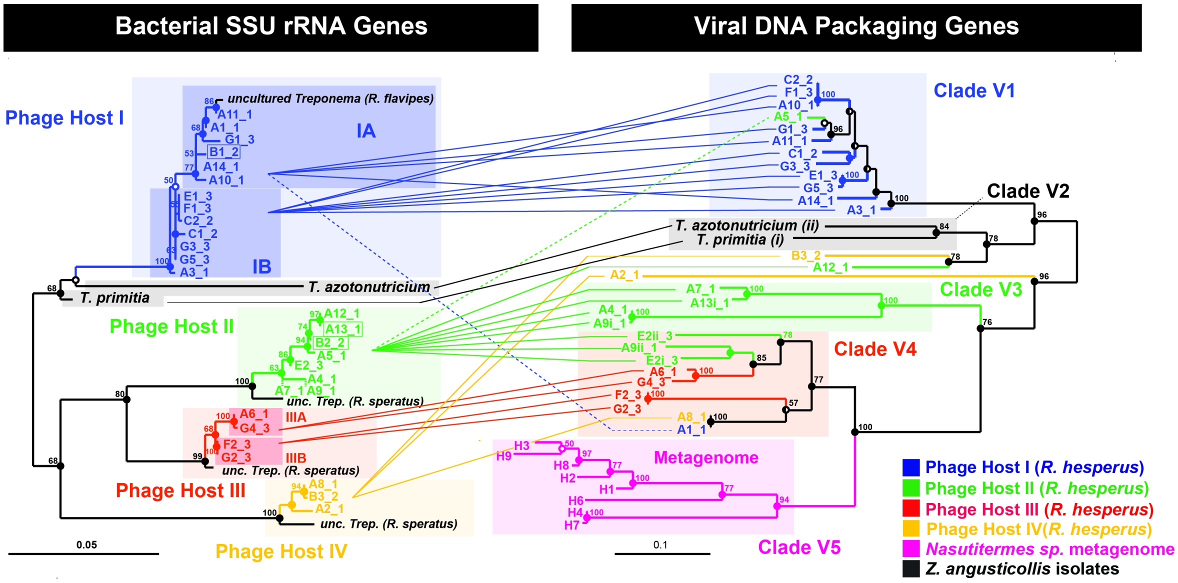


Figure 2.2. Phylogenetic relationship between cultured and uncultured bacterial host rRNA genes and their associated viral DNA packaging genes. Left: Maximum Likelihood (ML) tree of 898 unambiguous nucleotides of the SSU rRNA gene of ribotypes that repeatedly co-localized with the terminase gene, including the two isolated spirochetes *Treponema primitia* and *Treponema azotonutricium*. Shorter sequences (A7, 780bp and A9, 806bp) were added by parsimony (dashed branches). **Right:** ML tree of 705 unambiguous nucleotides of the large terminase subunit gene. Connecting lines represent co-localized pairs, revealing restricted mixing of terminase alleles between different bacterial hosts. For association of three additional recombinant sequences (boxed on the left) see Fig. 2.8. Statistically we estimate that an average of 0.6 co-localizations are false (~2% error (*19*)). The sequence error rate (*40*) for the rRNA and terminase genes was measured to be 0 (*n*=8) and <0.6±0.3% SD (*n*=9), respectively (*18*). Alleles are named by array (A–G) and retrieval index followed by an underscore and the colony number (colony 1 being sampled six months prior to colonies 2 and 3). Lower-case Roman numerals indicate multiple terminases per chromosome. Scale bars represent substitutions per alignment. For interpretation of node support refer to (*18*) and for accession numbers Table 2.11.

2.6 Phage-host cophylogeny

To elucidate the evolutionary relationship between the terminase alleles and their hosts we examined the phylogeny of the terminase genes associated with each bacterial host. Terminase alleles from *R. hesperus* formed separate clades from the clades of the two other termite species investigated in this study (Clades V2 and V5 in Fig. 2.2). Within *R.*

hesperus, different bacterial hosts exhibited different patterns of viral allelic diversity. Terminase sequences associated with Host I, for example, were highly clonal, with 11 out of 13 terminase alleles sharing 96.7 ± 1.7% SD identity (*n*=11, Clade V1) (*33*). Conversely, terminase alleles associated with Host II displayed marked diversity (79.1 ± 6.2% identity, *n*=11) (*33*), deep branches and divergent multiple alleles per bacterium for 3 out of 8 repetitions (with 15–31% divergence). The unique features of the terminase alleles associated with Host II compared with Host I may reflect a more ancient infection or possibly an infection by a phage replicating with a lower fidelity. Alternatively, Host II may be a more sensitive bacterial host susceptible to a wider range of phages. Overall, phage terminase alleles associated with different bacterial hosts were significantly divergent with only three exceptions (Table 2.8).

The tandem trees in Fig. 2.2 reveal multiple possible relations between bacterial hosts and terminase alleles: while Host I was associated almost exclusively with a single terminase clade (V1), Host II was associated with multiple terminase clades (primarily V3 and V4). Conversely, terminase Clade V1 was associated almost exclusively with Host I, while terminase Clade V4 was associated with all bacterial hosts. Overall, the terminase tree was highly structured and displayed specific bacterial host associated clades (e.g., Clades V1 and V3, see Fig. 2.8A). Applying the P Test (*34*) implemented in Fast UniFrac (*35*) to terminase alleles grouped by bacterial host indeed revealed significant differences between alleles associated with most pairs of hosts (Table 2.9). Grouping terminase alleles by colony, however, did not reveal significant differences between alleles (Table 2.10), indicating that sampling was not a factor in determining the observed host associated heterogeneity in terminase alleles. The highly non-random distribution of host associated terminase alleles therefore suggest that lateral gene transfer and/or host switching is limited in this system. This result, however, could also reflect the fact that the terminase gene does not appear to shuffle randomly among phages, possibly indicating a connection between DNA packaging and other characteristics of the phage (*36*). It remains to be seen whether other viral genes follow similar patterns.

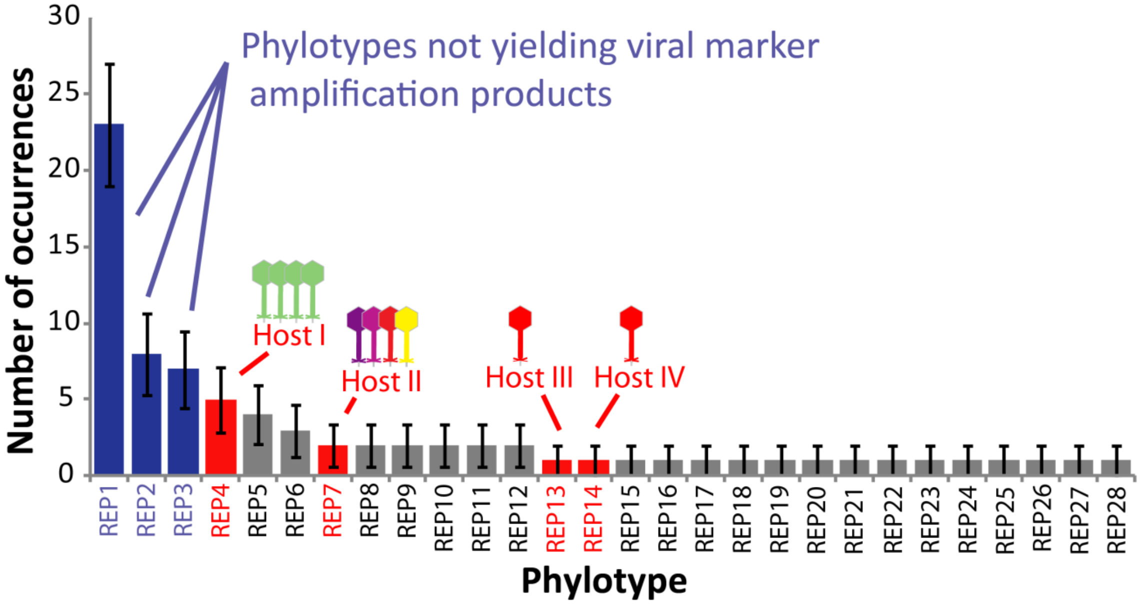


Figure 2.3. Rank abundance curve of free living *Treponema* spirochetes in *R. hesperus* termites identifying putative phage hosts. A library of 118 random chambers positive for the rRNA gene were retrieved, post-amplified, and sequenced. Of these, *n*=78 were related to the *Treponema* genus, corresponding to 28 different phylotypes using an operational taxonomical unit, OTU, cut-off set by DOTUR (*41*) at 3.1%. Here we show these 28 phylotypes, designated as *Reticulitermes* Environmental Phylotypes (REPs), ordered by their abundance. Phylotype abundance is expected to reflect true relative abundances in the gut, since single-cell amplification is not susceptible to primer bias or rRNA copy number bias. Phylotypes identified as phage hosts are marked by red bars (with the highly clonal marker associated with Host I depicted by green viruses and the divergent marker associated with Host II depicted by colored viruses). The most abundant free living *Treponema* in the gut — REPs 1, 2, and 3 (blue bars) were not associated with the viral marker. Remaining bars are gray. Error bars are estimated by the binomial SD. See Table 2.6 for OTU assignment. Note that the isolated spirochetes were not spanned by these REPs (see Fig. 2.7).

The fact that there was little mixing between terminase alleles associated with Host I (V1) and the more distantly related Hosts II (V3 and V4) and III (V4), whereas alleles of the more closely related Hosts II and III (Table 2.5) exhibited a certain degree of mixing (V4), supports the notion that the probability of cross-species transmission or lateral gene transfer decreases with the phylogenetic distance of the hosts (*37*). The rRNA gene of Hosts I through IV also exhibited patterns of microdiversity that may have physiological relevance (*38-39*), however, mirrored only by the terminase alleles of Host III. Host I and II terminase alleles appeared to be indifferent to the bacterial host at the sub-species level.

2.7 Conclusions

Our results show that, in a marked departure from classical phage enrichment techniques, specific viral-host relationships can be revealed in uncultivated cells harvested straight from the environment. We found that variants of a viral packaging gene appear to have infected bacterial hosts across an entire genus of bacteria. Furthermore, despite the significant potential for lateral gene transfer and/or host switching in this well-mixed, small-volume system, the terminase tree was highly structured and displayed specific bacterial host associated clades. It will be interesting to continue to monitor the host-virus interactions within this ecosystem as a function of space and time and across the termite community at large, shedding further light on host-virus co-evolution in this unique ecosystem. More broadly, the method we have developed enables a highly parallel analysis of host-virus interactions in environmental samples from virtually any environment in nature.

2.8 References and notes

1. C. Suttle, *Nat. Rev. Microbiol.* **5**, 801 (2007).
2. M. Sullivan et al., *Environ. Microbiol.* **12**, 3035 (2010).
3. D. Lindell et al., *Nature* **449**, 83 (2007).
4. F. Angly et al., *PLoS Biol* **4**, e368 (2006).
5. S. Williamson et al., *PLoS ONE* **3**, 1456 (2008).
6. P. Hugenholtz, *Genome Biol.* **3**, reviews0003 (2002).
7. R. Edwards, F. Rohwer, *Nat. Rev. Microbiol.* **3**, 504 (2005).
8. E. Dinsdale et al., *Nature* **452**, 629 (2008).
9. D. Kristensen, A. Mushegian, V. Dolja, E. Koonin, *Trends Microbiol.* **18**, 11 (2009).
10. A. Andersson, J. Banfield, *Science* **320**, 1047 (2008).
11. R. N. Zare, S. Kim, *Annu. Rev. Biomed. Eng.* **12**, 187 (2010).
12. E. Ottesen, J. Hong, S. Quake, J. Leadbetter, *Science* **314**, 1464 (2006).
13. Y. Marcy et al., *Proc. Natl. Acad. Sci. USA* **104**, 11889 (2007).
14. L. Warren, D. Bryder, I. L. Weissman, S. R. Quake, *Proc. Natl. Acad. Sci. USA* **103**, 17807 (2006).
15. S. Dube, J. Qin, R. Ramakrishnan, *PLoS ONE* **3**, doi:10.1371/journal.pone.0002876 (2008).
16. F. Rohwer, R. Edwards, *J. Bacteriol.* **184**, 4529 (2002).
17. T. Rose et al., *Nucleic Acids Res.* **26**, 1628 (1998).
18. See Materials and methods in the Appendix.
19. See Supporting text in the Appendix.
20. A. Tholen, B. Schink, A. Brune, *FEMS Microbiol. Ecol.* **24**, 137 (1997).

21. Y. Hongoh, M. Ohkuma, T. Kudo, *FEMS Microbiol. Ecol.* **44**, 231 (2003).

22. F. Warnecke et al., *Nature* **450**, 560 (2007).

23. J. Leadbetter, T. Schmidt, J. Graber, J. Breznak, *Science* **283**, 686 (1999).

24. T. Lilburn et al., *Science* **292**, 2495 (2001).

25. S. D. Moore, P. E. Prevelige Jr, *Curr. Biol.* **12**, R96 (2002).

26. V. Rao, M. Feiss, *Annu. Rev. Genet.* **42**, 647 (2008).

27. S. Chai et al., *J. Mol. Biol.* **224**, 87 (1992).

28. K. Eppler, E. Wyckoff, J. Goates, R. Parr, S. Casjens, *Virology* **183**, 519 (1991).

29. S. Casjens, *Mol. Microbiol.* **49**, 277 (2003).

30. M. Mitchell, S. Matsuzaki, S. Imai, V. Rao, *Nucleic Acids Res.* **30**, 4009 (2002).

31. S. Sun et al., *Cell* **135**, 1251 (2008).

32. K. Pruitt, T. Tatusova, D. Maglott, *Nucleic Acids Res.* **33**, D501 (2005).

33. Percent identity was measured across 235 unambiguous aligned amino acids.

34. A. P. Martin, *Appl. Environ. Microbiol.* **68**, 3673 (2002).

35. M. Hamady, C. Lozupone, R. Knight, *The ISME journal* **4**, 17 (2009).

36. S. Casjens et al., *J. Bacteriol.* **187**, 1091 (2005).

37. N. Wolfe et al., *Global Change & Human Health* **1**, 10 (2000).

38. L. Moore, G. Rocap, S. Chisholm, *Nature* **393**, 465 (1998).

39. J. Thompson et al., *Appl. Environ. Microbiol.* **70**, 4103 (2004).

40. S. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, M. Polz, *Appl. Environ. Microbiol.* **71**, 8966 (2005).

41. P. Schloss, J. Handelsman, *Appl. Environ. Microbiol.* **71**, 1501 (2005).

2.9 Appendix

2.9.0 Contents

2.9.1 Materials and methods

Termite collection

Reticulitermes hesperus specimens were collected from Chilao Flats Campground in the Angeles National Forest (Table 2.4). Throughout the experiment, starting in the field, different colonies were kept in separate tip boxes and never came in contact with each other. Colonies thereafter were maintained in the laboratory (*S1*). Microfluidic array experiments were carried out days to weeks (<4 weeks) thereafter.

PCR on the microfluidic array

Microfluidic array multiplex PCR reactions contained Perfecta multiplex qPCR master mix (Quanta Biosciences), 0.1% Tween 20 (Sigma Aldrich Incorporated), 100nM ROX (Quanta Biosciences). Universal 16S SSU rRNA primers and probes used were (*S1*): forward 357F 5'-CTCCTACGGGAGGCAGCAG-3' (300nM), reverse 1492RL2D 5'-TACGGYTACCTTGTTACGACTT-3' (300nM), 1389 probe HEX-GTGCCAGCMGCCGCGGTAA-BHQ1 HPLC purified (300nM). Unprobed terminase primers used were: forward ter7F 5'-CATTTGATTTGCCGTTACCGIGCYAARGAYGC-3' (200nM) and reverse ter5eR 5'-CICCWCCAGCCGGATCRCARTAMAC-3' (100nM). The probed terminase reverse primer used was: ter5eR.L 5'- CAGCCACACICCWCCAGCCGGATCRCARTAMAC-3' (100nM). The universal probe used for the terminase primer set was: Roche Universal Probe #5 (250 nM). The primers and the rRNA probe were ordered from Integrated DNA Technologies and resuspended in sterile TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8) filtered with a 0.02 μm sterile Anotop syringe filter (Whatman). Primers and probes were diluted in DEPC-treated sterile filtered water (Sigma) and then sterile filtered again (prior to dilution) with a 0.02 μm syringe filter.

Preparation of termite hindguts

In each experiment three *Reticulitermes hesperus* worker termites from the same colony (and same tip box) were incubated for several minutes at 4°C to immobilize the specimens and whole guts were subsequently extracted using sterilized forceps on a disposable sterile petri dish. Guts were resuspended in 897 µL of 4°C "synthetic gut fluid" (SGF) salt solution (*S2*) pre-filtered with a 0.02 µm sterile syringe filter containing 0.5 µg/mL final concentration of DNase free RNase (Roche) to prevent inhibition by ribosomal RNA. Guts were repeatedly disrupted with a sterile 1 ml filter pipette tip and suspensions were briefly vortexed and allowed to settle for 30 seconds to sediment large particles. Samples were then diluted to working concentrations using the SGF diluent. For microfluidic arrays C through G the resuspended gut fluid was further filtered with an Acrodisc 5 µm sterile syringe filter (Pall Life Sciences) to remove inhibiting large particles such as wood fragments and protists. Samples were then mixed 1:10 with the PCR reaction mix (above) for immediate loading onto the primed microfluidic array once the dilutions were completed. Termite bodies were frozen for later analysis of their COII sequences (see below).

Microfluidic array thermocycling and fluorescence analysis

BioMark 12.765P peelable microfluidic arrays from Fluidigm were loaded with the samples described above and PCR was performed using the BioMark system (Fluidigm Corporation) as recommended by Fluidigm. The cycling protocol was 95ºC 5 min, (95ºC 15 s, 60ºC 90 s) x 45, 10 min at 60ºC, 20ºC 10 sec. Amplification curves were evaluated using BioMark Digital PCR analysis software (Fluidigm, v.2.0.6) applying ROX normalization and a linear baseline correction. FAM fluorescence threshold was set to detect any increase in fluorescence, while the HEX threshold was set above the fluorescence leakage of the FAM channel into the HEX channel, detectable in both a no-16S rRNA-primer control panel (dedicated for this purpose) and the no-template-control panel. Both panels were included in every microfluidic array. To minimize diffusion from neighboring chambers after pressure release, only chambers displaying fluorescence in both channels that were flanked by chambers displaying no fluorescence in both channels were selected for retrieval. An example of end-point fluorescence of an array panel is shown in Fig. 2.1A. In this figure only fluorescence from within chambers is shown, detected based on the reference dye fluorescence measurement. To illustrate the nature of co-localizations, we mask the chambers in such a way that half of each chamber shows one fluorescence channel and the other half shows the other. This way the left half of each chamber showed only the FAM/viral channel fluorescence and the right half of each chamber showed only the HEX/SSU rRNA channel fluorescence. Fluorescence is shown on a logarithmic scale with background subtracted.

Sample retrieval

Microfluidic arrays were peeled shortly after the end of the PCR run and pressure in the arrays was released by depressing the pressure valves. Samples were retrieved into 10µl TE buffer (that was pre-filtered with a 0.02µm sterile Anotop syringe filter) using disposable sterile 30.5G needles (*S1*) (one disposable needle per chamber) and subsequently evaluated for the presence of target genes via conventional simplex PCR. In addition, for each array, with the exception of array B, at least five chambers were also retrieved from the no-template-control panel to test for possible cross-contamination (all control retrievals were negative - see below). The PCR reaction mix consisted of perfecta qPCR multiplex master mix with the SSU rRNA primers at 300nM concentration and terminase primers at 200nM concentration. The SSU rRNA probe, the Universal Probe #5 and the probe binding primer ter5eR.L were omitted from these reactions. The cycling protocol for conventional PCR for the simplex terminase reaction was 95ºC 3 min, (95ºC 15 s, 60ºC 60 s, 72ºC 60

s) x 40, 72ºC 10 min and for the simplex SSU rRNA reaction was the same but with 32 cycles of amplification to prevent amplification of contaminates associated with the Taq master mix. The presence or absence of product was evaluated using agarose gel electrophoresis. Samples that displayed a band at the expected fragment size for both simplex reactions were deemed successful.

The majority of successful retrievals from the microfluidic arrays were amplified for cloning and/or sequencing in two 30 μL reactions using 3.5 U of EXPAND high fidelity polymerase (Roche), Fail-Safe PCR PreMix D (Epicentre), and primers and cycling conditions as above. In the case of microfluidic array A, terminase sequences were amplified with Perfecta qPCR multiplex master mix instead. For each reaction 1.5 μL of retrieved sample was used. PCR products were purified using the Qiagen PCR purification kit, and sequenced using the terminase ter7F and ter5eR primers and SSU rRNA gene internal primers 1100R (3'-AGGGTTGCGCTCGTTG-5') and 533F (3'-GTGCCAGCMGCCGCGGTAA-5'). Sequencing reactions of microfluidic array amplicons were carried out by the USC DNA core facility (Los Angeles, CA) using an annealing temperature of 50 or 55ºC.

Sequences that contained a mixture of SSU rRNA sequences were discarded from further analysis. Sequences that contained a mixture of terminase sequences, or in which the trace quality was poor were cloned for sequencing using the TOPO TA cloning kit (Invitrogen). At least eight colonies from each cloning reaction were picked and used as templates for PCR reactions. PCR reaction mix included Fail-Safe PCR PreMix H (Epicentre), Taq polymerase (New England Biolabs) and standard T3/T7 primers at 250 nM. Cycling conditions were 95ºC 3 min, (95ºC 15 s, 55ºC 30 s, 72ºC 60 s) x 35, 72ºC 10 min. Sequences with different restriction fragment length polymorphism (RFLP) patterns were chosen for sequencing. For the RFLP analysis, 6 μl of each reaction was digested at 37ºC for 4 hr with 3 units HinPI1 from New England Biolabs followed by an inactivation step at 65ºC for 20 min. A representative of each RFLP type (with the correct product band) was sequenced with the high fidelity polymerase and standard T3 and T7 primers. PCR products were purified using the Qiagen PCR purification kit and sequenced with standard T3/T7 primers. Sequencing reactions for cloning were carried out by Laragen Inc. (Los Angeles, CA).

Identification of termite species

The mitochondrial cytochrome oxidase II (COII) gene was used to identify the termite specimens analyzed in this study (*S3*-*S5*). For each of the three colonies that were collected, either heads or bodies of three to five worker termites frozen on the day of the microfluidic array experiments were used as a template for a PCR amplification of the COII gene. Primers used were A-tLeu (5'-ATGGCAGATTAGTGCAATGG -3') and B-tLys (5'-GTTTAAGAGACCAGTACTTG-3')(*S6*-*S7*). For colonies 1 and 2 the PCR product was cloned and sequenced. For colony 3 the product was directly sequenced. Colonies 1, 2, and 3 shared 99.3% nt identity with 0 gaps (0.003% SD; *n*=3 over 680 unambiguous nt) and 100% amino acid identity (over 226 residues) with the COII sequence of *Reticulitermes hesperus* isolate LBL2 (accession# AY623445.1).

Sequence analysis

Sequence traces were converted into a nucleotide sequence using Lasergene SeqMan Pro v8.1.2. Representatives of the SSU rRNA nucleotide sequence of Hosts I through IV were then screened for chimeras using Pintail (*S8*) and Bellerophon (*S9*), the latter implemented in Greengenes (*S10*), returning negative results. All terminase sequences from all 41 co-localizations were also tested for amplification related chimeras using Bellerophon (*S9*). Cases where both chimera parents belonged to the same PCR batch (E2iii) were eliminated from further analysis.

SSU rRNA sequences were aligned by SILVA (*S11*) incremental aligner SINA and subsequently analyzed in ARB (*S12*) version 07.12.07org using SILVA release 100 (SSURef_100_SILVA_02_08_09_opt). jModelTest 0.1.1 (*S13-S14*) was used to find the optimal nucleotide substitution model for the rRNA sequences in Fig. 2.2, testing 40 different models on an alignment of 898 unambiguous nucleotides without gaps, estimating a maximum likelihood (ML) tree for each model. The optimal nucleotide substitution model (based on the AICc criterion with sample size set to the number of sites in the alignment) was a Tamura-Nei model (*S15*) +I+Γ with unequal base frequencies. A maximum likelihood tree was then computed for this alignment with PhyML 2.4.5 (*S14*) implemented in ARB using the Tamura-Nei model +I+Γ (nCat=4), with all parameters estimated from the data and with 1000 non-parametric bootstrap iterations. Other treeing methods such as Phylip DNAPARS v3.6a3 (*S16*) and Fitch-Margoliash (*S17*) distance method implemented in ARB predicted very similar topologies (Fig. 2.2). In Fig. 2.2 solid circles represent significant nodes supported by ML, parsimony (Phylip DNAPARS v3.6a3 (*S16*)), and distance (Fitch-Margoliash (*S17*)) methods. Half circles represent nodes supported by ML and either parsimony or distance methods. Open circles represent nodes supported by only ML. In addition, support values greater than 50% for 1000 bootstrap iterations are shown. We note that the topological relation between Phage Host clades I–IV appeared to be sensitive to the addition of other *Treponema* sequences from public databases, and to the particular outgroup chosen as well, and therefore the topology in Fig. 2.2, though robust, may not be definitive.

Nucleotide sequences of the large terminase subunit gene present in *R. hesperus*, Z. *angusticollis* and *Nasutitermes* sp. termites were translated in reading frame and aligned with ClustalW (*S18*) in MEGA4 (*S19*) (the alignment used in the analysis was straightforward and involved a single insertion event of a highly conserved five amino acid sequence in some of the sequences). Subsequently 705 unambiguous aligned nucleotides without gaps were tested for the presence of recombination with RDP3 v3.44 (*S20*). Methods used to scan for recombinant sequences included Geneconv (*S21*), Maxchi (*S22*), and RDP (*S23*) (as recommended in the RDP3 manual and shown to be the preferable tests for non-redundant sequences (*S24-S25*)) as well as the Bootscan method (*S26*). Since each recombination detection method individually is error prone (*S24-S25, S27*) several methods are required to explore recombination (*S24, S27*). Similar sequences ≤(3.3%) were removed prior to analysis as recommended in the RDP3 manual. The first two events found by RDP3 implicated by all four methods alleles A13ii and B1 as recombinants, confirmed by manual phylogenetic inspection in RDP3. A NeighborNet analysis with SplitsTree4 (*S28*) using optimal substitution parameters estimated by FindModel (*S29*) confirmed the reticulate nature of these alleles and consequently these alleles were excluded from the phylogenetic tree in Fig. 2.2 (see Fig. 2.8). The following two events detected by RDP3 (H5, B2) were only supported by Maxchi, however the NeighborNet network showed these putative recombinants were also associated with significant reticulate patterns, which were eliminated upon removal of these sequences. Consequently these two samples were also excluded from the phylogenetic tree. The remaining events detected by RDP3 with lower confidence exhibited either a small degree of local reticulate patterns or no reticulate patterns and were therefore kept in the analysis. Eliminating potential recombinant alleles resulted in a largely tree-like network suitable for phylogenetic analysis (Fig. 2.8B). A likelihood-mapping analysis (*S30-S31*) with TREE-PUZZLE 5.0 using 10000 quartets and the optimal model found by jModelTest (see below) showed that 95.7% percent of the quartets fell in the triangle corners (A_1,A_2,A_3) suggesting that a phylogenetic tree should fit the data (*S31*).

After recombinant sequences were removed, jModelTest was used to find the optimal nucleotide substitution model testing 40 different models, estimating a ML tree for each model. The optimal model (based on the AICc criterion as described above) was a Tamura-Nei model (*S15*) +I+Γ with the base frequencies having little effect on the AICc score. A ML tree was then computed with PhyML 2.4.5 implemented in ARB using the Tamura-Nei model with +I+Γ (nCat=4), with all parameters estimated from the data and with 1000 non-parametric bootstrap iterations. Other treeing methods such as DNAPARS v3.6a3 and Fitch-Margoliash distance method implemented in ARB predicted very similar topologies (Fig. 2.2). Tree topology was also similar to the ML estimated tree topology of the corresponding 235 amino acid residues, with the main differences being a slight repositioning of the higher termite clade and sequence A2. Since the terminase gene is comprised of two functional domains, an ATPase domain and a nuclease domain (see Fig. 2.6), we also compared the ML estimated topology of 495 unambiguous aligned nucleotides of the N-terminal domain of the gene (see Fig. 2.5 for alignment) with the nucleotide tree of the entire gene and found the topologies to be nearly identical. p-distances were measured in MEGA4 and standard deviations were calculated in Matlab.

Survey of SSU rRNA ribotypes on the microfluidic array

In order to assess the frequency of putative host ribotypes I through IV on the microfluidic array as well as the frequency of other rRNA ribotypes, we constructed a library of 118 randomly sampled rRNA hits from the microfluidic arrays. To this end, for two microfluidic arrays (F and G) and for every panel on these arrays (except the two control panels), 10 chambers for which the HEX (rRNA) fluorescence exceeded the detection threshold (irrespective of florescence in the FAM/terminase channel) were randomly selected for retrieval. The identities of the chambers for retrieval were obtained by a random number generator implemented in Matlab 7.4. These sequences were then post-amplified for sequencing using Perfecta multiplex qPCR master mix (Quanta Biosciences) as described in the Methods section. Sequencing was performed by the USC DNA core facility using internal SSU rRNA primers 533F and 1100R (see Methods). A total of 118 sequences were successfully sequenced and assembled using Lasergene SeqMan Pro v8.1.2. In Fig. 2.3 we plot the rank abundance curve of just *Treponema* phylotypes from the reference library. The frequency of each phylotype is given in Table 2.6. Each column in Fig. 2.3 can be thought of as a random variable sampled from a binomial distribution with mean $n \cdot p$ and standard deviation $SD = \sqrt{n \cdot p \cdot (1-p)}$, where p is the probability to sample this phylotype and n is the total number of trials (here n=78 trials). The error bars in Fig. 2.3 are ±SD, with p estimated for each phylotype as the number of occurrences of that phylotype divided by n.

Degenerate primer design and testing

Terminase phage primers were designed to target several conserved regions of the large terminase subunit gene found in the four prophage-like elements in *Treponema primitia* (ZAS-2) (*S32*) and *Treponema azotonutricium* (ZAS-9) (*S33*), and in 46 contigs found in the metagenome of a *Nasutitermes* species termite (*S5*). The primers were designed with CODEHOP (*S34*), selecting candidates with melting temperatures matching the all-bacterial SSU rRNA primer set (primer candidates were required to be different by at least five base pairs to be considered different candidates). The primer sequences in both the degenerate core region and the clamp region were manually tweaked to offer the best coverage for the conserved region (matching the codon bias in these sequences) and to minimize primer dimers. In addition, inosines were incorporated at certain positions instead of mixed bases to reduce primer degeneracy. Several forward and reverse primer candidates were chosen and the nucleotide regions were further adjusted to minimize forward/reverse primer-dimers and dimers with the all-bacterial primers and probe. Multiplex PCRs for various forward and reverse primers were performed on a dilution series of purified genomic DNA from ZAS-2 and ZAS-9. PCR products were analyzed by agarose gel electrophoresis and primers yielding the strongest bands and having the lowest detection limit (<100 copies) were selected. The chosen primers were further screened on genomic DNA extracted from *Zootermopsis nevadensis* by agarose gel electrophoresis.

To allow us to do quantitative PCR (qPCR) with these primers without having to design a degenerate probe we implemented a universal-template probe strategy first suggested by Zhang et al. (*S35*) and adapted for degenerate primers by Ottesen et al. (*S2*). In this method a short universal nondegenerate probe sequence is attached to the 5' end of the forward and/or reverse primers. The probe-binding sequence is incorporated into the amplicon during the first round of amplification, allowing the probe to detect amplification of that product. A short nondegenerate 8 base probe incorporating locked nucleic acids (LNAs) then binds to the probe-binding sequence and is subsequently cleaved by the DNA polymerase like in a standard TaqMan chemistry. The locked nucleic acids increase the melting temperature of the probe allowing usage of a very short probe. A probe yielding the minimal interaction with the SSU rRNA amplicon and other oligos in the master mix was chosen for this task. A linker sequence was incorporated between the probe-binding sequence and the degenerate primer to further reduce dimers.

Multiplex qPCR standard curves were obtained for all probe binding sequence combinations (probe binding sequence on the forward primer, probe binding sequence on the reverse primer and probe binding sequence on both the forward and the reverse primers) and for all the candidate primer sets. In all cases, primers with LNA probe binding sequences were mixed 50% with primers lacking the probe binding sequence as this seemed to enhance the PCR reaction. Primer sets yielding the best standard curves, highest end-point amplification for positive templates and highest Cts for the no-template-controls were selected. Primer sequences for the best candidates were fine tuned to further reduce dimers and then screened again using the same metric described above. The best candidates were then tested on ZAS DNA on the digital PCR microfluidic array. Primers yielding the best amplification curves, highest end-point amplification, and lowest number of no-template-control hits were selected. Finally, primer and probe concentrations were optimized on the microfluidic array for the chosen primer set. All benchtop qPCRs were performed on a Stratagene Mx3000P. Cycling conditions were as described in the Methods section.

Measures to prevent and test for contamination

To prevent contamination from the environment, from termites and from post-PCR products, several precautions were taken. Experiments were conducted in five different laboratories that were physically separated (different laboratories within the same building or different buildings). All PCR master mixes for dPCR runs, PCR master mixes for post-amplification of retrieved microfluidic array samples, and tubes loaded with 10 µl TE buffer for retrieved sample resuspension were prepared in laboratory #1 that never came in contact with termites or related samples thereafter. In addition, pipettes and benches were always thoroughly cleaned with EtOH or EtOH and bleach prior to setup. Termite handling and microfluidic array loading were conducted in laboratory #2, where each of these two procedures took place in well-separated designated areas. Sample retrieval was performed in a separate room within laboratory #2 using disposable syringes. Sample loading for post-amplification was performed in laboratory #3. Master mixes for cloning related PCR reactions were prepared in laboratory #3 (which was designated as a PCR cloning "clean area") and loading of samples for cloning-related PCR was performed in laboratory #4. All subsequent manipulations of samples or cloned PCR products (such as RFLP analysis, agarose gel electrophoresis, PCR purification, etc.) were performed in laboratory #5.

To test that no contamination occurred, every microfluidic array contained a no-template control panel and for each array (except B) at least five chambers from the no-template-control panel on the array were retrieved and processed with the rest of the samples to insure there was no cross-contamination during the retrieval process. No-template-control chambers retrieved for this purpose were selected such that these chambers and their flanking chambers on either of their sides did not exhibit fluorescence in both the FAM and HEX channels (this was done to prevent possible diffusion of targets from adjacent chambers into the sampled chamber after pressure release). All no-template-control samples that were retrieved from the microfluidic arrays were post-amplified with the rest of the retrievals and tested by agarose gel electrophoresis. All negative controls were always negative for both channels[1]. Background amplification in the no-template-control-panels never exceeded 2.6% of positive chambers for both channels (1.25 ± 0.75% SD for the terminase channel and 1.35 ± 0.7% SD for the SSU rRNA channel). Some background amplification using all-bacterial SSU rRNA primers is expected (*S1*) and is commonly attributed to DNA fragments present in commercial enzyme preparations (*S36*). The positive hits for the FAM channel in the microfluidic panels are expected to be

[1] One of the five SSU rRNA control chambers in array G was positive in a diagnostic post-amplification (not for sequencing), however this turned out to be an artifact of the diagnostic run as post-amplification of the same sample a second time was negative (with the positive control being positive).

a consequence of the modified TaqMan chemistry employed: since the universal LNA probe can spuriously bind to a terminase primer, primer-dimers will lead to amplification of a spurious product and fluorescence (similar to primer-dimers observed in SYBR Green assays), however no actual contaminating target is present, verified by agarose gel electrophoresis (see Fig. 2.10, Table 2.12, and supporting text for further discussion). Finally, every post-array amplification was always executed with several no-template-controls.

Measurement of PCR and cloning error rates

To measure the sequence error rate of samples retrieved from the microfluidic dPCR array, genomic DNA from ZAS-9 was used as a reference template in a microfluidic dPCR array. Vortexed genomic DNA from ZAS-9 was loaded onto a microfluidic dPCR array and cycled as described in the Methods section. Samples were then retrieved and the rRNA and terminase gene fragments were post-amplified using EXPAND high fidelity polymerase (Roche) as described in the Methods section. To measure the error rate, sequenced array retrievals were aligned against the known sequence of ZAS-9 rRNA and terminase genes. The error rate of the rRNA gene was 0 with 0 gaps (*n*=8, 905 ± 20bp SD) and the error rate of the terminase gene was 0 with 0 gaps (*n*=16, 711 ± 14bp SD). Post-amplification of the terminase gene fragment with the Quanta master mix resulted in a small number of ambiguous bases, however correcting these artifacts resulted in perfect matches. To test cloning associated errors, a retrieved ZAS-9 terminase sequence post-amplified with Roche high fidelity polymerase was cloned and several colonies were picked, amplified with the Roche high fidelity polymerase and sent for sequencing, as described in the Methods section. The measured error rate was 0.59 ± 0.29% SD (*n*=9, 759 ± 4bp SD) with 1 gap for 1 out of 9 cases. A similar cloning error rate was found when comparing the nucleotide sequences of 12 terminase amplicons in Fig. 2.2 sequenced directly from retrieved samples with their corresponding TOPO clones (0.55% ± 0.32% SD, *n*=12). In some cases single nucleotide deletions were also observed (see below). To check that clone errors were not sequencing related, five samples of the same terminase clone were amplified and sent for sequencing, however all sequences were found to be identical. To check that these errors are not introduced by *E. coli* during the growth phase, a single terminase colony was re-streaked and five colonies were amplified and sent for sequencing. All colonies yielded 100% identical sequences. Consequently, the origin of the terminase sequence errors appears to be the cloning step.

Out of 31 terminase sequences in Fig. 2.2, 10 were sequenced from the original retrieval, 12 were sequenced from a combination of the original retrieval and a TOPO clone, and 9 were sequenced from the TOPO clone alone. When sequences from the original retrieval were available and unambiguous, to minimize cloning errors these sequences were used in the consensus sequence in overlapping regions. Therefore for these sequences the error rate is expected to be lower. TOPO clones A9ii and E2i initially contained a frame shift mutation and E2i contained in addition an errant stop codon. These mutations were suspected to be cloning-related errors, confirmed by sequencing additional TOPO clones for each sample and calling base pairs by majority consensus. TOPO clone A11 also contained a frame shift mutation outside the alignment region considered in Fig. 2.2. This frame shift mutation also appears to be a cloning artifact as similar (though not identical) clones from the same retrieval did not contain this frame shift mutation. Consequently an N was inserted at this position. In the absence of TOPO clones, if an ambiguous base was declared (one such case) the degeneracy was arbitrarily broken to facilitate translation.

Measurement of primer efficiency

To measure SSU rRNA primer efficiency, five panels of a microfluidic dPCR array were loaded with ZAS-9 genomic DNA. Genomic DNA was titrated to achieve a final expected number of 400 (*n*=1), 300 (*n*=2), and 200 (*n*=2) SSU rRNA targets that were uniformly distributed across a panel containing 765 microfluidic chambers. Expected number of targets was estimated based on genomic DNA concentration measured using a Hoefer DynaQuant 200 fluorimeter. Digital PCR chemistry and cycling conditions were as described in the Methods section. The genomic DNA was vortexed upon extraction and therefore the genome is expected to be sheared to 10–20kb fragments. Since the two copies of the rRNA and terminase genes were located 689 kbs and 939 kbp apart, respectively, each genome was assumed to contribute two separate copies of each gene. After subtraction of noise, estimated from the no-template-control panels, the average rRNA and terminase primer efficiencies were calculated to be 59 ± 6% SD (*n*=5) and 74 ± 7% SD (*n*=5).

Selection pressure analysis

The program HyPhy 2.0 (*S37*) was used to estimate the relative rate of non-synonymous (β) and synonymous (α) substitutions ($\omega=\beta/\alpha$) for all 28 retrievals associated with Hosts I through IV using a maximum likelihood approach with a codon substitution model (*S38*). An alignment comprising 705 unambiguous nucleotides without gaps was used to generate a maximum likelihood (ML) tree with phyml assuming a TN93 (*S15*) nucleotide substitution model +Γ(nCat=4)+I+F. Given the above alignment and ML tree, HyPhy was used to find an optimal nucleotide substitution model out of all possible time-reversible models using the AIC criterion for selection. Finally, HyPhy was used to obtain the ML estimates of the independent model parameters of an MG94(*S39*)xREV_3X4(*S38*) substitution model with the optimal constraints found above (012032) assuming global parameters, the above ML tree, and the above in-frame alignment. Equilibrium frequencies were estimated from the partition. The global estimated ω was found to be 0.079. The 95% profile likelihood confidence interval was 0.071 to 0.088. This range is significantly lower than ω=1 (the case of neutral evolution) indicating that the terminase gene is under substantial negative selection pressure. A likelihood ratio test (LRT) comparing the null hypothesis model (ω=1) to the above alternative model strongly rejects the null hypothesis of neutral evolution with LR=754 and a P value (likelihood ratio test) predicted by HyPhy to be 0. In Table 2.7 the selection pressure was estimated for individual bacterial hosts using several additional methods and resulted in the same conclusion.

Analysis of viral genes in the metagenome

We were interested in finding the more abundant viral genes in the metagenome to identify a viral marker gene for this environment. In order to make this method widely accessible we designed an automated tool called MetaCAT that screens all gene objects in a metagenome and clusters them based on homology to genes in a reference database of known viral genes. The number of metagenome gene objects in a given cluster is then interpreted as the relative frequency of the corresponding known viral reference gene in the metagenome. This method is capable of assessing the relative frequency of viral-related metagenome gene objects in an annotation independent way. We refer to the implementation of this algorithm as the Metagenome Cluster Analysis Tool (MetaCAT), available upon request.

The MetaCAT algorithm is as follows: we first BLAST a list of known (viral) reference genes against all metagenome gene objects using BLAST v2.2.22+ (*S40*) (wrapped by Matlab) with a cutoff E value of 10^{-3}. As a reference list of known viral genes we use NCBI's viral RefSeq database v37 (*S41*). The number of metagenome gene objects homologous to each of the known reference genes is defined to be the *abundance* of that known reference gene in the metagenome. Since the list of known reference genes is long (~80,000 genes) we wished to filter this list based on several criteria. First, we retain only known reference genes whose best E value score is$\leq 10^{-7}$. This filtering step is performed to retain only known reference genes that yield reasonable alignments to metagenome gene objects. The second filtering step, implemented in Matlab, was designed to take out redundancy in the RefSeq database itself with respect to the metagenome using a dedicated clustering algorithm. For example, if two known reference genes are homologous to similar lists of metagenome gene objects, we would like to report only one of the two known reference genes, choosing the one with the lower E value. More generally, we wish to find for every known reference gene all the other known reference genes to which it is *related* (a known reference gene is always *related* to itself; see definition below). Therefore each known reference gene belongs to a *group* of *related* known reference genes. Finally, for each *group* of *related* known reference genes we only report the known reference gene with the lowest E value to represent that *group*. The combined list of reported known reference genes is then the final list of viral genes. The frequency of each reported viral gene is defined as the *abundance* of that known reference gene in the metagenome (see above). To complete the definitions: two known reference genes are said to be *related* if the *signatures* of both known reference genes is *similar*. A *signature* of a known reference gene is defined as the list of metagenome gene objects to which that known reference gene is homologous ($E \leq 10^{-3}$). Two signatures are then said to be *similar* if they share 50% of the elements in their lists. That is, if list A has L_i elements and list B has L_j elements, lists A and B are said to be similar if $50\% \geq 100 \cdot \min\left(L_i \cap L_j / L_i, L_i \cap L_j / L_j\right)$, with the symbol $\cap$ denoting the intersection between the two lists.

Note that the final reported known reference genes can still be *related*. Nevertheless, this filtering step is effective at removing a considerable amount of redundancy in the RefSeq database. A third manual filtering step is applied to retain only viral genes related to building a virion. Such genes are considered to be virus-specific genes (*S42*). Examples of such genes include capsid proteins, portal proteins, terminase proteins, tail proteins, baseplate proteins, and so on (*S42*). The list of the most abundant viral genes in the metagenome (*abundance* ≥10) is given in Table 2.2.

2.9.2 Supporting text

2.9.2.1 Statistical analysis of co-localization in digital PCR microfluidic arrays

Origin of a random co-localization component

We wish to see if *k* repeated co-localizations of a particular 16S rRNA ribotype with the terminase gene can be explained by chance co-localization on the microfluidic array (referred hereto as a "chip"). The reason there is a finite probability for chance co-localization is that typical array panels usually contain a certain fraction of FAM hits (the channel of the terminase marker) that are not co-localized with HEX hits (the channel of the 16S rRNA marker) as is shown in Fig. 2.9. If a fraction of these non-co-localized FAM hits contains the terminase target there is finite probability they may co-localize by random chance with a 16S rRNA gene and be mistaken for a true (host/terminase) co-localization. The number of these types of chance events determines the probability for false co-localization. Non co-localized FAM hits (which do not always contain an actual terminase product) can arise for several reasons:

(1) Since the universal LNA probe binds to a terminase primer, primer-dimers can lead to amplification and spurious fluorescence, i.e., fluorescence in the absence of a terminase target. These types of hits are apparent in the no-template-control panel and can account for roughly half of the non co-localized hits on a typical panel (see Table 2.12 and Table 2.14 discussed below). To verify that FAM hits in the no-template-control panel do not contain a target and are not the result of a contamination, four positive FAM chambers were retrieved from a no-template-control panel, post amplified for the terminase gene and analyzed by agarose gel electrophoresis, however no bands were detected. In addition, for several panels for two chips all FAM hits (both co-localized and non-co-localized) were retrieved, post amplified for the terminase gene and analyzed by agarose gel electrophoresis (Table 2.12). For each panel there were several samples that did not display any band (see Fig. 2.10 for a representative example), a finding that is consistent with the presence of spurious products observed in the no-template-control (NTC) panel. Furthermore, the average number of samples that did not display a band agreed well with the number of FAM hits in the no-template-control panels for these chips (Table 2.12), confirming that there is a noise component of spurious amplification on the panels similar to the no-template-control panel. For the seven chips in this study the average number of FAM hits in the no-template-control panel was 9.6±5.7. These types of non-co-localized FAM hits will not lead to chance co-localization with a 16S rRNA gene since there is no actual terminase target present.

(2) If the end-point fluorescence generated by a 16S rRNA target did not exceed the HEX threshold, this chamber would seemingly appear as a non-co-localized event (even though there is a 16S product present). Since the HEX threshold is set high enough to filter out cross-talk from the FAM channel into the HEX channel, some potential HEX hits may have been omitted. Indeed, when retrieving all FAM hits from a panel and amplifying all retrievals for the 16S rRNA gene, usually some wells whose HEX end point fluorescence did not pass the detection threshold did have a 16S rRNA band (data not shown). These types of non co-localized FAM hits should not contribute to false co-localization or contribute minimally because samples with mixed/chimera 16S rRNA traces are discarded from analysis and the probability of repeatedly amplifying the same wrong 16S rRNA is negligibly small (see discussion below).

(3) The 16S rRNA qPCR efficiency was measured to be ~60% for ZAS-9 genomic DNA (see Materials and methods). These types of events could potentially lead to false co-localization if a 16S rRNA amplification product is not generated (but the terminase gene in this cell was amplified) and this target co-localized by chance with another bacterial cell whose 16S rRNA gene was amplified. If an amplicon was generated (but for some reason fluorescence was inhibited) then these types of non-co-localized FAM hits will not contribute to false co-localization because samples with mixed 16S rRNA traces are discarded.

(4) Some cells may potentially prematurely lyse and their DNA may get sheared (for example when crushing the gut or during the loading process onto the chip). If this happens there is a possibility that free floating terminase targets are released into the mix.

(5) There may be assembled viruses present or free floating viral DNA, which can be regarded as free floating terminase targets.

As mentioned above, approximately half of the non co-localized FAM hits on a given panel can be explained by the spurious noise and do not contribute to random co-localization. Of the remaining non-co-localized FAM hits, the fraction relating to (2), if present, will not lead to false co-localization. Therefore the probability for false co-localization estimated below, which is based on fluorescence measurements alone, is an upper bound on the true probability for false co-localization.

Statistical model of random co-localization (P value estimation)

In Fig. 2.2 we see that certain 16S rRNA ribotypes are repeatedly co-localized, giving rise to 16S rRNA clades I–IV. The null hypothesis is that these 16S rRNA ribotypes are not true hosts and that the observed repeated co-localizations are due to chance associations, that is, these 16S rRNA ribotypes are simply co-localized many times by chance with free floating terminase targets. We therefore wish to estimate the probability (P value) that out of *n*=41 successful retrievals from the chip, i.e., retrievals that resulted in obtaining a 16S rRNA and terminase sequence after post-amplification, we will retrieve *k* or more instances of a particular ribotype ***S*** co-localized with a terminase (any terminase). This probability is given by

$$\begin{aligned}
&\text{Prob}\left(\text{number of chance co-localizations of } \boldsymbol{S} \text{ with a terminase } \geq k \,|\, n \text{ successful retrievals}\right) = \\
&= \sum_{k'=k}^{n} \text{Prob}\left(\text{number of chance co-localizations of } \boldsymbol{S} \text{ with a terminase } = k' \,|\, n \text{ successful retrievals }\right) = \\
&= 1 - \sum_{k'=0}^{k-1} \text{Prob}\left(\text{number of chance co-localizations of } \boldsymbol{S} \text{ with a terminase } = k' | \, n \text{ successful retrievals}\right) = \\
&= 1 - \sum_{k'=0}^{k-1} \text{Prob}\left(\text{succeed } \, k' \text{ times with probability } p_F | n \text{ trials}\right) = \\
&= 1 - \sum_{k'=0}^{k-1} \binom{n}{k'} p_F^{k'} \left(1 - p_F\right)^{n-k'} = 1 - \textbf{binocdf}(k-1, n, p_F)
\end{aligned}$$

where **binocdf** is the cumulative distribution function of the binomial distribution and p_F is the probability that when we successfully retrieve a co-localized well from a panel it contains the particular ribotype ***S*** and any terminase gene by pure chance. Given *k*, *n* and p_F (estimated below) the P value can be calculated. We find that the P values (*n*=41; one-tailed) for Hosts I–IV are all highly statistically significant ($P < 10^{-4}$; see Table 2.1 and Table 2.14) allowing us to reject the null hypothesis.

A model for a typical panel

Each panel loaded with a template is assumed to have the following species: Y HEX hits ("blue" hits), X FAM hits ("red" hits) out of which "*noise*" FAM hits are due to spurious amplification (no actual target). We assume that out of the X FAM hits there is a fraction of FAM hits that are free floating targets, that is a DNA fragment coding for a terminase gene but not for a 16S rRNA gene. The number of free floating targets is defined to be $X_T - noise$. These free floating targets would be the source of false co-localizations events. Thus co-localization events observed on the chip can be due to three possible causes: **(1)** genuine co-localization of a host SSU rRNA with its terminase, **(2)** chance co-localization of a free floating terminase gene with a 16S rRNA gene, **(3)** chance co-localization of a spurious FAM amplification (no actual terminase amplicon present) with an rRNA gene. See Table 2.13 for a definition of all the variables used in the model.

Estimation of p_F

To calculate the P value above, one must estimate p_F, i.e., the probability that a successful retrieval from a panel contains our particular ribotype ***S*** and any terminase gene by pure chance. This probability can be estimated as follows: let X_T be defined as the sum of the total number of free floating terminase targets and spurious targets leading to spurious FAM amplification (i.e., noise). We will see how to estimate X_T later on but for the time being let's assume it is given. The average number of free floating terminase targets to co-localize with a *particular* 16S rRNA ribotype ***S*** on a panel, defined as I_S, is given by multiplying the number of wells on a panel (765) by (a) the probability that a given well will contain a free floating terminase target p_{ter} and (b) the probability that that well will also contain ribotype ***S*.** The probability that a given well will contain a free floating terminase target is

(S1)
$$p_{ter} = \left(\frac{X_T - noise}{765} \right).$$

where *noise* is the number of FAM hits that are due to spurious amplification and are not associated with an actual terminase target. Thus $X_T - noise$ is the number of free floating terminase targets on the panel. Note that $X_T - noise$ will lead to an upper bound on the number of free floating terminase targets (leading to an upper bound on p_F) since $X_T - noise$ may include wells with a genuine 16S rRNA amplicon that simply did not pass the HEX detection threshold and are thus wrongly labeled as free-floating terminase targets (as described above). The value for *noise* can be estimated from the no-template-control panel for a given chip (see for example Table 2.12).

The average number of free floating terminase targets to co-localize with a *particular* 16S rRNA ribotype ***S*** on a panel is therefore given by

(S2a)
$$I_S = 765 \cdot p_{ter} \cdot \left(\frac{f_S \cdot Y}{765} \right) = p_{ter} \cdot f_S \cdot Y.$$

where *Y* is the total number of HEX hits on a panel, f_S is the frequency of ribotype ***S*** on the chip so that $f_S Y$ is the number of ribotypes ***S*** on a given panel. I_S is an estimate of the number of false co-localizations on a panel. This number is smaller than the number of observed co-localization on the panel, which we designate by *I* (=number of HEX and FAM intersections on a given panel). The number actual co-localizations on a panel of any 16S rRNA target with any terminase target (i.e., the total pool from which we draw successful retrievals) would be on average

(S2b)
$$I_{\text{all 16S-ter}} = I - \frac{noise \cdot Y}{765}.$$

taking out random co-localization of spurious FAM hits from *I*. The probability p_F is therefore given by the ratio of the number of random co-localization on a panel, I_S, and $I_{\text{all 16S-ter}}$, the number of actual co-localizations on the panel (i.e., of any 16S rRNA and any terminase target, both true and false co-localizations). Thus

(S3)
$$p_F = \frac{I_{\text{S}}}{I_{\text{all 16S-ter}}} = f_{\text{S}} \cdot \frac{p_{ter} \cdot Y}{I_{\text{all 16S-ter}}}.$$

Since $p_{ter} \cdot Y / I_{\text{all 16S-ter}}$ can vary somewhat from panel to panel, to calculate p_F we use Bayes' theorem:

$$p_F = P(\text{false}\,|\,\text{panel A})P(\text{panel A}) + P(\text{false}\,|\,\text{panel B})P(\text{panel B}) + \ldots$$

We therefore replace $p_{ter} \cdot Y / I_{\text{all 16S-ter}}$ in Eq. S3 by its panel averaged value, weighted by the number of times each panel was sampled (making at total of *n*=41 trials). The estimated values of p_F per host type are given in Table 2.14.

Estimation of X_T

Let us assume that a given panel has X FAM hits, Y HEX hits, and I intersections. The number of non-co-localized terminase hits is then $X_f = X - I$. X_T is slightly larger than X_f since some of the free floating targets or spurious targets may have co-localized with HEX hits. This difference ($X_T - X_f$) is estimated by multiplying the number of wells on a panel by (a) the probability that a well will contain a free floating target *or* a spurious target and (b) the probability that that well will contain *any* HEX hit. Thus $X_T - X_f = (765 \text{ wells})\left(\frac{X_T}{765}\right)\left(\frac{Y}{765}\right)$, or

$$X_T = X_f + (765 \text{ wells})\left(\frac{X_T}{765}\right)\left(\frac{Y}{765}\right).$$

Solving for X_T we find that

(S4) $$X_T = X_f\left(1 - \frac{Y}{765}\right)^{-1} = (X - I)\left(1 - \frac{Y}{765}\right)^{-1}.$$

Note that since typically Y~50, $X_T \approx X - I$.

Estimation of f_s

f_s, the frequency of ribotypes ***S*** on the chip, is estimated based on the number of the particular REP ribotypes that grouped with the corresponding host ***S*** (e.g., five REP4 ribotypes out of 118 grouped with Host I in Fig. 2.7, therefore f_s=5/118). Operational taxonomical units for REP/host clades were determined by a DOTUR analysis (Table 2.6 and Fig. 2.7).

Given f_S and X_T (Eq. S4) we can calculate p_F (Eq. S3), and given k (Table 2.1) we can calculate the P value. Table 2.14 summarizes the frequencies f_S, probabilities p_F and P values for Hosts I though IV. As mentioned in the beginning of this section, the P values calculated for Hosts I through IV were very small ($P < 10^{-4}$) allowing us to reject the null hypothesis, i.e., the repeated ribotypes I–IV cannot be explained by random co-localization of these ribotypes with free floating terminase targets.

Bound on false co-localization in the dataset

We would like to estimate the average number of retrievals where one of the observed hosts co-localized by chance with a terminase (resulting in either two terminases — the host's and the free floating terminase, or, in the case the host's terminase did not amplify or was not present, one wrong terminase). The probability that we retrieve from a given panel any of the host ribotypes with the wrong terminase is given by summing the individual false co-localization probabilities for each host -

$p_{F,tot} = \sum_{\text{host I-IV}} p_F = \left(p_{ter} \cdot \sum_{\text{host I-IV}} f_S \cdot Y \right) \Big/ I_{\text{all 16S-ter}}$. The average number of false co-localizations in a dataset of n=41 retrievals would therefore be

(S5) $$N_{false} = p_{F,tot} \cdot n.$$

We find that N_{false} =0.6. Thus out of 28 repeated co-localizations of our hosts, on average ~0.6 are expected to be false (an error of 2%). The fact that no co-localized pairs were retrieved with the most abundant phylotypes on the array (see Table 2.6 and Fig. 2.7) and that the three most abundant phylotypes on the array comprising 49% of all treponemes in only one out of 38 cases co-localized with an rRNA gene (see discussion on non-hosts below) confirms that erroneous co-localization was indeed very rare.

Numerical simulation to test the statistical model

To check our statistical analysis (Eq. S1-S7) we conducted a Monte Carlo simulation of retrieval from the microfluidic panels based on the model presented above (Fig. 2.11). The numerical simulation results were predicted precisely by the statistical model described above.

Model for Monte Carlo simulation

In the simulation Y rRNA templates were loaded randomly onto a panel of 765 chambers (Y~U[Y_{min},Y_{max}]). Each panel was also randomly loaded with *noise* spurious FAM hits (*noise*~U[$noise_{min}$,$noise_{max}$]) and *free* free floating terminase targets (*free*~U[$free_{min}$, $free_{max}$]). A fraction f (i.e., probability) of the Y rRNA templates was assumed to be genuine hosts (i.e., hosts that genuinely harbor a terminase gene). The terminase gene within these hosts was assumed to be amplified with probability e_{ter}. Each retrieval trial consisted of loading a single panel of 765 chambers with the above elements and retrieving one sample that contained both a 16S rRNA sequence and a terminase sequence. If the retrieval failed (i.e., the rRNA was co-localized with a spurious FAM target) a new retrieval trial would be attempted until successful (these mute trials would not be counted as successful iterations). For each successful retrieval trial it was registered if the retrieval was a false co-localization (i.e., a host 16S rRNA sequence was co-localized with a free floating terminase). In addition for each successful retrieval trial the probability of false co-localization p_F was calculated. This probability is given by the ratio of number of false-co-localizations on the panel (i.e., a 16S rRNA gene that co-localized with a free-floating terminase) and the total number of co-localization on the panel (any 16S and any terminase gene). A single Monte Carlo iteration ended when $N_{retrievals}$ (=41) successful retrievals were obtained. At the end of each Monte Carlo iteration, the total number of false co-localizations (N_{false}) was tallied and the average value for p_F was calculated. In total there were 1000 Monte Carlo iterations.

