UNIVERSITY OF CALIFORNIA SAN DIEGO

Unraveling Darwin's entangled bank in coevolution between bacteriophage lambda and its host *Escherichia coli*

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics

by

Animesh Gupta

Committee in charge:

Professor Justin R. Meyer, Chair
Professor Terence Hwa, Co-Chair
Professor Lin Chao
Professor Sergey Kryazhimskiy
Professor Jérémie Palacci

2020

The dissertation of Animesh Gupta is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2020

DEDICATION

To my parents,

whose principles and values have been the greatest source of learning.

EPIGRAPH

We will now discuss in a little more detail

the Struggle for Existence.

Charles Darwin

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Chapter 2 is currently being prepared for submission for publication of the material, as it may appear in *Ecology Letters*, 2021: A. Gupta, S. Peng, C. Y. Leung, J. M. Borin, J. S. Weitz, J. R. Meyer, "Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing".

Chapter 3 has been submitted for publication of the material, with minor edits, as it may appear in *Nature Ecology and Evolution*, 2021: A. Gupta, L. Zaman, H. Strobel, J. Gallie, A. R. Burmeister, B. Kerr, R. Kishony, E. Tamar, J. R. Meyer, "Coevolution promotes innovation through deformations in fitness landscapes".

Chapter 4 is a reprint of material with minor edits, as it appears in: A. Gupta, A. N. Soto, S. J. Medina, K. L. Petrie, J. R. Meyer, "Bacteriophage lambda overcomes a perturbation in its host–viral genetic network through mutualism and evolution of life history traits," *Evolution*, vol. 74, no. 4, pp. 764-774, 2020.

The author of the dissertation was the primary author of these papers.

VITA

2014	B. Tech. and M. Tech. (Dual Degree) in Engineering Physics with specialization in Nanoscience, Indian Institute of Technology Bombay, India
2020	Ph. D. in Physics, University of California San Diego

PUBLICATIONS

A. Gupta, A. N. Soto, S. J. Medina, K. L. Petrie, J. R. Meyer, "Bacteriophage lambda overcomes a perturbation in its host–viral genetic network through mutualism and evolution of life history traits," *Evolution*, vol. 74, no. 4, pp. 764-774, 2019.

W. N. Chaudhry, M. Pleška, N. N. Shah, H. Weiss, I. C. McCall, J. R. Meyer, A. Gupta, C. C. Guet, B. R. Levin, "Leaky resistance and the conditions for the existence of lytic bacteriophage," *PLOS Biology*, vol. 16, no. 8, pp. e2005971, 2018.

J. R. Meyer, D. T. Dobias, S. J. Medina, L. Servilio, A. Gupta, R. E. Lenski, "Ecological speciation of bacteriophage lambda in allopatry and sympatry," *Science*, vol. 354, no. 6317, pp. 1301-1304, 2016.

S. S. Soumya, A. Gupta, A. Cugno, L. Deseri, K. Dayal, D. Das, S. Sen, M. M. Inamdar, "Coherent motion of monolayer sheets under confinement and its pathological implications," *PLOS Computational Biology*, vol. 11, no. 12, pp. e1004670, 2015.

A. Gupta, A. Sohane, V. Kohar, K. Murali, S. Sinha, "Noise-free logical stochastic resonance," *Physical Review E*, vol. 84, no. 5, pp. 005201, 2011.

ABSTRACTS

A. Gupta, L. Zaman, J. Gallie, A. Burmeister, B. Kerr, R. Kishony, J. R. Meyer, "Changing fitness landscapes during host-parasite coevolution opens adaptive pathways to a key innovation in phage lambda," *American Society for Microbiology*, San Francisco, CA, USA, June, 2019

A. Gupta, "Changing fitness landscapes during host-parasite coevolution opens adaptive pathways to evolutionary novelty," *American Physical Society*, Los Angeles, CA, USA, March, 2018

A. Gupta, J. R. Meyer, "Changing fitness landscapes during host-parasite coevolution rescues lambda evolution," *Gordon Research Conference*, Andover, NH, USA, July, 2017

A. Gupta, A. N. Soto, K. L. Petrie, J. R. Meyer, "Viral-host interactome evolution to compensate for a host gene deletion," *Frontiers of Innovation Scholars Program*, San Diego, CA, USA, October, 2016

ABSTRACT OF THE DISSERTATION

Unraveling Darwin's entangled bank in coevolution between bacteriophage lambda and its host *Escherichia coli*

by

Animesh Gupta

Doctor of Philosophy in Physics

University of California San Diego, 2020

Professor Justin R. Meyer, Chair
Professor Terence Hwa, Co-Chair

Coevolution between species can drive complex evolutionary and ecological processes of the living world. However, evolution of multiple mutations and constantly changing selective pressures during coevolution creates a tangled web of interactions between species that makes it hard to investigate processes underlying the rich observations. Here we combine next-generation sequencing, quantitative experimental assays, and computational analyses to understand the true complexity coevolution causes by studying coevolution between bacteriophage λ and its host *Escherichia coli*.

In Chapter 2, we test the mechanistic underpinnings of the arms race dynamics model in host-parasite coevolution. By coupling large-scale phenotypic assays with whole genome sequencing, we showed that although λ and *E. coli* engaged in arms race at the phenotypic level, multiple lineages of both species coexisted and shifted in dominance during the coevolution, a feature of the complementary fluctuation selection dynamics model.

Chapter 3 uses fitness landscapes to test the role and extent of coevolution in promoting evolution of a new function. During coevolution, λ evolves to target *E. coli* through a new membrane receptor, OmpF, by fixing multiple mutations in its host-recognition protein J. By measuring λ's fitness landscapes at two stages of coevolution: before and after *E. coli* evolved resistance, we show that evolution of resistance in *E. coli* deformed λ's landscape such that it opened adaptive pathways for λ to evolve the new function. Multiple replays of the coevolution experiment with different starting conditions and a detailed examination of a coevolving laboratory population further confirmed that λ required both ancestral and resistant hosts from coevolution to evolve the OmpF-function.

Lastly, we studied how perturbations in the host-parasite interactome could lead to adaptions at the population level. When we deleted *E. coli*'s *dnaJ* gene that λ requires to initiate DNA replication, λ evolved mutations in genes unrelated to *dnaJ*. λ adapted by improving its adsorption rate and lysis timing, and evolving intracellular mutualism.

Altogether, this dissertation offers insights into mechanisms that structure ecological networks, demonstrates that coevolution of multiple species can promote innovation, and shows how parasites can adapt in unintuitive ways to counter genetic deficiencies in their host.

Chapter 1

Introduction

In nature, organisms seldom evolve in isolation. Instead, they evolve while interacting with other cohabitating species. Darwin noted the complexity of this species-interaction network in the final pages of *On the Origin of Species*, "It is interesting to contemplate an entangled bank. . . ". He envisioned the bank to be entangled because of the numerous ways different species in an ecological community could interact—from mutually beneficial interactions like cooperation and symbiosis [10, 49] to antagonistic interactions like parasitic and competition [139, 78]. A recent example of convoluted interactions that span multiple species is the interplay between human brains and the different microbes that inhabit humans' gut [98]. While different types of stressors to brains can affect composition and activity of gut-microbiota, researchers are now discovering new mechanisms by which gut microbes can affect anxiety and cognition. Even in simple two-species antagonistic interactions, selective pressures can result in complex resolutions such as adaptive radiation where species diversify into multiple ecologically distinct species [105]. The challenge for us then is to disentangle this bank in a methodical way that informs us about the underlying evolutionary processes.

Experimental coevolution in the laboratory provides us the best opportunity to do

so. There are two major reasons for this. First, it is easy to perturb the interaction-network between species in a controlled laboratory setting. These perturbations can vary from abiotic changes in the system such as varying the environmental conditions [94, 107], to more biotic ones such as changing mutation rates of species, genetically engineering mutations in their genomes or introducing a new species in the community [101, 59, 141]. By observing how these perturbations change the evolutionary fate of species, we can test longstanding coevolutionary hypotheses and reveal novel facets of adaptation.

Second, in addition to being amenable to design experiments, coevolution in lab is shown to produce rich dynamics that can be used to study the evolutionary processes operating in the natural world. For example, coevolution between a bacterium and an archaeon improved mutualistic interactions between them that led to higher growth rates and productivity for both species [72]. This helps shed light on the initial stages of evolution to mutualism. In contrasting experimental systems, antagonistic interactions between parasites and their hosts have been shown to drive biodiversity by accelerating molecular evolution and generating genetic divergence [114, 124, 74]. Host-parasite coevolution, specifically, has been greatly focused upon by evolutionary biologists because it produces complex coevolutionary adaptations [68, 24] and is considered to be a key driver of biological diversity [56, 96]. Hosts are constantly under pressure to evolve new defense-mechanisms to evade their parasites, while parasites are driven to counter these defenses to better infect their hosts. This struggle for survival between them results in different types of adaptive solutions. Hosts have been found to evolve higher number of spontaneous mutations and elevated mutation rates when they are evolved with parasites as compared to evolution without parasites [112]. In other instances, coevolution between a host and a parasite can promote speciation by driving populations to reproductive isolation [26, 102]. Such observations make experimental host-parasite coevolution ideal for researchers to tackle fundamental questions relating to biological evolution.

Here we studied host-parasite coevolution between the virus, bacteriophage λ, and its host bacterium, *Escherichia coli*. Microbes are well-suited for studying coevolution because they have high mutation rates, short generation times, and can be cultured in large numbers in laboratory flasks [81]. This allows us to observe their evolution in real-time. Moreover, microbial cultures can be frozen at any point during the coevolution and revived later to restart evolution, enabling us to answer questions relating to stochasticity and repeatability in evolutionary outcomes [17]. λ and *E. coli*, in particular, are known to engage in an arms race that leads to evolution of complex infection patterns [52] and novel phenotypes in λ [103]. Additionally, extensive literature on λ and *E. coli* [27, 41] allows us to test leading hypotheses in evolutionary biology by building upon this knowledge. In this dissertation, we untangle the processes leading to rich coevolutionary dynamics between λ and *E. coli* by combining next-generation sequencing, quantitative experiments and computational analyses.

The next three chapters of this dissertation study coevolving interactions between λ and *E. coli* in three different ways. Chapter 2 tests different models of coevolution by studying discrepancies between the observed interactions that evolve at phenotypic level and the hypothesized mechanisms that cause them. Chapter 3 explains how coevolving interaction between species can lead to evolution of new functions. And lastly, Chapter 4 shows how perturbations in coevolving interactions at interactome level can lead to novel adaptations seemingly unrelated to the perturbation.

Chapter 2

Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing

2.1 Introduction

Bacteria and their viruses (phages) are the two most abundant and genetically diverse groups of organisms on Earth [137, 32, 133]. Together they form ecological communities with complex networks of interactions whose structures have important implications that extend beyond the microbial world. For example, in oceans, phages are responsible for $10-40\%$ mortality of bacteria. When bacterial cells are lysed by viruses, their biomass is diverted away from the rest of the food web, reducing productivity of macroscopic organisms and the full ecosystem [132, 47, 23]. Thus, if interactions within the network were to substantially change, such as the bacteria evolving resistance to the phage, then productivity may

increase causing rippling effects throughput the marine ecosystem. Bacteria and phages are known to engage in arms races where bacteria evolve phage resistance [84], and phages evolve counter defenses [69]. This dynamic causes continual remodeling of the interaction network, which can impact ecosystem processes, the stability of ecological communities, and maintenance of microbial diversity [60, 19, 24, 117, 131]. Given this, there is a growing interest in characterizing the dynamics of phage-bacterial coevolution and to understand the molecular and ecological mechanisms that shape their coevolving networks [81].

A starting point to study phage-bacterial coevolution is to learn how their interactions change over time [13, 142]. Models of coevolution tend to predict two types of dynamics [4, 149]. The first is arms race dynamics (ARD) where bacteria evolve resistance to an increasing number of phages and phages counter by expanding their host-range. For the bacteria, this leads to an escalation where increasingly resistant bacteria replace their less-resistant predecessors. This causes rapid bacterial genomic divergence and leads to an imbalanced phylogenetic pattern with a single pronounced branch. Similarly, as the phage broadens its host range, the most derived type will supplant its predecessors, resulting in the formation of an imbalanced phylogeny. A second model is based on the evolution of interactions, often termed as *lock and key* interactions. In this model, as coevolution progresses, bacteria gain resistance to contemporary phages, but lose resistance to phages encountered in the past. Likewise, as phages evolve counter-defenses, they lose the ability to infect other host genotypes. Under this model, host genotypes rise and fall according to how abundant their corresponding parasite genotypes are, while parasite genotypes track the abundance of their hosts creating a feedback loop and fluctuating selection dynamics (FSD) [122, 62]. FSD produces negative frequency-dependent selection that promotes diversification and the formation of a balanced phylogeny with multiple branches. ARD and FSD represent two ends of a spectrum of possible coevolutionary dynamics and notably models have been created that span the space between the two end points.

One way to gain insight on whether phage and bacteria coevolve according to ARD or FSD is to quantify their interaction networks (phage-bacterial interaction networks; PBINs) and test for nonrandom nested and modular patterns [52]. Nestedness refers to the pattern of host-ranges where they align one within another like nesting Russian dolls. This pattern is produced by ARD since at each step the phage adds on to its existing host-range, expanding its range in a way that encapsulates its ancestors' ranges. Modularity arises in networks when interactions form into nonrandom clusters such that groups of phages and bacteria tend to interact more often within clusters. This pattern is consistent with FSD where interactions are highly specialized. The majority of PBINs are significantly nested supporting the prominence of ARD; however some PBINs are modular [52], and while rare, specialized interactions have been documented to evolve during phage-bacterial coevolution [66]. Surprisingly, nested patterns at short spatial scales can give way to modular patterns at large spatial scales [54], and ARD has been shown to give way to FSD during advanced stages of coevolution [67, 54]. Together, this variation and scale-dependence provide a glimpse at how complicated phage-bacterial coevolution can be.

While the phenotypic predictions for ARD and FSD are often tested, assessments of the phylogenetic predictions are not as common (ARD: imbalanced, FSD: balanced), and we are unaware of an example where PBINs are coupled with phylogenomic analyses. Given this, we decided to study a model phage-bacterial coevolutionary system; bacteriophage λ and its host, Escherichia coli. When these species are cultured under certain laboratory conditions, they rapidly coevolve with one another [103, 106]. *E. coli* is known to evolve resistance through mutation in the regulatory gene *malT* that suppress the host receptor, the outer-membrane protein LamB. λ counters this by evolving mutations in the binding domain of its host recognition protein that allows it to use a new receptor, OmpF. *E. coli* then evolves additional mutations in OmpF or in an inner-membrane protein complex,

ManXYZ, that transports λ DNA into the cytoplasm [50, 51]. While much is already known about the molecular details of their coevolution, a PBIN or phylogenomic analyses have yet to be performed on this pair.

For this study, we revived cryopreserved samples that were isolated from a coevolution experiment previously reported by Meyer *et al.* [103]. We focused our analyses on a single replicate; the first experimental community in which λ evolved to use OmpF. We isolated a total of 50 bacteria and 44 phages spread across multiple time points. Next, we constructed a PBIN of all combinations of pairwise phage and bacteria interactions and used multiple analyses to characterize their coevolution based on phenotypes. All three analyses suggested that they engage in ARD. Lastly, we sequenced the full genomes of each isolate and reconstructed the isolates' phylogenetic relationships. The genome sequences revealed a phylogenetic pattern that was inconsistent with the ARD model. Our study demonstrates that phenotypic analyses are not sufficient to test hypothesis on coevolutionary dynamics and reveals a new type of coevolutionary dynamic we refer to as Leapfrog Dynamics (LFD).

2.2 Materials and Methods

2.2.1 Details on the initial coevolution experiment previous published

Meyer *et al.* [103] performed the original coevolution experiment with Escherichia coli B strain REL606 and a lytic bacteriophage λ strain, cI26. This λ strain was chosen because it cannot enter lysogeny, a life cycle phase where λ confers immunity to additional λ infections. By choosing a lytic strain, we forced the bacteria to evolve genetic resistance. *E. coli* and λ were cocultured in a carbon-limited minimal glucose media at 37°C for 37

days [103]. At the end of each day, 1% of the community was transferred to new flasks with fresh media, and, weekly, 2 ml of community was preserved by adding ∼15% of glycerol and freezing the mixture at -80°C.

2.2.2 Isolation of host and phage clones

We randomly isolated ten host and eleven phage individuals from different timepoints from the cryopreserved samples. In total, 50 strains of *E. coli* and 44 strains of λ were isolated from days 8, 15, 22, 28 and 37 of the experiment (no phage were detected on day 37). Bacteria were isolated by streaking onto Luria-Bertani (LB) agar plates [121] and randomly picking 10 colonies. These colonies were re-streaked three times to remove phage particles and grown overnight in liquid LB to create stocks. Phages were isolated by plating an appropriate dilution of the population onto overlay plates [3] with the sensitive ancestral bacteria, REL606, and randomly picking 11 plaques. These plaques were grown overnight with REL606 in LBM9 medium and stocks were created using chloroform isolation technique [103]. All phage and bacteria stocks were stored at -80°C with the addition of 15% of glycerol.

2.2.3 Pairwise infection assays and efficiency of plaquing (EOP)

We performed quantitative, pairwise infection assays for all combinations of host strains and phage strains that were isolated. Specifically, seven serial 1/10th dilutions were made of each phage isolate. 2 μl of each dilution plus undiluted phage stock was spotted on top of different host strain lawns including ancestor REL606. Thus, a total of $8 * 44$ spots of phage were plated on 51 different types of bacterial lawns, leading to a total of 17,952 pairwise infections. This allowed us to measure how well each phage isolate infects host isolates from different timepoints. We quantified phage infectivity by calculating efficiency of plaquing (EOP), defined as the ratio of density of phage isolate calculated on a coevolved

isolate to the density of phage calculated on the REL606 ancestor.

2.2.4 Analysis of PBIN nestedness and modularity

BiMat was used to assess the nestedness of the PBIN [53]. The raw EOP value matrix was first binarized into 0 for EOP = 0 and 1 for EOP > 0, and then *BiMat* was run with default settings. Since the result was a highly nested pattern, we report here the statistics for the conservative analysis where the rows and columns that contained all zeros were removed from the matrix to reduce the bias towards nestedness.

2.2.5 Resistance and infectivity calculations

For a total number of n host samples and m phage samples, we denote the EOP value for the ith host sample against jth phage sample as e_{ij} where $i \in [1, n]$ and $j \in [1, m]$; $n = 50$ and $m = 44$. We denote the five checkpoint days of day 8, 15, 22, 28 and 37 for host by k, where $k = 1, 2, 3, 4, 5$, and the four checkpoint days of day 8, 15, 22 and 28 for phage by l where $l = 1, 2, 3, 4$. Host resistance for a host sample i is calculated as

$$r_i = \sum_{j=1}^{m} \mathbf{1}_{\{e_{ij}=0\}} \tag{2.1}$$

which measures the number of phage strains that the host is resistant to. The host range of a phage sample j is calculated as

$$h_j = \sum_{i=1}^{n} \mathbf{1}_{\{e_{ij}>0\}} \tag{2.2}$$

which measures the number of host strains that the phage can successfully infect. The resistance percentage of host for each day is calculated as

$$RP_k = \frac{\sum_{i\ in A_k} r_i}{m \times |A_k|} \tag{2.3}$$

where A_k denotes the range of the host sample that belongs to the kth day and $|A_k|$ denotes the cardinality of the set A_k, i.e. the number of host samples at the kth day. Likewise, the host-range percentage of phage for each day is calculated as

$$HP_l = \frac{\sum_{j \in B_l} h_j}{n \times |B_l|} \tag{2.4}$$

where B_l denotes the range of the phage sample that belongs to the lth checkpoint and $|B_l|$ denotes the cardinality of the set B_l, i.e. the number of phage samples at the lth checkpoint.

2.2.6 Genomic DNA preparation for sequencing

λ genome extraction for whole genome sequencing were previously reported in [102]. To summarize, λ particles were concentrated using PEG precipitation, the phage were treated with DNase I to remove free-floating DNA not protected by phage capsids, the DNase is denatured with heat, which also releases capsid-enclosed phage DNA. The DNA was extracted using Invitrogen's PureLink kit. *E. coli* genomic DNA was extracted and purified from a 1 ml sample of culture by using PureLink kit.

Genomic DNA was further processed by fragmenting the DNA and attaching adapters and barcodes using a method outline in [12]. Sequencing was done at UC San Diego IGM Genomics using paired-end Illumina HiSeq 4000 platform.

2.2.7 Construction of mutation profile tables

After collecting the raw sequencing reads, we removed the adapters using cutadapt [97] and performed quality control (QC) for each isolated strain using FastQC [1].

The QC filtered sequencing reads were then analyzed using the breseq (v0.32.1) [40]. We ran breseq in the consensus mode with default parameters except for the consensus-frequency-cutoff, which was set to 0.5. The breseq pipeline first aligns the reads to the reference genome using bowtie2 [86], and then analyzes the mapped reads to identify mutations based on the new junctions, missing coverage and read alignment evidences. Finally, breseq generates a summary table with list of mutations and their corresponding evidence. The same breseq settings were used to analyze both host and phage data.

2.2.8 Phylogenomics

Due to the prevalence of large insertions and deletions in the host genomes, conventional nucleotide substitution models were not suitable for estimating the host phylogenetic tree. However, such models were suitable for estimating the maximum-likelihood phylogenetic tree for phage genomes. As a result, two different approaches were taken to reconstruct the evolutionary trajectories of the host and virus.

To construct the phage phylogeny, multiple sequence alignments were performed for all recovered genomes and the ancestral genome using mafft (v7.305b) [75] with default settings except that 'retree' was set to 2 and 'maxiterate' was set to 1,000. A maximum likelihood tree was constructed using raxml-ng [128]. We performed root-to-tip regression analysis to confirm the existence of temporal signal in the maximum likelihood tree. This was done by regressing tip distance from the root against the sample time. The significance of correlation between tip distance from the root and the sample time was evaluated by comparing the observed with the null distribution of coefficient of determination (R^2). The null distribution of R^2 was generated by randomly permuting the sample 500 times. Finally, the TreeTime [120] program was used to generate the phylogenetic tree.

To reconstruct the hosts' evolution we constructed a hamming distance matrix to calculate genetic distances between different host isolates. Neighbor-joining (NJ) trees were

then built based on the hamming distance matrix using T-REX [95]. Finally, the TreeTime program was used to build the host phylodynamic tree.

2.2.9 Whole genome whole population sequencing

The new LFD coevolutionary model developed to explain our results predicts that cryptic genetic variation that arises early in coevolution fuels the arms race at the end of the study. To test this prediction, we sequenced the full population of phage and bacteria to 142-fold coverage for bacterial genomes and 3,726-fold converge for phage. This allowed us to uncover alleles in the bacterial and phage populations that existed at lower frequencies than we could detect by isolating individuals. To do this, λ and *E. coli* populations from Day 8 were revived by growing 120 μl of frozen stock of the whole community in the conditions from the original experiment [103]. Phage and bacteria were then separated, and their genomic DNA was extracted in the same manner as described before for clonal stocks. Genomic libraries were prepared using NexteraXT kit at UC San Diego IGM Genomics. IGM also sequenced the samples using 75 base single reads on the Illumina HiSeq 4000 platform. breseq v0.32.1 was used to analyze whole population sequencing data of Day 8. We ran breseq in polymorphism mode with default settings to construct the mutation profile tables.

2.3 Results

2.3.1 Phage-bacterial infection network

The pairwise interaction study revealed an incredible number of λ genotypes with phenotypically distinct host-ranges, and *E. coli* genotypes that vary in resistance (Figure 2.1a). In line with the ARD model, we found that the interactions were highly nested (Figure 2.1b) and had a low level of modularity. Also in line with ARD, *E. coli* evolved

increasing resistance (Figure 2.1c, $R^2_{adj} = 0.5051$, $F_{1,48} = 51.01$, $P = 4.453e-09$ for linear model: response $\sim$ time), and λ gained increasing host range and infectivity (Figure 2.1d, $R^2_{adj} = 0.8131$, $F_{1,42} = 188.1$, $P = 2.2e-16$).

To further test between ARD and FSD, we performed a time-shift analysis using efficiency of plating (EOP) values to determine how phages' infectivity varies when presented with past, contemporary, or future bacteria (Figure 2.2a) [61]. ARD predicts that phages will be able to infect past and contemporary, but not future hosts, while FSD predicts that phages will be best at infecting contemporary hosts. The time-shift analyses was conducted for each λ isolate by calculating its mean EOP value for all 10 bacterial isolates on each day (Figure 2.2b). This analysis was repeated for the bacteria using the same EOP data but by calculating levels of resistance to λs isolated from different time points (Figure 2.2c). In line with ARD, λ isolates from days 22 and 28 had higher infectivity on past hosts than contemporary or future hosts (Figure 2.2b). The analysis for day 15 phages was inconclusive because the EOP values across time were not statistically significant. The pattern for bacteria was also in line with ARD: isolates from days 8, 15 and 22, had lower resistance (higher EOP) for phage samples from the future versus the phage isolated from the same time or in the past (Figure 2.2c). A full time-shift analysis could not be conducted for isolates form days 28 and 37 since the phage went extinct between days 28 and 37, however the contemporary bacteria was not significantly more resistant than the past time points.

Note, all mean EOP values were zero for day 8 phages because all isolated hosts were resistant to all day 8 phages (Figure 2.1a and Figure 2.2a). Still, λ did not go extinct in the coevolution experiment because of a phenomenon known as 'leaky-resistance' [30]. In this, a small number of resistant hosts revert to sensitive, thereby sustaining phage in the population until they are able to evolve to gain access to OmpF [103].

2.3.2 Genome sequencing

The genome sequencing revealed 22 unique *E. coli* and 34 phage genomes. Among the *E. coli* strains, we found a total of 18 unique mutations; 6 missense mutations, 1 nonsense mutation, 1 intergenic point mutation, 7 deletions and 3 duplications (Figure 2.3a). The most abundant mutation that occurred in 38 out of 50 host genomes was a frameshift mutation caused by a 25-base duplication in the *malT* gene, in line with the original study [103]. Disruptions in *malT* interferes with the expression of LamB protein which ancestral λ needs to bind to *E. coli* cells. We also observed one isolate with a *lamB* mutation (1-base deletion) *in lieu* of the typical *malT* mutation. The most resistant *E. coli* strains on day 37 have a number of mutations that are expected to confer resistance; a *malT* deletion, a nonsynonymous change in *ompF*, and a deletion in *manZ* [103].

A new bacterial mutation we observed in 25 genomes that had not been uncovered previously was a 777 bp deletion caused by the excision of an IS element. None of the genes impacted by this deletion—*insB-22* which encodes for IS1 protein InsB, *insA-22* which encodes for IS1 protein InsA, and *ECB_02825* which encodes for a pyrophosphorylase—are known to affect λ infection or reproduction [99], and similar IS mutations are known to occur at a high rates [33]. This, along with the observation that it always occurred with the 25-base duplication mutation, suggested that the deletion may be a neutral genomic hitchhiker. However, in a side experiment that we report in the online materials, we show that when Δ777 co-occurs with *malT* mutations, it produces high levels of resistance through epistasis (Figure 2.5, section 2.5).

In the λ isolates, we found a total of 176 unique mutations; 53 nonsynonymous point mutations, 87 synonymous point mutations, 2 insertions, 3 deletions and 31 intergenic mutations (Figure 2.3b). While this level of molecular evolution may seem unbelievable for such a short-term experiment, similar levels have been observed for other phages evolving in the laboratory [151]. 116 of these mutations were in the host-recognition gene *J*. The J

protein is positioned at the end of the phage's tail, and initiates infection by binding to *E. coli*'s LamB protein. Some of these *J* mutations have been shown to increase adsorption rates to LamB and allow λ to exploit a novel receptor, OmpF [25, 93, 115]. Interestingly, the more extensive sequencing effort performed here revealed a mutation in another tail fiber protein called H (C→T substitution at nucleotide position 11,451). This mutation rises later and likely plays a role in expanding host range. *H* is called the tape-measure gene because it helps determine the length of λ's tail, and mutations in this gene have been shown to increase λ's host-range in other experiments [123].

2.3.3 Phylogenomic reconstruction of coevolution

Even though multiple analysis of the phenotypic data supported the ARD model for coevolution, the pattern produced by the phylogenies are in line with predictions of FSD (Figure 2.4). The phylogenies of both *E. coli* and λ show that multiple lineages coexist for weeks, rather than a single dominant branch. A second unexpected observation was that the bacteria that had acquired the highest level of resistance at the end of the experiment did not descend from the isolates at earlier time points, suggesting that a rare lineage with high levels of resistance leaped ahead of the dominant lineage. This observation inspired the name leapfrog dynamics (LFD). Similarly, for λ, we find that that the clade dominant at the final timepoint with the broadest host-range was more closely related to wildtype λ than the clade dominant at preceding timepoints. For both species, the clades that win out later in the arms race seemed to exist as cryptic subpopulations early in coevolution.

