Analysis of Protein Adaptation from High Throughput Mutagenesis Studies

Bryan L. Andrews

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Stanley Fields, Chair

Benjamin Kerr

Maitreya Dunham

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Analysis of protein adaptation from high throughput mutagenesis studies

Bryan L. Andrews

Chair of the Supervisory Committee:
Stan Fields,
Professor and Chair, Department of Genome Sciences
University of Washington

Proteins are sophisticated molecular machines, yet they arise through a simple process of mutation and selection. Understanding how proteins adapt to their ever-changing environments is one of the central challenges in molecular biology. In this dissertation, I first discuss the state of the molecular evolution field, which has recently been pushed forward by advances in deep sequencing, and I highlight some broad consensuses that have risen out of recent deep mutational scans. In chapters 2-4, I investigate three systems that highlight different aspects of protein adaptation. First, I characterize the mutational neighborhood of a bacterial protein that plays a dual role as a nutrient transporter and the receptor for a phage. By performing a deep mutational scan with respect to both properties, I show that specific resistance mutations are common relative to destabilizing mutations, and that these specific resistance mutations are spatially clustered around a particular structural feature, Loop L6. Second, I characterize the tail fiber tip of phage, which mediates attachment to its host and is the key player in overcoming host resistance. I challenged a library of phage variants to infect four hosts, expressing the wild type receptor or one of three resistance mutations. By comparing infectivity of variants across these hosts, I characterize two properties that underlie adaptation: specificity and promiscuity. Third, I aggregated data from 20 published deep mutational scans, showing that many proteins have a strong signature of genetic robustness. Across these datasets, most genes have more favorable mutational neighborhoods than would be expected

by chance. In most cases, this effect cannot be explained by codon bias alone – the positions in which codons are used is much more important than how frequently they are used. In the fifth and final chapter, I draw some connections between the different systems I have investigated, and I provide a perspective on the standing questions in the field. In particular, the role of stability in biasing evolutionary outcomes rises to prominence. As technology advances, further dissection of the links between genetic alterations and protein properties is likely to enhance our understanding of how proteins adapt.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Science is never done in a vacuum. Every good scientist needs at least three types of people around them: people who make your science better, people who make you a better scientist, and people who make you sane and happy enough to keep doing it. Those who have made my science better include all of the Fields lab, but especially Stephanie Zimmerman, Mike Dorrity, Ben Brandsen. and Russ Lo. Special thanks also to the Kerr lab, and in particular Olivia Kosterlitz, Ryan McGee, Katrina van Raay, and Hannah Jordt, for putting evolution in the center of my mind. In terms of those who have made me a better scientist, I need to thank my committee, each of whom is a rock star in their own right. Christine Queitsch, Maitreya Dunham, Doug Fowler and Harmit Malik have each pushed me in exactly the right direction at every turn. If I were going to build a committee today, I would pick the exact same people. Those that have kept me sane and happy makes a somewhat longer list, so I will not list them all, but only mention a few of the especially important. My family, especially my mom and Vanessa, have been a rock for me when I've sometimes felt adrift. My MCB friends have been adrift there with me, and I'm forever grateful to you: Abe, Abby, Amy, Andrea, Becca, Colby, Leah, Liesja. Taylor and John Wang, for nerding out with me, thank you. Last, my non-science friends: Angel, Joseph, Corinne, Kaitlyn, thanks for putting up with everything.

Finally, there are a few people fit in all three categories. Stan, despite his famously light touch, has really shaped who I am as a writer, as a colleague to the rest of the lab, and as an independent problem solver. Throughout all of it, he's encouraged me to maintain a healthy relationship with my work. We haven't always agreed on project directions but thank you for letting me make it my own. Ben Kerr is a paragon of compassion, sincerity, and kindness, *i.e.*, the kind of scientist we should all aspire to be. He's pushed me to think about questions I never would have considered, both at the bench and away from it. Brooke Angell shines her light on everything.

Chapter 1: Introduction

1.1 Proteins as molecular machines

Proteins are sophisticated molecular machines that perform nearly all of the molecular-level tasks necessary to sustain life. Every protein is encoded as a sequence of amino acids, which forms a chain (or chains) that spontaneously folds into the structure necessary to fulfill its function. Thus, proteins act as the most fundamental link between information encoded in the genome and the physical structures and activities that characterize living systems. Despite their diversity and complexity, all proteins arise from a simple process of random mutation and selection. A mutation that breaks protein functions often causes the cells containing those mutations to be less fit than their competitors, leading to the mutation being swept from the population. Conversely, a mutation that improves protein function relative to the goal of maximizing cellular or organismal fitness will generally increase in frequency within a population.

As with most machines, it is much more common for a random alteration to break a protein than for a random alteration to improve it. Until recently, however, most large-scale analyses of mutations were restricted to naturally-occurring mutations, which are mostly neutral or beneficial. Deleterious mutations generally do not persist in natural systems and must therefore be generated de novo to be studied. This restriction fundamentally limits the types of questions that can be asked about protein evolution. Recent technological advances, though, have allowed “massively parallel” strategies for generating and assaying thousands of random mutations in a single experiment. These strategies allow researchers to capture broad and unbiased pictures of the “space” of possible mutations, including deleterious mutations.