To compare with the statistical model above, after each Monte Carlo iteration, p_F and N_{false} were estimated based on Eq. S3 and Eq. S5 assuming f=f_S and given the random values for X, Y, I and *noise* generated for each of the 41 panels in the simulation. At the end of the simulation the average value of p_F and N_{false} (averaged over 1000 iterations) was compared to the predicted values of p_F and N_{false} based on the statistical analysis.

Simulation parameters

Simulation parameters were chosen to mimic the experiments in this study as closely as possible: $N_{retrievals}$=41; all hosts were assumed to be indistinguishable so that f_S was given as the sum of all the rates f_S in Table 2.14 (i.e., f_S=9/118, where 9 is the total number of occurrences of Hosts I–IV phylotypes in the reference library, and 118 is the size of the reference library — see Table 2.1). All other parameters followed the distributions in Table 2.14 with $Y \sim U(20,80)$, $noise \sim U(5,15)$, $free \sim U(0,20)$ and e_{ter}=0.74 (see Materials and Methods).

Simulation results

We found that the predictions for p_F (Eq. S3) and N_{false} (Eq. S5) closely matched the numerical simulation:

$$\begin{cases} p_F\text{(simulation)=0.014}\pm 0.011 \\ \hat{p}_F\text{(Eq. S3)=0.018}\pm 0.022 \end{cases} \quad \begin{cases} N_{false}\text{(simulation)=0.6}\pm 0.8 \\ \hat{N}_{false}\text{(Eq. S5)=0.7}\pm 0.9 \end{cases}$$

The errors are standard deviations. The simulation presented here shows that the statistical model presented above (Eq. S1-S7) is consistent with the numerical simulations.

Chambers with multiple cells

Since the average number of targets loaded per panel was small (~50), the chance of obtaining multiple cells in a given chamber was small (1.7 chambers out of 50 on average (*S43*)). However cells can also potentially "stick" together upon loading as well. If a chamber contains multiple 16S rRNA genes and more than one gene is amplified then the sequence trace will be mixed. Such samples were automatically discarded in this study. If a 16S rRNA chimera is formed, chimera products are screened with Pintail (*S8*) and Bellerophon (*S9*) and discarded from further analysis (no such chimeras were found in this study). The chance however that the same ribotypes would repeatedly co-localize and either form a chimera or amplify the wrong rRNA gene are extremely small. To estimate the chance for such an event, we shall consider the case where the host 16S rRNA gene, ***S***, repeatedly co-localized with the same rRNA gene ***S'***, and that the foreign 16S rRNA gene (***S'***) was amplified while the host 16S rRNA gene (***S***) was not amplified. The average number of such chance events per panel where the host terminase was also amplified is given by $I_{SS'} = \varepsilon_{ter}\varepsilon_{16S}(1-\varepsilon_{16S})\left(f_sY\right)\left(f_{s'}Y\right)/765$, where ε_{ter} and ε_{16S} are the amplification efficiencies of the terminase gene and the 16S rRNA gene, respectively (see Materials and methods for an estimation of these efficiencies), $f_{s'}$ is the frequency of the ***S'*** ribotype, and $\left(f_sY\right)\left(f_{s'}Y\right)/765$ is the number of chance co-localizations of ***S*** and ***S'*** cell types on a given panel. The probability therefore of retrieving such events is $p_F^{mixed} = I_{SS'}/I_{\text{all 16S-ter}}$. Assuming $f_{s'} \sim 0.2$ (corresponding to the worst case scenario of co-localizing with the most frequent ribotype on the chip, REP1) then based on Table 2.14 we have $p_F^{mixed} << p_F$ (where p_F is given in Eq. S3) and therefore these events can be neglected (the P values for such events would be much smaller than those in Table 2.1).

Uniformity of panel loading

On a few occasions, panels were loaded by the NanoFlex somewhat nonuniformly. This has the consequence of reducing the effective number of wells available for the cells. The samples affected for Host I were C2 and G1. The terminases of samples C2 and G1 fell in the main clade of Host I of highly similar terminases (Clade V1 in Fig. 2.2) lending support for these co-localizations. Sample G2 (Host III) was taken from a slightly nonuniform panel, however the terminase of sample G2 was 100% identical at the amino acid level (235 aa alignment) to F2 also associated with Host III, lending support for this co-localization. Samples affected for Host II were A4 and A7, however the terminase of A4 was 99.6% identical at the amino acids level (235 aa alignment) to the terminase of A9i also of Host II, lending support for this sample. The terminase of A7 was 95.3% identical at the amino acids level to the terminase of A13i also of Host II, lending support for this sample.

Estimation of the P value for putative *Treponema* non-host (REPs1-3)

The phylotypes REP1, REP2, and REP3 were highly repeated in the random rRNA reference library ($f_{\mathrm{S}} = 23/118, 8/118, 7/118$, respectively) but were never sampled in the co-localization library (*n*=41). The null hypothesis is therefore that ribotype ***S*** is a genuine host but was not sampled *n*=41 times by chance. We wish to calculate the probability for this event. The fraction of co-localizations in a given panel that contain host ***S*** is given on average by

(S6) $$p_S = \frac{\varepsilon_{ter} \cdot f_{\mathrm{S}} \cdot Y}{I_{\text{all 16S-ter}}}.$$

where ε_{ter} is the efficiency of amplification for the terminase gene (see Materials and methods), f_{S} the frequency of host ***S*** on the chip, *Y* the number of 16S rRNA hits on a given panel, and $I_{\text{all 16S-ter}}$ is the number co-localizations on a panel of a 16S rRNA target with an actual terminase target (Eq. S2b). Therefore $\varepsilon_{ter} \cdot f_{\mathrm{S}} \cdot Y$ is the number of expected genuine co-localizations for ribotype ***S***, and $\varepsilon_{ter} \cdot f_{\mathrm{S}} \cdot Y / I_{\text{all 16S-ter}}$ would be the probability to sample this co-localization. The probability (P value, one tailed, *n*=41) for not retrieving ***S*** (*k*=0) after *n*=41 trials is given by

$$\text{P value} = \text{Prob}\left(k = 0 \,|\, n = 41 \text{ successful retrievals}\right) = (1 - p_S)^n$$

where p_S is averaged using Bayes' theorem as described above (i.e., a panel-weighed average based on Table 2.14 for all 41 retrievals). For $\varepsilon_{ter} \approx 0.8$ (measured value) we find that the P value (one tailed test with n=41) for not retrieving a host with a frequency of $f_S \geq 7/118$ is $\leq 4.8 \cdot 10^{-20}$ allowing us to reject this hypothesis. If REPs-1, 2, and 3 are infected in only >5%, 14%, and 16% of the cases respectively, then the P value for not retrieving these infected strains is 0.01 (one tailed test with n=41). Therefore based on statistical grounds we conclude that the majority of REP1–3 cells are not infected. Furthermore 21 out of 23 REP-1 ribotypes, 8 out of 8 REP-2 ribotypes, and 7 out of 7 REP-3 ribotypes were not associated with a terminase hit on the microfluidic chips. Of the two positive hits for REP-1, post-amplification followed by agarose gel electrophoresis showed that just one of these samples contained a terminase target. Statistically, out of n=38 occurrences of REPs1-3, $p_{ter} \cdot n$ should randomly co-localize with a terminase target on the chip, or 0.4±0.2 random co-localizations, as observed. This is consistent with the hypothesis that REPs1-3 are indeed non-hosts.

2.9.2.2 The viral marker gene and its genetic context

Requirements for a viral marker gene

Since certain viral genes can be of bacterial origin, and some viral genes may not be associated with an actual functional virus, a genuine viral marker should satisfy certain requirements (*S42*). We were therefore interested in choosing as a viral marker a gene that (a) was unique to viruses, (b) was present in a larger viral context, (c) was prevalent in the ecosystem we were investigating, (d) contained multiple conserved regions that could be used to design degenerate primers, and (e) is active or has been active in recent evolutionary history in this system. The large terminase subunit chosen as a viral marker gene fulfilled all of the above requirements:

(1) The large terminase subunit is considered to be one of the most universally conserved phage genes and best phage identifiers (*S42*), exhibiting certain conserved residues and motifs (see Figs.

2.5 and 2.6). Furthermore, since typically different phages exhibit little overall sequence similarly (see main text), the terminase gene also appears to be system specific (*S44*), thereby potentially serving as a good differentiating marker (*S45*).

(2) Bioinformatic analysis of the ZAS-2 and ZAS-9 genomes revealed four prophage-like elements (two in each genome) that were related to tailed phages based on their sequence homology. The largest of these elements (ZAS-2A) spanned 43.5 kb, which is a typical size for tailed phages (*S46*). Furthermore, all four copies of the terminase gene in the ZAS genomes had homologs in the higher termite metagenome with 77–79% amino acid identity. The largest of these elements, ZAS-2A, appeared to be associated with the *Caudovirales* order: When BLASTing each of the 41 identified genes in this prophage-like element against NCBI's viral RefSeq (v37) database, 16 genes had significant hits ($E < 0.005$), with 15 out of the 16 genes being associated with homologs present in viruses belonging to the *Caudovirales* order. The viral genes also follow a typical tailed-phage gene organization pattern (*S47*). For example genes ZA3, ZA4, ZA5, ZA7, ZA8 are the head related genes (homologous to the small and large terminase subunit genes, portal protein gene, prohead protease gene, and capsid protein gene, respectively), whereas genes ZA32 and ZA33 towards the end of the cassette exhibited a weak homology to a tail fiber gene and a tail tape measure protein gene, respectively ($E = 0.16$, 0.29, respectively). Among the 15 hits above, 11 were associated with the *Siphoviridae* family, two with the *Podoviridae* and two with the *Myoviridae* family. The last four genes appear to be less diagnostic than the *Siphoviridae* related genes as they are not signature phage genes and the E value for three of these genes was low ($E \geq 0.001$). Although it is possible that the prophage-like elements are mosaics of

Caudovirales families (*S48*), based on the above analysis it appears that these elements are mostly closely related to the *Siphoviridae* family.

(3) Bioinfomratic analysis of the metagenome (Table 2.2) identified the large terminase subunit as one of the most abundant viral-unique genes in the metagenome (though this may not reflect absolute abundance in the sample due to assembler bias). In addition, more generally, the ZAS prophage-like elements appear to be ubiquitous to the termite environment as certain cassettes within the ZAS prophage-like elements were found to be abundant in the higher termite metagenome. For example, the large terminase subunit and its adjacent portal protein from ZAS-2A had a maximum percent amino acid identity of 78% and 70%, respectively, when BLASTed against the metagenome (Table 2.3) and were homologous to 46 and 43 metagenome gene objects respectively, (E ≤ 1e-5). Furthermore, these two genes, that are adjacent to each other in the ZAS genomes (a typical organization in viruses (*S42*)) were also found to be next to each other in the metagenome contigs.

(4) Alignment of the terminase alleles from the ZAS genomes and the higher termite metagenome revealed multiple conserved regions that could be used for primer design (Fig. 2.5).

(5) Viral-specific genes encoded by ZAS-2 and ZAS-9 prophage-like elements (the portal protein, the capsid protein, the large terminase subunit and the prohead protease protein) exhibited substantial negative selection pressure (data not shown). In addition, the terminase genes retrieved from *R. hesperus* specimens also exhibited substantial negative selection pressure (see Materials and methods and Table 2.7). This evidence suggests that the terminase gene in the termite system if not functional, has been functional in recent evolutionary history (see discussion below). In addition, there is some anecdotal evidence suggesting the terminase is part of an active viral entity. In one of the earlier experiments with the microfluidic arrays (prior to execution of arrays A through G from which samples were retrieved), where chilling of samples to 4°C was not strictly enforced, a dilution series of a *Zootermopsis nevadensis* termite hindgut fluid was loaded onto a microfluidic array. The panel on the array corresponding to the largest gut dilution exhibited 34.9 times the number of expected terminase hits (384 observed verus 11 expected), where the expected number of hits was estimated based on the number of hits from more concentrated dilutions loaded onto the same microfluidic array. At the same time, the rRNA channel displayed the expected number of hits (72 observed versus 74 expected) for this dilution. Since the degenerate terminase primers that were used in the qPCR chemistry were designed based on the terminase alleles in the ZAS-2 and ZAS-9 prophage-like elements (among other alleles), this induction event is specific to the terminase gene investigated in this study. This result indicates that a lytic event associated with the prophage-like element may have taken place in the tube containing the largest gut dilution, suggesting that this putative prophage is functional. We note that earlier experiments to induce the ZAS-2 and ZAS-9 cultures using mitomycin C were not successful, suggesting that mitomycin C may not be the inducing agent of this element.

Functionality of the terminase gene

Given the fact that the terminase gene is under negative selection pressure and in the absence of obvious frame shift mutations or errant stop codons in the alignment, there are several options regarding the nature of the prophage-like element in which it resides and the functionality of the terminase gene within these elements: **(1)** the terminase is part of an active prophage (for which there is some evidence, as discussed in point 5 above) **(2)** the terminase is part of a defective prophage but it remained functional because there was not enough time for point mutations to have accumulated. This can happen because "prophage-debilitating deletions can accumulate more rapidly than gene-inactivating point mutations" (*S42*). **(3)** The prophage indeed decayed and the terminase gene degraded over time, but was subsequently repaired by a recombination event with another phage that was likely functional (since it infected the cell in the first place)(*S42*). Finally, **(4)** the terminase was recruited by the bacterium because it confers on the bacterium some competitive advantage and is therefore under negative selection pressure.

To further elaborate on the last point (4), phage genes that are adopted by the cell are typically lysogenic conversion genes (*S42*) — genes that change the phenotype of the cell and confer some selective advantage to the cell. In this context, known possibilities may be (*S42*) tail-like bacteriocins and genetic transfer agents (GTAs). Bacteriocins are devices that kill other bacteria and some bacteria can produce bacteriocins that resemble phage tails (*S42, S49*). However since these entities do not have heads or package DNA it seems unlikely they would encode a terminase gene. For example, type F and type R tail-like bacteriocins of *Pseudomonas aeruginosa* PAO1 do not appear to encode a terminase gene or any other head related proteins (*S50-S51*). GTAs are tailed phage-like particles that encapsidate random fragments of the bacterial genome and can transfer them to other bacteria of the same species (*S42*). GTAs are thought to be adopted by the host cell to facilitate genetic exchange under the control of the host (*S52-S54*). The GTA coding region is typically short (~14–16 kb (*S54*)) and appears to contain the genes required for assembly of the GTA head and tail structures and the genes required for DNA packaging (including a terminase gene) (*S52, S54*). Phage DNA-specific replication functions and phage DNA-specific integration or excision functions are in principle not required by the GTA (*S52*). Although it cannot be ruled out that the terminase genes retrieved from *R. hesperus* are part of a GTA, this possibility appears to be unlikely since the predicted prophage-like element identified in ZAS-2 spans ~43.5 kb (a typical length for a functional phage), which is much longer than a typical GTA length (14–16 kb — see above). In addition, unlike GTAs, the ZAS-2 prophage-like element encodes both integration genes and several DNA replication machinery genes.

To summarize, the fact that the *R. hesperus* terminase alleles are under substantial negative selection pressure suggests that this terminase is either active or has been active in recent evolutionary history and was the direct or indirect result of a viral infection (options 1, 2, or 3 above). The possibility that the terminase was adopted by the cell and is part of a GTA appears to be unlikely. Thus the associations between the hosts and the terminase genes revealed by the microfluidic assay should be a valid proxy for interaction of these hosts with genuine infecting phages, reflecting either current or recent infections.

2.9.3 Supporting figures

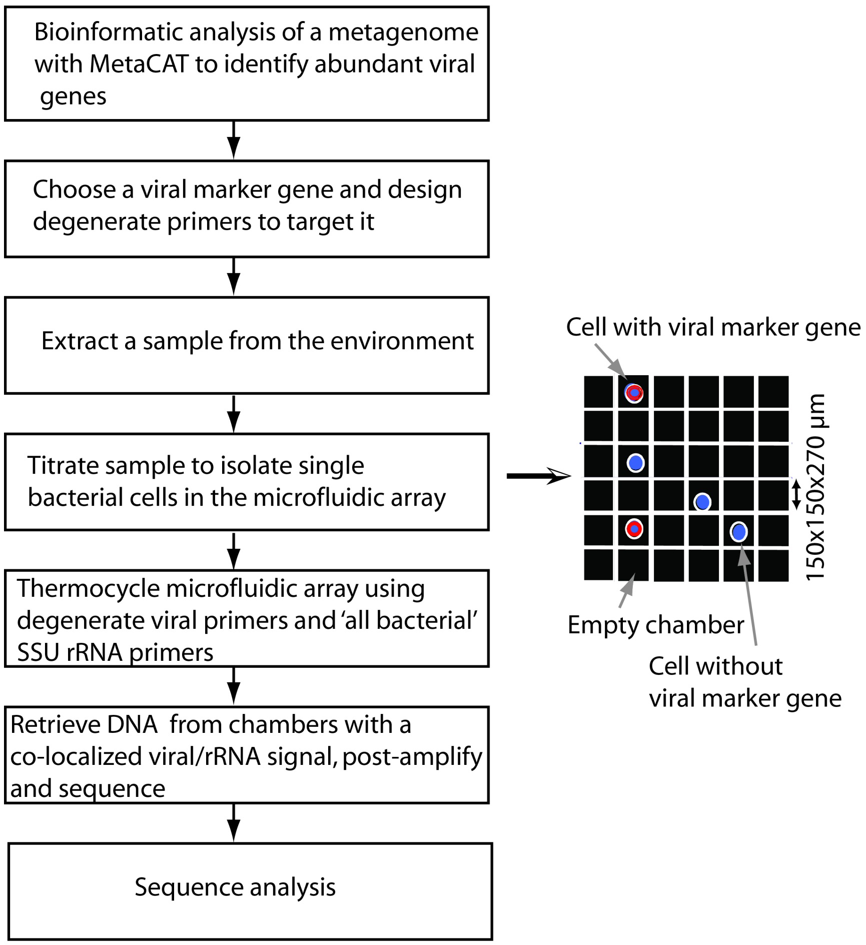

Bioinformatic analysis of a metagenome with MetaCAT to identify abundant viral genes
Choose a viral marker gene and design degenerate primers to target it
Extract a sample from the environment
Titrate sample to isolate single bacterial cells in the microfluidic array
Thermocycle microfluidic array using degenerate viral primers and 'all bacterial' SSU rRNA primers
Retrieve DNA from chambers with a co-localized viral/rRNA signal, post-amplify and sequence
Sequence analysis
Cell with viral marker gene
150x150x270 µm
Empty chamber
Cell without viral marker gene

Figure 2.4. Workflow using the microfluidic digital PCR array for host-virus co-localization in a novel environmental sample. See Materials and methods for further details.

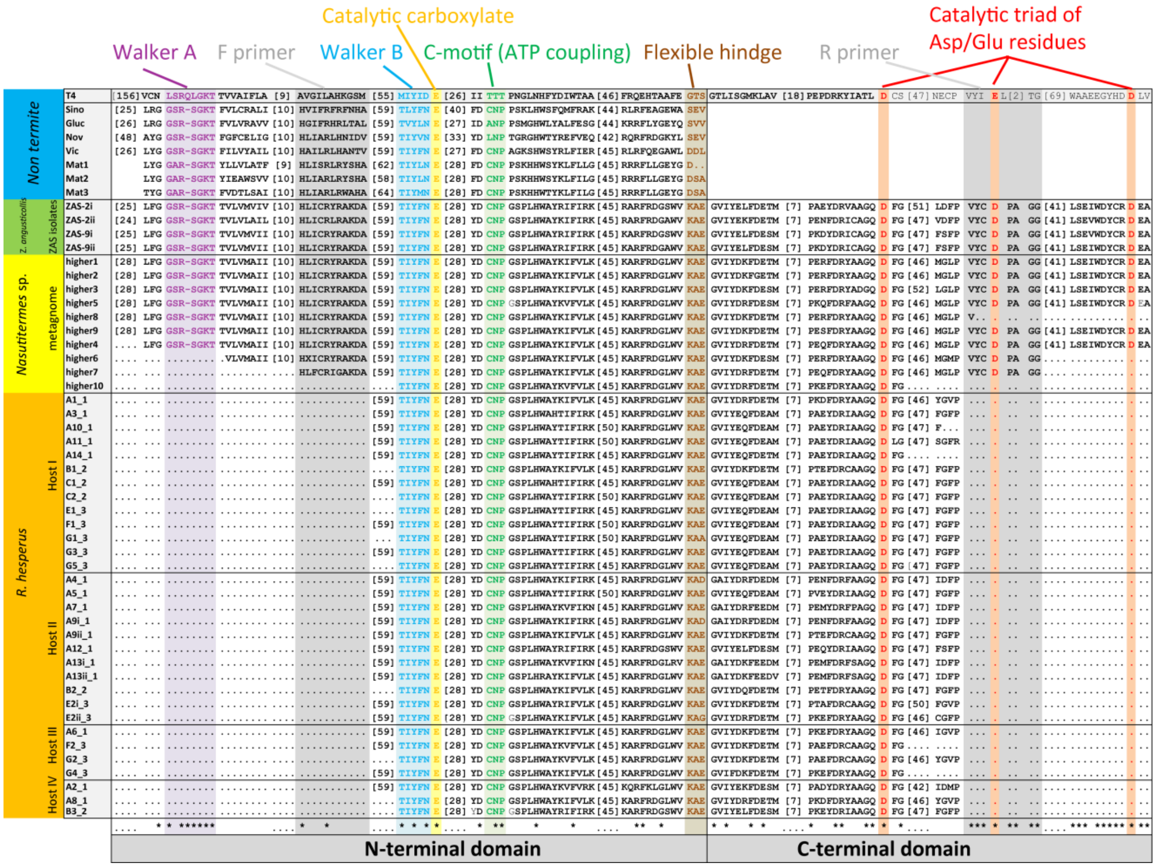


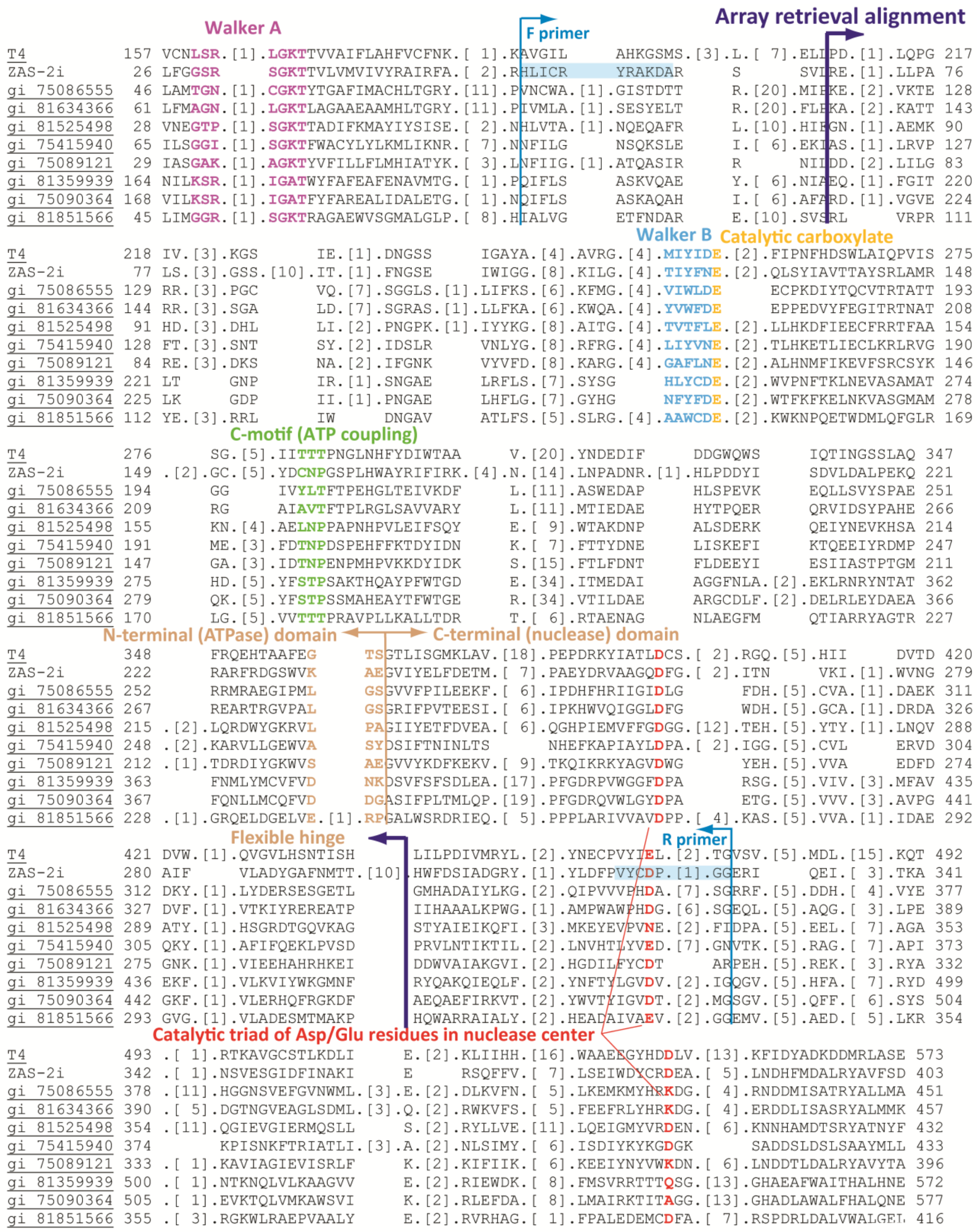
Figure 2.5. Multiple alignment of termite related terminase sequences and closest homologs. Here we show a multiple alignment of terminase genes of both termite and non-termite origin highlighting putative functional motifs. Terminase sequences included are (1) terminase sequences retrieved from *R. hesperus* termites using the digital PCR, (2) homologous terminases from the metagenome of a *Nasutitermes* sp. termite, (3) homologous terminases from *Treponema* isolates obtained from a *Z. angusticollis* termite, and (4) homologous terminases from non-termite related bacteria found in public databases (NCBI's protein RefSeq database and the Joint Genome Institute database). Also highlighted are putative conserved functional motifs for the N-terminal ATPase center and the C-terminal nuclease center (see Fig. 2.6). When searching for homologs for the ZAS2-i terminase gene in public databases, the N-terminal ATPase domain of this gene (amino acids 1-234 — see Fig. 2.6) appeared to be much more conserved (47% identity) than the entire gene (29% identity). Consistent with this fact, the ATPase domain of the large terminase subunit has been shown to be conserved in a wide variety of dsDNA (*S55*) viruses and even shows certain conserved motifs with the putative herpesvirus terminase (*S55-S56*) suggesting it is an ancient viral domain (*S55, S57-S58*). We therefore show here only the N-terminal domain alignment of non-termite homologous terminases.

N-terminal alignment: The boundary of the N-terminal domain for the terminase alleles was determined based on its location in T4 (residue 360)(*S59*) by aligning the amino acid sequences of the ZAS2-i terminase and all non-termite related terminases with RPS-BLAST against pfam03237 (*S59*) in the CDD (*S60*) (see Fig. 2.6 for ZAS2-i alignment). The N-terminal domain of other termite related sequences was then determined by a MUSCLE alignment to the ZAS2-i terminase (*S61*). All N-terminal domains were then MUSCLE aligned. **C-terminal alignment:** maximum length termite related terminases were MUSCLE aligned and then only their C-terminal regions were juxtaposed to the N-terminus alignment found above (the overlap with the N-terminus alignment was identical).

Functional motifs were identified based on an RPS-BLAST alignment of ZAS-2i against pfam03237 (Fig. 2.6). This figure demonstrates that the termite related terminase sequences exhibit terminase-like functional motifs. Putative functional motifs include (1) Walker A motif G/A-XXXXGK(T/S) (purple) with a single residue X deletion, (2) Walker B ZZZZD motif with D replaced by N — a relatively common substitution for this residue (blue), (3) catalytic carboxylate group motif — E (orange), (4) putative ATP coupling motif (green), and (5) catalytic Asp/Glu triad motif — here a conserved D (red)(*S62-S63*). Also highlighted is the putative flexible hinge motif (brown)(*S63*) based on the RPS-BLAST alignment. Numbers in brackets correspond to aligned residues not shown. Stars indicate conserved residues excluding T4. Dots indicate end of available sequence. X residues in the higher termite sequences are due to ambiguous base pairs in the nucleotide sequence. The RPS-BLAST ZAS2-i alignment with T4 (Fig. 2.6) was superimposed to guide the eye and was not part of the MUSCLE alignment. Also shown are the primer binding sites. The degenerate core region of the CODEHOP primers (*S34*) that is required to be conserved consists of 4 amino acids at the 3' end of the primer. Out of the 50 ZAS and higher termite gut alleles, 31 alleles included the forward primer motif and 26 alleles included the reverse primer motif. In all cases, the degenerate core region of the primers was strictly conserved. In one additional allele, the sequence began from the center Asp residue in the conserved catalytic Asp/Glu triad motif. This residue was mutated in this allele from an Asp residue to a Gly residue suggesting this partial allele encodes a nonfunctional terminase. Thus, all functional alleles of the terminase gene exhibited a strictly conserved degenerate core region. Note that the Walker A motif was not chosen for a forward primer binding site due to the high degeneracy involved with this amino acid sequence.

To check what diversity of terminase genes are expected to be amplified, we BLASTed the core region of the forward (ter7F) and reverse (ter5eR) terminase primers against all viral genes in NCBI's viral RefSeq database v37. Only the core region of the primer was used in the BLAST analysis (a more general search) because the primers are CODEHOP primers and therefore while the degenerate core region (11–12 bases in the 3' region of the primer) must base pair with the target, homology of the clamp region is less critical for initial amplification. We then crossed the list of hits for the forward and reverse primers searching for mutual hits present in the same gene within the same bacteriophages, however no such solutions were found. Based on this result we anticipate that the degenerate terminase primers target the unique diversity of terminase genes currently known to exist only in termite and possibly related insect species.

Non-termite related terminases (Vic, Sino, Gluc, and Nov) are gram negative isolates belonging to the Lentisphaerae and Proteobacteria phyla. These bacteria grow in a variety of habitats (human gut, soil, fresh water, plants, etc.) and can either be free living or symbiotic, anaerobic or aerobic. Mat1, Mat2, and Mat3 were found to be present in the metagenome of a hypersaline microbial mat from Mexico (see Table 2.11 for accession numbers).

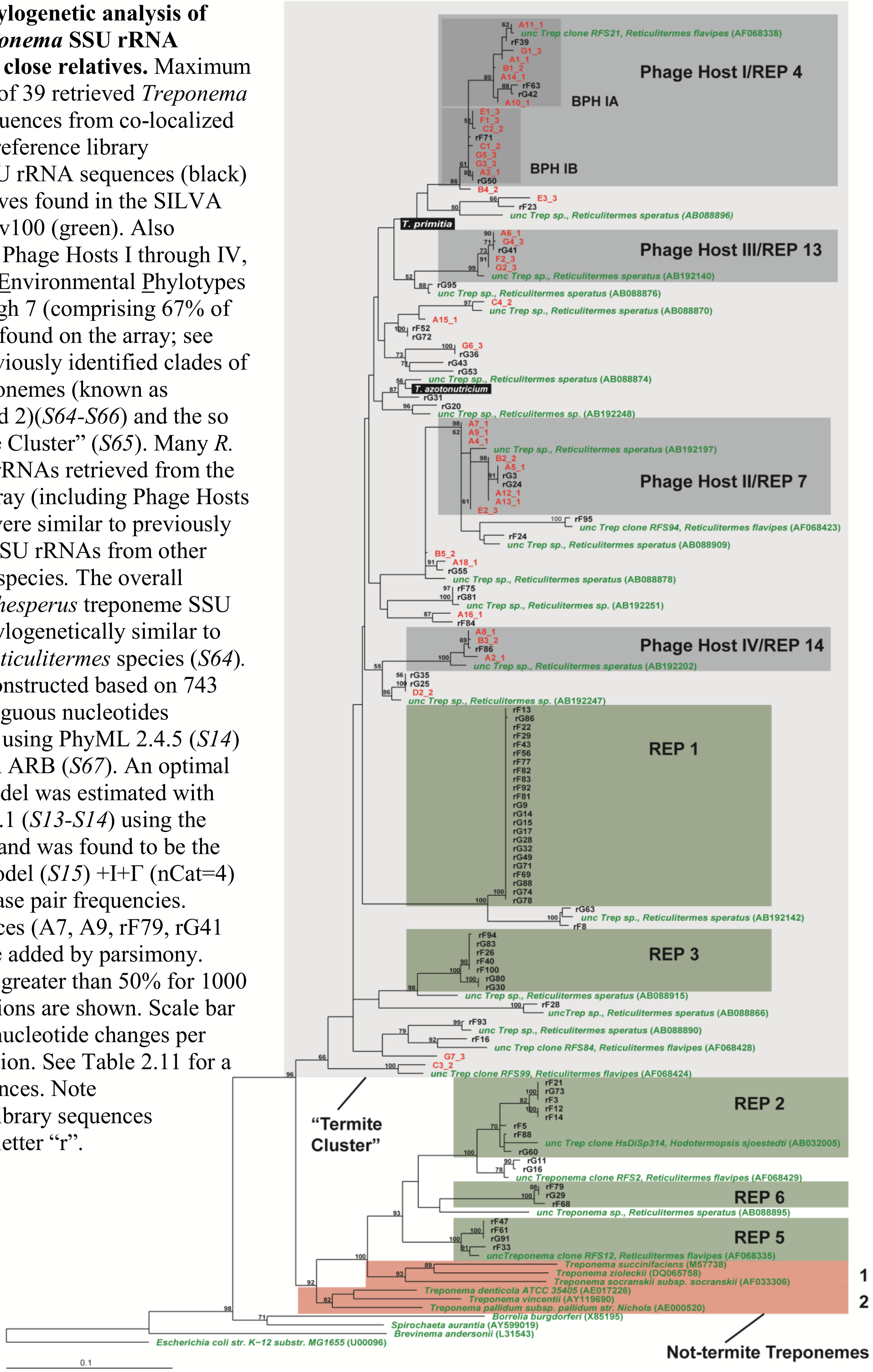

Walker A
F primer
Array retrieval alignment
Walker B
Catalytic carboxylate
C-motif (ATP coupling)
N-terminal (ATPase) domain
C-terminal (nuclease) domain
Flexible hinge
R primer
Catalytic triad of Asp/Glu residues in nuclease center

```
T4          157 VCNLSR.[1].LGKTTVVAIFLAHFVCFNK.[ 1].KAVGIL     AHKGSMS.[3].L.[ 7].ELIPD.[1].LQPG 217
ZAS-2i       26 LFGGSR     SGKTTVLVMVIVYRAIRFA.[ 2].RHLICR     YRAKDAR     S       SVIRE.[1].LLPA 76
gi 75086555  46 LAMTGN.[1].CGKTYTGAFIMACHLTGRY.[11].PVNCWA.[1].GISTDTT     R.[20].MIEKE.[2].VKTE 128
gi 81634366  61 LFMAGN.[1].LGKTLAGAAEAAMHLTGRY.[11].PIVMLA.[1].SESYELT     R.[20].FLEKA.[2].KATT 143
gi 81525498  28 VNEGTP.[1].SGKTTADIFKMAYIYSISE.[ 2].NHLVTA.[1].NQEQAFR     L.[10].HIEGN.[1].AEMK 90
gi 75415940  65 ILSGGI.[1].SGKTFWACYLYLKMLIKNR.[ 7].NNFILG     NSQKSLE     I.[ 6].EKIAS.[1].LRVP 127
gi 75089121  29 IASGAK.[1].AGKTYVFILLFLMHIATYK.[ 3].LNFIIG.[1].ATQASIR     R       NIIDD.[2].LILG 83
gi 81359939 164 NILKSR.[1].IGATWYFAFEAFENAVMTG.[ 1].PQIFLS     ASKVQAE     Y.[ 6].NIAEQ.[1].FGIT 220
gi 75090364 168 VILKSR.[1].IGATFYFAREALIDALETG.[ 1].NQIFLS     ASKAQAH     I.[ 6].AFARD.[1].VGVE 224
gi 81851566  45 LIMGGR.[1].SGKTRAGAEWVSGMALGLP.[ 8].HIALVG     ETFNDAR     E.[10].SVSRL      VRPR 111

T4          218 IV.[3].KGS      IE.[1].DNGSS     IGAYA.[4].AVRG.[4].MIYIDE.[2].FIPNFHDSWLAIQPVIS 275
ZAS-2i       77 LS.[3].GSS.[10].IT.[1].FNGSE     IWIGG.[8].KILG.[4].TIYFNE.[2].QLSYIAVTTAYSRLAMR 148
gi 75086555 129 RR.[3].PGC      VQ.[7].SGGLS.[1].LIFKS.[6].KFMG.[4].VIWLDE     ECPKDIYTQCVTRTATT 193
gi 81634366 144 RR.[3].SGA      LD.[7].SGRAS.[1].LLFKA.[6].KWQA.[4].YVWFDE     EPPEDVYFEGITRTNAT 208
gi 81525498  91 HD.[3].DHL      LI.[2].PNGPK.[1].IYYKG.[8].AITG.[4].TVTFLE.[2].LLHKDFIEECFRRTFAA 154
gi 75415940 128 FT.[3].SNT      SY.[2].IDSLR     VNLYG.[8].RFRG.[4].LIYVNE.[2].TLHKETLIECLKRLRVG 190
gi 75089121  84 RE.[3].DKS      NA.[2].IFGNK     VYVFD.[8].KARG.[4].GAFLNE.[2].ALHNMFIKEVFSRCSYK 146
gi 81359939 221 LT     GNP      IR.[1].SNGAE     LRFLS.[7].SYSG     HLYCDE.[2].WVPNFTKLNEVASAMAT 274
gi 75090364 225 LK     GDP      II.[1].PNGAE     LHFLG.[7].GYHG     NFYFDE.[2].WTFKFKELNKVASGMAM 278
gi 81851566 112 YE.[3].RRL      IW     DNGAV     ATLFS.[5].SLRG.[4].AAWCDE.[2].KWKNPQETWDMLQFGLR 169

T4          276       SG.[5].IITTTPNGLNHFYDIWTAA     V.[20].YNDEDIF     DDGWQWS     IQTINGSSLAQ 347
ZAS-2i      149 .[2].GC.[5].YDCNPGSPLHWAYRIFIRK.[4].N.[14].LNPADNR.[1].HLPDDYI     SDVLDALPEKQ 221
gi 75086555 194       GG     IVYLTFTPEHGLTEIVKDF     L.[11].ASWEDAP     HLSPEVK     EQLLSVYSPAE 251
gi 81634366 209       RG     AIAVTFTPLRGLSAVVARY     L.[11].MTIEDAE     HYTPQER     QRVIDSYPAHE 266
gi 81525498 155       KN.[4].AELNPPAPNHPVLEIFSQY     E.[ 9].WTAKDNP     ALSDERK     QEIYNEVKHSA 214
gi 75415940 191       ME.[3].FDTNPDSPEHFFKTDYIDN     K.[ 7].FTTYDNE     LISKEFI     KTQEEIYRDMP 247
gi 75089121 147       GA.[3].IDTNPENPMHPVKKDYIDK     S.[15].FTLFDNT     FLDEEYI     ESIIASTPTGM 211
gi 81359939 275       HD.[5].YFSTPSAKTHQAYPFWTGD     E.[34].ITMEDAI     AGGFNLA.[2].EKLRNRYNTAT 362
gi 75090364 279       QK.[5].YFSTPSSMAHEAYTFWTGE     R.[34].VTILDAE     ARGCDLF.[2].DELRLEYDAEA 366
gi 81851566 170       LG.[5].VVTTTPRAVPLLKALLTDR     T.[ 6].RTAENAG     NLAEGFM     QTIARRYAGTR 227

T4          348       FRQEHTAAFEG      TSGTLISGMKLAV.[18].PEPDRKYIATLDCS.[ 2].RGQ.[5].HII       DVTD 420
ZAS-2i      222       RARFRDGSWVK      AEGVIYELFDETM.[ 7].PAEYDRVAAGQDFG.[ 2].ITN       VKI.[1].WVNG 279
gi 75086555 252       RRMRAEGIPML      GSGVVFPILEEKF.[ 6].IPDHFHRIIGIDLG       FDH.[5].CVA.[1].DAEK 311
gi 81634366 267       REARTRGVPAL      GSGRIFPVTEESI.[ 6].IPKHWVQIGGLDFG       WDH.[5].GCA.[1].DRDA 326
gi 81525498 215 .[2].LQRDWYGKRVL      PAGIIYETFDVEA.[ 6].QGHPIEMVFFGDGG.[12].TEH.[5].YTY.[1].LNQV 288
gi 75415940 248 .[2].KARVLLGEWVA      SYDSIFTNINLTS      NHEFKAPIAYLDPA.[ 2].IGG.[5].CVL       ERVD 304
gi 75089121 212 .[1].TDRDIYGKWVS      AEGVVYKDFKEKV.[ 9].TKQIKRKYAGVDWG       YEH.[5].VVA       EDFD 274
gi 81359939 363       FNMLYMCVFVD      NKDSVFSFSDLEA.[17].PFGDRPVWGGFDPA       RSG.[5].VIV.[3].MFAV 435
gi 75090364 367       FQNLLMCQFVD      DGASIFPLTMLQP.[19].PFGDRQVWLGYDPA       ETG.[5].VVV.[3].AVPG 441
gi 81851566 228 .[1].GRQELDGELVE.[1].RPGALWSRDRIEQ.[ 5].PPPLARIVVAVDPP.[ 4].KAS.[5].VVA.[1].IDAE 292

T4          421 DVW.[1].QVGVLHSNTISH        LILPDIVMRYL.[2].YNECPVYIEL.[2].TGVSV.[5].MDL.[15].KQT 492
ZAS-2i      280 AIF     VLADYGAFNMTT.[10].HWFDSIADGRY.[1].YLDFPVYCDP.[1].GGERI     QEI.[ 3].TKA 341
gi 75086555 312 DKY.[1].LYDERSESGETL        GMHADAIYLKG.[2].QIPVVVPHDA.[7].SGRRF.[5].DDH.[ 4].VYE 377
gi 81634366 327 DVF.[1].VTKIYREREATP        IIHAAALKPWG.[1].AMPWAWPHDG.[6].SGEQL.[5].AQG.[ 3].LPE 389
gi 81525498 289 ATY.[1].HSGRDTGQVKAG        STYAIEIKQFI.[3].MKEYEVPVNE.[2].FIDPA.[5].EEL.[ 7].AGA 353
gi 75415940 305 QKY.[1].AFIFQEKLPVSD        PRVLNTIKTIL.[2].LNVHTLYVED.[7].GNVTK.[5].RAG.[ 7].API 373
gi 75089121 275 GNK.[1].VIEEHAHRHKEI        DDWVAIAKGVI.[2].HGDILFYCDT     ARPEH.[5].REK.[ 3].RYA 332
gi 81359939 436 EKF.[1].VLKVIYWKGMNF        RYQAKQIEQLF.[2].YNFTYLGVDV.[2].IGQGV.[5].HFA.[ 7].RYD 499
gi 75090364 442 GKF.[1].VLERHQFRGKDF        AEQAEFIRKVT.[2].YWVTYIGVDT.[2].MGSGV.[5].QFF.[ 6].SYS 504
gi 81851566 293 GVG.[1].VLADESMTMAKP        HQWARRAIALY.[2].HEADAIVAEV.[2].GGEMV.[5].AED.[ 5].LKR 354

T4          493 .[ 1].RTKAVGCSTLKDLI     E.[2].KLIIHH.[16].WAAEEGYHDDLV.[13].KFIDYADKDDMRLASE 573
ZAS-2i      342 .[ 1].NSVESGIDFINAKI     E     RSQFFV.[ 7].LSEIWDYCRDEA.[ 5].LNDHFMDALRYAVFSD 403
gi 75086555 378 .[11].HGGNSVEFGVNWML.[3].E.[2].DLKVFN.[ 5].LKEMKMYHRKDG.[ 4].RNDDMISATRYALLMA 451
gi 81634366 390 .[ 5].DGTNGVEAGLSDML.[3].Q.[2].RWKVFS.[ 5].FEEFRLYHRKDG.[ 4].ERDDLISASRYALMMK 457
gi 81525498 354 .[11].QGIEVGIERMQSLL     S.[2].RYLLVE.[11].LQEIGMYVRDEN.[ 6].KNNHAMDTSRYATNYF 432
gi 75415940 374       KPISNKFTRIATLI.[3].A.[2].NLSIMY.[ 6].ISDIYKYKGDGK       SADDSLDSLSAAYMLL 433
gi 75089121 333 .[ 1].KAVIAGIEVISRLF     K.[2].KIFIIK.[ 6].KEEIYNYVWKDN.[ 6].LNDDTLDALRYAVYTA 396
gi 81359939 500 .[ 1].NTKNQLVLKAAGVV     E.[2].RIEWDK.[ 8].FMSVRRTTTQSG.[13].GHAEAFWAITHALHNE 572
gi 75090364 505 .[ 1].EVKTQLVMKAWSVI     K.[2].RLEFDA.[ 8].LMAIRKTITAGG.[13].GHADLAWALFHALQNE 577
gi 81851566 355 .[ 3].RGKWLRAEPVAALY     E.[2].RVRHAG.[ 1].FPALEDEMCDFA.[ 7].RSPDRLDALVWALGEL 416
```

Figure 2.6. Multiple alignment of pfam03237 with a ZAS-associated terminase. Multiple sequence alignment of pfam03237 (Terminase_6) with the ZAS-2 terminase sequence (ZAS-2i) aligned with RPS-BLAST in the CDD (*S60*) (E value 1.2e-19). Conserved functional motifs (*S62-S63*) are indicated as well as the boundary between the N-terminal ATPase domain (T4: amino acids 1–360 (*S63*)) and C-terminal nuclease domain (T4: amino acids 361–610 (*S63*)) based on T4 (*S59, S62*). Conserved functional motifs for the N terminal ATPase center include (*S62-S63*) a Walker A motif G/A-XXXXGK(T/S) (purple), a Walker B motif ZZZZD where Z represents a

hydrophobic amino acid (blue), a catalytic carboxylate group motif (usually) Glu (orange), and an ATPase coupling motif (T/S-G/A-T/S(N)) (green). The functional motif for the C-terminal nuclease center is a catalytic triad of Asp/Glu residues (red)(*S62-S63*). The forward primer (upper light blue box) targeted a conserved region between the putative Walker A and Walker B motifs in the ATPase domain and the reverse primer targeted a conserved region that included the central aspartic acid residue in the catalytic triad (lower light blue box). Also indicated is the 235 residue alignment region (without gaps) used for phylogenetic analysis. The alignment shows the 10 most diverse members (out of 43) of the pfam with the T4 large terminase subunit gene gp17 being the representative sequence. Numbers in brackets are unaligned residues. ZA2-2i was chosen for the alignment because this gene was found to be present in the largest (43.5 kb) prophoage-like element of the ZAS genome (see supporting text).

Figure 2.7. Phylogenetic analysis of retrieved *Treponema* SSU rRNA sequences and close relatives. Maximum likelihood tree of 39 retrieved *Treponema* SSU rRNA sequences from co-localized pairs (red), 78 reference library *Treponema* SSU rRNA sequences (black) and close relatives found in the SILVA (*S11*) database v100 (green). Also highlighted are Phage Hosts I through IV, *Reticulitermes* Environmental Phylotypes (REPs) 1 through 7 (comprising 67% of all treponemes found on the array; see Table 2.6), previously identified clades of traditional treponemes (known as subgroups 1 and 2)(*S64-S66*) and the so called "Termite Cluster" (*S65*). Many *R. hesperus* SSU rRNAs retrieved from the microfluidic array (including Phage Hosts I through IV) were similar to previously characterized SSU rRNAs from other *Reticulitermes* species. The overall diversity of *R. hesperus* treponeme SSU rRNAs was phylogenetically similar to that of other *Reticulitermes* species (*S64*). The tree was constructed based on 743 aligned unambiguous nucleotides excluding gaps using PhyML 2.4.5 (*S14*) implemented in ARB (*S67*). An optimal substitution model was estimated with jModelTest 0.1.1 (*S13-S14*) using the AICc criterion and was found to be the Tamura-Nei model (*S15*) +I+Γ (nCat=4) with unequal base pair frequencies. Shorter sequences (A7, A9, rF79, rG41 and rG53) were added by parsimony. Support values greater than 50% for 1000 bootstrap iterations are shown. Scale bar represents 0.1 nucleotide changes per alignment position. See Table 2.11 for a list of all sequences. Note that reference library sequences begin with the letter "r".


Phage Host I/REP 4
BPH IA
BPH IB
A11_1
unc Trep clone RFS21, Reticulitermes flavipes (AF068338)
rF39
G1_3
A1_1
B1_2
A14_1
rF63
rG42
A10_1
E1_3
F1_3
C2_2
rF71
C1_2
G5_3
G3_3
A3_1
rG50
B4_2
E3_3
rF23
unc Trep sp., Reticulitermes speratus (AB088896)
T. primitia
Phage Host III/REP 13
A6_1
G4_3
rG41
F2_3
G2_3
unc Trep sp., Reticulitermes speratus (AB192140)
rG95
unc Trep sp., Reticulitermes speratus (AB088876)
C4_2
unc Trep sp., Reticulitermes speratus (AB088870)
A15_1
rF52
rG72
G6_3
rG36
rG43
rG53
unc Trep sp., Reticulitermes speratus (AB088874)
T. azotonutricium
rG31
rG20
unc Trep sp., Reticulitermes sp. (AB192248)
Phage Host II/REP 7
A7_1
A9_1
A4_1
unc Trep sp., Reticulitermes speratus (AB192197)
B2_2
A5_1
rG3
rG24
A12_1
A13_1
E2_3
rF95
unc Trep clone RFS94, Reticulitermes flavipes (AF068423)
rF24
unc Trep sp., Reticulitermes speratus (AB088909)
B5_2
A18_1
rG55
unc Trep sp., Reticulitermes speratus (AB088878)
rF75
rG81
unc Trep sp., Reticulitermes sp. (AB192251)
A16_1
rF84
Phage Host IV/REP 14
A8_1
B3_2
rF86
A2_1
unc Trep sp., Reticulitermes speratus (AB192202)
rG35
rG25
D2_2
unc Trep sp., Reticulitermes sp. (AB192247)
REP 1
rF13
rG86
rF22
rF29
rF43
rF56
rF77
rF82
rF83
rF92
rF81
rG9
rG14
rG15
rG17
rG28
rG32
rG49
rG71
rF69
rG88
rG74
rG78
rG63
unc Trep sp., Reticulitermes speratus (AB192142)
rF8
REP 3
rF94
rG83
rF26
rF40
rF100
rG80
rG30
unc Trep sp., Reticulitermes speratus (AB088915)
rF28
uncTrep sp., Reticulitermes speratus (AB088866)
rF93
unc Trep sp., Reticulitermes speratus (AB088890)
rF16
unc Trep clone RFS84, Reticulitermes flavipes (AF068428)
G7_3
C3_2
unc Trep clone RFS99, Reticulitermes flavipes (AF068424)
"Termite Cluster"
REP 2
rF21
rG73
rF3
rF12
rF14
rF5
rF88
unc Trep clone HsDiSp314, Hodotermopsis sjoestedti (AB032005)
rG60
rG11
rG16
unc Treponema clone RFS2, Reticulitermes flavipes (AF068429)
REP 6
rF79
rG29
rF68
unc Treponema sp., Reticulitermes speratus (AB088895)
REP 5
rF47
rF61
rG91
rF33
uncTreponema clone RFS12, Reticulitermes flavipes (AF068335)
Treponema succinifaciens (M57738)
Treponema zioleckii (DQ065758)
Treponema socranskii subsp. socranskii (AF033306)
1
Treponema denticola ATCC 35405 (AE017226)
Treponema vincentii (AY119690)
Treponema pallidum subsp. pallidum str. Nichols (AE000520)
2
Borrelia burgdorferi (X85195)
Spirochaeta aurantia (AY599019)
Brevinema andersonii (L31543)
Escherichia coli str. K−12 substr. MG1655 (U00096)
Not-termite Treponemes
0.1

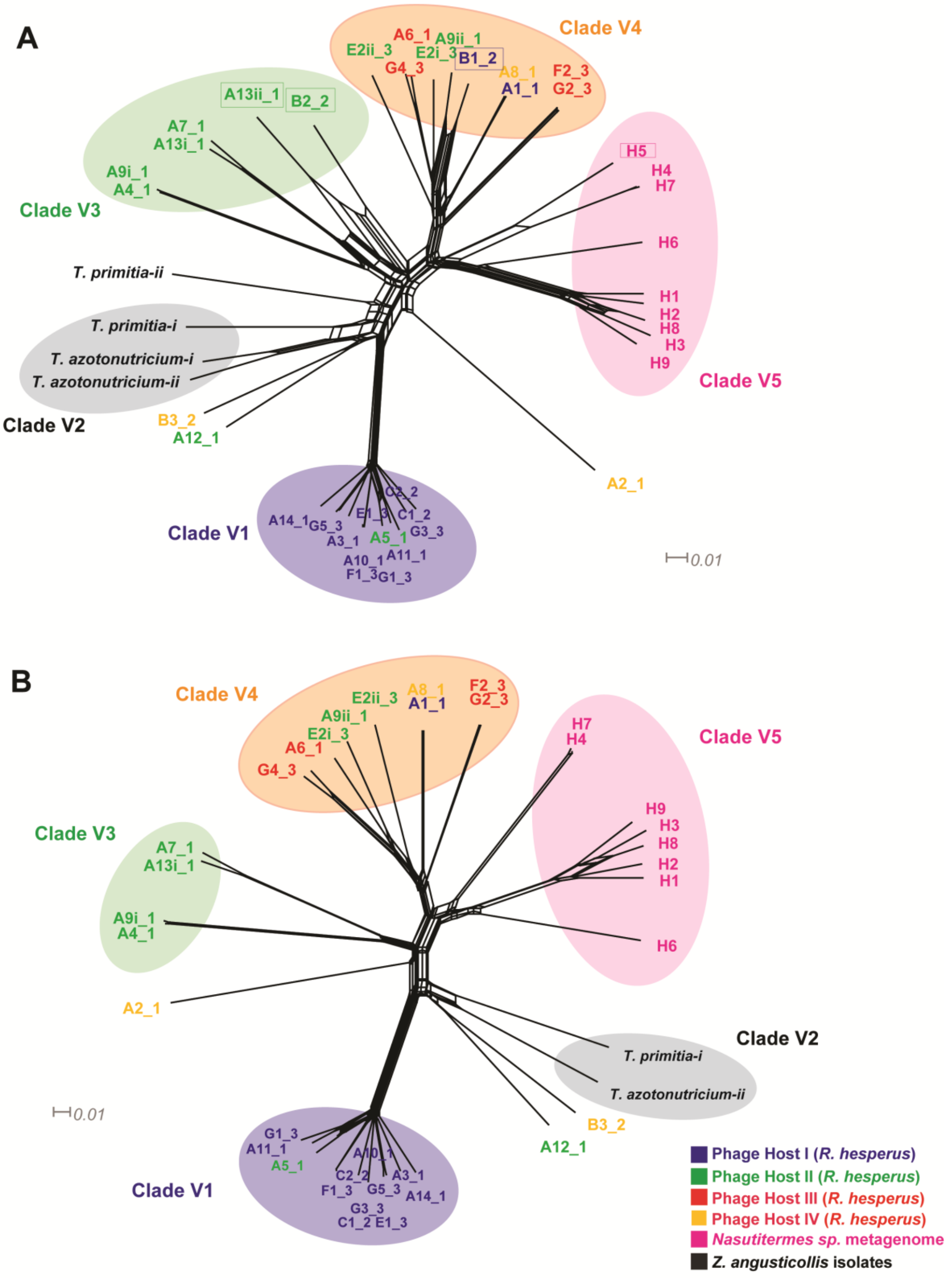


Figure 2.8. NeighborNet network of termite-related terminase alleles. (A) NeighborNet (*S68*) of **(1)** all terminase alleles that were retrieved with Phage Hosts I through IV, **(2)** terminases genes present in Z. *angusticollis* isolates, *Treponema primitia* (ZAS-2), and *Treponema azotonutricium* (ZAS-9), and **(3)** terminase alleles found in the metagenome of the hindgut of an *Nasutitermes sp.* termite. Boxed sequences are the first four events identified by RDP3 as recombinant (see Methods). **(B)** Same as (A) but excluding **(1)** RDP3 identified recombinant sequences, **(2)** ZAS terminases alleles associated with most likely defunct phage cassettes. ZAS-2 and ZAS-9 both have two copies of the terminase gene. Each copy resides in a region coding for

other viral genes, however only a single one of these copies in each genome appears to be present in a large enough contiguous region of putative viral genes (~36–43 kbp) that could constitute a viable phage and therefore only this copy was included. After removal of recombinant sequences (B1, B2, A13ii, H5) there remains some residual reticulate patterns at the base of the network, however the network largely appears to be tree-like (confirmed by likelihood mapping; see Methods). These sequences were used to generate the terminase tree in Fig. 2.2. The network structure shown here is consistent with the topology shown in Fig. 2.2. The network was calculated using SplitsTree4 (*S28*) on 705 aligned unambiguous nucleotides without gaps using the optimal model found by FindModel (*S29*), a K80 substitution model (*S69*) +Γ with $\alpha \simeq 0.5$. The LSfit score for networks A and B was 99.97% and 99.94%, respectively. Note that sample B1 associated with Host I in (A) was found by RDP3 to be a chimera of A1 (Host I) and A9ii (Host II), possibly indicating a lateral gene transfer event between these two distinct subpopulations of viruses. Alternatively, since only one such event was observed, it could also be due to an unlikely experimental artifact. Sample notation is as described in Fig. 2.2.

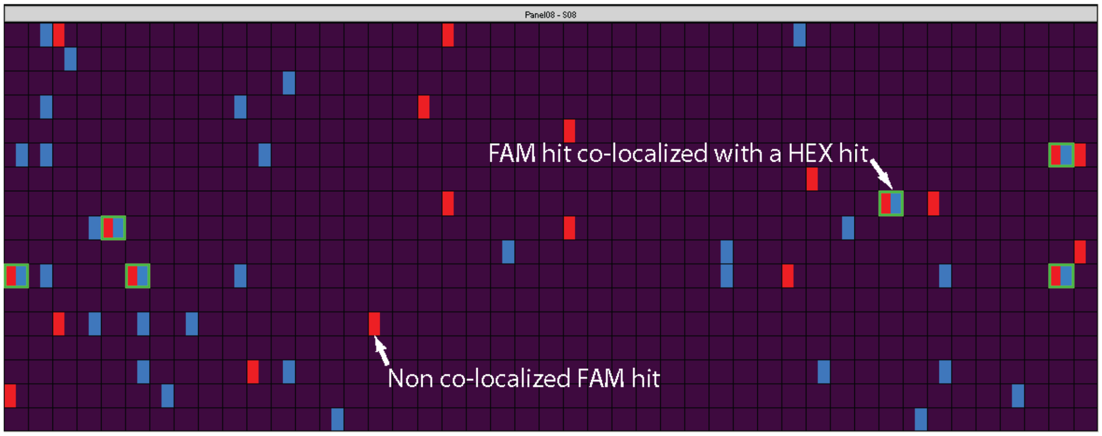


Figure 2.9. Example of microfluidic array panel readout after thresholding. Blue squares represent hits in the HEX/rRNA channel and red squares represent hits in the FAM/terminase channel. Co-localized hits are highlighted in green. In this example, spurious amplification is expected to account for ~50% of all non co-localized FAM hits based on the number of FAM hits in the no-template-control panel for this microfluidic array (7 hits).