2.3.4 Whole population sequencing at early stage of coevolution

To test the key prediction of LFD that cryptic lineages coexist with dominant lineages and can supply the genetic artillery used for later stages of the arms race, we sequenced full populations of *E. coli* and λ from day 8 and searched for mutations that rose

to prominence at the end of the study. Given the low mutation rates of *E. coli* (8.9×10^{-11} base^{-1} replication^{-1} [152]) and λ (7.7×10^{-8} base^{-1} replication^{-1} [44]), and their average population sizes (10^9 and 10^7, respectively [103]) it is unlikely that any single mutation would evolve twice, so the presence of these key mutations in an early sample would indicate that the lineages had established early in the arms race.

For *E. coli*, we were specifically searching for two mutations: a Δ16 bp deletion at position 1,882,915 in *manZ* and a non-synonymous mutation at 1,003,271 in *ompF*. These mutations are present in most of the day 37 isolates and are thought to confer resistance. The Δ16 bp deletion in *manZ* was detected, but not the OmpF mutation. We also found a 141 bp deletion in *malT* that cooccurs in the day 37 genomes with the *manZ* mutation. The *malT* deletion was at the same frequency, suggesting that these mutations were indeed linked and that they evolved sometime before day 8.

For λ, we focused on the mutation in *H* that rises to dominance between days 22 and 28. Indeed, this specific *H* mutation was present at day 8. Unlike *E. coli*, we did not find any other mutations present in the day 28 isolates, suggesting that the *H* mutation was the first adaptation to occur in this lineage.

Besides revealing the eventual winning lineages of the arms race, by sequencing populations we also discovered much more genetic diversity than through isolate sequencing. We found 52 unique mutations in *E. coli* and 38 mutations in λ from full population sequencing compared to 7 and 30 through isolate sampling, respectively. This shows that there is a significant amount of genetic diversity generated at the earliest phases of the arms race that can provide fuel for the later stages.

2.4 Discussion

Through large scale phenotypic assays and whole genome sequencing, we were able to test existing models of coevolution and learn that each were inadequate to explain λ and *E. coli*'s coevolutionary dynamics. Three complimentary phenotypic analyses in Figure 2.1 and Figure 2.2 suggested that coevolution between λ and *E. coli* followed ARD. However, the phylogenetic pattern revealed by whole genome sequencing was in line with the FSD (Figure 2.3). These observations lead us to develop a new model to characterize λ and *E. coli*'s coevolution—leapfrog dynamics (LFD). In this model, selection operates similarly to the ARD model, where parasite genotypes with ever-expanding host-ranges are selected and hosts with ever-increasing resistance are favored. However, the difference is that in the LFD model there is a genetically diverse pool of hosts and parasites that evolve early and on occasion, rare individuals are drawn from this pool with more advanced phenotypes and replace the dominant strains.

It is likely that ARD models fall short in making accurate predictions for the phylogenies because of assumptions about the genetics of host-range expansion and resistance. Evolutionary models tend to assume that mutations have small additive effects on phenotypes [149]. Applied to ARD, this would mean that the phage with the broadest host range is likely to expand its host-range faster than lagging genotypes, and for bacteria, the strain with the greatest resistance is most likely to acquire the next level of resistance and outcompete other strains. This genetic architecture favors the evolution of directed phylogenies with one dominant branch. Instead, λ J mutations are known to possess high-order epistasis and are nonadditive [102, 93]. This may allow rare lineages with an exceedingly beneficial combination of mutations to leap ahead. Non-additivity was also discovered in the bacteria in this study with respect to the interactions between *malT* – and the Δ777 mutations. A second problem with the genetics is the assumption of small effect-size mutations since mutations like the *malT* mutation can cause nearly complete

resistance to some λs (Figure 1a). Acquiring such large effect mutations could help lineages leap ahead.

One question left unanswered by this study is how multiple lineages persisted in this population for long durations. We hypothesized that trade-offs between host-range and other viral traits for phage, and resistance and competitive fitness for bacteria, could explain the evolution of genomic diversity. Trade-offs between host-range and λ stability were previously observed [115]; however, we were unable to detect trade-offs for resistance. A recent theoretical study showed that that there are many other mechanisms besides trade-offs that can select for genetic diversity during ARD, providing new hypothetical mechanism to pursue in the future [15].

Our findings also have important implications for understanding the mechanisms that structure nested PBINs. The nested pattern is ubiquitous in PBINs [52], as well as many other ecological networks [65, 11, 73], so it is important to understand the processes that produce this ordering. One hypothesis is that the structure is determined by the genetics of the interactions (gene-for-gene, see [149]). The second is ecological; nestedness emerges because of how the selection steadily shifts during an arms race to promote incremental increases in host-range and resistance. The conventional wisdom for which processes controls the arms races is the underlying genetics. This stems from a pervasive idea that genetic mutation and evolution happen more slowly than changes in ecology, so coevolutionary systems must be constrained by their access to genetic variation. The coevolution experiment studied here was initiated with small population size of isogenic stocks of λ and *E. coli*. This should have favored genetic control because all of the variation had to evolve *de novo*. However, a tremendous amount of genetic variation was generated at the earliest phase of the arms race that was not deployed for many generations, suggesting that the ecology controlled the dynamics, not the availability of genetic variation.

Lastly, our results provide a cautionary tale for over interpreting phenotypic data

based on the phage-bacteria infection networks. Before performing genomic analysis, we predicted that λ-*E. coli* coevolution under these laboratory conditions would fit the ARD model. Our error was that we did not anticipate the large amount of cryptic genetic variation and its role in driving late-stages of coevolution. Interestingly, this is the second time that we learnt about the limitations of PBIN data for making prediction. Based on the matrix in Figure 2.1a one would predict that the phage should go extinct on day 8, however leaky-resistance fueled by host cells too rare to be sampled allows λ to persist. Together, the LFD and leaky-resistance results show that host-phage dynamics revealed in the PBINs miss the rich dynamics played out at lower frequencies in the population and offer insufficient information to predict dynamics.

Altogether, we find it remarkable that such a diverse and complex community could evolve in a single population within five weeks. In studying this diversity with both phenotypic and genomic approaches we revealed that typical models of coevolution are insufficient to characterize the complex dynamics in this community and that phenotypic assays alone fall short in characterizing coevolutionary dynamics. We also showed that highly organized ecological pattern like nestedness can emerge from the extremely disordered population genomic pattern underlying them, demonstrating the power of selection to producing nonrandom ecological patterns.

2.5 Epistasis between the *malT* mutation and Δ777 bp mutation

We tested for non-additive interactive effects between the two mutations by comparing resistance to phage between *E. coli* genotypes that had only the individual mutations versus the isolate that had both. To do so, we first generated an *E. coli* strain that that only possessed the Δ777 mutation. We picked two isolates from the experiment that had

both *malT* and Δ777 mutations (isolates D37-7 and D37-9 from day 37) and selected for malT-revertants that only had the Δ777 mutation. The 25-base duplication malT mutation has a high rate of reversion [30], allowing us to select $malT^+$ revertants by growing on a medium where maltose is the only carbon source. We spread $\sim 10^8$ cells of each isolate on minimal maltose agar plates and incubated the plates for two days at 37°C. A single colony was picked from each plate and re-streaked on tetrazolium maltose (TM) indicator plates to confirm the reversion. Revertants that can metabolize maltose produce white colonies and are easily distinguishable from $malT^-$ strains that produce smaller red colonies [127]. The TM plates were incubated at 37°C for overnight, and this procedure was repeated once more from colonies on the TM plates to further purify the isolates. The resulting strains were labeled D37-7+ and D37-9+.

To test for resistance, we compared the growth of the bacteria with and without the phage in a 96-well flat bottom plate. We chose a phage isolate that had a broad host-range and high efficieny-of-plaquing (EOP) value to challenge the *E. coli* genotypes with (isolate D28-11 from day 28). Six *E. coli* genotypes were evaluated: ancestor without mutations, a strain with only 25-base malT duplication (reported in (Chaudhry et al. 2018)), D37-7+, D37-9+, D37-7 and D37-9. A plate for each genotype was incubated at 37 °C with ~ 100 bacterial cells and OD was measured every 5 minutes, just after the plate was shaken to resuspend any cells that may have settled. Four replicates were run for both treatments, with phage ($\sim 10^9$ λ particles) and without phage. Bacteria's growth were observed using a plate reader (Tecan Sunrise) that monitors optical density (OD) at wavelength 600 nm. LBM9 medium was used for this experiment and each well contained a final volume of 200 μl of liquid.

The growth patterns without the phage were similar for all *E. coli* strains (Figure 2.5). None of the strains grew with λ, except for D37-7 and D37-9, the strains with both mutations. To compare the suppressive effect of λ on the different *E. coli* strains, we quantified the

maximum OD achieved for each replicate with λ added. We ran a general linear model where maximum OD was predicted by the presence of $malT^-$, Δ777, and their interactions. A variable that accounts for difference in strains derived from isolate D37-7 or D37-9 was left out of the model because we found no significant difference between the maximum ODs of D37-7 and D37-9 or D37-7+ and D37-9+. The analysis was performed in R (version 3.6.1). Neither $malT^-$ or Δ777 conferred resistance: t = -0.098, $P = 0.923$, and $t = 0.045$, $P = 0.964$, respectively. However, the interaction term was highly significant: $t = 41.316$, $P < 0.001$). This shows that the *malT* mutation and Δ777 bp mutation have an epistatic interaction that confers resistance to λ.

2.5.1 Media recipe for plates

Minimal maltose agar plates: 5.34 g potassium phosphate dibasic anhydrous, 2 g potassium phosphate monobasic anhydrous, 1 g ammonium sulfate, 0.57 g sodium citrate dihydrate, 16 g agar, 4 g maltose per liter of water and supplemented to a final concentration of 1 mM magnesium sulfate, 0.0002% w/v thiamine, and 0.0002% w/v biotin.

Tetrazolium maltose plates: 10 g tryptone, 1 g yeast extract, 5 g sodium chloride, 16 g agar, 10 g maltose per liter of water and supplemented to a final concentration of 0.005% tetrazolium indicator dye TTC.

2.6 Acknowledgments

Chapter 2 is currently being prepared for submission for publication of the material, as it may appear in *Ecology Letters*, 2021: A. Gupta, S. Peng, C. Y. Leung, J. M. Borin, J. S. Weitz, J. R. Meyer, "Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing".

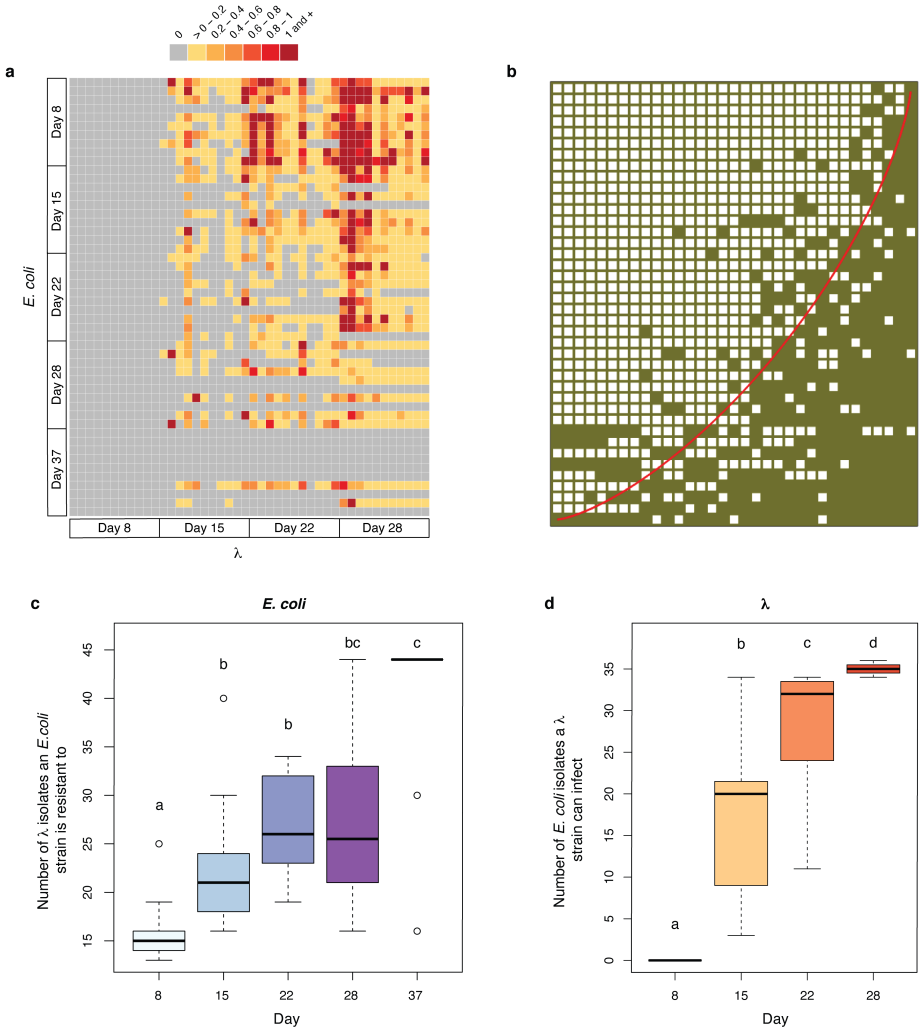

Figure 2.1: Host resistance and phage infectivity measured by pairwise plaque assay. a) Phage-bacteria infection network where the color of each cell is determined by the EOP values obtained for that host-phage interaction pair; grey cells represent no infection by λ on the given E. coli strain, yellow represents low infectivity and red represents high infectivity. b) The original network in a) but reassembled to maximize nestedness using the software BiMat. Filled squares indicate a combination of host and phage that result in successful interactions (EOP > 0), and the red line highlights the isocline using the NTC algorithm. The nestedness value of the network based on NODF algorithm was significantly greater than what null models with same number of filled cells would have based on 200 random shuffles (0.839 vs 0.638 $\pm$ 0.011). c) Boxplots showing the total number of λ isolates from all days that E. coli genotypes are resistant to across different sampling days. d) Boxplots showing the total number of E. coli isolates from all days that λ genotypes can infect across different sampling days. Lowercase letters in c) and d) denote significant difference between different days via Tukey's honest significance test: c) ANOVA: $F_{4,45} = 13.3$, $P = 3.11\text{e-}07$), d) ANOVA: $F_{3,40} = 67.05$, $P = 1.17\text{e-}15$. A simple linear regression model with time as the predictor variable was also used to test if E. coli evolved increasing resistance in c) and λ evolved increasing host range in d).

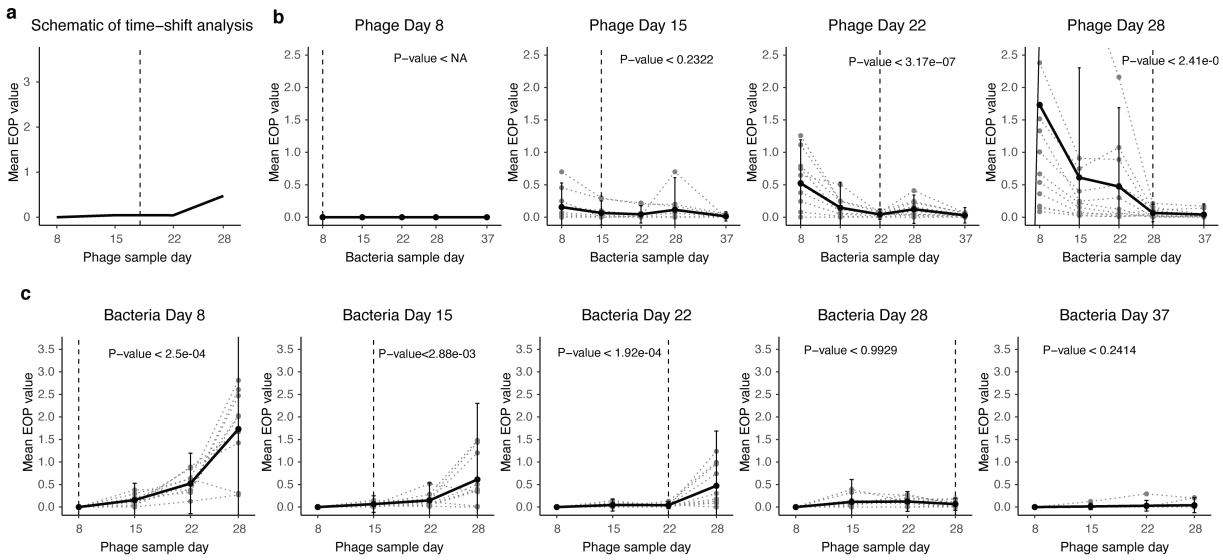

Figure 2.2: Time-shift analysis results from different checkpoints. a) Schematic for the time-shift analysis that compares the mean EOP from hosts or phages interacting with their counterparts from the past, contemporary and the future. b) Time-shift results from phage checkpoints day 8, 15, 22 and 28 respectively. The gray dotted line shows the time-shift curve for each individual phage and the black line shows the average. The vertical dashed line represents the phage sample day. The P-values shown here are the maximum P-value from one-sided paired t-tests comparing the initial checkpoints with each of the later checkpoints. c) Time-shift results from host checkpoints day 8, 15, 22, 28 and 37, respectively. The gray dotted line shows the time-shift curve for each individual host and the black line shows the average. The vertical dashed line represents the host sample day. The P-values shown here are the maximum P-value from one-sided paired t-tests comparing the final checkpoints with each of the previous checkpoints.

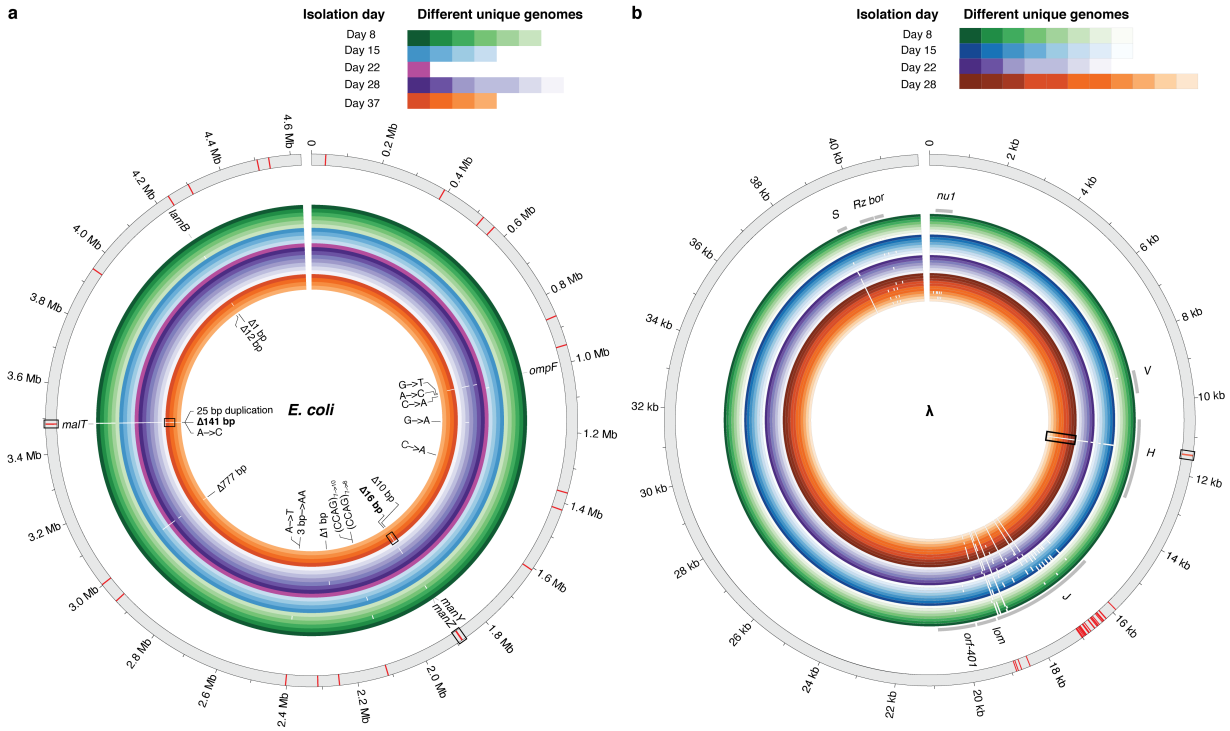

Figure 2.3: Genomic diversity in clones isolated from different days and full population sequencing for a) *E. coli* and b) λ. The outermost gray ring represents the reference genome and inner colored rings represents different unique genomes isolated from different days, with different shades of the same color represent same sampling day. White gaps in the genomic rings indicate the location of observed mutations. All mutations found in clonal isolates have been identified for *E. coli* in a), however due to large number of mutations in λ, only the genes that harbors the mutation have been shown in gray in b). The red bars in the outer gray ring represents mutations discovered from full population sequencing on day 8. The mutations that become dominant at later stages of coevolution and were also found in day 8 population sequencing have been highlighted with rectangular boxes.

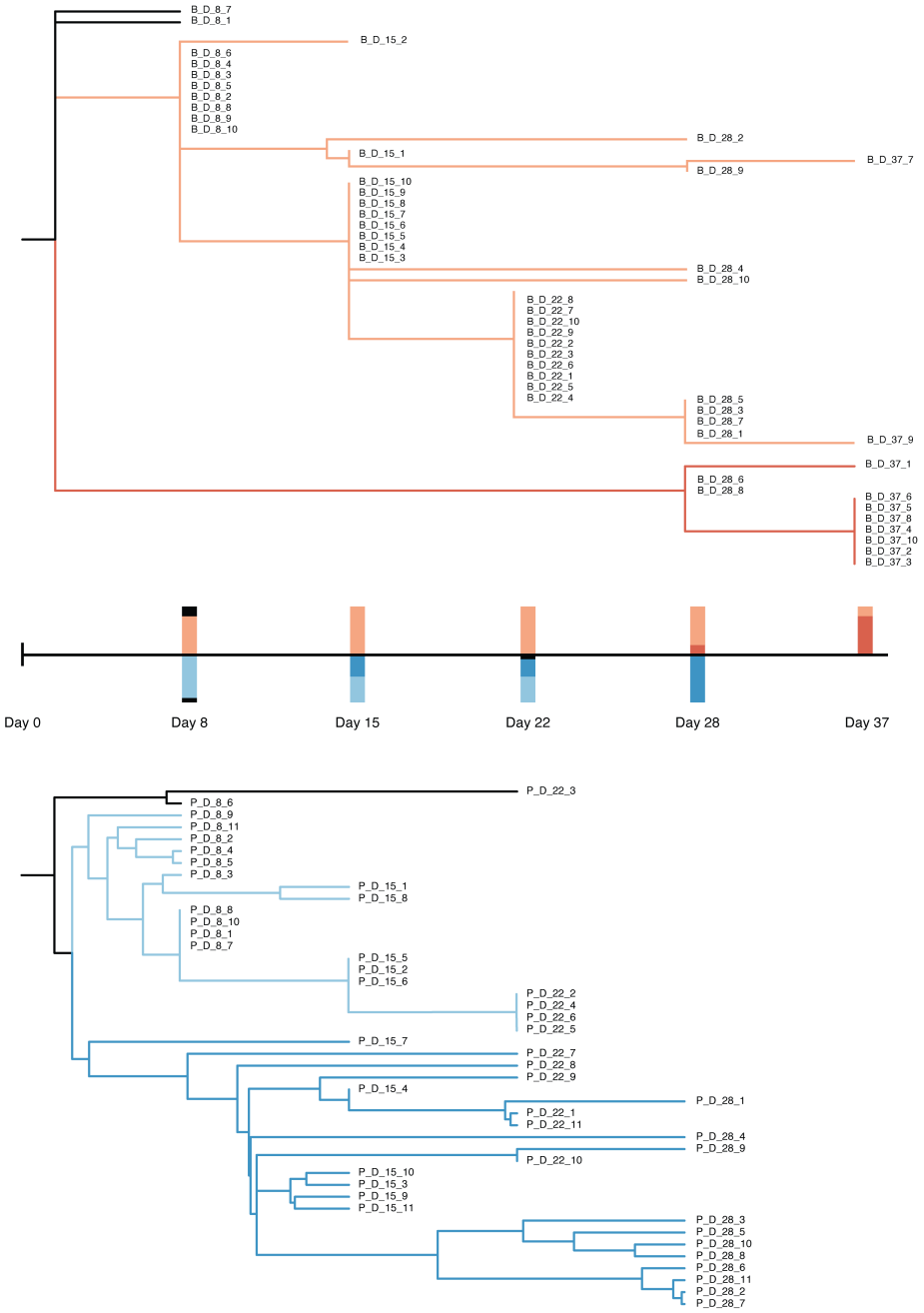

Figure 2.4: Reconstructed phylodynamic trees of the hosts and phage. (A) The host phylodynamic tree reconstructed based on host mutation profiles. All super-resistant host strains are located on the red branch. The bar above the time scale represents the proportion of host strains from each colored branch across different checkpoints. (B) The phage phylodynamic tree reconstructed based on the phage mutation profiles. All day 28 phage strains are located on the dark blue branch. The bar below the time scale represents the proportion of host strains from each colored branch across different checkpoints.

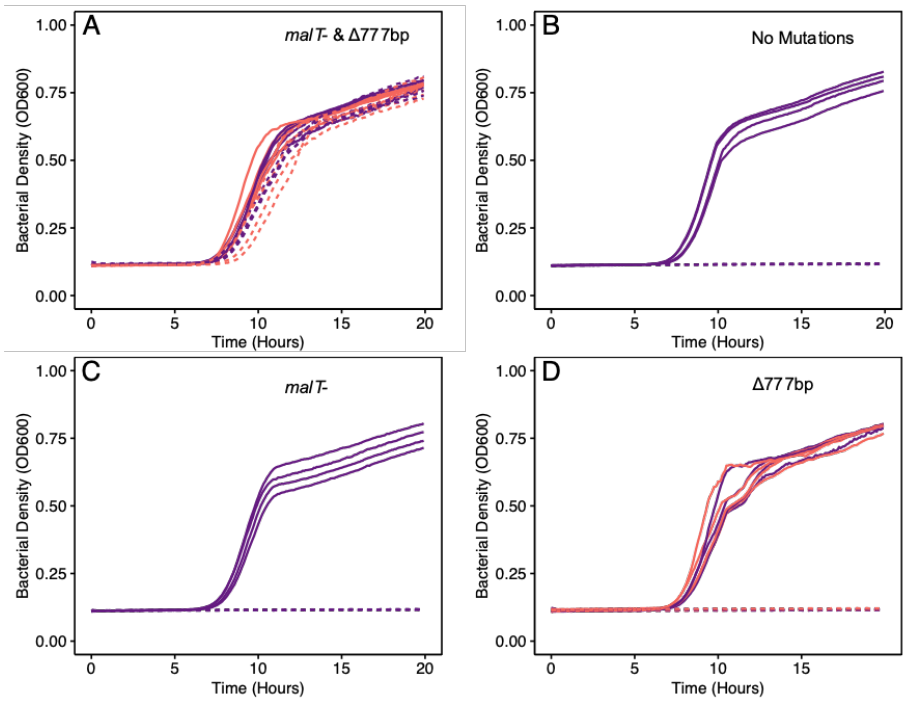

Figure 2.5: Growth trajectories of bacteria with different combinations of *malT* and Δ777 mutations growing in the absence (solid line) or presence (dashed line) of a coevolved phage from day 28 in wells of a 96-well plate. A) Day 37 isolates D37-7 (purple) and D37-9 (pink), which have both *malT* and Δ777 mutations, B) ancestor REL606 which has neither malT nor Δ777 mutations, C) $malT^{-}$ strain which only has the 25-base duplication *malT* mutation, and D) strains D37-7+ and D37-9+ derived from D37-7 and and D37-9 (colors as in Panel A) which have reverted their malT mutation from MalT^{-} to MalT^{+}, and thus, have only the Δ777 mutation.

Chapter 3

Coevolution promotes innovation through deformations in fitness landscapes

3.1 Introduction

A starting point for understanding how populations evolve is to assume that they exist in an unchanging world where they can adapt towards optimality [111, 63]. However, even in static environments, populations never reach optimality because their circumstances continuously change as neighbouring species coevolve with them [143]. This more dynamic view of the evolutionary process opens up the potential for unbounded evolution and creates new opportunities for evolutionary innovation [43, 154, 136, 134, 108]. Darwin recognized this potential in the final pages of *On the Origin of Species* where he wrote that, "It is interesting to contemplate an entangled bank" of organisms evolving with one another to produce such a variety of forms and functions [35]. But he also realized the empirical challenges created by the richness of species interactions within ecological communities

in his further description of "these elaborately constructed forms, ... dependent on each other in so complex a manner..." [35]. The complexity arises because an organism's fitness is a function of its interactions with other species and these interactions continuously change as they coevolve. Furthermore, the coevolving traits of the organisms are encoded within genomes by mutations that can interact with one another through epistasis [148]. This means that interactions at all levels must be considered; from mutation-by-mutation within a species (classical epistasis), to mutation-by-mutation between species (interspecific epistasis), and higher order phenomena such as classic and interspecific epistasis combined.