In this chapter, I will survey some of the basic questions at the heart of molecular biology, the massively parallel strategies that researchers have developed to address these questions, and the broad conclusions that have subsequently emerged. Lastly, I will pay special

attention to cases in which multiple proteins and/or organisms come into evolutionary conflict with each other. Because both partners simultaneously fare well, these conflicts highlight the boundaries of what is possible to evolve.

1.2 Multiplex methods for assaying thousands of variants

Over the last several years, multiple labs have independently developed methods for assaying thousands of genetic variants in parallel (Fowler et al., 2010; McLaughlin et al., 2012; Roscoe et al., 2014; Weile et al., 2017). We will restrict this discussion to protein sequences, which are the most commonly studied, although DNA and RNA sequences can be subjected to similar methods. These methods, while varying in the experimental details, share a common general architecture (Figure 1.1). First, mutations are introduced to a genetic sequence of interest. Typically, mutagenesis is done through a process involving the stochastic misincorporation of either individual nucleotides or whole codons. Second, the variant sequences are expressed in a cell that has been designed so that functional sequences are discernible from nonfunctional sequences. In some cases, such as 'complementation' assays, the variant protein is simply expressed in a cell lacking a native copy of that protein, and functional variants confer much faster growth rate to the cell than non-functional variants (Roscoe et al., 2014; Weile et al., 2017; Mighell et al., 2018). In other cases, more sophisticated cell engineering is required, for example to link a particular protein-protein binding interaction to the ability of a yeast cell to produce histidine using a yeast two-hybrid platform (Starita et al., 2015). Third, the functional and non-functional variants are selected or sorted based on the phenotype of interest. Fourth, pools of cells before and after selection or sorting are sequenced using next-generation sequencers to determine the relative abundance of each variant in each population. Fifth, each variant is scored based on its change in abundance, and these scores are compared between different variants to draw inferences about the genetic sequence of interest. This general architecture has been referred to by various names, such as Deep

Mutational Scanning (DMS) (Fowler and Fields., 2014), Multiple Analysis of Variant Effect (MAVE) (Weile et al., 2017), or Site-Saturation Mutagenesis (SSM) (Siloto et al., 2012). We will use 'DMS' throughout, recognizing that these terms overlap but are not always interchangeable.

Although DMS is fairly well defined in terms of the experiments and techniques that are required, it provides a tremendous degree of flexibility in terms of the questions that can be addressed. Common applications include categorizing clinically actionable human genetic variants, identifying antibody epitopes or other important sites on proteins, analyzing fitness landscape topology, and determining protein structure based on epistatic pairs of mutations. The main restriction to this approach is that the function of interest must correspond to a selectable or sortable phenotype at the level of a single whole cell or virus.

1.3 General protein properties and specific protein functions

Most proteins are hundreds of amino acids in length, but only a few of those residues have a direct functional contact with a substrate or binding partner. In some cases (Roscoe et al., 2014; Weile et al., 2017), the presence of deleterious mutations at a particular position is a strong predictor that that position makes a direct functional contact. However, mutations at non-contact positions can also drive strong changes to protein function (Wrenbeck et al., 2017; Rocklin et al., 2018). These effects occur because proteins, in addition to their specific functions, must also fulfill the general requirements for being a protein: they must fold into a stable (or functionally unstable) conformation, they must be trafficked to the right subcellular location, they must eventually be degraded, etc. Because these general properties are prerequisites for the specific functions of the protein, mutations that affect these general properties can also affect specific functions. In particular, many labs have focused on protein stability as a mediator of the effects of mutations (Matreyak et al., 2018; Rocklin et al., 2018; Pokusaeva et al., 2019), in part because of its universality.

Various experimental approaches have been developed to disentangle mutations that affect general protein properties from those that affect specific protein functions. These can generally be grouped into three approaches: assaying a single protein for multiple specific functions, directly assaying for general properties like stability, and assaying multiple mutations co-occurring on the same protein molecule to infer latent variables related to stability (Figure 1.2).

1.4 Assaying single proteins for multiple functions

The first and most straightforward strategy to disentangle types of mutations is to assay a single protein for multiple functions (Figure 1.2a). Mutations that affect common protein properties are presumed to affect all functions of that protein similarly, in contrast to mutations that affect one specific function. The specific functions need not be fully orthogonal. For instance, Melnikov et al. (2014) measured the ability of Kka2 variants to confer resistance to kanamycin and several kanamycin-related compounds, such as G418 and Neomycin. Ranked preferences for different amino acids at each position were highly correlated between different conditions, with Spearman's ρ varying from ~0.6-0.8. This global correlation is consistent with the idea that the majority of mutations are affecting general protein properties. However, this correlation breaks down at specific positions, in some cases inverting to a negative correlation. At these positions, there is no one optimal amino acid for all functions, but instead a tradeoff in which different amino acids are preferred in different contexts. Many of these positions map to direct contacts with the small molecule substrates that vary between conditions. However, not all direct contacts have different amino acid preference between conditions, perhaps because they contact an invariable part of the substrate. Thus, while differences in preferences between conditions are probably sufficient to conclude a position is involved in specific protein functions, the lack of such differences is not sufficient to conclude that a position is not involved in specific protein functions. The inability to conclusively demonstrate that a given position solely affects

general protein properties is one of the notable downsides of assaying multiple functions on a single protein. Regardless, it is probably safe to assume that the vast majority of mutations that cause similar phenotypes across conditions are affecting general protein properties.