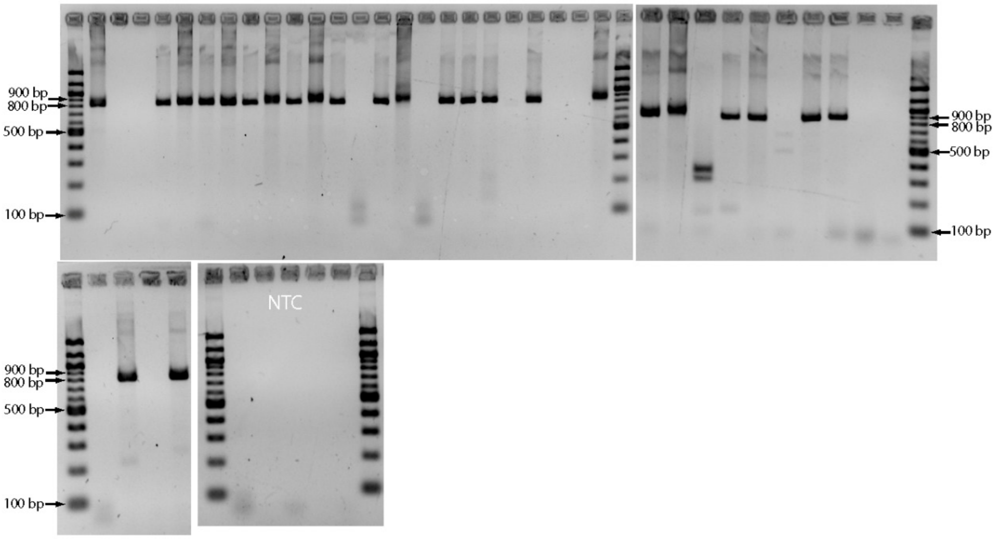


Figure 2.10. Agarose gel electrophoresis analysis of all FAM hits in a microfluidic array panel. All 38 FAM hits in panel #7 of chip B were post-amplified and analyzed by agarose gel electrophoresis. Also shown are the five no-template-control (NTC) samples for this PCR reaction. The expected amplicon size is ~820 bp (compared to a 100 bp ladder). Out of 38 reactions, 13 were negative for the template. This value is consistent with the number of FAM hits in the no-template-control panel for this microfluidic array, which was 16. The gel image was inverted, brightness was linearly scaled to maximize contrast and size was proportionally scaled to fit the figure. The microfluidic array was analyzed with the BioMark Digital PCR analysis software (Fluidigm, v.2.0.6) using a FAM threshold 0.2 and linear baseline correction.

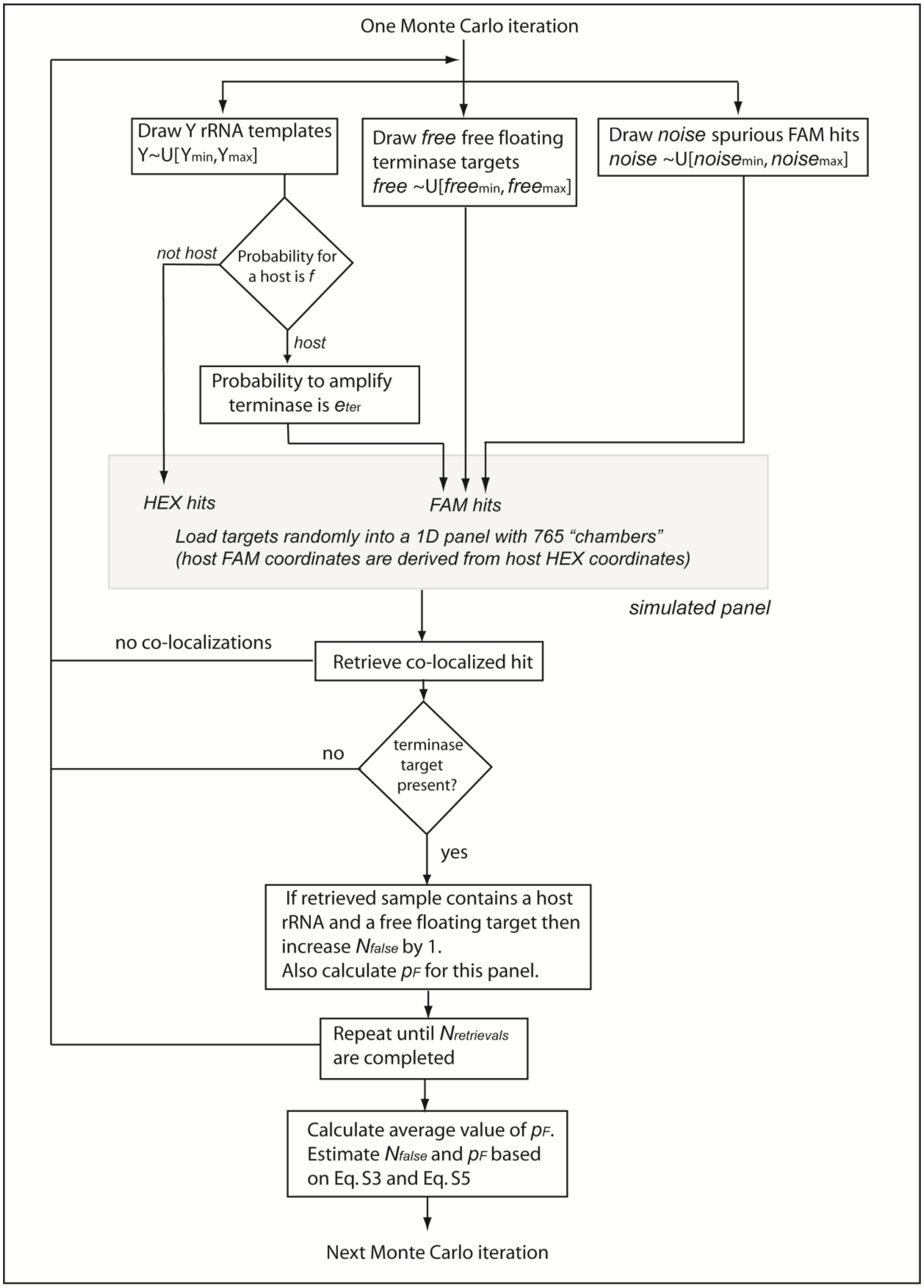
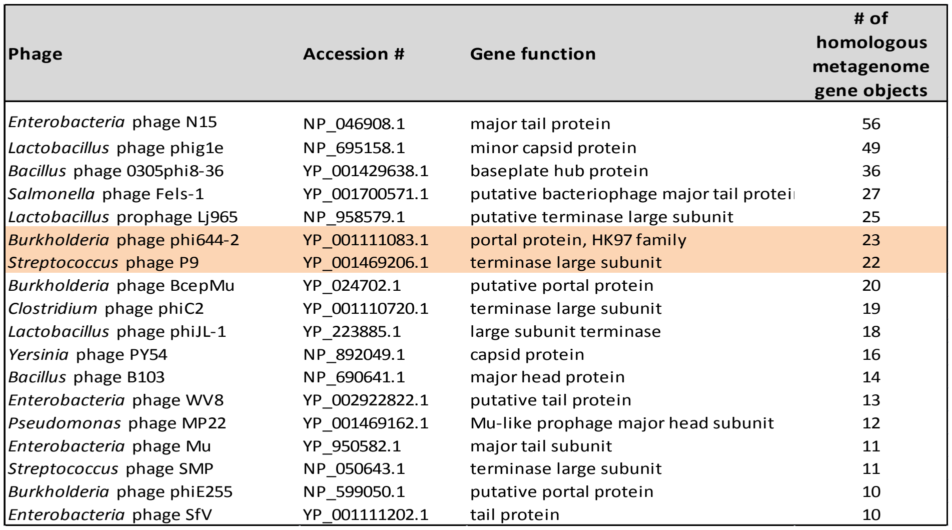

One Monte Carlo iteration
Draw Y rRNA templates Y~U[Y_{min},Y_{max}]
Draw *free* free floating terminase targets *free* ~U[$free_{min}$, $free_{max}$]
Draw *noise* spurious FAM hits *noise* ~U[$noise_{min}$, $noise_{max}$]
not host
Probability for a host is *f*
host
Probability to amplify terminase is e_{ter}
HEX hits
FAM hits
Load targets randomly into a 1D panel with 765 "chambers" (host FAM coordinates are derived from host HEX coordinates)
simulated panel
no co-localizations
Retrieve co-localized hit
no
terminase target present?
yes
If retrieved sample contains a host rRNA and a free floating target then increase N_{false} by 1. Also calculate p_F for this panel.
Repeat until $N_{retrievals}$ are completed
Calculate average value of p_F. Estimate N_{false} and p_F based on Eq. S3 and Eq. S5
Next Monte Carlo iteration

Figure 2.11. Schematic diagram of a Monte Carlo simulation of microfluidic array loading and sampling. See supporting text for further details.

2.9.4 Supporting tables

Table 2.2. Abundance of homologs of known viral genes in the higher termite metagenome. This table describes the number (or *abundance*, see definition in Materials and methods) of metagenome gene objects in the higher termite metagenome that were homologous to the indicated viral phage genes (E value ≤ 0.001, *abundance* ≥ 10 metagenome gene objects). This list constitutes the most abundant viral-specific genes in the metagenome (i.e., viral genes related to building a virion), using the viral RefSeq database v37 (*S41*) as a reference for known viral genes. The two highlighted rows are the portal protein and terminase protein that were found to have homologs in the ZAS prophage-like elements.



Phage	Accession #	Gene function	# of homologous metagenome gene objects
Enterobacteria phage N15	NP_046908.1	major tail protein	56
Lactobacillus phage phig1e	NP_695158.1	minor capsid protein	49
Bacillus phage 0305phi8-36	YP_001429638.1	baseplate hub protein	36
Salmonella phage Fels-1	YP_001700571.1	putative bacteriophage major tail protei	27
Lactobacillus prophage Lj965	NP_958579.1	putative terminase large subunit	25
Burkholderia phage phi644-2	YP_001111083.1	portal protein, HK97 family	23
Streptococcus phage P9	YP_001469206.1	terminase large subunit	22
Burkholderia phage BcepMu	YP_024702.1	putative portal protein	20
Clostridium phage phiC2	YP_001110720.1	terminase large subunit	19
Lactobacillus phage phiJL-1	YP_223885.1	large subunit terminase	18
Yersinia phage PY54	NP_892049.1	capsid protein	16
Bacillus phage B103	NP_690641.1	major head protein	14
Enterobacteria phage WV8	YP_002922822.1	putative tail protein	13
Pseudomonas phage MP22	YP_001469162.1	Mu-like prophage major head subunit	12
Enterobacteria phage Mu	YP_950582.1	major tail subunit	11
Streptococcus phage SMP	NP_050643.1	terminase large subunit	11
Burkholderia phage phiE255	NP_599050.1	putative portal protein	10
Enterobacteria phage SfV	YP_001111202.1	tail protein	10

Table 2.3. Similarity analysis of the termite-associated terminase gene and portal protein gene with close homologs. The following table describes the result of a BLAST analysis of the large terminase subunit gene (411 aa in length) and the portal protein gene (396 aa in length) found in *T. primitia's* prophage-like element with close homologs. Close homologs were searched for in: (1) the larger prophage-like element present in the genome of *T. azotonutricium*, (2) the metagenome of the hindgut of a *Nasutitermes sp*. termite, and (3) the viral RefSeq database v37 (*S41*). The table demonstrates that the alleles of the termite-associated phage genes were very similar to each other and highly divergent from their closest homologs found among all currently known viral genomes. Alignments were performed on the amino acid sequences.

Large terminase subunit gene	**% identity** *	**% similarity***	**Gaps***	**E value**
	T. azotonutricium			
	363/411 (89%)	385/411 (94%)	4/411 (0%)	0
	Higher termite metagenome			
	317/407 (78%)	359/407 (89%)	5/407 (1%)	0
	Viral RefSeq database *(Lactobacillus johnsonii* prophage Lj771)			
	107/415 (25%)	177/415 (42%)	64/415 (15%)	4.00E-19
Portal protein	**% identity**	**% similarity**	**Gaps**	**E value**
	T. azotonutricium			
	309/382 (81%)	348/382 (92%)	3/382 (0%)	0
	Higher termite metagenome			
	273/392 (70%)	324/392 (83%)	11/392 (2%)	1.00E-167
	Viral RefSeq database *(Streptomyces* phage mu1/6)			
	99/382 (25%)	156/382 (40%)	52/382 (13%)	6.00E-17

*Numbers divided by a forward slash correspond to the number of amino acids in each pair-wise alignment ("identity/total", "similarity/total", and "gaps/total", depending on the column).

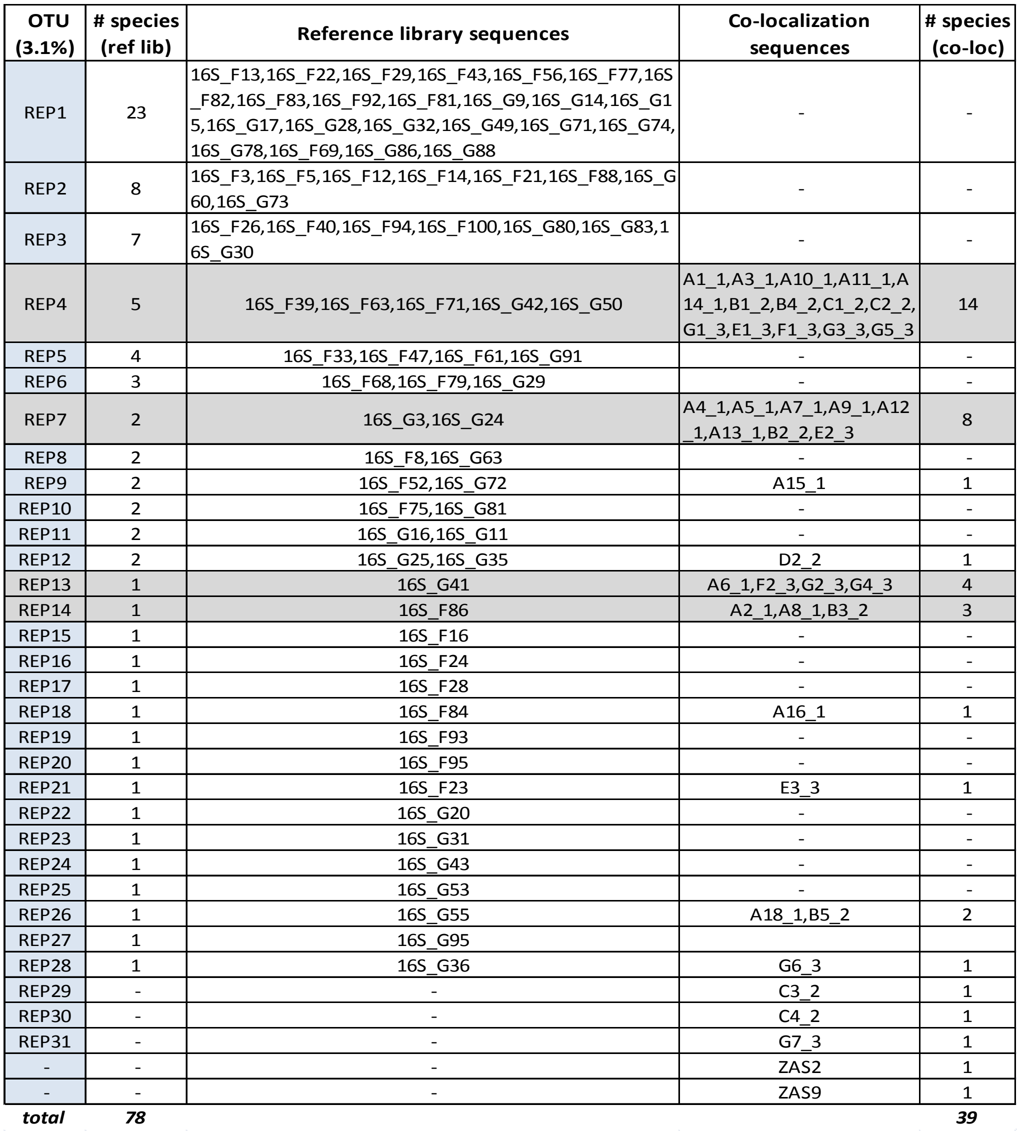
Table 2.4. Sample collection and analysis information. Collection dates, collection sites, and dPCR execution dates for the *R. hesperus* specimens. The different colonies were on average 120 meters apart. The microfluidic array and colony labels noted here were used to label the samples throughout this report.

Chip ID	Chip designation in trees	Termite collection date	Date of chip execution	Colony	GPS coordinates
1151065015	A	11/13/2008	11/25/2008	1	34 19' 25.6"N/ 118 0' 17.9"W
1151065011	B	5/27/2009	5/29/2009	2	34 19' 31"N/118 00' 20.8"W
1151065010	C	5/27/2009	6/6/2009	2	"
1151065012	D	5/27/2009	6/7/2009	2	"
1151065017	E	5/27/2009	6/21/2009	3	34 19' 28"N/118 00' 17.5"W
1151065018	F	5/27/2009	6/22/2009	3	"
1151065019	G	5/27/2009	6/24/2009	3	"

Table 2.5. Estimated evolutionary distance between bacterial host SSU rRNA phylotypes. The number of base substitutions per site from averaging over all sequence pairs within and between host groups is shown. With the exception of samples A7 and A9 (that were composed of 784 and 810 nucleotides respectively) the SILVA (*S11*) -based alignment contained 898 unambiguous nucleotides. Distances were calculated using the Jukes-Cantor (*S70*) nucleotide substitution model in MEGA4 (*S19*). The number of repetitions appearing in Table 2.1 are based on an Operational Taxonomical Unit (OTU) cutoff of 2% assigned by DOTUR (*S71*) with the furthest neighbor sequence assignment method. The next significant OTU cutoff was 2.5%, adding a more divergent member (B4) to Host I, however due to the larger divergence and single instance of this event it cannot be statistically validated and therefore it was not included in this analysis. The distance matrix used by DOTUR was based on the above alignment and calculated in ARB (*S67*) using the Jukes-Cantor substitution model. Each bacterial host was less than 0.9% divergent on average. The maximum divergence was observed between Host III and ZAS-9 where the corrected evolutionary distance across their deduced rRNAs was measured to be 9.3%.

	Host I *(n=13)*	**Host II** *(n=8)*	**Host III** *(n=4)*	**Host IV** *(n=3)*	**ZAS-2** *(n=1)*	**ZAS-9** *(n=1)*
Host I	0.0084					
Host II	0.0822	0.0083				
Host III	0.0685	0.0544	0.005			
Host IV	0.0817	0.0841	0.087	0.0075		
ZAS-2	0.0396	0.0678	0.06	0.0712	-	
ZAS-9	0.073	0.086	0.0933	0.0865	0.0603	-

Table 2.6. Retrieved *Treponema* phylotypes from the microfluidic arrays



OTU (3.1%)	# species (ref lib)	Reference library sequences	Co-localization sequences	# species (co-loc)
REP1	23	16S_F13,16S_F22,16S_F29,16S_F43,16S_F56,16S_F77,16S_F82,16S_F83,16S_F92,16S_F81,16S_G9,16S_G14,16S_G15,16S_G17,16S_G28,16S_G32,16S_G49,16S_G71,16S_G74,16S_G78,16S_F69,16S_G86,16S_G88	-	-
REP2	8	16S_F3,16S_F5,16S_F12,16S_F14,16S_F21,16S_F88,16S_G60,16S_G73	-	-
REP3	7	16S_F26,16S_F40,16S_F94,16S_F100,16S_G80,16S_G83,16S_G30	-	-
REP4	5	16S_F39,16S_F63,16S_F71,16S_G42,16S_G50	A1_1,A3_1,A10_1,A11_1,A14_1,B1_2,B4_2,C1_2,C2_2,G1_3,E1_3,F1_3,G3_3,G5_3	14
REP5	4	16S_F33,16S_F47,16S_F61,16S_G91	-	-
REP6	3	16S_F68,16S_F79,16S_G29	-	-
REP7	2	16S_G3,16S_G24	A4_1,A5_1,A7_1,A9_1,A12_1,A13_1,B2_2,E2_3	8
REP8	2	16S_F8,16S_G63	-	-
REP9	2	16S_F52,16S_G72	A15_1	1
REP10	2	16S_F75,16S_G81	-	-
REP11	2	16S_G16,16S_G11	-	-
REP12	2	16S_G25,16S_G35	D2_2	1
REP13	1	16S_G41	A6_1,F2_3,G2_3,G4_3	4
REP14	1	16S_F86	A2_1,A8_1,B3_2	3
REP15	1	16S_F16	-	-
REP16	1	16S_F24	-	-
REP17	1	16S_F28	-	-
REP18	1	16S_F84	A16_1	1
REP19	1	16S_F93	-	-
REP20	1	16S_F95	-	-
REP21	1	16S_F23	E3_3	1
REP22	1	16S_G20	-	-
REP23	1	16S_G31	-	-
REP24	1	16S_G43	-	-
REP25	1	16S_G53	-	-
REP26	1	16S_G55	A18_1,B5_2	2
REP27	1	16S_G95		
REP28	1	16S_G36	G6_3	1
REP29	-	-	C3_2	1
REP30	-	-	C4_2	1
REP31	-	-	G7_3	1
-	-	-	ZAS2	1
-	-	-	ZAS9	1
total	***78***			***39***

All reference library sequences (*n*=118; 876 ± 71 bp SD) were initially classified with RDB (*S72*) and *Treponema* phylotypes (66.1%, *n*=78 with 99–100% confidence) were subsequently aligned by the SILVA incremental aligner SINA (*S11*). A distance matrix was calculated in ARB (*S67*) for the 78 reference library *Treponema* species, the 39 co-localized *Treponema* species, and ZAS-2 and ZAS-9 (*n*=119). Note that REP4 was co-localized 14 times, however one of these co-localizations, B4, was more divergent than the other ribotypes of this group (see Table 2.5) and was therefore not regarded as a repeated co-localization of Host I in Table 2.1 and Table 2.5. The distance matrix was calculated based on 780 unambiguous nucleotides (with the exception of A7, A9, rF79, rG41, rG53 that were in the range of 624–767 nucleotides) using the Jukes-Cantor (*S70*) method. Operational taxonomical units (OTUs) were then determined by DOTUR (*S71*) based on the furthest neighbor sequence assignment method using an OTU cutoff of 3.1%. This cutoff is slightly higher than the OTU cutoff used to identify the repeated co-localizations (2%) in Fig. 2.2 in order to make the statistical test for repeated co-localization more stringent. REPs corresponding to putative bacterial hosts are highlighted in gray. All *Treponema* sequences were also screened with Bellerophon v3 (*S9*) on Greengenes (*S10*) for chimeras and were found to be negative. The remaining phyla indentified by RDB to be present in the reference library were Proteobacteria (13.6%, 100% confidence), Firmicutes (6.8%, Clostridia 53–100% confidence), Tenericutes (5.9%, Mycoplasmataceae with 77–90% confidence), Bacteroidetes (3.4%, 100% confidence), Actinobacteria (3.4%, 100% confidence) and Planctomycetes (0.8%, 100% confidence). All these phyla have been observed previously in SSU rRNA libraries of *Reticulitermes speratus* (*S73*). However, from the number of rRNA targets observe in the no-template-control panels we anticipate that background amplification (see Materials and methods) should contribute to 34.1 ± 18.4% SD of the reference library sequences due to sparse loading of the panels (increasing the fraction of background amplification products). Based on retrieval of rRNA sequences from the no-template-control panel (not shown) we expect the major contributor to this fraction to be bacteria from the Proteobacteria phylum. The finding that free living prokaryotes in the termite hindgut are dominated by spirochetes is consistent with electron microscope observations showing that spirochetes can account for over 50% of the gut microbes in some termites (*S74*). The absence of bacteria belonging to the TG-1 phylum (*S75*) is an indication that large flagellates were successfully filtered out by the 5 μm pre-filter and did not lyse in this process (see Methods).

Table 2.7. Selection pressure analysis of the terminase gene. Codon-based test of purifying (negative) selection for Hosts I through IV excluding suspected recombinant sequences (B1, B2 and A13ii). d_S and d_N are the number of synonymous and nonsynonymous substitutions per number of synonymous and nonsynonymous sites respectively obtained from averaging over all sequence pairs within a given group. d_S and d_N were calculated by various methods: **NG86** — Nei-Gojobori method (*S76*) with the Jukes-Cantor (*S70*) nucleotide substitution model, **Modified NG86** (*S77*) — NG86 method with the Jukes-Cantor nucleotide substitution model, **LWL85** — Li-Wu-Luo method (*S78*), **PBL85** — Pamilo-Bianchi-Li method (*S79*), and **Kumar** — Kumar method (*S80*). For the modified NG86 method, the ratio of transitional to transversional distances per site (R) was calculated by averaging over all sequence pairs within each group using the 3rd codon position based on the Kimura 2-parameter method (*S69*). All results are based on the pairwise analysis of 235 unambiguous codon positions without gaps. Standard error estimates were obtained by a bootstrap procedure with 1000 replicates. The distribution of the test statistic (*D*) is approximated to be normal since the number of nucleotides contributing to d_S and d_N were sufficiently large (>10), allowing to test the null hypothesis using a one-tailed (Z > 0) Z test (*S80*). The P value (one-tailed Z test) for observing Z > 0 ($d_S > d_N$) by chance is shown in the table. Z is shown to be greater than zero in a statistically significant manner (P < 10^{-7} for Hosts I–III and P < 0.025 for Host IV) indicating negative selection was statistically significant. n/c denotes cases in which it was not possible to estimate evolutionary distances. All analyses were carried out with MEGA4 (*S19*).

Host	Method	d_S (± S.E.)			d_N (± S.E.)			d_N/d_S	$D = d_S - d_N$ (± S.E.)			$Z = D/std(D)$	P value
I	NG86 (R=0.5)	0.57	±	0.08	0.04	±	0.01	0.08	0.53	±	0.07	7.58	1.7E-14
(*n* =12)	Modified NG86 (R=2.02)	0.33	±	0.03	0.05	±	0.01	0.15	0.28	±	0.03	8.94	0.0E+00
	LWL85	0.49	±	0.06	0.04	±	0.01	0.09	0.45	±	0.06	7.52	2.8E-14
	PBL93	0.44	±	0.05	0.04	±	0.01	0.10	0.40	±	0.05	7.42	5.7E-14
	Kumar	0.37	±	0.04	0.04	±	0.01	0.12	0.32	±	0.04	8.28	0.0E+00
II	NG86 (R=0.5)	1.50	±	0.12	0.17	±	0.02	0.11	1.34	±	0.12	11.18	0.0E+00
(*n* =9)	Modified NG86 (R=1.44)	0.99	±	0.07	0.18	±	0.02	0.18	0.81	±	0.07	11.32	0.0E+00
	LWL85	1.48	±	0.11	0.17	±	0.02	0.11	1.31	±	0.11	11.69	0.0E+00
	PBL93	1.49	±	0.10	0.17	±	0.02	0.11	1.32	±	0.10	13.33	0.0E+00
	Kumar	1.14	±	0.08	0.16	±	0.02	0.14	0.97	±	0.08	12.25	0.0E+00
III	NG86 (R=0.5)	0.72	±	0.13	0.06	±	0.01	0.08	0.66	±	0.12	5.35	4.3E-08
(*n* =4)	Modified NG86 (R=1.80)	0.50	±	0.06	0.06	±	0.01	0.12	0.44	±	0.06	6.92	2.2E-12
	LWL85	0.70	±	0.09	0.05	±	0.01	0.08	0.64	±	0.10	6.75	7.2E-12
	PBL93	0.62	±	0.09	0.06	±	0.01	0.09	0.56	±	0.09	6.26	1.9E-10
	Kumar	0.55	±	0.07	0.05	±	0.01	0.10	0.50	±	0.08	6.63	1.7E-11
IV	NG86 (R=0.5)	n/c	±	n/c	0.19	±	0.02	n/c	n/c	±	n/c	n/c	n/c
(*n* =3)	Modified NG86 (R=1.97)	1.53	±	0.20	0.21	±	0.03	0.14	1.32	±	0.19	6.76	6.8E-12
	LWL85	2.30	±	1.06	0.20	±	0.09	0.09	2.10	±	0.99	2.11	1.7E-02
	PBL93	1.65	±	0.82	0.20	±	0.09	0.12	1.45	±	0.73	1.98	2.4E-02
	Kumar	1.94	±	0.62	0.17	±	0.07	0.09	1.76	±	0.57	3.09	9.9E-04

Table 2.8. Similar terminase sequences associated with different bacterial hosts. Terminase alleles associated with different bacterial hosts having less than 10% difference between their nucleotide sequences.

Sequence 1	Sequence 2	% p-distance (705 bp)
A1_1 (Host I)	A8_1 (Host IV)	0
G1_3 (Host I)	A5_1 (Host II)	3
B1_1* (Host I)	A9ii_1 (Host II)	6.5

*Identified by RDP3 as a recombination between A9ii_1 (Host II) and A1_1 (Host I). See also Fig. 2.8.

Table 2.9. P values for the P Test comparing terminase alleles by bacterial host. The P Test (*S97*) estimates the similarity between communities as the number of parsimony changes that would be required to explain the distribution of sequences between the different samples in the tree (samples here were grouped by bacterial host). The P value is the fraction of trials in which the true tree requires fewer changes than trees in which the sample assignments have been randomized (*S98*). The P test was implemented in Fast UniFrac (*S99*) selecting the “P Test Significance” option, comparing “Each pair of samples” using *n*=1000 random permutations. The analysis was performed on the phylogenetic tree in Fig. 2.2 applying midpoint rooting. P values shown have been corrected for multiple comparisons using the Bonferroni correction.

	Host II	Host III	Host IV
Host I	≤0.001	≤0.001	0.024
Host II	-	0.018	1
Host III	-	-	0.204

Table 2.10. P values for the P Test comparing terminase alleles by colonies. Samples here were grouped by termite colony. P values shown have been corrected for multiple comparisons using the Bonferroni correction. *n*=1000 random permutations were used to calculate P Values. See Table 2.9 for further details.

	Colony 2	Colony 3
Colony 1	0.399	0.927
Colony 2	-	0.537

Table 2.11. Sequences analyzed in this study. Accession numbers of the uncultured treponemes associated with Phage Host I through IV in Fig. 2.2 were AF068338, AB192197, AB192140, and AB192202, respectively.

Clone ID	Termite/bacterium species	Location/Source	Method	Accession (NCBI/JGI)	Figure	Reference
Terminase gene – isolates						
ZAS2i	*Z. angusticollis /T. primitia*	California	Isolate		2.2,2.5,2.6,2.8	this study
ZAS2ii	*Z. angusticollis /T. primitia*	California	Isolate		2.5,,2.8	this study
ZAS9i	*Z. angusticollis /T. azotonutricium*	California	Isolate		2.5,,2.8	this study
ZAS9ii	*Z. angusticollis /T. azotonutricium*	California	Isolate		2.2,2.5,2.6	this study
Terminase gene - co-localization						
A1_1	*Reticulitermes hesperus*	California	Digital PCR	HQ202808	2.2,2.5,2.6	this study
A3_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187752	2.2,2.5,2.6	this study
A10_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187760	2.2,2.5,2.6	this study
A11_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187761	2.2,2.5,2.6	this study
A14_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187765	2.2,2.5,2.6	this study
B1_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187766	2.5,,2.8	this study
C1_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187769	2.2,2.5,2.6	this study
C2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187770	2.2,2.5,2.6	this study
E1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187771	2.2,2.5,2.6	this study
F1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187774	2.2,2.5,2.6	this study
G1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187776	2.2,2.5,2.6	this study
G3_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187778	2.2,2.5,2.6	this study
G5_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187780	2.2,2.5,2.6	this study
A4_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187753	2.2,2.5,2.6	this study
A5_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187754	2.2,2.5,2.6	this study
A7_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187756	2.2,2.5,2.6	this study
A9i_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187758	2.2,2.5,2.6	this study
A9ii_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187759	2.2,2.5,2.6	this study
A12_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187762	2.2,2.5,2.6	this study
A13i_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187763	2.2,2.5,2.6	this study
A13ii_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187764	2.5,,2.8	this study
B2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187767	2.5,,2.8	this study
E2i_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187772	2.2,2.5,2.6	this study
E2ii_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187773	2.2,2.5,2.6	this study
A6_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187755	2.2,2.5,2.6	this study
F2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187775	2.2,2.5,2.6	this study
G2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187777	2.2,2.5,2.6	this study
G4_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187779	2.2,2.5,2.6	this study
A2_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187751	2.5,,2.8	this study

A8_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187757	2.5,,2.8	this study
B3_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187768	2.5,,2.8	this study
Terminase gene - close relatives						
H1	*Nasutitermes sp.*	Costa Rica	Metagenome	2004118157	2.2,2.5,2.8	(*S5*)
H2	*Nasutitermes sp.*	Costa Rica	Metagenome	2004126816	2.2,2.5,2.8	(*S5*)
H3	*Nasutitermes sp.*	Costa Rica	Metagenome	2004144277	2.2,2.5,2.8	(*S5*)
H4	*Nasutitermes sp.*	Costa Rica	Metagenome	2004144007	2.2,2.5,2.8	(*S5*)
H5	*Nasutitermes sp.*	Costa Rica	Metagenome	2004132071	2.5,,2.8	(*S5*)
H6	*Nasutitermes sp.*	Costa Rica	Metagenome	2004107522	2.2,2.5,2.8	(*S5*)
H7	*Nasutitermes sp.*	Costa Rica	Metagenome	2004111244	2.2,2.5,2.8	(*S5*)
H8	*Nasutitermes sp.*	Costa Rica	Metagenome	2004124547	2.2,2.5,2.8	(*S5*)
H9	*Nasutitermes sp.*	Costa Rica	Metagenome	2004134785	2.2,2.5,2.8	(*S5*)
H10	*Nasutitermes sp.*	Costa Rica	Metagenome	2004136622	2.5,,2.8	(*S5*)
Terminase gene - non termite related						
T4	Phage isolate		Isolate	NP_049776.1	2.5,2.6	(*S81*)
Vic	*Victivallis vadensis ATCC BAA-548*	Feces, human	Isolate	ZP_06243301.1	2.5	-
Sino	*Sinorhizobium medicae WSM419*	Plant root, Soil (Sardinia)	Isolate	YP_001327565.1	2.5	(*S82*)
Gluc	*Gluconobacter oxydans 621H*	Fruits, Plants, Wine (Germany)	Isolate	YP_191628.1	2.5	(*S83*)
Nov	*Novosphingobium aromaticivorans DSM 12444*	Fresh water, Soil (S. Carolina)	Isolate	YP_497986.1	2.5	-
Mat1	Hypersaline mat	Mexico	Metagenome	2004359243	2.5	(*S84*)
Mat2	Hypersaline mat	Mexico	Metagenome	2004346681	2.5	(*S84*)
Mat3	Hypersaline mat	Mexico	Metagenome	2004362568	2.5	(*S84*)
SSU rRNA gene – isolates						
ZAS2	*Z. angusticollis /T. primitia*	California	Isolate	AF093252	2.2,2.7	(*S32*)
ZAS9	*Z. angusticollis /T. azotonutricium*	California	Isolate	AF320287	2.2,2.7	(*S33*)
SSU rRNA gene - co-localization and reference library						
A1_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187712	2.2,2.7	this study
A3_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187722	2.2,2.7	this study
A10_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187713	2.2,2.7	this study
A11_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187714	2.2,2.7	this study
A14_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187717	2.2,2.7	this study
B1_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187729	2.2,2.7	this study
C1_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187734	2.2,2.7	this study
C2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187735	2.2,2.7	this study
E1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187739	2.2,2.7	this study
F1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187742	2.2,2.7	this study
G1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187744	2.2,2.7	this study
G3_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187746	2.2,2.7	this study

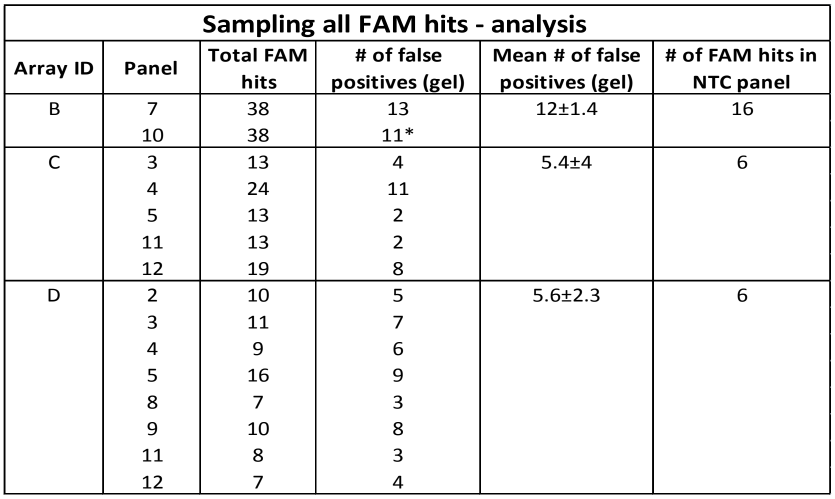
G5_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187748	2.2,2.7	this study
A4_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187723	2.2,2.7	this study
A5_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187724	2.2,2.7	this study
A7_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187726	2.2,2.7	this study
A9_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187728	2.2,2.7	this study
A12_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187715	2.2,2.7	this study
A13_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187716	2.2,2.7	this study
B2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187730	2.2,2.7	this study
E2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187740	2.2,2.7	this study
A6_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187725	2.2,2.7	this study
F2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187743	2.2,2.7	this study
G2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187745	2.2,2.7	this study
G4_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187747	2.2,2.7	this study
A2_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187721	2.2,2.7	this study
A8_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187727	2.2,2.7	this study
B3_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187731	2.2,2.7	this study
A15_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187718	2.7	this study
D2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187738	2.7	this study
A16_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187719	2.7	this study
E3_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187741	2.7	this study
A18_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187720	2.7	this study
B5_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187733	2.7	this study
G6_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187749	2.7	this study
C3_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187736	2.7	this study
C4_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187737	2.7	this study
G7_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187750	2.7	this study
B4_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187732	2.7	this study
rF100	*Reticulitermes hesperus*	California	Digital PCR	HQ187634	2.7	this study
rF12	*Reticulitermes hesperus*	California	Digital PCR	HQ187635	2.7	this study
rF13	*Reticulitermes hesperus*	California	Digital PCR	HQ187636	2.7	this study
rF14	*Reticulitermes hesperus*	California	Digital PCR	HQ187637	2.7	this study
rF16	*Reticulitermes hesperus*	California	Digital PCR	HQ187638	2.7	this study
rF21	*Reticulitermes hesperus*	California	Digital PCR	HQ187639	2.7	this study
rF22	*Reticulitermes hesperus*	California	Digital PCR	HQ187640	2.7	this study
rF23	*Reticulitermes hesperus*	California	Digital PCR	HQ187641	2.7	this study
rF24	*Reticulitermes hesperus*	California	Digital PCR	HQ187642	2.7	this study
rF26	*Reticulitermes hesperus*	California	Digital PCR	HQ187643	2.7	this study
rF28	*Reticulitermes hesperus*	California	Digital PCR	HQ187644	2.7	this study
rF29	*Reticulitermes hesperus*	California	Digital PCR	HQ187645	2.7	this study
rF3	*Reticulitermes hesperus*	California	Digital PCR	HQ187646	2.7	this study
rF33	*Reticulitermes hesperus*	California	Digital PCR	HQ187647	2.7	this study

rF39	*Reticulitermes hesperus*	California	Digital PCR	HQ187648	2.7	this study
rF40	*Reticulitermes hesperus*	California	Digital PCR	HQ187649	2.7	this study
rF43	*Reticulitermes hesperus*	California	Digital PCR	HQ187650	2.7	this study
rF47	*Reticulitermes hesperus*	California	Digital PCR	HQ187651	2.7	this study
rF5	*Reticulitermes hesperus*	California	Digital PCR	HQ187652	2.7	this study
rF52	*Reticulitermes hesperus*	California	Digital PCR	HQ187653	2.7	this study
rF56	*Reticulitermes hesperus*	California	Digital PCR	HQ187654	2.7	this study
rF61	*Reticulitermes hesperus*	California	Digital PCR	HQ187655	2.7	this study
rF63	*Reticulitermes hesperus*	California	Digital PCR	HQ187656	2.7	this study
rF68	*Reticulitermes hesperus*	California	Digital PCR	HQ187657	2.7	this study
rF69	*Reticulitermes hesperus*	California	Digital PCR	HQ187658	2.7	this study
rF71	*Reticulitermes hesperus*	California	Digital PCR	HQ187659	2.7	this study
rF75	*Reticulitermes hesperus*	California	Digital PCR	HQ187660	2.7	this study
rF77	*Reticulitermes hesperus*	California	Digital PCR	HQ187661	2.7	this study
rF79	*Reticulitermes hesperus*	California	Digital PCR	HQ187662	2.7	this study
rF8	*Reticulitermes hesperus*	California	Digital PCR	HQ187663	2.7	this study
rF81	*Reticulitermes hesperus*	California	Digital PCR	HQ187664	2.7	this study
rF82	*Reticulitermes hesperus*	California	Digital PCR	HQ187665	2.7	this study
rF83	*Reticulitermes hesperus*	California	Digital PCR	HQ187666	2.7	this study
rF84	*Reticulitermes hesperus*	California	Digital PCR	HQ187667	2.7	this study
rF86	*Reticulitermes hesperus*	California	Digital PCR	HQ187668	2.7	this study
rF88	*Reticulitermes hesperus*	California	Digital PCR	HQ187669	2.7	this study
rF92	*Reticulitermes hesperus*	California	Digital PCR	HQ187670	2.7	this study
rF93	*Reticulitermes hesperus*	California	Digital PCR	HQ187671	2.7	this study
rF94	*Reticulitermes hesperus*	California	Digital PCR	HQ187672	2.7	this study
rF95	*Reticulitermes hesperus*	California	Digital PCR	HQ187673	2.7	this study
rG11	*Reticulitermes hesperus*	California	Digital PCR	HQ187674	2.7	this study
rG14	*Reticulitermes hesperus*	California	Digital PCR	HQ187675	2.7	this study
rG15	*Reticulitermes hesperus*	California	Digital PCR	HQ187676	2.7	this study
rG16	*Reticulitermes hesperus*	California	Digital PCR	HQ187677	2.7	this study
rG17	*Reticulitermes hesperus*	California	Digital PCR	HQ187678	2.7	this study
rG20	*Reticulitermes hesperus*	California	Digital PCR	HQ187679	2.7	this study
rG24	*Reticulitermes hesperus*	California	Digital PCR	HQ187680	2.7	this study
rG25	*Reticulitermes hesperus*	California	Digital PCR	HQ187681	2.7	this study
rG28	*Reticulitermes hesperus*	California	Digital PCR	HQ187682	2.7	this study
rG29	*Reticulitermes hesperus*	California	Digital PCR	HQ187683	2.7	this study
rG3	*Reticulitermes hesperus*	California	Digital PCR	HQ187684	2.7	this study
rG30	*Reticulitermes hesperus*	California	Digital PCR	HQ187685	2.7	this study
rG31	*Reticulitermes hesperus*	California	Digital PCR	HQ187686	2.7	this study
rG32	*Reticulitermes hesperus*	California	Digital PCR	HQ187687	2.7	this study
rG35	*Reticulitermes hesperus*	California	Digital PCR	HQ187688	2.7	this study

rG36	*Reticulitermes hesperus*	California	Digital PCR	HQ187689	2.7	this study
rG41	*Reticulitermes hesperus*	California	Digital PCR	HQ187690	2.7	this study
rG42	*Reticulitermes hesperus*	California	Digital PCR	HQ187691	2.7	this study
rG43	*Reticulitermes hesperus*	California	Digital PCR	HQ187692	2.7	this study
rG49	*Reticulitermes hesperus*	California	Digital PCR	HQ187693	2.7	this study
rG50	*Reticulitermes hesperus*	California	Digital PCR	HQ187694	2.7	this study
rG53	*Reticulitermes hesperus*	California	Digital PCR	HQ187695	2.7	this study
rG55	*Reticulitermes hesperus*	California	Digital PCR	HQ187696	2.7	this study
rG60	*Reticulitermes hesperus*	California	Digital PCR	HQ187697	2.7	this study
rG63	*Reticulitermes hesperus*	California	Digital PCR	HQ187698	2.7	this study
rG71	*Reticulitermes hesperus*	California	Digital PCR	HQ187699	2.7	this study
rG72	*Reticulitermes hesperus*	California	Digital PCR	HQ187700	2.7	this study
rG73	*Reticulitermes hesperus*	California	Digital PCR	HQ187701	2.7	this study
rG74	*Reticulitermes hesperus*	California	Digital PCR	HQ187702	2.7	this study
rG78	*Reticulitermes hesperus*	California	Digital PCR	HQ187703	2.7	this study
rG80	*Reticulitermes hesperus*	California	Digital PCR	HQ187704	2.7	this study
rG81	*Reticulitermes hesperus*	California	Digital PCR	HQ187705	2.7	this study
rG83	*Reticulitermes hesperus*	California	Digital PCR	HQ187706	2.7	this study
rG86	*Reticulitermes hesperus*	California	Digital PCR	HQ187707	2.7	this study
rG88	*Reticulitermes hesperus*	California	Digital PCR	HQ187708	2.7	this study
rG9	*Reticulitermes hesperus*	California	Digital PCR	HQ187709	2.7	this study
rG91	*Reticulitermes hesperus*	California	Digital PCR	HQ187710	2.7	this study
rG95	*Reticulitermes hesperus*	California	Digital PCR	HQ187711	2.7	this study
SSU rRNA gene - close relatives and other termite related						
unc Trep clone RFS84	*Reticulitermes flavipes*	Michigan	PCR	AF068428	2.7	(*S64*)
unc Trep clone RFS99	*Reticulitermes flavipes*	Michigan	PCR	AF068424	2.7	(*S64*)
unc Trep clone RFS94	*Reticulitermes flavipes*	Michigan	PCR	AF068423	2.7	(*S64*)
unc Trep clone RFS21	*Reticulitermes flavipes*	Michigan	PCR	AF068338	2.7	(*S64*)
unc Trep clone RFS12	*Reticulitermes flavipes*	Michigan	PCR	AF068335	2.7	(*S64*)
unc Trep clone RFS2	*Reticulitermes flavipes*	Michigan	PCR	AF068429	2.7	(*S64*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB192140	2.7	(*S85*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB192197	2.7	(*S85*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB192202	2.7	(*S85*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB192142	2.7	(*S85*)
unc Trep sp.	*Reticulitermes sp.*	Asia	PCR	AB192251	2.7	(*S85*)
unc Trep sp.	*Reticulitermes sp.*	Asia	PCR	AB192248	2.7	(*S85*)
unc Trep sp.	*Reticulitermes sp.*	Asia	PCR	AB192247	2.7	(*S85*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088870	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088896	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088915	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088876	2.7	(*S73*)

unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088895	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088866	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088874	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088890	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088878	2.7	(*S73*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088909	2.7	(*S73*)
unc Trep clone HsDiSp314	*Hodotermopsis sjoestedti*	Asia	PCR	AB032005	2.7	(*S86*)
SSU rRNA gene - non termite related						
Treponema vincentii (D2A-2)		Oral cavity	isolate	AY119690	2.7	(*S87*)
Treponema denticola (ATCC 35405)		Oral cavity	isolate	AE017226	2.7	(*S88*)
Treponema pallidum (Nichols)		Human genital tract	isolate	AE000520	2.7	(*S89*)
Treponema zioleckii (kT)		Sheep rumen	isolate	DQ065758	2.7	(*S90*)
Treponema socranskii (socranskii)		Oral cavity	isolate	AF033306	2.7	(*S91*)
Treponema succinifaciens		Pig colon	isolate	M57738	2.7	(*S92*)
Brevinema andersonii		Shrews and mice	isolate	L31543	2.7	(*S93*)
Borrelia burgdorferi (DK7)		Ticks, deer and humans	isolate	X85195	2.7	(*S94*)
Spirochaeta aurantia (M1)		Fresh water	isolate	AY599019	2.7	(*S95*)
Escherichia coli K-12 MG1655		-	isolate	U00096	2.7	(*S96*)

Table 2.12. Analysis of all FAM hits for a number of microfluidic array panels. For several microfluidic array panels, all chambers exhibiting amplification in the FAM fluorescence channel were retrieved, post-amplified and analyzed by agarose gel electrophoresis. In this table we show the total number of chambers that exhibited FAM fluorescence on the given panel ("Total FAM hits"), the number of false positives based on analysis by agarose gel electrophoresis ("# of false positives"), the mean number of false positives per array ("Mean # of false positives"), and the average number of chambers that exhibited FAM fluorescence in the no-template-control panel on the same array ("# of FAM hits in NTC panel"). The mean number of false positive hits agrees well with the number of hits in the corresponding no-template-control panel indicating the latter is a good predictor of the former. See supporting text for further details.



Sampling all FAM hits - analysis

Array ID	Panel	Total FAM hits	# of false positives (gel)	Mean # of false positives (gel)	# of FAM hits in NTC panel
B	7	38	13	12±1.4	16
	10	38	11*		
C	3	13	4	5.4±4	6
	4	24	11		
	5	13	2		
	11	13	2		
	12	19	8		
D	2	10	5	5.6±2.3	6
	3	11	7		
	4	9	6		
	5	16	9		
	8	7	3		
	9	10	8		
	11	8	3		
	12	7	4		

* 3 retrievals were not tested due to an experimental problem

Table 2.13. Definition of variables used in the microfluidic array statistical model. See supporting text for further details.

Variable	Definition	Estimation method
X	Number of FAM hits per panel	Measured
Y	Number of HEX hits per panel	Measured
I	Number of wells per panel with both a FAM hit and a HEX hit (i.e. co-localization)	Measured
noise	Number of FAM hits that are due to spurious amplification	Measured
f_S	Frequency of ribotype ***S*** on the chip	Measured
$\varepsilon_{ter/16S}$	Terminase/16S primer efficiency	Measured
X_f	Number of non co-localized FAM events	$X_f = X - I$
p_{ter}	The probability that a given well will contain a free floating terminase target	Eq. S1
I_{S}	Average number of free floating terminase targets to co-localize with a *particular* 16S rRNA ribotype ***S*** on a panel	Eq. S2a
$I_{\text{all 16S-ter}}$	Average number of any terminase target to co-localize with any 16S rRNA target on a panel	Eq. S2b
p_F	Probability that a successful retrieval from a panel contains a particular ribotype ***S*** and any terminase gene by chance	Eq. S3
X_T	Sum of the total number of free floating terminase targets and spurious targets	Eq. S4
N_{false}	Expected number of false co-localizations in the dataset	Eq. S5
p_S	Probability that a successful retrieval will contain host ***S***	Eq. S6

Table 2.14. Statistics for all sampled panels. This table lists for each ribotype in Fig. 2.2 the panel from which the ribotype was retrieved, the number of FAM hits X on that panel, the number of HEX hits Y on that panel, their intersection I, the number of FAM hits found in the no-target-control-panel for the microfluidic array containing the given panel (*noise*), the frequency of this host in the reference rRNA library, f_s (based on Table 2.6), the estimated probability for false co-localization p_F (Eq. S3), and the P value (one-tailed test, n=41) for each host for obtaining at least the number of observed co-localizations by chance (based on the data in Table 2.1). The statistical test to determine the P value is explained in the supporting text. Chip analysis was performed using the Fluidigm Digital PCR Analysis software v.2.1.1 with the linear baseline correction. See supporting text for further details.

#	Retrieval ID (*n* =41)	Host	chip	panel	X (FAM)	Y (HEX)	I	*noise* (FAM)	X_T-*noise*	p_{ter}	$I_{all16S\text{-}ter}$	f_S	p_F	P value (*n* =41)
1	A1_1	I	A	3	22	38	2	15	6.0	7.9E-03	1.3	4.2%	7.77E-03	5.45E-18
2	A3_1	I	A	5	33	66	8	15	12.4	1.6E-02	6.7			
3	A10_1	I	A	8	40	59	12	15	15.3	2.0E-02	10.8			
4	A11_1	I	A	9	34	46	9	15	11.6	1.5E-02	8.1			
5	A14_1	I	A	10	30	46	11	15	5.2	6.8E-03	10.1			
6	B1_2	I	B	10	42	52	5	20	19.7	2.6E-02	3.6			
7	C1_2	I	C	11	13	55	3	6	4.8	6.2E-03	2.6			
8	C2_2	I	C	5	13	69	4	6	3.9	5.1E-03	3.5			
9	E1_3	I	E	2	14	21	2	5	7.3	9.6E-03	1.9			
10	F1_3	I	F	3	22	32	2	7	13.9	1.8E-02	1.7			
11	G1_3	I	G	3	12	51	4	6	2.6	3.4E-03	3.6			
12	G3_3	I	G	8	17	33	2	6	9.7	1.3E-02	1.7			
13	G5_3	I	G	11	14	26	1	6	7.5	9.7E-03	0.8			
14	A4_1	II	A	6	54	79	10	15	34.1	4.5E-02	8.5	1.7%	3.11E-03	7.63E-13
15	A5_1	II	A	6	54	79	10	15	34.1	4.5E-02	8.5			
16	A7_1	II	A	8	40	59	12	15	15.3	2.0E-02	10.8			
17	A9_1	II	A	8	40	59	12	15	15.3	2.0E-02	10.8			
18	A12_1	II	A	10	30	46	11	15	5.2	6.8E-03	10.1			
19	A13_1	II	A	10	30	46	11	15	5.2	6.8E-03	10.1			
20	B2_2	II	B	10	42	52	5	20	19.7	2.6E-02	3.6			
21	E2_3	II	E	2	14	21	2	5	7.3	9.6E-03	1.9			
22	A6_1	III	A	7	40	66	8	15	20.0	2.6E-02	6.7	0.9%	1.55E-03	5.65E-07
23	F2_3	III	F	8	21	34	6	7	8.7	1.1E-02	5.7			
24	G2_3	III	G	4	20	53	3	6	12.3	1.6E-02	2.6			
25	G4_3	III	G	10	19	36	2	6	11.8	1.5E-02	1.7			
26	A2_1	IV	A	4	46	129	17	15	19.9	2.6E-02	14.5	0.9%	1.55E-03	3.83E-05
27	A8_1	IV	A	8	40	59	12	15	15.3	2.0E-02	10.8			
28	B3_2	IV	B	10	42	52	5	20	19.7	2.6E-02	3.6			
29	A15_1	-	A	4	46	129	17	15	19.9	2.6E-02	14.5	-	-	-
30	A16_1	-	A	5	33	66	8	15	12.4	1.6E-02	6.7	-	-	-
31	A17_1	-	A	10	30	46	11	15	5.2	6.8E-03	10.1	-	-	-
32	A18_1	-	A	11	27	84	7	15	7.5	9.8E-03	5.4	-	-	-
33	B5_2	-	B	7	46	53	11	20	17.6	2.3E-02	9.6	-	-	-
34	B4_2	-	B	7	46	53	11	20	17.6	2.3E-02	9.6	-	-	-
35	C3_2	-	C	11	13	55	3	6	4.8	6.2E-03	2.6	-	-	-
36	C4_2	-	C	11	13	55	3	6	4.8	6.2E-03	2.6	-	-	-
37	D1_2	-	D	4	9	24	1	8	0.3	3.4E-04	0.7	-	-	-
38	D2_2	-	D	3	11	26	1	8	2.4	3.1E-03	0.7	-	-	-
39	E3_3	-	E	11	12	24	2	5	5.3	7.0E-03	1.8	-	-	-
40	G6_3	-	G	4	20	53	3	6	12.3	1.6E-02	2.6	-	-	-
41	G7_3	-	G	4	20	53	3	6	12.3	1.6E-02	2.6	-	-	-

2.9.5 References

*S*1. E. Ottesen, J. Hong, S. Quake, J. Leadbetter, *Science* **314**, 1464 (2006).

*S*2. E. Ottesen, PhD thesis, California Institute of Technology (2008).

*S*3. J. Austin, A. Szalanski, B. Cabrera, *Ann. Entomol. Soc. Am.* **97**, 548 (2004).

*S*4. M. Ohkuma et al., *Mol. Phylogenet. Evol.* **31**, 701 (2004).

*S*5. F. Warnecke et al., *Nature* **450**, 560 (2007).

*S*6. K. Maekawa, N. Lo, O. Kitade, T. Miura, T. Matsumoto, *Mol. Phylogenet. Evol.* **13**, 360 (1999).

*S*7. H. Liu, A. Beckenbach, *Mol. Phylogenet. Evol.* **1**, 41 (1992).

*S*8. K. Ashelford, N. Chuzhanova, J. Fry, A. Jones, A. Weightman, *Appl. Environ. Microbiol.* **71**, 7724 (2005).

*S*9. T. Huber, G. Faulkner, P. Hugenholtz, *Bioinformatics* **20**, 2317 (2004).

*S*10. T. DeSantis et al., *Appl. Environ. Microbiol.* **72**, 5069 (2006).

*S*11. E. Pruesse et al., *Nucleic Acids Res.* **35**, 7188 (2007).

*S*12. W. Ludwig et al., *Nucleic Acids Res.* **32**, 1363 (2004).

*S*13. D. Posada, *Mol. Biol. Evol.* **25**, 1253 (2008).

*S*14. S. Guindon, O. Gascuel, *Syst. Biol.* **52**, 696 (2003).

*S*15. K. Tamura, M. Nei, *Mol. Biol. Evol.* **10**, 512 (1993).

*S*16. J. Felsenstein, *Phylogeny inference package (PHYLIP), Version 3.6 a3* (2002).

*S*17. W. Fitch, E. Margoliash, *Science* **155**, 279 (1967).

*S*18. J. Thompson, D. Higgins, T. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).

*S*19. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol. Biol. Evol.* **24**, 1596 (2007).

*S*20. L. Heath, E. Van Der Walt, A. Varsani, D. Martin, *J. Virol.* **80**, 11827 (2006).

*S*21. M. Padidam, S. Sawyer, C. Fauquet, *Virology* **265**, 218 (1999).

*S*22. J. Smith, *J. Mol. Evol.* **34**, 126 (1992).

*S*23. D. Martin, E. Rybicki, *Bioinformatics* **16**, 562 (2000).

*S*24. D. Posada, *Mol. Biol. Evol.* **19**, 708 (2002).

*S*25. D. Posada, K. Crandall, *Proc. Natl. Acad. Sci. USA* **98**, 13757 (2001).

*S*26. M. O. Salminen, J. K. Carr, D. S. Burke, F. E. McCutchan, *AIDS Res. Hum. Retroviruses* **11**, 1423 (1995).

*S*27. M. Salminen, D. Marin, in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing,* P. Lemey, M. Salemi, A. Vandamme, Eds. (Cambridge University Press; 2nd edition, 2010), pp. 519-548.

*S*28. D. Huson, D. Bryant, *Mol. Biol. Evol.* **23**, 254 (2006).