Many advances have been made over recent decades that enable us to tackle this combinatorial problem. Efficient genetic engineering methods permit the construction of genetic libraries with combinatorial sets of mutations that can be used to measure epistasis [82, 57]. Also available are convenient approaches to measure Darwinian fitness of the mutant libraries [148, 113, 80, 31]. Coupling these two technologies allows the creation of extensive genotype-to-fitness maps, or fitness landscapes [153], that provide information important for predicting adaption [38, 39, 87]. However, these maps alone are often not sufficient to predict evolution because their topologies can depend on abiotic environmental conditions [110, 91, 55, 129] and biotic interactions [28, 58]. Here we take two significant steps forward in fitness landscape research. First, we build on the observation that landscape structures depend on species interactions by studying the interdependence of two species' landscapes and how they shift during coevolution. Secondly, we test whether these shifts facilitate the evolution of a key innovation.

As a model system of coevolution, we studied the host-parasite interactions among bacteriophage λ and its host, Escherichia coli, because of the extensive background research completed on their coevolution and the availability of well-developed molecular tools [103, 93]. When λ and *E. coli* are cocultured in the laboratory, one quarter of the λ populations evolve to use a new receptor [103]. λ's native receptor is *E. coli*'s outer-membrane protein

LamB, but through mutations in its host-recognition gene J, λ evolves to use a second receptor protein, OmpF. While only four mutations are necessary for OmpF$^+$ function [93], more J mutations typically evolve along the way [103]. λ gains this new function after *E. coli* evolves resistance through malT mutations [103] that cause reduced LamB expression [21]. Thus, it was hypothesized that the evolution of resistance in *E. coli* deformed λ's fitness landscape in ways that promoted λ's innovation [135]. In line with this, it was previously shown that some λ genotypes have higher relative fitness when cultured with resistant $malT^-$ cells than ancestral cells [25], suggesting that the host's coevolution may promote key steps in λ's evolution. However, not all genotypes were consistent with this pattern, and overall, too few genotypes were assayed to test the hypothesis. Here we build on this study with high throughput technologies capable of measuring the fitness of hundreds of λ genotypes in the presence of each host. This allows us to establish the contours of λ's adaptive landscape and to determine whether host-induced deformations that naturally arise during coevolution promote OmpF$^+$ evolution.

3.2 Construction of λ's fitness landscapes

To construct λ's fitness landscape, we focused on ten J mutations that were a subset of mutations λ repeatedly evolved on its path to OmpF use [103] (Figure 3.1). Together they form a ten-dimensional genotype space with a total of 1,024 (210) unique variants of different combinations of the mutations including the wild type allele configuration. Using Multiplexed Automated Genome Engineering (MAGE) [144], we were successful at engineering a library of 671 genotypes out of the possible 1,024 (section 3.8.2). To measure the fitness of each genotype in this library, we competed the full library en masse and monitored their frequency changes using next generation sequencing (Figure 3.2) [79]. The fitness of each genotype was then computed by comparing its change in frequency relative

to the non-engineered ancestor (section 3.8.2). Fitness was measured in four replicate competitions for both the ancestral host and $malT^-$ host. To reduce the effect of sequencing errors and to overcome other methodological pitfalls, we modified the MAGE protocol by introducing neutral watermark mutations in the library construction and developed a high-throughput competition assay that yielded reproduceable results (see Materials and Methods, Figure 3.17, section 3.2.1). Overall, we were able to measure the fitness of 580 λ genotypes cocultured with ancestral *E. coli* and 131 genotypes with $malT^-$ (Figure 3.11). The reduction in the number of genotypes we were able to measure out of 671 in the initial library was mainly due to unfit genotypes' frequencies falling below our limit of detection during the competitions.

3.2.1 Fitness effect of synonymous mutations used as watermarks in MAGE

We designed two watermark mutations for each focal mutation. Watermark mutations are synonymous mutations that fall within a few nucleotides of the focal mutation. These edits improved our ability to detect the presence of focal mutations within the MAGE library (Figure 3.16). In principle, these watermark mutations should not affect the fitness of λ because the introduced mutations do not change the amino acids. However, we designed two different watermarks in order to test whether or not they influence λ fitness. To test for fitness effects of the synonymous mutations, we evaluated the fitness of genotypes with just a single focal mutation, and then we split the count data from the competition experiments into two, one for watermark '1' and a second pool for '2'. We calculated the fitness of each group and compared their means using a t-test. The analysis was limited to just fitness measurements made for the treatment with the ancestral host because the single mutants did poorly on $malT^-$ host and many dropped below our limit of detection during the competitions. We were also only able to run the analysis on 7 out of

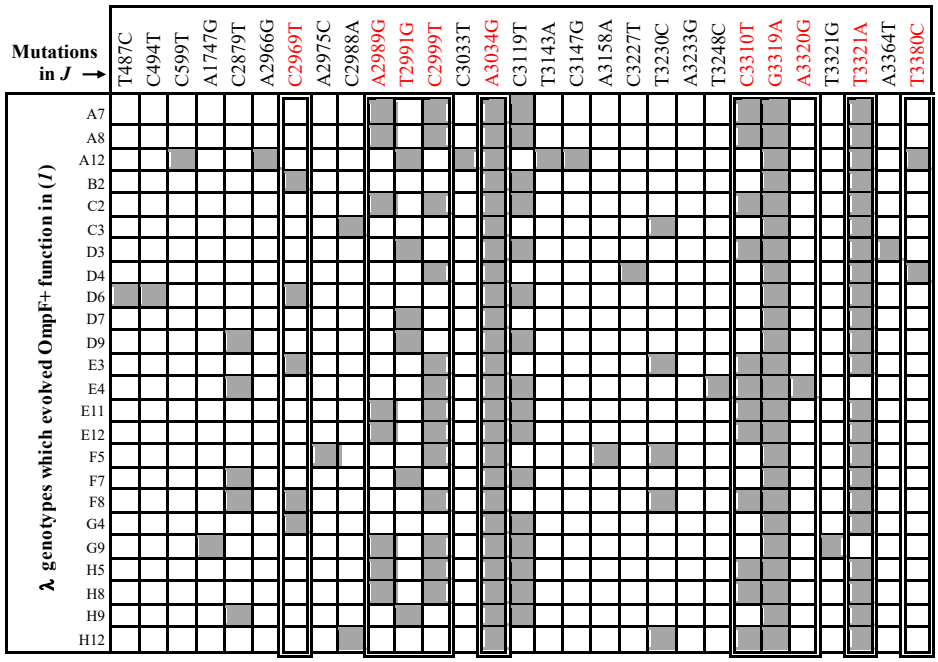

Figure 3.1: Reconstruction of the list of mutations in 24 phage isolates that evolved to target the OmpF receptor in the large-scale coevolution experiment by Meyer *et al.* [103]. Highlighted mutations were chosen to form the genotype space for the 10-dimensional fitness landscape in Figure 3.5. The first five and last five mutations fall within two 100-nucleotide windows for Mi-Seq 100 base paired-end sequencing.

10 mutations because in three cases only one watermark was represented in the final counts data. We found that the fitness did not depend on the neutral marker for all seven using a Bonferroni corrected alpha value of 0.0071. However, one of the seven was significant based on the uncorrected alpha value of 0.05. The effect on the selection rate was estimated to be 0.9947 with 95% confidence interval of 0.4758 to 1.5136. This value falls well within the error associated with fitness estimates (Figure 3.12). Because of our uncertainty of whether there was an effect of this synonymous mutation, and because the effect falls within normal levels of error, we did not take steps to correct for this possible effect.

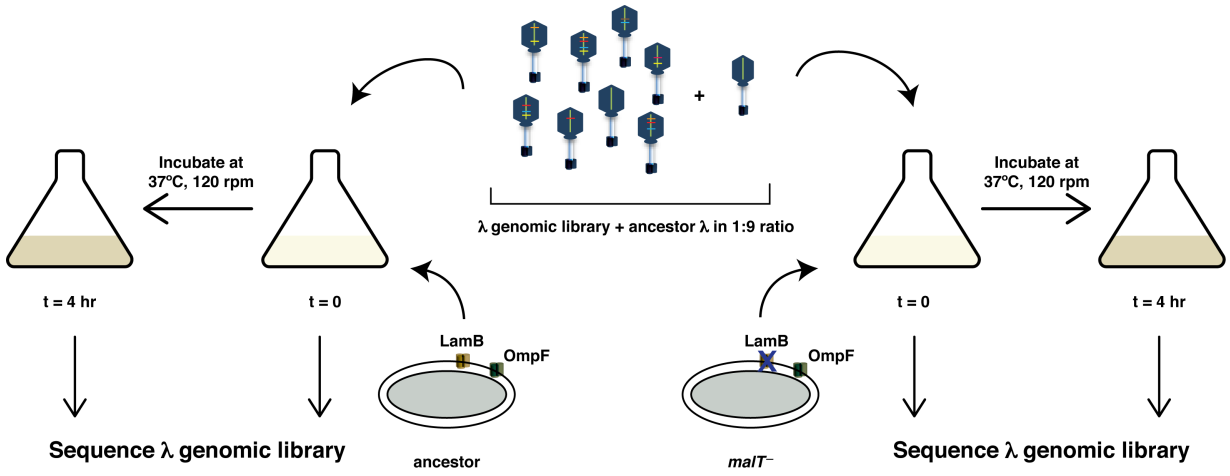

Figure 3.2: Schematic illustration showing how fitness was measured for the combinatorial λ library in order to construct fitness landscapes. The genomic library was first mixed with the ancestor in $1:9$ ratio, incubated with different types of host for four hours at 37°C, and then sequenced at both initial and final timepoints to calculate the selection rates of individual genotypes.

3.3 Structural differences between the two landscapes

Visual inspection of the two fitness landscapes reveal host-dependent structures; the landscape with the ancestral host has a standard diminishing return pattern [80, 31, 92, 83], while the landscape with $malT^{-}$ host has an atypical sigmoidal shape that plateaus at a higher fitness than the first (Figure 3.5). The nonlinear relationship between mutation number and fitness suggests the presence of epistasis (mutation-by-mutation interactions), the differences in the magnitude of fitness effects between landscapes suggests mutation-by-host interactions, and different shapes suggest host-dependent epistasis (mutation-by-mutation-by-host interactions). To determine how much variation in fitness is explained by these interactions, we performed multiple linear regression analyses (see Methods). We found pervasive epistasis in both landscapes (Figure 3.3), significant mutation-by-host interactions (Figure 3.3), and sizeable effects of the host-dependent epistasis (21 terms were significant out of 45 and 12.62% of the total variance in the data were attributable to these terms, Figure 3.4). The last three-way interaction term measures the extent to which

the landscape structure is transformed by host evolution and suggests that λ's evolutionary trajectory could depend on its host's genotype.

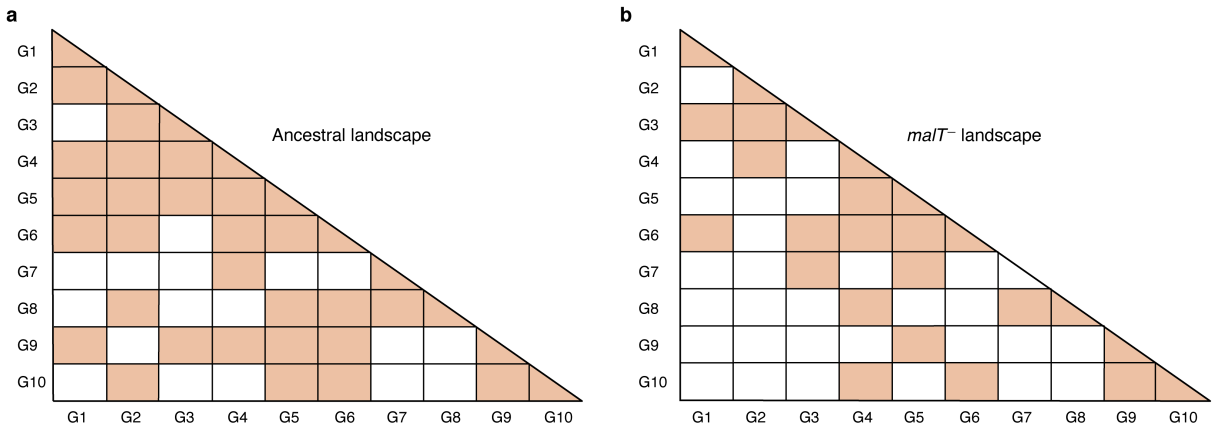

Figure 3.3: Statistical analysis of direct and interactive effects of mutations in **a)** ancestral landscape and **b)** $malT^-$ landscape. Colored cells represent statistically significant terms determined by multiple regression analysis after correction for multiple hypothesis testing (see Materials and Methods). The diagonal elements of the matrix represent single mutation effects and all the off-diagonal terms represent pairwise epistatic interactions. See Supplementary Table 5 for identity of mutations corresponding to different G_i. For the ancestral landscape, 58.66% of the variation was explained by the identity of the mutations and 24.69% by pairwise interactions by the model ($R^2_{adj} = 0.8712$, $F_{55,439} = 39.97$, $P < 0.0001$). Similarly, 48.35% of the variance in the $malT^-$ landscape was explained by the direct mutation effects and 27.61% by the interaction terms ($R^2_{adj} = 0.7072$, $F_{55,252} = 14.48$, $P < 0.0001$). Note that we cannot compare the different epistatic (GxG) terms directly between the two landscapes because of the different number of data points collected between the two landscapes. To understand how epistasis changes with the host type, we regressed another linear model (see Fig. 3.4) that includes mutation-by-mutation-by-host (GxGxE) interaction terms.

3.4 Simulation results of λ's evolution on fitness landscapes

To test whether changes in the structure of λ's landscape opened trajectories to OmpF function, we simulated λ's evolution on the landscapes using a modified Wright-Fisher model (see Materials and Methods). Before running the simulations, we imputed

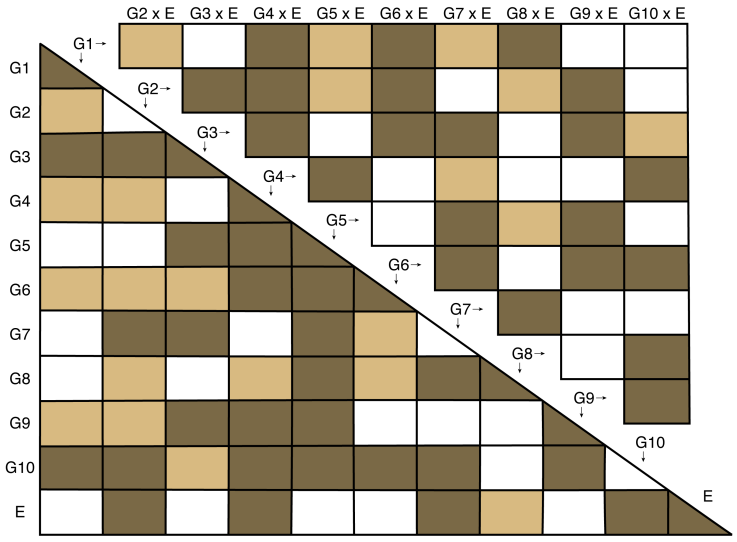

Figure 3.4: Statistical test of whether the two landscapes varied in topology. Light colored cells indicate terms present in the final AIC-optimized model out of the full-factorial model ($F_{76,726} = 37.45$, $P < 0.0001$), and dark colored cells indicate statistically significant terms after controlling for rate of false positives (see 'Statistical analysis of fitness landscapes' in Materials and Methods). This is analogous to the analysis represented in Figure 3.3; however, this analysis integrates information from both landscapes. The diagonal in the first ten rows in the lower-left matrix indicates direct effects of the mutations, whereas the off-diagonal cells indicate genotype by genotype interactions. The last row of this matrix indicates environment (host) by genotype effects. The upper matrix provides information on genotype by genotype by environment interactions. See Figure 3.16 for identity of mutations. In the final AIC-optimized model, 42.45% of the total variance in the data was explained by all single-effect terms, 24.60% by all two-way interaction terms and 12.61% by all three-way interaction terms.

the missing λ genotypes' fitnesses to complete the landscapes. We did this by successively choosing missing genotypes at random and assigning them the average fitness of their nearest neighbors. The simulations were run based on conservative estimates of the number of generations and population sizes from the experiment (960 generations; $\sim 6.3 \times 10^9$ λ particles, Figure 3.13) and λ's intrinsic mutation rate ($\sim 7.7 \times 10^{-8}$ base-1 replication-1) [44]. We predicted that λ would be more likely to evolve OmpF function in simulations that accounted for coevolution by shifting the population from one landscape to the next. We ran trials where λ evolved on only one landscape at a time to establish a baseline for the frequency of OmpF$^+$ evolution without coevolution. Next, we ran nine shifting landscape

scenarios where we varied how many generations λ evolved on the ancestral host landscape before switching to the $malT^-$ landscape. As anticipated, the switching protocol increased the frequency of OmpF$^+$ evolution in all 9 treatments above the single host simulations, but only 7 out of 9 treatments were found to be significantly higher (Figure 3.5; ANOVA: F-ratio $= 6.14$, $d.f. = 99$, $P < 0.0001$). This result was robust to changes in population size and total number of generations. (Figure 3.6).

3.4.1 Simulation results when only using genotypes present in both the ancestral and $malT^-$ fitness landscapes

One possible complication with the fitness landscape analyses is that the $malT^-$ landscape is based on a subset of the ancestral landscape (131 versus 580 out of 1,024 possible genotypes).

This could lead to potentially spurious comparisons of evolution on the two hosts. To control for this, we re-ran the simulations starting with only the genotypes that were present in both the landscapes. While the frequencies of OmpF$^+$ populations did shift with this new analysis, the main result that switching landscapes enhances the frequency of OmpF$^+$ evolution remained statistically significant (Figure 3.6d). The increased frequency in ancestor-only treatment stems from the increased stochasticity in estimating the landscape since more genotypes are imputed in panel d than a-c. This stochasticity leads to more OmpF$^+$ evolution because the increased randomness in landscape formation increases the chances of producing viable pathways to OmpF use. Interestingly, the ancestral landscape OmpF$^+$ frequency is equivalent to $malT^-$ in panel d. This similarity suggests that the frequency of OmpF evolution for $malT^-$ may be artificially high, which would mean that in reality there may be greater differences between this treatment and the fluctuating treatments.

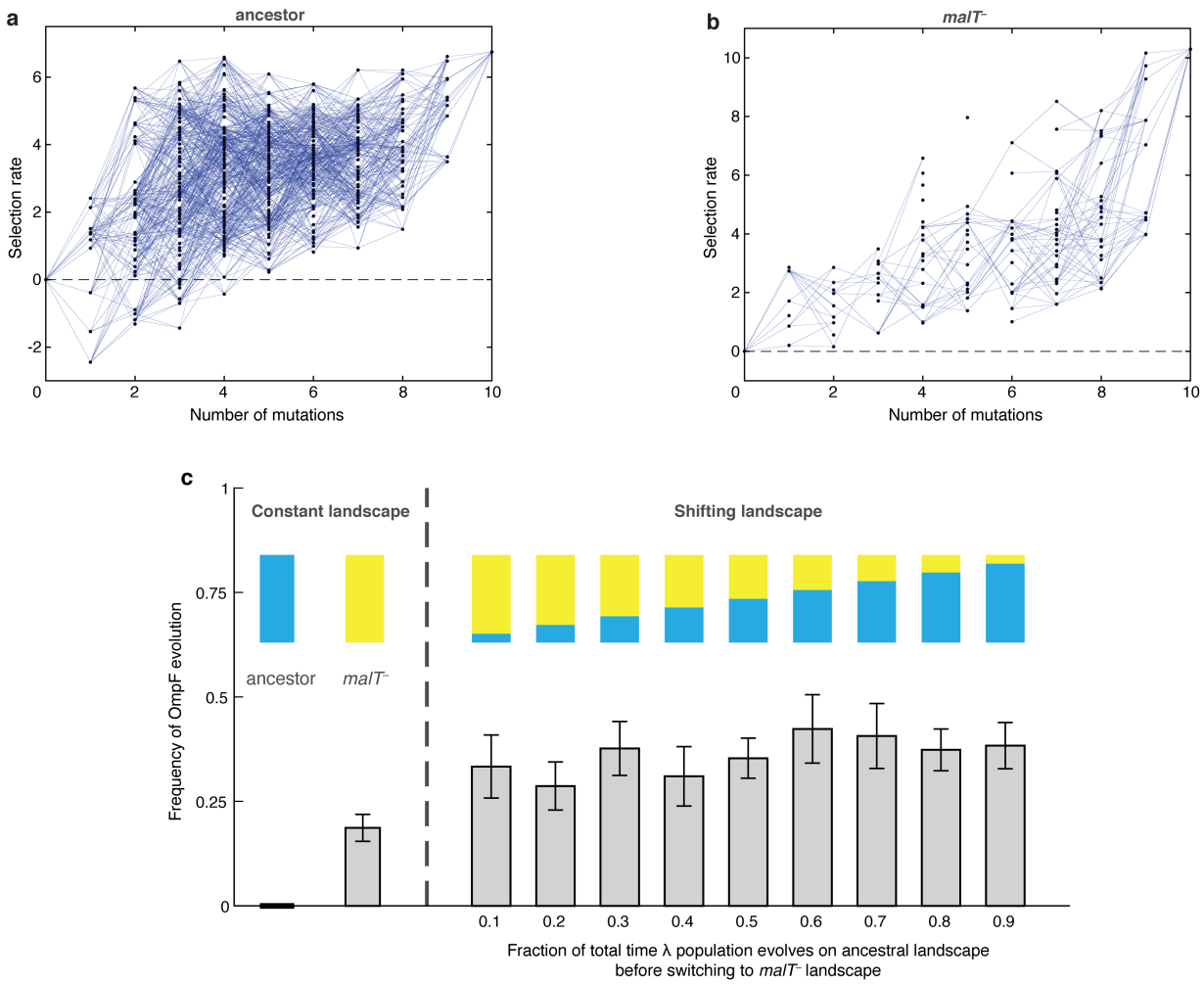

Figure 3.5: Empirical fitness landscapes of λ when infecting the **a)** ancestral host and **b)** $malT^-$. Each node represents a unique genotype and two nodes are connected by edges if the corresponding genotypes are separated by one mutation. The node at zero mutations is ancestral λ. Selection rate (per four-hour competition experiment) is the difference of Malthusian growth rates of a given genotype i to ancestral λ over four hours, calculated as $\ln(\lambda_{i,4}/\lambda_{i,0}) - \ln(\lambda_{anc,4}/\lambda_{anc,0})$ where $\lambda_{i,t}$ denotes the density of the given genotype at time t. c) Frequency of OmpF evolution observed when the evolution of λ population is simulated on different landscapes. Each bar represents an average of 300 simulation runs. Error bars indicate 95% confidence intervals. The only two switching treatments which are not significantly higher than constant $malT^-$ landscape are 0.2 and 0.4.

3.4.2 Simulation results when shifting is done between landscapes of the same host

Each time we impute fitness values of missing genotypes in a simulation run, a slightly different landscape is produced. This procedure raises the question of whether the increased frequency of OmpF$^+$ in the shifting landscape simulations is due to structural differences between the two landscapes, or the random differences created by the imputation technique. To test this, we repeated simulations as previously discussed; however, this time we shifted to a newly generated landscape created from the same host-landscape data. OmpF$^+$ did not evolve when the shift was made between two ancestral landscapes. λ did evolve OmpF function in some replicates when the switching was made between two $malT^-$ landscapes; however, the frequency was significantly less than when the shift was made from ancestor to $malT^-$ (Figure 3.6e and 3.6f). Lastly, to control for the sparser landscape of $malT^-$ compared with the ancestor, we shifted between two ancestral host landscapes that were generated using only fitness values of genotypes present in both landscapes. We found the same qualitative result as for the $malT^-$ to $malT^-$ landscape shift (Figure 3.6g and 3.6h). These results show that the structural differences between the two host landscapes encourage OmpF$^+$ evolution above and beyond what results from the noise associated with the imputation procedure.

3.5 Detailed examination of a single coevolving laboratory population

The simulation results suggest that the shifting landscape encourages λ's evolution to OmpF function. In particular, the simulations show that the first steps along the path to the innovation are more likely if λ adapts to the ancestral bacterium, meanwhile

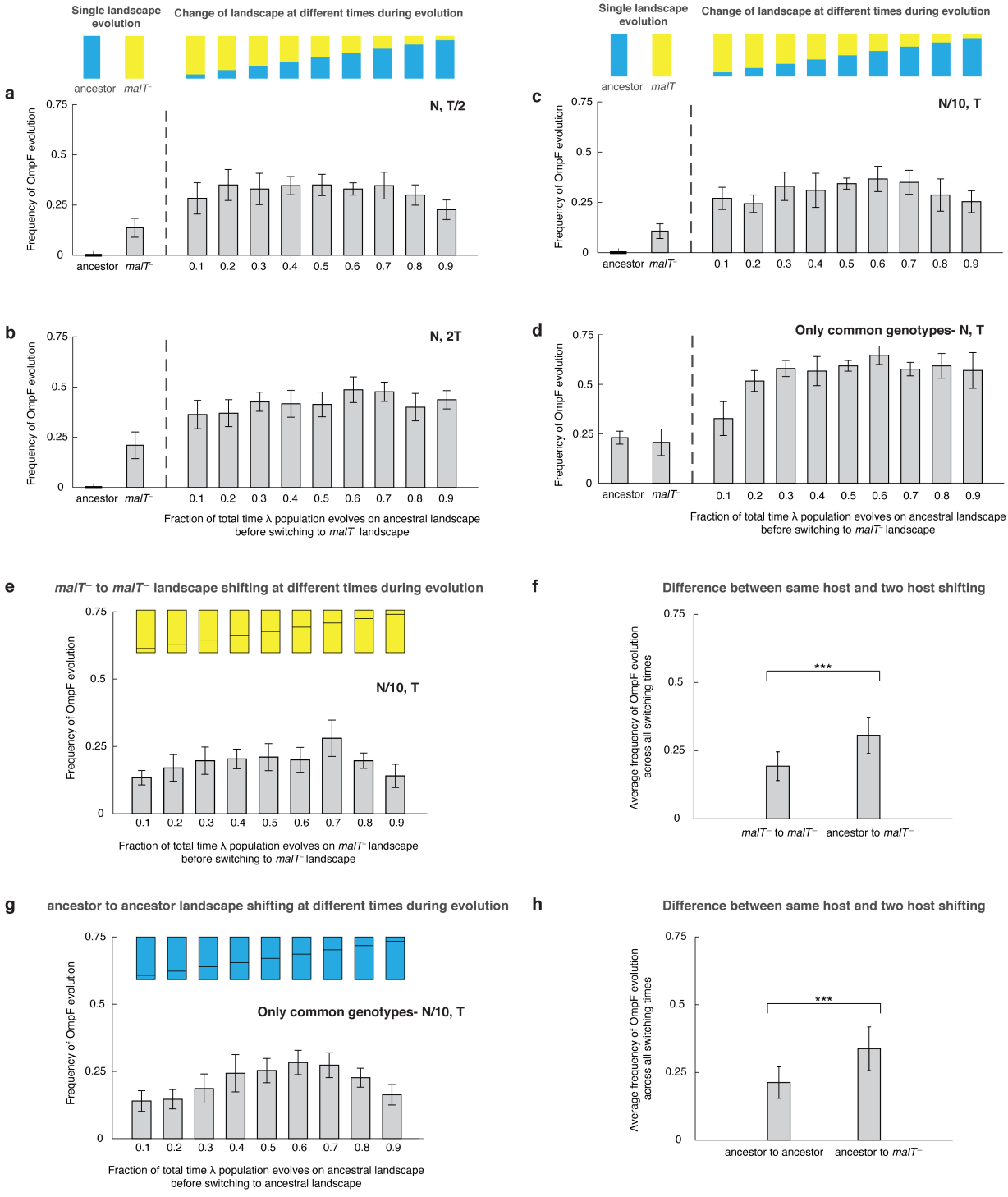

Figure 3.6: (Continued on the following page.)

Figure 3.6: Additional simulations to verify that shifting landscapes promote OmpF$^+$evolution. Frequency of OmpF$^+$evolution when λ population is evolved on different landscapes with different controls as compared with Fig. 1c— **a)** half the total number of generations, **b)** twice the total number of generations, **c)** with one-tenth of the population size, and **d)** starting with only the genotypes whose fitness values were measured for both the landscapes. For all switching landscape treatments except two (switching fraction-time of 0.9 in (a), and 0.1 in (d)), the frequency of OmpF$^+$ evolution was significantly higher (adjusted P-value¡0.05) than the constant landscape treatments (ANOVA: F-ratio = 6.62, $d.f. = 99$, $P < 0.0001$ for (a), F-ratio = 8.22, $d.f. = 99$, $P < 0.0001$ for (b), and F-ratio = 8.12, $d.f. = 99$, $P < 0.0001$ for (c), and F-ratio = 36.04, $d.f. = 109$, $P < 0.0001$ for (d)). Each bar represents an average of 300 simulation runs and error bars indicate 95% confidence interval. The frequency of OmpF$^+$ evolution in our simulations depends mainly on whether λ spends any time on the ancestral landscape before shifting to $malT^-$ and there is little evidence of a relationship between the amount of time spend on each landscape and the probability of evolving OmpF+. However, we can reason that as the time spent on the ancestral landscape approaches zero, the frequency of OmpF+ will drop and approach the frequency caused by the constant $malT^-$ landscape. This transition must happen sometime between 0 and 10%. e) and g) show the frequency of OmpF$^+$ evolution when shifting is done between the same hosts, and f) and h) show difference between this switching treatment, and the varying host treatment. The bars are the averages taken across all switching times. Statistics for two-sample t-tests: f) $t = 9.10$, $d.f. = 169$, $P < 0.0001$; h) $t = 8.54$, $d.f. = 161$, $P < 0.0001$.

the final steps are more likely to occur if the host coevolves resistance. To verify this result with laboratory experiments, we analyzed the path λ took to OmpF$^+$ in a single population cryopreserved from the previous coevolution study (population 'D7' in Meyer et al. 2012 [103]; Figure 3.1). We sampled λ strains from different timepoints and sequenced their J gene (Figure 3.7a, Figure 3.14). Next, we ran pairwise competition experiments between λ genotypes at different stages of evolution on the two hosts. We found that the first mutation on the line of descent to OmpF$^+$ required ancestral *E. coli* to evolve, while the second mutation required $malT^-$ *E. coli* (Figure 3.7b and 3.7c). In addition, the OmpF$^+$ genotype with five J mutations only outcompeted the genotype with two mutations when provided with $malT^-$ hosts (Figure 3.8). These findings show that the path λ took in population D7 required it to sequentially adapt to both host types and that λ's fitness landscape changed during coevolution in a way that facilitated an innovation.