More recent studies following this strategy have largely borne out the observations in Melnikov et al. (2014) that most fitness-changing mutations affect general protein properties rather than specific functions, that a small subset of positions have different preferences between conditions, and that these different preferences can often be interpreted in light of structural contacts between the protein and substrate. Notable additions to this framework were made by Stiffler et al. (2015) and Wrenbeck et al. (2017). Stiffler et al. showed that adaptive mutations that confer cefotaxime resistance in Tem-1 do not necessarily impair the ability of Tem-1 to degrade its native substrate, ampicillin. However, very strong selection on ampicillin causes more mutations to be deleterious, including some cefotaxime-resistance mutations. Thus, the strength of selection can reveal (or mask) trade-offs for different functions. Wrenbeck et al. showed that substrate-specific mutations in AmiE can occur at positions far from the enzyme's active site (~9–24Å). These mutations are interpreted to interact with other amino acids, causing cascading effects that subtly alter the shape of the active site. Alternatively, these mutations may alter general protein properties that are more relevant to certain specific functions than to others. For example, if catalysis is rate-limited for one substrate by diffusion of the substrate into the cell but is rate-limited for another substrate by enzyme abundance, then we should expect enzyme stability to affect growth differently on each substrate.

1.5 Directly measuring effects on protein stability

There has recently been significant interest in developing general methods that can assess mutational effects across any protein of interest. Since most mutations are thought to primarily affect protein stability, efforts have focused on assaying stability using a fusion between a protein of interest and a reporter protein (Figure 1.2b). Matreyak et al. (2018)

developed VAMP-seq, in which a protein of interest (here, PTEN) is fused to GFP, and cells expressing the protein are sorted based on fluorescence. Abundance and stability are, under this framework, taken to be proxies of each other, with abundance being the direct antecedent to cellular phenotype. VAMP-seq-derived abundance scores are weakly anti-correlated with other predictors of mutational effect like evolutionary conservation (PTEN, r = -0.27) and computationally predicted ΔΔG scores (PTEN, r = -0.35), and they are weakly correlated with mutational scores derived from functional assays of the same protein (PTEN, r = 0.39). The modesty of these correlations may reflect the technical limitations of sorting and sequencing human cells, or they may reflect complexity in the relationship between protein abundance and cellular phenotype. Mutations that cause loss of PTEN function but not destabilization are much less abundant than destabilizing mutations that maintain PTEN function, suggesting that these differences may reflect real biology. However, more work is needed to understand the relationship between abundance and phenotype across many proteins.

Using a conceptually similar but experimentally distinct strategy, Rocklin et al. (2018) assayed thousands of computationally designed proteins and mutational libraries of three natural proteins for resistance to protease degradation, another proxy for stability. Under this approach, the protein of interest was fused to a c-Myc tag and displayed on the surface of a yeast cell. The proteins were then subjected to degradation by trypsin or chymotrypsin, releasing the c-Myc tag if they were cut. To determine whether the tag was still present, yeast cells were stained with a fluorescent antibody against c-Myc and sorted. A major difference between this research and others is that it focused primarily on designed proteins and very small natural protein domains with the goal of identifying compact structural elements that contribute to stability. However, the authors were also able to identify functional residues by looking for sites of discordance between mutational effects on stability and evolutionary conservation. Positions with high evolutionary conservation but where mutations do not significantly alter stability are presumed to mediate a specific function that relies on a particular

amino acid. In principle, this method is amenable to any particular set of protein variants, including more typical mutational libraries of longer proteins.

Both of the above methods rely on linking a protein variant to a cellular phenotype, then sorting cells. Going forward, selecting populations of protein molecules in a lysate may provide higher accuracy and precision. Technological advances in mass spectrometry have allowed the multiplexed measurement of melting temperature for thousands of proteins in a single experiment (Mateus et al., 2017). However, these strategies are not directly applicable to mutational libraries, because detection of rare mutant peptides amongst a population of similar wild type peptides is very difficult. Similar difficulties in sequencing rare mutations within mutational libraries have been overcome by barcoding DNA variants with short random sequences that do not affect function. It remains to be seen whether barcoding proteins with protein barcodes (Wroblewska et al., 2018; Egloff et al., 2019) or covalently linked nucleotide barcodes might prove similarly fruitful.

1.6 Inferring latent properties from multiple mutations

Some mutations have little observable effect on protein functions but do exhibit effects when in combination with other mutations. The non-additive effects of multiple mutations within a protein is called ‘epistasis.’ In general, epistasis can be divided into specific and nonspecific epistasis (Starr and Thornton, 2016; Husain and Murugan, 2020). Specific epistasis is usually the result of direct contacts between pairs of amino acids. When one amino acid in the pair changes, the optimal amino acid at the other position changes as well. Non-specific epistasis, on the other hand, occurs when each mutation affects some latent (*i.e*., unobservable) property of the protein, and this latent property has a non-linear relationship with the observable functions of the protein. For example, if two mutations each mildly destabilize a protein, but leave it stable enough to function, they might individually have little effect on fitness. However,

when both mutations occur on the same protein, they can destabilize it sufficiently to cause the protein to denature, leading to a strong effect on fitness.