*S*29. N. Tao et al., *FINDMODEL: a tool to select the best-fit model of nucleotide substitution*, M.S. thesis, University of New Mexico (2005).

*S*30. K. Strimmer, A. Von Haeseler, *Proc. Natl. Acad. Sci. USA* **94**, 6815 (1997).

*S*31. H. Schmidt, A. von Haeseler, in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing,* P. Lemey, M. Salemi, A. Vandamme, Eds. (Cambridge University Press; 2nd edition, 2010), pp. 181-209.

*S*32. J. Leadbetter, T. Schmidt, J. Graber, J. Breznak, *Science* **283**, 686 (1999).

*S*33. T. Lilburn et al., *Science* **292**, 2495 (2001).

*S*34. T. Rose et al., *Nucleic Acids Res.* **26**, 1628 (1998).

*S*35. Y. Zhang et al., *Nucleic Acids Res.* **31**, doi: 10.1093/nar/gng123 (2003).

*S*36. C. Corless et al., *J. Clin. Microbiol.* **38**, 1747 (2000).

*S*37. S. Pond, S. Muse, *HyPhy: hypothesis testing using phylogenies*. Statistical Methods in Molecular Evolution (Springer, 2005), pp. 125-181.

*S*38. S. Pond, S. Muse, in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing,* P. Lemey, M. Salemi, A. Vandamme, Eds. (Cambridge University Press; 2nd edition, 2010), pp. 419-490.

*S*39. S. Muse, B. Gaut, *Mol. Biol. Evol.* **11**, 715 (1994).

*S*40. S. Altschul et al., *Nucleic Acids Res.* **25**, 3389 (1997).

*S*41. K. Pruitt, T. Tatusova, D. Maglott, *Nucleic Acids Res.* **33**, D501 (2005).

*S*42. S. Casjens, *Mol. Microbiol.* **49**, 277 (2003).

*S*43. S. Dube, J. Qin, R. Ramakrishnan, *PLoS ONE* **3**, doi:10.1371/journal.pone.0002876 (2008).

*S*44. L. Black, *Bioessays* **17**, 1025 (1995).

*S*45. S. Casjens et al., *J. Bacteriol.* **187**, 1091 (2005).

*S*46. H. Ackermann, *Adv. Virus Res.* **51**, 135 (1999).

*S*47. S. R. Casjens, *Res. Microbiol.* **159**, 340 (2008).

*S*48. J. Lawrence, G. Hatfull, R. Hendrix, *J. Bacteriol.* **184**, 4891 (2002).

*S*49. M. Daw, F. Falkiner, *Micron* **27**, 467 (1996).

*S*50. K. Nakayama et al., *Mol. Microbiol.* **38**, 213 (2000).

*S*51. Y. Michel-Briand, C. Baysse, *Biochimie* **84**, 499 (2002).

*S*52. A. Lang, J. Beatty, *Proc. Natl. Acad. Sci. USA* **97**, 859 (2000).

*S*53. A. Lang, J. Beatty, *Arch. Microbiol.* **175**, 241 (2001).

*S*54. A. Lang, J. Beatty, *Trends Microbiol.* **15**, 54 (2007).

*S*55. M. Mitchell, S. Matsuzaki, S. Imai, V. Rao, *Nucleic Acids Res.* **30**, 4009 (2002).

*S*56. A. Davison, *Virology* **186**, 9 (1992).

*S*57. E. Koonin, T. Senkevich, V. Dolja, *Biol. Direct* **1**, 29 (2006).

*S*58. M. Baker, W. Jiang, F. Rixon, W. Chiu, *J. Virol.* **79**, 14967 (2005).

*S*59. S. Kanamaru, K. Kondabagil, M. Rossmann, V. Rao, *J. Biol. Chem.* **279**, 40795 (2004).

*S*60. A. Marchler-Bauer et al., *Nucleic Acids Res.* **33**, D192 (2005).

*S*61. R. Edgar, *BMC bioinformatics* **5**, 113 (2004).

*S*62. V. Rao, M. Feiss, *Annu. Rev. Genet.* **42**, 647 (2008).

*S*63. S. Sun et al., *Cell* **135**, 1251 (2008).

*S*64. S. Lilburn, *Environ. Microbiol.* **1**, 331 (1999).

*S*65. J. Breznak, J. Leadbetter, *Prokaryotes* **7**, 318 (2006).

*S*66. B. Paster et al., *J. Bacteriol.* **173**, 6101 (1991).

*S*67. W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, *Nucleic Acids Res.* **32**, 1363 (2004).

*S*68. D. Bryant, V. Moulton, *Mol. Biol. Evol.* **21**, 255 (2004).

*S*69. M. Kimura, *J. Mol. Evol.* **16**, 111 (1980).

S70. T. Jukes, C. Cantor, *Mammalian protein metabolism* **3**, 21 (1969).

S71. P. Schloss, J. Handelsman, *Appl. Environ. Microbiol.* **71**, 1501 (2005).

*S*72. J. Cole et al., *Nucleic Acids Res.* **37**, doi: 10.1093/nar/gkn879 (2008).

*S*73. Y. Hongoh, M. Ohkuma, T. Kudo, *FEMS Microbiol. Ecol.* **44**, 231 (2003).

*S*74. B. Paster et al., *Appl. Environ. Microbiol.* **62**, 347 (1996).

*S*75. W. Ikeda-Ohtsubo, M. Desai, U. Stingl, A. Brune, *Microbiology* **153**, 3458 (2007).

*S*76. M. Nei, T. Gojobori, *Mol. Biol. Evol.* **3**, 418 (1986).

*S*77. J. Zhang, H. Rosenberg, M. Nei, *Proc. Natl. Acad. Sci. USA* **95**, 3708 (1998).

*S*78. W. Li, C. Wu, C. Luo, *Mol. Biol. Evol.* **2**, 150 (1985).

*S*79. P. Pamilo, N. Bianchi, *Mol. Biol. Evol.* **10**, 271 (1993).

*S*80. M. Nei, S. Kumar, *Molecular evolution and phylogenetics* (Oxford University Press, USA, 2000).

*S*81. E. Miller et al., *Microbiol. Mol. Biol. Rev.* **67**, 86 (2003).

*S*82. W. Reeve et al., *Standards in Genomic Sciences* **2**, 77 (2010).

*S*83. C. Prust et al., *Nat. Biotechnol.* **23**, 195 (2005).

*S*84. V. Kunin et al., *Mol. syst. biol.* **4**, 198 (2008).

*S*85. Y. Hongoh et al., *Appl. Environ. Microbiol.* **71**, 6590 (2005).

*S*86. T. Iida, M. Ohkuma, K. Ohtoko, T. Kudo, *FEMS Microbiol. Ecol.* **34**, 17 (2000).

*S*87. A. Edwards, D. Dymock, M. Woodward, H. Jenkinson, *Microbiology* **149**, 1083 (2003).

*S*88. R. Seshadri et al., *Proc. Natl. Acad. Sci. USA* **101**, 5646 (2004).

*S*89. C. Fraser et al., *Science* **281**, 375 (1998).

*S*90. M. Piknova et al., *FEMS Microbiol. Lett.* **289**, 166 (2008).

*S*91. B. Paster, F. Dewhirst, B. Coleman, C. Lau, R. Ericson, *Int. J. Syst. Evol. Micr.* **48**, 713 (1998).

*S*92. T. Stanton et al., *Int. J. Syst. Bacteriol.* **41**, 50 (1991).

*S*93. D. Defosse, R. Johnson, B. Paster, F. Dewhirst, G. Fraser, *Int. J. Syst. Evol. Micr.* **45**, 78 (1995).

*S*94. M. Theisen et al., *J. Bacteriol.* **177**, 3036 (1995).

*S*95. R. McLaughlin, D. Secko, C. Paul, A. Kropinski, *Can. J. Microbiol.* **50**, 967 (2004).

*S*96. W. Kang, T. Icho, S. Isono, M. Kitakawa, K. Isono, *Mol. Gen. Genet.* **217**, 281 (1989).

S97. A. P. Martin, *Appl. Environ. Microbiol.* **68**, 3673 (2002).

S98. C. Lozupone, M. Hamady, R. Knight, *BMC bioinformatics* **7**, 371 (2006).

S99. M. Hamady, C. Lozupone, R. Knight, *The ISME journal* **4**, 17 (2009).

Chapter 3

MetaCAT — Metagenome Cluster Analysis Tool

3.1 Introduction

Much of what we know in biology today is the result of careful studies of cultivated pure cultures over decades. Yet today it is apparent that natural microbial communities look nothing like microbes cultivated *in vitro*. Microbial communities in nature can be vastly complex assemblies, often containing hundreds of species of bacteria or more, with many forming intricate associations. One example is the higher termite hindgut, which contains a complex microbial community that specialized in lignocellulose degradation [1]. A similar complex microbial community resides in the human gut, with every human harboring about 150 bacterial species, with most bacteria and genes shared among humans [2]. Similar levels of complexities were found in many other environments, from marine to soil to deep sea "whale fall" carcasses [3,4]. Thus, communities are far more complex that any single organism in a pure culture. However, traditional microbiological techniques are limited in their capability to study these communities since it has been estimated that >99% of microbes in nature cannot be cultured *in vitro* [5]. As a result, the field of metagenomics came to the fore. Metagenomics is the study of genetic material recovered directly from environmental samples [6,7]. The field is also referred to as environmental genomics, or community genomics. With the advent of next-generation sequencing, metagenomics enables direct extraction, cloning and sequencing of DNA from their natural environment with protocols optimized to capture unexplored microbial diversity [8,9]. Recently, the viral fraction of microbial communities has come into the spot light, revealing the vast complexity of viral diversity on Earth [10,11,12].

Today, researchers can, at a reasonable cost and effort, obtain a metagenome of the environment they are interested in. Thus, currently the problem has shifted from sampling the diversity in a given environment to developing the bioinformatic tools to analyze this complexity and the enormous amounts of data generated by such studies. Current metagenome analysis tools such as MEGAN [13], CAMERA [14], and MG-RAST [15] focus on the annotation and classification of the gene fragments present in metagenomes. While these tools are essential, they do not provide an annotation-independent census of the various genes present in a metagenome. Thus, there is no available tool that we are aware of that can produce a ranked list of the abundance of all genes in a metagenome automatically, grouping genes based on some similarity criterion without relying on annotation. Such a census can be useful in comparing relative abundances of genes in a given community in a way that does not rely on annotation. Current methods of achieving this goal involve searching for keywords and summing hits manually. Such methods are both not rigorous and depend on the quality of annotation. Furthermore, annotation-based methods do not group genes based on a similarity criterion (indicating homology), thus potentially significantly biasing results. Another use for an annotation-independent census of a metagenome is to collect a diversity of genes for primer design, especially for genes that exhibit significant diversity such as viral genes (see Chapter 2).

Here we present a new tool called MetaCAT (metagenome cluster analysis tool) that is designed to calculate an "abundance spectrum" of known genes present in a given metagenome. This spectrum can be used to quickly ascertain the “major players” in terms of genes present/genes being expressed in the given environment.

The process of annotation of a metagenome involves BLASTing each metagenome gene object against a database of “known reference genes” and searching for the best hit. MetaCAT generates a census by reversing this procedure and BLASTing a database of “known reference genes” against the metagenome, counting the number of hits for each reference gene. The “known reference genes” that MetaCAT uses as a reference when constructing its spectrum can be in principle all currently known genes or a particular subset of these genes, such as all known viral genes, all known mammalian genes, all known plastid genes, and so on. Although there is no restriction on the library of known genes that MetaCAT can use, MetaCAT was designed with NCBI’s RefSeq database in mind. The RefSeq database is intended to be a “stable, consistent, comprehensive, non-redundant database of genomes” [16] and can be downloaded in full or for certain major taxonomic or other logical groups. MetaCAT, in addition to specifying which known genes are present in the metagenome and their abundances, also gives additional information regarding the known genes. This additional information includes a detailed description of the genes and a description of the lineage of the source organism in which these genes appeared. In the cases of viruses for example, such information can be useful in obtaining a quick snapshot of the classes of viruses that may be abundant in the given environment (e.g., tailed phages versus filamentous phages).

The “abundance spectrum” for the metagenome is calculated by inverting the normal BLAST process. In this scheme every known reference gene is BLASTed against the metagenome (instead of vice versa) and the number of significant hits in the metagenome is counted. Thus in principle every gene in the reference library is given a score which is the number of significant hits that that gene received in the metagenome. This list is, in general, very long but can be significantly compressed with little loss of information by noticing that the reference library often contains many reference genes that yield similar “signatures” in the metagenome, making this list partly redundant. Furthermore, many genes in the reference library yield only tenuous homologies and can be discarded by placing an E value threshold for the best alignment of a given reference gene. MetaCAT therefore compresses the reference gene library for a given metagenome by taking advantage of both these factors, thus generating a spectrum containing a tractable list of genes.

3.2 The MetaCAT algorithm

3.2.1 Overview

The objective of MetaCAT, loosely speaking, is the following: (1) given a metagenome, find in an annotation-independent manner all of the clusters of homologous genes within the metagenome, and (2) for each cluster that was found report the number of members (i.e., gene objects) within the cluster and in addition report the best mach to this cluster among all genes present in reference database, such as the RefSeq database. Fig. 3.1 illustrates the end goal of a MetaCAT analysis of a metagenome.

Typically when one executes a BLAST analysis one BLASTs an unknown gene, such as a gene object from a metagenome, against a reference database of known genes, such as the

RefSeq database (Fig. 3.2A). This procedure results in a list of known genes that pass a certain E value cutoff. If one is interested in mapping a metagenome, this process is repeated for every gene object in the metagenome. The key idea behind MetaCAT is that instead of BLASTing every gene object in the metagenome, MetaCAT BLASTs every known gene in reference database against the metagenome (Fig. 3.2B).

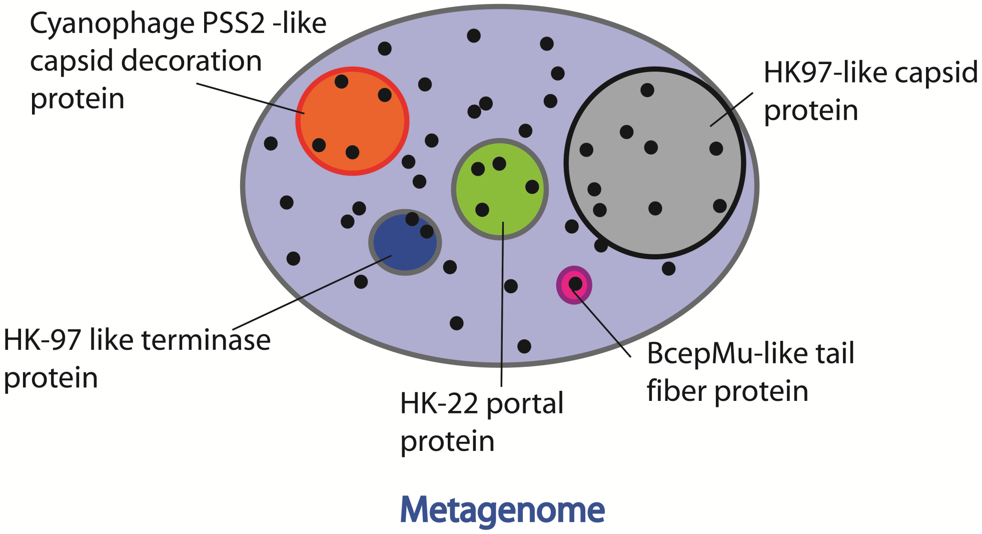

Cyanophage PSS2 -like capsid decoration protein
HK97-like capsid protein
HK-97 like terminase protein
HK-22 portal protein
BcepMu-like tail fiber protein
Metagenome

Figure 3.1 Ideal clustering of gene objects in a metagenome. Each dot represents a gene object in a metagenome, with the entire metagenome depicted by the blue oval. Similar genes are grouped into clusters (circles of different colors) and each cluster is represented by a single gene from a known reference database. In this schematic description the distance between dots is interpreted in an abstract manner and does not correspond to a rigorous metric.

We refer to this "inverse" BLAST analysis as an iBLAST transformation. Each known gene that is iBLASTed results in a list of metagenome gene objects that pass a certain E value threshold. This list is referred to as the "coverage" of the particular known gene in the metagenome. The number of gene objects in this list is interpreted as the abundance of the particular known gene in the metagenome. This process is repeated for every known gene in the reference database. The result is a rather long table that describes for every gene in the reference database its abundance in the metagenome.

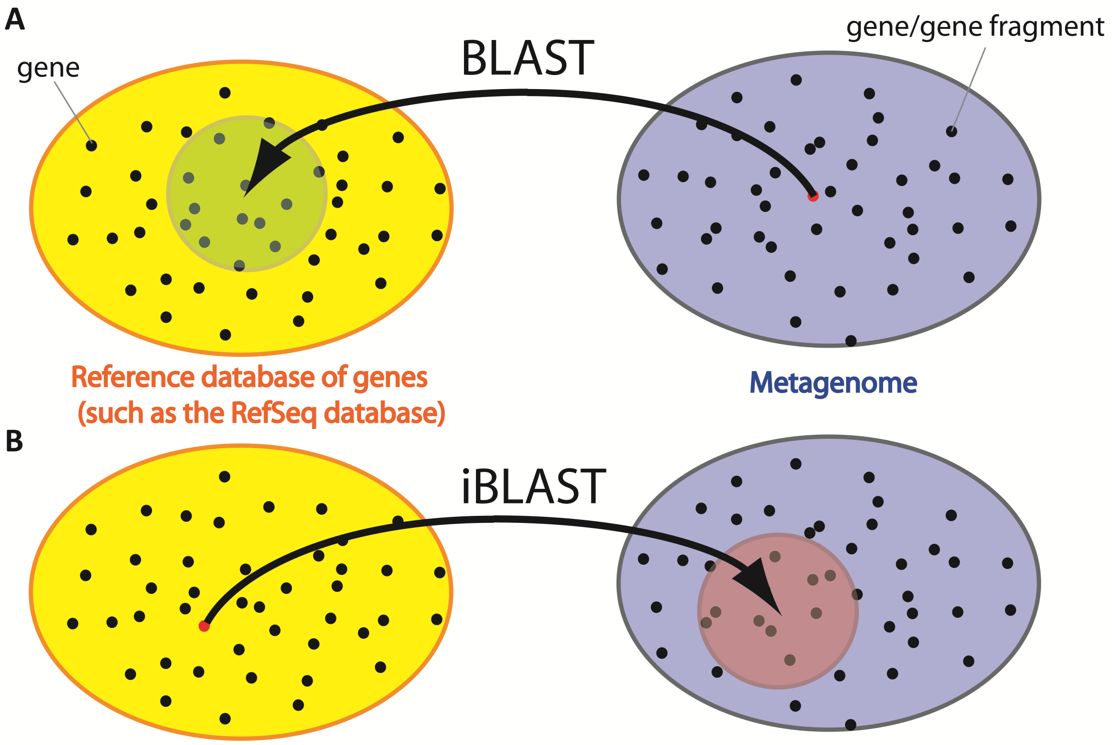


Figure 3.2 An illustration of a BLAST and an iBLAST analysis. A. Typically when performing a BLAST analysis a novel gene object from the metagenome (blue oval) is BLASTed against a reference database of known genes, such as the RefSeq database (yellow oval). The result of the BLAST analysis is a list of "hits" that pass a certain E value threshold. **B.** In an inverse BLAST analysis ("iBLAST" for short) a gene from a known database is BLASTed against the metagenome. The corresponding list of "hits" that pass a certain E value threshold is defined as the "coverage" or "abundance" of the particular known gene in the metagenome.

The list of known genes can be quite long. For example, there are currently approximately 80,000 known viral genes in the RefSeq viral database. MetaCAT compresses this list in two stages. The first filtering stage to impose an E value thresholds for the iBLAST: MetaCAT rejects a gene from the reference database if the *lowest* E value is not low enough. In this way known genes that are remote homologs of every metagenome gene object are automatically discarded from further analysis. The second filtering step has to do with removing redundancy from the reference database. Many genes in this reference database can be close homologs. Close homologs can correspond to very similar clusters of gene objects in the metagenome (Fig. 3.3). Ideally MetaCAT would report and reject all of the close homologs in the reference database and report a single gene, whose E value is the lowest with respect to the metagenome gene objects. MetaCAT identifies related genes in the reference database (red dots in yellow oval in Fig. 3.3) by comparing the overlap of the metagenome gene object (overlapping red circles in the blue oval in Fig. 3.3). If the overlap exceeds a certain threshold the genes in the reference database are said to be "related". Once all related genes are found, only one gene is chosen to represent the group. Though the resulting list of genes from the reference library may still have residual redundancy (see below), this step significantly removes a great deal of redundancy from this database. Note that declaring that two known genes are related by comparing their overlap in the metagenome database is more general that comparing the homology of the genes directly, since in the latter case the reduction is performed independently of the metagenome, and therefore there is information loss. An illustration of a final clustering of genes by MetaCAT is given in Fig. 3.4.

The MetaCAT algorithm is described in detail in the following Section, and a guide for how to use this tool and interpret results is presented in Section 3.3. Installation instructions are given in Section 3.4. An example of a MetaCAT run is given in Chapter 2, Table 2.2.

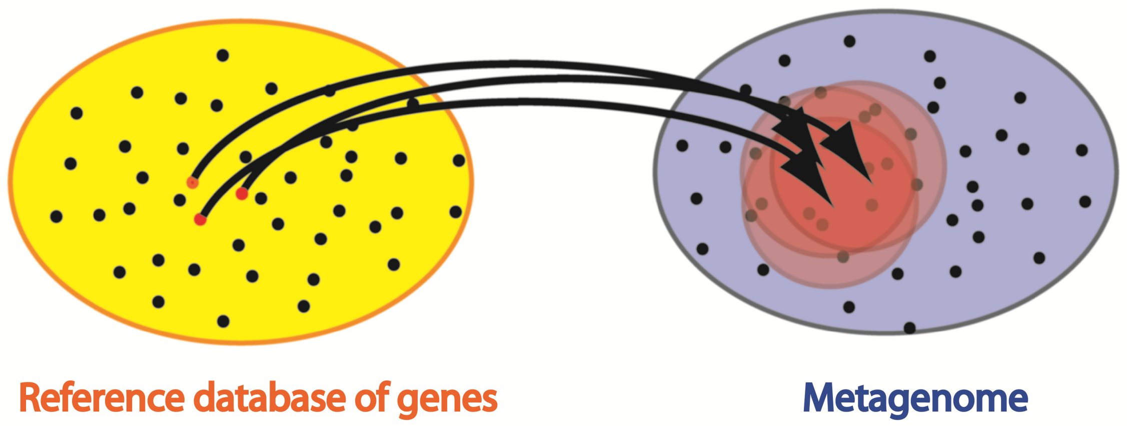


Figure 3.3 Coverage overlap in a metagenome. Similar genes in the reference database (e.g., closely related homologous genes) can have an overlapping coverage in the metagenome. MetaCAT can identify this overlap and will consequently report only one of the reference genes (the one with the lowest E value).

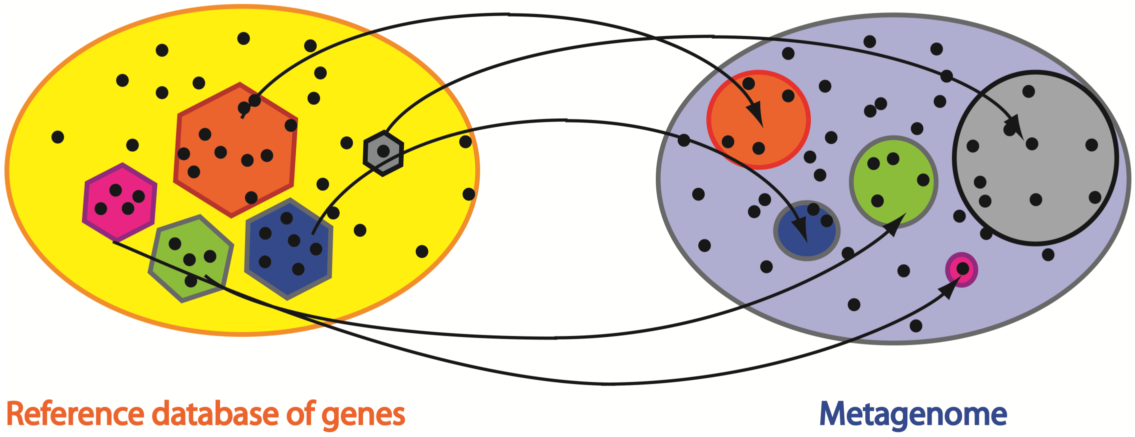


Figure 3.4 Illustration of a final MetaCAT analysis. At the end of the analysis, MetaCAT has identified all known genes that result in overlapping coverage, reporting one representative for each such group. In this manner the apparent redundancy of the reference database is significantly compressed (though the current algorithm used by MetaCAT does not remove this redundancy totally — see discussion below).

3.2.2 The MetaCAT algorithm in detail

MetaCAT analyzes a metagenome through the following sequence of steps:

1. BLAST analysis. The amino acid sequence of every known gene from a reference database K_i (*i=1..N*) — which can be for example a RefSeq database [16] — is BLASTed using NCBI's blastp (v2.2.22+) [17] against the amino acid sequences of all gene objects in the metagenome M_j (*j=1..M*). All BLAST hits must be lower than a maximal E value threshold of 10^{-3}. All alignment information is stored by blastp in a table. This step involves *N*X*M* alignments.

2. Extracting best E value scores and abundances of known reference genes in metagenome. MetaCAT reads the resulting BLAST table and for each known reference gene (KRG) finds the lowest E value score $E_i^{\min}$ and the number of different metagenome gene objects (MGO) n_i (*i=1..N*) that were equal or lower than this E value score. n_i is said to be the *footprint* of K_i in the metagenome. Each KRG, K_i, is therefore homologous to n_i MGOs. The list of n_i elements will hereafter be designated as the *signature* of K_i in the metagenome. A *footprint* is therefore defined as the number of elements in a *signature*.

3. E value filtering of the known reference gene database. We wish to keep only the KRGs that yielded reasonable alignments to MGOs, since we do not want to be concerned with KRGs yielding tenuous similarities. We therefore discard all KRGs whose best E value exceeded a maximal E value threshold (with a default threshold of 10^{-7}), i.e., we require that $E_i^{\min} \le E_{th} = 10^{-7}$. This filtering step will reduce the number of KRGs from *N* to *N'*.

4. Clustering known reference genes. Two KRGs with similar *signatures* are said to be *related* (the measure of similarity will be defined below). Once we identify for given KRG all of the other KRGs to which it is *related*, instead of declaring the given KRG, MetaCAT will declare out of all the *related* KRGs, the KRG with the lowest E value. This filtering method is conservative in the sense that every KRG that is omitted from the final list of declared reference genes is represented by a different KRG that has a very similar metagenome *signature* but has a better E value score, thus there is essentially no loss of information. The outcome of this process is reduction of the number of reference genes that MetaCAT reports by one if the reported (or "declared") KRG is different from the given KRG.

More specifically: for each remaining KRG, K_i (*i=1..N'*), MetaCAT finds all *related* KRGs $\{K_i\}$. Two KRGs are said to be *related* if their corresponding signatures in the metagenome share $P_{th} = 50\%$ of their elements, i.e., if L_i is the number of MGOs homologous to K_i, and L_j is the number of MGOs homologous to K_j, then K_i and K_j are said to be related if and only if P_{th} ≤ $100 \cdot \min\left(L_i \cap L_j / L_i , L_i \cap L_j / L_j\right)$. The stringent *min* function ensures that the overlap between MGOs is normalized by the length of the longer list of MGOs. This prevents relating two KRGs in situations for example where one signature list is very short and is included in a second, very long signature list (which would yield 100% if the *max* function was used). This stringent definition of overlap ensures that all related genes have roughly similar footprints in the metagenome.

Note that the group $\{K_i\}$ will always include K_i since K_i is always related to itself by definition. Each KRG of the remaining KRGs (*i=1..N'*) is then said to "declare" one element of the group $\{K_i\}$ to represent this group. The element that is chosen to represent the group is the KRG with the lowest E value. In case more than one element has the same E value then the following criteria are tested sequentially: highest percent identity, highest number of identical amino acids, highest percent of gene length aligned. If all measures are equal (that can happen if the KRGs have identical amino acid sequences) then to prevent dependence on the order of the genes in the reference library, the KRGs are sorted by their FASTA gene name and the first one is selected.

This clustering step involves *N'* X *N'* comparisons. The group of KRGs that are "declared" K_i *(i=1..N'')* are the final list of reference genes reported by MetaCAT that are said to be found in the metagenome. If two or more KRGs declare the same gene, then that gene appears only once in the final output (this is the compression stage). The total number of KRG declared by MetaCAT will therefore satisfy *N''<N'<N* and typically *N''<<N*. The abundance of each declared KRG K_i in the metagenome is then simply its *footprint* n_i .

Predictably of the algorithm to changing E value thresholds

The algorithm behaves in a predictable fashion when changing E_{th}. For example, if E_{th} is increased from E_1 to, say, E_2, more KRGs are declared in the final list, however these additional KRGs will simply add to the previous declared list when $E_{th} = E_1$. This is because while increasing E_{th} may add additional KRGs to the list of *related* genes $\{K_i\}$ of a given KRG, these additional *related* genes will (by definition) have lower best E value scores, and therefore will not be declared. Thus, increasing the E value threshold can only expand the set of declared KRGs.

3.3 Future directions

Redundancy in the final list of declared reference genes

Note that the final list of declared KRGs can still be redundant. That is, two declared KRGs can still be related, i.e., share > P_{th} of their signature elements. This can happen for example if K_i declares K_j, but K_j, which happens to be related also to K_k, declares K_k. K_j and K_k are thus both declared, yet they are related. Therefore there may still be redundancy present in the final list of declared MGOs that needs to be removed.

We have just seen that how MetaCAT generated a list of *N''* KRGs that can still be redundant, i.e., some KRGs can still have overlapping signatures. We will therefore repeat the clustering algorithm (step 4) until all redundancy is removed. As long as there is redundancy left the clustering will continue to remove nodes. The algorithm will therefore cease when all redundancy is removed (see proof below).

The iterative clustering algorithm discards KRGs at each step, however the KRGs that are discarded have an overlapping signature with one of the remaining KRGs, and therefore in principle no information is lost by this compression.

When does the compression algorithm cease to remove KRGs? The compression algorithm will cease when every node declares itself, a state where by definition there are no more related KRGs and thus no more redundancy. In other words, all redundancy is removed when every node is a local minimum of E value. To see this we note the following: if there are any related KRGs in the final list of declared KRGs (Fig. 3.5) then some node Y will not be at a local E value minimum and will have to declare a node which isn't itself (because if for example node X is a local E value minimum and declares itself, then any node connected to X, like Y, will certainly not declare itself since the E value of X is lower). There are two possibilities at this point. If no other node declares Y then Y will not be declared and the algorithm has compressed the list of KRGs by 1 (Y is not declared). The second possibility is that some other node Z with a higher E value than Y will declare Y. In this case the question would be, is there another node with a higher E value than Z that declares Z. If Z is not declared by any other node than the list has been compressed by 1 (Z is not declared). If there is a node with a higher E value that declares Z we can continue the chain, each time increasing the E value, however at some point we will reach a node which has the highest E value (i.e., a maximum) and therefore will never be declared, resulting in compression of the KRG list. Therefore, as long as there is connectivity in the KRG list, the KRG list will be compressed in the current iteration.

Compression will cease once all connectively is eliminated, that is when each KRG is at local minimum of E values and therefore declares itself.

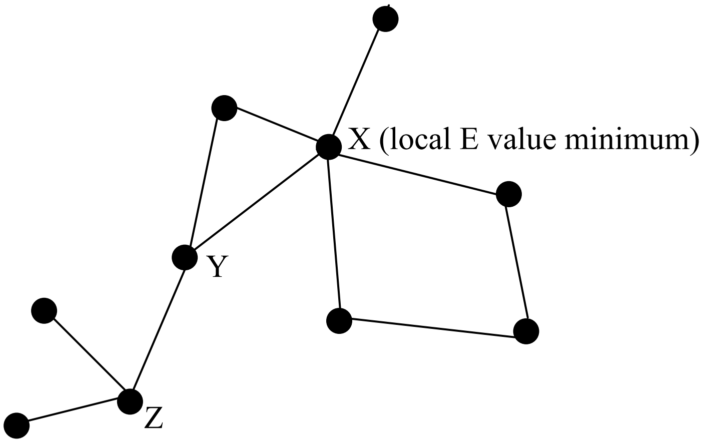


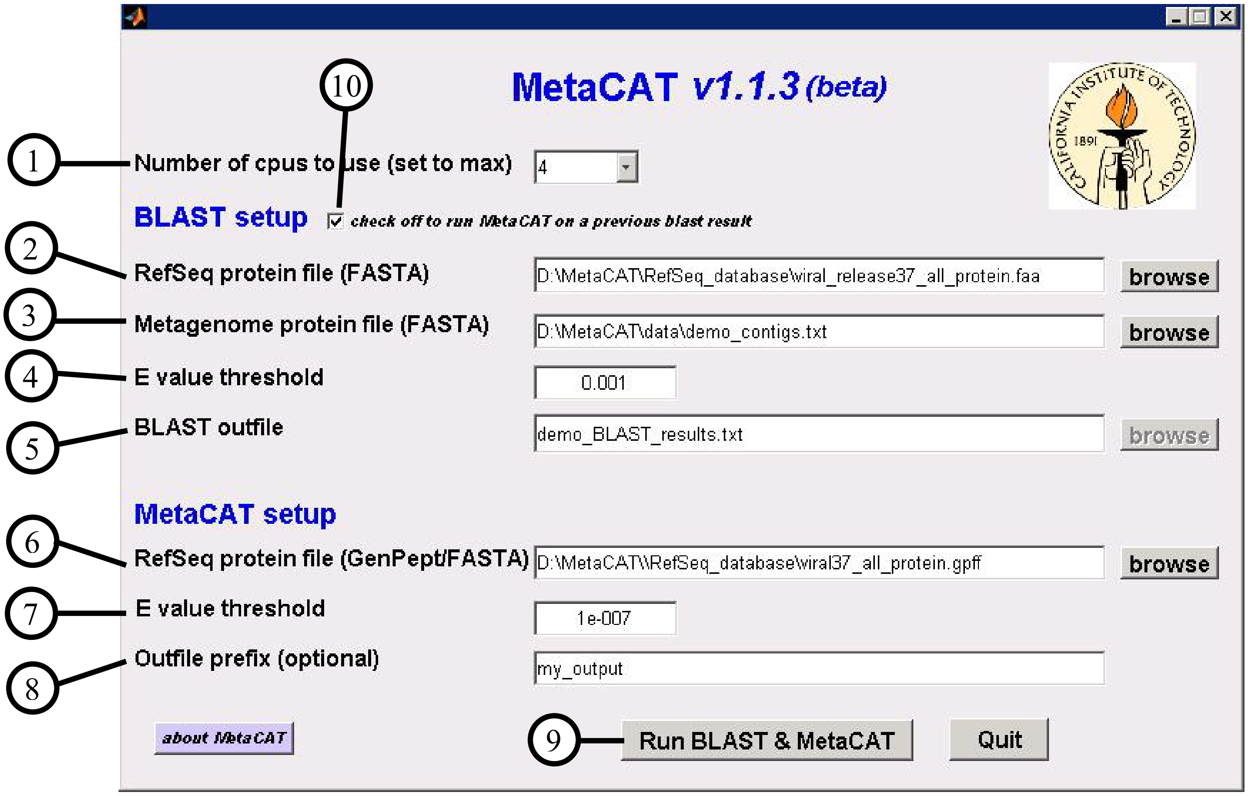
Figure 3.5 Example of list of connected KRGs at one of the clustering iterations

Sample abundance versus metagenome abundance

MetaCAT currently is written to work on the amino acid sequences of metagenome gene objects. These gene objects are the result of reads that have been assembled and translated. The true abundance of a gene object in nature is proportional to the number of reads found in the database and not the number of times this gene object appears in the metagenome. The reason for this discrepancy is that in principle, each gene object should appear only once in the metagenome, since the assembler should automatically remove identical gene objects. For this reason it would be advisable to incorporate into MetaCAT information regarding the abundance of reads. One way to accomplish this would be to use available programs to map reads to gene objects, and then count the number of hits per gene object. For the cases of viruses our prediction is that assembler collapse of contigs did not introduce significant bias to abundances because the viral genes tend to naturally mutate (being part of quasi-species). These mutated gene objects are most likely not collapsed by the assembler, especially given horizontal gene transfer that can affect neighboring genes. This appeared to be the situation in the case of the higher termite metagenome. Terminase alleles in the metagenome were quite divergent (see for example Fig. 2.2). Nevertheless it would be interesting to compare MetaCAT's report of abundances with the abundances corrected for assembler bias.

3.4 Software operation

3.4.1 First-time run on a metagenome



Number of cpus to use (set to max) 4

BLAST setup

☑ check off to run MetaCAT on a previous blast result

RefSeq protein file (FASTA) D:\MetaCAT\RefSeq_database\viral_release37_all_protein.faa

Metagenome protein file (FASTA) D:\MetaCAT\data\demo_contigs.txt

E value threshold 0.001

BLAST outfile demo_BLAST_results.txt

MetaCAT setup

RefSeq protein file (GenPept/FASTA) D:\MetaCAT\\RefSeq_database\viral37_all_protein.gpff

E value threshold 1e-007

Outfile prefix (optional) my_output

Figure 3.6. Meta main interface.

Computing resources parameters:

If the pull-down menu is activated this means you have the Matlab Parallel Processing Toolbox installed. The default number is set to the number of processors on your computer. It is recommended you utilize all cpus for faster execution, however you can use the pull-down menu to restrict the number of cpus used.

BLAST input parameters:

1. Enter the protein FASTA file for your known reference database. Any FASTA file can be used, however we recommend using NCBI's RefSeq database. This database is a comprehensive, non-redundant database curated by NCBI. Each RefSeq FASTA file issued by NCBI is accompanied by a corresponding GenPept file that includes comprehensive information about the gene and its origin. This GenPept file can be parsed by MetaCAT (see description of output below). For demonstration purposes the "`RefSeq_database`" folder includes the file: "`viral_release37_all_protein.faa`" which is release 37 (Sep. 2009) of all RefSeq viral genes, spanning 2386 distinct species.
2. Enter the FASTA protein file for the metagenome of interest. For demonstration purposes the file "`demo_contigs.txt`" is provided in the "`data`" folder.
3. Enter the E value threshold for BLAST (the default is 0.001).
4. Enter a name for the BLAST output file that will be generated in the "`output`" folder.

MetaCAT input parameters:

5. Enter the GenPept file corresponding to FASTA file entered in option #2. If this file is not available use the FASTA file entered in option #2. For demonstration purposes the “RefSeq_database” folder contains the GenPept file “viral_release37_all_protein.gpff” that corresponds to the RefSeq FASTA file “viral_release37_all_protein.faa”.
6. Enter the E value threshold for the MetaCAT algorithm (default is 1e-7).
7. Enter a string that will be included in all MetaCAT output files for your reference.
8. Click “Run BLAST & MetaCAT” to run MetaCAT.

3.4.2 Output files generated

MetaCAT generated files

At the end of MetaCAT’s run six output files will be generated to the ‘output’ folder:

- The BLAST output file
- <blast file name><MetaCAT output string>_MetaCAT_output0_params.txt
- <blast file name><MetaCAT output string>_MetaCAT_output1_AllGenes.txt
- <blast file name><MetaCAT output string>_MetaCAT_output2_AllGenesFilt.txt
- <blast file name><MetaCAT output string>_MetaCAT_output3_RelatedGenes.txt
- <blast file name><MetaCAT output string>_MetaCAT_output4_ShortTable.txt

These output files can be opened in EXCEL. The “Filter” option in EXCEL can be used to filter these output files. The last file, ‘…Output4_ShortTable’, is the final output of MetaCAT. See the Section 3.6 for a description of output files 0, 1, 2, 3 and 4.

The final output of MetaCAT: the *output4_ShortTable file

This file lists all the RefSeq genes whose lowest E value was equal to or lower than the MetaCAT E value threshold, after removing *related* RefSeq genes. The list is sorted according to the number of metagenome gene objects (MGOs) that are homologous to the given RefSeq gene such that those with the largest number appear at the top of the list. Each line includes additional information about the given RefSeq gene and its alignment with the MGO that yielded the lowest E value. A detailed description of the fields included in this file is given in Section 3.6.4.

3.4.3 Subsequent runs of MetaCAT

The program can run in two modes — a BLAST analysis followed by MetaCAT analysis (the default mode) and just a MetaCAT analysis. The latter mode is useful for analyzing a previous BLAST run, for example with a different E value threshold. To toggle between the modes click the check box (option 10 in Fig. 3.6). This switch will inactivate all fields related to the BLAST run, and will allow the user to select a previous BLAST output file generated by MetaCAT.

3.5 Installation instructions

3.5.1 System requirements

1. Operating system: Windows (32bit/64bit), Linux (32 bit/64 bit), Mac OS (32bit/64bit)
2. Matlab 7.4 and higher
3. To enable parallel processing Matlab Parallel Processing Toolbox v4.2 is required.

For best performance:

The program is computationally cpu intensive and is capable of utilizing multiple cpus using Matlab's Parallel Processing Toolbox. For optimal performance we recommend computers with multiple fast processors. On a Dell Precision T3500 with Quad Core Xeon X5550 (eight 2.66 GHz processors) analysis of a metagenome of 80,000 gene objects with a reference viral library of 80,000 genes takes 3 hours.

3.5.2 Installation

Note: On Linux/Mac systems you may need to have root privileges. It is recommended that you install this software as a root/administrator for these operating systems.

Installation requires an internet connection.

1. Download the compressed sources for MetaCAT and extract locally .
2. Start Matlab.
3. Change directory to the `'bin'` folder of MetaCAT.
4. Run MetaCAT by typing `MetaCAT_EXE` in the Matlab command prompt.
5. Click "Automatic installation (recommended)".
6. The program automatically downloads and installs BLAST vs. 2.2.22+ from NCBI[1]. This process may take a few minutes. Once the installation of BLAST starts click "next" and accept all of the default entries.
7. Once the main interface loads you may start using MetaCAT.

[1] If you already have blast 2.2.22+ installed in the default MetaCAT directory this step is automatically skipped. Other blast versions are ignored by MetaCAT. If you have blast 2.2.22+ installed in a different directory MetaCAT gives you the option to select the folder in which you previously installed this program. Simply select the "Please proceed to manually install blast 2.2.22+" option in the first menu.

3.5.3 Troubleshooting

1. For Linux/Mac OS it is recommend you have root privileges.
2. If there is a problem downloading BLAST make sure your firewall does not block the ftp port. Alternatively you can download blast v2.2.22+ manually from the NCBI website (choosing the appropriate software version appropriate for your OS): ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.22/.

   Installation instructions for blast can be found here ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.22/user_manual.pdf.

   Once blast is installed, start MetaCAT, click "Locate sources on computer" and locate the "bin" folder of blast installation.
3. If there is a problem running blast there may be a previous installation of blast from another utility which is interfering with MetaCAT. If you are running Windows and Windows is not installed on the c:\ drive, check in the windows folder on your computer for a file called 'ncbi.ini'. If this file exists, temporarily change its name.

3.5.4. Downloading and combining RefSeq files

The RefSeq database is continuously updated. The most up-to-date release of RefSeq files is available on the NCBI ftp server at ftp://ftp.ncbi.nih.gov/refseq/release/. You can select to download the entire database or subsets of this database for particular taxonomic or other logical groups. FASTA files have an *.faa extension and GenPept files have a *.gpff extension. Since the RefSeq databases are large, these databases have been separated into numerous smaller files to facilitate downloading from the web and handling. These files can be downloaded in-bulk using one of the many free ftp software programs available on the web. Simply log on to the ftp site ftp://ftp.ncbi.nih.gov as an anonymous user and browse to the `refseq/release/` directory.

To combine all of the RefSeq files into one large FASTA/GenPept file one option is to use a MetaCAT utility:

1. Make sure MetaCAT's "`Combine_RefSeq_files`" folder *contains only one file:* `util_concat_all_files_in_folder.m`.
2. Copy all the files you wish to combine into one file (and only these files) into the above folder.
3. In Matlab change directories to the "`Combine_RefSeq_files`" folder.
4. Type `util_concat_all_files_in_folder` at the prompt.

This utility will combine all of the files in the given directory expect for the Matlab source into one file named "`combined_all`". For large files this may take some time. After the run is finished change the name of this file to whatever name you choose.

3.5.5 MetaCAT folders

The following folders are installed with MetaCAT:

`bin`	location of MetaCAT Matlab execution source MetaCAT_EXE.m
`msrc`	folder with MetaCAT source code
`blast`	folder where blast is installed
`data`	folder to store metagenome files
`RefSeq_database`	folder to store the RefSeq FASTA/GenPept files
`output`	folder to which all output files are written
`Combine_RefSeq_files`	folder that contains a Matlab source code for concatenating all the files in this folder (see §6)

3.5.6 Known bugs

NCBI blast 2.2.22+ appears to have an intrinsic bug where BLASTing a single FASTA record against a large database in some rare cases can add/miss some hits compared to the case where this record is embedded in a very large FASTA file. This can lead to very small differences in the BLAST output depending on the number of cpus used, as the parallelization requires splitting the RefSeq FASTA file into *n* smaller files, *n* being the number of cpus used as defined by the user. This bug appears to be very rare and doesn't affect results significantly, for example slightly affecting the number of hits (+/- 1 or 2) for 10 out of ~6500 records that passed the BLAST E value threshold.

3.6 Description of additional output files

Description of additional output files:

Output0_params

This file contains:

- Statistics on the run such as execution time, date and time of run, version of MetaCAT used, etc.
- The command line(s) used to run BLAST (if appropriate)
- A summary of the parameters/file names used to run MetaCAT
- The list of output files generated by this run.

Output1_AllGenes

Information parsed from the BLAST output file. The file lists all of the RefSeq genes that passed the BLAST E value threshold. Each RefSeq gene is followed by the list of all metagenome gene objects (MGOs) that passed the BLAST E value threshold. The following additional information is provided for the MGO yielding the lowest E value:[2]

1. Index

Counter of RefSeq gene in table.

2. RefSeq gene

RefSeq gene identification as it appears in the RefSeq FASTA file definition line (extracted by BLAST).

3. RefSeq gene definition

RefSeq gene definition as it appears in the RefSeq FASTA file definition line (N/A for the AllGenes output file).

4. Metagenome gene object ID with lowest E value

Identification of MGO that yielded the lowest E value for the given RefSeq gene.

5. # of metagenome gene objects similar to this RefSeq gene

Number of MGOs that yielded an E value equal or lower than the BLAST E value threshold when aligned against the given RefSeq gene.

6. % identity

Percent identity of the given RefSeq gene and the MGO.

7. # of identical amino acids

Number of identical amino acids between the given RefSeq gene and the MGO.

8. E value

E value for the alignment between the given RefSeq gene and the MGO.

[2] To display the information for just the best MGOs use EXCEL and filter the first column by the string “table”.

9. Alignment length (amino acids)

Number of amino acids in the alignment.

10. RefSeq gene length (amino acids)

Number of amino acids in the RefSeq gene (N/A for the AllGenes output file).

11. % of RefSeq gene length aligned

The percent of the RefSeq gene length that appears in the alignment — i.e., the ratio of (9) and (10) times 100 (N/A for the AllGenes output file).

12. aa sequence

The amino acids sequence of the RefSeq gene (N/A for the AllGenes output file).

Output2_AllGenesFilt

Same as the 'Output1_AllGenes' file but showing only RefSeq genes whose lowest E value was equal to or lower than the MetaCAT E value threshold.

Output3_RelatedGenes

List of RefSeq genes whose lowest E value was equal to or lower than the MetaCAT E value threshold, after removing related RefSeq genes (i.e., the list of group representatives)[3]. Following each group representative is the list of related genes (i.e., group members). The list is sorted according to the number of MGOs homologous to the given RefSeq gene (highest number at the top of the list). The following additional information is given for every RefSeq gene:

1. Index

Counter of RefSeq gene in table.

[3] To display just this list in EXCEL, filter the first column by the string "table1".

2. RefSeq gene

RefSeq gene ID as it appears in the RefSeq FASTA file definition line (extracted by BLAST).

3. RefSeq gene definition

RefSeq gene “Definition” field as it appears in the GenPept file (or if a RefSeq FASTA file was supplied, the RefSeq gene definition as it appears in the FASTA definition line).

4. Min % of shared metagenome gene objects

The overlap between the MGO list of the given RefSeq gene and the MGO list of the group representative in units of percent (i.e., the number of MGOs shared between both RefSeq genes divided by the larger number of MGOs of both genes, in units of percent).

5. # of metagenome gene objects similar to this RefSeq gene

Number of MGOs that had an E value equal to or lower than the BLAST E value threshold when BLASTed against the given RefSeq gene.

6. % identity

Percent identity between the MGO with the lowest E value and the given RefSeq gene.

7. # of identical amino acids

Number of identical amino acids in the alignment of the MGO with the lowest E value and the given RefSeq gene.

8. E value

E value in the alignment of the MGO with the lowest E value and the given RefSeq gene.

9. Alignment length (amino acids)

Length of alignment in amino acids between the MGO with the lowest E value and the given RefSeq gene.

10. RefSeq gene length (amino acids)

Number of amino acids of the given RefSeq gene.

11. % of RefSeq gene length aligned

The percent of the RefSeq gene length that appears in the alignment — i.e., the ratio of (9) and (10) times 100.

12. GenPept Features

RefSeq gene Features field as it appears in the GenPept file.

Output4_ShortTable

This file is the main output of MetaCAT. This file contains the following fields:

1. Index

Counter of the RefSeq gene in the table.

2. RefSeq gene

RefSeq gene identification as it appears in the RefSeq FASTA file definition line (extracted by BLAST).

3. Metagenome gene object ID with lowest E value

Identification of the MGO that yielded the lowest E value for the given RefSeq gene.

4. # of metagenome gene objects similar to this RefSeq gene

Number of MGOs that had an E value equal to or lower than the BLAST E value threshold when the given RefSeq gene was BLASTed against the metagenome.

5. tot # of metagenome gene objects associated with this RefSeq gene group

Combined number of homologous MGOs of the given RefSeq gene and all its related RefSeq genes (i.e., group members).

6. # of related RefSeq genes

Number of RefSeq genes related to the given RefSeq gene, including the given RefSeq gene, i.e., number of group members including the group representative.

7. % identity

Percent identity between the MGO with the lowest E value and the given RefSeq gene.

8. # of identical amino acids

Number of identical amino acids in the alignment of the MGO with the lowest E value and the given RefSeq gene.

9. E value

E value for the alignment of the MGO with the lowest E value and the given RefSeq gene.

10. Alignment length (amino acids)

Length of alignment in amino acids between the MGO with the lowest E value and the given RefSeq gene.

11. RefSeq gene length (amino acids)

Number of amino acids for the given RefSeq gene.

12. % of RefSeq gene length aligned

The percent of the RefSeq gene length that appears in the alignment — i.e., the ratio of (10) and (11) times 100.

13. aa sequence

Amino acid sequence of the given RefSeq gene.

14. RefSeq gene definition

The “Definition” field for the given RefSeq gene as it appears in the GenPept file (or if a RefSeq FASTA file was supplied, the RefSeq gene definition as it appears in the FASTA definition line).

> **Definition field** — “Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as ‘complete cds’.” [18]

15. GenPept GenBank division

GenBank division field for the given RefSeq gene as it appears in the GenPept file.

> **GenBank division field** — “The GenBank division to which a record belongs is indicated with a three-letter abbreviation. The GenBank database is divided into 18 divisions:
>
> 1. PRI — primate sequences
> 2. ROD — rodent sequences
> 3. MAM — other mammalian sequences
> 4. VRT — other vertebrate sequences
> 5. INV — invertebrate sequences
> 6. PLN — plant, fungal, and algal sequences
> 7. BCT — bacterial sequences
> 8. VRL — viral sequences
> 9. PHG — bacteriophage sequences
> 10. SYN — synthetic sequences
> 11. UNA — unannotated sequences
> 12. EST — EST sequences (expressed sequence tags)
> 13. PAT — patent sequences
> 14. STS — STS sequences (sequence tagged sites)

15. GSS — GSS sequences (genome survey sequences)
16. HTG — HTG sequences (high-throughput genomic sequences)
17. HTC — unfinished high-throughput cDNA sequencing
18. ENV — environmental sampling sequences”

16. GenPept molecule type

Molecule type field for the given RefSeq gene as it appears in the GenPept file.

> **Molecule type field** — “The type of molecule that was sequenced. Each GenBank record must contain contiguous sequence data from a single molecule type. The various molecule types are described in the Sequin documentation and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA.” [18]

17. GenPept source

Source field for the given RefSeq gene as it appears in the GenPept file.

> **Source field** — “Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type.” [18]

18. GenPept classification

Organism field for the given RefSeq gene as it appears in the GenPept file.

> **Organism filed** — “The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database. If the complete lineage of an organism is very long, an abbreviated lineage will be shown in the GenBank record and the complete lineage will be available in the Taxonomy Database.” [18]

19. GenPept comments

Comments field for the given RefSeq gene as it appears in the GenPept file.

> **Comments field** — “A COMMENT identifying the RefSeq Status is provided for the majority of the RefSeq records. This comment may include information

about the RefSeq status, collaborating groups, and the GenBank records(s) from which the RefSeq is derived. The RefSeq COMMENT is not provided comprehensively in this release… Additional COMMENTS are provided for some records to provide information about the sequence function, notes about the aspects of curation, or comments describing transcript variants." [19]

20. GenPept Features

Features field for the given RefSeq gene as it appears in the GenPept file.

Features field — "Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.

Source: Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter.

Taxon: A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database.

CDS: "Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons)." [18]

Protein Names: Protein names may be provided by a collaborating group, may be based on the Gene Name, or for some records, the curation process may identify the preferred protein name based on that associated with a specific EC number or based on the literature.

Protein Products: Signal peptide and mature peptide annotation is provided by propagation from the GenBank submission that the RefSeq is based on, when provided by a collaborating group, or when determined by the curation process.

Domains: "Domains are computed by alignment to the NCBI Conserved Domain Database database for human, mouse, rat, zebrafish, nematode, and cow. The best hits are annotated on the RefSeq. For some records, additional functionally significant regions of the protein may be annotated by the curation staff. Domain annotation is not provided comprehensively at this time." [19]

3. 7 References

1. Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson T, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450: 560-565.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf K, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66.
4. Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554.
5. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: reviews0003.
6. Handelsman J, Tiedje J, Alvarez-Cohen L, Ashburner M, Cann I, et al. (2007) The New Science of metagenomics: revealing the secrets of our microbial planet. National Academy of Sciences, Washington, D.C.
7. Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annual Review of Genetics 38: 525-552.
8. Singh J, Behal A, Singla N, Joshi A, Birbian N, et al. (2009) Metagenomics: Concept, methodology, ecological inference and recent advances. Biotechnology journal 4: 480-494.
9. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiology and Molecular Biology Reviews 72: 557.
10. Edwards R, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504-510.
11. Dinsdale E, Edwards R, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.

12. Kristensen D, Mushegian A, Dolja V, Koonin E (2009) New dimensions of the virus world discovered through metagenomics. Trends in Microbiology 18: 11-19.
13. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Research 17: 377.
14. Sun S, Chen J, Li W, Altintas I, Lin A, et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Research 39: D546.
15. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics 9: 386.
16. Pruitt K, Tatusova T, Maglott D (2005) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33: D501-D504.
17. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389.
18. http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
19. RefSeq release notes ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/

Chapter 4

The Biophysics of Prokaryotic and Viral Diversity in Aqueous Environments

4.1 Abstract

Recent advances in techniques for enumerating viruses have led to a plethora of measurements of viral and bacterial abundances in nature that beckon for both qualitative and quantitative explanation. Here we propose a biophysical model that describes the interaction between bacteria and their lytic viruses in aqueous environments that combines both predator-prey relations and a diffusion-based transport model of viruses. In addition we postulate that the burst size is proportional to the volume ratio of the host cell and its infecting virion, for which there is empirical support for cell radii < ~1μm. We find that the concentration of a given bacterial species approximately scales with the radius of the cell *r*, as r^{-4}, suggesting that, within the context of a predator-prey model, the size of a bacterium is the most critical parameter determining its fixed point concentration. To extend our model to the community level, we postulated that there is no selection pressure on bacterial radii, i.e., *a priori*, all bacterial radii are equally probable. Given this hypothesis we predict that the size spectrum of marine bacteria follows a power law with slope -1, close to the observed average spectrum. We proceed to derive expressions for the total concentration of bacteria and viruses in the environment, reproducing for typical marine systems a virus-to-bacterium ratio (VBR) of ~10. We show that the VBR is primarily determined by the average net growth rate of bacteria (growth minus predation), the average viral decay rate and, interestingly, the radius of the minimum viable bacterium. We next

derive a simple expression for the number of species in a given environment per unit volume, and predict that for offshore waters, where there are ~10^5 bacteria per ml, there should be ~10^2 to ~10^3 prokaryotic species in at most ~10^2 to ~10^4 liters of water, consistent with current empirical estimates of species richness. Thus, any given marine environment can only pack a finite degree of diversity. We use this observation to calculate an absolute lower and upper bound on the total number of active bacterial species in the ocean water column (excluding sediment), by considering the case of completely homogenous oceans and maximally heterogeneous oceans. We find that the number of species in the ocean water column should lie in the range of 10^4–10^{21}. We conclude by considering further experiments to test the validity of the proposed model.

4.2 Introduction

It was only in the late 1980s that the first quantitative estimates of viral abundance in the oceans using transmission electron microscopes revealed the existence of as many as millions of viral particles per milliliter of seawater [1]. Subsequently, more reliable counting methods based on epifluorescence imaging of stained nucleic acids came to the fore [1,2,3]. These methods were simple to execute even in field conditions and led to an explosion of measurements of viral and prokaryote concentrations in many environments [2,3]. In marine and fresh water ecosystems, these types of studies revealed that as a rule of thumb, viral concentrations typically exceed bacterial concentrations by one order-of-magnitude [1,2,4].

We were interested in understanding the basic processes in play that determine phage-host interactions in aqueous environments, and how these processes affect the bacterial and viral community composition. We therefore sought to identify the key variables that determine the virus and bacterium concentrations in the environment and to formulate a simple toy model that is capable of making reasonable order of magnitude predictions and that can qualitatively explain the observed trends. Predator-prey models for host-virus interaction have earlier been examined by Campbell [5], Levin et al. [6], Lenski [7], Beretta et al. [8] and Thingstad et al. [9,10]. However, in these models the biophysical process of virus transport, which governs the contact rates between viruses and bacteria, was not considered. Stent [11] and Murray et al. [12] considered transport processes of viruses in aqueous environments but not in the context of a predator-prey model in an ecological setting. Our starting point is a simple toy model that incorporates virus transport within the context of a predator-prey model. We begin by examining the case of a particular isolated phage-host system with the goal of identifying the key variables that govern this system. We then extend our model to the community scale by hypothesizing the simplest evolutionary scenario that there is no selection pressure on bacterial radii, i.e., *a priori*, all bacterial radii are equally probable. We derive basic relations for the total concentration of bacteria and their viruses in the environment, and a basic relation for the total prokaryotic mass in the environment. Based on these results we explore questions such as, what are the critical parameters governing the system and how do variables scale with respect to these parameters? What determines the virus-to-bacterium ratio? What determines the number of species in a given environment? In what volume of water should we find this diversity? What are the bounds on the total diversity of species in Earth's oceans? Where possible we compare our predictions to observations and conclude with suggestions for experiments to further test our model.