At this point, we shifted the focus of our studies to testing whether *E. coli*'s resistance evolution was also impacted by λ's evolution. Before reconstructing J evolution, it was believed that *E. coli* evolved resistance and then J mutations evolved in response [103]. However, J evolved within a day, while it was previously shown that $malT^-$ mutations take about a week to evolve [103]. The timing suggests that λ improved infectivity and then applied pressure on *E. coli* to evolve resistance. To test whether λ evolution promoted host resistance evolution, we ran competition experiments between ancestral and $malT^-$ hosts in the presence of phages isolated from four different time points. We found that $malT^-$ was not significantly more fit than the ancestral *E. coli* in the presence of the ancestral λ, but it was more fit in the presence of the evolved λs (Figure 3.7d). This result combined with the others suggests that there is an intricate coevolutionary feedback at play between λ and *E. coli*: λ evolves J mutations that better exploit *E. coli*, which in turn applies pressure on *E. coli* to evolve resistance. Once resistance evolves, new adaptive pathways become available to λ that encourage the innovation. For the computer simulations, we arbitrarily chose timepoints to switch from one host to the other; however, in reality, the dynamics of the switch are dictated by the host-parasite coevolution.

3.6 Evolutionary replay experiments

To further test the role of coevolutionary processes at driving λ's innovation, we ran replays of the coevolution experiment. We initiated 12 populations with $malT^-$ host that already possessed resistance, and 12 populations with ancestral host where λ and *E. coli* would coevolve normally. The former treatment should hinder the evolution of OmpF function because it denies λ the opportunity to evolve first with ancestral *E. coli*. In line with our expectations, 0 of 12 replicates evolved OmpF use in the $malT^-$ treatment, meanwhile 3 out of 12 evolved the innovation in the ancestral treatment (Figure 3.9a

and 3.9b). By sequencing J alleles of the resulting λ genotypes, we found that fewer mutations evolved with $malT^-$ despite evolving for the same length of time. This suggests that λ's evolution was stymied by starting with the resistant host (Figure 3.10), and by disrupting the coevolutionary process we interfered with λ's ability to innovate.

Lastly, we tested whether $malT^-$ would promote the evolution of OmpF use if the replay experiments were initiated with λ genotypes positioned further along the path to OmpF$^+$. We initiated two more replays: one with a λ strain that was just a single mutation away from OmpF$^+$, and another that was two mutations away (see Figure 3.21 for J alleles). 16 of 24 λ populations evolved OmpF$^+$ showing that whether $malT^-$ promotes OmpF$^+$ evolution depends on where the λ genotypes are located in the fitness landscape (Figure 3.9c and 3.9d).

3.7 Conclusion

Taken together, our studies show that the fitness of a parasite depends on complex genetic interactions within its own genome and with the genomes of interacting hosts. These interdependencies result in highly contingent evolution where λ is unlikely to evolve an innovation unless it participates in a particular sequence of coevolutionary steps with its host. Despite the stochasticity that is expected to arise in systems with substantial historical contingency [64, 18], λ's evolution to use a new receptor is repeatable because the sequence is coordinated by coevolutionary feedbacks. While coevolution may yield tangled banks of interactions, we demonstrate how high throughput technologies can be used to untangle them and to predict evolution despite their complexity. With this approach, we were able to provide direct experimental evidence that fluctuating landscapes, also known as fitness seascapes [100], can promote evolutionary innovations.

3.8 Materials and Methods

3.8.1 General Laboratory techniques

Bacterial and phage strains

The ancestral phage λ strain in our study is cI26; it is a strictly lytic strain that was used in the previous coevolution study on which this paper builds [103]. λ has two life cycles; lytic and lysogenic. In the lytic life cycle, λ infects its host, creates multiple copies of its genome, and lyses the cell to produce new viral particles. This is unlike the lysogenic life cycle of λ, where it stably integrates its genome into the host genome and is replicated along with the host [70]. The λ strain cI26 has a frameshift mutation in the cI gene that disrupts the regulatory protein cI required by λ to switch to a lysogenic life cycle [103]. This renders cI26 obligatory lytic. The other λ strain we used was cI857 (provided by Ing-Nang Wang, State University of New York at Albany) to construct our λ genomic library using MAGE (discussed in section 3.8.2). The advantage to using cI857 over cI26 is that it is able to form a lysogen and as a genomic complex with *E. coli*'s genome, its genome can be easily edited with *E. coli* engineering methods. cI857 has an advantage over typical lysogenic λ because it has a mutation in the cI gene that makes the repressor protein cI unfold at high temperatures [102]. This enables us to induce lytic life cycle of cI857 from lysogens using heat shock treatments, which has fewer side effects compared to using mutagens that the typical strain requires for induction.

We used *Escherichia coli B* strain REL606 for the ancestral-sensitive host and its derivative EcC4 for the evolved-resistant ($malT^{-}$) host to construct fitness landscapes (in Figure 3.5) and run competition assays (in Figure 3.7). REL606 was the ancestral host in the coevolution experiments performed by Meyer et al. [103], and EcC4 was isolated from one of the coevolving populations that has a single mutation, a nonsense mutation (C→T) at genome location 3,482,567 in *malT* gene (Table S5 in [103]). MalT is a positive

regulator of *lamB*, so a disruption in transcription of *malT* inhibits LamB expression [21]. Resistance to λ in EcC4 due to this nonsense *malT* mutation has been shown to yield high levels of resistance, the equivalent of a nonsense mutation in lamB (Fig. S5 in [30]). We used another $malT^-$ mutant of REL606, named LR01, for our coevolutionary replay experiments (Fig. 3). This strain has a 25-bp duplication at genome location 3,482,677 causing a frameshift in *malT*. LR01 was also isolated from a coevolving population of λ and *E. coli* (Fig. S3 in1), however, unlike EcC4, LR01 has a high reversion rate to $malT^+$ that results in "leaky resistance" to λ [30]. This allows λ populations to sustain serial dilution transfers, and thus was critical for the success of the coevolutionary replay experiments (Figure 3.9).

A number of *E. coli* strains were used for culturing λ. The most often used was DH5α, a $lacZ\alpha^-$ derivative of *E. coli* K-12, because it is permissive to all λ genotypes and it lacks *lacZ* which is used as a genetic marker to distinguish phage genotypes in competition assays (additional information in 'Phage Competition Experiments' section). Two $lamB^-$ mutants of *E. coli* with nonfunctional LamB were used interchangeably to culture OmpF^+ λ on Petri dishes. One strain was a derivative of REL606 that has a 1-bp insertion of nucleotide T between base positions 610 and 611 in lamB [103, 104], and the other strain was an *E. coli* K-12 derivative from the Keio collection ($lamB^-$ JW3996) [9]. We found no difference between the efficiency of plaquing on two strains.

Multiplexed Automated Genome Engineering (MAGE) was performed in an *E. coli* K12 strain, HWEC106, provided by Harris Wang, Columbia University. The strain's *mutS* gene is deleted and it possesses the pKD46 plasmid with an inducible λ-red recombineering system [36]. cI857 successfully integrated into this strain's canonical ATTB site located genomically near the *galK* gene [103, 93, 115].

Media

We performed most experiments in conditions identical to the initial coevolution experiment1. For competition assays (low and high throughput) and two of the four coevolutionary replay experiments (Figure 3.9a and 3.9b), bacteria and phage were co-cultured in modified M9 Glucose (47.7 mM disodium phosphate, 22.0 mM potassium phosphate monobasic, 18.7 mM ammonium chloride, 8.6 mM sodium chloride, 0.1 mM calcium chloride, 10 mM magnesium sulfate and 5.55 mM glucose) as used in Meyer *et al.* [103]. A new medium was used for the remaining two coevolution replays (Figure 3.9c and 3.9d) that we call Tris DM Glucose: 1.6 mM potassium phosphate monobasic, 0.59 mM potassium phosphate dibasic, 0.2 mM of calcium chloride, 50 mM Tris base (pH 7.4), 10 mM of magnesium sulphate, 7.5 μM thiamine, 3.2 mM ammonium sulphate, and 5.55 mM glucose. This medium is improved over M9 Glucose because it is less prone to magnesium precipitation, and fortunately has no noticeable effect on the coevolution.

For isolation, estimating densities and initial culturing of bacteria and phage, four more media were used according to the following specifications— a) LB (Lennox Broth): 10 g tryptone, 5 g yeast extract, and 5 g sodium chloride per liter of water, b) LBM9: 20 g tryptone, 10 g yeast extract, 12.8 g sodium phosphate heptahydrate, 3 g potassium phosphate monobasic, 0.5 g sodium chloride, 1 g ammonium chloride, 1.2 g magnesium sulfate, 22 mg calcium chloride per liter of water, c) LB agar: 10 g tryptone, 5 g yeast extract, 5 g sodium chloride, and 16 g agar per liter of water, d) Soft agar: 10 g tryptone, 1 g yeast extract, 8 g sodium chloride, 7 g agar, 0.1 g glucose per liter of water, supplemented with a final concentration of 2 mM calcium chloride. The soft agar was also supplemented with 10 mM magnesium sulphate to improve plaquing of phage particles.

Isolation and culturing techniques

Most strains were grown at 37°C, except for HWEC106, which was grown at 30°C because pKD46 has a temperature sensitive origin of replication. To ensure uniform aeration and nutrient availability, 4 ml cultures were shaken at 220 rpm and 10 ml cultures at 120 rpm; different rpm was chosen to ensure equivalent aeration in cultures (10 ml cultures were grown in 50-ml Erlenmeyer flasks and 4ml in glass tubes). We used LBM9 to grow phage cultures, and LB to grow bacteria cultures, unless otherwise indicated. All bacteria and phage stocks were stored by freezing 1 ml of culture with 15% glycerol at -80°C. We revived phage from freezer stocks by growing $\sim 2\mu$l of frozen stocks on 100 μl of DH5α overnight culture in 4 ml of LBM9. To harvest phage from this, cells were killed and separated by adding 100 μl of chloroform and centrifuging the solution at 3900 rpm for 10 min. The supernatant containing phage lysate was then stored with $\sim 2\%$ chloroform at 4°C. Bacteria were revived from -80°C by growing $\sim 2\mu$l of the frozen stocks over night in 4 ml of LB.

Phage strains were isolated from a population by infusing phage particles into bacterial lawns of DH5α cells. We made these infused plates by mixing a small volume (between 10 and 100 μl) of diluted phage and $\sim 25 \times 10^8$ host cells to a 4 ml of molten soft agar at 55°C, and pouring the mixture over an LB agar plate [121]. The soft agar was allowed to solidify and then incubated overnight at 37°C for phage particles to form plaques. A plaque is a near-circular clearing that forms on the bacterial lawn when nearby cells are killed by an infection that is initiated by a single phage particle. Phage dilutions were made in saline solution (8.5 g/L NaCl) and implemented to yield between 30 and 300 plaques. A single plaque was picked from the infused plates and clonal phage stocks were made from it by culturing it overnight with DH5α and extracting phage lysate from it using techniques similar to frozen λ stocks.

Estimation of phage densities

Infused plates with appropriate dilutions of phage in saline solution (8.5 g/L NaCl) were used to estimate phage densities for phage competition assays. We controlled the number of plaques on an infused plate so that individual plaques could be identified and counted. These counts were used to back-calculate phage growth rate.

To estimate the phage densities in the coevolutionary replay experiments, serial dilutions of phage were spotted on a bacterial lawn of DH5α. We made the bacterial lawns by adding $\sim 5 \times 10^9$ cells of DH5α to 10 ml of molten soft agar at 55°C, and pouring it over a 150-mm diameter Petri dish of LB agar base11. After the soft agar solidified, 2 μl of eight different phage dilutions was added onto the surface, let dry, and incubated overnight at 37°C. The plaques were counted from the dilution where we could identify individual plaques. We then used these counts to calculate phage density in pfu per ml, where pfu is plaque forming unit.

Sequencing for analyses of mutations in J

We routinely sequenced the J gene of λ to verify the identity of stocks and to determine whether J evolved in the re-play experiments. DNA samples were PCR amplified using Q5® High-Fidelity 2X master mix (New England Biolabs) and primers described in Figure 3.15. Unpurified PCRs were submitted to Genewiz (La Jolla, CA) for Sanger sequencing.

Calculation of selection rate using Malthusian parameter

Selection rate (s) was used to quantify the difference in fitness between pairs of strains (say X and Y) in a given environment. It is calculated as the difference of their Malthusian parameters:

$$s = \frac{\ln \frac{X_T}{X_0} - \frac{Y_T}{Y_0}}{T} \tag{3.1}$$

where X_t and Y_t are the densities of strains X and Y, respectively, at time t, and T is the period of time over which strains are assessed (and the assay starts at $t = 0$). Selection rate has units of inverse time, however, in all our figures, we report selection rate as per unit time period (T) of the assays. We reported the values this way because λ does not strictly grow exponentially and it can be misleading to report normalized rates after dividing by T because the rates cannot be extrapolated for different times in a straightforward way like for exponential growths. That being said, for readers who want the per hour rates, we have provided the value of T in hours within each figure legend.

3.8.2 Fitness landscapes

Construction of λ genomic library using Multiplexed Automated Genome Engineering (MAGE)

Our goal was to measure the fitness of a significant number of λ genotypes to establish the basic structure of the landscape. To accomplish this, we generated a combinatorial library of genotypes made with 10 J mutations previously observed to evolve en route to OmpF$^+$ (Figure 3.1). These mutations were chosen because they fell within two ¡100 base frames where adaptive mutations tended to evolve. We used a genetic engineering technique called Multiplexed Automated Genome Engineering (MAGE) to construct the library [144]. This technique employs the λ-red recombineering system that can efficiently recombine single stranded DNA into *E. coli*'s genomic DNA. To engineer λ, λ's genome is integrated into *E. coli*'s genome creating a lysogen. The λ genome in the lysogen (a prophage) becomes dormant and can be treated as any other *E. coli* gene. Gene edits in this λ lysogen are made by expressing λ-red from a plasmid (pKD46) within *E. coli* cells8

and electroporating synthetic oligos with specific mutations written into the sequences. Multiple oligos can be combined into a single experiment in order to create a diverse library of different combinations of the mutations. The output of this procedure is a single sample with a library of genotypes mixed together.

We previously reported our MAGE protocol in Maddamsetti *et al.* [93]. The oligos used are in Table S1 of the manuscript listed under the subheading '10-mutation library'. There are a few important aspects of the oligo and protocol design that we discuss here, but for more detailed information see the original paper.

Our goal was to construct a genetic library with each combination of mutations represented equally, this means a 50/50 split of mutant and wild type states at each site. To achieve this, we modified the typical MAGE protocol. First, when mutations were clustered near each other such that a single 90-mer used for MAGE editing encompassed multiple mutations, we designed multiple oligos with different combinations of the mutations at these nearby sites. Without doing this, edits would be correlated, creating an imbalance in the network representation, and limiting our ability to determine the effects of individual mutations. Secondly, we designed oligos with the wild type state at the 10 positions. Including these oligos slowed the efficiency of MAGE to introduce mutations because sometimes they would overwrite an edit, however, it safeguarded the procedure from saturating the library with the 10 mutations. A third strategy was employed to even out the variation at these ten sites. This was to perform MAGE starting from two orthogonal points in the genetic network; λ that had none of the 10 mutations and one that we engineered in all 10 [93]. Additionally, we ran MAGE on three separate replicates from each starting point in order to enhance the potential to explore more genetic space. Lastly, we ran the MAGE protocol for 50 cycles, which is the number of cycles a computer simulation of MAGE with an efficiency of 10% recombination efficiency per cycle predicted we would need to construct all variants. In the end, we mixed all six MAGE libraries together in

equal frequencies to maximize diversity in the library.

Our design also included measures to improve our ability to detect rare variants in the library. The next step of the fitness landscape protocol is amplicon sequencing with an Illumina MiSeq. This technology has relatively low rates of error ($\sim 10^{-3}$ per base sequenced), however, we anticipated that some of the genotypes in our library would fall below this frequency, making it impossible to distinguish between false positives and true reads. To improve our resolution, we edited a synonymous mutation adjacent to each focal mutation. We called this a watermark mutation because it helped us distinguish between a true genome edit and error in sequencing. By only recording the presence of a focal mutation if it occurred alongside the watermark, we decreased our detection limit from $\sim 10^{-3}$ to $\sim 10^{-6}$.

The watermark strategy we employed was slightly more complicated than this. Most synonymous mutations at the C-terminal end of proteins, like our watermark mutations, have no fitness effects [79]. Despite this, we designed a way to test for neutrality. For each focal mutation we edited one of two different watermark mutations (see Figure 3.16 for list of mutations). This allowed us to compare equivalent genotypes with distinct watermarks in order test whether one of the synonymous mutations impacted λ's fitness when compared to the other. We found that there was no significant fitness effect of the neutral mutations on the phage genotype (see section 3.2.1).

Empirical fitness landscape resolved through one-pot competition experiment

After creating the λ genomic library, we measured the relative fitness of genotypes with respect to the wild type λ. To do this, we cocultured λ genomic library and ancestral λ in a 1:9 ratio. This ratio was used so that the most abundant competitor remains the ancestor throughout the competition experiment. This means that the competitive fitness we measured is with respect to a single genotype. If we had not done this, then as the

community of engineered λ shifts during the competition, and more fit genotypes become enriched, they will change the mean fitness of the population and cause mildly fit genotypes to begin to decline, making them appear unfit. Flooding the flask with ancestral λ and running competitions for a short time period (4 hours) solves this problem.

Having a disproportionate number of ancestral λ has a pitfall. We measured λ fitness by comparing the frequency changes of each genotype using amplicon sequencing (more information in the section 3.8.2). The problem is that most of our sequencing effort would be spent on sequencing a single genotype; the ancestor. To avoid this, we engineered an ancestral genome with 6 synonymous mutations where the reverse primer binds. These edits interfered with primer binding and caused selective amplification of the library λs. We edited 6 'wobble' positions; 3381 (g→t), 3384 (c→a), 3387 (c→a), 3390 (c→t) 3393 (g→a), and 3396 (c→a). The engineering was done using MAGE with the oligo provided in Figure 3.18. These edits were made at the end of the protein (3,399 nucleotide) and so they likely did not have an effect on λ fitness. We did not test this because this engineered strain acted as a standard competitor for all competition experiments and so it should have no effect on the relative fitness comparisons.

We ran eight competitions, four in the presence of ancestral *E. coli* and four with *malT*$^{-}$ *E. coli*. Competitions were inoculated with $\sim 10^7$ total λ particles and $\sim 2 \times 10^8$ cells into 10 ml of M9 glucose in 50 ml flasks. The cells were preconditioned in M9 glucose for 24 hours before the competition. Flasks were cultured for four hours at 37°C and shaking at 120 rpm. 1 ml samples were removed from the library before the competitions, and after the competition for processing. Phage particles were concentrated using PEG precipitation [121]. The pellet was resuspended in 25 μl of molecular grade water and then a two-step PCR reaction was performed in order to amplify just the region of interest, to attach barcodes to the amplicons so that we could multiplex samples into a single Illumina run, and attach adapters required for Illumina sequencing. The protocol was published

by Kelsic et al. [79]. The primers we used are provided in Figure 3.19. As described in Kelsic et al. [79], we ran two separate PCR reactions for each sample and used a unique barcode on each. This allowed us to test for amplification bias, which we did not detect (Figure 3.17).

Sequencing

Amplicon samples were pooled together and sequenced using an Illumina Mi-Seq maintained in the Systems Biology Department at Harvard Medical School. 100 base paired end reads were run. We were able to extract on average 107,839 high quality reads per sample with standard deviation of 26,889. This provided considerable coverage to reliably estimate fitness for even rare genotypes.

Post-sequencing analysis and construction of fitness landscape

Amplicon sequences were analyzed to calculate selection rates for each engineered genotype. These values were later used to construct the fitness landscapes. We used a combination of custom Python and MATLAB (version R2019b) scripts to read and concatenate the raw paired-end reads, identify the genotypes based on the focal mutations they possessed, and to count their abundances. All focal mutations were called only if either of the two corresponding watermark mutations were also present. We removed any reads that contained more than one mutation other than the focal and watermark mutations. This essentially acted as a quality filter and we did not have to filter based on the Illumina provided Q-score. For each competition, a mean of counts corresponding to the two barcodes used in PCR amplifications was taken for all the genotypes. If the counts for both, initial and final timepoints were available for a genotype, they were used to estimate the genotype's Malthusian growth rate, given by the calculation $\left(\ln \frac{\lambda_{i,T}}{\lambda_{i,0}}\right)/T$ where $\lambda_{i,t}$ is the density of the given genotype at time t and T is the total time of the competition. Note

that most genotypes performed poorly on $malT^{-}$ host and fell below detection limit at the final timepoint. We quantified the differences in fitness of all genotypes by calculating their selection rate with respect to the wild type λ (see 'Calculation of selection rate using Malthusian parameter' section). The final fitness landscapes (Figure 3.5a and 3.5b) were constructed by taking the mean of the selection rates obtained from the four replicate competition trials performed on the same host.

Statistical analysis of fitness landscapes

We used multiple linear regression models to quantify the genetic interactions in the fitness landscapes. We regressed fitness values on each landscape against individual effects of the mutations and the pairwise epistatic interactions between them (genotype-by-genotype or GxG interactions). This allowed us to estimate the contribution of epistasis towards explaining the observed fitness data. The fitness of each genotype in the regression model was thus described by 55 predictor variables (10 main effect terms for the ten mutations + 45 interaction terms):

$$y = \beta_{**} + \sum_{i} \beta_{i*} G_i + \sum_{i<j} \beta_{ij} G_i G_j \tag{3.2}$$

where β_{**} is the intercept, all other βs are regression coefficients, contribution of an individual mutation i is described by the term G_i, and the effect of pair of mutations i and j are captured by terms $G_i G_j$, where G_i is an indicator variable that is equal to 0 when the mutation i is absent and 1 when present. After fitting the model, we used Benjamini-Hochberg procedure [14] to control for false discovery rates and identify statistically significant terms in the model (see Figure 3.3). All analyses were performed in R version 3.6.1 [2]. The error in fitness estimates of the genotypes did not depend on the mean fitness value of the genotypes (Figure 3.12).

To understand how λ's landscape differed between the two hosts, we incorporated the host genotype as a predictor variable in the linear model and regressed the combined

fitness data of both landscapes together. This allowed us to test if a mutation significantly interacted with the host (genotype-by-environment or GxE interaction) and whether interaction between pairs of mutations changed with the host genotype (genotype-by-genotype-by-environment or GxGxE interaction). The combined full-factorial model consisted of a total of 111 terms (10 G for individual mutation + 1 E for host + 45 GxG + 10 GxE + 45 GxGxE terms),

$$y = \beta_{***} + \sum_i \beta_{i**} G_i + \beta_{**+} E + \sum_i \beta_{i*+} G_i E + \sum_{i<j} \beta_{ij*} G_i G_j + \sum_{i<j} \beta_{ij+} G_i G_j E \qquad (3.3)$$

where additional E is an indicator variable for host type and other notations follow same scheme as in Eq. (2). Since the fit of a model generally improves as the number of predictor variables increases, we tested for overfitting using Akaike information criterion (AIC) [5]. AIC is a penalized-likelihood criterion which penalizes a model for increasing number of parameters in it. The model with the greatest relative likelihood is considered to be the one with minimum AIC value. We minimized AIC value using the step function in R [2] to uncover subsets of predictor variables that have high predictive power. This resulted in a more parsimonious model (77 terms out of a total 111, see Figure 3.4). A complimentary adjusted R-squared analysis came to the same conclusion, $R^2_{adj} = 0.775$ for the reduced model and $R^2_{adj} = 0.769$ for the full-factorial model (higher values indicate more predictive and parsimonious models). After model selection, we used Benjamini-Hochberg procedure to determine which variables were significantly predictive (Figure 3.3). This procedure only controls for false positives and is susceptible to false negatives. In this way, it is conservative for our purposes.

Simulation of λ's evolution on the landscapes

To test if the changes in λ's fitness landscape facilitated its evolution to infect via OmpF, we simulated λ's evolution on the two landscapes we measured and recorded whether OmpF$^+$ genotypes arose to a high enough frequency that we would have detected them in the original laboratory evolution. We initiated λ populations with no mutations and allowed them to evolve and mutate at any of the ten focal sites used to construct the landscapes (Figure 3.16). For each treatment (Figure 3.5c) we simulated 300 separate trials using a modified Wright-Fisher model with discrete generations and a fixed population size.

For each simulation, a new fitness landscape was constructed by assigning fitness values to all the genotypes ($2^{10} = 1024$ genotypes). This was done in a way to account for error in estimating fitness and the error associated with imputing the values of missing data points. For the genotypes that had empirical fitness data available (Fig 3.5a and 3.5b), we did not simply average the values of the replicate fitness measurements, but instead we performed a bootstrapping protocol in order to account for the error in the fitness estimate. We randomly resampled from the four replicate measurements (with replacement) and then computed the mean of the four. Some genotypes were not present in all four replicates, in these cases we resampled as many times as there were replicates. Genotypes with only one replicate data point were, thus, assigned the same corresponding fitness values in all the runs of the simulations. Next, we imputed fitness of the genotypes that were missing from the empirical landscape. A genotype with a missing fitness value was randomly chosen and assigned the mean of the fitness values of its nearest neighbors (one-mutation away genotypes) present in the landscape. This was iterated until the full genotypic space was complete. Note that the order in which genotypes are chosen can affect the value that is estimated for a given genotype. This is because as the landscape is filled in, each genotype will have more neighbors to draw values from. This means that if a genotype is randomly chosen early, its fitness will be based on fewer neighbors than if it were chosen later, and

its value will be slightly different. Since missing genotypes are randomly chosen in each iteration, the order will vary from one simulation to the next. This method introduces an extra source of variation in the simulation runs and captures uncertainty associated with the imputation of fitness values.

After constructing a complete fitness landscape, we evolved a λ population through repeated cycles of reproduction, selection, and mutation. For each generation, reproduction in the population was simulated by a multinomial sampling where the number of trials was equal to the population size N (set to $\sim 6.3 \times 10^9$ based on Figure 3.13) and the success probability associated with a genotype i was given by $p_i = n_i w_i / \sum_i n_i w_i$, where n_i is the abundance of the genotype i, and w_i is defined as the exponential of selection rate used in fitness landscape. Thus, the probability of $(k_1, k_2, k_3, \cdots k_m)$ offsprings for genotypes $1, 2, 3, \cdots m$ would be:

$$\binom{N}{k_1, k_2, k_3, \cdots k_m} \prod_{i=1}^{m} \left(\frac{n_i w_i}{\sum_j n_j w_j} \right)^{k_i} \tag{3.4}$$

To incorporate mutations, all genotypes whose frequencies increased were mutated as per λ's mutation rate (7.7×10^{-8} per base per replication [44]). Consider a genotype (say i) that increased in abundance and let the number of additional individuals produced by this genotype be denoted by z_i (with $z_i = k_i - n_i$). Each of these z_i individuals retained its parent's genotype with a probability $e^{-\mu}$ (assuming a Poisson distribution for mutations). Otherwise the individual was assigned to a random neighboring (i.e. mutant) genotype. Given λ's mutation rate, the probability of multiple mutations is very small and was ignored here. We simulated this modified Wright-Fisher cycle of reproduction and mutation for 960 generations. It is difficult to know how many generations phages undergo because the evolved phages have a high spontaneous death rate10 and can also experience other sources of mortality. Given this, we decided to run the simulation for a somewhat arbitrary amount of time, 960 generations which corresponds to two doublings per hour for the 20-day

experiment we are trying to replicate Meyer *et al.* [103]. This is likely an overestimate, which is unintuitively conservative for our purposes because more cycles will cause additional evolution and exploration of the fitness landscape, enhancing the possibility of OmpF$^+$ evolution in the negative controls, and reducing our ability to detect treatment differences.