By measuring each mutation in combination with many other mutations, it is possible to estimate the effect of the focal mutation on the latent property that drives non-specific epistasis, which is generally taken to correspond to stability (Figure 1.2c). This strategy was first used by Araya et al. (2012), who described the latent variable as a 'partner potentiation score,' that is, the ability to improve the mean effect of other mutations. The authors isolated several mutations with high partner potentiation scores and showed that they correspond to increased thermal denaturation temperatures. Thus, there seems to be a fundamental link between protein stability and the mean fitness effect of a random mutation, where increased stability buffers weakly deleterious mutations. While mutations that directly alter function are often restricted to a small subset of substrate-contacting residues, stabilizing mutations are more broadly distributed over the protein structure, increasing the number of potentially useful mutations.

This latent variable framework was more explicitly modeled by Pokusaeva et al. (2019), who trained a machine learning algorithm to estimate 'fitness potential,' which can be considered roughly equivalent to a partner potentiation score, though it is calculated differently. By comparing measured fitness values to fitness potential, the authors describe a 'cliff' function where a critical threshold of fitness potential divides high-fitness and low-fitness variants. Individual mutations change 'fitness potential,' but they do not strongly affect fitness unless they cause the whole sequence to cross this critical threshold. Similar to partner potentiation scores, fitness potential correlates strongly to protein thermal stability. Although Pokusaeva et al. focused on a single protein, studies of other proteins have observed similar cliff-like functions for nonspecific epistasis (Schmiedel et al., 2019; Starr and Thornton, 2016). Additional value can be gained by analyzing specific epistasis, non-additive interactions that deviate from this cliff-like function, especially in determining amino acid contacts (Rollins et al., 2019; Schmiedel et al., 2019). However, specific epistasis definitionally depends on the interaction between the

residues in question rather than the effects of those mutations individually. Therefore, specific epistasis may not be informative in clarifying the effect of each component mutation in isolation.

1.7 Generalist and specialist strategies

Many proteins have some degree of flexibility in the functions they carry out. For instance, a DNA-binding protein might have a preferred motif of DNA sequence but tolerate certain permutations in that motif. However, proteins vary in this degree of flexibility, with some following a generalist strategy and others following a specialist strategy. Heat shock proteins and chaperonins, for example, are often generalists, weakly binding many similar client proteins. Antibodies, on the other hand, must bind a particular foreign epitope while minimizing off-target binding. Generalism may be a product of a flexible or dynamic protein structure, with the optimal conformation for binding a particular client (or catalyzing a particular reaction) being stabilized by interaction with that client. However, it unclear whether this flexibility corresponds to decreased thermal stability. Directed evolution studies have consistently found that stabilizing a protein scaffold increases the potential for that scaffold to evolve a new function (Bloom et al., 2006; Tokuriki and Tawfik, 2009a), suggesting that stability is positive associated with the potential to diversify function. However, the connection between the diversity of functions within a given protein (*i.e.*, generalism) and the potential for a protein to evolve new diverse functions has not been well explored.

1.8 Tradeoffs between multiple protein functions

Tradeoffs between different protein functions exist when optimizing for one function negatively affects the other function. The presence of tradeoffs would imply that specialist strategies may be selected for not only because of selection against undesired interactions (as is the case for antibodies), but also because selection for a novel function might be costly with respect to an ancestral function. Individual cases of such tradeoffs are well documented. For

example, in E. coli, Tem-1 breaks down some beta-lactam drugs such as ampicillin, but it is ineffective against newer synthetic beta-lactams like cefotaxime. The G238S mutation, which was first observed in clinical isolates, allows Tem-1 to break down cefotaxime, but comes at a significant cost to activity against ampicillin (Stiffler et al., 2015; Salverda et al., 2010). However, this observation does not imply that all mutations that improve activity on one drug decrease activity on the other. G238A, for instance, also improves Tem-1 activity against cefotaxime, but does not substantially affect activity against ampicillin (Stiffler et al., 2015). That G238S has observed in clinical isolates while G238A has not (Salverda et al., 2010) may reflect that selection for ampicillin resistance is relatively unimportant in the context in which Tem-1 is evolving in the wild.

More generally, the existence of mutations that are beneficial for one function and deleterious for another does not imply that evolution, which samples many mutations, is constrained by tradeoffs. As described above, DMS studies of single proteins for multiple functions have found that most mutations affect different functions similarly, but the exceptions to this rule can correspond to residues that directly contact the ligand. Given that mutations distant to the ligand-binding site can strongly affect fitness (Wrenbeck et al., 2017), further work is needed to determine the extent to which evolution is constrained by tradeoffs at ligand-binding residues.

1.9 Robustness and Evolvability

Many biological systems exhibit robustness, the tolerance to slight perturbations without disrupting the system at large. Proteins, too, exhibit robustness in that many mutations can be tolerated without significant deleterious effects, and the degree of robustness varies between proteins. It remains an open question whether evolution leads to higher or lower robustness than one would expect for a hypothetical non-evolved protein. Theoretical models and evolution of ‘digital organisms’ predict that robustness could be selected for in populations with high

mutation rates and large effective population sizes, because descendants of non-robust sequences would be more likely to harbor deleterious mutations (Wilke, 2001). In agreement with these predictions, some RNA viruses appear to be selected for sequence robustness (Burch and Chao, 2000; Lauring et al., 2012; Lauring et al., 2013). However, this form of selection for robustness is not generally thought to apply to organisms, such as humans, where the per-site mutation rate, μ, is much lower than the inverse of effective population size, 1/Neff (Plotkin et al., 2006).