We will further claim that the precise definition of a species lies outside the scope of our biophysical model. Consequently in Chapter 5 we will consider an evolutionary model for the generation of bacterial and viral species consistent with the definition of a species used in this chapter. The evolutionary model is a first step in connecting the predictions of the biophysical model described in this chapter with empirical observations of diversity in the environment.

4.3. General assumptions

4.3.1 Decoupling phage-host systems

Typically a given environment will contain many species of bacteria and viruses. However, under certain assumptions, the microbial and viral communities can be treated as a set of decoupled phage-host systems [9]. Such an approximation will be valid if the following conditions are satisfied: (1) different bacterial species function independently of each other. Thus symbiotic relationships are prohibited. (2) Each viral species infects a single bacterial species and (3) each bacterial species is infected by a single viral species. The second assumption is a decent approximation given that phages characteristically exhibit species or subspecies [2,13,14] (although some exceptions, such as certain broad host range cyanophages, exist [2]). The third assumption may seem odd given that the most familiar example, *E. coli*, is known to be infected by many lytic viruses (e.g., the T-series). *E. coli*, however, is a commensal organism that lives in the intestines of animals and humans. Since the guts of animals/humans are physically separated, in principle at least, different species of phages that infect *E. coli* could have evolved in different guts. Aqueous environments on the other hand are diffusible and generally topologically connected and therefore of a different nature. Thus host range observations regarding *E. coli,* or any other gut bacterium, may not apply to marine ecosystems. That said, biogeography may play a role in marine ecosystems when considering very distant regions (e.g., the same cyanobacterium species in remote regions may be infected by different phages). Therefore both assumptions (2) and (3) may perhaps be relaxed by requiring them to be satisfied locally.

Assumption (3) is consistent with assumption (2) in the sense that two viruses cannot control the same bacterial species indefinitely, since such a system is unstable (Section 4.6). The opposite is also true, two bacterial species cannot be controlled by the same virus indefinitely, thus assumption (2) is consistent with assumption (3).

As a result, we begin our discussion by considering a simple phage-host system consisting of a single phage species infecting a single bacterial species, henceforth denoted by the index i. In Section 4.4.2 we will consider multiple independent phage-host systems.

4.3.2 Host mortality

Causes of mortality

It is generally accepted that bacterial host mortality is primarily due to either protist grazing or viral predation [4,15,16,17], both appearing to contribute about equally to microbial mortality [15,18,19]. In surface waters for example, viruses are thought to be responsible for ~10–50% of the total bacterial mortality, whereas in environments in which protists do not thrive, such as low-oxygen lake waters, viruses are thought to be responsible for 50–100% of bacterial mortality [4]. Thus it appears that the two likely fates of a bacterial cell in the ocean are either to be eaten by a protist or be lysed a virus.

Lysogenic versus lytic viruses

The process of viral predation can be mediated either through infection by lytic viruses or through induction of temperate viruses. In the case of temperate viruses, the infecting virus either enters a lytic phase and kills its host or is integrated into the genome of the host and may be induced at a later stage in response to an induction event (e.g., exposure to a mutagenic agent

[2]). In the oceans however it appears that lysogenic induction is rare [2,4,20], occurring either sporadically or at a low level [4]. Though this matter has still not been completely settled [2], it has been suggested that the majority of viruses observed in sea water are the result of successive lytic infections [4]. Other forms of infection such as chronic infection and pseudolysogeny [2] do not lead to host death and are therefore not considered to contribute to viral predation in this context. We shall therefore assume in our toy model that viral predation is exclusively the result of infection by lytic viruses.

Protists versus viruses

When comparing the effect of protist grazing to virus lysis on bacteria, there is a fundamental difference between these two predators that has to do with their host range. As a first-order approximation [10], protists can be regarded as omnivorous, i.e., they are not host selective [10,17]. On the other hand, viruses are known be highly selective, displaying species or subspecies (strain) specificity [2,13,14]. Therefore, protists would control the total concentration of bacteria while viruses would control the individual concentration of bacterial species [10,17]. In a resource rich environment, there is evidence to suggest that because protists themselves are preyed upon, bacterial growth is determined by competition for resources and not by protist grazing [17]. Regardless of the mechanism that controls the total concentration of bacteria, in our model we simply assume that the total concentration of bacteria is fixed by some process and refer to this limiting factor as the “carrying capacity” of the environment.

Thus we will assume that every bacterial “species” is under viral control. Since grazing is thought to be complex non-passive hydrodynamical process owing to the currents induced by the motion of the flagella drawing its prey in [17], we account for this process by means of an effective grazing rate denoted by $\gamma^{(i)}_{non-viral}$ (see Table 4.1 for a list of notation). Another potential source for bacterial mortality is autolysis or programmed cell death in response to, for example, radiation damage [17]. Here all non-viral mediated mortality can be included effectively in $\gamma^{(i)}_{non-viral}$.

What is a bacterial species?

Note that we have not precisely defined what a bacterial "species" is or what a viral "species" is. What is the definition of a bacterial species? Similarly, what is the definition of the viral "species" that infects this bacterial "species"? We will argue that the precise definition of these concepts lies outside the scope of a biophysical model of phage-host interaction and requires a "higher" theory that probes the genetic complexity of these species (i.e., an evolutionary theory). In Chapter 5 we will propose an evolutionary mechanism that can be used to define a bacterial "species" and a viral "species", and by which new bacterial and viral "species" co-emerge through a process of co-speciation. We will also show that when this evolutionary model is viewed in a genetic coarse-grained way, the evolutionary model converges to our current biophysical model. Our conclusion will be that while a bacterial species interacts with just one viral species, and vice versa, each of these species is comprised of strains (which are emerging new species) that are part of interaction networks with more than one viral strain. The key result that we derive is that although the species are independent, we need to multiply their concentration by (roughly) the number of strains per species to get the total concentration of a species (with a strain defined as an entity distinguishable in a consistent and clear way from all other strains). Thus, while a "strain" would have been our intuitive definition *a priori* for a "species", we find that strains are not independent elements (they are part of networks), and one needs to consider a more complex structure called a "species" to achieve independent phage-host systems.

4.3.3 Virus decay

Viral decay is thought to be mainly due to environmental damage from sunlight, temperature effects, and interaction with certain substances such as heat-labile colloidal dissolved organic matter [2,4,21]. These events lead to a certain rate of viral decay which we denote by $\gamma^{(i)}_{virus\ decay}$.

Though protists can also potentially lead to viral removal by ingestion [21], grazing is generally not considered to be a significant factor leading to loss of viruses [2].

4.3.4 The physiological state of the host

The physiological state of bacteria in nature is generally unknown and is the subject of current research [2]. Generally speaking, bacteria appear to be growing slowly in marine environments. For example, in the cold waters of the Barents Sea in the Arctic ocean, growth rates were estimated to lie between 0.05 and 0.25 day^{-1} [22] whereas in the warmer coastal seawater near Santa Monica growth rates were measured to be higher, ~1–3 day^{-1} [18]. In our simple toy model we will assume that the environment is ideal in the sense that the bacteria are in a state of exponential growth. Though many environments are most likely not ideal, the notion of an "ideal environment" can be a useful construct that can at the very least serve as a null hypothesis for a given environment. In the context of our model, the growth rate of the i^{th} bacterial species is denoted by $\alpha^{(i)}$ = (doubling rate)·ln2. This growth rate is thus species specific and is determined by the availability of nutrients required by the given species.

4.3.5 Bacterial and viral abundance distribution

The virus-to-bacterium ratio (VBR) in marine systems is typically measured to be in the range of 5–25 [1,2,4,19] and in the deep waters of the Atlantic Ocean this ratio often exceeds 100 [1]. Particular phage-host systems have also been shown to exhibit VBRs as high as 8 (and locally even as high as 30 — see example discussed later on) [23]. We shall therefore assume in our model that the VBR for the i^{th} bacterial species satisfies $VBR^{(i)} >> 1$. We will also assume that local spatial inhomogeneities in free virion concentration due to, for example, burst events [24], diffuse over time without inducing lysis in neighboring hosts. Thus, synchronized lysing (a possible mechanism for bloom termination [1,13,25]) is not accounted for by our model. Since blooms appear to be the exception rather than the rule [13], we do not expect this to affect the applicability our model to most ecological settings. The spatial nonuniformity of viruses will be further discussed below.

4.4 A biophysical model of phage-host interaction

4.4.1 Model development part I: A single phage-host system

4.4.1.1 Viral diffusion and infection rate

We begin by estimating the infection rate of a certain bacterial species given that its viruses are freely diffusing in the medium. Let $N_{bacteria}^{(i)}$ and $N_{virus}^{(i)}$ be the number of bacteria and viruses respectively associated with the i^{th} bacterial species in a given volume *V*. We wish to estimate the absorption rate of the viruses to their hosts, denoted by $I_{virus}^{(i)}$ (in units of s^{-1}), given that $N_{virus}^{(i)} >> N_{bacteria}^{(i)}$. We will assume the bacterium is stationary and is described by a simple spherical geometry with an effective radius $R_{bact}^{(i)}$. The approximation that the bacterium is stationary is supported by the following facts: Based on the Stokes-Einstein relation (see below), the diffusion constant of a typical *E. coli*-like bacterium is expected to be roughly 30 times smaller than the diffusion constant of a typical phage particle in the same environment, thus bacteria are diffusing very slowly in comparison to their viruses. Even if a bacterium is engaged in swimming, its contact rate with viruses is relatively unaffected by the swimming motion of the bacterium [12]. Bacteria attached to marine snow may encounter enhanced viral contact rates due to the fast motion of the sinking particles [12], however these are thought to constitute a small fraction of the overall population of bacteria and should therefore not contribute much to the overall number of bacterium-viral contacts [12].

Table 4.1. Variables and parameters used in the discrete phage-host interaction model

Variables	Definition	Units
$c_{bacteria}^{(i)}$	Concentration of bacteria belonging to the *i*th bacterial species	(number)/*m*^{3}
$N_{bacteria}^{(i)}$	Number of bacteria belonging to the *i*th bacterial species in volume *V*	dimensionless
$c_{virus}^{(i)}$	Concentration of viruses infecting the *i*th bacterial species	(number)/*m*^{3}
$N_{virus}^{(i)}$	Number of viruses infecting the *i*th bacterial species in volume *V*	dimensionless
$I_{virus}^{(i)}$	Absorption rate of viruses onto the *i*th bacterium	s^{-1}
$VBR^{(i)}$	Virus-to-bacterium ratio of the *i*th phage-host system = $c_{virus}^{(i)} \big/ c_{bacteria}^{(i)}$	dimensionless
Parameters		
i	Index of the *i*th bacterial species. Parameters that depend on *i* can be interpreted as random variables drawn from a certain distribution	Dimensionless
$\alpha^{(i)}$	Specific growth rate= $\mu^{(i)}$ln 2, where $\mu^{(i)}$ is the doubling rate	s^{-1}
$\gamma_{non-viral}^{(i)}$	Bacterial mortality rate due to non-viral mediated processes such as grazing	s^{-1}
$\gamma_{viral\ decay}^{(i)}$	Viral decay rate	s^{-1}
$R_{virus}^{(i)}$	Effective radius of the virus	*m*
$R_{bact}^{(i)}$	Effective radius of the bacterium	*m*
$b^{(i)}$	Burst size	Dimensionless
$\beta^{(i)}$	Volume fraction of host cell occupied by virions	Dimensionless
$D_{virus}^{(i)}$	Diffusion constant of the virus	m^2/s
η	Viscosity of the environment	$kg \cdot m^{-1} s^{-1}$
τ	Latency period	*s*
η	Viscosity of the environment	$kg \cdot m^{-1} s^{-1}$
k_B	Boltzmann constant	$kg \cdot m^2\ s^{-2} K^{-1}$
T	Temperature of the environment	*K*

To estimate the infection rate we assume that viruses anchor to the cell surface, and that consequently the bacterium can be regarded as a perfect absorber. We then solve the diffusion equation for the virions at steady-state. We assume the bacterium is placed at the origin and that the boundary conditions are given by $c_{virus}^{(i)}(r = R_{bact}^{(i)}) = 0$ and $c_{virus}^{(i)}(r = \infty) = c_{virus}^{(i)}(\infty)$, where $c_{virus}^{(i)}(\infty)$ $(= N_{virus}^{(i)} \big/ V)$ is the far-field concentration of the i^{th} viral species. Solving the diffusion equation at steady-state and calculating the transport flux across the boundary of the sphere gives us the steady-state absorption rate of viruses onto the bacterium ($I_{virus}^{(i)}$) [11,26]

$$I_{virus}^{(i)} = 4\pi D_{virus}^{(i)} R_{bact}^{(i)} c_{virus}^{(i)}\left(\infty\right). \tag{1}$$

where $D_{virus}^{(i)}$ is the diffusion constant of the i^{th} viral species. Thus the average time until the i^{th} bacterial species is infected is $1/I_{virus}^{(i)}$. The assumption of a perfect absorber means that once viruses make contact with the cell, they are "absorbed" (i.e., infect the cell). Berg and Purcell [26] showed that the net flux to a cell with a small number of receptors is almost as large as the net flux into a perfectly absorbing cell. For example, fewer than 500 phage receptors are necessary for λ phage to attain half the maximum absorption rate [26,27] where *E. coli* typically has between 30 to 10,000 receptors per cell depending on the growth medium [27]. Therefore the assumption of a perfect absorber requires a small correction factor that we shall ignore in our simple toy model.

Because the perfect absorber leads to a steady-state gradient in viral concentration, the distribution of viruses is spatially nonuniform. For the case of a single absorber at the origin, the steady-state concentration of viruses is given by $c_{virus}^{(i)}(r) = c_{virus}^{(i)}(\infty)\left(1 - R_{bact}^{(i)}/r\right)$ [26], where $c_{virus}^{(i)}(\infty)$ is the far field concentration of the viruses infecting the i^{th} bacterial species. Thus, if we assume that the mean spacing between cells of a given bacterial species is significantly larger than $R_{bact}^{(i)}$ (i.e., $\left(c_{bact}^{(i)}\right)^{-\frac{1}{3}} >> R_{bact}^{(i)}$), then any given bacterial host of this species will lie in the far-field range of adjacent hosts of the same species. Thus under these conditions, to a first-order approximation, each bacterium can effectively be thought of as an isolated perfect absorber. These conditions are typically satisfied for marine ecosystems. For example, for typical marine ecosystems $c_{bact} \le 10^6$ ml^{-1}, or $\left(c_{bact}\right)^{-\frac{1}{3}} \le 100\,\mu$m. In the extreme (and unlikely) scenario where the entire bacterial population consists of a single species, then as long as the radius of this species is <~10 μm this condition is satisfied. Since we will see that larger bacteria are rarer (Section 4.4.2), the error incurred for larger radii will be weighted down when integrating over all radii.

4.4.1.2 Predator-prey relations

We next wish to calculate the total rate of virus infection in the population. The fraction of bacteria $\Delta N_{\text{infected}}^{(i)} / N_{bacteria}^{(i)}$ that are infected during the time Δt, where Δt satisfies $\Delta t << 1/I_{virus}^{(i)}$ is given by $\Delta t / (1 / I_{virus}^{(i)})$. Therefore the fraction of infected cells during Δt is given by $\Delta N_{\text{infected}}^{(i)} / N_{bacteria}^{(i)} = \Delta t / (1 / I_{virus}^{(i)})$, or

$$\frac{dN_{\text{infected}}^{(i)}}{dt} = I_{virus}^{(i)} N_{bacteria}^{(i)}.$$

In principle, not every virion absorption event will lead to successful infection and host lysis. However, at least in the case of T4 infecting *E. coli* this fraction appears to be close to one [11]. We will therefore assume in our toy model that each absorption event leads to host lysis. Building upon this result, we take into account bacterial growth and bacterial death due to non-viral mediated processes and obtain the following bacterial rate equation

$$\frac{dN_{bacteria}^{(i)}(t)}{dt} = \alpha^{(i)} \; N_{bacteria}^{(i)}(t) - \gamma_{non-viral}^{(i)} N_{bacteria}^{(i)}(t) - I_{virus}^{(i)}(t) N_{bacteria}^{(i)}(t).$$

where the first term is due to bacterial growth, the second term is due to non-viral mediated cell mortality, and the third term is due to viral predation leading to host mortality. Dividing by the system volume *V* and inserting Eq. 1 we obtain the following rate equation for the bacterium

(2) $$\frac{dc_{bacteria}^{(i)}(t)}{dt}=\left(\alpha^{(i)}-\gamma_{non-viral}^{(i)}\right)c_{bacteria}^{(i)}(t)-4\pi D_{virus}^{(i)}R_{bact}^{(i)}c_{virus}^{(i)}(t)c_{bacteria}^{(i)}(t).$$

The corresponding rate equation for the i^{th} viral species is given by

$$\frac{dN_{virus}^{(i)}(t)}{dt}=b^{(i)}\cdot I_{virus}^{(i)}(t-\tau)N_{bacteria}^{(i)}(t-\tau)-\gamma_{virus\ decay}^{(i)}N_{virus}^{(i)}(t)-I_{virus}^{(i)}(t)N_{bacteria}^{(i)}(t).$$

where the first term is due to viral production (with τ being the latency period and $b^{(i)}$ being the average burst size of the i^{th} viral species, i.e., the number of virions released per cell into the extracellular environment), the second term is due to virion decay, and the third term is due to viral loss upon absorption (which is negligible since typically $b^{(i)}$>>1) [5]. Dividing by the system volume *V* and inserting Eq. 1 we obtain the following rate equation for the viruses

(3) $$\begin{aligned}\frac{dc_{virus}^{(i)}(t)}{dt}&=b^{(i)}\cdot 4\pi D_{virus}^{(i)}R_{bact}^{(i)}c_{virus}^{(i)}(t-\tau)c_{bacteria}^{(i)}(t-\tau)-\gamma_{virus\ decay}^{(i)}c_{virus}^{(i)}(t)+\\&-4\pi D_{virus}^{(i)}R_{bact}^{(i)}c_{virus}^{(i)}(t)c_{bacteria}^{(i)}(t).\end{aligned}$$

Equations 2 and 3 together form a predator-prey dynamical system. In the simple case where τ=0, Eq. 2 and Eq. 3 form an ideal Lotka-Volterra model. This system exhibits small oscillations with a period of $\left[\left(\alpha \ -\gamma_{non-viral}\right)\gamma_{virus\ decay}\right]^{-\frac{1}{2}}$ ~hours to days around the non-trivial fixed point determined below (see Table 4.3 for typical parameters). Since the steady-state of the viral diffusion equation is achieved on the order of $t >> R_{bact}^{2}/D_{virus} << \sim 1\text{sec}$, Eqs. 2 and 3 can be interpreted as describing the slow dynamics of the far-field viral concentration with the viral diffusion equation at pseudo steady-state.

We are interested in the non-trivial fixed point solutions for this system obtained by setting $d/dt = 0$. Since the steady-state solutions are time invariant we have $c_{virus}^{(i)}(t) = c_{virus}^{(i)}(t-\tau) = \text{const}$ and $c_{bacteria}^{(i)}(t) = c_{bacteria}^{(i)}(t-\tau) = \text{const}$. Solving for these two constants we find that the non-trivial fixed point solutions for this system are given by

$$
\begin{aligned}
c_{virus}^{(i)} &= \frac{\alpha^{(i)} \ -\gamma_{non-viral}^{(i)}}{4\pi D_{virus}^{(i)} R_{bact}^{(i)}}. \\
c_{bacteria}^{(i)} &= \frac{\gamma_{viral\ decay}^{(i)}}{\left(b^{(i)}-1\right)\cdot 4\pi D_{virus}^{(i)} R_{bact}^{(i)}}.
\end{aligned}
\tag{4}
$$

where we implicitly assume that $\alpha^{(i)} > \gamma_{non-viral}^{(i)}$. Note that equating the rate equation for bacteria to zero leads to the following condition:

$$\alpha^{(i)} \equiv I_{virus}^{(i)} + \gamma_{non-viral}^{(i)}.$$

i.e., total bacterial production equals total bacterial mortality. Though solutions to predator-prey models are typically time dependent, here we are mainly concerned with understanding the scaling of the fixed point solutions, which we take as a proxy for the time averaged response of the system.

4.4.1.3 The virus diffusion constant

Since the shape of the virus appears to have little effect on the expected viral transport rate to the bacterium [12] we will follow Murray and Jackson and model the viruses as spheres. For a sphere of radius $R_{virus}^{(i)}$ (the effective radius for the virus) the Stokes-Einstein relation for the diffusion constant is given by

$$D_{virus}^{(i)} = k_B T \big/ \left(6\pi\eta R_{virus}^{(i)} \right)$$

where k_B is the Boltzmann coefficient, T the temperature, and η the viscosity of the medium. Substituting this expression into the fixed point solution given in Eq. 4 we find that

(5A) $$c_{virus}^{(i)} = \tfrac{3}{2} \frac{\eta}{k_B T} \left(\frac{R_{virus}^{(i)}}{R_{bact}^{(i)}} \right) \left(\alpha^{(i)} - \gamma_{non-viral}^{(i)} \right).$$

(5B) $$c_{bacteria}^{(i)} = \tfrac{3}{2} \frac{\eta}{k_B T} \left(\frac{R_{virus}^{(i)}}{R_{bact}^{(i)}} \right) \frac{\gamma_{viral\ decay}^{(i)}}{\left(b^{(i)} - 1 \right)}.$$

Eq. 5A makes the prediction that the larger the factor $D_{virus}^{(i)} R_{bact}^{(i)}$ in Eq. 2 (resulting in a larger viral infection rate – Eq. 1), the lower the concentration of viruses needs to be in order for the overall lysis rate (second term in Eq. 2) to match the net bacterial production rate (first term in Eq. 2). This explains the dependence of Eq. 5A on viscosity, temperature, and the virus-to-bacterium radii ratio. Eq. 5A also predicts that the fixed point concentration of viruses does not depend on their decay rate or burst size. This paradoxical behavior is explained by the fact that viruses need to keep the net bacterial growth in check (leading to the dependence on the growth rate; first term in Eq. 2) irrespective of the viral decay rate or burst size.

Similarly, Eq. 5B predicts that the higher the factor $b^{(i)}D^{(i)}_{virus}R^{(i)}_{bact}$ in the viral production rate term (Eq. 3), the lower the fixed point concentration of bacteria needs to be in order to match viral production (first term in Eq. 3) with viral decay (second term in Eq. 3), explaining the dependence on viscosity, temperature, the virus-to-bacterium radii ratio, and the burst size. Eq. 5B also makes the intuitive prediction that the faster viruses decay, the higher the concentration of bacteria will be. This result holds because the faster viruses degrade (second term in Eq. 3) the more viruses are required to be produced (first term in Eq. 3) to sustain this degradation, and therefore more bacteria are required for viral production (since bacteria are the sources of viruses). Here too we find the paradoxical situation where the fixed point concentration of bacteria does not depend on their net growth rate. The reason for this paradoxical behavior is that as long as bacteria grow — no matter how fast — their fixed point concentration need only be high enough so that viral production (which is proportional to the bacterium concentration — first term in Eq. 3) matches viral decay (second term in Eq. 3).

4.4.1.4 The virus-to-bacterium ratio for a given phage-host system

To obtain the virus-to-bacterium ratio for the i^{th} species we divide $c^{(i)}_{virus}$ by $c^{(i)}_{bacteria}$ obtaining the simple relation

$$(6) \qquad VBR^{(i)} = \frac{c^{(i)}_{virus}}{c^{(i)}_{bacteria}} \cong b^{(i)} \frac{\alpha^{(i)} - \gamma^{(i)}_{non-viral}}{\gamma^{(i)}_{virus\ decay}}.$$

where for simplicity we use the approximation $b^{(i)} - 1 \cong b^{(i)}$ since typically $b^{(i)} >> 1$. Below we shall derive the expression for the VBR for the entire community in a given environment, i.e., for all phage-host systems.

4.4.1.5 Correlation between burst size and host/virus dimensions

Since we are interested in average scaling laws, it is worthwhile to consider the relation between burst size and the dimensions of the host and its virus, as these two quantities may be statistically highly correlated. Lytic viruses typically pack the host cytoplasm with virions upon replication, suggesting that perhaps one can make the assumption that the number of virus progeny per cell is correlated with cell volume and inversely correlated with the volume of the infecting virus. Indeed, Weinbauer et al. found that in ~50% of the visibly infected rods and spirillae and in more than 80% of the cocci found in the northern Adriatic Sea, the entire cell was occupied by mature phages (as opposed to displaying a non-uniform distribution) with the difference between cocci and other morphologies possibly explained by a shorter time span between the appearance of the first mature phages and lysis in cocci cells due to their smaller burst size [28]. Weinbauer et al. also note that almost all bacteria observed in the disruption stage were completely filled with phages [29]. That said, in 18% of the infected bacteria the phage was concentrated in two or three defined areas of the host and did not occupy the entire cell [28]. Furthermore, some bacteria may lyse prematurely [2]. Nevertheless, it has been found empirically that burst size is approximately linearly correlated with cell size for cells with a radius of ~0.2μm to ~1 μm, and larger phages have been found to produce less progeny [2,19]. For example, Weinbauer et al.

[29] found a linear correlation between burst size and host cell volume, with the cell size being the only measured parameter that could account for the distribution of burst sizes [29]. In addition, Weinbauer et al. also found an inverse correlation between burst size and capsid size [28].

We will therefore assume in our toy model that to a first-order approximation the burst size is proportional to the volume ratio of the bacterium and its virus, namely, $b^{(i)} = \beta^{(i)} \cdot \left(R_{bact}^{(i)} / R_{virus}^{(i)} \right)^3$, where $\beta^{(i)}$ is a positive proportionally factor ≤ 1. Note that $\beta^{(i)}$ can be interpreted as the volume fraction of the cell occupied by viruses since: $\beta^{(i)} = b^{(i)} \cdot \left(R_{virus}^{(i)} / R_{bact}^{(i)} \right)^3 = b^{(i)} \cdot V_{virus}^{(i)} / V_{bact}^{(i)}$.

Inserting this correlation into Eq. 5B and approximating $b^{(i)} - 1 \cong b^{(i)}$ we obtain

(7)
$$c_{bacteria}^{(i)} = \tfrac{3}{2} \frac{1}{\beta^{(i)}} \frac{\eta}{k_B T} \left(\frac{R_{virus}^{(i)}}{R_{bact}^{(i)}} \right)^4 \gamma_{viral\ decay}^{(i)} .$$

Further implications of the model and Eq. 7 are discussed in the following section.

We wish to estimate $\beta^{(i)}$ based on experimental observations. In Fig. 4.1 we reproduce the data of Weinbauer et al. [28], who measured in the northern Adriatic sea the burst size $b^{(i)}$ as a function of the cell volume $V_{bact}^{(i)}$ for bacterial radii ranging from ~0.2 to ~0.9μm and for two groups of capsid diameters: 30-60nm (group A; blue) and 60-110nm (group B; red). Note that in Fig. 4.1 the y-axis is plotted as the burst size times the average volume of a capsid for that group.

Therefore, if indeed $b^{(i)}V_{virus}^{(i)} = \beta^{(i)}V_{bact}^{(i)}$ with $\beta^{(i)} \equiv \beta \equiv \mathsf{const}$, we would expect the slope in Fig. 4.1 to be the same for both size classes. Indeed we estimate very similar values for β for both size classes: $\beta \approx 0.005$. Furthermore, when consolidating both size groups and assuming an average capsid diameter of 60nm (the peak value found in nature) we obtain $\beta \approx 0.0049$, in agreement with the previous results. We therefore find that over a wide range of bacterial sizes, 0.5% of the cell volume is occupied with viruses upon lysis. Closer inspection of the correlation suggests however that for small cell volumes (< ~1 μm^3, corresponding to a radius < ~0.6μm), the burst size is underestimated based on our simple linear formula given the above estimate for β. In a different work, Weinbauer et al. [29] studied the correlation between burst size and cell volume for small cells (<0.3 μm^3) in Lake Plußsee. Based on this correlation we find that cells with a volume of ~0.3 μm^3 had a burst size of ~90 while cells with a cell volume of ~0.05 μm^3 had a burst size of ~35. Assuming a typical capsid diameter of 60nm, this corresponds to β ~0.03 for the former case and β ~0.08 for the latter case. Since we will find however that small cells tend to be much more abundant than large cells, an underestimation of the burst size at small volumes may bias results. We will therefore assume that β is bounded in the range of ~0.5% to ~5%. It is less certain how well this relation will hold for bacterial radii >~1μm. For unicellular eukaryotes with radii in the range of ≈2 to ≈7μm we indeed estimate values of β in the range of 0.1% to 3.4% (Table 4.2), consistent with the above bounds, suggesting that our empirical correlation may hold for larger cells as well. Below we will discuss the case of extremely large bacteria that have massive cell inclusions that reduce the effective cytoplasm volume (and thus the value for β). Thus, for very large cells β may behave anomalously.

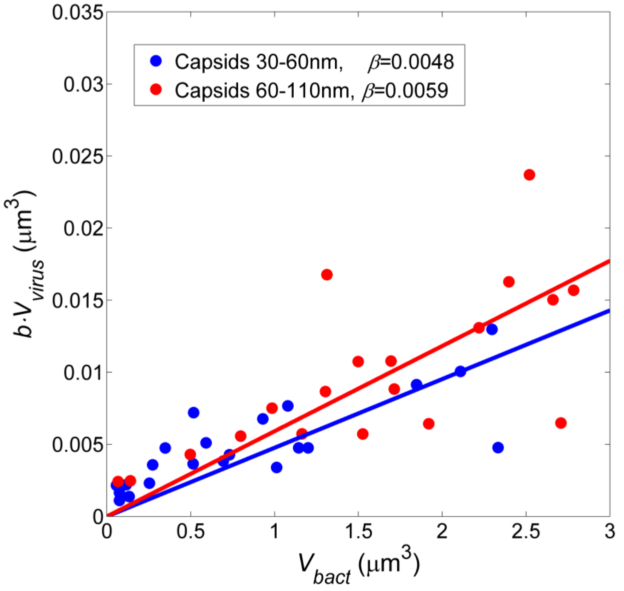


Figure 4.1. Correlation between burst size and cell volume. Here we reproduce the data from Weinbauer et al. [28] for the correlation between burst size and cell volume for two capsid diameter classes; 30-60nm (group A) and 60-110nm (group B). The y-axis was plotted as the burst size times the average capsid volume for that group. The average volume of a capsid for group A was calculated assuming a capsid diameter of (30+60)/2 = 45nm, while for group B the average capsid diameter was assumed to be (60+110)/2 = 85nm. The straight line is a least squares fit a line with a zero constant. The Pearson correlation coefficient for data points of group A was ρ=0.79, and for data points of group B was ρ=0.71.

Table 4.2. Estimation of virus volume fraction, β, for unicellular eukaryotes.

Eukaryote	Approx. radius	Burst size	Virus diameter	Ref.	β
E. huxleyi	≈2.3µm	400-1000 (mean 620)	≈170nm	[30]	≈3.1%
H. akashiwo	~5-7 µm*	~10^5	≈30nm	[30,31]	~0.1-0.3%
C. ericina	~7 µm*	1800-4100	≈155nm	[30,32]	~3.4%

*Size estimated based on different strain of this species

4.4.1.6 Dependence of host concentration on bacterium size

The most striking feature of Eq. 7 is the dependence of the concentration of the bacteria on the fourth power of the ratio $R_{virus}^{(i)} / R_{bact}^{(i)}$. Thus a bacterium that is twice as large is predicted according to this model to be $(1/2)^4$=1/16 times less abundant. This effect is both because larger bacteria have a larger cross section for diffusing viruses and because larger viruses produce more virions, thus for larger cells, fewer bacteria are needed for viral production to match viral decay.

Comparing the radii of viruses and their hosts, it appears that the radii of bacteria are much more variable in natural environments. The range of the dimensions of prokaryotes in nature is tremendous, ranging from a diameter of 0.2 to 750 μm [33,34] spanning over three orders of magnitude. When raised to the power of four this variable spans an astonishing 14 orders of magnitude. On the other hand, the diameter range of heads of tailed phages (that constitute about 96% of all phages examined to date via electron microscope [35]) is very narrow and lies between 34 and 160 nm, peaking sharply at 60 nm [36]. If we use the simple rule of thumb that the dimensions of viruses is fixed at 60 nm, then for a given environment defined by η and T we can plot the fixed point concentration of bacteria as a function of the size of the bacterium (Fig. 4.2). Since *a priori* we have no reason to believe there is a correlation between $\gamma^{(i)}_{viral\ decay}$ and $R^{(i)}_{bact}$, we will regard $\gamma^{(i)}_{viral\ decay}$ as a constant. We will further approximate $\beta^{(i)}$ as a constant that is uncorrelated with $R^{(i)}_{bact}$, though we are less certain how well this assumption will hold for larger bacteria (further discussed below). From Fig. 4.2 we see that small bacteria are predicted to achieve significantly higher concentrations and lower VBRs. Thus, according to this model, the size of a bacterium appears to be the most important factor determining its fixed point concentration.

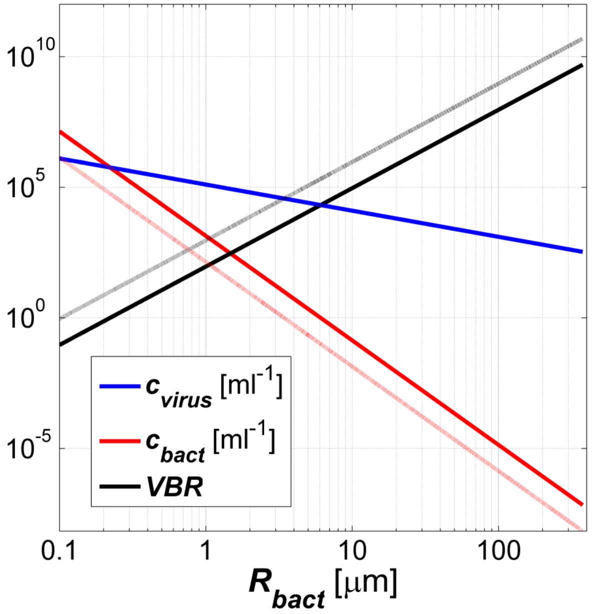


Figure 4.2. Scaling of the virus concentration, the bacterium concentration and the VBR with the radius of the bacterium for a single phage-host system. Virus concentration was calculated based on Eq. 5A and bacterium concentration was calculated based on Eq. 7. Parameters used for these equations, which are typical for marine systems, are given in Table 4.3, with $\gamma_{viral\ decay} = 2$ day^{-1} chosen to represent an offshore marine environment [23]. Solid lines are for β=0.005 while dotted lines are for β=0.05.

4.4.1.7 Large bacteria are rare

Fig. 4.2 demonstrates that very large bacteria should exist at extremely low concentrations in the ocean. The largest bacterium known to date, *Thiomargarita namibiensis*, with a diameter of up to 750 μm, found in marine sediments [37], is predicted to occur at a frequency of 1 cell per ~$1.5 \cdot 10^4$ liters of water (Fig. 4.2). For very large bacteria however, our assumption of a constant β most likely breaks down. Many of the large bacteria harbor massive cell inclusions that reduce the volume of the metabolically active cytoplasm [38]. In the case of *T. namibiensis* for example, its cytoplasm is restricted to a thin ~1 μm outer layer that surrounds a large central vacuole [37]. This inclusion therefore leads to a reduction in $\beta^{(i)}$ by two orders of magnitude:

$\beta^{(i)} \to \beta^{(i)}\left(1-\left(374\mu m/375\mu m\right)^3\right) \cong 10^{-2}\beta^{(i)}$. Thus, the effective concentration of this bacterium would be predicted to be higher by two orders of magnitude. Consequently, inclusions have the beneficial effect of increasing the abundance of the host by reducing its effective burst size. However, even with this large inclusion, assuming $\beta^{(i)} \approx 5\cdot 10^{-5}$, the host cell will still be very rare, with only one cell per ~100 liters of ocean water. Thus free-floating large cells can easily go undetected. It is therefore not surprising that *T. namibiensis* was discovered in sediments where it was found to be highly enriched [37] and not free floating in the ocean. Since large bacteria are predicted to be very rare in the open ocean and easily missed, it is worth noting that the viruses infecting such large cells are predicted to be relatively abundant, with several hundreds of virions per ml. Thus marine phages, according to this model, may be very sensitive proxies to rare, large bacterial cells.

Table 4.3. Typical parameters for phage-host systems in aquatic environments

Parameter	Value	Aquatic region	Reference
R_{virus}	≈30 nm	Many environments	[36]
R_{bact}	≈0.1-0.2 μm	Open ocean	[34]
α	~2 day^{-1}	Coastal	[18]
$\gamma_{non\text{-}viral}$	~1 day^{-1}	Coastal	Inferred
$\gamma_{viral\ decay}$	~0.1 to ~10 day^{-1}	Various marine	[19]
β	~0.5% to ~5%	Marine and lake	[28,29] (Inferred)
η	≈10^{-3} kg/(m·s)	-	
k_BT (T=24°C)	4.1·10^{-21} m^2·kg·s^{-2}	-	
$r_{\boldsymbol{max}}$	375 μm	Sediments	[34,38]
$r_{\boldsymbol{min}}$	0.1 μm	Open ocean	[34,38]
$D_{\boldsymbol{virus}}$	≈5 (μm)^2/s	λ phage	[39] [39]

4.4.1.8 Application of the model to environmental systems

The *Synechococcus* phage-host system in the Gulf of Mexico

Eqs. 5A and 5B are very powerful in the sense that they predict the absolute equilibrium concentration of hosts and their viruses from basic parameters describing the environment, the bacterium, and the virus that infects it. We wish to see how the model predictions of the concentration of particular phage-host systems compare with measurements of specific phage-host systems in nature. One particular system of interest is cyanobacteria, which has been studied extensively. The concentration of cyanophages in coastal waters and off shore waters in the Gulf of Mexico infecting *Synechococcus* (1.5 µm in diameter) peaked at $\cdot 140^{5}$ ml^{-1} at the ocean surface [23] with a VBR for this phage-host system measured to be as high as 8 [23]. Based on the depth profiles in this study we computed the average concentration of *Synechococcus*, the average concentration of cyanophages infecting *Synechococcus* strain DC2 and the average VBR (Table 4.4). We wish to compare these observations to model predictions.

Virus concentration

For the *Synechococcus* case study $R_{bact}^{(i)} = 0.75\,\mu\text{m}$. In a related study, capsid diameters of virions infecting a *Synechococcus* host were found to be in the range of 50–65 nm [40]. Thus we assumed that $R_{virus}^{(i)} \approx 30\,\text{nm}$ [36]. The average growth rate of bacteria in coastal waters is on the order of $\alpha^{(i)} \sim 2$ day^{-1} [18] (Table 4.3). At steady-state this growth rate equals the sum of the lysis rate and non-viral mediated mortality rate (see above). There is evidence to suggest that grazing and lysis contribute equally to microbial mortality [15,18,19], though this matter is still the subject of debate [19]. Nevertheless, as a first-order approximation we will assume that bacterial production is roughly halved by grazing so that $\gamma_{non-viral}^{(i)} \sim 1$ day^{-1}. Thus, given that for water at 24°C $\eta \approx 10^{-3}$ Pa·s, Eq. 5A predicts that $c_{virus}^{(i)} \approx 1.7 \cdot 10^5 ml^{-1}$. This prediction is of the same order of magnitude as the measurements described above (see Table 4.4).

Alternatively, using the Stokes-Einstein relation we can calculate that a virus with an effective sphere diameter of 60 nm in water at 24°C should have a diffusion constant of 7.25 $(\mu m)^2/s$. This value is close to the measured diffusion constant of λ phage at the same temperature, which is 4.97 $(\mu m)^2/s$ [39]. Using the diffusion constant of the virus, one can also calculate the fixed point concentration of viruses using Eq. 4 and obtain the same solution.

Host concentration

To calculate the concentration of the bacteria one needs to know the viral decay rate and the burst size. The viral decay rate was measured in this study to be 0.1 day^{-1} inshore and 2 day^{-1} offshore [23]. Given our earlier estimate of β=0.005 we find a burst size of $b^{(i)} = \beta \left(\frac{R_{bact}^{(i)}}{R_{virus}^{(i)}} \right)^3 \cong 80$. In a one-step growth experiment for a different strain of *Synechococcus*, the burst size was measured to be 250, in rough agreement with our simple linear model prediction. The one-step growth experiment burst size is most likely an overestimate since burst sizes of isolated phage-host systems are known to be consistently higher than those found in the environment since cells growing in culture are larger and thus produce more progeny and/or are better adapted to high nutrient concentrations [2]. Thus, assuming β=0.005, $\gamma_{virus\ decay}^{(i)}$~0.1 – 2 day^{-1} and $R_{virus}^{(i)}$~30 nm (see above) with the remaining parameters taken from Table 4.3, then based on Eq. 7 (or 5B) we find that $c_{bacteria}^{(i)} \sim$ 200 ml^{-1} to 4.3·10^3 ml^{-1}. These predictions are consistent with the range of observed concentrations of *Synechococcus* cells (Table 4.4). Note however that a true test of the model predictions would require comparing with seasonal averages and not with one time measurements.

Table 4.4. Measured concentration of *Synechococcus* and the cyanobacteria infecting it in the Gulf of Mexico versus model predictions

Variable	**Observed (*n*=21)[a]**		**Prediction[b]**
	Mean ± S.D.	**Range**	
c_{virus} *(cyanophages)*	$(5.6\pm8.6)\cdot10^{4}$ ml^{-1}	150 ml^{-1} to $2.5\cdot10^{5}$ ml^{-1}	$1.7\cdot10^{5}$ ml^{-1}
$c_{bacteria}$ (*Synechococcus*)	$(1.8\pm2.9)\cdot10^{4}$ ml^{-1}	3.0 ml^{-1} to $9.2\cdot10^{4}$ ml^{-1}	200 ml^{-1} to $4.3\cdot10^{3}$ ml^{-1}
VBR	6.3±8.1	0.2 to 30.7	40 to 780[c]

[a] Measurements based on depth profiles measured by Suttle and Chan [23]. Virus concentration corresponds to viruses infecting *Synechococcus* strain DC2.

[b] Predictions were made based on Eq. 5 and 7. See text for further details.

[c] A better prediction could be made if the data for each station was analyzed separately as there were significant differences between stations.

4.4.2 Model development part II: Non-interacting phage-host systems

4.4.2.1 A stochastic interpretation of bacterial and viral parameters

Thus far we have considered the case of an isolated phage-host system and have treated quantities that depend on the species index *i* as deterministic quantities, i.e., every index *i* corresponds to a different phage-host system with a different set of parameters. When considering a natural environment, many different species – i.e., phage-host systems – co-exist. One can therefore imagine a hypothetical “species sample space” comprised of many phage-host systems, where each time we draw a phage-host system with index *i* we obtain a set of values for all model parameters based on some joint density function. Hence, all parameters can be thought of as random variables drawn from some joint distribution. Since the concentration of viruses and bacteria are functions of these parameters, these variables can be thought of as random variables themselves.

Of all the parameters that $c_{bacteria}^{(i)}$ depends on, $R_{bact}^{(i)}$ has the widest range of values, spanning over three orders of magnitude. Furthermore, given that $R_{bact}^{(i)}$ is raised to the fourth power, it is by far the most sensitive parameter in Eq. 7 (see above). For comparison, the distribution of phage capsid diameters $2R_{virus}^{(i)}$ peaks sharply at 60 nm (see above). $\gamma_{viral\ decay}^{(i)}$ varies by about two orders of magnitude across environments [19], however, we expect that for a given environment, where all phages are subject to the same conditions, the range of $\gamma_{viral\ decay}^{(i)}$ will be more restricted. In addition, $c_{bacteria}^{(i)}$ is only linearly dependent on $\gamma_{viral\ decay}^{(i)}$. Finally, $\beta^{(i)}$ also appears to display limited variability (see above).

Therefore, if we assume, to a first-order approximation, that the random variables $R_{virus}^{(i)}$, $\gamma_{virus\ decay}^{(i)}$, and $\beta^{(i)}$ are statistically independent of the random variable $R_{bact}^{(i)}$, then we can average out these parameters by taking their expected value. If these parameters are also statistically independent of each other then we have

(8)
$$c_{bacteria}(R_{bact}=r) \approx \tfrac{3}{2}\frac{\eta}{k_B T} E\beta^{-1} E R_{virus}^{4} E\gamma_{viral\ decay} r^{-4} = \text{const}\cdot r^{-4}.$$

where E denotes the expectation operator. Thus, our hypothetical species sample space reduces to a single random variable, $R_{bacteria}^{(i)}$, drawn from some distribution $f_R(r)$, the functional form of which we do not know. Eq. 8 predicts what would be the average concentration of a particular bacterial species with radius r were it to exist in a given environment. In practice, the number of bacteria of a given radius present per ml of water in a given environment per radius,

$\rho_{environment}(r)$ (in units of (number)/m^4 see Table 4.5), depends on which bacteria happen to be in the given environment to begin with. Let's assume that a given environment contains $N_{species}$ different bacterial species i=1...$N_{species}$, with each species characterized by its own radius $R^{(i)}_{bacteria}$, where the subscript i labels the species. Thus, for a given realization of this environment, the distribution of observed bacterial radii would be given by

$$\rho_{environment}(r) = c_{bacteria}(r)\sum_{i=1}^{N}\delta(r - R^{(i)}_{bacteria}). \tag{9}$$

where $\delta(r)$ denotes the Dirac delta function (in units of m^{-1}) and where $R^{(i)}_{bacteria}$ are $N_{species}$ i.i.d.[1] random variables drawn from a distribution $f_R(r)$. Note that $f_R(r)$ is the probability density that a bacterium with radius $R_{bact} = r$ *a priori* exists in the environment, whereas $\rho_{environment}(r)$ is the actual concentration of bacteria observed in the environment per bacterial radius. Thus $\rho_{environment}(r)$ is one realization of the distribution of bacteria in the given environment. To obtain the ensemble average of $\rho_{environment}(r)$, averaging over many realizations of the given environment (making the simplifying assumption that in each realization there are always $N_{species}$ different species) one should calculate the expectation value of $\rho_{environment}(r)$ with respect to the $N_{species}$ random variables $R^{(i)}_{bacteria}$. In Section 4.7 we show that this ensemble average is given by

$$\langle \rho_{environment}(r) \rangle = N_{species} c_{bact}(r) f_R(r) \tag{10}$$

[1] Independent and identically distributed

Table 4.5. Variables and parameters used in the continuous phage-host interaction model

Variables	Definition	Units
$\rho_{environment}(r)$	Concentration of bacteria with radius *r* per radius, predicted to exist in a given environment	(number)/*m*^{4}
$N_{species}$	Total number of prokaryote species that exist in any given realization of the environment	dimensionless
V_{env}	Volume to find one cell of the largest bacterium (*r*=*r*$_{max}$), defining the effective size of the environment	*m*^{3}
$\rho_{species}$	Concentration of *species* (bacterial and viral) in the environment ($= N_{species} / V_{env}$)	(number)/*m*^{3}
$f_R(r)$	Probability density function from which the radius *r* of a bacterial species is drawn (also defined as the density of bacterial species)	(probability)/*m*
$f_\rho(r)$	Probability density function of radii measured in a given environment (empirically, the histogram of measured bacterial radii in a given environment)	(probability)/*m*
c_{bact}^{tot}	Concentration of all prokaryotes in a given environment	(number)/*m*^{3}
c_{virus}^{tot}	Concentration of all phages in a given environment	(number)/*m*^{3}
$m_{bact}(r)$	Wet mass of bacterium of radius *r*	*kg*
$M_{bact}(r)$	Wet mass density of prokaryotes of radius *r* per radius in a given environment	*kg/m*^{4}
M_{bact}^{tot}	Wet mass density of all prokaryotes in the environment	*kg/m*^{3}
VBR	Virus-to-bacterium ratio in the environment = $c_{virus}^{tot} / c_{bact}^{tot}$	Dimensionless
Parameters		
ρ_{cell}	Wet mass density of a cell	*kg*/m^3
r_{min} /r_{max}	Minimum/maximum radius of viable bacterium in nature	m
m_{min} /m_{max}	Minimum/maximum wet mass of viable bacterium in nature	*kg*

4.4.2.2 A simple evolutionary scenario

In the simplest evolutionary scenario we assume that there is no selection pressure on bacterial radii, i.e., bacteria of all sizes are equally adapted to survive and therefore can all have equal probability to exist *a priori* in a given environment. Consequently evolution did not evolve more small bacterial species than large bacterial species, and hence the density of bacterial species per radii is constant. This hypothesis therefore implies that all bacterial radii are equally probable to exist and therefore the random variables $R_{bacteria}^{(i)}$ should be drawn from a uniform distribution: $R_{bacteria}^{(i)} \sim U\left(r_{\min}, r_{\max}\right)$, where $r_{\min}$ and $r_{\max}$ are the minimum and maximum radii for a viable bacterium, respectively, and where $U\left(a,b\right)$ denotes a uniform continuous distribution in the range [*a*, *b*], thus

$$f_R\,(r) = \begin{cases} \left(r_{\max} - r_{\min}\right)^{-1} & r_{\max} \le r \le r_{\min} \\ 0 & \text{otherwise} \end{cases} . \tag{11}$$

Thus $f_R\,(r)$ can be interpreted as the density of bacterial species, perhaps analogous to the density of states in statistical mechanics, and reflects the evolutionary history of bacteria in the given environment. If the radii of all bacterial species that have adapted to survive in the given environment were known, one could, in principle, calculate $f_R\,(r)$ directly. Given this scenario, using Eq. 8 and Eq. 10, we find that the ensemble average of the concentration of bacteria expected to exist in a given environment is given by

$$\left\langle \rho_{environment}(r) \right\rangle \approx \tfrac{3}{2} N_{species} \frac{\eta}{k_B T} E\beta^{-1} ER_{virus}^4 E\gamma_{virus\ decay} r_{\max}^{-1} r^{-4} = \text{const} \cdot r^{-4} \tag{12}$$

where we have assumed that $r_{\min} << r_{\max}$.

4.4.2.3 The size spectra of bacteria in aqueous environments

To calculate the size spectra of bacteria in the environment we first derive the probability density function (pdf) of observed radii in the environment. This function is obtained by normalizing $\left\langle \rho_{environment}(r) \right\rangle$ given in Eq. 12:

$$f_\rho(r) = \frac{\left\langle \rho_{environment}(r) \right\rangle}{\int_{r_{\min}}^{r_{\max}} \left\langle \rho_{environment}(r) \right\rangle dr} = 3\left(r_{\min}^{-3} - r_{\max}^{-3}\right)^{-1} r^{-4} \approx 3 r_{\min}^{3} r^{-4} \tag{13}$$

where $r_{\min} \le r \le r_{\max}$ and where we assumed that $r << r_{\max}$. Note that $f_{\rho}(r)$ is the pdf of $\langle \rho_{environment} \rangle$ where as $f_R(r)$ is the pdf of R_{bact}. Thus $f_{\rho}(r)dr \approx 3r_{\min}^{3} r^{-4} dr$ is the probability of observing bacteria with radii between *r* and *r+dr* in a given environment. The probability that a bacterium of random volume *V* is greater than or equal to a given volume, *v*, would then be given by

$$\begin{aligned}(14)\quad & \text{Prob}(V \ge v) = \text{Prob}(R \ge r) = \int_{r}^{\infty} f_{\rho}(r')dr' = \int_{r}^{R_{\max}} 3\left(r_{\min}^{-3} - r_{\max}^{-3}\right)^{-1} r'^{-4} dr' \\ & = \left(\frac{r_{\min}}{r_{\max}}\right)^{3} \left[\left(r_{\max}/r\right)^{3} - 1\right] \approx \left(\frac{r_{\min}}{r}\right)^{3} = \frac{v_{\min}}{v}.\end{aligned}$$

assuming that $r << r_{\max}$ (see Section 4.7 for further details). When plotting log(Prob(*V*≥*v*)) against log(*v*) one obtains a power law with slope -1. In 2001 Cavender-Bares *et al.* [41] measured the size spectra of microbes up to a diameter of ~5 μm (i.e., from bacteria to nanophytoplankton) in the western north Atlantic Ocean. The researchers found that when plotting log(Prob(*V*≥*v*)) versus log(*v*), measurements fell on a straight line with a slope ranging between -1 and -1.4. The ensemble average of all environments was well described by a power law of slope -1.2. When expanding their dataset to include microzooplankton the slope was corrected to a value close to -1. A slope of -1 was also found earlier by Sheldon et al. [42].

Eq. 14 also predicts that the power law behavior with slope -1 is an intrinsic scaling property of the biophysical/biological dynamics of phages and their hosts and of $f_R(r)$, and therefore should remain unchanged under perturbations (irrespective of the functional form of $f_R(r)$). Thus perturbations increasing the viral decay rate or increasing bacterial growth rate, etc., should not have an effect on this power law. This prediction was validated in IronEx II [41], an iron enrichment experiment in the equatorial Pacific, where it was shown that the slope of the power law for samples taken from outside and inside fertilized waters over the course of the experiment differed by little [41]. In both cases the power law was measured to be in the range of -1.1 to -1.2 [41].

4.4.2.4 Possible deviation from a uniform distribution

If we take into account that β tends to decrease with r, we would expect a weaker slope for the size spectra. This result may indicate that a more realistic evolutionary scenario would be one in which larger bacteria are less probable, i.e., the density of bacterial species is higher for small radii. Indeed, small cells may have certain advantages over larger cells. For example, since small cells are more numerous, their population explores collectively more mutations allowing them to adapt more quickly to changing environments and allows them to more easily exploit new habitats [34]. In addition, the high surface-to-volume ratio of cells with smaller radii allows them more efficient exchange of nutrients and higher specific metabolic rates [34,38] possibly giving them a selective advantage.

4.4.2.5 Total bacterial concentration

To obtain the total concentration of bacteria in a given environment we integrate $\langle \rho_{environment}(r) \rangle$ (Eq. 10) over the range of viable bacteria sizes

$$c_{bact}^{tot} = \int_{r_{\min}}^{r_{\max}} \langle \rho_{environment}(r) \rangle dr = N_{species} \int_{r_{\min}}^{r_{\max}} c_{bacteria}(r) \cdot f_R\ (r) dr. \tag{15}$$

Note that Eq. 15 can also be rewritten as $c_{bact}^{tot} = N_{species} \cdot Ec_{bacteria}$, that is the total concentration of bacteria in a given environment equals the total number of species in a given environment, $N_{species}$, times the mean concentration of a single bacterial species. Inserting the population average of $c_{bacteria}^{(i)}$ given in Eq. 8, and assuming again a uniform distribution $f_R(r) = (r_{\max} - r_{\min})^{-1}$ for $r_{\min} \le r \le r_{\max}$ we find that

(16A)
$$c_{bact}^{tot} \cong \tfrac{1}{2} N_{species} \frac{\eta}{k_B T} E\beta^{-1} ER_{virus}^4 E\gamma_{viral\ decay} r_{\max}^{-1} r_{\min}^{-3}.$$

4.4.2.6 Species richness

Given a known total concentration of bacteria (determined either by protist grazing or nutrient availability), Eq. 16A can be reversed to predict the number of species in the given environment:

(16B)
$$N_{species} \cong 2 \frac{k_B T}{\eta} \frac{1}{E\gamma_{viral\ decay} ER_{virus}^4 E\beta^{-1}} c_{bact}^{tot} r_{\max} r_{\min}^3.$$

The largest bacterium found to date has a diameter of 750 μm (see above) and the smallest bacterium has a diameter of 0.2 μm (Table 4.3), close to the theoretical lower limit thought to be 0.14 μm [43]. Thus, given a typical marine scenario (such as the open ocean) in which direct observation reveals ~10^5 bacterial cells per ml [44] (i.e., $c_{bact}^{tot} \sim 10^5$ ml^{-1}), then based on the parameters in Table 4.3, which are typical for marine systems, and assuming a viral decay rate of $\gamma_{viral\ decay} \sim 2$ day^{-1} for offshore ecosystems [23], we find via Eq. 16B that the total number of bacterial species in any given realization of the environment is $N_{species}$ = 82 to 820, thus $N_{species}$ ~10^2 to ~10^3. $N_{species}$ is thus the number of species (i.e., phage-host systems) that any realization of the environment must contain in order to reach the observed total bacterial concentration. Thingstad & Lignell [10] also calculated the number of species in the environment given a fixed total concentration of bacteria, however in their model the authors assumed that all hosts were identical, ignoring their distribution in the environment. Eq. 16B is expected to be a more realistic estimate since we take into account the distribution of species in the environment. Because many species can be very rare (i.e., have a low concentration due to a large radius — Fig. 4.2) the total predicted diversity is expected to be much higher.

4.4.2.7 What is a species?

Since our model is capable of predicting the number of species in a given environment, we should ask ourselves, what precisely are we counting? What is the definition of a "species" according to our model? This question has practical meaning because we would like to test our prediction against observation. However, there are many "cutoff" values for genetic diversity. For example, is a "species" equivalent to a "species" in biology? Is it equivalent to a "strain"? Does one mutation constitute a new "species"?

In the context of our model here, a "species" of a bacterium is defined by (a) a set of random variables (e.g., the size of the bacterium, its growth rate, etc.) (b) having a unique association with a "viral species" independent of all other phage-host systems, and finally, (c) there is an equal number of "bacterial species" as "viral species". However, this definition is not sufficient. If, for example, two hosts have exactly the same parameters, they could still have totally different genomes, and thus constitute distinguishable entities that should be counted separately. Thus, to say that two hosts with the same parameters are identical would be wrong. All that our biophysical model predicts is the number of independent phage-host systems that can be accommodated in a given environment. It does not define how these phage-host systems are different. Therefore a more detailed definition of what a "species" is lies outside the scope of the present model, necessitating us to dig deeper.

An analogy to physics. This paradoxical situation is often encountered in physics. To draw on a physics analogy, our biophysical model's description of a species is analogous to nuclear physics' description of a nucleus, which makes the abstraction that the nucleolus is comprised of protons and neutrons. In nuclear physics, protons and neutrons are regarded as point particles defined by certain quantum numbers (like our random variables describing a "species"). Within the framework of this theory though, it is meaningless to ask what is the internal structure of these particles. Likewise, within the context of this biophysical model it doesn't make sense to ask what the structure of a "species" is. To better understand what a proton and neutron is, a more sophisticated model was required, called the standard model, which showed that protons and neutrons are made out of quarks held together by gluons. In Chapter 5 we propose the "standard model" of phage-host interaction, which allowed us to probe the "internal" structure of a "species". The model proposed in Chapter 5 is a speciation model describing how new *species* of both bacteria and viruses are generated in nature, leading to a "world" of non-interacting phage-host systems, consistent with the present biophysical model. Drawing on evolution, the new model adds another metric to our description of these organisms, which is the evolutionary distance metric. Therefore in Chapter 5 we will be able to describe a model where a *species* is comprised of many *strains*, and explain how *strains* evolve into *species*. Again, drawing on our physics analogy, our *strains* will be the "quarks and the gluons" that comprise our *species*. We will therefore revisit the question of "what is a species" in Section 5.2.3 after developing our evolutionary model. The final answer we will arrive at is that $N_{species}$ is (to within a small correction factor between) the total number of consistently distinguishable genomes (termed *strains*), a very intuitive result, with the twist that the "species" defined in our coarse-grained model are actually comprised of a collection of *strains*.