We evaluated whether λ evolved OmpF$^+$ by examining the population for genotypes that had the necessary mutations to be OmpF$^+$. Previous studies revealed that OmpF genotypes must have four mutations [93]. They all possess two specific changes (A3034G) and (G3319A) and a third change that can occur at positions 3320 or 3321. For the 10 mutations we studied, T3321A is the only third mutation that satisfies this requirement, meaning all OmpF genotypes must have three specific changes (A3034G, G3319A and T3321A). Many J mutations satisfy the last requirement [93], so we implemented what we call the '3+1' rule where genotypes are designated as OmpF$^+$ if they have the three necessary J mutations plus any additional mutation (see Figure 3.16). If any such genotype crossed the threshold of 5,000 λ particles during the course of a simulation, the λ population in that run was marked to have evolved OmpF-function. We based the threshold value on the detection limit of OmpF$^+$ genotypes in the original laboratory coevolution experiments ($\sim$ 500 pfu/ml, see section 3.8.5).

We implemented host switching by first evolving λ on the landscape measured with ancestral *E. coli*, stopping the simulation early and quantifying the frequency of each λ genotype. Next, we would initiate evolution on the $malT^-$ for the remainder of the time but starting with genotypes at the frequencies recorded on the ancestral landscape. A total of 11 different coevolution treatments were run by varying how long λ evolved on the ancestral host landscape and $malT^-$ landscape (Figure 3.5c). 300 simulation trials were run for each treatment. We calculated error in the simulations in order to detect significant treatment differences by batching the runs into 30 and estimating the frequency of OmpF$^+$ for each subset. This allowed us to calculate a 95% confidence interval using a Student's t

distribution for the frequency of OmpF$^+$ evolution in each treatment, and then to use an ANOVA coupled with Tukey's multiple comparison test to test for treatment differences.

3.8.3 Phage competition assays

To test whether λ's path to the innovation requires sequential adaptation to host genotypes isolated from different stages of their coevolution, we studied the dynamics of λ evolution in much more depth than previously reported. We focused on a single experimental replicate reported on in Meyer et al. [103]; 'D7' that evolved to be OmpF$^+$ on the 12th day. Daily samples of the population were preserved so we were able to revive phages from the full time series. We isolated 66 phages in total and sequenced the full-length J gene from each in order to reconstruct the evolutionary dynamics (Figure 3.7a).

Next, we ran head-to-head competition experiments from key genotypes along the path to OmpF$^+$, the wildtype (WT) and λ with a single J mutation (A), double (AB), and quintuple (ABCDE) (Figure 3.7b and 3.7c, Figure 3.8). Our first step was to mark key strains with a gene that caused the plaques to be visually distinguishable from unmarked phage. The gene is lacZα, which in the presence of a 5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside (X-gal), Isopropyl-β-D-thiogalactoside (IPTG), and a host cell like DH5α that lacks lacZα, produces blue plaques. To incorporate lacZα into λ's genome, we used a gene fusion of lacZα with λ's R gene [125, 25]. The fusion readily recombines into λ's genome by the phage's endogenous recombination system, λ-red [48]. The procedure is straightforward, infect an *E. coli* strain that has a plasmid with the lacZ fusion (strain: SYP042; plasmid: pSwtRlaczalphaZalpha+RZ x11 Blue Amp Blue provided by Ing-Nang Wang, Albany, NY) and some fraction of the λ produced will have recombined with the plasmid. The recombinants are isolated by picking blue plaques. WT and AB were each marked. The marker has been shown to have a slight fitness effect, however, this would not significantly influence our measurements of fitness since we observed such large differences

between strains [125, 25].

We competed three pairs of phages, WT_{lacZ} versus A, A versus AB_{lacZ}, and AB_{lacZ} versus ABCDE, under two conditions; with ancestral *E. coli* or $malT^{-}$. Competitions were run for a single 24-hour period under identical conditions in which the phage evolved [103]. Initial λ density was between $10^4 - 10^5$ particles per ml and the relative frequency of the strains was sometimes skewed in order to start with more of the unfit genotype. Fitness differences were so large that the less fit genotype would be overwhelmed and its frequency undetectable if they did not start with a numerical advantage. $\sim 5 \times 10^6$ exponentially growing cells were inoculated to each flask (5×10^5 per ml). The cells were preconditioned by growing them overnight in the competition medium. Three replicate competitions were run for each treatment. Phage densities of the unmarked and marked λ were determined by plating in typical soft agar plates, where we added 0.5 mg/ml of X-gal and 0.25 mg/ml IPTG to the molten soft agar. Densities were measured at $t = 0$ and $t = 24$ h.

3.8.4 Host competition assay

This competition experiment tested the role of λ's evolution plays in promoting the evolution of host resistance. We competed the ancestral host (REL606) against the evolved-resistant host (EcC4) in the presence of different λ genotypes that had increasing numbers of J mutations (Figure 3.7; WT (cI26), A (1-mutation), ABC (3-mutation) and A**C (4-mutation)). We initiated 3 replicate populations for each phage treatment. The competition was performed with identical conditions as the original evolution experiment and run for 4 h to prevent λ from evolving during the competition assay. Cells were preconditioned in modified M9 glucose for 24 hours before the competition. As with λ, fitness was measured by observing the change in frequency of the competitors over time. Frequencies were determined by plating a subsample of the populations onto tetrazolium maltose (TMal) agar plates [127]. REL606 produces white colonies on the plates, while EcC4

produces smaller red colonies. Before plating, the phages were removed by centrifugation of the cells and then removing the supernatant that possessed the phage. Cells were resuspended in saline solution and the centrifugation and resuspension was repeated once more.

3.8.5 Coevolutionary replay experiments

Would λ's evolution to OmpF innovation proceed without initially evolving on ancestral host? We tested this by replaying λ-*E. coli* coevolution but starting out with a host that already evolved resistance via a $malT^-$ mutation (strain LR01). We initiated 12 replicate populations with $malT^-$, and another 12 with REL606 as a positive control. The replay experiment was run nearly identically to the coevolution experiment performed by Meyer *et al.* [103] with cI26 as the ancestor phage, and daily sampling was done to detect presence of OmpF$^+$ phage by spotting the phage on bacterial lawn of $llamB^-$ cells (detection limit ~ 500 pfu per ml) (Figure 3.20). The two differences were that the study was run longer (31 days) and daily estimates of λ populations were made (Figure 3.13). 12 replicates were chosen in order to have enough statistical power to determine whether or not a treatment reduces the chances of OmpF$^+$ evolution. In the previously reported study, only a quarter of populations (24 out of 96) evolved to use OmpF1; this result has also been replicated in our lab on two separate instances where 3 out of 12 populations gained OmpF-function. Assuming a binomial distribution and a true success rate of 0.25 for λ's OmpF$^+$ evolution with ancestral host, the probability of observing no replicate evolving to be OmpF$^+$ in 12 replicate populations is 0.03167. Thus, we can conclude that $malT^-$ treatment reduced λ's ability to evolve the innovation as compared to when coevolution is initiated with ancestral host. Two isolates from each population were sampled at day 26 and the reactive region of their J genes were sequenced using the Sanger sequencing method.

We ran two additional replay experiments as a positive control for any unintended side effects of initiating an experiment with $malT^-$ host (LR01). A prediction from the fitness landscape simulation results was that if we started a replay with $malT^-$ and phage isolated from a later-stage of the coevolution, then the phage should evolve OmpF use. To test this, we genetically modified an OmpF$^+$ λ by removing one mutation required for OmpF use (λ^{-1}), and then a second (λ^{-2}). The OmpF$^+$ genotype was a lysogen (cI857) that we had previously edited in 7 J mutations (Figure 3.21), which is reported about in Petrie *et al.* [115]. The two new edits were made using MAGE and the oligos are reported in Figure 3.18.

The replay experiments were run identically, except because of an oversight, these replays were run in a slightly different minimal glucose medium called Tris DM (compared to original coevolution experiment run by Meyer *et al.* [103]). The key difference in the medium is the buffer, it is a Tris buffer, not a phosphate buffer. Both mediums impose carbon limitation and have the same concentration of the single carbon source, glucose, and so population and evolutionary dynamics should not have been affected. We confirmed in an additional replay experiment that the medium does not affect the timing or repeatability of OmpF$^+$ evolution.

3.9 Acknowledgements

Chapter 3 has been submitted for publication of the material, with minor edits, as it may appear in Nature Ecology and Evolution, 2021: A. Gupta, L. Zaman, H. Strobel, J. Gallie, A. R. Burmeister, B. Kerr, R. Kishony, E. Tamar, J. R. Meyer, "Coevolution promotes innovation through deformations in fitness landscapes". The author of the dissertation was the primary author of this paper.

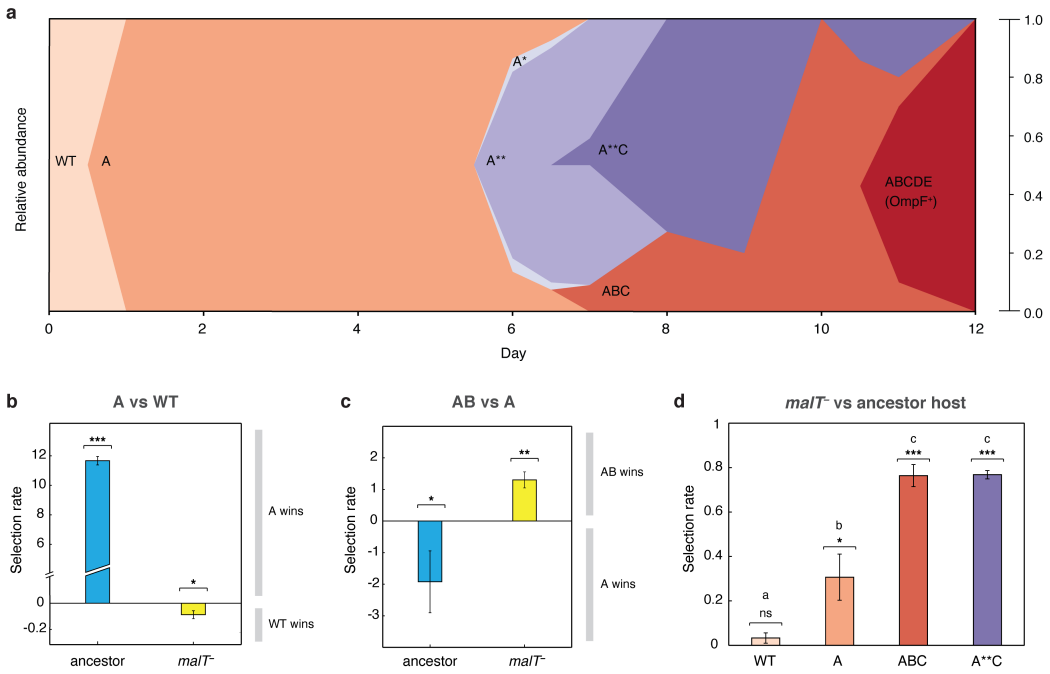

Figure 3.7: J evolution in λ and tests of the interdependency of J mutant and *E. coli* genotype fitness. **a)** Phylogenetic reconstruction and relative abundance of λ genotypes isolated through time from a previously coevolved community [103]. Each letter and star indicate a non-synonymous mutation in J (see Figure 3.14 for labels' corresponding mutations). A genotype's relative abundance on a given day is denoted by the fraction of the total height of the y-axis that it occupies (e.g. on Day 9, frequency of ABC is 0.2 and A∗∗C is 0.8). The lineage WT-A-ABC-ABCDE eventually evolves OmpF function and fixes in the population; resistance in *E. coli* through $malT^-$ rises to high frequencies between days five and eight31. **b) & c)** Selection rates (per 24 h) of phage genotypes on the two hosts. Each bar represents the mean of three experimental replicates. While mutation (A) is favored over wildtype λ in the presence of the ancestral host and not $malT^-$, AB only outcompetes A in the presence of $malT^-$ and not the ancestral host. One tailed t-tests to test if the mean selection rate is significantly greater (or less) than zero: A vs WT with ancestor host- $t = 98.76$, $d.f. = 2$, $P < 0.0001$; A vs WT with $malT^-$ - $t = -4.99$, $d.f. = 2$, $P = 0.0190$; AB vs A with ancestor- $t = 3.4$, $d.f. = 2$, $P = 0.0383$; AB vs A with $malT^-$ - $t = -8.88$, $d.f. = 2$, $P = 0.0062$. **d)** Selection rate (per 4 h) of $malT^-$ *E. coli* relative to its ancestor in the presence of λ from different stages of coevolution. Each competition was replicated three times. Lowercase letters denote significance via Tukey's honest significance test (ANOVA: F-ratio = 111.22, $d.f. = 11$, $P < 0.0001$). One tailed t-tests were also used to test if the selection rate of $malT^-$ was greater than zero- WT: $t = 2.44$, $d.f. = 2$, $P = 0.676$; A: $t = 5.12$, $d.f. = 2$, $P = 0.0181$; ABC: $t = 26.59$, $d.f. = 2$, $P = 0.0007$; A∗∗C: $t = 71.67$, $d.f. = 2$, $P < 0.0001$. This shows that $malT^-$ is unlikely to evolve in the presence of WT λ but it becomes progressively more likely as λ gains mutations. Asterisks over all the competitions indicate significance level corresponding to the P-values. Error bars in all bar graphs represent one sample standard deviation.

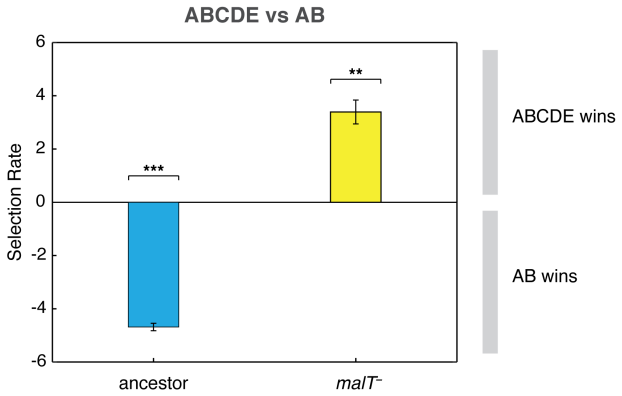

Figure 3.8: Competition assay of λ isolates from population D7. Relative fitness of λ with two mutations (AB) to an OmpF$^+$ λ with five mutations (ABCDE) on two host genotypes. This shows that AB outcompetes ABCDE in the presence of ancestral *E. coli* while ABCDE is favored over AB with $malT^-$ E. coli. Selection rate (per 4 h) is the difference in Malthusian growth rates of the competitors over one day with a value of 0 indicating no difference in fitness. Each bar represents mean of three replicate trials and error bars indicate one sample standard deviation. (One-tailed t-test to test if the selection rate is greater (or less) than zero— ABCDE vs AB with ancestor host: $t = -58.53$, $d.f. = 2$, $P < 0.0001$, ABCDE vs AB with $malT^-$: $t = 13.09$, $d.f. = 2$, $P = 0.0029$. Asterisks over the bar graphs indicate significance level corresponding to the P-values).

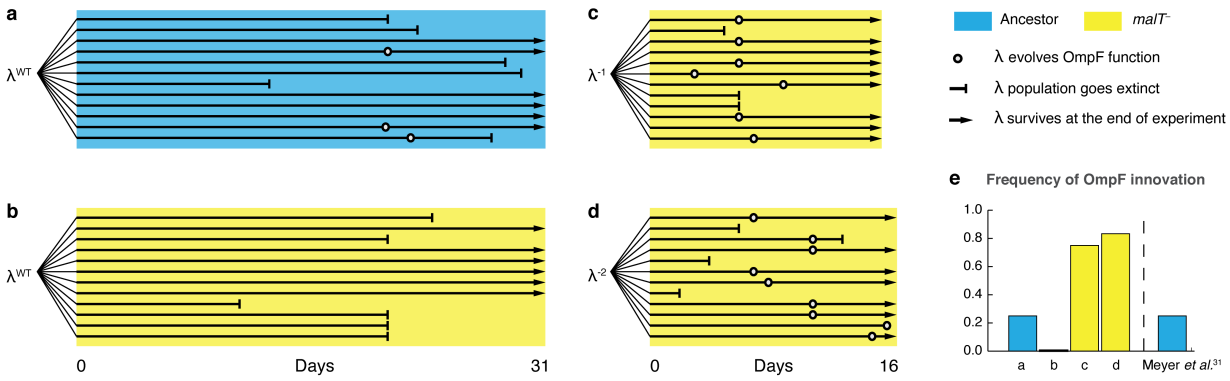

Figure 3.9: Evolutionary replay experiments where time-shifted phage and host pairs were cocultured. **a)** Wildtype λ with ancestral host, **b)** wildtype λ with $malT^-$, **c)** λ one mutation removed from evolving OmpF function with $malT^-$, and **d)** identical setup as (c), but with λ two mutations removed (see Figure 3.10 for identity of mutations). **e)** The bar graph provides the frequency of OmpF$^+$ evolution compared to the frequency observed by Meyer *et al.* [103]. Given λ's established 1 in 4 rate of OmpF$^+$ with ancestral host, the probability of observing no OmpF$^+$ evolution in 12 replicate populations is ~ 0.03. Thus, no positives for OmpF evolution in (b) shows that λ's evolution to OmpF function is significantly hindered when coevolution is initiated with $malT^-$ host. However, $malT^-$ does not impede OmpF$^+$ evolution when the coevolution is initiated with already evolved λs (P-values for Fisher's exact test: between (b) and (c)- $P = 0.0261$; between (b) and (d)- $P = 0.0122$). Notably, some λ populations went extinct which is common for these experiments and was previously shown to be caused by the evolution of resistance mutations in *E. coli*'s ManXYZ protein complex [103].

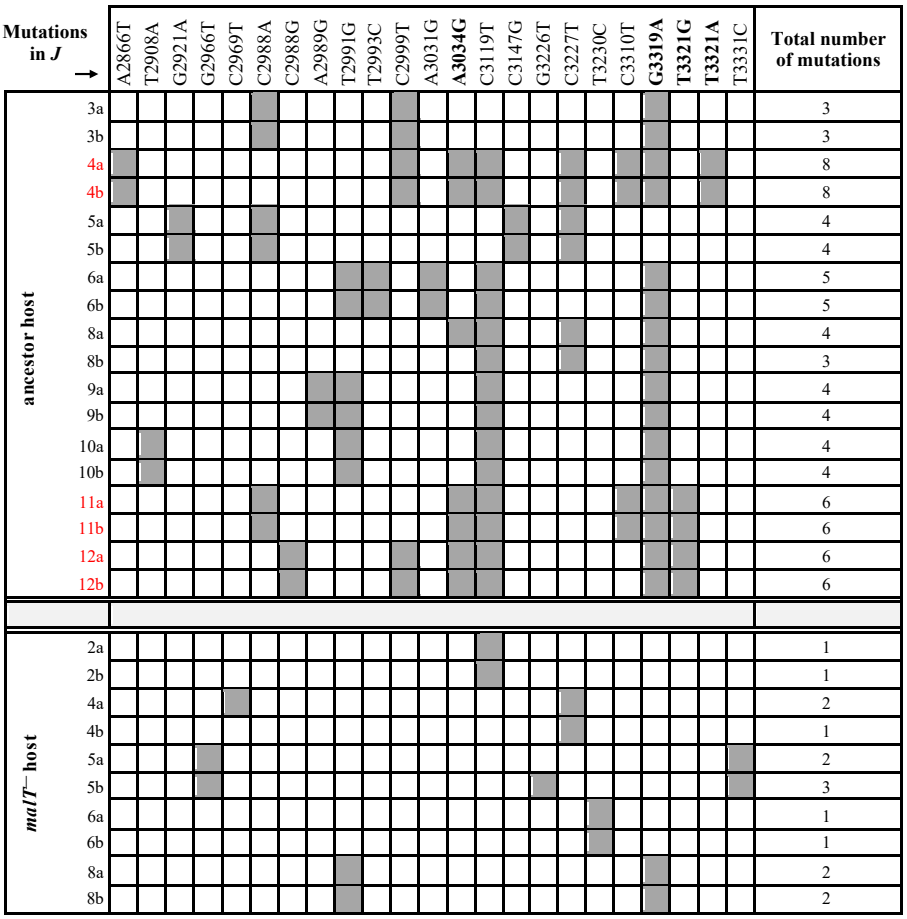

Figure 3.10: Mutations present in λ isolates from day 26 of the coevolutionary replay experiment initiated with **a)** ancestor host and **b)** $malT^-$ host (corresponding to Figure 3.9a and 3.9b). Two strains (a and b) were isolated from each population and the active region of J (roughly between nucleotide position 2,600 and the end) sequenced. Replicates marked in red indicate populations that evolved OmpF-function. Canonical mutations for evolution of OmpF function are bolded. λ evolved more of both the total number of mutations and total number of canonical mutations in replicate populations initiated with ancestor host than with the evolved $malT^-$ host (statistics for difference in total number of mutations: $t = 5.37$, $d.f. = 11$, $P = 0.0002$, statistics for difference in total number of canonical mutations: $t = 3.18$, $d.f. = 11$, $P = 0.0088$; both were tested using two-sample t-test with unequal variances assumed).

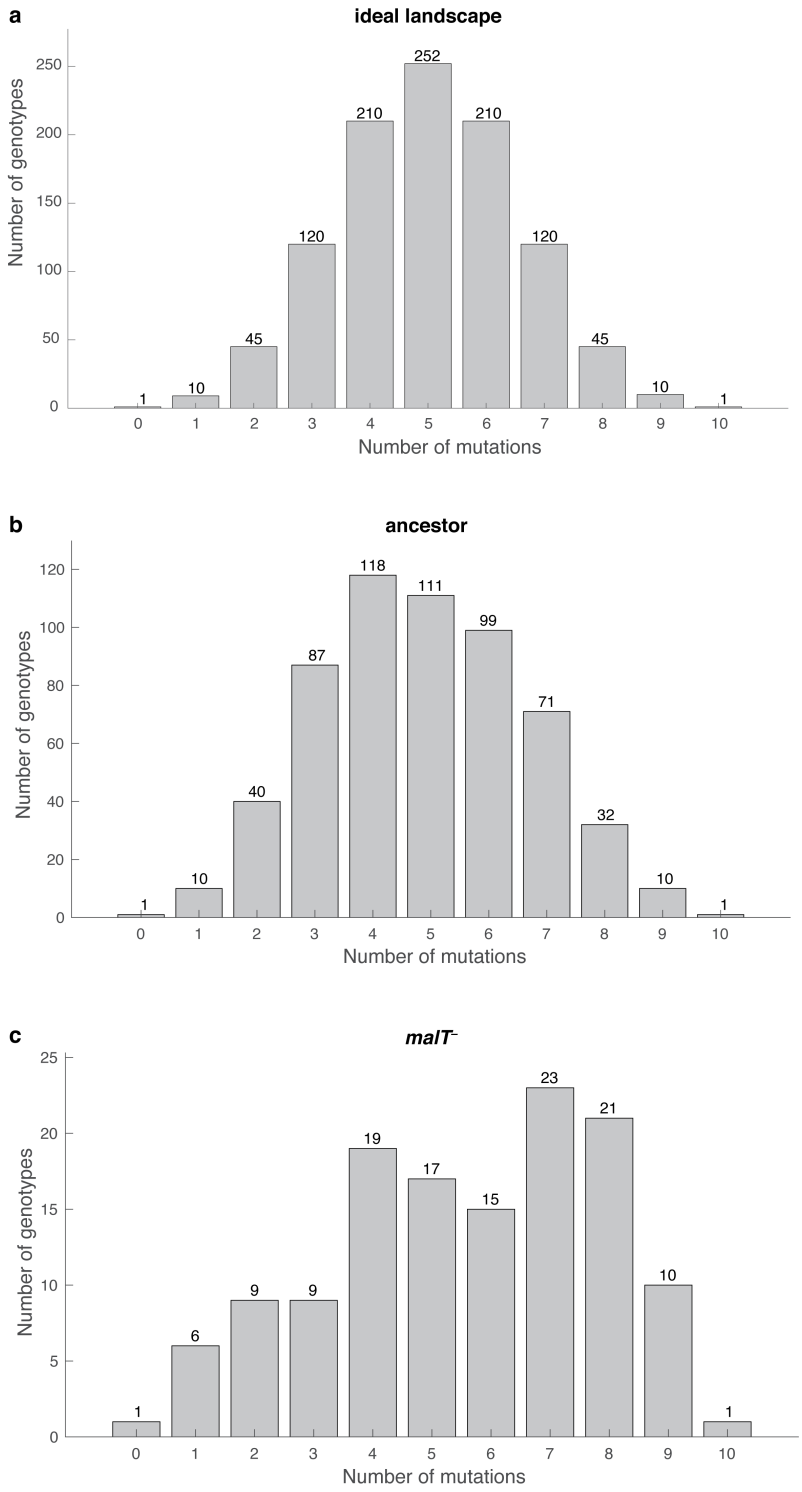

Figure 3.11: Distribution of λ genotypes present with respect to the number of mutations they possess. **a)** An ideal fitness landscape where all combinations of ten mutations are present. **b)** and **c)** Empirical fitness landscapes with ancestral host and $malT^{-}$ host, respectively.

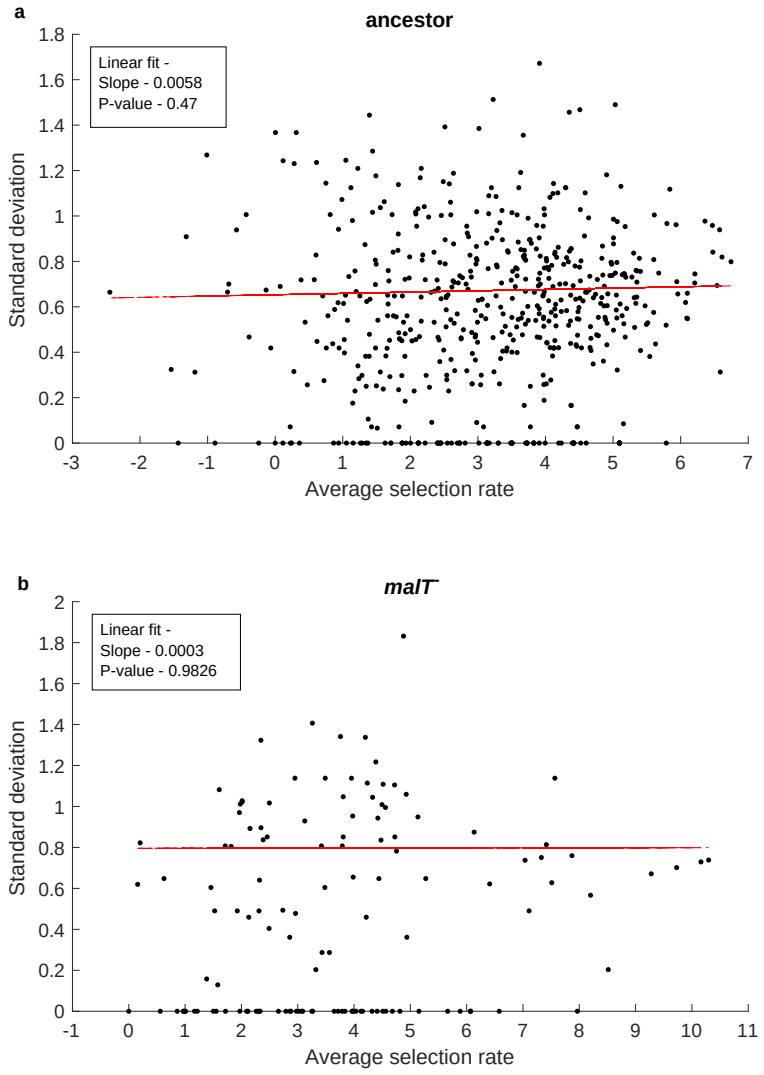

Figure 3.12: Test of whether the error in our estimates of selection rates (per 4 h) is influenced by the magnitude of the estimate. Standard deviation does not correlate with mean fitness of a genotypes for both a) ancestor host, and b) $malT^{-}$ host; P-values in the insets indicate that the slope is not significantly different than 0. Uniform variance in fitness values permit the performance of analysis of variance to detect genotype by genotype by host interactions. Points with zero standard deviation indicate presence of only one replicate population with a fitness value for that genotype.