In principle, robustness could also be selected for as a byproduct of other protein properties. For instance, more stable proteins are likely to tolerate more mutations, so selection for proteins that are able to withstand high temperatures or other stresses could produce mutationally robust proteins. Additionally, errors other than mutations, such as translational errors, occur regularly within cells and are orders of magnitude more common than mutations. Selection for proteins that are able to tolerate these more common errors could also produce mutationally robust proteins (Drummond and Wilke, 2008). Though theoretically plausible, this type of coincident selection for robustness has not been well explored empirically.

While robustness describes the tendency to maintain an existing function, evolvability describes the potential to evolve a new useful function. Whether proteins are more evolvable than one would expect is a more contentious question than whether proteins are robust, as it more explicitly evokes a 'future-oriented' view of evolution. However, fluctuating selection could in principle lead to evolvability. Consider a case where the future environment and past environment are very similar, but both different from the present environment. The protein sequence at present might be surprisingly close to a sequence optimal for a future environment, owing only to selection in a past environment. For example, consider the Bordetella phage BPP-1, whose host periodically alternates between a plus-trophic phenotype and a minus-trophic phenotype. To accommodate these switches, BPP-1 encodes dedicated machinery to hypermutate its tail fiber, allowing it to rapidly adapt to the new host tropism (Guo et al., 2014).

Without considering evolvability, it would be very surprising that adding several random mutations would produce a tail fiber capable of binding an arbitrary new host. However, the evolutionary history of BPP-1 has included selection on both tropisms of the host, likely causing the tail fiber scaffold to be well suited for adapting to either host tropism. Additionally, the BPP-1 tail fiber forms a highly stable C-lectin type fold, which may allow it to tolerate more mutations than less-structured tail fibers (Dai et al., 2010).

Absent these possibly rare cases of predictably changing environments, is selection for evolvability possible? For example, do proteins that must evolve rapidly display properties of being able to generally evolve to a broad set of potential environments? For some viral proteins, very rapid evolution is driven by the need to escape host antibodies. Influenza hemagglutinin displays hypervariable loops that are more highly immunogenic than the conserved stalk domains and can tolerate most mutations without losing function (Xu et al., 2013). These loops can therefore be seen as structures whose primary purpose is to increase evolvability by decoupling antibody escape mutations from the functional portions of the protein. Outside of rapidly evolving viruses and some bacteria (Burch and Chao, 2000; Merrikh and Merrikh, 2018; Woods et al., 2011), there has been little direct evidence for selection acting on evolvability, although this may change with the explosion of mutational data available for human proteins.

1.10 Evolutionary conflicts

Evolutionary conflicts, exemplified by host-pathogen interactions, set up interesting cases for studying the limits of evolution, because the opposed organisms cannot both ‘win.’ Hosts are generally under strong selective pressure to escape from pathogens, while pathogens are under strong selective pressure to overcome that resistance or expand to a new host population. In the simple case, the host protein acts as an escapee, seeking to minimize interactions while maintaining normal function, and the pathogen protein acts as a pursuer, seeking to maximize interactions. However, these roles can also be reversed, for instance when

a pathogen protein must escape from host antibodies or restriction factors. This escapee-pursuer dynamic is fundamentally asymmetrical because mutations that disrupt binding are much more common than mutations that increase binding, giving the escapee an ostensible advantage. To compensate, the pursuer may take advantage of higher mutation rates and shorter generation times (e.g., many viruses), or employ dedicated diversification machinery (e.g., antibody V(D)J recombination). The escapee may also be limited by tradeoffs between reducing pathogenic interactions and maintaining the normal functions of the evolving protein. Spatially decoupling pathogenic interactions and protein function, for example by occluding the functional site with more mutationally tolerant residues, may help alleviate these tradeoffs. Despite the fundamental asymmetry between escapees and pursuers, many host-pathogen systems appear stable, with hosts and pathogens both maintaining large population sizes indefinitely. This observation has prompted some researchers to question the importance of 'arms races,' in resistance and counter-resistance each rapidly escalate, because this type of escalation cannot go on forever if resistance or counter-resistance incurs any non-zero costs to growth rate (Koskella and Brockhurst, 2014). Phage-bacteria coevolution studies that take pains to mimic natural environments have shown that in complex environments with scarce resources, arms race dynamics are replaced by 'fluctuating selection dynamics' (Gomez and Buckling, 2011; Lopez-Pascua et al., 2008). Under fluctuating selection dynamics, bacteria and phages adapt to resist / counter-resist the strains in their local milieu, but rapidly lose these adaptations once their milieu changes. These experiments suggest that resistance or counter-resistance mutations do not generally accumulate, but functionally revert once they are no longer useful, only to be replaced by new resistance or counter-resistance mechanisms. Furthermore, the reversion of resistance and counter-resistance implies that they are costly in some way (Poullain et al., 2007). Although fluctuating selection dynamics have been well demonstrated in natural-like environments, there is little in the way of described molecular mechanisms. It is not known, for instance, whether the observed resistance mechanisms are driven by missense

mutations in the receptor or, perhaps, by changes to transcriptional regulation of the receptor. In general, unification is needed to link the observations of forward genetic screens and experimental evolution studies with the reverse genetics approaches like DMS that have become widely used in recent years.