A second question that arises from our calculation is in what volume, according to our model, should we find these species? We will answer this question in the next section, and by answering this question we will be able to calculate the density of *species* in the ocean, from which we will be able to calculate an upper bound on the total bacterial diversity in the oceans.

4.4.2.8 Volume of diversity

The minimum volume that needs to be sampled to detect the $N_{species}$ species is determined by the lowest predicted concentration of bacteria, namely the concentration of the largest bacteria. This volume is given by

$$V_{env} = \left[\tfrac{3}{2} \frac{1}{\beta^{(\max)}} \frac{\eta}{k_B T} \left(\frac{R_{virus}^{(\max)}}{r_{\max}} \right)^4 \gamma_{viral\ decay}^{(\max)} \right]^{-1} .$$

where the index "max" corresponds to the bacterial species with the largest diameter. Taking the model at face value, given $\beta^{(\max)}$=0.005, $\gamma_{viral\ decay}^{(\max)}$=2 day^{-1} (offshore waters) and with the remaining parameters taken from Table 4.3 we find that at least ~15,000 liters of water are required to detect the largest known bacterium (see above). In other words, ~1.5·10^4 liters of water must contain ~10^2–10^3 species of bacteria in order to account for ~10^5 cells per ml. In practice this volume may be two orders of magnitude smaller due to an uncertainly in β for very large cells (see Section 4.4.1.7). Thus, anywhere between ~10^2 liters to ~10^4 liters of water are required to be sampled in order to observe the predicted number of prokaryotic species.

4.4.2.9 Species density

Dividing $N_{species}$ from Eq. 16B by V_{env} we obtain the following expression for the "species density":

$$\rho_{species} \triangleq \frac{N_{species}}{V_{env}} = 3\frac{1}{E\beta^{-1}} c_{bact}^{tot} \left(\frac{r_{\min}}{r_{\max}}\right)^3 \frac{1}{\beta^{(\max)}} \approx 5\cdot 10^{-9} c_{bact}^{tot}.$$

.

where we have assumed that $E\gamma_{viral\ decay} \approx \gamma_{viral\ decay}^{(\max)}, \left(R_{virus}^{(\max)}\right)^4 \approx ER_{virus}^4$, $r_{\max}/r_{\min} = 3750$ (Table 4.3), and, based on *T. namibiensis,* $\beta^{(\max)}\left(E\beta^{-1}\right)^{-1} \sim 10^{-2}$ (see above).

4.4.2.10 Observed species diversity in nature

4.4.2.10.1 Estimates of microbial diversity

How do these predictions compare with the measured prokaryotic diversity in marine systems? In a metagenome study of the Sargasso Sea, where the concentration of bacteria was indeed measured to be ~10^5 ml^{-1} [45], it was estimated that each sample, consisting of 170–340 liters of ocean water, contained a minimum of 300 species per sample [45]. A model based on assembly depth coverage estimated between 1800 and 48,000 species [45]. A "species" in this study is defined as "a clustering of assemblies or unassembled reads more than 94% identical on the nucleotide level", which is "roughly comparable to the 97% cutoff traditionally used for the rRNA" [45]. In terms of rRNA diversity, in the combined study there were 1412 distinct small rRNA sequences spanning different prokaryotic phyla. Applying a similarity cutoff of 99% reduced this number to 643 strains and applying a 97% similarity cutoff reduced his number further to 148 phylotypes [45]. Given a bacterial concentration of ~10^5 ml^{-1} for the open sea observed in this study [45] and an offshore viral decay rate of 2 day^{-1} [23], our model predicts $N_{species}$~10^2-10^3 species (see above). Although the predicted value for $N_{species}$ is in agreement with the observed rRNA diversity/microdiversity and in rough agreement with the observed number of species, it is not entirely obvious how to compare $N_{species}$ with the observed diversity. If a species is defined as a "distinguishable" genetic entity, it is not clear that the rRNA is the correct indictor for the number of species, as in principle two genomes can be "distinguishable" but have identical rRNAs. However, it is not clear that the number of "distinguishable" genetic entities is the correct measure to compare $N_{species}$ with, since in Section 5.2.3 we will see that "strains", which are defined to be "distinguishable" genetic elements, may not be under the sole control of a single viral species, and therefore "take up" less concentration. Thus, a certain similarity cutoff seems to be required. However, it is currently not clear how to translate this observation into an effective cutoff.

4.4.2.10.2 Viral diversity

Our model is constructed so that the number of viral species equals the number of bacterial species. Therefore we can also compare this estimate to estimate of viral diversity in the oceans. In another metagenome study, a viral metagenome was obtained from 200 liters collected from the surface seawater from Scripps Pier and a second sample was collected from Mission May, San Diego [46]. From these samples marine viruses were isolated using a combination of differential filtering and density-dependent gradient centrifugation. Several mathematical models based on the observed number of contigs predicted between 374 and 7114 viral types. Assuming a concentration of ~10^6 ml^{-1} for coastal waters and a decay rate of 0.1 day^{-1} for inshore waters [23], our model predicts that $N_{species}$ = ~10^4 to ~10^5, within rough agreement of these estimates. Here too, it is not clear what should be the correct "species" cutoff and an overestimation of diversity is not necessarily incorrect (see Section 4.4.2.10.1).

4.4.2.11 Bounds on global marine diversity

Given the expression for $N_{species}$ and $\rho_{species}$ we can attempt to estimate the minimum and maximum number of species in the Earth's oceans.

Lower bound on diversity — the case of a homogeneous ocean

To estimate the minimum bound on diversity we will assume the ocean is completely homogenous, and therefore extrapolate from one region in the ocean (of the highest diversity) to the entire ocean (Fig. 4.3B). Assuming a low (onshore) decay rate for viruses of ~0.1 day^{-1}, and a concentration of ~10^6 ml^{-1} we obtain an estimate of $N_{species} \sim 10^4 - \sim 10^5$. Though the sediment contains ~10^3 more cells per ml, and viral decay rates are comparable to the surface of the ocean [47,48], the particles in this region are probably not modeled well by free diffusion and therefore we will not use this region of the ocean to calculate our lower bound. The actual number of species must be higher than this since the ocean contains different regions with unique species adapted only to that region.

Upper bound on diversity — the case of a maximally heterogeneous ocean

An upper bound can be obtained by assuming that every volume V_{env} in the ocean contains a different sample of species, assuming each volume V_{env} contains the typical diversity found in the ocean (Fig. 4.3C). Thus $N_{species} \sim \rho_{species} V_{ocean} \sim 10^{-8} c_{bact}^{tot} V_{ocean}$. Given that V_{ocean}~10^{24} ml [44], and for the open sea $c_{bact}^{tot} \sim 10^{5}\,\mathrm{ml}^{-1}$, the maximum number of species in the ocean would be $N_{species}$~ $10^{-8} \cdot 10^{5} \cdot 10^{24} = 10^{21}$. Note that $\rho_{species}$ scales as $c_{bact}^{tot} \left(r_{\min} / r_{\max} \right)^{3}$. Thus, the upper bound on the total diversity in the oceans essentially depends only on the "carrying capacity" of the ocean and on the size of the smallest and largest viable bacteria.

Thus the total number of actively replicating prokaryotic cells with a *distinguishable* genome in Earth's oceans is predicted to lie somewhere between 10^{4} to 10^{21}. Although 10^{21} is a large number, it is exceedingly smaller than the total number of possible bacterial strains, and 7 orders of magnitude lower than the total number of bacterial cells in the ocean, estimated to be ~10^{29} [44]. This upper bound is of course a gross overestimate since adjacent volumes of water exchange cells constantly, and therefore the overlap in species between adjacent "volumes of diversity" will be very large. Note that using this approach one could obtain much tighter bounds on species diversity for smaller volume ecosystems such as lakes.

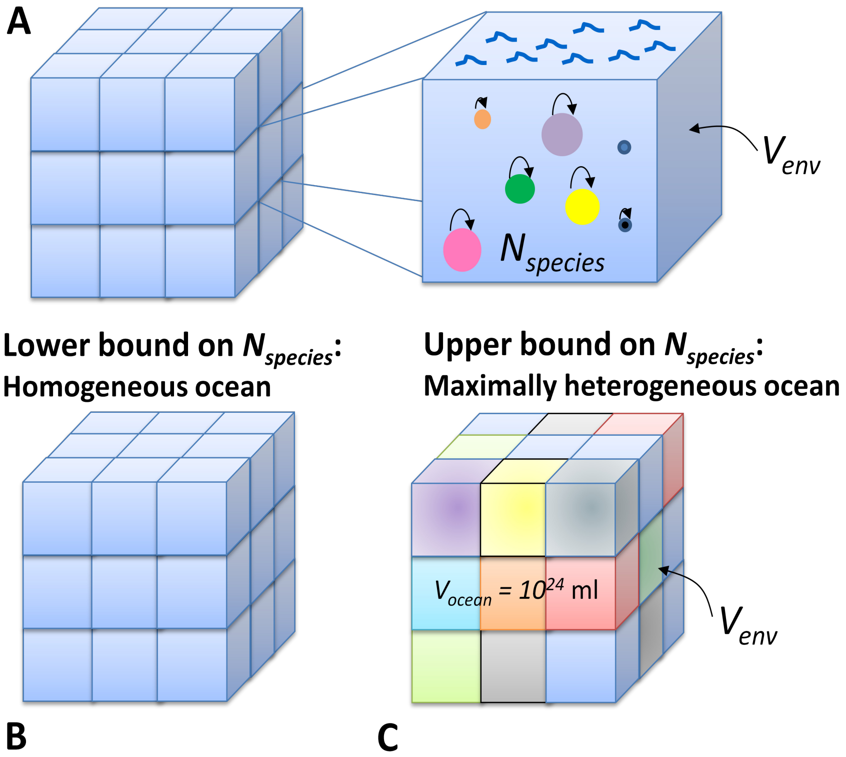

A
V_{env}
$N_{species}$
Lower bound on $N_{species}$:
Homogeneous ocean
Upper bound on $N_{species}$:
Maximally heterogeneous ocean
V_{ocean} = 10^{24} ml
V_{env}
B
C

Figure 4.3. Illustration of lower and upper bounds on $N_{species}$. A. Each volume of diversity, V_{env}, contains $N_{species}$ *species* given by Eq. 16B. The concentration of each *species* is controlled by its lytic virus. **B.** In the case of a homogenous ocean scenario, all volumes of diversity contain the same *species*, resulting in a lower bound on the total diversity in the oceans **C.** In the case of a maximally heterogeneous ocean, every volume of diversity contains a different set of *species*, resulting in an upper bound on the total diversity in the oceans.

Current observed diversity in public databases

How do these values compare with the current estimates of diversity? If one uses the small subunit rRNA gene as a proxy for genetic diversity, then one can compare our range estimates for the total number of species in the oceans' water column with the total number of rRNA sequences that are >1% divergent. The Silva SSU Ref NR 106 [49] released in April 2011 is a non-redundant SSU rRNA database with an operational taxonomical unit (OTU) cutoff of 1%. According to this database there are 2.19 ^{5} bacterial and archeal non-redundant SSU rRNA sequences. While an OTU of 1% will give us a lower bound on the number of "distinguishable" genomes (see Chapter 5), taken at face value, this comparison suggests that the ocean appears to be more homogenous than heterogeneous.

4.4.2.12 Factors determining species richness

Nutrient availability

Eq. 16B leads to some interesting predictions regarding the diversity of species in different environments. An eutrophic environment for example, which can sustain a higher concentration of bacteria (assuming total bacterial concentration is not determined by grazers [17]), is predicted via Eq. 16B to harbor a larger number of species. Increasing the total concentration of bacteria by a factor of ten will lead via Eq. 16B to ten times the number of species and, as will be discussed below, also ten times the concentration of viruses. In fact, the increase in species diversity can even be significantly higher than a factor of ten since $N_{species}$ is proportional to $r_{\min}^3 r_{\max}$, and cell size often increases with growth rate, which in turn increases with nutrient availability. Thus, an increase in nutrient availability may lead to an explosion in species diversity (and also, possibility a significant increase in the VBR, discussed below). Conversely, oligotrophic environments, where the concentration of bacteria can be lower, are predicted to harbor fewer species. Thus, a direct prediction of our model is that eutrophic environments harbor a larger diversity of species compared with oligotrophic environments, given similar temperature conditions and similar viral decay rates.

Viral decay rate

Another interesting parameter that comes into play is the virus decay rate. The more viruses are allowed to thrive (i.e., decay more slowly, thus having a lower $\gamma_{viral\ decay}$), the lower the concentration of any phage-host system will be (Eq. 8), thus requiring more species to reach a given carrying capacity (Eq. 16B). Thus viruses directly contribute to bacterial species diversity, and so "what's good for the virus is good for the bacterium". Since the number of bacterial species must equal the number of viral species, generating bacterial diversity also means generating viral diversity. In Section 4.5 we prose closed and open mesocosm experiments, where we show that in both cases, increasing the viral decay rate should lead to a decrease in the number of species.

The reciprocal relationship between bacterial diversify and viruses has been proposed in the past [9,10], however here we have expanded this concept by combing several ideas: (a) we have taken into account the biophysical nature of phage-host interaction, which allowed us to describe quantities in terms of physical parameters such as temperature, viscosity, the size of the virus, and the size of the host. (b) We have taken into account the observed correlation between burst size and the physical dimensions of the host and its virus. Finally, (c) we have introduced the notion of the "density of bacterial species" that was used in a statistical fashion to make the transition from a single isolated phage-host system to a community distribution. These concepts have led us to derive a realistic prediction for the number of species given in terms of physical measurable parameters and also define a physical volume associated with the predicted diversity.

Temperature

Eq. 16B predicts that given the same carrying capacity, warmer environments will contain more species. Overall this effect is not very large however. The difference between an environment just above freezing and 40°C will lead to only a 15% increase is species diversity ((273 + 40)/273=1.15), unless temperature will have an effect on r_{min} and r_{max} through its effect on growth rates (see above). The quantitative predictions of Eq. 16B may however be biased for extreme temperature environments since the selection pressure in such environments may skew the bacterial radius density function, violating our assumption of uniformity.

Extreme bacteria

We have already noted that above that $N_{species} \propto r_{\min}^3 r_{\max}$. Thus, halving the size of the minimum viable bacterium would lead to decreasing the total diversity in the environment by about one order of magnitude (2^3), due to the great abundance of small bacteria (thus reaching the carrying capacity more quickly). On the other hand, doubling the size of a largest bacterium would only lead to a modest doubling of the total bacterial diversity in the given environment since we are adding rare species with low concentrations that do not contribute much to the total concentration, thus necessitating more species to reach a given carrying capacity. One possible way r_{min} and r_{max} may be influenced is through nutrient availability, as discussed above.

4.4.2.13 The total concentration of viruses and the VBR in the environment

Total concentration viruses

In a similar fashion we can calculate the predicted total concentration of viruses in the environment. Since the average concentration of viral species is $Ec_{virus}^{(i)}$, the total concentration of viruses is simply $c_{virus}^{tot}(r) = N_{species} Ec_{virus}^{(i)}$. Inserting Eq. 5 we find that

$$(17) \qquad c_{virus}^{tot}(r) = N_{species} Ec_{virus}^{(i)} = \tfrac{3}{2} N_{species} \frac{\eta}{k_B T} ER_{virus}^{(i)} \left(E\alpha^{(i)} - E\gamma_{non-viral}^{(i)} \right) E\left(\frac{1}{R_{bact}^{(i)}} \right).$$

Assuming once again a uniform distribution for the bacteria (Eq. 11) we find that

$$E\left(\frac{1}{R_{bact}^{(i)}}\right)=\int_{r_{\min}}^{r_{\max}} r^{-1} f_R(r)dr \cong r_{\max}^{-1}\int_{r_{\min}}^{r_{\max}} r^{-1}dr = r_{\max}^{-1}\ln\left(\frac{r_{\max}}{r_{\min}}\right). \tag{18}$$

thus

$$c_{virus}^{tot} = \tfrac{3}{2} N_{species}\frac{\eta}{k_B T} ER_{virus}^{(i)}\left(E\alpha^{(i)} - E\gamma_{non-viral}^{(i)}\right) r_{\max}^{-1}\ln\left(\frac{r_{\max}}{r_{\min}}\right). \tag{19}$$

The concentration of viruses in the ocean

For an offshore ecosystem with $\gamma_{viral\ decay}$ ~ 2 day^{-1} and $c_{bact}^{tot} \sim 10^5 ml^{-1}$ we previously found that $N_{species}$ ~10^2 to ~10^3 species. To calculate the total concentration of viruses in the environment we use the above estimates and the remaining parameters from Table 4.3 and find that c_{virus}^{tot} = ~2·10^5 to ~2·10^6 ml^{-1}, or ~10^5 to ~10^6 ml^{-1}. This prediction falls exactly in the range of observed concentrations: virus concentrations in offshore surface waters are typically in the range of 10^5–10^6 ml^{-1} [2].

The VBR in a given environment

With the total concentration of bacteria at hand we can now calculate using Eq. 16A and Eq. 19 the VBR in the environment:

$$VBR = \frac{c_{virus}^{tot}}{c_{bact}^{tot}} = 3\frac{\left(E\alpha^{(i)} - E\gamma_{non-viral}^{(i)}\right)}{E\beta^{-1}E\gamma_{viral\ decay}}\left(\frac{r_{\min}}{R_{virus}}\right)^3 \ln\left(\frac{r_{\max}}{r_{\min}}\right) \tag{20}$$

where we have assumed that $ER_{virus}^4 \approx \left(ER_{virus}^{(i)}\right)^4 \approx R_{virus}$. Given the parameters in Table 4.3 for a typical marine system, with an offshore viral decay rate of 2 day^{-1} [23], Eq. 20 predicts that *VBR*

~2 to ~20, precisely as observed for typical marine systems [1]. The key observation predicted by this formula is that VBR is essentially controlled by the following parameters: (1) the net average growth of bacteria (growth minus predation), (2) the decay rate of viruses, and (3) the minimum viable bacteria (which may be related to nutrient availability). β and R_{virus} have a relatively narrow distribution and the effect of r_{max} is subdued due to the log. Thus, this basic equation can be used to predict both qualitatively and quantitatively the VBR in any aqueous environment.

Examples for environmental VBRs

VBR in nutrient-rich versus nutrient-poor environments

It has been observed that the VBR is higher for nutrient-rich, productive environments compared with nutrient-poor environments [19]. This has been attributed to the fact that "bacterioplankton host populations produce greater numbers of viruses under environmental conditions favoring fast growth and high productivity" [19]. Eq. 5 indeed predicts that — all things being equal — the higher the average growth rate of the bacterium, the higher the concentration of viruses will be, and consequently the higher the VBR. This prediction is also apparent from Eq. 20 for the VBR, where it is shown that the VBR is directly proportional to the average net growth rate of bacteria in the environment. In addition, cell size often increases with growth rate, which increases with the availability of nutrients. Since based in Eq. 20 the VBR is proportional to $r_{\min}^3$, even a modest increase in r_{min} would lead to a significant increase in the VBR.

VBR in oceans versus lakes

In a recent study it has been shown that the VBR is higher in marine systems than in freshwater systems [1,50]. In the surface waters of the Pacific and Arctic oceans for example, the VBRs are ~40 and ~10 respectively, while in lakes the average VBR was measured to be less than 5 [50]. Though the reasons for these differences are unknown, it has been suggested that this is related to possible higher loss rates of virus particles in freshwater environments that may be related to the presence of clays and chemicals from the terrestrial environment, which are known to contribute to viral decay [1,50]. This hypothesis is consistent with the prediction of Eq. 20, namely that the VBR should decrease with increased viral decay rate.

4.4.2.14 Total prokaryotic biomass concentration

The predicted slope of -1 (Eq. 14) for the size spectra of bacteria suggests that on average there is a tendency toward a uniform distribution of mass among size classes in aquatic ecosystems [41,42]. This result follows from Eq. 13: let $m_{bact}(r) = \frac{4}{3}\pi r^3 \rho_{cell}$ be the mass of a bacterium of radius *r* having a cellular mass density of ρ_{cell}. The total mass concentration per cell radius, given Eq. 12 for $\langle \rho_{environment}(r) \rangle$, scales as $M_{bact}(r) = \langle \rho_{environment}(r) \rangle m_{bact}(r) \sim r^{-1}$. Therefore the total mass concentration between radius r_1 and r_2 is given by

$$\text{(21)} \qquad \text{Mass}(r_1 < r < r_2) = \int_{r_1}^{r_2} M_{bact}(r)dr \propto \int_{r_1}^{r_2} r^{-1}dr = \ln\left(\frac{r_2}{r_1}\right).$$

Thus the total mass of prokaryotes between *r* and 10*r* equals the total mass of prokaryotes between 10*r* and 100*r* and so on. Integrating over all viable bacterial radii (using Eq. 12) we obtain the total mass of prokaryotes per unit volume

$$\text{(22A)}\qquad M_{bact}^{tot} = \int_{r_{\min}}^{r_{\max}} M_{bact}(r)dr \cong 2\pi\rho_{cell} N_{species} \frac{\eta}{k_B T} E\beta^{-1} ER_{virus}^{4} E\gamma_{viral\ decay} r_{\max}^{-1} \ln\left(\frac{r_{\max}}{r_{\min}}\right).$$

where we have assumed again that $r_{\min} << r_{\max}$. Combining Eq. 16A and Eq. 22A we find that

$$\text{(22B)}\qquad M_{bact}^{tot} \cong c_{bact}^{tot} m_{\min} \ln\left(\frac{m_{\max}}{m_{\min}}\right)$$

where $m_{\min}$ and $m_{\min}$ are the mass of minimum and maximum viable bacteria. Thus Eq. 22A predicts the total prokaryotes mass concentration in the ocean in terms of basic parameters such as: environmental parameters (viscosity and temperature of the water), viral parameters (average radius, average decay rate, and volume fraction within host cell) and host parameters (mass — or water — density, number of bacterial species, and minimum and maximum radii of viable bacteria). Assuming $\rho_{cell} \approx 1$ g/ml then $m_{\min} = \frac{4}{3}\pi\rho_{cell} r_{\min}^{3} = 4.2 \cdot 10^{-15} g$, and the total mass density of prokaryotes is $M_{bact}^{tot} = 10$ mg/m^3 (including cytoplasmic water). This mass can be compared with the following simple order-of-magnitude estimate. The typical radii of bacteria in the open ocean is 0.1–0.2 μm (Table 4.3). The mass of such a bacterium is given by $m_{bact}(r \approx 0.2$ μm$)$ = $\frac{4}{3}\pi r^3 \rho_{cell}$~$10^{-11}$ mg. Assuming that all 10^5 cells per ml have a radius of 0.2 μm, then the total mass of cells in 1 m^3 would be 10^{-11} mg × (10^{11} cells per m^3) = 1 mg. Thus, most of the mass contribution, according to Eq. 22, comes from the larger, rarer bacteria, and not the more abundant small bacteria. This can also be appreciated by noting that, whereas the total number of prokaryotes up to radius *r* scales as ~ r^{-3} (Eq. 14), thus decaying very fast, the total mass of prokaryotes up to radius *r* scales much more slowly as $\sim \ln(r)$.

4.5 Conclusions and further experiments

We developed a simple biophysical model that describes the interaction of an isolated phage-host system leading us to conclude that the single most important parameter determining the abundance of bacteria in the ocean is their size. We then extended our model to an ecological scale by making the assumption that the *a priori* distribution of bacterial radii in the environment is uniform, i.e., there is no selection pressure shaping this distribution. Given these basic ingredients we derive a model that makes reasonable predictions for the size spectra of bacteria, the VBR and the number of bacterial/viral species in the environment that largely seem to be consistent with observations. To further test our model we propose the following experiments:

4.5.1 *In vitro* investigation of phage-host systems

By choosing a particular phage-host system such as T4 and *E. coli*, one can analyze infected cultures *in vitro* as different model parameters are perturbed. To prevent total lysis of the hosts one should include an ecological factor leading to virus degradation (perhaps by introducing some organic substance that is innocuous to bacteria but would inactivate virions). Alternatively, a chemostat may be sufficient. Once a sustainable infection can be established, one can vary parameters such as growth rate, viral decay rate, temperature, and viscosity, thus testing predictions of Eqs. 5–7. Other phage-host systems can be chosen as well. Of particular interest are hosts of significantly different size. Alternatively, the growth medium of *E. coli* can be changed, thus affecting its size. The timescale of this experiment needs to be shorter than the timescale for *E. coli* and/or T4 to start evolving in a way that affects their interaction (see Section

5.4). Since the small oscillations of this system around the fixed point occur with a period of $\tau \sim \left[(\alpha \ - \gamma_{non-viral})\gamma_{virus\ decay}\right]^{-\frac{1}{2}}$ (assuming the latent period=0), the viral decay rate needs to be high enough to prevent large fluctuations from steady state. In addition, a high viral decay rate will prevent the fixed point bacterial concentration from becoming too low, circumventing possible bottle neck affects that can lead to *in vitro* evolution.

4.5.2 Investigating phage-host systems in nature

Our model makes many assumptions regarding viruses and their hosts. For example, we assume that bacteria are in a state of exponential growth, that radii are uniformly distributed, that the virus-host systems are independent and so on. It is therefore crucial to test our model in natural environments. One way to do this is to analyze culturable phage-host systems directly in nature, where hosts are selected to cover a wide spectrum of sizes. Of particular interest are phage-host systems involving giant bacteria. Giant bacteria are predicted by our model to have a very low density (Eq. 7), even when correcting for massive cell inclusions. However, viruses of giant bacteria are predicted by the model to be quite numerous (Eq. 5A), with as many as hundreds of virions per ml of water (see above). By designing primers against phages of giant bacteria and using quantitative assays such as quantitative PCR and/or digital PCR, one can test a direct and extreme prediction of this model, namely that phages of giant bacteria are numerous in nature (with their density predicted by Eq. 5A) and should be detected even in the absence of the host. The absence of the host can be confirmed with SSU rRNA sequencing. If the genome of a lytic phage infecting the giant bacteria cannot be obtained and the host has been sequenced, CRISPR sequences can be crossed with a viral metagenome from the environment of interest to detect phage genes for primer design.

4.5.3 Closed mesocosm experiments

Decay rate perturbation

Closed mesocosm experiments can be used to test total bacterial and viral abundances when perturbing parameters such as viral decay rate, growth rate (through nutrient availability), temperature, and viscosity. These types of experiments can be used to test the predictions of total bacterial concentration and total viral concentration (Eqs. 16 and 19). Note that in closed system experiments, the total number of species $N_{species}$ cannot increase since we cannot create species de novo. As a control, $N_{species}$ can be measured under every perturbation via a SSU rRNA library to check this assumption.

One can also test the ratio between quantities. For example, if the decay rate is changed without affecting bacterial growth (e.g., by introducing some organic chemical that decreases viral lifetime but does not affect bacterial growth or by filtering out UV bands that damage phages, assuming growth rate is not affected) then Eq. 16A makes the simple prediction that

$$\text{(23A)} \qquad \frac{c_{bact}^{tot}\left(\mathsf{UV}\right)}{c_{bact}^{tot}\left(\mathsf{no\ UV}\right)} = \frac{N_{species}\left(\mathsf{UV}\right)}{N_{species}\left(\mathsf{no\ UV}\right)} \frac{E\gamma_{viral\ decay}\left(\mathsf{UV}\right)}{E\gamma_{viral\ decay}\left(\mathsf{no\ UV}\right)}.$$

where for concreteness we designate high decay rate as UV and low decay rate as no UV. If we constrain that $N_{species} = \mathsf{const}$ then we obtain the result that $c_{bact}^{tot}\left(\mathsf{UV}\right) > c_{bact}^{tot}\left(\mathsf{no\ UV}\right)$. Although this result on the one hand makes intuitive sense (viruses that degrading faster lead to more bacteria) it is counterintuitive in the sense that if the environment has the capacity to sustain a higher concentration of bacteria, then why wasn't this capacity utilized by species $i = N_{species} + 1$?

Thus a more logical alternative would be that as the viral decay rate increases, the number of species *decreases*, as some species die, allowing other species to increase in concentration (via Eq. 7) such that $c_{bact}^{tot}\text{=const}$. Thus, increasing the rate of virus degradation leads to a *decrease* in the diversity of the mesocosm by a factor of $E\gamma_{viral\ decay}\left(\text{no UV}\right)/E\gamma_{viral\ decay}\left(\text{UV}\right)$. This solution is pleasing in the sense that there are no undetermined degrees of freedom left. In addition, from Eq. 19 we predict that

$$\text{(23B)}\qquad \frac{c_{virus}^{tot}\left(\text{UV}\right)}{c_{virus}^{tot}\left(\text{no UV}\right)}=\frac{N_{species}\left(\text{UV}\right)}{N_{species}\left(\text{no UV}\right)}.$$

If species die in the mesocosm, then when increasing the decay rate of viruses the total concentration of viruses should *decrease*. The VBR is predicted to decrease when increasing the decay rate of viruses:

$$\frac{VBR(\text{UV})}{VBR(\text{no UV})}=\frac{E\gamma_{viral\ decay}(\text{no UV})}{E\gamma_{viral\ decay}(\text{UV})}<1.$$

Nutrient perturbation

Another critical test of the model would be an enrichment experiment on a nutrient-limited closed mesocosm. Based on Eq. 16A we have

$$\text{(24)}\qquad \frac{c_{bact}^{tot}\left(\text{enriched}\right)}{c_{bact}^{tot}(\text{poor})}=\frac{N_{species}\left(\text{enriched}\right)}{N_{species}(\text{poor})}\frac{r_{\max}\left(\text{poor}\right)}{r_{\max}(\text{enriched})}\left(\frac{r_{\min}\left(\text{poor}\right)}{r_{\min}(\text{enriched})}\right)^{3}.$$

When adding nutrients to our mesocosm we do not expect species to die since there are more resources present in the environment. However, since $N_{species}$ also cannot grow (since there is no available reservoir for species) we conclude that $N_{species} = \mathsf{const}$. Since bacterial size is expected to increase with nutrients, we anticipate that the total concentration of bacteria upon enrichment will *decrease*. The explanation for this paradoxical behavior is apparent from Eq. 3: as nutrients are added and the growth rate of bacteria increases, so does their radius (and thus burst size). Thus the viral production term in Eq. 3 (first term) increases, necessitating the bacterial density to decrease owing to a constant viral decay rate (second term in Eq. 3). We will see that in an open mesocosm experiment exactly the opposite response is anticipated.

Spiking approach

In another approach, a non-indigenous culturable host can be "released" into the mesocosm with its lytic virus allowing one to track host and virus concentrations upon various perturbations. The concentration of the bacterium can be monitored by a quantitative PCR (qPCR) assay targeting the SSU rRNA gene of the organism. The virus concentration can also be monitored via qPCR if there is genetic information on the virus. The advantage of this method is that one can use molecular techniques to precisely gauge the abundance of the host and its virus (instead of measuring pfus or cfus). This approach assumes however that in the time course of the experiment, primer binding sites have not mutated in the evolving viral quasispecies. This assumption can be checked by attempting to amplify plaques with the viral primers and analyzing the success rate statistically.

4.5.4 Open mesocosm experiments

Decay rate perturbation

In open mesocosm experiments the number of species is not constrained as new species can diffuse or swim into our mesocosm and existing species can diffuse or swim out. Repeating the perturbation experiment for the viral decay rate in an open mesocosm system we would predict once more Eq. 23A and 23B and, as before, there is an undetermined degree of freedom. Increasing the viral degradation rate should lead an increase in the concentration of each bacterial species (Eq. 7). However, the total concentration of bacteria should not be allowed to increase upon perturbation, since if the mesocosm could have sustained a higher concentration of bacteria, some new species would have taken advantage of this and stayed in this volume by means of chemotaxis. Thus, we conclude that upon an increase in viral decay rate the number of species will decrease, as some species will die allowing other species to increase in concentration to sustain a constant total concentration of bacteria. Thus, either in an open or closed mesocosm, it appears that increasing the decay rate of viruses should lead to a decrease in species diversity.

Note that when testing predictions of diversity, it is not sufficient to change the viral load, as this will only affect the transient response of the system. In order to observe a steady-state effect one should change the fundamental parameters governing the system, such as the viral decay rate.

Nutrient perturbation

Repeating the enrichment experiment in a nutrient limited open mesocosm Eq. 24 still holds. Here again, the concentration of any given species will decrease due to the increase in radii (Eq. 7), thus there is room for more species. Since new species entering this region can stay in the region by means of chemotaxis, we expect the total number of species to significantly increase and with the total number of bacteria either constant or increasing.

Size spectra perturbation

Our model predicts that the size spectra of bacteria is the result of viral predation and that the slope of the resulting power law should be independent of, for example, nutrient availability, viral decay rate, temperature, medium viscosity, and so on. These predictions can be directly tested in a mesocosm, similar to the IronEx II perturbation experiments. Furthermore, removal of the lytic viral fraction should result in a certain decrease in the slope of the spectrum (more positive), with the new slope being determined presumably by nutrient availably.

Systematic mapping of prokaryotic species diversity in different aquatic zones

One of the interesting predictions of the model deals with species diversity (Eq. 16B). Species diversity changes in a very predictable manner dictated by the total bacterial concentration, viral decay rate, temperature, and so on. By systemically sampling different environments on Earth (e.g., eutrophic versus oligotrophic zones, photic versus the aphotic zones, epipelagic zones in tropical versus polar regions, marine versus freshwater ecosystems, etc.) and measuring the concentration of bacteria, the temperature, the viral decay rate, and the number of species (via SSU rRNA libraries) one can directly test the predicted number of species (Eq. 16B).

4.5.5 Investigate host range in nature

To test our assumption that a host in a given region is infected with a single viral species one can isolate different phages infecting the same host species using conventional plaque assays. Phages that appear to be morphologically different via EM can be sequenced and their genomes compared. To test our assumption that phages have a species or subspecies host range, one can perform host-independent co-localization experiments (via, for example, digital PCR – see Chapter 2) using as a viral marker a gene of a lytic virus from the environment.

4.6 Relation between number of bacterial species and number of viral species

We would like to show that a system of *n* bacterial *species* must be associated with exactly *n* viral *species* or else the system will be overdetermined, driving excess *species* into extinction. The proof is the following: Let's assume there are *n* bacterial species infected by *n* viral species. There are therefore 2*n* rate equations, *n* for *n* bacteria and *n* for *n* viruses:

$$\text{(A1)} \qquad \begin{cases} \dfrac{dB_1}{dt} = \alpha_1 B_1 - k_{11} B_1 V_1 - \ldots - k_{1n} B_1 V_n \\ \vdots \\ \dfrac{dB_n}{dt} = \alpha_n B_n - k_{n1} B_n V_1 - \ldots - k_{nn} B_n V_n \end{cases} \qquad \begin{cases} \dfrac{dV_1}{dt} = -\gamma_1 V_1 + b_1 k_{11} B_1 V_1 + \ldots + b_1 k_{n1} B_n V_1 \\ \vdots \\ \dfrac{dV_n}{dt} = -\gamma_n V_n + b_n k_{1n} B_1 V_n + \ldots + b_n k_{nn} B_n V_n \end{cases}$$

where we have allowed the most general interaction network between the viruses and the bacteria. At steady-state we obtain the following 2*n* linear relations:

$$\text{(A2)} \qquad \text{bacterial rate equations} \rightarrow \begin{cases} \alpha_1 - k_{11} V_1 - \ldots - k_{1n} V_n = 0 \\ \vdots \\ \alpha_n - k_{n1} V_1 - \ldots - k_{nn} V_n = 0 \end{cases} .$$

$$\text{(A3)} \qquad \text{viral rate equations} \rightarrow \begin{cases} -\gamma_1 + b_1 k_{11} B_{11} + \ldots + b_1 k_{n1} B_{n1} = 0 \\ \vdots \\ -\gamma_{nn} + b_n k_{1n} B_{1n} + \ldots + b_n k_{nn} B_{nn} = 0 \end{cases} .$$

Now let's assume we introduce bacterium n+1. If we write the rate equation for this bacterium, then at steady-state we will obtain the n+1 equation for (A2), however there are only n variables V_i i=1..n. The system is therefore overdetermined and therefore some species will become extinct in the transient solution. The same rational applies if we add the n+1 viral species. In this case we will have n+1 steady-state equations for the viruses (A3), yet we have only n variables for B_i i=1..n, again obtaining an overdetermined set of equations. If we remove one bacterial species or one viral species, we again find the same situation: the reciprocal variable will be overdetermined. Thus, the only solution which is not overdetermined is if we have n bacterial species being infected by n viral species.

For the special case of one virus species with a wide host range infecting two bacterial species the proof is the following: Let's imagine we have a closed system containing two different *distinguishable* hosts of concentration $c_{bacteria}^{(1)}$ and $c_{bacteria}^{(2)}$, both infected with the same virus of concentration c_{virus}. According to Eq. A2 and Eq. A3 the set of differential equations governing the interaction of these three species is

$$\begin{cases} \dfrac{dc_{virus}}{dt} \cong b^{(1)}k^{(1)}c^{(1)}_{bacteria}c_{virus} + b^{(2)}k^{(2)}c^{(2)}_{bacteria}c_{virus} - \gamma_{virus\ decay}c_{virus} \\ \dfrac{dc^{(1)}_{bacteria}}{dt} = \tilde{\alpha}^{(1)}c^{(1)}_{bacteria} - k^{(1)}c_{virus}c^{(1)}_{bacteria} \\ \dfrac{dc^{(2)}_{bacteria}}{dt} = \tilde{\alpha}^{(2)}c^{(2)}_{bacteria} - k^{(2)}c_{virus}c^{(2)}_{bacteria} \end{cases} \tag{A4}$$

where $k^{(i)} = 4\pi D_{virus}R^{(i)}_{bact}$. At steady-state, this system is, however, overdetermined since the solutions $c_{virus} = \tilde{\alpha}^{(1)}/k^{(1)}$ and $c_{virus} = \tilde{\alpha}^{(2)}/k^{(2)}$ cannot be mutually satisfied. The only consistent steady-state solutions would be $c^{(1)}_{bacteria} \equiv 0$ or $c^{(2)}_{bacteria} \equiv 0$ or $c^{(1)}_{bacteria} = c^{(2)}_{bacteria} \equiv 0$, unless the two hosts have precisely the same radius and growth rate. Thus, only bacteria with the same radius and same growth rate can be infected with the same virus and sustain a population. The slightest difference and, with enough time, one species will be driven to extinction.

Similarly, if we have two viral species with a specific host range, infecting the same bacteria, we would again run into an overdetermined system of equations:

$$\begin{cases} \dfrac{dc^{(1)}_{virus}}{dt} \cong b^{(1)}k^{(1)}c_{\ bacteria}c^{(1)}_{virus} - \gamma^{(1)}_{virus\ decay}c^{(1)}_{virus} \\ \dfrac{dc^{(2)}_{virus}}{dt} \cong b^{(2)}k^{(2)}c_{\ bacteria}c^{(2)}_{virus} - \gamma^{(2)}_{virus\ decay}c^{(2)}_{virus} \\ \dfrac{dc_{\ bacteria}}{dt} \cong \alpha_{\ bacteria}c^{(1)}_{virus} - k^{(1)}c_{\ bacteria}c^{(1)}_{virus} - k^{(2)}c_{\ bacteria}c^{(2)}_{virus} \end{cases} . \tag{A5}$$

Thus at steady-state we would find that from the first equation $c_{\,bacteria} = \frac{\gamma^{(1)}_{virus\ decay}}{b^{(1)}k^{(1)}}$ while from the second equation $c_{\,bacteria} = \frac{\gamma^{(2)}_{virus\ decay}}{b^{(2)}k^{(2)}}$, thus the system is overdetermined. It is intuitively clear that two viruses cannot control the same species.

4.7 Power law derivation

4.7.1 The distribution of bacteria in the environment

According to Eq. 9, the concentration of bacteria of a given radius r per radius in a given realization of an environment containing $N_{species}$ bacterial species is given:

(B1)
$$\rho_{environment}(r) = c_{bacteria}(r) \sum_{i=1}^{N_{species}} \delta(r - R_{bacteria}^{(i)}).$$

where $R_{bacteria}^{(i)}$ are $N_{species}$ i.i.d. random variables drawn from a distribution $f_R(r)$ and where $\delta(r)$ is the Dirac delta function. To obtain the ensemble average of $\rho_{environment}(r)$, averaging over many realizations of a given environment one should calculate the expectation value of $\rho_{environment}(r)$ with respect to the N random variables $R_{bacteria}^{(i)}$:

$$\langle \rho_{environment}(r) \rangle = c_{bacteria}(r) \sum_{i=1}^{N_{species}} \int_{R^{(i)}} dR_{bacteria}^{(1)} ... dR_{bacteria}^{(N)} f_{R^{(1)},...,R^{(N_{species})}} \left(R_{bacteria}^{(1)}, ..., R_{bacteria}^{(N_{species})} \right) \delta(r - R_{bacteria}^{(i)}).$$

Since $R_{bacteria}^{(i)}$ are i.i.d. we have

$$\langle \rho_{environment}(r) \rangle == c_{bacteria}(r) \sum_{i=1}^{N_{species}} \int_{R_{bacteria}^{(i)}} dR_{bacteria}^{(i)} \left[f_{R^{(1)}} \left(R_{bacteria}^{(1)} \right) \cdot ... \cdot f_{R^{(N_{species})}} \left(R_{bacteria}^{(N_{species})} \right) \right] \delta(r - R_{bacteria}^{(i)}) =$$
$$= c_{bacteria}(r) \sum_{i=1}^{N_{species}} \int_{R_{bacteria}^{(i)}} dR_{bacteria}^{(i)} f_R \left(R_{bacteria}^{(i)} \right) \delta(r - R_{bacteria}^{(i)}) =$$
$$= c_{bacteria}(r) \sum_{i=1}^{N_{species}} f_{R^{(i)}}(r) = N \cdot c_{bacteria}(r) \cdot f_R(r).$$

Thus the average distribution of bacterium sizes in a given environment is given by Eq. B2:

(B2)
$$\left\langle \rho_{environment}(r) \right\rangle = N_{species} \cdot c_{bacteria}(r) \cdot f_R\ (r).$$

To test this equation we performed the following Monte Carlo simulation: We draw $N_{species}$=100 radii R_i (i=1.. 100) for bacteria according to a specified probability density function (pdf) $f_R(r)$. The concentration of each bacterial species as a function of its radius is given by the hypothetical distribution $c_{bacterium}(r) = r.$ We then construct an empirical discrete distribution function for $\rho_{environment}(r)$ such that $\rho_{environment}(r = r_i) = c_{bacterium}(r_i) = r_i$ for i=1..100. Finally we average this distribution over many Monte Carlo simulations (M=10000), simulating many realizations of this environment to obtain $\left\langle \rho_{environment}(r) \right\rangle$. The ensemble average that we compute, $\left\langle \rho_{environment}(r) \right\rangle$, should converge according to Eq. B2 to $\left\langle \rho_{environment}(r) \right\rangle = N_{species} \cdot r \cdot f_R\ (r).$. Examples of two pdfs for $f_R(r)$ are shown in Fig. 4.4.

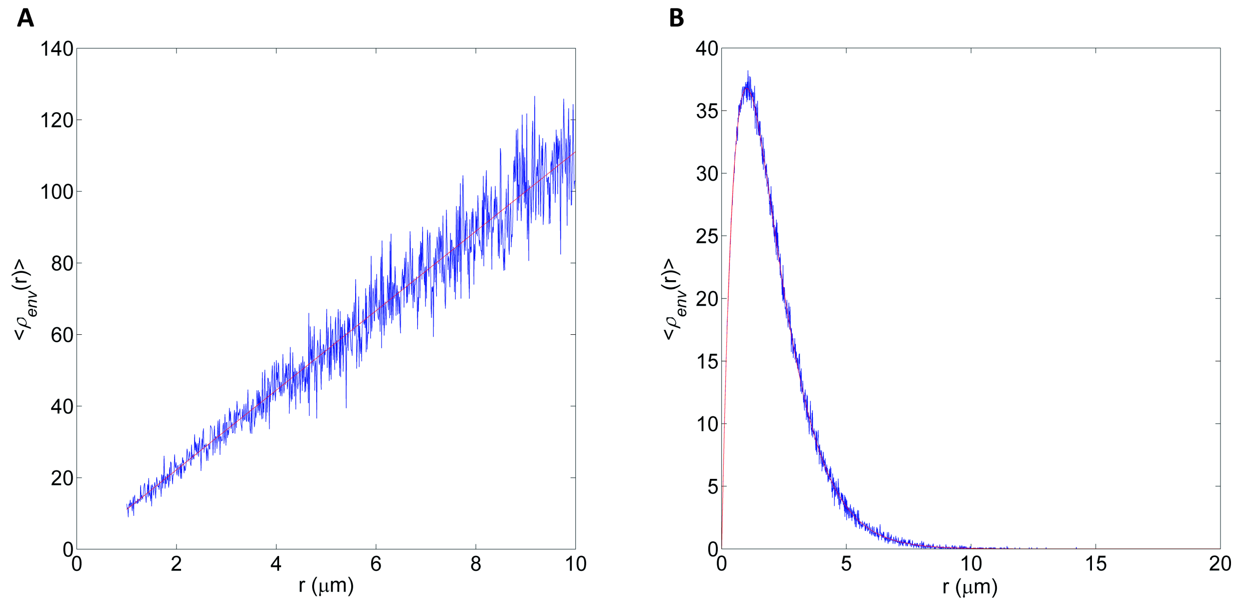


Figure 4.4. Monte Carlo simulation of a hypothetical distribution of bacteria in a given environment. In each of M=10^4 Monte Carlo iterations, $N_{species}$=100 bacterial radii were drawn such that in **(A)** R was exponentially distributed with rate λ=1 and in **(B)** R was uniformly distributed between r_{min}=1 and r_{max}=10. The empirical distribution of bacteria $\rho_{environment}(r)$ in both cases was calculated assuming the hypothetical relation $c_{bacterium}(r) = r$. That is, for each radius R_i drawn in a given iteration we update the empirical distribution function in the following way: $\rho_{environment}(r = r_i) = c_{bacterium}(r_i) = r_i$ (see Eq. B1). Then finally we average M=10^4 calculated empirical distribution functions $\rho_{environment}(r)$ to obtain the ensemble average of $\rho_{environment}(r)$, which we denote by $\langle \rho_{environment}(r) \rangle$. Based on Eq. B2 we expect that for (A) $\langle \rho_{environment}(r) \rangle = \lambda \cdot N_{species} \cdot r \cdot e^{-\lambda r}$ and for (B) $\langle \rho_{environment}(r) \rangle = N_{species} \cdot r / (r_{\max} - r_{\min})$. The figure demonstrates that in both cases the calculated value for $\langle \rho_{environment}(r) \rangle$ based on the Monte Carlo simulation (blue) converged precisely to the theoretical prediction (red) describe above.

4.7.2 The predicted size spectra of bacteria in the environment

In the main text we derived the probability that a bacterium of random volume V, is greater than or equal to a given volume, v (Eq. 14). Here we test Eq. 14 in the following Monte Carlo simulation: We assumed that $c_{bacterium}(r) = r^{-4}$, $R \sim U(r_{\min}, r_{\max})$ and we computed $\langle \rho_{environment}(r) \rangle$ as explained above (see Fig. 4.4). We then normalized the computed function $\langle \rho_{environment}(r) \rangle$ to obtain the empirical pdf $f_\rho(r)$ and calculated Prob(V≥v). The Monte Carlo simulations should converge to $\left\langle \rho_{environment}(r) \right\rangle = N_{species} \cdot r^{-4} \Big/ \left(r_{\max} - r_{\min} \right)$ (following Eq. B2) and $\text{Prob}(V \geq v)$ should converge to $\text{Prob}(V \geq v) = \left(\dfrac{r_{\min}}{r_{\max}} \right)^3 \left[\left(r_{\max} / r \right)^3 - 1 \right]$ (Eq. 14). Results are shown in Fig. 4.5 and demonstrate that the simulation converged precisely to the theoretical predictions.

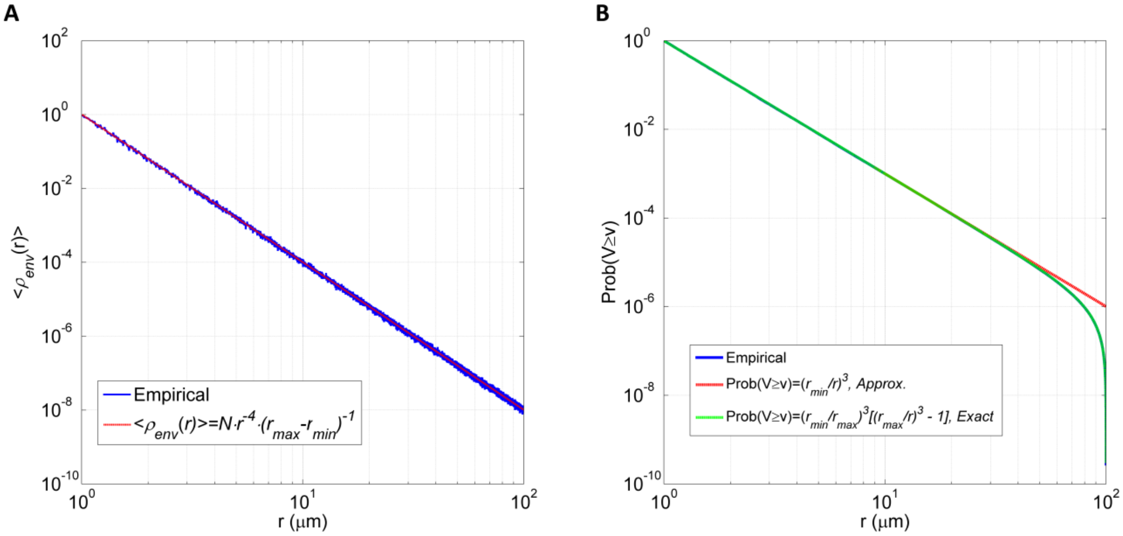


Figure 4.5. Monte Carlo simulation of the predicted size spectra of bacteria in a given environment. Monte Carlo simulation assuming the predicted concentration of a bacterium with radius r obeys $c_{bacterium}(r) = r^{-4}$ and that bacteria radii are drawn from the uniform distribution. **(A)** Theoretical prediction (red) for the ensemble average of the distribution of bacteria in the given environment $\left\langle \rho_{environment}(r) \right\rangle = N_{species} \cdot r^{-4} \Big/ \left(r_{\max} - r_{\min} \right)$ (Eq. B2) versus Monte Carlo simulation (blue) with M=10^4 iterations (see caption of Fig. 4.4 for simulation details). **(B)** The numerical estimate of $\left\langle \rho_{environment}(r) \right\rangle$ was normalized to obtain an empirical pdf, which was used to calculate $\text{Prob}(V \geq v)$. The result of the Monte Carlo simulation (blue) was compared with the theoretical prediction for $\text{Prob}(V \geq v)$ (Eq. B1; red in A, green in B). The figure demonstrates that the numerical simulation converged precisely to the theoretical prediction for both (A) and (B).

4.8 References

1. Suttle C (2007) Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5: 801-812.
2. Weinbauer M (2004) Ecology of prokaryotic viruses. FEMS Microbiology Reviews 28: 127-181.
3. Noble R, Fuhrman J (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquatic Microbial Ecology 14: 113-118.
4. Fuhrman J (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541-548.
5. Campbell A (1961) Conditions for the existence of bacteriophage. Evolution 15: 153-165.
6. Levin B, Stewart F, Chao L (1977) Resource-limited growth, competition, and predation: a model and experimental studies with bacteria and bacteriophage. American Naturalist 111: 3-24.
7. Lenski RE (1988) Dynamics of interactions between bacteria and virulent bacteriophage. Advances in microbial ecology 10: 1-44.
8. Beretta E, Kuang Y (1998) Modeling and analysis of a marine bacteriophage infection. Mathematical Biosciences 149: 57-76.
9. Thingstad T (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography: 1320-1328.
10. Thingstad T, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquatic Microbial Ecology 13: 19-27.
11. Stent GS (1963) Molecular Biology of Bacterial Viruses. San Francisco: Freeman.
12. Murray A, Jackson G (1992) Viral dynamics: a model of the effects of size, shape, motion and abundance of single-celled planktonic organisms and other particles. Marine ecology progress series Oldendorf 89: 103-116.
13. Suttle C (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. Viral Ecology: Academic Press. pp. 247–296.
14. Kutter E, Sulakvelidze A (2005) Bacteriophages: biology and applications: CRC Press.
15. Suttle C (2005) Viruses in the sea. Nature 437: 356-361.
16. Paul J, Kellogg C (2000) Ecology of bacteriophages in nature. Viral Ecology: 211–246.
17. Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. Nature Reviews Microbiology 3: 537-546.
18. Fuhrman J, Noble R (1995) Viruses and protists cause similar bacterial mortality in coastal seawater. Limnology and Oceanography 40: 1236-1242.
19. Wommack K, Colwell R (2000) Virioplankton: viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews 64: 69.
20. Wilcox R, Fuhrman J (1994) Bacterial viruses in coastal seawater: lytic rather than lysogenic production. Marine Ecology-Progress Series 114: 35-35.
21. Suttle CA, Chen F (1992) Mechanisms and rates of decay of marine viruses in seawater. Applied and Environmental Microbiology 58: 3721.
22. Howard-Jones M, Ballard V, Allen A, Frischer M, Verity P (2002) Distribution of bacterial biomass and activity in the marginal ice zone of the central Barents Sea during summer. Journal of marine systems 38: 77-91.

23. Suttle CA, Chan AM (1994) Dynamics and distribution of cyanophages and their effect on marine Synechococcus spp. Applied and Environmental Microbiology 60: 3167.

24. Seymour J, Seuront L, Doubell M, Waters R, Mitchell JG (2006) Microscale patchiness of virioplankton. Journal of the Marine Biological Association of the UK 86: 551-561.

25. Bratbak G, Egge J, Heldal M (1993) Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms. Marine Ecology Progress Series.

26. Berg H, Purcell E (1977) Physics of chemoreception. Biophysical journal 20: 193-219.

27. Schwartz M (1976) The adsorption of coliphage lambda to its host: Effect of variations in the surface density of receptor and in phage-receptor affinity* 1. Journal of molecular biology 103: 521-536.

28. Weinbauer M, Peduzzi P (1994) Frequency, size and distribution of bacteriophages in different marine bacterial morphotypes. Marine Ecology Progress Series 108: 11-20.

29. Weinbauer M, Hoefle M (1998) Size-specific mortality of lake bacterioplankton by natural virus communities. Aquatic Microbial Ecology 15: 103-113.

30. Castberg T, Thyrhaug R, Larsen A, Sandaa RA, Heldal M, et al. (2002) Isolation and characterization of a virus that infects< i> Emiliania huxleyi</i>(Haptophyta).

31. Lawrence JE, Chan AM, Suttle CA (2001) A novel virus (HaNIV) causes lysis of the toxic bloom-forming alga Heterosigma akashiwo (Raphidophyceae). Journal of Phycology 37: 216-222.

32. Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G (2001) Isolation and characterization of two viruses with large genome size infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae). Virology 290: 272-280.

33. Schultz H, Jorgensen B (2001) Big bacteria. Annual Review of Microbiology 55: 105-137.

34. Madigan MT, Martinko JM (2006) Brock biology of microorganisms: Upper Saddle River, NJ, USA: Pearson Prentice Hall.

35. Ackermann H (2006) Classification of bacteriophages. The bacteriophages 2: 8-17.

36. Ackermann H (1999) Tailed bacteriophages: the order Caudovirales. Advances in virus research 51: 135-202.

37. Schulz H, Brinkhoff T, Ferdelman T, Mariné MH, Teske A, et al. (1999) Dense populations of a giant sulfur bacterium in Namibian shelf sediments. Science 284: 493.

38. Schultz H, Jorgensen B (2001) Big bacteria. Annu Rev Microbiol 55: 105-137.

39. Dubin S, Benedek G (1970) Molecular weights of coliphages and coliphage DNA: II. Measurement of diffusion coefficients using optical mixing spectroscopy, and measurement of sedimentation coefficients. Journal of molecular biology 54: 547-556.

40. Suttle CA, Chan AM (1993) Marine cyanophages infecting oceanic and coastal strains of Synechococcus: abundance, morphology, cross-infectivity and growth characteristics. Marine Ecology-Progress Series 92: 99-99.

41. Cavender-Bares K, Rinaldo A, Chisholm S (2001) Microbial size spectra from natural and nutrient enriched ecosystems. Limnology and Oceanography 46: 778-789.

42. Sheldon R, Prakash A, Sutcliffe Jr W (1972) The size distribution of particles in the ocean. Limnology and Oceanography 17: 327-340.

43. Maniloff J (1997) Nannobacteria: size limits and evidence. Science 276: 1773.

44. Whitman W, Coleman D, Wiebe W (1998) Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences 95: 6578.

45. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66.
46. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99: 14250.
47. Mei M, Danovaro R (2004) Virus production and life strategies in aquatic sediments. Limnology and Oceanography 49: 459-470.
48. Danovaro R, Corinaldesi C, Luna GM, Magagnini M, Manini E, et al. (2009) Prokaryote diversity and viral production in deep-sea sediments and seamounts. Deep Sea Research Part II: Topical Studies in Oceanography 56: 738-747.
49. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35: 7188.
50. Clasen J, Brigden S, Payet J, Suttle C (2008) Evidence that viral abundance across oceans and lakes is driven by different biological factors. Freshwater Biology 53: 1090-1100.

Chapter 5

An evolutionary model of phage-host interaction

5.1 Introduction

In Section 4.4.2.6 we predicted the total number of “species” in a given environment. It is therefore of interest to define what a “species” is. In the biophysical model a “species” of bacterium or virus is defined by a set of random variables drawn from some distribution. However, as explained previously, two “species” with the same parameters can be totally different organisms and should be counted separately. Therefore the biophysical model described in Section 4.4 cannot provide us with an adequate definition of a “species” that would be useful for testing the predictions of our model. The problem lies in the fact that at that level of abstraction, bacteria and viruses are the equivalent of “point particles” without internal structure. In the present section we will attempt to go one step further and define an evolutionary model, which when viewed at a coarse-grained level, would be equivalent to the description of the biophysical model in Section 4.4.

In order to understand what a species is in the context of our biophysical model, we propose definitions for both bacterial and viral species that ensure that the assumptions of the biophysical model are respected. These assumptions include: (1) each bacterial species was associated with a single viral species and vice versa (i.e., there is no cross interaction between phage-host systems), and (2) each species (bacterial or viral) was unique and distinguishable from all other species. Based on these definitions, we will construct an evolutionary model for the emergence of new bacterial and viral "species" in nature. While this model is equivalent to our biophysical model when viewed in a genetic coarse-grained way, the evolutionary model leads to the prediction that bacterial "strains" are part of interaction networks with viral "strains", whereas bacterial "species" form a unique association with a single viral "species" and vice versa. Furthermore, in order for new bacterial and viral species to emerge as independent elements, the emerging viral species needs to abandon the parental bacterial strain that it previously controlled in favor of the new emerging species. We propose that the "arms race" between bacteria and viruses may lead to a "positive feedback" mode of evolution, that both enables the emerging viral species to switch hosts, and enables the emerging bacterial strain to evolve at an accelerated pace through selection sweeps to form a new species. Thus, the arms race that bacteria and viruses are locked in is perhaps the engine driving bacterial and viral co-speciation, with selection pressure arising from the environment biasing the direction of evolution. In addition we show that for the simple case of a "butterfly" 2x2 strain interaction network the total concentration of the parental and emerging strains doubles when speciation is complete. We then generalize this result the case of $N_{strains}$ 2x2 interaction networks, with $N_{strains}$ defined as the number of strains per species. Finally we conclude by suggesting an experiment to test our hypothesis regarding "positive feedback evolution".