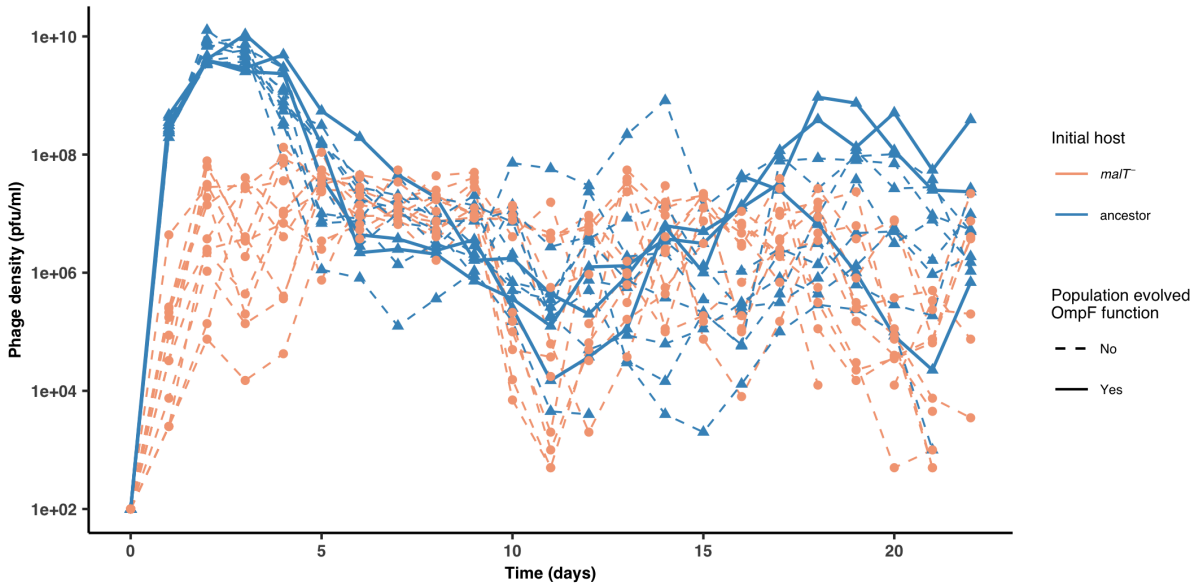

Figure 3.13: Phage densities during the coevolutionary replay experiment performed in Figure 3.9. Timeseries of phage density of 12 replicates measured at the end of each day for cultures initiated with wildtype λ and ancestor host or $malT^-$ host (corresponding to Figure 3.9a and Figure 3.9b). The three λ populations that evolved OmpF function in the ancestor host treatment are marked by solid lines. For the simulations, we used a populations size of 6.3×10^9 which is lower than the peak population density observed ($\sim 10^{10}$), but above the average population density ($\sim 6.3 \times 10^8$).

Mutations / λ Isolates	T2991G (A)	A3031G (A*)	A3034G (E)	G3319A (B)	A3320G (A**)	T3321A (D)	T3380C (C)
WT							
A	■						
A*	■	■					
A**	■	■			■		
AB	■			■			
ABC	■			■			■
ABCDE	■		■	■		■	■

Figure 3.14: Mutations and their corresponding labels in λ genotypes isolated from population D7 in Meyer *et al.* [103]. Red asterisks indicate the particular mutation in a genotype's description in Figure 3.7a.

PCR primers	
Forward Primer (5'-3')	CCTGCGGGCGGTTTTGTCATTTA
Reverse Primer (3'-5')	CGCATCGTTCACCTCTCACT

Figure 3.15: PCR primers for sequencing J gene in λ.

Focal *J* Mutation	Mutation ID used for linear regression analysis	Amino Acid Change	Neutral 1	Neutral 2
C2969T	G1	A → V	G2970C	G2970T
A2989G	G2	I → V	G2985T	C2988T
T2991G	G3	I → M	A2994C	A2994G
C2999T	G4	A → V	G3000A	G3000C
A3034G	G5	S → G	C3033T	T3036A
C3310T	G6	H → Y	G3309A	G3309T
G3319A	G7	D → N	G3315C	G3315A
A3320G	G8	D → G	A3318C	A3318G
T3321A	G9	D → E	T3324A	T3324C
T3380C	G10	L → P	G3378A	G3378C

Figure 3.16: List of focal mutations with their corresponding two watermark mutations. The three canonical mutations required for OmpF-function [93] are marked in bold.

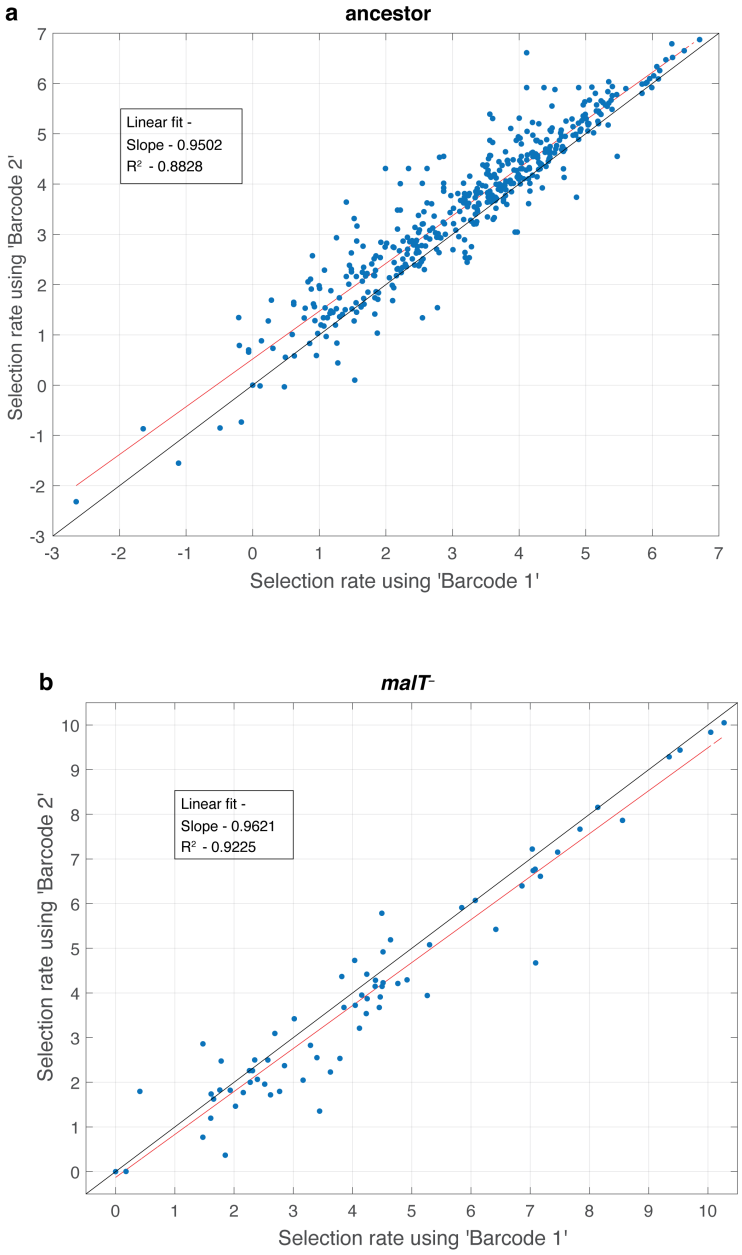

Figure 3.17: Test to determine whether amplification distorts measurements of fitness. For every replicate population in the competition (total eight, four for each host) we performed two separate amplifications and used a unique barcode for each. This allowed us to compare fitness calculations based on independent amplifications and to determine if the amplification step is repeatable and unbiased. Each point represents a unique λ genotype presented in the fitness landscapes reported in Figure 3.5. The red line is a linear least-square fit. The regression fitting parameter and coefficient of determination are provided in the inset. The black line represents the vector where fitness estimates from each barcode are equal.

Mutations introduced	Sequence (5'-3')
A3321T	CATCGCTGGCAAACGTATACGGCGGAATaTTTGCCGAATACCGTGT GGACGTAAGCGTGAACGTCAGGATCACGTTTCCCCGACCCGCTG
G3034A	CATCGGTCACGGTGACAGTACGGGTACCTGACGGCCAGTCCACACt GCTTTCACGCTGGCGCGGAAAAGCCGCGCTCGCCACCTTTACAA
6 'wobble' edits- G3381T C3384A, C3387A, C3390T, G3393A, and C3396A	TAAAACGCCCGTTCCCGGACGAACCTCTGTAACACACTCAtACtACa CTtATtCCaAGCGCCTGTTTCTTAATCACCATAACCTGCACAT

Figure 3.18: Oligos used in MAGE to insert *J* mutations in λ.

Primer label	Nucleotide sequence
J Mage for 2949.2968 6N illum	CCTACACGACGCTCTTCCGATCTNNNNNNgataaacggtacgctgaggg
J Mage for 2949.2968 5N illum	CCTACACGACGCTCTTCCGATCTNNNNNgataaacggtacgctgaggg
J Mage for 2949.2968 4N illum	CCTACACGACGCTCTTCCGATCTNNNNgataaacggtacgctgaggg
J Mage rev 3381.3400 2N illum	GAGTTCAGACGTGTGCTCTTCCGATCTNNctcagaccacgctgatgccc
J Mage rev 3381.3400 1N illum	GAGTTCAGACGTGTGCTCTTCCGATCTNctcagaccacgctgatgccc
J Mage rev 3381.3400 0N illum	GAGTTCAGACGTGTGCTCTTCCGATCTctcagaccacgctgatgccc

Figure 3.19: PCR primers used to generate *J* amplicons. These are custom primers designed for the first PCR reaction that uses λ genomic DNA as the template. The second PCR step uses standard primers listed in Kelsic *et al.* [79]. Each primer is broken up into three sections. The first, capitalized nucleotides, are the annealing region for the second set of PCR primers. The second, N's of variable length, improve our ability to multiplex amplicons since they cause reading frame shifts so that when PCRs mixed together originating from different primers there will be variability among the clusters on the Illumina flow cell, allowing the machine to more easily distinguish clusters and reduce sequencing errors. The last segment of lowercase letters indicates the chromosome annealing region. The numbers in the primer label (2949.2968 or 3381.3400) indicate which nucleotides the primer anneals to in *J*. Note, *J* is only 3,399 nucleotides long, so 3400 is one nucleotide beyond its reading frame

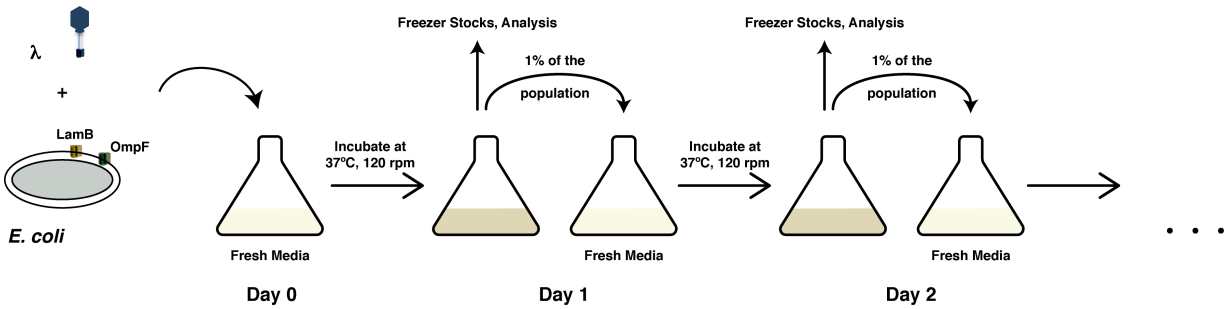

Figure 3.20: A schematic overview of the coevolution experiments performed in this study and by Meyer *et al.* [103]. Coevolution is initiated by coculturing λ and *E. coli* in glucose media for 24 hours at 37°C. At the end of each day, 1% of the total population is transferred to fresh media to continue the coevolution and the spent media is discarded. A fraction of the community is also sampled for long-term storage at -80°C, to determine population densities and to check the status of λ's OmpF use.

λ genotype	**Mutations**
7-mut (OmpF$^+$)	C2999T, **A3034G**, T3230C, C3310T, **G3319A**, **T3321A**, A3364T
λ^{-1} (OmpF$^-$)	C2999T, **A3034G**, T3230C, C3310T, **G3319A**, A3364T
λ^{-2} (OmpF$^-$)	C2999T, T3230C, C3310T, **G3319A**, A3364T

Figure 3.21: List of mutations in 7-mut OmpF$^+$ cI857 lysogen and the two engineered OmpF$^-$ genotypes; λ^{-1} and λ^{-2} (see section 3.8.5 for details on construction). Bolded mutations represent the three canonical mutations for OmpF$^+$ function.

Chapter 4

λ overcomes a perturbation in its host–viral genetic network through mutualism and evolution of life history traits

4.1 Introduction

Viruses are molecular hackers that have evolved to reprogram the goal of their host's genetic code from Darwinian survival and reproduction to viral replication and transmission. They achieve this by injecting their host cells with genetic instructions that coerce cells into replicating viral DNA and peptides instead of the host's own molecules. The result is that viruses are able to take control of massive host genetic networks using just the information from a handful of viral genes. How they evolve this level of control is an enigma and provides a unique opportunity to learn how gene networks evolve. Here we take advantage of a viral-host pair (bacteriophage λ and *Escherichia coli*) that share a well-characterized

genetic interaction network [16, 99] and readily evolve in the laboratory [103, 116, 146] to study how genetic networks evolve.

We reduced the difficulty of studying the evolution of host-viral genetic networks by focusing on one element of the network, λ's DNA replication complex. The complex consists of *E. coli* proteins DnaJ, DnaK, and DnaB, λ proteins O, P and the λ genome's origin of replication (Figure 4.1) [6, 88, 89, 155]. We chose this complex for a number of reasons. First, the genes for dnaJ and dnaK can be removed from *E. coli* with little impact on its growth, but their loss prohibits λ replication. Therefore, we could challenge λ with host cells that it cannot use but could possibly adapt to exploit by repairing this five-gene network. The second reason we chose this network was to understand how viruses evolve to expand their host-ranges when cells have incompatibilities in the cytoplasm. Most viral host-range evolution experiments focus on molecular interactions that happen on the outer membrane; however, viral infections can also be thwarted by incompatibilities that occur intracellularly [37].

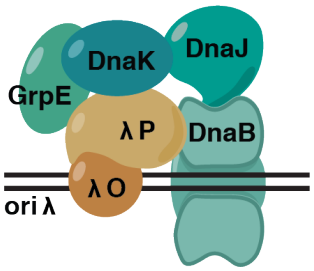

Figure 4.1: Schematic of bacteriophage λ replication complex. λ gene products are shown in orange and *E. coli* gene products are in green.

We disrupted the complex by deleting the gene for DnaJ and then challenged the virus to evolve new mechanisms to initiate DNA replication. DnaJ is an *E. coli* chaperone that the bacterium typically uses to correct protein misfolding [77]. The protein can be deleted with little effect on *E. coli* [9], but the deletion prohibits λ replication. λ has coopted DnaJ's ability to interact with DNA replication proteins [89, 155] to help assemble

the λ replication complex. λ O binds to the λ ori site, and recruits λ P. λ P binds the host helicase DnaB, simultaneously recruiting it to the origin and holding it in an inactive state. Host proteins DnaJ and DnaK then join the complex, with DnaJ binding DnaB, and DnaK binding λ P. DnaK hydrolyzes ATP, which triggers a conformational change in DnaK [90] and triggers the release of λ P. When λ P is free, DnaB unwinds the λ DNA. There is evidence that DnaJ plays a role in recruiting DnaK [89] and accelerating DnaK's rate of ATP hydrolysis [118]. For readers peripheral to the phage molecular biology field, the most important aspect to understand of how the complex works is that DnaJ does not directly replicate λ's DNA, but its role is through modifying other proteins. This opens the door to indirect molecular adaptation.

Knowing which λ proteins are members of the DNA replication complex, we predicted that λ O and λ P would acquire mutations in order to compensate for the missing DnaJ. However, these proteins did not evolve in our study. Instead, λ evolved in ways that have indirect effects on the DNA replication complex. Our failed prediction was based on an overly simple interpretation of the protein complex's mechanics. Rather than understanding the molecular process as a chain of reactions connected by causal links, the system should be viewed as one dynamically balanced between helicase initiation and inhibition, where slight cellular changes can result in some probability of spontaneously initiating DNA replication. Supporting this notion is the observation that the relative amount of DnaK in the replication complex can shift this balance one-way or another to trigger DNA replication initiation [7, 155]. A more nuanced molecular model suggests that λ could adapt in indirect ways to influence the concentration of components in its replication complex, or by lengthening the time the complex has to spontaneously initiate.

4.2 Materials and Methods

4.2.1 Evolution experiment

Twelve populations of an obligatorily lytic strain of λ (cI26) [101] were cultured for 20 growth cycles. Six populations were grown on a 90/10 mix of the *dnaJ*$^-$ *E. coli* (KEIO collection JW0014), and a related strain without a gene deletion (JW0100) [9]. We laced in 10% of this unmodified host because λ was unable to reproduce on *dnaJ*$^-$ cells alone. An additional six populations with just JW0100 were cultured as a positive control. 5×10^8 cells and $\sim 10^4$ λ particles were used to initiate each population, which were cultured in 50 ml Erlenmeyer flasks with 10 ml of modified LBM9 (20 g tryptone, 10 g yeast extract, 12.8 g sodium phosphate heptahydrate, 3 g potassium phosphate monobasic, 0.5 g sodium chloride, 1 g ammonium chloride, 1.2 g magnesium sulfate, 22.2 mg calcium chloride per L water) with 12.5 μg per ml kanamycin. Flasks were incubated for 4 hours at 37 °C with shaking at 120 rpm. To isolate phage at the end of the growth cycle, a 1 ml sample from each flask was removed, 100 μl of chloroform was added to kill the bacteria, and the remaining cell debris was removed with centrifugation. Phage samples were stored at 4 °C until the next cycle. To continue the experiment on another day, the same four-hour protocol was used except 100 μl of the previous cycle's phage was added to the flasks.

Phage densities were measured at the end of the 1st, 5th, 10th, and 20th cycles by counting plaques that formed on a lawn of a highly phage sensitive *E. coli*, strain DH5α (Invitrogen). When some λ populations stopped producing plaques, we confirmed the absence of phage DNA, and thus phage, by PCR with primers that amplify an essential λ gene, J (Supplementary Table 1 for DNA sequences). The PCR test was negative for the predicted extinct populations but positive for the others, confirming the reliability of the plaque assay.

At the end of the experiment, a single λ genotype was isolated from each of the two

remaining $dnaJ^-$ populations (labeled DNAJ1 or DNAJ2), as well as one from each of the six wildtype-only populations (labeled WT1 - WT6).

4.2.2 Competitve fitness assays

To measure the Darwinian fitness of the evolved strains relative to their ancestor, we ran head-to-head competition experiments between DNAJ1 or DNAJ2 against a genetically marked version of the ancestor. cI26 (ancestor λ) was marked by fusing the alpha subunit of the *E. coli* gene lacZ with λ's R gene [25, 126]. LacZ reacts with bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and releases a blue dye when it is metabolized. Evolved and ancestral-marked phage can be identified in mixed populations by plating and counting clear and blue plaques, respectively. The phages were plated with 0.5 mg per ml of X-gal and 0.25mg per ml Isopropyl β-D-1-thiogalactopyranoside on DH5α, which has lacZα deleted, so that the only source of blue coloration was from the marked phage and not *E. coli*. Previous studies with this marker have shown a slight fitness cost of the marker. To know how much to correct for fitness costs of the marker in our experiments, we ran control competitions of the unmarked and marked ancestor side by side with each focal competition. We observed that the marker effect was not statistically significant when competing for either host (P-values of t test for effect of the marker: 0.1556 for $dnaJ^-$ cells, and 0.0667 for wild type cells), so no correction was made.

Fitness was calculated as the difference in the Malthusian parameters *[ln(final density/initial density)]/time* of the evolved genotype and the ancestor [138]. Fitness was measured over four hours for each isolate in the conditions they were evolved in. Fitness was measured in two different host contexts, with 100% $dnaJ^-$ cells or 100% wild type ($\sim 5 \times 10^8$ cells from overnight culture). Recall, the evolution condition has a mixture of both hosts, so these competitions will inform whether λ adapted to one or the other, or both.

4.2.3 Growth rate assays

To determine whether λ had adapted to infect *E. coli* lacking dnaJ, we measured λ growth rates for the ancestor (cI26) and evolved strains (DNAJ1 and DNAJ2) with 100% $dnaJ^-$ cells, and for comparison, 100% unmodified *E. coli*. Growth rates are quantified as the change in population size over time, and thus, they are a measurement of the combined effects of birth and death. Many measurements we made are negative because $dnaJ^-$ cells cause λ mortality. λ injects its DNA into the cell, where it is not replicated nor incorporated into a new particle. Since the parent only has one copy of its genome, it is unable to attempt additional infections, causing a decline in phage population size.

Rates were measured with the same medium, four-hour growth cycle, and initial bacterial density used during the evolution (5×10^7 cells per ml). Rates were estimated by measuring the change in phage densities over the four-hour incubation period by plaque assay on *E. coli* strain DH5α. Malthusian parameters were calculated to estimate growth rates. Growth is typically density-dependent, so we measured the growth rates at eight densities spread across $5 \times 10^5 - 1 \times 10^{10}$ particles per ml (MOI of 0.01 to 200).

Note, growth rate assays in liquid culture were the only means we had to observe positive growth on $dnaJ^-$ cells. A more common assay is a plaque assay where phages are plated on lawns of the focal host cells, $dnaJ^-$ for this study. However, even the evolved λs are too poor at infecting $dnaJ^-$ cells to produce plaques, so this more sensitive liquid assay was required.

Surprisingly, we observed positive density dependent patterns for λ growing on $dnaJ^-$ cells. To confirm this pattern, we repeated the density-dependent growth experiment with higher resolution experiments. We honed our efforts by focusing on DNAJ2 and on three low densities where the λ population declined (about 1.1×10^5, 3.1×10^6, and 5.3×10^6 per ml) and three high densities where they had positive growth (8.8×10^7, 2.8×10^8, and 1.2×10^9 per ml). We ran three replicates at each density and tested for density dependence

using an ANOVA and Tukey's significance test.

We also measured growth rates for three genetically engineered λ particles and the unmodified phage (construction details follow). Each genotype's growth was measured in a separate block of the experiment. To control for block effects, the unmodified λ was included in each block. Growth rates were measured for two initial densities of λ and a single density of bacterial cells (5×10^7 cells per ml). Data were analyzed using multiple linear regression to determine the effects of the J and S mutations alone, their interaction with density, and their interaction with each other. Each term's significance was determined by a Chi-squared test. Block effects were initially included in the model, but later removed because they were not significant. Analyses were carried out by JMP software version 10.

4.2.4 Mutualism assays

The growth rate assays revealed a positive density dependent pattern of growth for the ancestor, DNAJ1, and DNAJ2 on $dnaJ^-$ cells. This growth pattern is indicative of mutualism, since higher densities yield more growth. To test whether the evolved λ had enhanced ability to cooperate, we measured whether the ancestor had greater fitness in the presence of DNAJ2 compared to other individuals of the ancestor. Adaptation under typical resource competition conditions should cause DNAJ2 to outcompete the ancestor for host cells making the ancestor's fitness decline in the presence of the evolved competitor. However, if DNAJ2 evolved enhanced ability to cooperate, the opposite pattern would be observed. We set up the experiments identically to the competitions previously described, but this time with just the $dnaJ^-$ cell treatment, and the density of λ was 2.3 times higher than the bacteria, which increases the chance that cells will be multiply infected. This improves the opportunity for intracellular interactions; either competitive or cooperative. *E. coli* cells can express upwards of 30,000 LamB monomers on their outer membrane, so λ did not compete for receptors on the outer membrane, but could compete intercellularly [109].

We ran two types of competitions; DNAJ2 versus marked ancestor and ancestor versus the marked ancestor. The key readout of the assays was the growth rates of the marked ancestor. If λ evolved enhanced cooperation, then the marked ancestor should survive better in the presence of DNAJ2 than in the presence of the unmarked ancestor.

In this assay, we manipulated the initial phage densities such that the marked ancestor would nearly always infect a cell that was coinfected with at least one other unmarked λ particle (DNJA2 or ancestor) but not with another marked ancestor. Thus, our measurements of cross-strain cooperation were not confounded by the effects of interactions between two marked ancestors or the growth of marked ancestor alone in a singly infected host cell. We achieved this by using a total phage-to-host ratio (or multiplicity of infection, MOI) of 2.3, and a ratio of $\sim 1:2.3$ for the marked ancestor to the other unmarked genotype for both competitions (see caption of Supplementary Fig. 1 for the value of densities). These specific ratios were chosen because the number of λ particles infecting any given host cell should follow a Poisson distribution with a mean equal to the MOI. Given this distribution, a MOI of 2.3 ensures that few cells will be singly infected or triply infected. Having the unmarked genotype make up 70% of the population meant that the marked genotype was much more likely to coinfect cells with the unmarked genotype, and not the same genotype.

4.2.5 Mutation discovery

We sequenced λ O and λ P genes from DNAJ1 and DNAJ2 using traditional Sanger sequencing techniques (see Figure 4.11 for primer sequences). Next, we sequenced the full genomes of DNAJ1 and DNAJ2 and a single isolate from each of the six wildtype-only evolved populations sampled after 20 cycles of evolution. The wild type isolates allowed us to determine whether the mutations evolved with $dnaJ^-$ were unique to this treatment. Genomic DNA was prepared using methods adapted from Baym *et al.* 2015 [12]

and analyses run on the software BRESEQ [40]. For specific protocols, see Meyer *et al.* 2016 [102]. IGM Genomics Center, University of California, San Diego, La Jolla, CA performed the sequencing.

4.2.6 Mutation reconstruction

The two mutations uncovered in DNAJ2 were engineered into λ to determine their effects. Three genomes were constructed, one for each mutation, and then a third with both mutations. We edited strain cI857 (supplied by I. N. Wang, University of Albany) because this strain naturally integrates into *E. coli*'s genome (lysogeny), where it remains dormant and is amenable to genome editing. We also chose to introduce the mutations into a lysogen because estimating lysis timing is less error prone with lysogens than lytic strains (see the following 'λ Lysis timing' section). cI857 and cI26 are mutants of the original λ strain and are closely related except for 30 mutations scattered throughout their genomes, none of which are in the genes that evolved in the experiment (Table S1 in Meyer *et al.* 2012 [103]). Comparisons of the effects of mutations were made between strains of cI857 with different combinations of the mutations, and never between cI857 and cI26, in order to control for strain-specific differences. Modifications were made using Multiplexed Automated Genome Engineering (MAGE) following identical protocols used in Meyer *et al.* 2016 [102]. See Supplementary Table 1 for oligo sequences used for transformation and sanger sequencing verification.

4.2.7 Adsorption assay

The J mutation found in DNAJ1 and DNAJ2 was predicted to increase the rate at which λ binds to the cell (adsorption rate). We tested this by comparing the adsorption rates of cI857 to the modified version with the J mutation. These measurements were made in triplicate identically to the methods described in Meyer *et al.* 2016 [102]. We used wild

type cells for this assay.

4.2.8 λ Lysis timing

DNAJ2 has a mutation upstream of the gene S that we predicted would delay cell lysis during infection. We also predicted that the S mutation in DNAJ1 would have similar effects, however, we only tested the effects of this single S mutation. To test this, we triggered the cell lysis program by heat shocking lysogens. High temperatures cause the λ cI repressor protein in this strain to unfold, initiating S expression and starting the count-down to lysis. This method allows for precise measurements of lysis timing since all cells in the population are triggered to initiate the clock at the exact same time. A heat-shocked test tube will transition from highly turbid to clear within minutes. In contrast, lysis by lytic strains is not as synchronized since the timing is impacted by a number of other infection processes, making lysogens less error prone for detecting timing differences. Lysogen cell densities were observed for two hours after the heat shock. If the S mutation slows lysis, then the cI857 lysogen modified with the S mutation will maintain a higher OD for longer compared to the unmodified lysogen. Heat shock was performed according to Meyer *et al.* 2016 [102]. After heating the culture, a 200 μl subsample was removed and placed in a Costar flat bottom 96-well microtiter plate. Nine samples from cI857 and nine from the S mutant were analyzed. The optical densities of the wells were measured every two minutes for ~ 2 hours using a Tecan Sunrise with an Absorbance 600 nm, Infinite F200 PRO/F500, NIP filter. The instrument was set to incubate at 37 °C, and to shake for 5 seconds just before each measurement.

4.3 Results and discussion

4.3.1 Population Dynamics

Four out of six populations cultured with 90% $dnaJ^-$ cells did not survive to the end of the experiment, demonstrating the significant challenge $dnaJ^-$ cells pose for λ infection (Figure 4.2). $dnaJ^-$ hosts are not just difficult to infect, they cause phage mortality. When a λ particle injects its DNA into any *E. coli* cell, the particle is rendered inert and cannot infect another cell. In normal hosts, the DNA typically yields a successful infection and the production of new particles. However, the majority of DNA injections into $dnaJ^-$ cells do not produce progeny, causing λ populations to decline.

For the two surviving populations, the first indication that λ gained some function on $dnaJ^-$ cells was the three orders of magnitude increase in phage density (Figure 4.2), which is roughly the proportion of $dnaJ^-$ cells multiplied by λ's burst size [145]. During typical serial transfer experiments, where 1/100th of the population is transferred to the next round, λ replication keeps pace with dilution between rounds, and densities barely change from one day to the next, as observed by the modest gains in the control populations (Figure 4.2). Expansions are typically an indication that the population has evolved access to a new resource [17].