1.11 Concluding remarks

Over the last decade, protein biology has shifted from a perspective based on single protein sequences toward a view that encompasses the local spaces of possible protein sequences (still centered around a single protein sequence as the parent). In doing so, we have learned a great deal about the process of protein evolution. A few ideas have arisen separately in many studies and have effectively reached consensus in the field. First, many mutations, if not most, are neutral or have very small effect. Most proteins have some positions that can tolerate any amino acid. Second, deleterious mutations usually affect general protein properties, like stability, rather than changing anything specific about how the protein interacts with its ligand(s). When proteins have multiple functions, a typical deleterious mutation will affect all of those functions to a similar degree. Third, a small subset of mutations can alter functions of the protein in a specific way. These mutations often correspond to direct contacts with the ligand but can be more spatially distributed in other cases. Fourth, while some mutations that promote one function come at a cost to a different function, there is only limited evidence to suggest that evolution is strongly constrained by tradeoffs between functions. Mutations that improve multiple functions, or improve one function while being neutral toward others, are reasonably common. Fifth, proteins vary in intrinsic qualities such as generalism (the ability to perform many functions), robustness (the ability to tolerate mutations without losing function), and evolvability (the ability to acquire new useful functions). These qualities may affect evolutionary outcomes, but further research is needed to determine whether selection acts directly on these properties in typical proteins. Sixth, coevolving systems reveal limits to evolution, and suggest that proteins

that confer resistance (or counter-resistance) may rapidly revert once resistance (or counter-resistance) is no longer selected for. For rapidly evolving proteins, it is therefore plausible that the protein's evolutionary history has involved selection for functions that no are no longer performed by the protein.

The set of protein sequences in extant species is much smaller than the space of functional proteins. Moreover, extant proteins are not a random sampling of that space – they are shaped by their exploration through it. As recent technologies have allowed us to survey the space of possible protein sequences, rather than individual sequences, we are learning about the processes that drive this exploration. A major hurdle that remains is that many of the protein properties that are important, such as stability or generalism, are not directly measured but must be inferred from changes to cellular fitness. Technological strategies to directly measure protein properties in bulk pooled formats are being developed, and they are badly needed to further dissect the space of protein sequences. Furthermore, while edge cases with very high rates of evolution, such as BPP-1, have revealed much about the limits of evolution, it is not always obvious how well they represent more typical proteins. Unified technological and theoretical frameworks for deep mutational scans would greatly benefit the field.

1.12 Figures and Tables

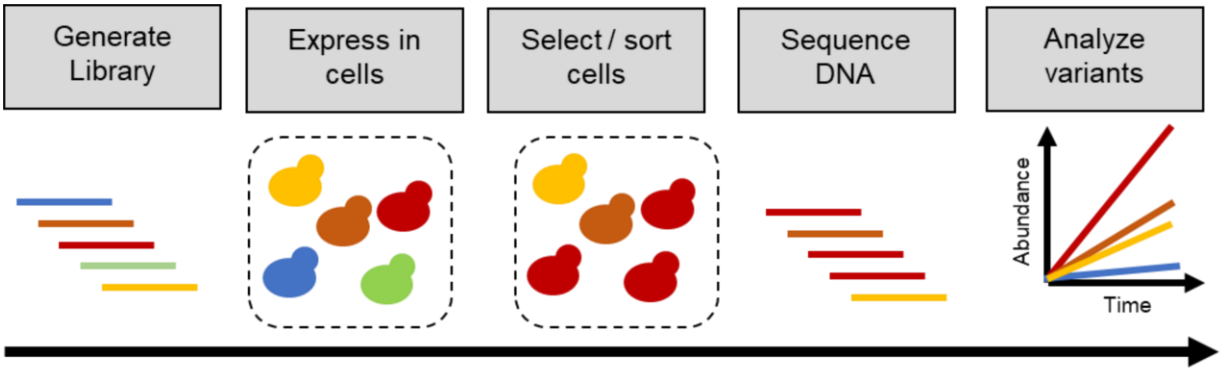

Figure 1.1 Deep Mutational Scanning workflow.
Deep Mutational Scanning is a flexible experimental approach for assessing the effects of thousands of genetic variants in parallel. First, the library is generated, typically by making random mutations to a single parental DNA sequence. Second, the library is expressed in a cell system that is chosen such that the variants will confer a selectable or sortable phenotype. Third, the cells are selected or sorted. Fourth, the input and output libraries (and sometimes intermediate timepoints) are sequenced to determine the relative frequency of each variant. Last, an analytical framework is applied that scores variants according to their phenotypic effects.