Summary of findings: Our biophysical model is consistent with an evolutionary model where (1) a bacterial "species" is comprised of bacterial "strains" and where a viral "species" is comprised of viral "strains" (with a "strain" = a quasispecies). (2) New bacterial "species" co-emerge with new viral "species" and vice versa. (2) A bacterial "species" interacts with just one viral "species", however a bacterial "strain" generally interacts with many viral "strains" as it is part of a network of bacterial-viral interactions. (3) The host range of viruses should be (mostly) species or strain specific. (4) The evolutionary arms-race between phages and hosts (such as the CRISPR warfare) may be a critical part of the stage where bacterial and viral "species" co-emerge out of their parental "strains" by (a) accelerating the evolution of the bacterial strain through selective sweeps and at the same time (b) accelerating the evolution of the virus to switch hosts.

The road map: We will begin by describing the critical features of the biophysical model described in Section 4.4, which our evolutionary model must reproduce when viewed at a coarse-grained level. We will then define the concept of a "strain" and a "species" in such a way that when placed in an evolutionary context produces a "bacterial and phage world" that when coarse-grained is equivalent to the description of our biophysical model. Thus we will use the biophysical model to guide us in selecting a good evolutionary model.

5.2 Definition of a bacterial and viral *strain* and *species*

Critical features of the biophysical phage-host model

The following are the critical features of the biophysical phage-host model described in Section 4.4:

1. **Growth:** All bacteria and viruses are actively replicating in the environment.
2. **Viral control:** Each bacterium is associated with a lytic virus that controls its concentration.
3. **Uniqueness:** Each phage-host system is comprised of a bacterium and a virus that can be distinguished (in some measurable way) from all other bacteria and viruses in the environment. We therefore say that each bacterium belongs to a unique "species" denoted by the index *i*, and each virus belongs to a unique "species", denoted by the same index.

4. **Symmetry:** There are equal numbers of bacterial and viral "species" (both denoted with the index *i*) .
5. **Independence:** All phage-host systems in a given environment are independent of each other, i.e., there is no cross interaction between one system and another.

An evolutionary model that satisfies these five conditions will be consistent with our biophysical model. We can then use the evolutionary definitions of a bacterial species and a viral species to interpret the meaning of the species in our biophysical model.

Bacteria "take up" concentration: Bacterial "species" in the biophysical model have one additional consequence. A bacterial "species" has the property that it "takes up" concentration in the environment, with the concentration being given by Eq. 7. The reason we say it "takes up" concentration is that any environment has finite resources that can accommodate a finite concentration of cellular organisms (this is how we obtained the number of *species* in the environment, $N_{species}$). Thus, only elements that "take up" concentration contribute to the diversity of the system. A viral "species" also has a concentration, however viruses are not limited by resources and therefore there is no upper bound on the number of viral "species" in a given environment. Therefore viral "species" do not "take up" concentration. Drawing again on an analogy to physics, in this respect, bacteria are like fermions and viruses are like bosons — one can pack an infinite number of bosons into a negligibly small volume, whereas fermions take up volume due to their quantum charges. This is why $N_{species}$ was obtained from c_{bact}^{tot} and not c_{virus}^{tot}, the former has an upper bound whereas the latter does not.

Definition of a bacterial and viral *strains*

We seek to define a bacterial *species*[1] and a viral *species* in such a way that, when placed in an evolutionary context, we are able to reproduce the essential characteristic of the biophysical model described above. To define a *species* we first need to define an auxiliary term, which is a *strain*.

> **Definition of a *strain*:** A genetic element (bacterium or virus) is considered a new *strain* if and only if this genetic element is *distinguishable* from all other *strains* (the first cell is by default a *strain*). To be *distinguishable*, a genetic element needs to have a measurable property that sets that element apart from all other existing *strains*. This measurable property should give consistent results over time despite the mutation load of the genetic element.

This definition of a *strain* is consistent with the biologically intuitive definition of a "strain". Here we have also defined a viral *strain*. A viral *strain* can also be interpreted as a viral quasispecies [1] since each genome in the quasispecies is not *distinguishable* from other elements comprising the quasispecies.

Definition of a bacterial *species*

> **A bacterial *species:*** A bacterial cell constitutes a new *species* if and only if (1) it is actively replicating in the environment; (2) It can be classified as a new *strain* in the environment; (3) It forms a stable association with a virus that can be classified as a new *strain* in the environment.

Criterion 1 is necessary in order to distinguish actively replicating cells that have a finite growth rate from spore cells or inactive (possibly dead) cells [2]. The latter, although possibly alive, cannot be part of a phage-host system since they are not actively growing. Criterion 2 simply

[1] We use italics to distinguish the terms defined in the current model from the colloquial use of these words or from the terms used in the biophysical model.

ensures that the new bacterial *strain* is *distinguishable* from other pre-existing bacterial *strains* in the environment, and thus should receive a new index. Criteria 1+2 define an active strain in the intuitive sense, not a species in the intuitive sense. Why is it that to complete the definition of a bacterial *species* one must talk about its viruses (criterion 3)? The reason is the following: If an environment contains *n* bacterial *strains* with *n* infecting viral *strains*, adding a new bacterial *strain* (*strain* # *n*+1) without a new viral *strain* will lead to an overdetermined system of equations in which one or more bacterial *strains* will become extinct (Section 4.6). Thus, to add a new bacterial *species* one must also introduce a new viral *strain* into the system. This rule can be stated in a more general way (Section 4.6):

> A system of *n* bacterial *species* must be associated with exactly *n* viral *species* otherwise the system will be overdetermined, driving excess *species* to extinction.

This definition of a bacterial "species" satisfies the properties of: **bacterial growth** (the bacterial *species* must be growing); **viral control** (each bacterial *species* is associated with a virus); and **bacterial uniqueness** (each bacterial *species* is a new *strain*).

Definition of a viral species

The definition of a viral species is analogous to the definition of a bacterial species:

> **Definition of a viral *species:*** A virus constitutes a new *species* if and only if (1) it is actively replicating in the environment; (2) It can be classified as a new *strain* in the environment; (3) It forms a stable association with a host that can be classified as a new *strain* in the environment.

Criterion 1 is to ensure that we are considering a virus that is active and not a decayed or an inactivated virus. Criterion 2 ensures that the new viral *strain* is *distinguishable* from other pre-existing viral *strains* in the environment, and thus should receive a new index. Criterion 3 is, as before, required because the system should always have equal number of bacterial and viral *species* otherwise excess *species* will be driven to extinction (Section 4.6).

This definition of a viral "species" satisfies the properties of: **viral growth** (the viral *species* must be replicating); **viral uniqueness** (each viral *species* is a new *strain*); and **symmetry** (if each bacterial *species* is associated with a viral *species* and each viral *species* is associated with a bacterial *species*, there should be equal number of bacterial and viral *species*). The only property that has yet to be satisfied is independence. By constructing an evolutionary model that satisfies this property we will be able to understand the relation between *species* and *strains*.

Note that the definitions of a bacterial and viral *species* suggest that the formation of a new bacterial *species* is linked to the formation of a new viral *species* and vice versa. In the next section we will explain an evolutionary mechanism for this process.

5.3 A model for bacterial-viral co-speciation

5.3.1 Description of the evolutionary model

Stage 1: One bacterial *strain*, one viral *strain* (Fig. 5.1A). Let's assume our environment contains a bacterial *species* (species 1) comprised of a single *strain* (strain 1), and that this bacterial *species* is under the control of a viral *species* (species A), comprised of a single viral *strain* (strain A) (Fig. 5.1A). The concentration of bacterial strain 1 is dictated by Eq. 7, thus viral species A controls bacterial species 1 (the arrow in Fig. 5.1A). Bacterial strain 1 is said to "take up" concentration in the environment.

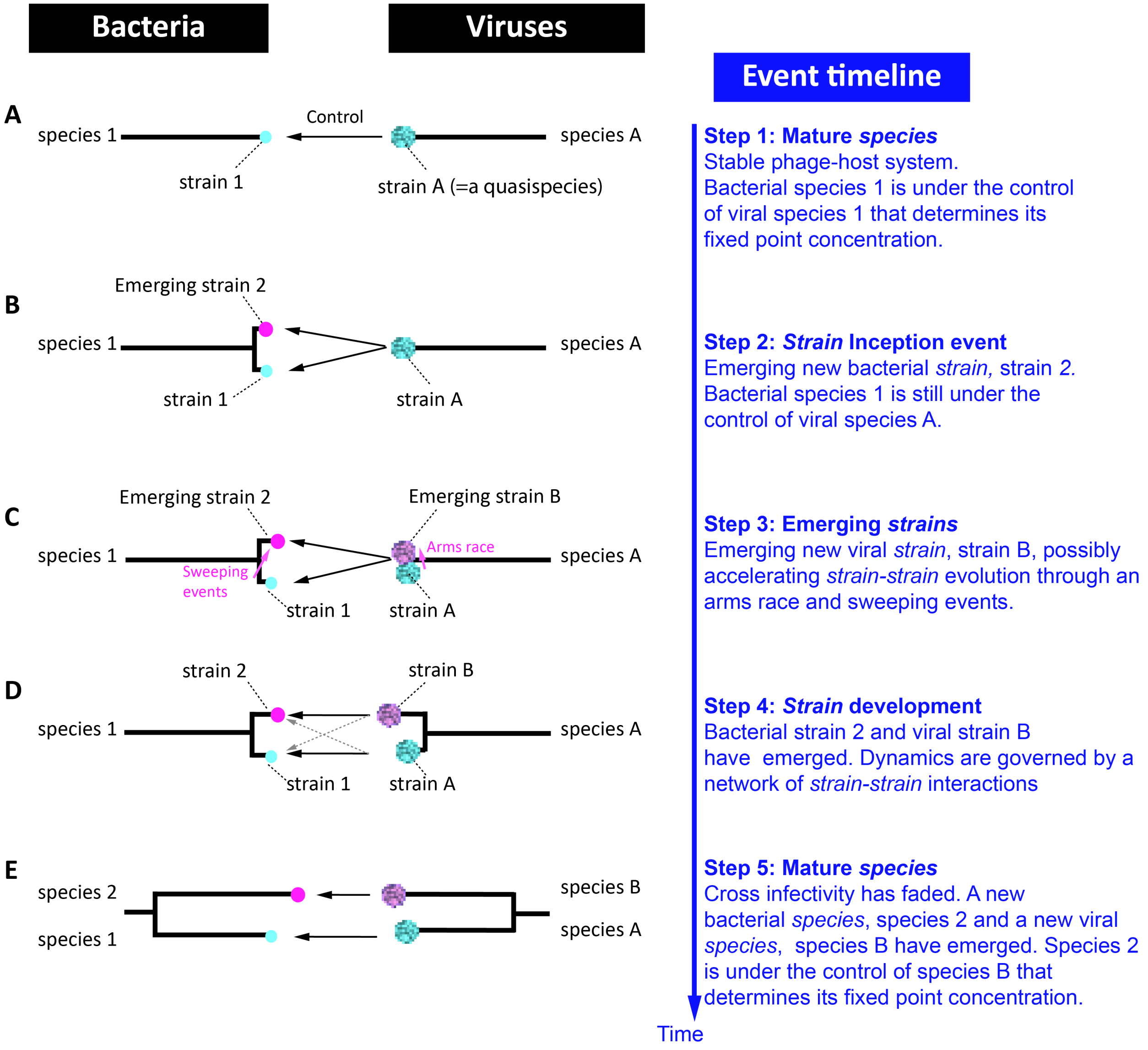
Stage 2: An incipient bacterial *strain* emerges (Fig. 5.1B). Now let's assume that through some genetic event (e.g., a transposon, a deletion/insertion/inversion event, a recombination event, a new plasmid, etc.), bacterial strain 1 begins to evolve a new bacterial *strain* that is on the verge of becoming *distinguishable* from strain 1 (Fig. 5.1B). The incipient bacterial strain 2 is under the growth control of viral strain A, and will not be "allowed" to take up concentration on its own, independent of bacterial strain 1 — i.e., it will not be allotted a status of a *species* and therefore will not contribute to the diversity of the system (i.e., increase $N_{species}$). Bacterial strain 2 will continue to undergo evolution with time and accumulate more mutations in its process of maturing into a new *strain*. During all this time bacterial strain 1 is under the control of viral s*train* A (Fig. 5.1B).

Stage 3: An incipient viral *strain* emerges (Fig. 5.1C). As the incipient bacterial strain 2 evolves, so does the viral *strain* that infects it (initially viral *strain* A). This viral *strain* (i.e., viral quasispecies) will begin to form a new cluster that will eventually mature into viral strain B. The incipient viral strain B (not yet *distinguishable* from viral strain A) both tracks the evolution of bacterial strain 2 and also drives the evolution of bacterial strain 2. This hypothesis is supported by the following observations. It has been suggested that viruses and bacteria are in a constant state of an "arms race" [3]. Perhaps the best example of this arms race is the CRISPR bacterial defense system. Bacteria continuously acquire CRISPR spacer sequences from viruses to evade these viruses, while viruses rapidly evolve by mutation, homologous recombination, and deletion of the target sequences to evade new acquired spacers [4]. Conversely, CRISPR repeats and their associated proteins undergo evolution to escape shut-down mechanism for the CRISPR system encoded by the phage [3]. There is also evidence that the bacterial population undergoes sweeping selection events, where potentially only one cell survives (the only cell that had the right spacer) [4]. Such bottlenecks will accelerate the evolution of the emerging bacterial *strain*, driving its evolution forward. This example illustrates how by a process of positive feedback between the new bacterial *strain* (strain 2) and the new viral *strain* (strain B) both elements track each other and push each other to further evolve (Fig. 5.2). **The bacterial-viral "arms race" may therefore be a critical step in forming (or at least accelerating) the formation of new bacterial *species* and new viral *species* from the parental *strains*.** Indeed, CRISPR sequences have been found in nearly half of all sequenced bacterial genomes [3]. While the CRISPR mechanism may contribute to the arms race, it may not be an essential component. Luria and Delbrück have shown that a bacterial strain grown from a single cell will mutates naturally (without interaction with the phage) so that a subpopulation of bacteria will become immune to the virus [5]. Thus, even without a CRISPR system the bacterium can evade the virus. Therefore, this "arms race" may be a fundamental mechanism of evolution to generate new bacterial and viral *species*. Given our interpretation, these events are not a disadvantage in terms of reduction in diversity, as previously proposed [4], since they may provide the mechanism for new *strains* to emerge. Thus ultimately these mechanisms generate diversity.

Stage 4: New bacterial and viral *strains* emerge (Fig. 5.1D). The incipient bacterial strain 2 is now *distinguishable* from strain 1 and can be defined as a new *strain*. The incipient viral strain B emerged as a new viral *strain* (strain B) that initially infects both bacterial strains 1 and 2 (Fig. 5.1C). At this stage, a 2x2 network like interaction emerges. This network can, in principle, persist indefinitely, and as the evolutionary distance between *strain* 1 and *strain* 2 grows, this could lead to the formation of viruses with a wide host range. If the system is stable over time, the two bacterial *strains* would be under control of two viral *strains* and both would "take up" concentration, thus increasing the diversity of the system ($N_{species}$ in Eq. 16B increases). However, it seems more plausible that as the distance between *strain* 1 and *strain* 2 grows, the cross affinity (B→1 and A→2) will decrease , leading to the emergence of two independent associations (A→1 and B→2). This is because the bacterial strain 2 is driving the evolution of viral strain 2, and it is expected that at some point this virus will lose its ability to infect the parental bacterial strain 1 (Fig. 5.2). Furthermore, it would seem that a 2x2 network of interacting *strains* would not be stable in an open environment for long, since if one of the viral *strains* drifts off, leaving bacterial *strains* 1+2 under the control of the remaining viral *strain*, one or both bacterial *strains* will be driven into extinction over time (Section 4.6). Numerical simulations would be required to see if a network of 2x2 interacting *strains* (*n*x*n* in the more general case) are indeed less stable than two 1x1 associations (or generally *n* 1x1 associations) under loss of a viral *strain*. The fact that in nature, phages typically display a narrow "species" or "strain" level host range [6,7,8] favors the interpretation that indeed independent phage-host system arise. That said, there are a few exceptions and some phages have been found to display a wide host range [8], however this does not seem to be the general case. Therefore we hypothesize that over time, as bacterial strain 2 and viral strain B continue to evolve, the cross infectivity B-1 and A-2 naturally fades, and we will define *strains* 2 and B as new *species* when this cross affinity disappears. Note that this hypothesis ensures that the last property of **independence** is satisfied since we require that emerging phage-host systems lose their dependence on the parental strains to which they were linked initially (discussed further below).

Stage 5: New bacterial species and viral species emerge (Fig. 5.1E). Bacterial strain 2 and viral strain B have evolved sufficiently that cross infectivity has completely faded. At this point bacterial strain 2 is under the exclusive control of viral strain B via Eq. 7 and can "take up" concentration. The association between bacterial strain 2 and viral strain B is stable and lasting. Bacterial strain 2 now answers the definition of a *species* (it is a replicating *strain* stably associated with a viral *strain*) and can be regarded as a new species (species 2). Viral strain B now also answers the definition of a *species* (it is a replicating *strain* stably associated with a bacterial *strain*). At this stage the process can begin again and a new *species* can emerge.

The conclusion from this model is that new bacterial *species* must emerge with a new viral *species* and vice versa. While it has been shown in many experiments that bacteria can evolve in the absence of viruses, this model proposes that in the presence of lytic viruses, the process of evolution may be accelerated.



Step 1: Mature *species*
Stable phage-host system.
Bacterial species 1 is under the control of viral species 1 that determines its fixed point concentration.

Step 2: *Strain* Inception event
Emerging new bacterial *strain,* strain 2.
Bacterial species 1 is still under the control of viral species A.

Step 3: Emerging *strains*
Emerging new viral *strain*, strain B, possibly accelerating *strain-strain* evolution through an arms race and sweeping events.

Step 4: *Strain* development
Bacterial strain 2 and viral strain B have emerged. Dynamics are governed by a network of *strain-strain* interactions

Step 5: Mature *species*
Cross infectivity has faded. A new bacterial *species*, species 2 and a new viral *species*, species B have emerged. Species 2 is under the control of species B that determines its fixed point concentration.

Figure 5.1. A possible evolutionary process of bacterial and viral co-speciation. If species 1 and species 2 have the same size and growth rate, then stage E "takes up" twice the concentration as stage A, with the intermediate states somewhere in between.

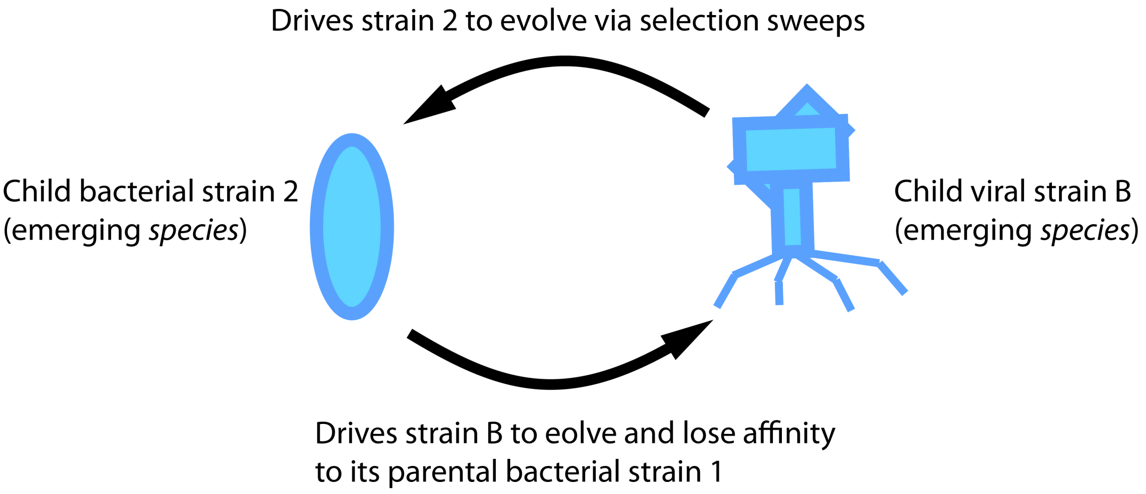

Drives strain 2 to evolve via selection sweeps
Child bacterial strain 2 (emerging *species*)
Child viral strain B (emerging *species*)
Drives strain B to eolve and lose affinity to its parental bacterial strain 1

Figure 5.2 Positive feedback evolution model for emerging bacterial and viral *species*. We propose that the arms race between bacteria and viruses may be a critical step in the formation of a new bacterial and viral *species*. This process is critical in order to allow viral strain B to relinquish its control of its parental bacterial strain (strain 1) while at the same time gaining control over the new bacterial strain (strain 2). Therefore this "arms race" may allow the two emerging *species* to form a one-to-one association, leading to the result that viral species have a narrow (*species*) host range. This process may also be critical for the bacterium, where by selective sweeps it drives the bacterium to evolve away from its original parental strain. This positive feedback model may amplify initially "noise". Thus, the process of co-speciation is perhaps equivalent to "amplification of noise" and therefore potentially a chaotic effect. The random trajectory in the genome space may be biased by selection pressure due to environmental factors such as available nutrients, competition and so on. Consequently, phylogenetic trees may have a fractal quality to them, though branches may be biased by selection pressure. Covering the genome space at such an exponential rate may be required in order to converge to a solution on a practical timescale, especially given the fact that bacteria are much less efficient at exploring this space than diploid organisms. Thus, the arms race may be an equivalent solution of bacteria to sexual reproduction (possibly a good enough solution for a smaller genome size).

5.3.2 A coarse-grained view of the evolutionary model satisfies all the properties of the biophysical model

We have seen that all the properties of the biophysical model except for independence were satisfied by the definition of *strains* and *species* that we use. The key point of this model is how the property of independence arises. According to Fig. 5.1, a bacterial *species* is born out of a single parental bacterial *strain*. Initially the new bacterial *strain* is under the control of both a parental viral *strain* and a new viral *strain* (the latter also born out of a parental viral *strain*). Once the new bacterial *strain* has evolved sufficiently from its parental s*train*, it loses its association with the parental viral *strain*. At the same time, the new viral *strain* loses its control over the parental bacterial *strain* (the transition from Fig. 2D to Fig. 2E). Thus, once both bacterial and viral *species* become an independent pair they are defined to be a new *species*. Therefore, the model described in Fig. 5.1 leads to a "world" where every bacterial *species* is controlled by just one viral *species*, and vice versa. If we view our evolutionary model in a coarse-grained way, and ignore the "structure" within each *species* (shown in Fig. 5.3 and discussed below), we obtain a model where each pair of interacting bacterial and viral *species* are independent of all other pairs (the condition of **independence** is satisfied). Therefore the two models become equivalent in the limit of describing organisms at a low genetic resolution, where the subtle differences between different *strains* within *species* (be it bacterial or viral) are lost. By interpreting the properties of the *species* defined in the current model, we can now answer the question raised in Section 4.4 of what is a "species"?

5.3.3 Revisiting the question of what is a "species"?

5.3.3.1 "Quark-gluon" model of a species

In one of the intermediate stages in the formation of a new bacterial *species* there is a state where a 2x2 network of interactions forms between the new and parental bacterial *strains* and the new and parental viral *strains* (Fig. 5.1D). In the general case, bacterial *strains* are continuously emerging from parental *strains*. Thus in the general case (applying our conservation rule that the number of bacterial *strains* must always equal the number of viral *strains*) we obtain a network of *n* bacterial *strains* infecting *n* viral *strains* (Fig. 5.3). These *n* bacterial *strains* are defined to be a bacterial *species*. Therefore, at any given point in time, a bacterial *species* in nature is comprised of *strains* that are in the process of maturing into new *species* (Fig. 5.1D). Likewise, the *n* viral *strains* can be technically classified as a viral *species*. Thus, a viral *species* is in essence a collection of viral strains, i.e., a collection of viral quasispecies, infecting the s*trains* of a given bacterial *species* in a network-like fashion (Fig. 5.3).

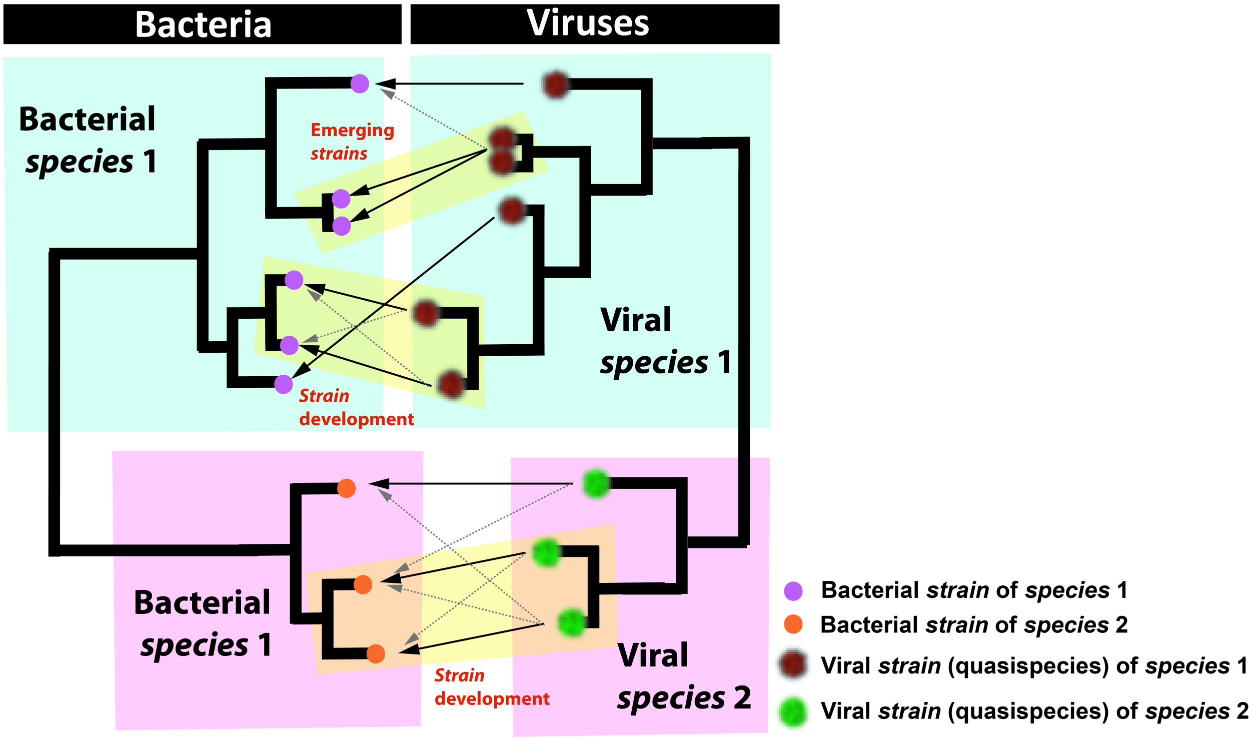

Bacteria
Viruses
Bacterial *species* 1
Emerging *strains*
Viral *species* 1
Strain development
Bacterial *species* 1
Strain development
Viral *species* 2
Bacterial *strain* of *species* 1
Bacterial *strain* of *species* 2
Viral *strain* (quasispecies) of *species* 1
Viral *strain* (quasispecies) of *species* 2

Figure 5.3 The "Quark and gluon" model of a *species*. Hypothetical phylogenetic tree of a conserved bacterial gene (left), revealing two bacterial *species*, paired with a phylogenetic tree of a conserved viral gene (right), revealing two corresponding viral *species*. Each clade of a *species* is comprised of the *strains* of that *species*. Yellow boxes highlight bacterial-viral *strains* in different stages of maturation. The arrows show which bacterial *strain* is infected (i.e., controlled) by which viral *strain*. Solid lines represent primary targets, whereas dashed lines represent secondary, weaker targets. The biophysical model that we propose lumps all *strains* within each *species* clade into one class.

5.3.3.2 The meaning of $N_{species}$

The case of a *species* comprised of one parental *strain*

To understand which entities contribute to $N_{species}$ we need to ask ourselves who "takes up" concentration in the model presented in Fig. 5.1 (or the more realistic view in Fig. 5.2). Let's begin by considering Fig. 5.1 again. Let's assume stage 1 "takes up" concentration x. Stage 2 is still under the control of one virus, so it "takes up" a concentration x as well. We will skip stage 3 for a moment. Stage 4 however is different. In this stage we have two *strains* in a 2x2 "butterfly" network configuration (Fig. 5.4). For simplicity let's assume that the coupling constants (i.e., infection rates) are $k_{11} = k_{22} = \alpha$ and $k_{12} = k_{12} = \beta$. Initially when bacterial strain 2 just emerges (the "child" strain), we have $\beta = \alpha$. This is because the child viral strain B also has just emerged and it is barely *distinguishable* from its parent viral strain A. At this stage we anticipate that both bacterial *strains* (parent 1 + child 2) will contribute together a concentration of x because both are under the control of one viral *strain* (parent A + child B). As the child bacterial strain 2 and child viral strain B evolve, the parent-child coupling constants are hypothesized to fade, and so $\beta \to 0$. When $\beta = 0$ a new *species* of bacteria and viruses has emerged. At this stage, we expect both new bacterial *strains* (*species*) to contribute together 2x to the concentration.

This effect can be readily appreciated by solving the butterfly network: Let B_i be the concentration of bacterial *strain i*, and V_i the concentration of the viral *strain i* , where *i*=1 are the parental strains and *i*=2 are the child strains (Fig. 5.4). The rate equations for the viral *strains* are given in the general case by

$$\frac{dV_1}{dt} = \alpha b V_1 B_1 + \beta b V_1 B_2 - \tilde{\gamma} V_1$$
$$\frac{dV_2}{dt} = \beta b V_2 B_1 + \alpha b V_2 B_2 - \tilde{\gamma} V_2$$

where b is the burst size (assumed to be equal for the two *strains*). Assuming steady-state conditions (to obtain the fixed point concentrations), after some algebra (defining $\gamma \triangleq \tilde{\gamma}/b$), we find that

$$B_{tot} = B_1 + B_2 = \frac{2\gamma}{\alpha+\beta} = \frac{2\gamma}{\alpha}\frac{1}{1+\beta/\alpha} = \frac{2\gamma}{\alpha}\frac{1}{1+\kappa}.$$

where we have defined the normalized parent-child coupling constant $\kappa \triangleq \beta/\alpha$.

Thus, initially, when $\kappa = 1$ we have $B_{tot} = \frac{\gamma}{\alpha}$, and when $\kappa = 0$ we have $B_{tot} = \frac{2\gamma}{\alpha}$. Thus, exactly as we predicted, the total concentration "taken up" by bacterial strains 1+2 increases from $\frac{\gamma}{\alpha}$ to $2\frac{\gamma}{\alpha}$ during the maturation process of the new *species*. We can parameterize this uncertainty with a "maturation factor" μ:

$$B_{tot} = \mu B_{\text{species}}, \quad \text{where} \quad 1 \le \mu \le 2$$

where B_{species} is the concentration one would obtain if one were to coarse grain the system to a *species* level ignoring *strains*. Therefore, in the case of a 2x2 network, if we were to coarse grain bacteria to a *species* level (say an OTU of 3%), we would be underestimating the concentration taken up by the species by a factor anywhere from μ=1 to μ=2 (Fig. 5.5). Now let's see what happens in a more realistic scenario when a species is comprised of *n* strain (where in reality *n* can be very large since it probes the microdiversity of a *species*).

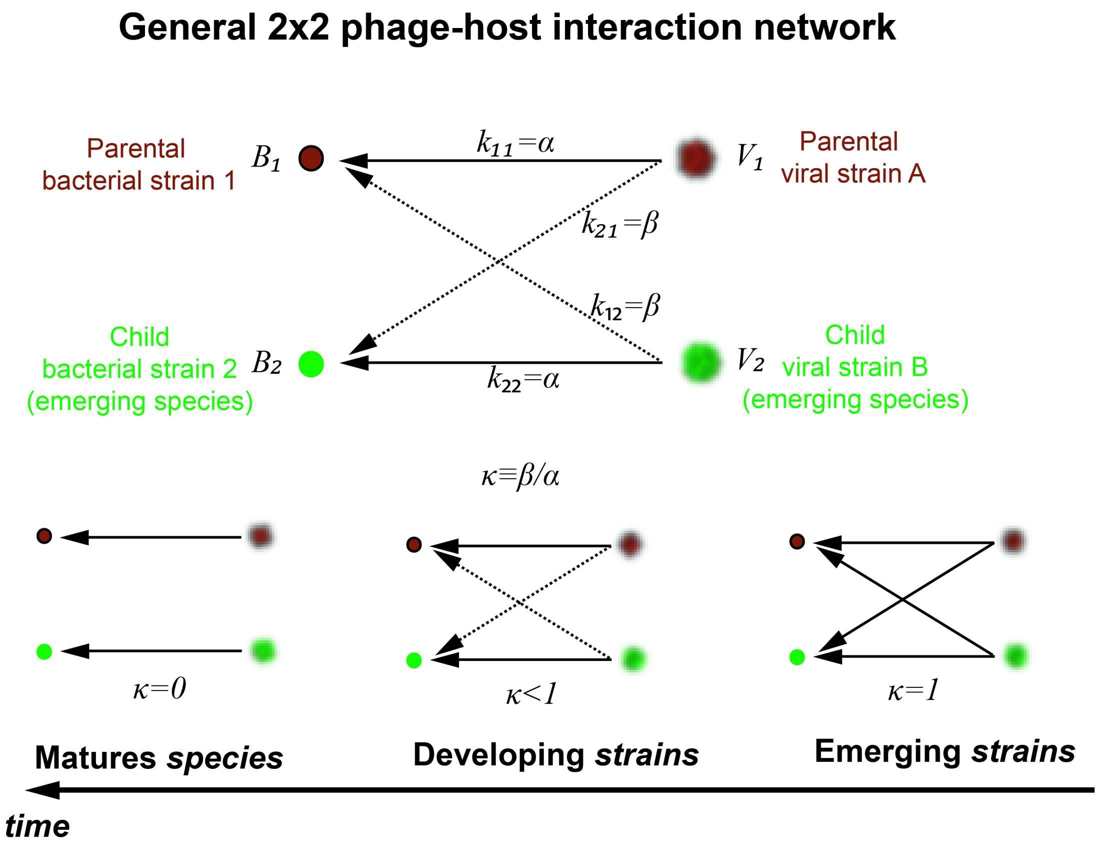


Figure 5.4 A 2x2 phage-host interaction network with event timeline. This diagram is a general 2x2 interaction network between two viral *strains* — a parental viral *strain* (strain A) and the emerging viral *species* (strain B), that are controlling the parental bacterial *strain* (strain 1) and the emerging bacterial *species* (strain 2). The timeline shows the hypothesized evolutionary trajectory of these four *strains*. Initially, as the new (child) *strains* have just emerged, the coupling constants are equal. As the child *strains* evolve, the parent-child coupling constants decrease (dashed lines). Finally the child *strains* have evolved enough so that the parent-child coupling constants are 0 and new *species* of bacteria and viruses have emerged.

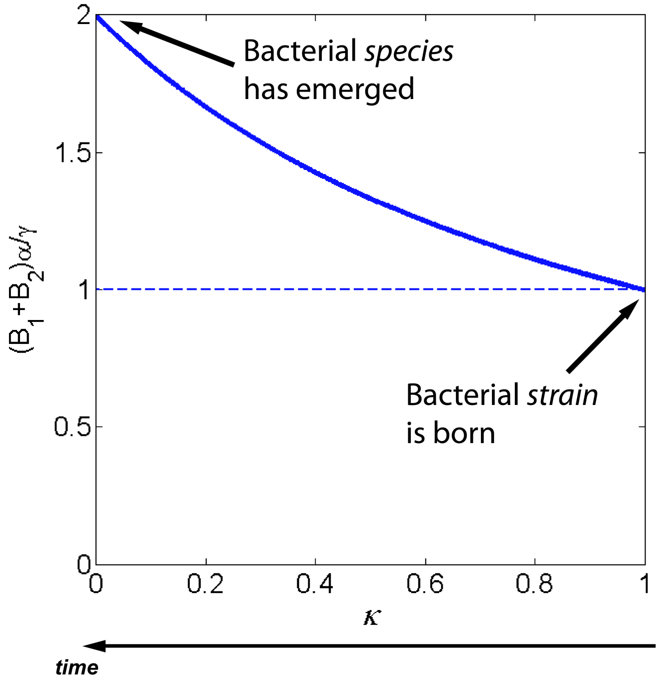


Figure 5.5 Total concentration "taken up" by parent and child bacterial *strains* as child bacterial *strain* evolves towards a new *species*. Here we show how the sum concentration of both parent and child bacterial *strains* changes with time, as the parent-child coupling constant κ goes to 0. Initially, when the bacterial child *strain* is born, it is under the control of the parental viral strain and the parent-child coupling constant is maximal (κ=1). The total concentration at this point is that of a single bacteria *strain* (=1 in normalized units). When the bacterial child *strain* is fully evolved, the parent-child coupling constant equals 0 and a new bacterial *species* under the control of a new viral *species* has emerged. The total concentration at this point has doubled because the new bacterial *species* is allowed by its controlling virus to "take up" a concentration =1 (in normalized units).

The case of a *species* comprised of *n* parental *strains*

In the general case (Fig. 5.3) a bacterial *species* will be comprised of N_{strain} parental *strains*. Each of these parental *strains* is anywhere in the stage between emerging a new *strain* to having a fully emerged *species* (thus the total number of strains will be anywhere between N_{strain} and $2N_{strain}$). We make the approximation that each one of these parental strains is part of a butterfly network with coupling constant β, which is anywhere between $\beta = \alpha$ to $\beta = 0$. If all *strains* were in a state of $\beta = \alpha$ then the total concentration "taken up" by this *species* would be

$$B_{tot} = \sum_{i=1}^{N_{strain}} B_1^{(i)} + B_2^{(i)} = \frac{\gamma}{\alpha} N_{strain} .$$

If all *strains* were in a state of $\beta = 0$ the total concentration "taken up" by this *species* would be

$$B_{tot} = \sum_{i=1}^{N_{strain}} B_1^{(i)} + B_2^{(i)} = \frac{2\gamma}{\alpha} N_{strain} .$$

Therefore,

$$B_{tot} = \mu N_{strain} B_{\text{species}}, \qquad \text{where } 1 \le \mu \le 2$$

where B_{species} , once again, is the concentration one would obtain if one were to coarse grain the system to a *species* level ignoring *strains*. Therefore the number of "species" in Eq. 16B is given by

$$N_{\text{"species"}} = \mu N_{strain} \approx N_{strain} .$$

Thus, our conclusion from this analysis is very simple and logical. Even though the total number of actual independent phage-host systems is equal to the number of *species* we need to multiply each *species* by a factor which approximately equals the number of strains in that *species*. Thus by probing the "structure" of a *species* (which is the assumed construct in the biophysical model) we came to the conclusion that one needs to weigh each species approximately by the number of *strains* in that *species*. Since strains are *distinguishable*, indeed each strain should contribute to the total concentration between ×1 and ×2.

5.3.3.3 The dynamics of speciation

The process of *speciation* (i.e., co-formation of new bacterial and viral *species*) is inherently stochastic since a bacterial *strain* can easily become extinct if a viral *strain* is lost, as the system becomes unstable (Section 4.6). We therefore envision the process of *speciation* as one in which new bacterial *strains* continually emerge from extant *strains* (the microclades in Fig. 5.3), with some *strains* evolving to become *species*, and with other *strains* being lost (Fig. 5.6). In principle, one should be able to calculate the rate at which bacterial *species* are formed in the oceans, possibly yielding better bounds on the total diversity in the oceans.

5.3.3.4 Analogy to the conventional concepts of a "species" and "strain"

The intuitive notion of a bacterial "strain" has been familiar to biologists for many years. Indeed genetic microdiversity below the species level has been observed in nature [9,10]. We too have observed such microdiversity in treponeme cells found in the termite hindgut ("Host I" and "Host III" in Fig. 2.2). The concept of a bacterial "species" comprised of "strains" is also well known and widely used by biologists, though the empirical identity thresholds used for classification of new species are somewhat questionable given the lack of a rigorous definition of a species. The concept of a *strain* of viruses is also familiar, this is the well-known quasispecies proposed by Eigen [1]. The definition of viral "species" on the other hand has been quite elusive [11]. If the model we propose proves to be valid, then it would seem that a host-range-based taxonomy [11] should lead to a meaningful organization of viral species, at least for marine ecosystems. In principle, according to our model, the true classification of marine life-forms (bacteria + viruses) requires both to be classified simultaneously. For example, when two marine bacterial "species" seem very similar (using "species" in the colloquial meaning), then according to our proposed model, if these "species" are infected with different non-overlapping viruses they should be classified as different *species*.

5.3.3.5 The insight for the coarse-grained model

When considering the coarse-grained biophysical model, the most natural definition for a "species" would be "a cell that can be distinguished reproducibly from all other cells", i.e., the definition of a *strain*. The evolutionary model has shown that this is not the case, as one needs a more complex structure, defined here as a *species*, in order to obtain a "world" of non-interacting phage-host systems. Thus, the "species" in the biophysical model are equivalent to the *species* defined in our evolutionary model, however, the concentration of each *species* needs to be multiplied by a weight of $\sim N_{strain}$, which is the number of *strains* in each *species*. This conclusion also leads to a clear distinction between the concept of a bacterial *strain* and a bacterial *species*. While a bacterial *species* interacts with just one viral *species* and vice versa, a bacterial *strain* interacts with several viral *strains* and is not an independent entity.

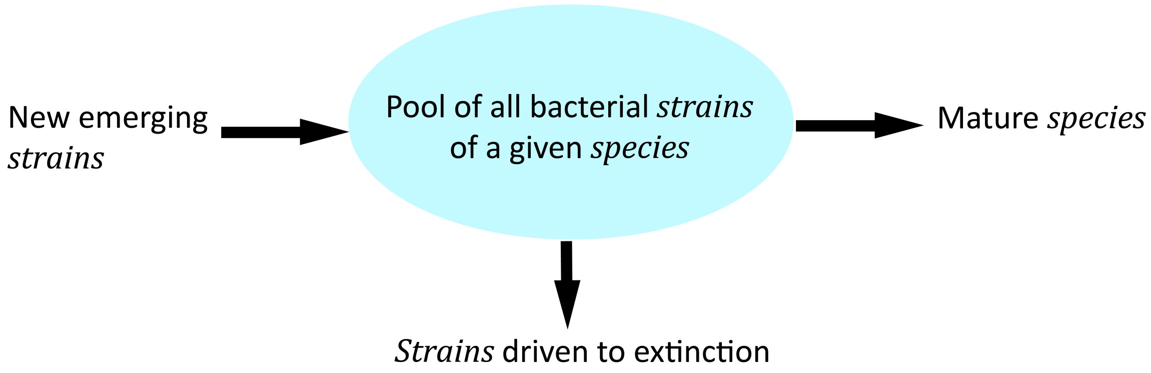

New emerging *strains*
Pool of all bacterial *strains* of a given *species*
Mature *species*
Strains driven to extinction

Figure 5.6. Flux of *strains* in the process of bacterial speciation. According to our evolutionary model, bacterial *strains* of a given *species* are cells that are *distinguishable* for all other cells in the population, but do not form a stable (i.e., unique) association with viral *strain* (see Fig. 5.1D). A bacterial *strain* matures into a *species* if it forms a one-to-one association with a viral *strain*. The pool in this figure is the sum of all bacterial *strains* comprising a *species*. The flux into this pool comes from new emerging *strains* (Fig. 5.1B & D). The flux out of this pool is due to either *strains* that have gone extinct (e.g., since the viral network in which they were in became destabilized), or *strains* that have matured into species (Fig. 5.1E).

5.3 Why do phages typically have a narrow host range?

It is a known fact that most phages are species or strain specific (although a few exceptions have been found) [6,7,8]. Naïvely, this observation seems peculiar given that all cellular life forms encode and read information in virtually the same manner (e.g., human genes can be expressed in bacteria). Generally speaking, the genome of phage A could be expressed in many very divergent species, yet phages tend to infect a single species. Why is this the case?

The evolutionary scheme we propose here in fact predicts that phages should have (in the majority of cases at least) a *species*- or *strain*-level host range. According to our model, any given viral *species* is expected to infect a single bacterial *species* (Fig. 5.1 and Fig. 5.3). Thus, the viral *strains* associated with a given viral *species* will infect some (or all) of the bacterial *strains* within a given bacterial *species* (Fig. 5.3). Our model therefore predicts that viruses will have either a “species”-specific host range (infecting all *strains* of a given bacterial *species*) or a “strain”-specific host range, infecting a subset of bacterial *strains* (or at the very minimum a single bacterial *strain*).

Mechanisms to generate a wide host range. A viral *species* could in principle evolve to infect another bacterial *species* in addition to its original host (and thus the former bacterial *species* will be susceptible to more than one viral *species*). As long as the viral species is part of an *n*x*n* network of associations, the dynamics are stable (see Section 4.6). However, such a scenario seems to be the exception since in open systems at least, *species* are not spatially constrained. Therefore, if a *species* drifts off, the network will become imbalanced (i.e., *n*x*m* where *n≠m*) leading to unstable dynamics and, over time, extinction events. This leads to a prediction that in closed systems (for example the gut) there will be more viruses with a wide host range than in open systems. Indeed, phages isolated from sewage appear to display a wide host range [12].

Another possibility for a wide host range is the following: if the cross-species infection in Fig. 2D does not fade away with time as we hypothesized, then in a closed system it is possible to have a lytic viral *species* with a wide host range if it is part of an *n*x*n* network of hosts and viruses. However, in an open environment, where *species* are not spatially constrained, again the system may become unstable as described above. Thus, unless the environment is constrained to a closed volume, it seems that generally a more robust and stable solution (and therefore more likely scenario) would be for phages to have a narrow host range. That said, the scheme we have presented here does not preclude the possibility that a given viral *species* happens to be successful in infecting many bacterial *species* that are not present in the given environment (e.g., they happen to have the same membrane receptor). Such coincidental events should also be kept in mind.

5.4 Testing the evolutionary model: evolution experiment of a phage-host system

One possible way to test our model is to perform a Lenski-type evolution experiment of a phage-host system (similar to the evolution experiments of Rainey [13]). One choice would be T4 and *E. coli* . To prevent total annihilation of the bacteria, we should add a degradation factor for the phages (or perhaps a chemostat would be sufficient?). *E. coli* is a good choice since its CRISPR system has been investigated [14]. After *n* generations would expect at least two new bacterial strains to co-emerge with an equal number of viral strains. After enough generations the *n* emerging strains should be distinguishable (measurable by sequencing). Furthermore, we should observe a decrease in parent-child cross affinity between the new evolving viral strain(s) and the original viral strain. In a different experiment, one can evolve a strain of *E. coli* with a mutation in one of the *cas* proteins inactivating the CRISPR array defense mechanism. We expect that either we will not observe the emergence of new strains, or that it will take a much longer time to obtain the same evolutionary distance between strains.

5.5 References

1. Eigen M (2006) Viral quasispecies. Evolution: a Scientific American reader: 114.
2. Ducklow H (2000) Bacterial production and biomass in the oceans. Microbial ecology of the oceans 1: 85-120.
3. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology 6: 181-186.
4. Banfield J, Young M (2009) Variety—the Splice of Life—in Microbial Communities. Science 326: 1198.
5. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28: 491.
6. Suttle C (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. Viral Ecology: Academic Press. pp. 247–296.
7. Kutter E, Sulakvelidze A (2005) Bacteriophages: biology and applications: CRC Press.
8. Weinbauer M (2004) Ecology of prokaryotic viruses. FEMS Microbiology Reviews 28: 127-181.
9. Moore L, Rocap G, Chisholm S (1998) Physiology and molecular phylogeny of coexisting. Nature 393: 465.
10. Thompson J, Randa M, Marcelino L, Tomita-Mitchell A, Lim E, et al. (2004) Diversity and dynamics of a North Atlantic coastal Vibrio community. Applied and Environmental Microbiology 70: 4103.
11. Lawrence J, Hatfull G, Hendrix R (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. Journal of bacteriology 184: 4891-4905.
12. Jensen EC, Schrader HS, Rieland B, Thompson TL, Lee KW, et al. (1998) Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. Applied and Environmental Microbiology 64: 575.
13. Buckling A, Rainey PB (2002) Antagonistic coevolution between a bacterium and a bacteriophage. Proceedings of the Royal Society of London Series B: Biological Sciences 269: 931.
14. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960.

Chapter 6

A Kinetic Model for Stress Fiber Contraction and Relaxation

6.1 Abstract

High-resolution measurement of the force generated by a single fibroblast cell during perturbation with Cytochalasin D (CD) or recovery reveals that the force is quantized. The origin of the force is thought to be a ventral stress fiber in the fibroblast. The magnitude of the quantized jumps in the force (~1 nN) appears to be consistent with a model where individual sarcomeres abruptly assemble or disassemble. Here we consider the dynamics of this process: the measured temporal profile for stress fiber contraction and relaxation and the duration of the force steps. We show that the observed dynamics are consistent with a simple stochastic model in which the time it takes for an individual sarcomere to abruptly assemble or disassemble is exponentially distributed with some characteristic constant rate $1/\tau$. The model is based on three parameters: the number of sarcomeres in the stress fiber N_S, the force step size f_0, and the above timescale τ. The stochastic model makes the following predictions: (1) the total force generated by a stress fiber should follow on average an exponential temporal profile, (2) the length of time between two subsequent force steps should be on average inversely proportional to the number of remaining sarcomeres (yet to assemble/disassemble) and therefore increases with time. The above three parameters (N_S, τ, f_0) were estimated for each measured

profile. Model predictions appear to be in good quantitative agreement with the fibroblast dataset and transition times were found to be statistically consistent with an exponential distribution (applying the Pearson’s chi-square test statistic). These findings support the proposed interpretation that sarcomeres assemble abruptly upon polymerization and disassemble abruptly upon depolymerization of actomyosin complexes.

6.2 Introduction

Actin filaments in association with myosin filaments form contractile structures in both muscle and non-muscle cells. One type of actin filament structure found in many types of non-muscle cells involved in adhesion, motility, and morphogenesis, are known as contractile actin stress fibers (*1-2*). Structurally, stress fibers are bundles of actin filaments containing bipolar periodic arrays of myosin II between consecutive α-actinin-foci (*1*) (a schematic diagram of a stress fiber is shown in Fig. 6.1). Stress fibers are thus structurally similar to muscle myofibrils and can also similarly produce contractile force, however unlike myofibrils, these structures are less organized (*1*). Furthermore, despite the similarity to myofibrils, stress fibers typically do not display repeatable contraction and relaxation on relatively short time scales but rather contract continuously with occasional relaxing or stretching (*1*).

There are three types of stress fiber that are known to exist: dorsal stress fibers, ventral stress fibers, and transverse arcs (*1-2*). Dorsal stress fibers, which typically attach to focal adhesions only at one end, are known to display uniform polarity and therefore may not be contractile structures at all (*2*). Indeed, these structures have never been observed to contract (*1*). Transverse arcs are curved acto-myosin bundles that do not directly attach to focal adhesions. Since these structures are not anchored to the plasma membrane, it is unclear if they can transmit force (*2*), although this may perhaps be achieved indirectly, as these fibers are often observed to interact with dorsal stress fibers, which in turn anchor them to the substrate (*1*). Finally the most commonly observed type of stress fibers are ventral stress fibers. These contractile structures are tethered at both ends to focal adhesions and are thus capable of generating tension under constant length (*1-2*). Indeed, the majority of the contractile force that a fibroblast applies to the substrate is aligned with the direction of ventral stress fibers (*2*). Thus ventral stress fibers may be the most effective generators of contractile force in these cell types. These type of stress fibers are thought to be responsible for tail retraction as well as other cell shape changes occurring due to increased contraction (*1*). In addition these fibers are thought to work against membrane tension at cell borders (*1*).

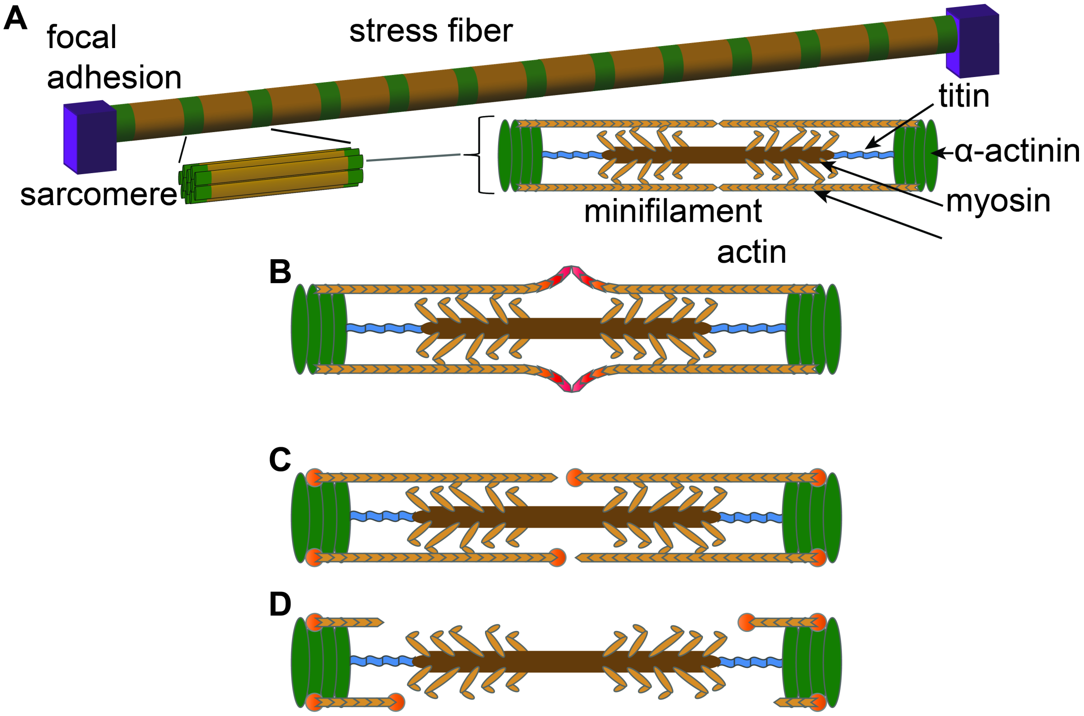


Figure 6.1. Cartoon model of a ventral stress fiber. A. Each stress fiber is thought to be composed of a serial sequence of sarcomeres where each sarcomere is composed of approximately 50 minifilaments in parallel, and each minifilament is composed of a bipolar myosin filament and opposing actin filaments held together by α-actinin and titin. **B.** Cartoon of unperturbed (closed) minifilament, actin polymerization pressure is highlighted in red. In the proposed model we assume that each closed sarcomere generates a discrete unit of force f_0. **C.** Cartoon of minifilament after initial exposure to CD (orange) has stopped actin polymerization. **D.** Cartoon of "open" minifilament after actomyosin complexes have disassembled due to CD exposure. Figure courtesy of Blake Axelrod (Roukes lab, Caltech).

Here we analyze data collected by Blake Axelrod at Caltech, a member of the Roukes group, who measured with the highest resolution to date the force as a function of time of a single fibroblast cell. For this purpose, Blake constructed an instrument that utilizes a polymer Nano-Electro-Mechanical Systems (NEMS) force sensor with integrated piezoresistive strain sensing to measure the force generated by adherent cells (see experimental setup in Fig. 6.2). To precisely control the aqueous media surrounding the cell, the NEMS chips are encapsulated in multi-layer PDMS microfluidics with pneumatically actuated valves. This device allows the experimenter to change the growth medium of cell while it is attached to the force sensor.

In the actual experiment, a single NIH-3T3 fibroblast cell attached itself to a platform adjacent to the force sensor, making contact with the force sensor. An image of this cell in contact with the NEMS device is shown in Fig. 6.3. While in the recovery medium, the calibrated force sensor registered a force of about 20 nN (see force profile in Fig. 6.3). To perturb the cell, Cytochalasin D (CD) was flowed in. CD forms stronger bonds to the barbed ends of the actin filaments than to the pointed ends (*3-4*) leading to depolymerization of the actin filaments, disassembly of the actomyosin complex, and loss of contraction. CD belongs to a class of substances called cytochalasins, that are fast-acting and reversible disruptors of contractile force that have been used extensively over the past 40 years to study the role of actin and contractile forces in various cellular processes (*5*). Despite extensive use and study, how the known molecular mechanisms of CD cause the reversible disruption of stress fiber force generation without dismantling stress fibers remains unclear. Upon exposure to CD, Blake indeed observed a decrease in force, following an exponential-like profile (Fig. 6.3). Indeed when flowing the recovery medium the force was generated again, presumably owing to repolymerization of the actomyosin complex. This process was repeated for several iterations (Fig. 6.8).

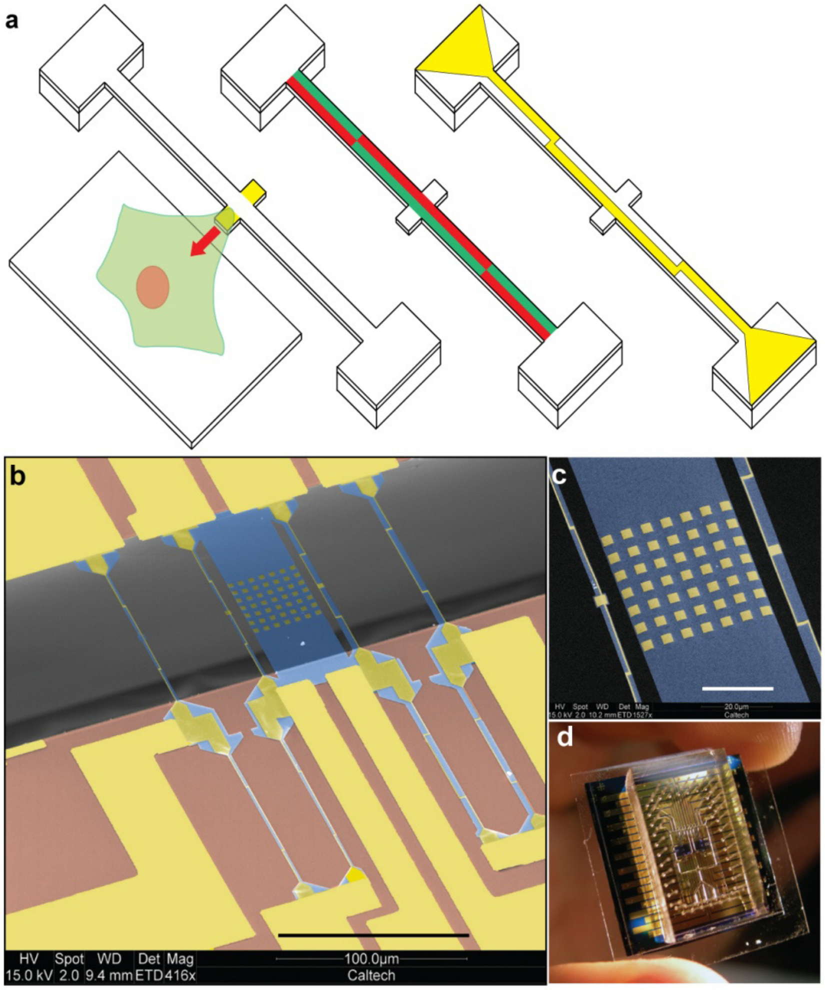


Figure 6.2. NEMS force sensor. a) At left, diagram of the cell on the platform adjacent to the force sensor, cell is spread and attached to the sensor which is coated with fibronectin. At center, contraction generates tensile (green) and compressive (red) strains in the beam. At right, the piezoresistor (yellow) is patterned across the beam to couple only to the tensile strains. **b)** Colorized SEM image of two force sensors, the left one is 2 µm wide, the right is 4 µm wide, both are 100 µm long. The stable platform for the cell is between them. Scale bar is 100 µm. **c)** Colorized SEM image showing close up of two force sensors and platform, gold surface pads and the gold piezoresistors are visible in yellow. Scale bar is 20 µm. **d)** Microfluidics encapsulated chip held between thumb and forefinger. Figure and caption courtesy of Blake Axelrod (Roukes lab, Caltech).