4.3.2 Competitive and absolute fitness

Through direct head to head competitions, we found that DNAJ1 and DNAJ2 had gained fitness on $dnaJ^-$ and on wild type cells (Figure 4.3a, 4.3b). There does not appear to be a trade-off in adapting to each cell type, rather, λ's adaptations provide correlated benefits. A closer examination of the underlying growth rate calculations used to compute the competitive fitness revealed that both evolved λs had adapted to better survive $dnaJ^-$ cells rather than gain a positive growth rate (Figure 4.3a, 4.3c). In other words, their

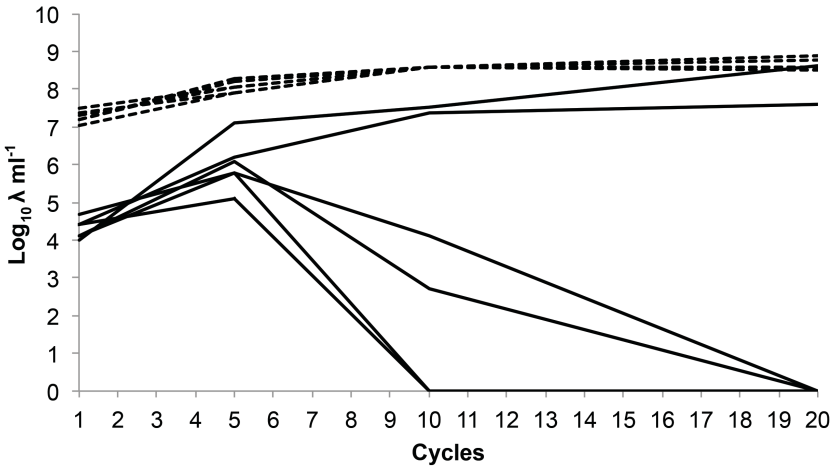

Figure 4.2: Population dynamics of λ serially propagated with unmodified *E. coli* (dashed) and a $90:10$ mixture of $dnaJ^-$ and unmodified *E. coli* (solid). Y values of '0' indicate when no phages were detected.

improved fitness relative to the ancestor came via a reduced death rate on $dnaJ^-$ cells. The fitness of both evolved λs on wild-type cells also improved, but this gain came from increased growth rates on wild type cells (Figure 4.3b, 4.3d). These observations raised an important question: If λ had only evolved to survive better on $dnaJ^-$ cells, rather than reproduce on them, then how did the two surviving populations in the $dnaJ^-$ treatment expand by 10^3-fold? The increased growth on wild type and decreased death on $dnaJ^-$ could only explain a roughly 10-fold increase.

A hint to the answer came from pilot studies where we noticed that λ's growth rate on $dnaJ^-$ cells varied between trials, sometimes positive and other times negative. We hypothesized the unintended differences in λ's starting density may have caused growth rate differences. To test this, we set up a series of growth experiments at different densities for the ancestral λ and DNAJ1 and DNAJ2 on both wild type and $dnaJ^-$ cells. In line with the competition experiments, all three genotypes displayed positive growth rates on wild type cells; however, their growth rates displayed a negative density-dependent growth pattern (Figure 4.4a). Analysis of the data revealed that DNAJ2 grows faster than DNAJ1,

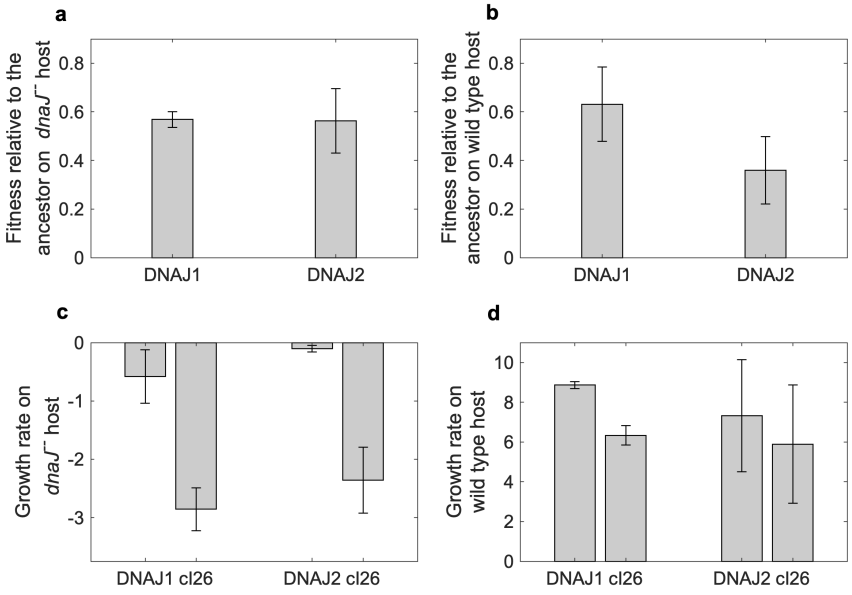

Figure 4.3: Fitness of evolved strains compared to their ancestor. Relative fitness calculated as the difference in growth rates for one cycle of the evolution experiment on **a)** $dnaJ^-$ cells and **b)** wild type cells. **c)** and **d)** show growth rates of genotypes used to calculate relative fitness on different hosts in panels a) and b), respectively. Error bars indicate one standard deviation.

which grows faster than the ancestor (general linear model where growth rate = log(density) + genotype; genotypic effect $F = 27.5629$, $P = 0.0003$ and a Tukey's Test $\alpha = 0.05$ indicates all three genotypes are significantly different). Negative density-dependent growth is a typical pattern. At low densities viruses have access to more cells and have enhanced potential to reproduce, whereas at high densities λs compete for limited numbers of host cells, which minimizes their per capita growth rate. The pattern for growth on $dnaJ^-$ cells was surprising (Figure 4.4b, Figure 4.5). DNAJ1 and DNAJ2 had slightly negative growth rates at low densities, with a sudden uptick in growth at $\sim 10^8$ particles per ml and a decline at $\sim 10^9$ particles per ml. The ancestor was much worse off, with a very high death rate for most λ densities and a reduction in mortality at $\sim 10^8$ particles per ml. This experiment showed that at certain densities, the evolved λs were able to reproduce on $dnaJ^-$ cells they previously were unable to use. This positive growth on $dnaJ^-$ cells

explains how two λ populations in $dnaJ^-$ treatment were able to reach such high densities (Figure 4.2).

Positive density-dependent growth, or the Allee-effect, can arise for many reasons. The explanations range from strict cooperation to unconscious facilitation among conspecific individuals [8, 34, 130]. In such cases, an increase in the density of mutualists increases beneficial social interactions and growth rates. While it may seem counterintuitive that viruses could have social lives, they do. Many researchers have documented social interactions between viruses [22, 85], and behaviors like cheating have even been observed to evolve during similar laboratory experiments to our own [119, 140]. Viruses are able to influence each other's growth when multiple particles infect the same cells. Intracellularly, they can interact to alter gene expression, share proteins and structural components, stifle host immune systems, or influence each other in other ways to help or hurt replication [46].

4.3.3 Mutualism

Seeking further evidence that the evolved λs were interacting mutualistically, we tested whether the evolved λ benefited the ancestral λ. The ability to transfer the benefit to other genotypes or species is a defining feature of mutualistic interactions [150]. We predicted that the ancestor λ particle will have a better chance of survival when it is paired intracellularly with the evolved λ particle rather than another ancestor λ particle. This prediction is orthogonal to the typical expectation in a purely competitive scenario where the better adapted virus should out-compete the ancestor and reduce its survival. We found that marked cI26 particles had a higher growth rate in the presence of DNAJ2 particles than in the presence of other cI26 particles (Fig. 5, $T = 4.07$, $d.f. = 6$, $P = 0.0066$, see Figure 4.7 for growth rates of the unmarked strains). This experiment provides further evidence that λ adapted to $dnaJ^-$ perturbation through mutations that enhance mutualism, since the evolved λ strain transfers its benefit to its ancestor.

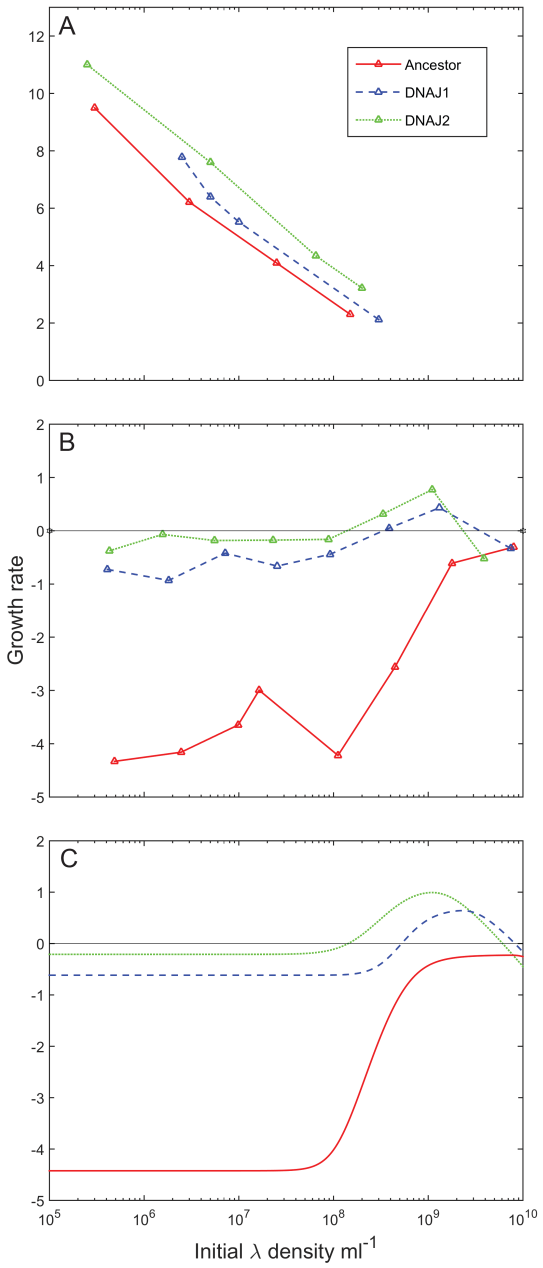

Figure 4.4: Density-dependent phage growth. Growth rate of ancestral λ (red solid), DNAJ1 (green dotted), and DNAJ2 (blue dashed) on **(A)** unmodified, and **(B)** *dnaJ*$^{-}$ *E. coli*. **(C)** Model predictions of λ's density-dependent growth rate simulated using MATLAB with the following parameters: ancestor $p = 0.985$ $t = 5$ $b = 70$ $r = 0.8$; DNAJ1 $p = 0.73t = 5b = 50r = 0.8; DNAJ2p = 0.27t = 4b = 35r = 3$. Parameter values fall within the range previously reported [145]. Growth rates are reported as Malthusian parameters computed for the entire growth cycle.

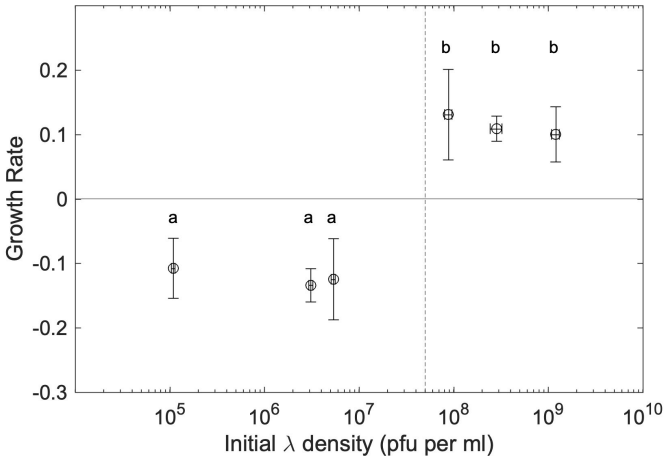

Figure 4.5: DNAJ2 positive density-dependent growth rate (per cycle). The vertical dashed line gives the density of bacterial cells in all experimental trials, this is the transition where MOI shifts from ¡1 to ¿1 and when the phages have an opportunity to cooperate intracellularly. At low densities, where the number of cells out number phages, DNAJ2 mortality overrides growth. At high densities, when λ outnumber E. coli cells, DNAJ2 population sustains growth. There is a significant effect of initial λ density on growth rate (ANOVA: $d.f. = 5$, F-ratio $= 21.56$, $P < 0.0001$), and a Tukey's significance test ($p = 0.05$) shows that λ populations initiated with MOI<1 have significantly lower growth rates than λ populations with MOI>1. Letters indicate Tukey's clusters. Error bars signify one standard deviation.

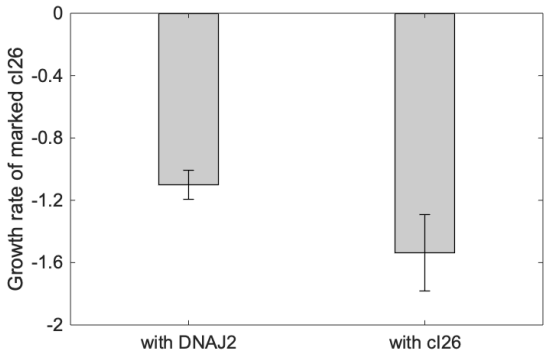

Figure 4.6: Ancestor λ (cI26) has a higher chance of survival when it coinfects a $dnaJ^-$ host cell with evolved λ (DNAJ2) as compared to another cI26 particle. Error bars indicate 95% confidence intervals.

How does this mutualism work mechanistically? We have two hypotheses. First, it has been observed that when *E. coli* is infected by multiple phages, the lysis timing

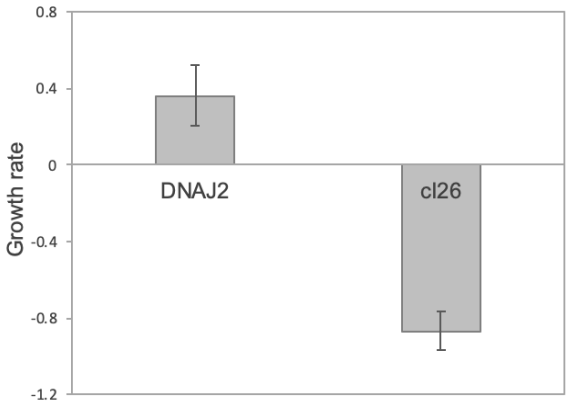

Figure 4.7: Growth rate of DNAJ2 and cI26 when grown with marked cI26, with initial densities of $\sim 7.68 \times 10^7$ phage per ml and $\sim 8.32 \times 10^7$ phage per ml respectively. Initial host density in this experiment was $\sim 5 \times 10^7$ cells per ml; initial density of marked cI26 was $\sim 3.48 \times 10^7$ phage per ml. These data are included as further confirmation that at crucial λ densities, evolved λ (DNAJ2) achieves positive growth on $dnaJ^-$ cells.

can be delayed [145]. Delayed lysis would give the λs more time to await spontaneous DnaK hydrolysis and successful DNA replication. The second is that increased λ proteins, namely O and P, could increase the chance of successful DNA replication, as increased DnaK, and DnaJ concentrations have been shown to do [6, 7, 155]. It is important to note that whatever the mechanism is, it appears to be an ancestral trait that is enhanced by the DNAJ1 and DNAJ2 adaptations. We will return to these hypotheses in the discussion of λ mutations.

4.3.4 Modeling Population growth patterns

The pattern of growth rate's density dependence on $dnaJ^-$ is unlike one that we have observed in the literature. Rather than a simple linear relationship, there is no relationship between density and growth rates below 10^8 particles per ml, and then a unimodal relationship with density for DNAJ1 and DNAJ2, or an increasing relationship

for the ancestor. We hypothesized that these abnormal patterns could be explained by two processes, intracellular mutualism and competition. To determine if this was a viable hypothesis, we constructed a mathematical model that predicted λ growth rates at different λ initial densities given parameters related to their ecological interactions. More specifically, the model predicted λ population size, given by D_{final}, after a single round of phage infection, replication, and cell burst. Phage particles (total number n) in the model can have one of three replication potentials that depend on how many other particles co-infect the same cell. Particles in the cells with a low multiplicity of infection do not cooperate and are penalized. This penalty (p) can be converted to express the probability that a cell infected by a single particle will produce offspring $(1-p)$. The value of p can range from 0 to 1, where 1 is complete death and 0 is no penalty. $1-p$ is applied to the reproductive potential (r) of each phage particle. r is an intrinsic property of each particle and is the maximum number of progenies a phage can produce under ideal mutualistic and noncompetitive conditions. Particles found in host cells that have more than a certain threshold (t) of phages cooperate and are not penalized, as long as there are not so many phages that their combined reproductive potential is greater than the cell's production limit or burst size (b). Above this threshold phages divide progeny production evenly among them with each phage particle producing b/i phages, where i is the number of coinfecting particles.

$$D_{final} = \sum_{i=1}^{n} \begin{cases} (1-p)r \times N(i) & \text{if } i < t \\ r \times N(i) & \text{if } t \leq i < \frac{b}{r} \\ \frac{b}{i} \times N(i) & \text{if } i \geq \frac{b}{r} \end{cases} \tag{4.1}$$

$N(i)$ is a function describing the number of phages found in the infection class i,

$$N(i) = h\Big[\frac{e^{-\alpha}\alpha^{i}}{i!}\Big] \times i \tag{4.2}$$

where h is the total number of host cells, and the bracketed term is the Poisson distribution of average number of phage particles available for infection per host cell, $\alpha = n/h$. $N(i)$ is then computed by calculating the probability of getting i infections per host cell, given by the Poisson distribution, and multiplying it by a total available host cells (h) and co-infecting particles (i) to give the total number of phage present in the infection class i. A simplifying assumption of this model is that all phage particles successfully infect a host cell by the end of the growth cycle.

This model is, indeed, able to produce the density-dependent relationship for λ growth on $dnaJ^{-}$ cells for DNAJ1, DNAJ2, and the ancestral strain (Figure 4.4c), confirming that the tension between mutualism and competition is one explanation for these unintuitive patterns. An important caveat about the modeling results is that the variable values were fit to the data and we did not measure them directly. The fit values are biologically reasonable, but they have not been directly measured (Figure 4.4c). The goal of this effort was to verify whether our hypothesis that the growth patterns could be produced from intracellular mutualism and competition was reasonable. Indeed, it was.

One unexpected result of the model was that the flat line pattern in the region of low multiplicity of infection could only be produced if the penalty, p, was less than 1. This means that even though most infections of $dnaJ^{-}$ cells with a single phage particle are aborted, some small fraction lead to a successful infection (the model predicts 1.5% for the ancestor and 27% for DNAJ1 and DNAJ2). Perhaps DnaK has a low background rate of hydrolysis so that in the absence of DnaJ some infected cells successfully initiate λ DNA replication. This finding is important because it suggests that the hurdle to adapt to $dnaJ^{-}$ cells may be easier to overcome than we first anticipated. λ adaptation could be achieved by tweaking cellular processes to enhance the frequency of this rare event, rather

than evolving changes to the protein network.

4.3.5 Mutations responsible for adaptation

We predicted that λ would evolve mutations in genes O and P since they encode λ proteins that directly interact with DnaJ. Sanger sequencing of these two loci in DNAJ1 and DNAJ2 revealed no changes. Full genome sequencing of DNAJ1 and DNAJ2 uncovered mutations in three other loci (J, S, and orf-314) that we did not expect to evolve (Figure 4.12). J encodes the tail fiber proteins, S affects cell lysis timing, and orf-314 is a vestigial gene no longer used by this strain of λ, so the mutation in it is likely neutral [71]. Neither J or S are thought to impact λ DNA replication, however, the mutations must be responsible for λ's adaptation since they are the only substitutions in DNAJ2, and they evolved in parallel in both populations.

Analysis of the wild type evolved λs revealed that J mutations also evolved in five out of six populations, and one of those populations also evolved a mutation in S (Supplementary Table 2). This suggests that the adaptations that facilitated replication on $dnaJ^-$ cells were not specific adaptations to the gene deletion. This is in line with the observation that DNAJ1 and DNAJ2 are more competitive on wild type cells than their ancestor (Figure 4.3 and Figure 4.4). The mutations have host-specific effects such as enhancing mutualism on $dnaJ^-$ cells, but their benefits are generic to both cell types.

Next, we tested the phenotypic effects of the J and S mutations alone and in combination by genetically engineering three genomes with one or both of the DNAJ2 mutations.

J is the host recognition gene and is responsible for λ adsorption to host cell's outer membrane [76]. A previous study showed that J mutations can enhance adsorption rates [25], so we measured adsorption rates of the strain with the single J mutation introduced and compared it to the unmodified λ. We found that the J mutation nearly

doubles λ's adsorption rate (Figure 4.3, $T = 11.413$, $d.f. = 4$, $P = 0.0040$). This specific J mutation has evolved in another λ evolution experiment with hosts that have fully functional DNA replication complexes [103], further indicating that this mutation provides general benefits to λ that are not specific to infecting $dnaJ^-$ cells.

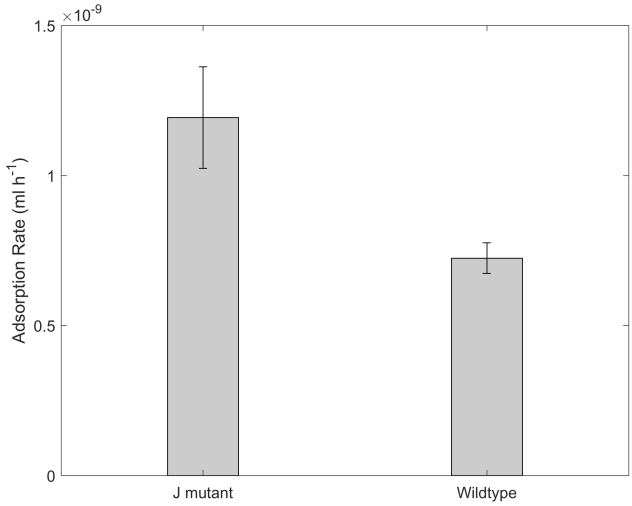

Figure 4.8: J mutation increased adsorption rate. Error bars indicate 95% confidence intervals.

S is a holin protein that determines the length of infection by triggering degradation of the inner cell membrane at the optimal time [147]. Mutations in both the regulatory and protein encoding regions of the gene have been shown to lengthen or shorten the infection cycle. While neither of the mutations uncovered in our experiment have been previously studied, mutations at nearby sites are known to delay lysis [29]. We found the upstream S mutation found in DNAJ2 significantly changed lysis dynamics (Figure 4.9). Cells with the unmodified λ began to lyse at 70 minutes and the culture completely cleared by 100 minutes (Figure 4.9). The cells with the upstream S mutation begin to lyse at around 85 minutes, but their decline is not as pronounced. In fact, the cell density remains relatively stable at optical density (OD) 0.7 until the end of the experiment. *E. coli* typically reach an OD of 1.4, indicating that the mutated S gene is still functional, and the cells are lysing.

The data suggest that the S mutation delays lysis and may even introduce stochasticity into the regulation as indicated by the lack of a synchronous clearing.

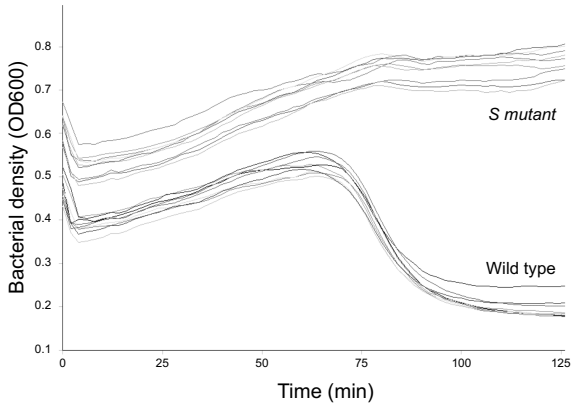

Figure 4.9: S regulatory mutation slows cell lysis timing. Nine independent trials were performed for each genotype.

How do these two mutations affect λ growth rates? More specifically, which of the mutations explain the changes in the density-dependent growth pattern? To answer this, we grew all four phages (unmodified, J-only, S regulatory-only, and J & S-regulatory) at low and high phage densities. We found that the J mutations increased growth rate at both low and high densities (Figure 4.10, $\chi^2 = 5.164,\ P = 0.0231$), providing an overall boost in the evolved λ's fitness. The upstream S mutation improves growth, but much more so at low λ densities (Figure 4.10, $\chi^2 = 11.429,\ P = 0.0007$), which explains why the largest difference in growth between the evolved strains and ancestor occur at low λ densities. Lastly, there is not an interactive effect (epistasis) between J and the upstream S mutation (Figure 4.10, $\chi^2 = 1.247,\ P = 0.2640$).

Why is the benefit of the S regulatory mutation more pronounced at low densities? The answer to this question may also help us understand the mechanism for mutualism. Recall that one of the hypotheses for how mutualism works is by delaying cell lysis. If both mutualism and the S regulatory mutation are beneficial through delaying lysis, then

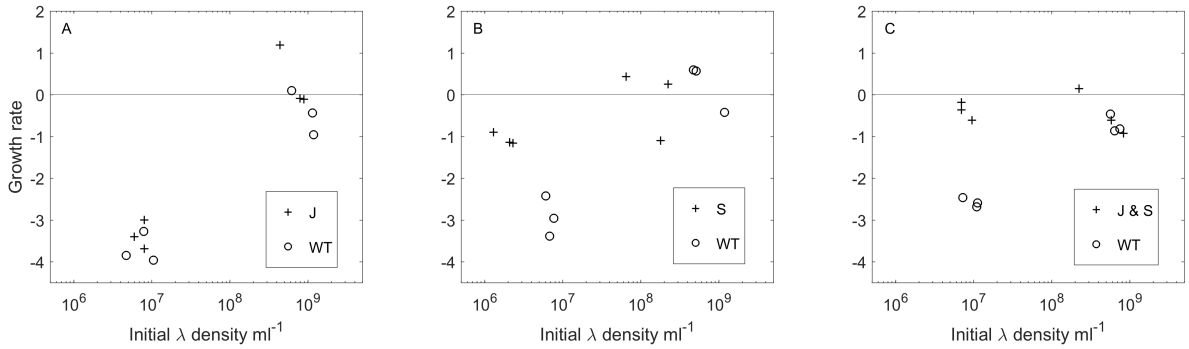

Figure 4.10: Density-dependent growth of the genetically modified λs compared to the unmodified strain. The J mutation subtly improves growth rate at low and high densities (A), while the S mutation improves growth at low densities (B), and the effects of the two mutations are additive (C). Growth rates are reported as Malthusian parameters computed across the entire growth cycle.

they may partially override each other, causing the S mutation to have less of an impact when cells are multiply infected. Taken together, these results point to a likely answer to the question of how λ copes with a dysfunctional genetic network. λ finds mechanisms—genetic or ecological—that allow it to spend more time within its host cell waiting for the dysfunctional protein complex to spontaneously work. This is potentially a short-term evolutionary patch, that λ could evolve a more reliable solution to in the future.

4.3.6 Host-range expansion and the evolution of novelty

Bacteriophage routinely evolve to infect new host genotypes and species during laboratory experiments, providing a unique opportunity to investigate the evolutionary mechanisms that promote innovation and niche expansions [42]. These studies typically focus on how phage host-recognition genes evolve to overcome barriers to infection on the outer-membrane of host cells [45, 103, 20]. Our work provides an important addition to this literature because it shows how viruses can overcome molecular barriers to infection that arise internally within cells. As shown here, processes within the host cell can also prohibit expansion and the nature of these adaptations appear to be distinct. Adaptations

to overcome outer membrane obstacles typically occur in the host-recognition proteins and promote new or augmented protein molecular interactions. However, more subtle augmentation of life history adaptations and social interactions are able to overcome obstacles within cellular protein networks.

Whether the adaptation we observed constitutes an evolutionary innovation is debatable. From an ecological perspective, this adaptation would constitute an innovation since λ evolved to use a new resource that triggered a population expansion. However, from a molecular perspective, λ did not evolve a new structure, new protein-protein interaction, or to facilitate novel chemical reaction, rather it altered existing properties to facilitate the expansion to DnaJ deficient cells. Even if this adaptation does not satisfy all requirements to qualify as evolutionary novelty, we believe the study has important implications for understanding how populations evolve new functions. The leading challenge in understanding how novelty evolves is understanding how populations explore new habitats when they are initially unfit in the new environment. In the case of λ and $dnaJ^{-}$ cells, the phage pays a very high cost to infecting the cells. This cost, however, can be largely overcome through mutualistic social interactions. This result suggests that social interactions may help mitigate the costs of exploring new resources and likely play an underappreciated role in the evolution of innovations.

4.4 Conclusion

λ was able to rapidly adapt to a terminal perturbation in its genetic network through relatively few mutations and through mutually beneficial behavior. The mutations were not in genes for proteins that directly interact with the missing network element but were in genes that impact broader features of the virus's life cycle and reproduction. Positive λ growth on $dnaJ^{-}$ cells was only achieved at high λ densities, when λs coinfected cells

and took advantage of intracellular mutualistic behavior. Taken together, we submit that gene network approaches may provide a useful conceptual model to develop evolutionary hypotheses, however, investigations of viral adaptation must incorporate a broader systems level perspective that includes an understanding of the organism's life history and ecology.

4.5 Acknowledgments

Chapter 4 is a reprint of material with minor edits, as it appears in: A. Gupta, A. N. Soto, S. J. Medina, K. L. Petrie, J. R. Meyer, "Bacteriophage lambda overcomes a perturbation in its host–viral genetic network through mutualism and evolution of life history traits," *Evolution*, vol. 74, no. 4, pp. 764-774, 2020. The author of the dissertation was the primary author of this paper.