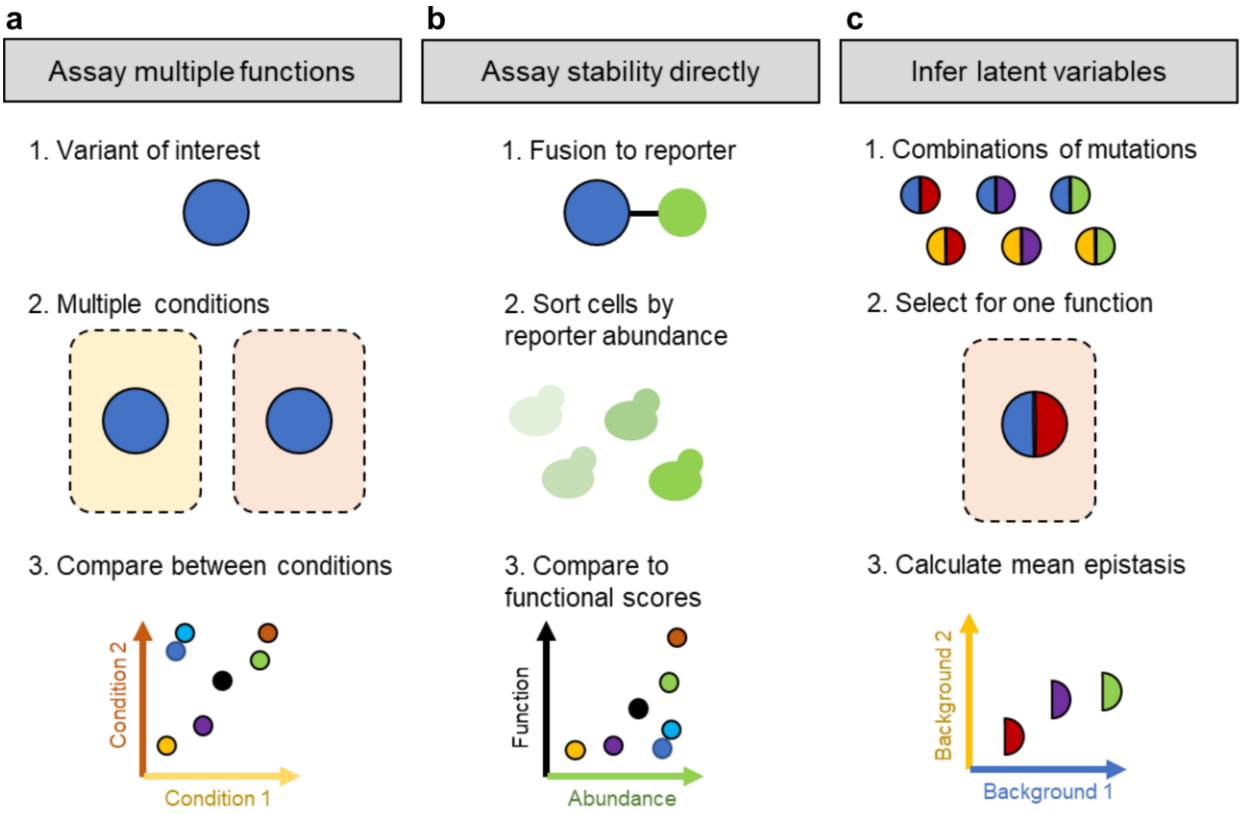

Figure 1.2 Three approaches to separating general and specific effects of mutations. **(a**) The simplest approach to decoupling general and specific mutational effects is to assay the same variants in multiple conditions. Variants that behave differently (relative to other variants) in each condition confer specific effects. However, variants that behave similarly in more than one condition cannot be assumed to confer general effects, as there may be some other condition that would reveal specificity. **(b)** By directly assaying for stability, it is possible to separate functional effects that are mediated by destabilization from those that are not. Current methods rely on fusing the variants of interest to fluorescent reporter, and they may have significant noise. **(c)** When multiple mutations co-occur on the same molecule, the effect of one mutation on stability can be inferred by the average effect of all mutations on the background of that one mutation. Mutations that confer mild destabilization tend to sensitize the protein to other slightly deleterious mutations, while stabilizing mutations do the opposite.

Chapter 2: Distinct Patterns of Mutational Sensitivity for λ Resistance and Maltodextrin Transport in *Escherichia coli* LamB

Bacteria can evade cohabiting phages through mutations in phage receptors, but these mutations may come at a cost if they disrupt the receptor's native cellular function. To investigate the relationship between these two conflicting activities, we generated sequence–function maps of E. coli LamB with respect to sensitivity to phage λ and transport of maltodextrin. By comparing 413 missense mutations whose effect on both traits could be analyzed, we find that these two phenotypes were correlated, implying that most mutations affect these phenotypes through a common mechanism such as loss of protein stability. However, individual mutations could be found that specifically disrupt λ-sensitivity without affecting maltodextrin transport. We identify and individually assay nine such mutations, whose spatial positions implicate Loop L6 of LamB in λ binding. Although missense mutations that lead to λ-resistance are rare, they were approximately as likely to be Mal^+ as Mal^-, implying that E. coli can adapt to λ while conserving the receptor's native function. We propose that in order for E. coli and λ to stably cohabitate, selection for λ-resistance and maltose transport must be spatially or temporally separated.

A version of the chapter has previously been published as:

Andrews, Bryan, and Stanley Fields. "Distinct patterns of mutational sensitivity for λ resistance and maltodextrin transport in Escherichia coli LamB." Microbial Genomics 6.4 (2020): e000364.

2.1 Introduction

Individual proteins frequently carry out multiple distinct activities. When pathogenesis is involved, two activities can be in conflict, with one benefitting the host cell and the other the

pathogen. For instance, in the case of some host cell surface proteins, one activity may be to transport a nutrient while another is to serve as the receptor for a virus (Gurnev et al., 2006; Bertozzi Silva et al., 2016). The cell can evade infection via a disruptive mutation in the gene for the receptor, but such mutations also abolish the native function of the receptor (Shoval et al., 1998). A solution for the cell might be a specific missense mutation that disrupts interaction with the pathogen while maintaining native function, but mutations of this type are typically much rarer than general loss of function mutations (Hottes et al., 2013).