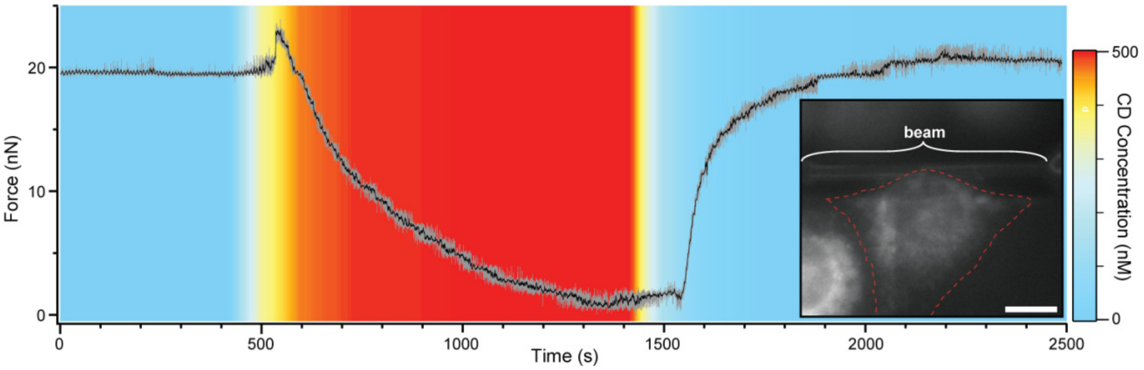


Figure 6.3. Force time time response to CD perturbation. Typical force response to 500 nM CD: switch from conditioned media to CD at 400 sec results in initial contraction. Following the initial contraction, the force drops as expected. At 1400 sec the flow is switched back to conditioned media and the cell re-establishes normal contractile force. The black line is a 1 sec running average, the grey line is the raw data (100 ms integration time, ~200 pN force noise). The background color depicts the CD concentration as estimated using finite elements simulations (COMSOL). There is transit time for the CD solution to reach the cell through the microfluidics after actuating the microfluidic valves at 400 sec and the ~1 nL/sec flow rate takes additional time to displace the conditioned media from the 150 nL chamber. Inset shows fluorescent image of the cell attached to the beam taken immediately before data acquisition, scale bar is 10 µm. Figure and caption courtesy of Blake Axelrod (Roukes lab, Caltech).

When examining the temporal profiles closely, these profiles revealed quantized steps in force during both CD induced relaxation and the post-CD contraction. A close up of the force profiles is shown in Fig. 6.4. When a step-fitting algorithm was applied (*6*), a histogram of the computed step sizes revealed that these steps were remarkably uniform with average step size 1.08 nN ± 0.18 nN (N=96, S.D.). In the following chapter we present our analysis of Blake’s experimental data and present a stochastic kinetic model for stress fiber assembly and disassembly that is capable of accounting for the observed temporal dynamics.

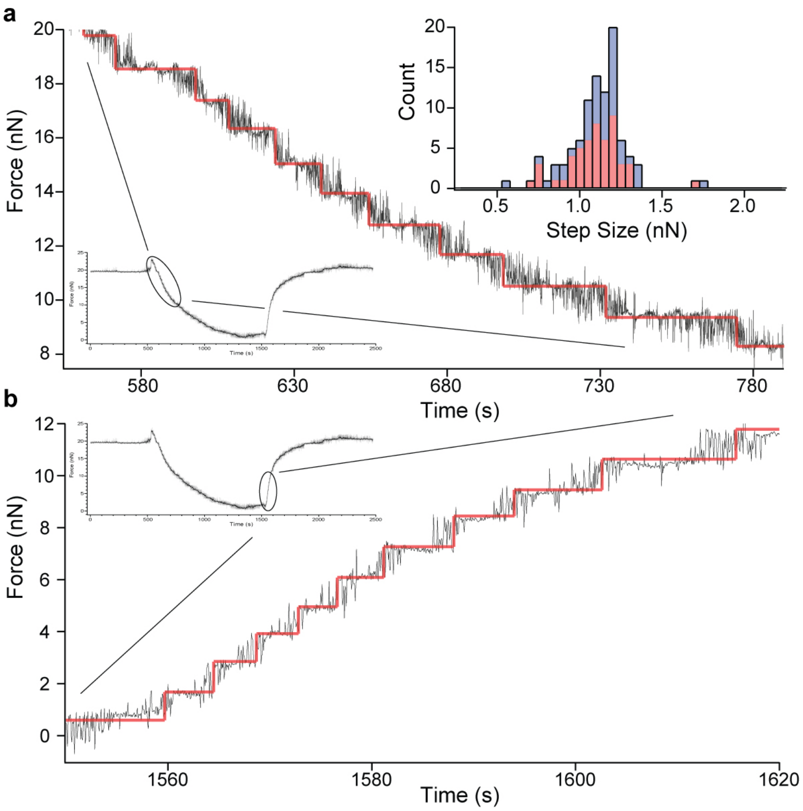


Figure 6.4. Close up of force steps. a) Steps during CD induced force collapse, red is output from step fitting algorithm. Inset is step size histogram for 3 force collapse and recovery cycles. The average step size is 1.08 nN ± 0.18 nN (N=96, error is standard deviation). **b)** Steps during force recovery following return to conditioned media, red is output from step fitting algorithm. Figure and caption courtesy of Blake Axelrod (Roukes lab, Caltech).

6.3 Model development

We assume that each stress fiber is composed of N_S sarcomeres, with each individual sarcomere capable of generating a unit of force f_0, as depicted in Fig. 6.1. The fact that the total force changes in quantized steps suggests that generation of a unit of force by a sarcomere can be approximated by an all-or-none event. Assuming that the quantized steps reflect dynamics of individual sarcomeres it follows that the assembly or disassembly (i.e., failure) of an individual sarcomere is an all-or-none event with amplitude f_0. Our model is based on the following assumptions: (1) Each sarcomere assembles or disassembles abruptly and irreversibly (i.e., the rate for the reverse reaction to occur is negligible). (2) Sarcomeres assemble or disassemble independently of each other. (3) The time until a sarcomere assembles or disassembles is exponentially distributed (reflecting a certain constant probability rate for this event to occur). (4) The time constant for assembly is the same for all sarcomeres. Similarly, the time constant for disassembly is the same for all sarcomeres. The model is illustrated for the case of disassembly in Fig. 6.5.

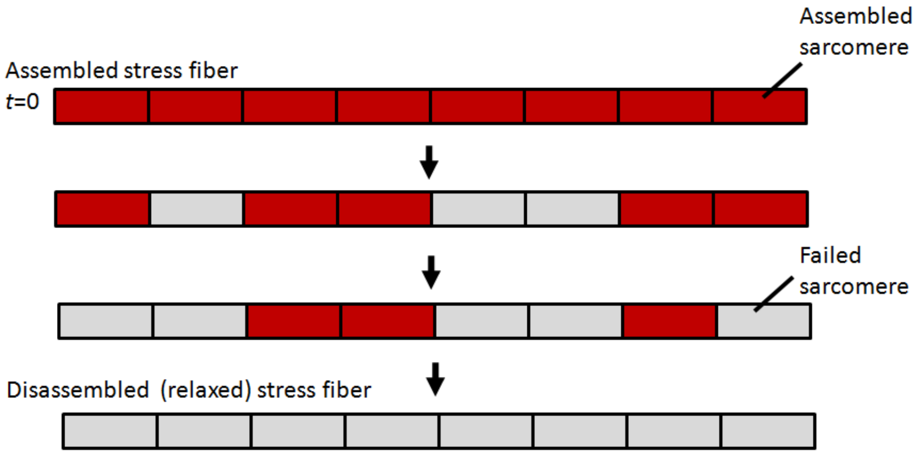


Figure 6.5. Schematic model for stress fiber relaxation

According to our model, we assume that the probability distribution for the time it takes a single sarcomere to assemble or disassemble is given by the exponential distribution

(1) $$f(t) = \lambda e^{-\lambda t}.$$

where (henceforth) $\lambda = \lambda_A = \tau_A^{-1}$ for assembly and $\lambda = \lambda_D = \tau_D^{-1}$ for disassembly. The probability that a single sarcomere will assemble or disassemble in the interval *0<t'<t* is therefore given by:

(2) $$p(t) = \int_{t'=0}^{t} \lambda e^{-\lambda t'} dt' = 1 - e^{-\lambda t}.$$

Hence the average number of sarcomeres to assemble or dissemble by time *t* is

(3) $$E\mathbf{n}(t) = N_S p(t) = N_S \left(1 - e^{-\lambda t}\right).$$

where $\mathbf{n}(t)$ is the number of sarcomeres that have assembled by time t $\left(0 \le \mathbf{n}(t) \le N_S\right)$ and where *E* represents the expectation value. The standard deviation of $\mathbf{n}(t)$ is given by the binomial standard deviation

$$\sqrt{\operatorname{var}\left(\mathbf{n}(t)\right)} = \sqrt{N_S pq} = \sqrt{N_S e^{-\lambda t}\left(1 - e^{-\lambda t}\right)}.$$

Considering the case of assembly first, if one assumes that each sarcomere contributes a unit force of f_0 to the total force generated by the stress fiber, then the stochastic force as a function of time would be given by

(4) $$\mathbf{F_{SF}}(t) = f_0 \mathbf{n}(t).$$

Taking the expectation value of $\mathbf{F_{SF}}(t)$ and using Eq. 3 we obtain the expressions for the average and standard deviation of the force of a stress fiber as a function of time:

(5) $$E\mathbf{F_{SF}}(t) = f_0 E\mathbf{n}(t) = f_0 N_S \left(1 - e^{-\lambda_A t}\right) = f_{\max}\left(1 - e^{-\lambda_A t}\right)$$

$$\sqrt{\operatorname{var} \mathbf{F_{SF}}(t)} = \sqrt{\operatorname{var} f_0 \mathbf{n}(t)} = f_0 \sqrt{N_S e^{-\lambda t}\left(1 - e^{-\lambda t}\right)}.$$

where $f_{\max} \equiv f_0 N_S$ is the maximum force generated by the stress fiber. Alternatively, the stochastic force can be written as a sum of N_S steps with an amplitude of f_0, where each step occurs at random times denoted by $\boldsymbol{\tau}_s$ (s=1..N_S):

(6) $$\mathbf{F_{SF}\boldsymbol{\tau}}\left(\mathbf{t}\right)=\sum_{s=1}^{N_S} f_0 u\left(t-\quad _s\right).$$

where $\boldsymbol{\tau}_s \sim \exp(\lambda)$ and $u(t)$ is a step function ($u(t \geq 0)=1,\ u(t<0)=0$). Note that Eq. 5 can be derived directly from Eq. 6 by taking the expectation value of $\mathbf{F_{SF}}\left(\mathbf{t}\right)$. The proof is as follows: let's rewrite $u\left(t-\boldsymbol{\tau}_s\right)$ in the following form:

$$u\left(t-\boldsymbol{\tau}_s\right)=\int_{t'=-\infty}^{\infty} d\,'tu\left(t'\right)\delta\left(t-\boldsymbol{\tau}_s-t'\right).$$

Taking the expectation value $u\left(t-\boldsymbol{\tau}_s\right)$ we find that

$$Eu\left(t-\boldsymbol{\tau}_s\right)=\int_{t'=-\infty}^{\infty} d\,'tu\left(t'\right)E\delta\left(t-\boldsymbol{\tau}_s-t'\right)$$

$$E\delta\left(t-\boldsymbol{\tau}_s-t'\right)=\int_{\tau_s=-\infty}^{\infty} d\tau_s f_{\boldsymbol{\tau}}(\tau_s)\delta\left(t-t'-\boldsymbol{\tau}_s\right)=f_{\boldsymbol{\tau}}(t-t')=\begin{cases}\lambda e^{-\lambda(t-t')} & t-t'\geq 0\\ 0 & t-t'<0\end{cases}$$

where $f_{\boldsymbol{\tau}}(t)$ is the probability density function of $\boldsymbol{\tau}_s: f_{\boldsymbol{\tau}}(t\geq 0)=\lambda e^{-\lambda t}, f_{\boldsymbol{\tau}}(t<0)=0$. Thus the integrand in $Eu\left(t-\boldsymbol{\tau}_s\right)$ is nonzero in the range $0\leq t'\leq t$:

$$Eu\left(t-\boldsymbol{\tau}_s\right)=\int_{t'=-\infty}^{\infty} dt'u\left(t'\right)f_{\tau}(t-t')=\lambda\int_{t'=0}^{t} dt'e^{-\lambda(t-t')}=e^{-\lambda(t-t')}\Big|_{t'=0}^{t}=1-e^{-\lambda t}$$

and thus

$$E\mathbf{F_{SF}\boldsymbol{\tau}}\left(\mathbf{t}\right)=E\sum_{s=1}^{N_S} f_0 u\left(t-\boldsymbol{\tau}\ _s\right)=f_0\sum_{s=1}^{N_S} Eu\left(t-\quad _s\right)=f_0\sum_{s=1}^{N_S}\left(1-e^{-\lambda t}\right)=f_0 N_S\left(1-e^{-\lambda t}\right)\therefore$$

For sarcomere disassembly we have

(7) $$\mathbf{F_{SF}}\left(t\right)=f_{\max}-f_0\mathbf{n}\left(t\right).$$

Taking the expectation value of $\mathbf{F_{SF}}\left(t\right)$ and using Eq. 3 we find that

(8) $$E\mathbf{F}_{\mathbf{SF}}\left(t\right)=f_{\max}-f_{0}E\mathbf{n}\left(t\right)=f_{\max}e^{-\lambda_{D}t}.$$

In Fig. 6.6A we plot the average stochastic force (Eq. 5) for $N_S = 20$ and λ_A*=1/200 sec*$^{-1}$. Fig. 6.6A also shows a simulation of the stochastic force (Eq. 6). To simulate the stochastic force, the assembly times for $N_S = 20$ sarcomeres were drawn from an exponential distribution with λ_A*=1/200 sec*$^{-1}$. From Fig. 6.6A we see that even though the average time until an individual sarcomere to assemble is long (200 sec), multiple independent sarcomeres will generate steps that can be much shorter than 200 sec, on average following the exponential distribution above. Note that the stochastic model predicts that the waiting time between assembly events generally increases with time. This is shown in Fig. 6.6A for the case of assembly and in Fig. 6.6B for the case of disassembly. Such increasing step times are also observed in the experimental data (e.g., Fig. 6.4). This effect will be discussed in more detail in section 6.7. In Fig. 6.6C we compare the model prediction of the expectation value and standard deviation for the stochastic force with a Monte Carlo simulation in which we averaged 1000 numerical time traces of the stochastic force. The figure shows that the Monte Carlo simulation converged precisely to the analytical expression in Eq. 5.

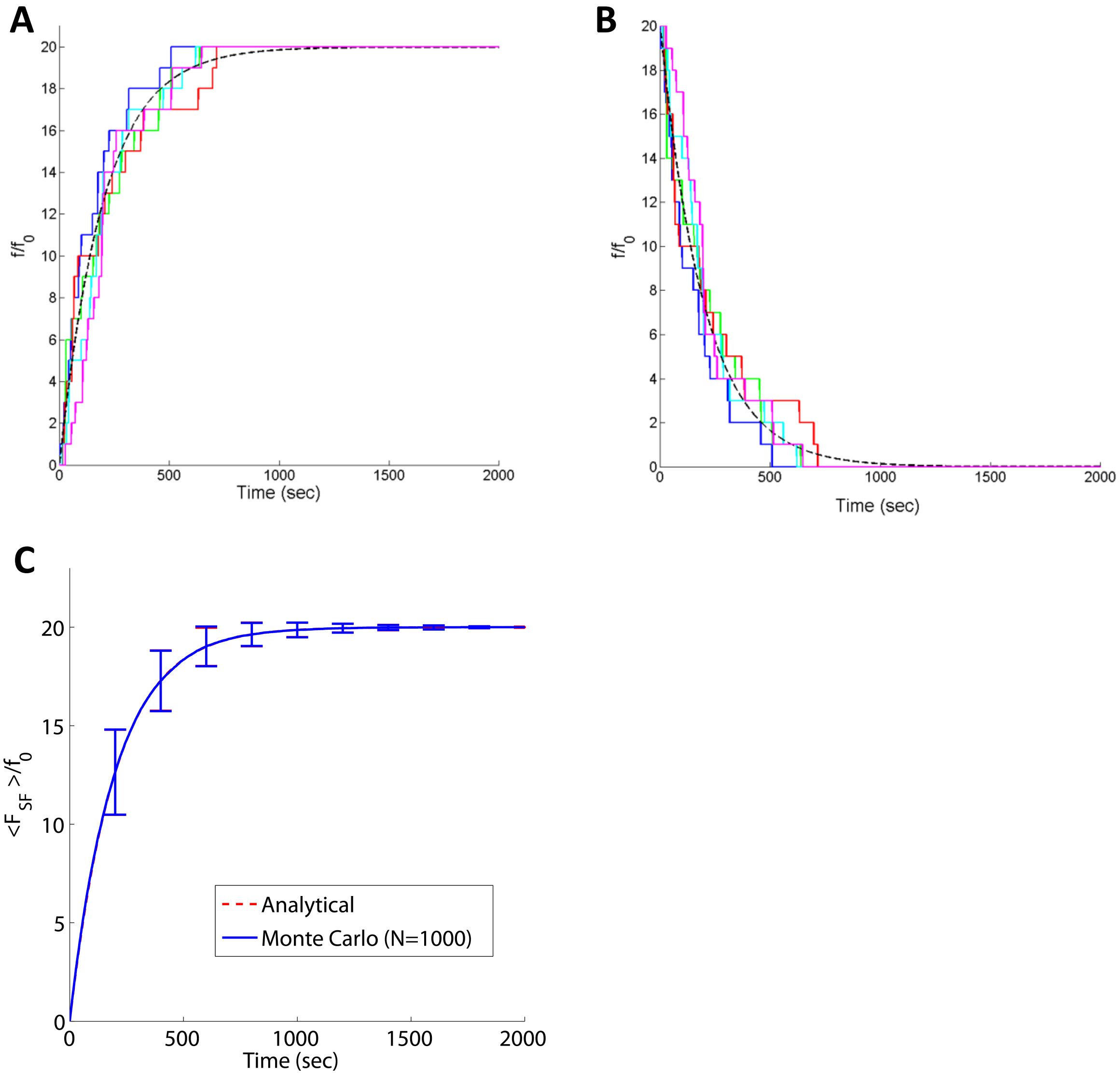


Figure 6.6. Model predictions for stress fiber contraction and relaxation. Analytical predictions are compared with five numerical simulations of stress fiber contraction (**A**) and relaxation (**B**). Each numerical simulation consisted of drawing N_S=20 exponentially distributed variables with rate parameter $\lambda_{A/D}$=1/200 sec^{-1} corresponding to the times that individual sarcomeres assemble (A) or disassemble (B). The analytical predictions for the average force during stress fiber contraction and relaxation are given in Eq. 5 and Eq. 8, respectively. **C.** Monte Carlo simulation of stress fiber contraction, averaging 1000 numerical time traces of the stochastic force (Eq. 6). This ensemble average is compared with the analytical prediction given in Eq. 5. The error bars are standard deviations (both analytical and from the simulation). Analytical predictions and simulation overlap.

6.4 Estimation of stress fiber assembly/disassembly rate

Let t_i be the time for the *i*-th sarcomere to assemble/dissemble (i=1..N_S), where t=0 is the time of perturbation. We wish to estimate the rate constant $\lambda = \tau^{-1}$. According to our model, $t_i \sim \exp(\lambda)$, therefore $\tau = Et_i$. The maximum likelihood estimator of τ for a given profile is therefore $\hat{\tau} = \frac{1}{N_S}\sum_{i=1}^{N_S} t_i$ (see proof in section 6.10). However, since the exact time of perturbation is not known due to the gradual increase in CD concentration (and there can also be an intrinsic unknown delay in the response of the cell) we cannot estimate the times t_i accurately. Alternatively, if we time order the transition times such that $t_1 < t_2 < \ldots < t_{N_S}$ then it follows that $T_k = t_k - t_m$ for $k>m$ are independent and exponentially distributed random variables with the same rate λ (see proof in section 6.10). The maximum likelihood estimator of τ for m=1 is therefore given by

(9A)
$$\hat{\tau}_{ML} = \frac{1}{N_S - 1}\sum_{k=2}^{N_S} T_k .$$

and standard error in the estimation of the mean is given by

(9B)
$$S.E. = \sqrt{\operatorname{var}\left(\frac{1}{N_S - 1}\sum_{k=2}^{N_S} T_k\right)} = \frac{\sigma_T}{\sqrt{N_S - 1}} = \frac{\tau}{\sqrt{N_S - 1}}.$$

where for an exponential distribution $\sigma_T = \tau$. A Monte Carlo simulation of Eq. 9 is presented in Fig. 6.12 (section 6.10).

6.5 A statistical test for the distribution of step times

We wish to test the hypothesis that the times $T_k = t_k - t_1$ for a given profile are exponentially distributed with rate $\hat{\lambda} = 1/\hat{\tau}_{ML}$ (estimated for each profile independently). We divide the time axis into three equiprobable regions: [0,0.4τ], (0.4 τ,1.1τ],(1.1τ,∞). The probability for each region is given by:

$$p_1 = \int_0^{0.4\tau} dt\lambda e^{-\lambda t} = 1 - e^{-0.4} = 0.3297$$

$$p_2 = \int_{0.4\tau}^{1.1\tau} dt\lambda e^{-\lambda t} = e^{-0.4} - e^{-1.1} = 0.3374$$

$$p_3 = \int_{1.1\tau}^{\infty} dt\lambda e^{-\lambda t} = e^{-1.1} = 0.3329$$

We then calculate the Pearson's chi-square statistic $X^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}$, where O_i and E_i represent the number of observed and expected counts for each bin, respectively, and where for each curve the rate constant was estimated based on Eq. 9. X^2 has a χ^2 distribution with *m-k-1*=3-1-1=1 degrees of freedom, where in this case *m*=3 bins were used and *k*=1 parameters were estimated. Table 6.1 shows the results for three curves, two post-CD contraction events (C1, C2), and one C- induced relaxation (R1)[1]. We see that the *p*-values are high suggesting that statistically the transition times are consistent with an exponential distribution. Other test statistics such as the likelihood ratio test

[1] The cell was cycled between CD and conditioned media 3 times, producing 3 relaxation curves (labeled R1–3) and 3 contraction curves (labeled C1–3). However, due to experimental errors, the CD concentration on one of the cycles was less by an unknown amount; those two cycles were omitted from all of the kinetic model analysis. The Pearson test requires at least 15 step events, thus the Pearson test could be applied to only 3 of the 4 remaining curves (C1, C2, and R1). Curves C1–C3 and R1, R3 are presented in Fig. 6.10 and Fig. 6.11.

statistic $-2\log\Lambda = 2\sum_{i=1}^{3} O_i \log\left(\frac{O_i}{E_i}\right)$ and the Power divergence test statistic with $\lambda = 2/3 : \frac{9}{5}\sum_{i=1}^{3} O_i \left[\left(\frac{O_i}{E_i}\right)^{\frac{2}{3}} - 1\right]$ (expected to give good results for small sample sizes), yield p-values identical to the Pearson p-value to within 3% (*7*) .

Table 6.1. The Pearson's chi-square test statistic for the step times T_i given a null hypothesis of an exponential distribution

	Contraction curve C1 (*n*=16) $\hat{\tau}$ =106.53±33.91 S.E.		**Contraction curve C2** (*n*=18) $\hat{\tau}$ =77.47 ±16.73 S.E.		**Relaxation curve R1** (*n*=18) $\hat{\tau}$ =231.15±44.94 S.E.	
Interval	**Observed**	**Expected**	**Observed**	**Expected**	**Observed**	**Expected**
[0,0.4τ]	7	5.27	5	5.93	5	5.93
(0.4 τ,1.1τ]	4	5.4	8	6.07	6	6.07
(1.1τ,∞)	5	5.33	5	5.99	7	5.99
X^2	0.947		0.922		0.318	
p-value	0.331		0.337		0.573	

6.6 Comparing the stochastic model to experimental data

Here we compare the measured force traces C1, C2 and R1, R3 with the stochastic model predictions. For each of these four profiles, τ, f_0, and N_S were estimated and the average force was calculated based on Eq. 5 or Eq. 8 depending on the scenario. In Fig. 6.7 we plot the measured force traces and compare them with the profiles predicted by the model. Note that N_S was allowed to change from profile to profile since the number of disassembled or assembled sarcomeres at the beginning of the perturbation is uncertain, and in some cases (curve R3) the profile is given from the middle. Fig. 6.7 shows that the measured traces follow the predicted ensemble averages and are contained within two standard deviations. Note that the model predicts that any single trace of the force is not exponential (e.g., Fig. 6.6), however the ensemble average of many traces does follow an exponential profile. For this reason deviations between the measured force traces and the predicted ensemble average are expected. In Fig. 6.8 we collapse the transition times for all contraction (or relaxation curves) by scaling the time axis $t \to t/\tau$ and scaling the force axis $F \to F/f_{\max}$. Both contraction curves and both relaxation curves appear to collapse to the same corresponding profiles, as predicted by the model.

6.7 Analysis of step durations

We wish to find the waiting time for the *n*-th sarcomere to assemble (or disassemble) given that *n*-1 sarcomeres have already assembled (or disassembled). For concreteness we will examine the case of assembly. The case for disassembly is analogous. After *n*-1 sarcomeres have assembled, there are $N_S - n + 1$ sarcomeres remaining to assemble. The next event will therefore occur after a time $\mathbf{\Delta t_n}$, which is the minimum time of N_S-n+1 i.i.d. exponential variables. Since it is known that the minimum of *N* i.i.d. exponentially distributed random variables with rate λ is also exponentially distributed with rate *N*λ, then the time the next event will occur, $\mathbf{\Delta t_n}$, will be exponentially distributed with a rate $\lambda\left(N_S - n + 1\right)$. Intuitively, the larger the pool of remaining assembled sarcomeres, the less time one needs to wait until one of the sarcomeres from this pool will fail. Thus initially, when the pool of remaining assembled sarcomeres is large, steps occur in rapid succession, and, with time, as the pool of assembled sarcomeres diminishes, steps also increase in duration.

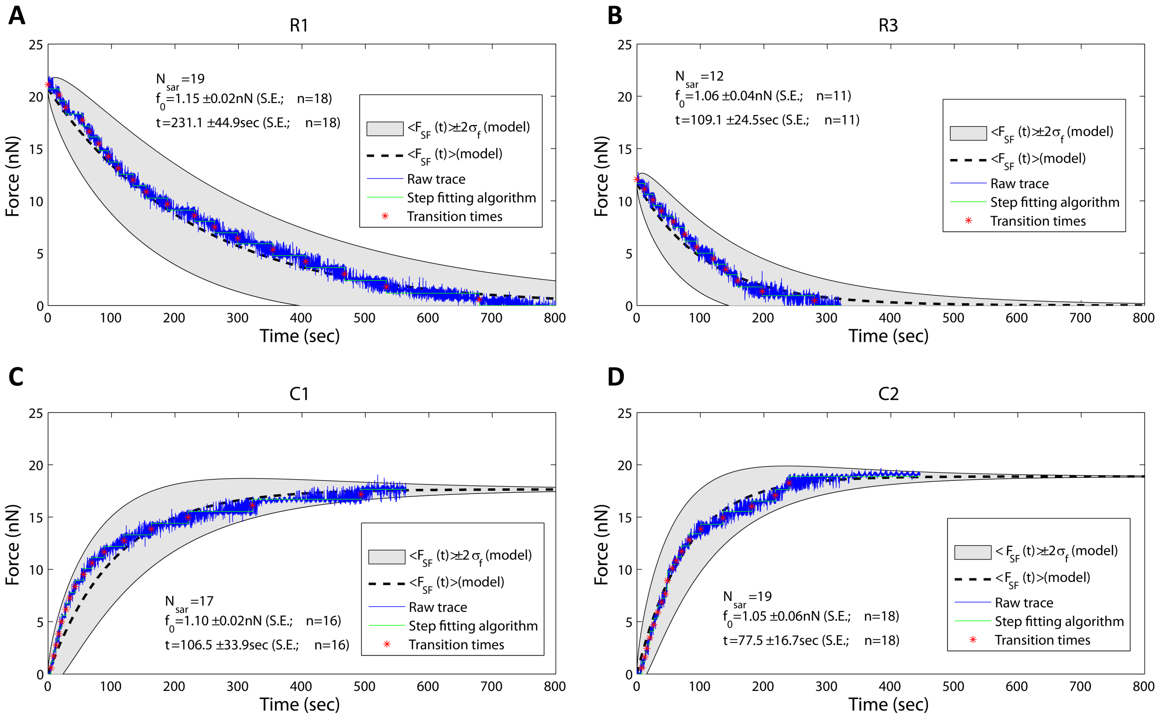


Figure 6.7. Force/time relations predicted by the stochastic model versus experimental data. Model predictions for the average force versus time measurement of a fibroblast perturbed by CD (Eq. 8) (**A & B**) or incubated in the recovery medium (Eq. 5) (**C & D**). Data is blue, output from step fitting algorithm is red, the discrete step events used to count N_S and calculate f_0 and τ for the kinetic model are yellow. For each profile, τ was estimated based on Eq. 9, f_0 was estimated by averaging the step size of the force predicted by the step fitting algorithm and the number of sarcomeres, N_{sar}, was estimated by the step fitting algorithm as the number of detected steps (note that since the origin was set at the first transition, $N_S = N_{sar} - 1$). Based on these three parameters the expected mean of the force versus time was calculated (dashed curve). This theoretical curve represents the average of many individual stochastic profiles. The shaded area represents the area bounded by the mean plus and minus two standard deviations. C1 and C2 are two contraction profiles with well defined initial conditions and R1 and R3 are two relaxation profiles with well defined initial conditions (see section 6.9).

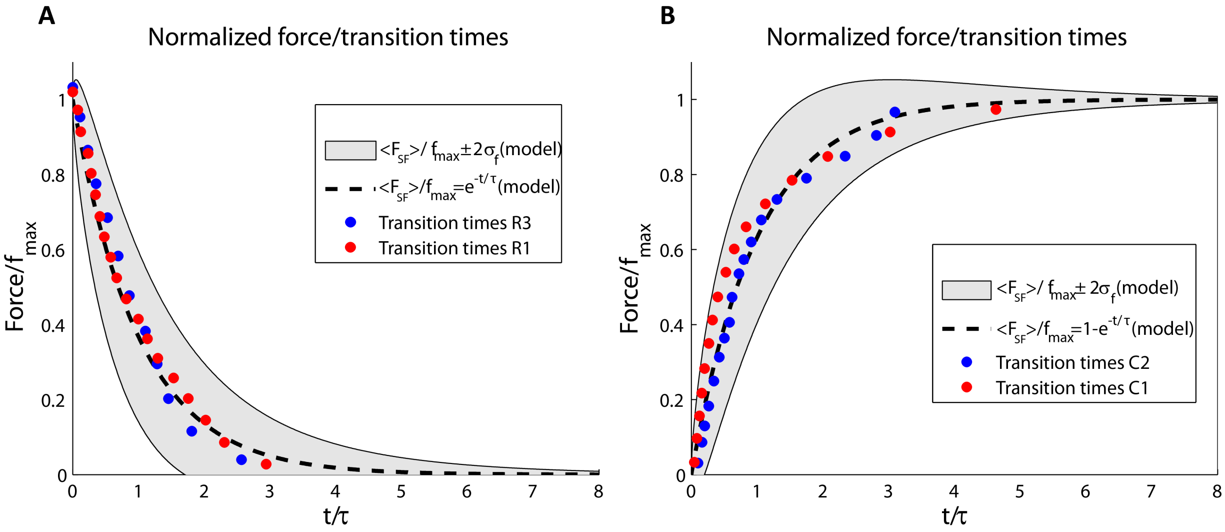


Figure 6.8. Rescaled force/time traces for contraction and relaxation profiles. Force versus step times for contraction (**A**) and relaxation (**B**) scenarios were collapsed by scaling the time axis $t \longrightarrow t/\tau$ and the force axis $F \to F / f_{\max}$ (where $f_{\max} = f_0 N_S$) for each profile. For Methods see caption of Fig. 6.7. The shaded area represents area bounded by the mean plus and minus two standard deviations assuming N_S=18 (giving a lower bound on deviations).

Another way to see this is the following: the probability that a single sarcomere will assemble in the next t seconds is (see Eq. 2) $p(0<t'<t)=1-e^{-\lambda_A t}$. The probability that it will not assemble is therefore $1-p(0<t'<t)=e^{-\lambda_A t}$. The probability that none of the remaining N_S-n+1 sarcomeres assemble during the interval $0<t'<t$ is $\left(e^{-\lambda_A t}\right)^{N_S-n+1}$. Therefore the probability that at least one sarcomere will assemble during the interval $0<t'<t$ is $F(t)=1-e^{-\lambda_A t(N_S-n+1)}$. Hence the probability density function for the time to wait until at least one sarcomere assembles after n-1 sarcomeres have already assembled is $F'(t)=\lambda_A\left(N_S-n+1\right)e^{-\lambda_A(N_S-n+1)t}$, or $\mathbf{\Delta t_n} \sim \exp\left(\lambda_A\left(N_S-n+1\right)\right)$.

The average waiting time for the *n*-th sarcomere to assemble/disassemble given that *n*-1 sarcomeres have already assembled (or disassembled) is therefore just the expectation value of an exponential distribution with a rate $\lambda_A \left(N_S - n + 1\right)$:

(10) $$E\mathbf{\Delta t_n} = \left[\lambda\left(N_S - n + 1\right)\right]^{-1}.$$

where *n*=1..N_S. The standard deviation of $\mathbf{\Delta t_n}$ is given by the exponential standard deviation: $\sqrt{\mathrm{var}(\mathbf{\Delta t_n})} = \left[\lambda\left(N_S - n + 1\right)\right]^{-1}$. Note that the inverse of $E\mathbf{\Delta t_n}$ is linear in *n*. In Fig. 6.9 we plot Eq. 10, as well as a Monte Carlo simulation for the waiting times. As can be seen from this figure, the numerical simulation converges to the analytically derived expectations value. Standard deviations converge as well (data not shown).

Testing prediction of duration of force steps against experimental data

Waiting times between force steps were estimated for each of the contraction and relaxation experimental curves based on the times extracted by the step-fitting algorithm. In Fig 6.10 we plot the inverse of the experimental waiting times versus the prediction of the stochastic model for these rates (Eq. 10). τ in Eq. 10 was estimated for each profile based on Eq. 9 and N_S was estimated for each profile as the number of steps detected by the step-fitting algorithm. The correlation between the data points and the model prediction was measured by Pearson correlation coefficient to be *ρ*= 0.72 to 0.95, suggesting that the model predictions were strongly reflected in the data. Note that we anticipate a high level of noise since the standard deviation equals the predicted rates (see above).

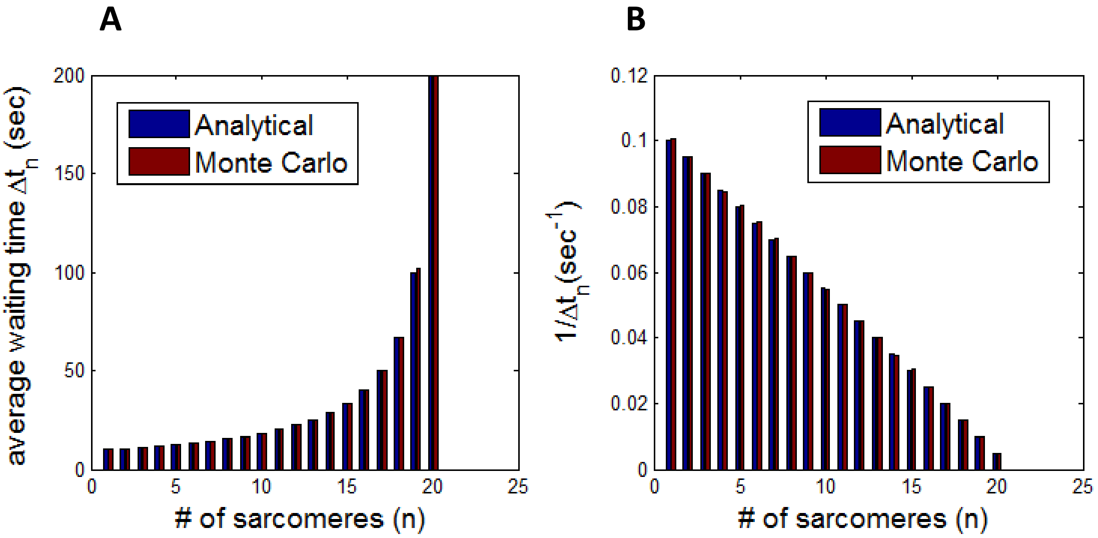


Figure 6.9. Force step durations predicted by the stochastic model compared with a Monte Carlo simulation. A. Each bar represents the average time $E\mathbf{\Delta t_n}$ until the next sarcomere fails, given that *n* sarcomeres have already failed, thus resulting in *n* failed sarcomeres. **B.** Predictions for $1/E\mathbf{\Delta t_n}$. Blue bars show analytical predictions based on Eq. 10. Red bars show Monte Carlo simulation results averaging 10^4 repetitions. Each iteration in the Monte Carlo simulation consisted of drawing N_S=20 exponentially distributed times with $\lambda = 1/200\ \text{sec}^{-1}$ and calculating the waiting times between temporally adjacent events. Times were then averaged for all 10^4 iterations.

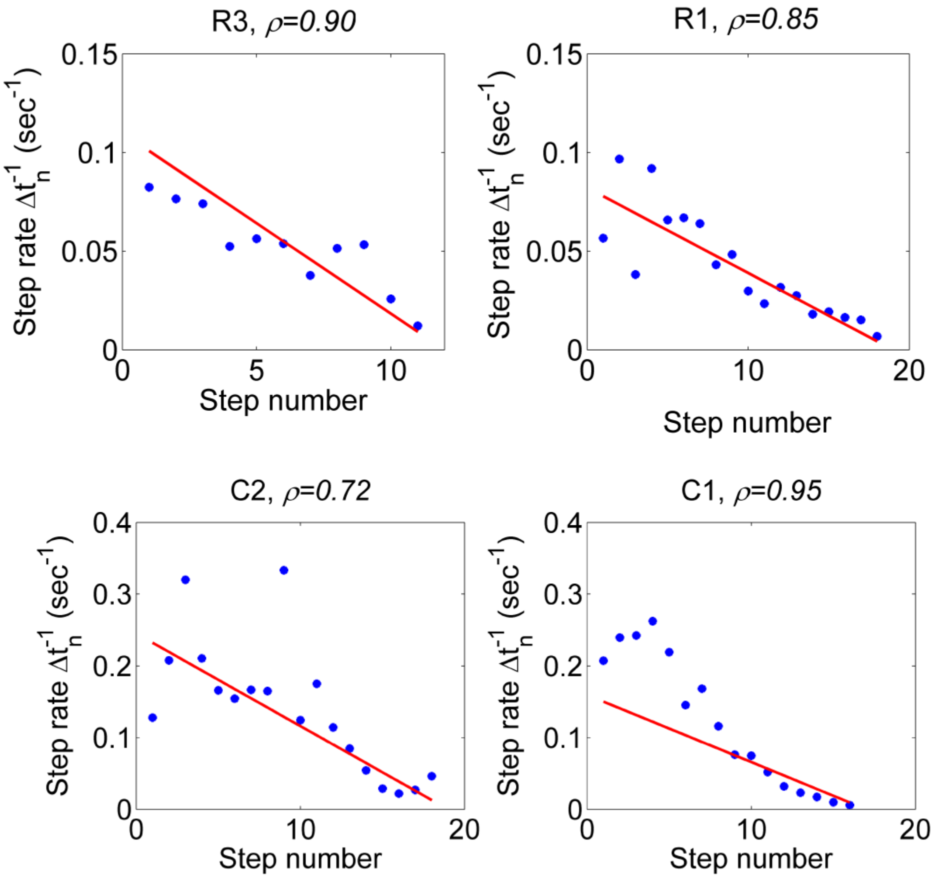


Figure 6.10. Stochastic model prediction of force step durations versus experimental data. The stochastic model prediction for the inverse of the force step durations (red curve) was calculated based on Eq. 10 with τ estimated based on Eq. 9. The experimental data points (blue dots) were calculated as the inverse of the time difference between adjacent steps extracted by the step-fitting algorithm. N_S, the number of observed steps, was determined by the step-fitting algorithm. ρ is the Pearson correlation coefficient measuring the strength of the correlation. C1 and C2 are two contraction profiles with well defined initial conditions and R1 and R3 are two relaxation profiles with well defined initial conditions (see section 6.9).

6.8 Conclusions

We have suggested a simple stochastic model for stress fiber contraction and relaxation that is capable or explaining both qualitatively and quantitatively the response of a single fibroblast cell that was subjected to cyclic perturbation of CD and recovery medium. The model we propose predicts stochastic force/time profiles that are qualitatively similar to the curves observed in the experiment: both predicted and observed curves exhibit an exponential-like profile with step times of increasing length. When rescaled, observed profiles collapse to similar curves that qualitatively follow the ensemble average predicted by the model. Finally, the step onset times are statistically consistent with an exponential distribution and the inverse of the step durations exhibit a high degree of correlation with the linear response predicted by the model. To further substantiate the proposed model it is recommended to repeat this experiment at least one more time, as currently, all data was taken from a single session observing a single cell. An effort to reproduce this data is currently underway.

6.9 Appendix A — Selection criteria of profiles

Figure 6.11 shows the complete measurement of force versus time for the fibroblast cell under investigation, subject to various perturbations. In our analysis we choose to focus on profiles that have the following characteristics: (1) the cell was subject to well-defined perturbations: growth medium (blue) or cytochalasin D (pink), and (2) the initial condition of the cell was well defined, approximating either full contraction or full relaxation of the stress fiber. The profiles that conform to these conditions are: relaxation profiles R3 and R1, which are preceded by contraction profiles C4 and C2, respectively, that leave the stress fibers in a fully contracted condition; contraction profiles C3, C2, and C1, which are preceded by relaxation profiles R3, R2, R1, respectively, that leave the stress fibers in a fully relaxed condition. Profile R2 was not considered for analysis because the CD perturbation was achieved here passively via diffusion, attaining an unknown, yet lower concentration of CD. Profile C3 was rejected due to the back-stepping (this curve is cut short by the step-fitting algorithm, leaving only ~150 sec of data). Since the curve is cut short, estimation of τ using Eq. 9 would be biased. Profile C4 is not considered for analysis because the cell did not appear to respond to the CD perturbation in the preceding period, and therefore the condition of the cell at the onset of C4 was not well defined.

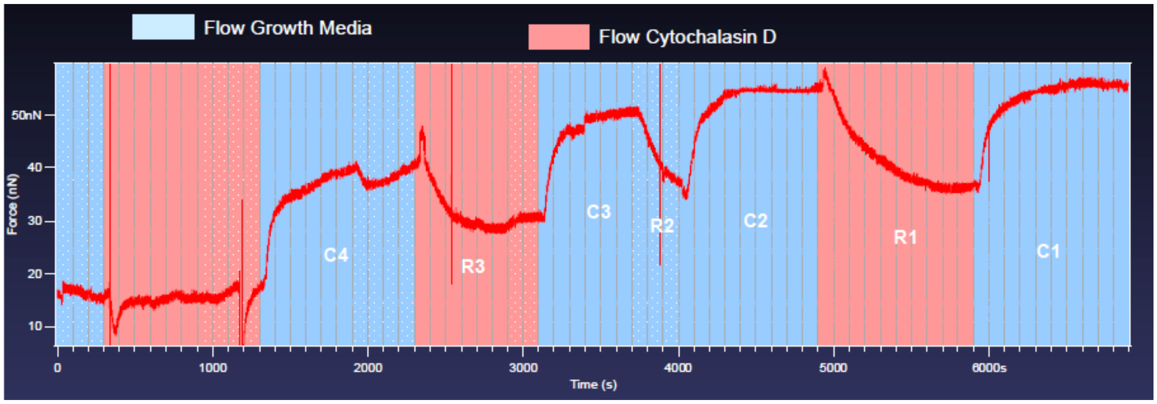


Figure 6.11. Complete force versus time measurement of a single cell. Blue areas represent exposure to growth media, pink areas represent exposure to cytochalasin D, white dots indicate where flow was stopped.

6.10 Appendix B — Time-ordered exponential variables

Lemma 1

Let X_i be N iid exponential random variables (RVs) with rate λ ($1 \le i \le N$). We define

$T_{(1)},\ldots,T_{(m)}$ to be the first m time-ordered RVs (m<N). We wish to show that for $X_i > T_{(m)}$,

$Z_j = X_i - T_{(m)}$ ($j = m+1 .. N_S$) are also iid exponential RVs with rate λ, that is

$$f(z_{m+1},\ldots,z_N) = \lambda e^{-\lambda z_{m+1}} \cdot \lambda e^{-\lambda z_{m+2}} \cdot \ldots \cdot \lambda e^{-\lambda z_N} \quad .$$

Proof:

We define Y_i ($i > m$) to be the N-m RVs that satisfy $X_i > T_{(m)}$

0 $\quad t_{(1)}$ $\quad \ldots$ $\quad t_{(m)}$ $\quad y_{m+1}$ $\quad \ldots$ $\quad y_N$

|------------|--------------------------|------------|----------------|-----› .

$dt_{(1)}$ $\quad \ldots$ $\quad dt_{(m)}$ $\quad dy_{m+1}$ $\quad \ldots$ $\quad dy_N$

The joint pdf of $T_{(i)}$ and Y_i is given by

$$\begin{aligned}&P(t_{(1)},...,t_{(m)},y_{m+1},...,y_N)=\\&=P\Big(one\ \ X_i\in dt_{(1)},...,one\ X_i\in dt_{(m)},one\ X_i\in dy_{m+1},...,one\ X_i\in dy_N\Big)=\\&=N!\cdot\frac{1}{(N-m)!}\cdot\underbrace{\Big(\lambda e^{-\lambda t_{(1)}}dt_{(1)}\cdot...\cdot\lambda e^{-\lambda t_{(m)}}dt_{(m)}\Big)}_{m}\cdot\underbrace{\Big(\lambda e^{-\lambda y_{m+1}}dy_{m+1}\cdot...\cdot\lambda e^{-\lambda y_N}dy_N\Big)}_{N-m}\end{aligned}$$

where the normalization factor stems from the fact that there are *N!* ways of ordering *N* variables, however since the Y_i's are not time ordered we need to remove the degeneracy of $(N-m)!$. Thus

$$f(t_{(1)},...,t_{(m)},y_{m+1},...,y_N)=\underbrace{(N-1)\cdot...\cdot(N-m+1)}_{m}\cdot\Big(\lambda e^{-\lambda t_{(1)}}\cdot...\cdot\lambda e^{-\lambda t_{(m)}}\Big)\cdot\Big(\lambda e^{-\lambda y_{m+1}}\cdot...\cdot\lambda e^{-\lambda y_N}\Big).$$

Now,

$$\begin{aligned}&f(t_{(1)},...,t_{(m)},y_{m+1},...,y_N)=\\&=f(t_{(1)},...,t_{(m)},y_{m+1}=z_{m+1}+t_{(m)},...,y_N=z_N+t_{(m)})=\\&=N\cdot...\cdot(N-m+1)\cdot\Big(\lambda e^{-\lambda t_{(1)}}\cdot...\cdot\lambda e^{-\lambda t_{(m)}}\Big)\cdot\underbrace{\Big(\lambda e^{-\lambda\left(z_{m+1}+t_{(m)}\right)}\cdot...\cdot\lambda e^{-\lambda\left(z_N+t_{(m)}\right)}\Big)}_{N-m}=\\&=N\cdot...\cdot(N-m+1)\cdot\Big(\lambda e^{-\lambda t_{(1)}}\cdot...\cdot\lambda e^{-\lambda t_{(m-1)}}\cdot\lambda e^{-\lambda t_{(m)}(N-m+1)}\Big)\cdot\Big(\lambda e^{-\lambda z_{m+1}}\cdot...\cdot\lambda e^{-\lambda z_N}\Big).\end{aligned}$$

Intergrating out $t_{(m)}$:

$$\begin{aligned}&f(t_{(1)},...,t_{(m-1)},z_{m+1},...,z_N)=\int_{t_{(m-1)}}^{\infty}d\,t_{(m)}f(t_{(1)},...,t_{(m)},z_{m+1},...,z_N)=\\&=N\cdot...\cdot(N-m+1)\cdot\underbrace{\Big(\lambda e^{-\lambda t_{(1)}}\cdot...\cdot\lambda e^{-\lambda t_{(m-1)}}\Big)}_{m-1}\cdot\Big(\lambda e^{-\lambda z_{m+1}}\cdot...\cdot\lambda e^{-\lambda z_N}\Big)\cdot\int_{t_{(m)}=t_{(m-1)}}^{\infty}dt_{(m)}\lambda e^{-\lambda t_{(m)}(N-m+1)}=\\&=\underbrace{N...\cdot(N-m+2)}_{m-1}\cdot\underbrace{\Big(\lambda e^{-\lambda t_{(1)}}\cdot...\cdot\lambda e^{-\lambda t_{(m-1)}(N-m+2)}\Big)}_{m-1}\cdot\Big(\lambda e^{-\lambda z_{m+1}}\cdot...\cdot\lambda e^{-\lambda z_N}\Big).\end{aligned}$$

We can now recursively integrate out $t_{(m-1)}$ and so on:

$$f(t_{(1)},\ldots,t_{(m-2)},z_{m+1},\ldots,z_N)=$$

$$=N\cdot\ldots\cdot(N-m+2)\cdot\underbrace{\left(\lambda e^{-\lambda t_{(1)}}\cdot\ldots\cdot\lambda e^{-\lambda t_{(m-2)}}\right)}_{m-2}\cdot\left(\lambda e^{-\lambda z_{m+1}}\cdot\ldots\cdot\lambda e^{-\lambda z_N}\right)\cdot\int_{t_{(m-1)}=t_{(m-2)}}^{\infty}dt_{(m-1)}\lambda e^{-\lambda t_{(m-1)}(N-m+2)}=$$

$$=\underbrace{N\cdot\ldots\cdot(N-m+3)}_{m-2}\cdot\underbrace{\left(\lambda e^{-\lambda t_{(1)}}\cdot\ldots\cdot\lambda e^{-\lambda t_{(m-2)}(N-m+3)}\right)}_{m-2}\cdot\left(\lambda e^{-\lambda z_{m+1}}\cdot\ldots\cdot\lambda e^{-\lambda z_N}\right).$$

Repeating this process *m*-2 more times, (integrating $t_{(1)}$ from 0 to ∞) for $t_{(m-2)}\ldots t_{(1)}$, we obtain $f(z_{m+1},\ldots,z_N)=\lambda e^{-\lambda z_{m+1}}\cdot\ldots\cdot\lambda e^{-\lambda z_N}$ $\quad\therefore$

Simulations supporting theory

In Fig. 6.12 we show the results of a Monte Carlo simulation for the case of *N*=20 and *m*=15, showing that $\{Z_i\}_{i=16..20}$ have properties of an exponential distribution (mean, standard deviation, and probability density function). In Fig. 6.13 we calculate the average “force” step profile and standard deviation of *N*=20 exponential independent RVs when taking into account only the last fifteen RVs measured with respect to the fifth time-ordered RV. This simulation is a precise test of the way in which the experimental data is analyzed. We see that simulations perfectly agree with theory.

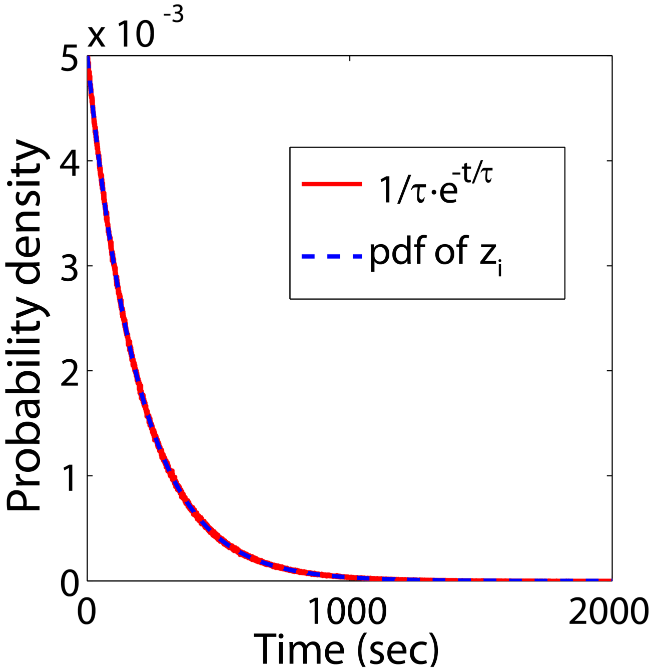


Figure 6.12. Monte Carlo simulation demonstrating lemma 1. This Monte Carlo simulation consisted of 10^7 iterations where in each iteration we drew N=20 exponentially distributed numbers (τ=200 sec) and time ordered them to obtain t_i. We then calculated $z_i = t_i - t_{15}$ for $16 \le i \le 20$ (five RVs in total, corresponding to the case where m=15) and wished to see if z_i indeed behave as i.i.d. exponential RVs with a rate parameter 1/200 sec^{-1} .We therefore calculated the sample mean of z_i for each iteration: $\hat{\mu} = \frac{1}{5}\sum_{i=16}^{20} z_i$ and estimated the mean and standard deviation of $\hat{\mu}$ over all iterations. Based on Eq. 9A and 9B (generalizing these equations to an arbitrary m: $\hat{\tau}_{ML} = \frac{1}{N_S - m}\sum_{k=1+m}^{N_S} T_k, S.E. = \sqrt{\text{var}(\hat{\tau}_{ML})} = \tau \big/ \sqrt{N_S - m}$) we should find that $\hat{\mu}$ converges to τ with a S.E. of $\tau \big/ \sqrt{5}$, i.e.: $E\hat{\mu} = 200 \pm 200 \big/ \sqrt{5} = 200 \pm 89.4427$. The simulation yielded 199.98±89.4457 in precise agreement with theory. We also estimated the pdf of z_i by collecting all z_i ($16 \le i \le 20$) and calculating the empirical pdf (red) versus the theoretical pdf (blue). Both curves indeed overlapped.

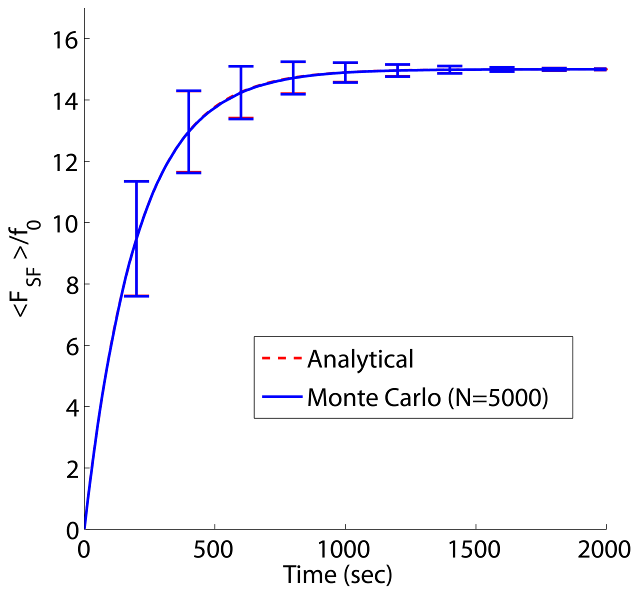


Figure 6.13. Monte Carlo simulation of stress fiber contraction. In each iteration N_S=20 exponentially distributed RVs X_i with λ =1/200 sec^{-1} are drawn. We then time order these 20 RVs, find the fifth time ordered RV, $X_{(5)}$, and form 15 new RVs: $T_j = X_{(j)} - X_{(5)}, j > 5$. Thus T_j represent the times until sarcomeres assemble, measured with respect to the time that the fifth sarcomere assembled (thereby generalizing the case discussed in Fig. 6.6C for m > 0). We then construct the stochastic force time trace (Eq. 6), where the step functions occur at times $t = T_j$. Finally we average 5000 of these force traces. We superimpose the theoretical prediction based on Eq. 5 (assuming 15 steps). The error bars are the theoretical and simulated standard deviations. We see that analytical predictions and simulation overlap.

Lemma 2

Let T_i be N i.i.d. exponential RVs with rate λ ($1 \le i \le N$). The ML estimator of τ=1/ λ is

$$\hat{\tau}_{ML} = \tfrac{1}{N}\sum_{i=1}^{N} t_i \, .$$

Proof:

$$f(t_1,t_2,...,t_N) = \lambda e^{-\lambda t_1} \cdot \lambda e^{-\lambda t_2} \cdot ... \cdot \lambda e^{-\lambda t_N} = \lambda^N e^{-\lambda(t_1+...t_N)}$$

$$\log\left(f(t_1,t_2,...,t_N)\right) = N \log \lambda - \lambda(t_1+...t_N) = -N \log \tau - \frac{1}{\tau}(t_1+...t_N)$$

$$\frac{\partial \log\left(f(t_1,t_2,...,t_N)\right)}{\partial \tau} = \frac{-N}{\tau} + \frac{1}{\tau^2}(t_1+...t_N)$$

$$\frac{-N}{\hat{\tau}_{ML}} + \frac{1}{\hat{\tau}_{ML}{}^2}(t_1+...t_N) = 0$$

$$\hat{\tau}_{ML} = \frac{1}{N}(t_1+...t_N)$$

$$\therefore$$

6.11 References

1. P. Naumanen, P. Lappalainen, P. Hotulainen, *J. Microsc.* **231**, 446 (2008).

2. S. Pellegrin, H. Mellor, *J. Cell Sci.* **120**, 3491 (2007).

3. J. A. Cooper, *Journal of Cell Biology* **105**, 1473 (1987).

4. P. Sampath, T. D. Pollard, *Biochemistry* **30**, 1973 (Feb, 1991).

5. J. R. Peterson, T. J. Mitchison, *Chemistry & Biology* **9**, 1275 (Dec, 2002).

6. J. W. J. Kerssemakers *et al.*, *Nature* **442**, 709 (Aug, 2006).

7. T. R. C. Reed, N. A. C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, Ed., Springer Series in Statistics (Springer, 1988), p. 211.