PCR primer oligos

Gene	Forward Primer (5'-3')	Reverse Primer (5'-3')
J	CGCATCGTTCACCTCTCACT	CCTGCGGGCGGTTTGTCATTT
O	AGCTGTGGGCGTTGATAAG	CTTCGTTCTGGTCACGGTTAG
P	AACTCCGCGATAAGTGGACC	TTTAGCCGCGGCCTGATTTA
S	CTCGTTGTGGTAGTGAGATGAA	GCAGCGAAGCGTTTGATAAG

MAGE oligos

Gene	Sequence (5'-3')
J	AGTCCACACTGCTTTCACGCTGGCGCGGAAAAGCCGCGCTCGCCaCCTTTACAATGTCCCCG ACGATTTTTTCCGCCCTCAGCGTACCGT
S	GGTCATGTTTTTCTGGCATCTTCATGTCTTACCCCCAATAAGGGtATTTGCTCTATTTAATTA GGAATAAGGTCGATTACTGATAGAACA

Figure 4.11: Nucleotide sequences used for PCR and MAGE. Lower case nucleotides in the MAGE section indicate the evolved version of the base to be edited into the genome.

Phage	Genome location	Mutation	Gene	Product
DNAJ1	18,503	A1000V (GCG→GTG)	*J*	tail: host specifici
	21,594	G189V (GGT→GTT)	*orf-314*	Tail fiber
	45,208	D8G (GAC→GGC)	*S*	antiholin
DNAJ2	18,503	A1000V (GCG→GTG)	*J*	tail: host specifici
	45,166	intergenic [+351/-20] (C→A)	*orf*-64 → / → *S*	hypothetic protein/antih
WT1	116	intergenic [–/-74] (A→G)	– / → *nu1*	–/DNA packagin
	9,067	R38R (CGT→CGC)	*V* →	tail compon
	18,503	A1000V (GCG→GTG)	*J* →	tail: host specifici
	18,884	L1127P (CTG→CCG)	*J* →	tail: host specifici
	20,200	E184G (GAA→GGA)	*orf-401* →	Tail fiber pro
WT2	18,503	A1000V (GCG→GTG)	*J* →	tail: host specifici
	18,823	D1107N (GAT→AAT)	*J* →	tail: host specifici
WT3	No mutations were detected			
WT4	116	intergenic (–/-74)	– / → *nu1*	–/DNA packagin
	10,636	G32G (GGT→GGC)	*H* →	tail compon
	18,503	A1000V (GCG→GTG)	*J* →	tail: host specifici
	18,884	L1127P (CTG→CCG)	*J* →	tail: host specifici
	22,852	L12F (TTA→TTC)	*int* ←	integration pr
	39,182	intergenic (+363/-8)	*orf-64* → / → S	hypothetic protein/anti-h
WT5	18,503	A1000V (GCG→GTG)	*J* →	tail: host specifici
	18,884	L1127P (CTG→CCG)	*J* →	tail: host specifici
WT6	18,503	A1000V (GCG→GTG)	*J* →	tail: host specifici
	18,884	L1127P (CTG→CCG)	*J* →	tail: host specifici
	39,198	M3I (ATG→ATA)	*S* →	anti-holin

Figure 4.12: Substitutions revealed by whole-genome sequencing.

Bibliography

[1] Fastqc: a quality control tool for high throughput sequence data.

[2] R: A language and environment for statistical computing.

[3] M. H. Adams. *Bacteriophages*. Interscience Publishers, Inc., New York, 1959.

[4] A. Agrawal and C. Lively. Infection genetics: Gene-for-gene versus matching-alleles models and all points in between. *Evolutionary Ecology Research*, 4, 2002.

[5] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[6] C. Alfano and R. McMacken. Heat shock protein-mediated disassembly of nucleoprotein structures is required for the initiation of bacteriophage lambda dna replication. *J Biol Chem*, 264(18):10709–18, 1989.

[7] C. Alfano and R. McMacken. Ordered assembly of nucleoprotein structures at the bacteriophage lambda replication origin during the initiation of dna replication. *J Biol Chem*, 264(18):10699–708, 1989.

[8] W. C. Allee. *Principles of animal ecology*. Saunders Co., Philadelphia, 1949.

[9] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol*, 2:2006.0008, 2006.

[10] A. C. Baker. Flexibility and specificity in coral-algal symbiosis: Diversity, ecology, and biogeography of symbiodinium. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):661–689, 2003.

[11] J. Bascompte, P. Jordano, C. J. Meli · n, and J. M. Olesen. The nested assembly of plant-animal mutualistic networks. *Proc Natl Acad Sci U S A*, 100(16):9383–7, 2003.

[12] M. Baym, S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, and R. Kishony. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*, 10(5):e0128036, 2015.

[13] S. J. Beckett and H. T. Williams. Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. *Interface Focus*, 3(6):20130033, 2013.

[14] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[15] A. Best, B. Ashby, A. White, R. Bowers, A. Buckling, B. Koskella, and M. Boots. Host–parasite fluctuating selection in the absence of specificity. *Proceedings of the Royal Society B: Biological Sciences*, 284(1866):20171615, 2017.

[16] S. Blasche, S. Wuchty, S. V. Rajagopala, and P. Uetz. The protein interaction network of bacteriophage lambda with its host, escherichia coli. *J Virol*, 87(23):12745–55, 2013.

[17] Z. D. Blount, C. Z. Borland, and R. E. Lenski. Historical contingency and the evolution of a key innovation in an experimental population of escherichia coli. *Proc Natl Acad Sci U S A*, 105(23):7899–906, 2008.

[18] Z. D. Blount, R. E. Lenski, and J. B. Losos. Contingency and determinism in evolution: Replaying life's tape. *Science*, 362(6415), 2018.

[19] B. Bohannan and R. Lenski. Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters*, 3(4):362–377, 2000.

[20] L. M. Bono, C. L. Gensel, D. W. Pfennig, and C. L. Burch. Competition and the origins of novelty: experimental evolution of niche-width expansion in a virus. *Biol Lett*, 9(1):20120616, 2013.

[21] W. Boos and A. Bohm. Learning new tricks from an old dog: Malt of the escherichia coli maltose system is part of a complex regulatory network. *Trends Genet*, 16(9):404–9, 2000.

[22] A. L. Borges, J. Y. Zhang, M. F. Rollins, B. A. Osuna, B. Wiedenheft, and J. Bondy-Denomy. Bacteriophage cooperation suppresses crispr-cas3 and cas9 immunity. *Cell*, 174(4):917–925.e10, 2018.

[23] J. R. Brum, B. L. Hurwitz, O. Schofield, H. W. Ducklow, and M. B. Sullivan. Seasonal time bombs: dominant temperate viruses affect southern ocean microbial dynamics. *The ISME Journal*, 10(2):437–449, 2016.

[24] A. Buckling and P. B. Rainey. The role of parasites in sympatric and allopatric host diversification. *Nature*, 420(6915):496–9, 2002.

[25] A. R. Burmeister, R. E. Lenski, and J. R. Meyer. Host coevolution alters the adaptive landscape of a virus. *Proc Biol Sci*, 283(1839), 2016.

[26] C. Bérénos, P. Schmid-Hempel, and K. M. Wegner. Antagonistic coevolution accelerates the evolution of reproductive isolation in *Tribolium castaneum*. *The American Naturalist*, 180(4):520–528, 2012.

[27] S. R. Casjens and R. W. Hendrix. Bacteriophage lambda: Early pioneer and still relevant. *Virology*, 479-480:310–330, 2015.

[28] H. Cervera, J. Lalić, and S. F. Elena. Effect of host species on topography of the fitness landscape for a plant rna virus. *J Virol*, 90(22):10160–10169, 2016.

[29] C. Y. Chang, K. Nam, and R. Young. S gene expression and the timing of lysis by bacteriophage lambda. *J Bacteriol*, 177(11):3283–94, 1995.

[30] W. N. Chaudhry, M. Pleška, N. N. Shah, H. Weiss, I. C. McCall, J. R. Meyer, A. Gupta, C. C. Guet, and B. R. Levin. Leaky resistance and the conditions for the existence of lytic bacteriophage. *PLOS Biology*, 16(8):e2005971, 2018.

[31] H. H. Chou, H. C. Chiu, N. F. Delaney, D. Segre, and C. J. Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034):1190–2, 2011.

[32] M. R. J. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, 2011.

[33] V. S. Cooper, D. Schneider, M. Blot, and R. E. Lenski. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of escherichia coli b. *Journal of bacteriology*, 183(9):2834–2841, 2001.

[34] F. Courchamp, T. Clutton-Brock, and B. Grenfell. Inverse density dependence and the allee effect. *Trends Ecol Evol*, 14(10):405–410, 1999.

[35] C. Darwin. *The Origin of Species*. John Murray, London, 1859.

[36] K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in ¡em¿escherichia coli¡/em¿ k-12 using pcr products. *Proceedings of the National Academy of Sciences*, 97(12):6640, 2000.

[37] P. A. de Jonge, F. L. Nobrega, S. J. J. Brouns, and B. E. Dutilh. Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol*, 27(1):51–63, 2019.

[38] J. A. de Visser and J. Krug. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet*, 15(7):480–90, 2014.

[39] J. A. G. de Visser, S. F. Elena, I. Fragata, and S. Matuszewski. The utility of fitness landscapes and big data for predicting evolution. *Heredity (Edinb)*, 121(5):401–405, 2018.

[40] D. E. Deatherage and J. E. Barrick. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol*, 1151:165–88, 2014.

[41] E. Denamur, O. Clermont, S. Bonacorsi, and D. Gordon. The population genetics of pathogenic escherichia coli. *Nature Reviews Microbiology*, 2020.

[42] J. J. Dennehy. Bacteriophages as model organisms for virus emergence research. *Trends Microbiol*, 17(10):450–7, 2009.

[43] M. Doebeli. *Adaptive diversification*. Monographs in population biology. Princeton University Press, Princeton, N.J., 2011.

[44] J. W. Drake. A constant rate of spontaneous mutation in dna-based microbes. *Proc Natl Acad Sci U S A*, 88(16):7160–4, 1991.

[45] S. Duffy, C. L. Burch, and P. E. Turner. Evolution of host specificity drives reproductive isolation among rna viruses. *Evolution*, 61(11):2614–22, 2007.

[46] S. L. DÌaz-MuÒoz, R. Sanju · n, and S. West. Sociovirology: Conflict, cooperation, and communication among viruses. *Cell Host & Microbe*, 22(4):437–441, 2017.

[47] M. Décima, M. Stukel, J. Brum, and J. Morris. *Mortality in the oceans: Causes and consequences*, pages 16–48. 2014.

[48] H. M. Ellis, D. Yu, T. DiTizio, and D. L. Court. High efficiency mutagenesis, repair, and engineering of chromosomal dna using single-stranded oligonucleotides. *Proceedings of the National Academy of Sciences*, 98(12):6742, 2001.

[49] A. M. Ellison, E. J. Farnsworth, and R. R. Twilley. Facultative mutualism between red mangroves and root-fouling sponges in belizean mangal. *Ecology*, 77(8):2431–2444, 1996.

[50] B. Erni, B. Zanolari, and H. P. Kocher. The mannose permease of escherichia coli consists of three different proteins. amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda dna. *J Biol Chem*, 262(11):5238–47, 1987.

[51] M. Esquinas-Rychen and B. Erni. Facilitation of bacteriophage lambda dna injection by inner membrane proteins of the bacterial phosphoenol-pyruvate: carbohydrate phosphotransferase system (pts). *J Mol Microbiol Biotechnol*, 3(3):361–70, 2001.

[52] C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, and J. S. Weitz. Statistical structure of host-phage interactions. *Proc Natl Acad Sci U S A*, 108(28):E288–97, 2011.

[53] C. O. Flores, T. Poisot, S. Valverde, and J. S. Weitz. Bimat: a matlab package to facilitate the analysis of bipartite networks. *Methods in Ecology and Evolution*, 7(1):127–132, 2016.

[54] C. O. Flores, S. Valverde, and J. S. Weitz. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME Journal*, 7(3):520–532, 2013.

[55] K. M. Flynn, T. F. Cooper, F. B. Moore, and V. S. Cooper. The environment affects epistatic interactions to alter the topology of an empirical fitness landscape. *PLoS Genet*, 9(4):e1003426, 2013.

[56] S. E. Forde, J. N. Thompson, R. D. Holt, and B. J. M. Bohannan. Coevolution drives temporal changes in fitness and diversity across environments in a bacteria–bacteriophage interaction. *Evolution*, 62(8):1830–1839, 2008.

[57] D. M. Fowler and S. Fields. Deep mutational scanning: a new style of protein science. *Nat Methods*, 11(8):801–7, 2014.

[58] I. Fragata, A. Blanckaert, M. A. Dias Louro, D. A. Liberles, and C. Bank. Evolution in the light of fitness landscape theory. *Trends Ecol Evol*, 34(1):69–82, 2019.

[59] V.-P. Friman and A. Buckling. Effects of predation on real-time host–parasite coevolutionary dynamics. *Ecology Letters*, 16(1):39–46, 2013.

[60] J. A. Fuhrman. Marine viruses and their biogeochemical and ecological effects. *Nature*, 399(6736):541–8, 1999.

[61] S. Gaba and D. Ebert. Time-shift experiments as a tool to study antagonistic coevolution. *Trends Ecol Evol*, 24(4):226–32, 2009.

[62] S. Gandon, A. Buckling, E. Decaestecker, and T. Day. Host-parasite coevolution and patterns of adaptation across time and space. *J Evol Biol*, 21(6):1861–6, 2008.

[63] S. Gavrilets. *High-dimensional fitness landscapes and speciation*, book section 3, pages 45–79. MIT Press, Cambridge, Mass., 2010.

[64] S. J. Gould. *Wonderful life : the Burgess Shale and the nature of history*. W.W. Norton, New York, 1st edition, 1989.

[65] J. Guimaraes, P. R., C. Sazima, S. F. dos Reis, and I. Sazima. The nested structure of marine cleaning symbiosis: is it like flowers and bees? *Biol Lett*, 3(1):51–4, 2007.

[66] P. Gómez and A. Buckling. Bacteria-phage antagonistic coevolution in soil. *Science*, 332(6025):106–9, 2011.

[67] A. R. Hall, P. D. Scanlan, A. D. Morgan, and A. Buckling. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol Lett*, 14(7):635–42, 2011.

[68] W. D. Hamilton, R. Axelrod, and R. Tanese. Sexual reproduction as an adaptation to resist parasites (a review). *Proceedings of the National Academy of Sciences*, 87(9):3566, 1990.

[69] H. G. Hampton, B. N. J. Watson, and P. C. Fineran. The arms race between bacteria and their phage foes. *Nature*, 577(7790):327–336, 2020.

[70] R. W. Hendrix. *Lambda II*. Cold Spring Harbor monograph series. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1983.

[71] R. W. Hendrix and R. L. Duda. Bacteriophage lambda papa: not the mother of all lambda phages. *Science*, 258(5085):1145, 1992.

[72] K. L. Hillesland and D. A. Stahl. Rapid evolution of stability and productivity at the origin of a microbial mutualism. *Proceedings of the National Academy of Sciences*, 107(5):2124, 2010.

[73] A. James, J. W. Pitchford, and M. J. Plank. Disentangling nestedness from models of ecological complexity. *Nature*, 487(7406):227–230, 2012.

[74] A. Kashiwagi and T. Yomo. Ongoing phenotypic and genomic changes in experimental coevolution of rna bacteriophage qβ and escherichia coli. *PLOS Genetics*, 7(8):e1002188, 2011.

[75] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[76] I. Katsura. Structure and function of the major tail protein of bacteriophage lambda. mutants having small major tail protein molecules in their virion. *J Mol Biol*, 146(4):493–512, 1981.

[77] S. Kedzierska, M. Staniszewska, A. Wegrzyn, and A. Taylor. The role of dnak/dnaj and groel/groes systems in the removal of endogenous proteins aggregated by heat-shock from escherichia coli cells. *FEBS Lett*, 446(2-3):331–7, 1999.

[78] N. T. Keen. A century of plant pathology: A retrospective view on understanding host-parasite interactions. *Annual Review of Phytopathology*, 38(1):31–48, 2000.

[79] E. D. Kelsic, H. Chung, N. Cohen, J. Park, H. H. Wang, and R. Kishony. Rna structural determinants of optimal codons revealed by mage-seq. *Cell Syst*, 3(6):563–571 e6, 2016.

[80] A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–6, 2011.

[81] B. Koskella and M. A. Brockhurst. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS microbiology reviews*, 38(5):916–931, 2014.

[82] S. Kosuri and G. M. Church. Large-scale de novo dna synthesis: technologies and applications. *Nature Methods*, 11(5):499–507, 2014.

[83] S. Kryazhimskiy, D. P. Rice, E. R. Jerison, and M. M. Desai. Microbial evolution. global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522, 2014.

[84] S. J. Labrie, J. E. Samson, and S. Moineau. Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, 8(5):317–327, 2010.

[85] M. Landsberger, S. Gandon, S. Meaden, C. Rollie, A. Chevallereau, H. Chabas, A. Buckling, E. R. Westra, and S. van Houte. Anti-crispr phages cooperate to overcome crispr-cas immunity. *Cell*, 174(4):908–916.e12, 2018.

[86] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[87] J. M. Lee, J. Huddleston, M. B. Doud, K. A. Hooper, N. C. Wu, T. Bedford, and J. D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proc Natl Acad Sci U S A*, 115(35):E8276–E8285, 2018.

[88] K. Liberek, C. Georgopoulos, and M. Zylicz. Role of the escherichia coli dnak and dnaj heat shock proteins in the initiation of bacteriophage lambda dna replication. *Proc Natl Acad Sci U S A*, 85(18):6632–6, 1988.

[89] K. Liberek, J. Osipiuk, M. Zylicz, D. Ang, J. Skorko, and C. Georgopoulos. Physical interactions between bacteriophage and escherichia coli proteins required for initiation of lambda dna replication. *J Biol Chem*, 265(6):3022–9, 1990.

[90] K. Liberek, D. Skowyra, M. Zylicz, C. Johnson, and C. Georgopoulos. The escherichia coli dnak chaperone, the 70-kda heat shock protein eukaryotic equivalent, changes conformation upon atp hydrolysis, thus triggering its dissociation from a bound target protein. *J Biol Chem*, 266(22):14491–6, 1991.

[91] H. A. Lindsey, J. Gallie, S. Taylor, and B. Kerr. Evolutionary rescue from extinction is contingent on a lower rate of environmental change. *Nature*, 494(7438):463–7, 2013.

[92] R. C. MacLean, G. G. Perron, and A. Gardner. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in pseudomonas aeruginosa. *Genetics*, 186(4):1345, 2010.

[93] R. Maddamsetti, D. T. Johnson, S. J. Spielman, K. L. Petrie, D. S. Marks, and J. R. Meyer. Gain-of-function experiments with bacteriophage lambda uncover residues under diversifying selection in nature. *Evolution*, 72(10):2234–2243, 2018.

[94] G. D. Maddox and N. Cappuccino. Genetic determination of a plant susceptibility to an herbivorous insect depends on environmental context. *Evolution*, 40(4):863–866, 1986.

[95] V. Makarenkov. T-rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–8, 2001.

[96] M. F. Marston, F. J. Pierciey, A. Shepard, G. Gearin, J. Qi, C. Yandava, S. C. Schuster, M. R. Henn, and J. B. H. Martiny. Rapid diversification of coevolving marine ¡em¿synechococcus¡/em¿ and a virus. *Proceedings of the National Academy of Sciences*, 109(12):4544, 2012.

[97] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, 17(1):3, 2011.

[98] E. A. Mayer, R. Knight, S. K. Mazmanian, J. F. Cryan, and K. Tillisch. Gut microbes and the brain: Paradigm shift in neuroscience. *The Journal of Neuroscience*, 34(46):15490, 2014.

[99] N. D. Maynard, E. W. Birch, J. C. Sanghvi, L. Chen, M. V. Gutschow, and M. W. Covert. A forward-genetic screen and dynamic analysis of lambda phage host-dependencies reveals an extensive interaction network and a new anti-viral strategy. *PLoS Genet*, 6(7):e1001017, 2010.

[100] D. J. Merrell. *The adaptive seascape : the mechanism of evolution*. University of Minnesota Press, Minneapolis, 1994.

[101] J. R. Meyer, A. A. Agrawal, R. T. Quick, D. T. Dobias, D. Schneider, and R. E. Lenski. Parallel changes in host resistance to viral infection during 45,000 generations of relaxed selection. *Evolution*, 64(10):3024–34, 2010.

[102] J. R. Meyer, D. T. Dobias, S. J. Medina, L. Servilio, A. Gupta, and R. E. Lenski. Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science*, 354(6317):1301–1304, 2016.

[103] J. R. Meyer, D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, and R. E. Lenski. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, 335(6067):428–32, 2012.

[104] J. R. Meyer, I. Gudelj, and R. Beardmore. Biophysical mechanisms that maintain biodiversity through trade-offs. *Nature Communications*, 6(1):6278, 2015.

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

[106] J. R. Meyer and R. E. Lenski. *Subtle Environmental Differences have Cascading Effects on the Ecology and Evolution of a Model Microbial Community**, pages 273–288. Springer International Publishing, Cham, 2020.

[107] S. E. Mitchell, E. S. Rogers, T. J. Little, and A. F. Read. Host-parasite and genotype-by-environment interactions: Temperature modifies potential for selection by a sterilizing pathogen. *Evolution*, 59(1):70–80, 2005.

[108] J. R. Nahum, J. West, B. M. Althouse, L. Zaman, C. Ofria, and B. Kerr. Improved adaptation in exogenously and endogenously changing environments. *The 2019 Conference on Artificial Life*, (29):306–313, 2017.

[109] F. Neidhardt, J. Ingraham, A. S. f. Microbiology, and R. Curtiss. *Escherichia Coli and Salmonella Typhimurium: Cellular and Molecular Biology*. American Society for Microbiology, 1987.

[110] C. B. Ogbunugafor, C. S. Wylie, I. Diakite, D. M. Weinreich, and D. L. Hartl. Adaptive landscape by environment interactions dictate evolutionary dynamics in models of drug resistance. *PLoS Comput Biol*, 12(1):e1004710, 2016.

[111] H. A. Orr. The genetic theory of adaptation: a brief history. *Nat Rev Genet*, 6(2):119–27, 2005.

[112] C. Pal, M. D. Maciá, A. Oliver, I. Schachar, and A. Buckling. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature*, 450(7172):1079–1081, 2007.

[113] A. C. Palmer, E. Toprak, M. Baym, S. Kim, A. Veres, S. Bershtein, and R. Kishony. Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nature Communications*, 6(1):7385, 2015.

[114] S. Paterson, T. Vogwill, A. Buckling, R. Benmayor, A. J. Spiers, N. R. Thomson, M. Quail, F. Smith, D. Walker, B. Libberton, A. Fenton, N. Hall, and M. A. Brockhurst. Antagonistic coevolution accelerates molecular evolution. *Nature*, 464(7286):275–8, 2010.

[115] K. L. Petrie, N. D. Palmer, D. T. Johnson, S. J. Medina, S. J. Yan, V. Li, A. R. Burmeister, and J. R. Meyer. Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science*, 359(6383):1542–1545, 2018.

[116] D. Refardt and P. B. Rainey. Tuning a genetic switch: experimental evolution and natural variation of prophage induction. *Evolution*, 64(4):1086–97, 2010.

[117] F. Rodriguez-Valera, A.-B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pašić, T. F. Thingstad, F. Rohwer, and A. Mira. Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7(11):828–836, 2009.

[118] R. Russell, A. Wali Karzai, A. F. Mehl, and R. McMacken. Dnaj dramatically stimulates atp hydrolysis by dnak: insight into targeting of hsp70 proteins to polypeptide substrates. *Biochemistry*, 38(13):4165–76, 1999.

[119] J. L. Sachs, U. G. Mueller, T. P. Wilcox, and J. J. Bull. The evolution of cooperation. *Q Rev Biol*, 79(2):135–60, 2004.

[120] P. Sagulenko, V. Puller, and R. A. Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042–vex042, 2018.

[121] J. Sambrook and D. W. Russell. *Molecular cloning : a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd edition, 2001.

[122] A. Sasaki. Host-parasite coevolution in a multilocus gene-for-gene system. *Proc Biol Sci*, 267(1458):2183–8, 2000.

[123] D. Scandella and W. Arber. Phage λ dna injection into *Escherichia coli* pel- mutants is restored by mutations in phage genes v or h. *Virology*, 69(1):206–215, 1976.

[124] R. D. Schulte, C. Makus, B. Hasert, N. K. Michiels, and H. Schulenburg. Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. *Proceedings of the National Academy of Sciences*, 107(16):7359, 2010.

[125] Y. Shao and I. N. Wang. Bacteriophage adsorption rate and optimal lysis time. *Genetics*, 180(1):471–82, 2008.

[126] Y. Shao and I. N. Wang. Bacteriophage adsorption rate and optimal lysis time. *Genetics*, 180(1):471–82, 2008.

[127] H. A. Shuman and T. J. Silhavy. The art and design of genetic screens: Escherichia coli. *Nat Rev Genet*, 4(6):419–31, 2003.

[128] A. Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–3, 2014.

[129] B. Steinberg and M. Ostermeier. Environmental changes bridge evolutionary valleys. *Sci Adv*, 2(1):e1500921, 2016.

[130] P. A. Stephens, W. J. Sutherland, and R. P. Freckleton. What is the allee effect? *Oikos*, 87(1):185–190, 1999.

[131] A. Stern and R. Sorek. The phage-host arms race: Shaping the evolution of microbes. *BioEssays*, 33(1):43–51, 2011.

[132] C. A. Suttle. Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol*, 5(10):801–12, 2007.

[133] R. Thomas, L. Berdjeb, T. Sime-Ngando, and S. Jacquet. Viral abundance, production, decay rates and life strategies (lysogeny versus lysis) in lake bourget (france). *Environ Microbiol*, 13(3):616–30, 2011.

[134] J. N. Thompson. *The geographic mosaic of coevolution*. Interspecific interactions. University of Chicago Press, Chicago, 2005.

[135] J. N. Thompson. Evolution. the role of coevolution. *Science*, 335(6067):410–1, 2012.

[136] J. N. Thompson and B. M. Cunningham. Geographic structure and dynamics of coevolutionary selection. *Nature*, 417(6890):735–8, 2002.

[137] V. Torsvik, L. Øvreås, and T. F. Thingstad. Prokaryotic diversity–magnitude, dynamics, and controlling factors. *Science*, 296(5570):1064–6, 2002.

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

[139] O. Tully and D. T. Nolan. A review of the population biology and host–parasite interactions of the sea louse lepeophtheirus salmonis (copepoda: Caligidae). *Parasitology*, 124(7):165–182, 2002.

[140] P. E. Turner and L. Chao. Prisoner's dilemma in an rna virus. *Nature*, 398(6726):441–3, 1999.

[141] C. Tétard-Jones, M. Kertesz, P. Gallois, and R. Preziosi. Genotype-by-genotype interactions modified by a third species in a plant-insect system. *The American Naturalist*, 170(3):492–499, 2007.

[142] S. Valverde, S. F. Elena, and R. Solé. Spatially induced nestedness in a neutral model of phage-bacteria networks. *Virus Evol*, 3(2):vex021, 2017.

[143] L. Van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30, 1973.

[144] H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest, and G. M. Church. Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, 460(7257):894–898, 2009.

[145] I. N. Wang. Lysis timing and bacteriophage fitness. *Genetics*, 172(1):17–26, 2006.

[146] I.-N. Wang, D. E. Dykhuizen, and L. B. Slobodkin. The evolution of phage lysis timing. *Evolutionary Ecology*, 10(5):545–558, 1996.

[147] I. N. Wang, D. L. Smith, and R. Young. Holins: the protein clocks of bacteriophage infections. *Annu Rev Microbiol*, 54:799–825, 2000.

[148] D. M. Weinreich, N. F. Delaney, M. A. Depristo, and D. L. Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–4, 2006.

[149] J. S. Weitz, T. Poisot, J. R. Meyer, C. O. Flores, S. Valverde, M. B. Sullivan, and M. E. Hochberg. Phage-bacteria infection networks. *Trends Microbiol*, 21(2):82–91, 2013.

[150] S. A. West, A. S. Griffin, and A. Gardner. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J Evol Biol*, 20(2):415–32, 2007.

[151] H. A. Wichman, J. Millstein, and J. J. Bull. Adaptive molecular evolution for 13,000 phage generations. *Genetics*, 170(1):19, 2005.

[152] S. Wielgoss, J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R. E. Lenski, and D. Schneider. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with ¡em¿escherichia coli¡/em¿. *G3: Genes—Genomes—Genetics*, 1(3):183, 2011.

[153] S. Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. na, 1932.

[154] L. Zaman, J. R. Meyer, S. Devangam, D. M. Bryson, R. E. Lenski, and C. Ofria. Coevolution drives the emergence of complex traits and promotes evolvability. *PLOS Biology*, 12(12):e1002023, 2014.

[155] M. Zylicz, D. Ang, K. Liberek, and C. Georgopoulos. Initiation of lambda dna replication with purified host- and bacteriophage-encoded proteins: the role of the dnak, dnaj and grpe heat shock proteins. *EMBO J*, 8(5):1601–8, 1989.