Phage λ is one of many phages that bind to a nutrient transporter, recognizing and infecting through the maltose-specific porin LamB on the E. coli outer membrane (Berkane et al., 2006; Hendrix and Casjens, 2006; Szmelcan et al., 1976). When E. coli is cultured with λ in rich media, mutations accrue in the maltose regulon, most commonly large and/or frameshifting deletions in the lamB gene or its transcriptional activator malT (Lederberg, 1955; Thirion and Hofnung, 1972). LamB is a trimeric β-barrel pore that spans the outer membrane of E. coli and facilitates the diffusion of maltose and maltose-derived oligosaccharides like maltodextrin into the periplasm. Loss-of-function mutations in lamB lead to cells that are incapable of transporting these nutrients, therefore incurring a cost in some environments, although rare missense mutations have been identified in lamB that confer λ resistance and maintain the ability of LamB to transport maltose (Charbit et al., 1988, Heine et al., 1988).

The structure of LamB consists of 18 transmembrane β-sheets, each linked to the adjacent sheets by a short linker on the periplasmic face and a long or short loop on the extracellular face (Schirmer et al., 1995). The extracellular loops are the most rapidly evolving portions of the protein and play a dual role in specifying the transported substrates and forming binding sites for proteins of pathogens, including λ (Benz et al., 1987; Zheng et al., 2004). In particular, the extracellular Loops L4, L5, L6, and L9 are highly diverged within Enterobacteriaceae (Figure 2.1), likely due to selective pressure to avoid phages. Most variation

within these loops is missense, as opposed to indels or structural variation, suggesting that missense mutations drive much of the long-term evolution of this protein.

In this study, we set out to explore the space of possible missense mutations in lamB and how they affect both maltodextrin transport and λ infectivity. By scoring both activities, we observe global correlation between λ-sensitivity and maltodextrin transport, suggesting that most mutations act through a mechanism like destabilization that affects all LamB activities. Mutations that affect the two activities occurred throughout the LamB structure. At a finer scale, individual mutations that specifically affect a single activity could be identified, and they strongly implicate the Loop L6 in determining λ-sensitivity.

2.2 Results

2.2.1 A sequencing-based approach to map the effects of mutation on lamB

We first sought to determine the fitness consequences for E. coli of lamB expression under different selective conditions. Deletion of lamB confers a large fitness advantage in LB media containing λ (Figure 2.2a), but a large cost in M9+maltodextrin, media in which maltose-derived oligosaccharides are the sole carbon source (Figure 2.2b). We therefore decided to use these conditions to categorize variants with respect to λ-sensitivity and maltodextrin transport using a pooled assay, in which variants compete for growth either in the presence of λ (in glucose media) or in maltodextrin as sole carbon source, the frequency of each mutation in selected and control conditions determined by high throughput DNA sequencing (Figure 2.2c) (Araya et al., 2012; Fowler and Fields, 2014).

To generate variants for pooled assays, we produced two libraries, each consisting of greater than 1 million variants of lamB, by amplifying the 1,341 bp coding sequence of lamB with error-prone PCR with different amounts of starting template. The libraries had per-base error rates of 0.52% and 0.063%, as determined by comparing the number of mismatches in successfully aligned fragments to the number of matched aligned bases in the control libraries.

Initial tests on 11 random variants from each library showed only moderate loss of λ-sensitivity (6/11 from each library retained sensitivity), indicating that even with a high mutational burden, a large fraction of variants produced functional protein (Figure 2.3). Because both libraries had many λ-sensitive variants, we used the higher mutation rate library (7 mutations/variant) in the λ selection. However, we elected to use the lower mutation rate library (0.84 mutations/variant) for the maltodextrin selection to get a better estimate of the effects of single mutations.

We induced expression of the plasmid-borne variant libraries by IPTG in the lamB-deficient strain DH10B(lamB$^{\Delta}$), and diluted the cultures 1:100 into control media (LB for the λ selection or M9+glucose for the maltodextrin selection) and selective media (LB+λ for the λ selection or M9+ maltodextrin for the maltodextrin selection) in biological triplicate. In the λ selection, we used the sam7 strain of λ, which carries a lysis gene interrupted by an amber stop codon. Because this strain infects susceptible cells without releasing λ progeny into the culture, it prevents the phage population from dramatically increasing over the course of the experiment, and λ cannot evolve adaptations to the variation in lamB. After overnight growth of the cultures, we isolated plasmids from each and amplified the lamB coding sequence and ~200 bp flanking sequence with 12 cycles of PCR.

Rather than attempting to measure the frequency of each full-length lamB variant, we instead measured the frequency of each mutation, disregarding other co-occurring mutations. We assume that the average function of variants containing a given mutation, compared to variants not containing it, will reflect the functional effect of that mutation on the wild type protein (Doud and Bloom, 2016; Weile et al., 2017). To correct for sequencing errors, we developed a strategy of randomly fragmenting the gene with Tn5, and using the start site, stop site, and mutations of each fragment as an identifier by which to construct a consensus sequence, using a minimum count of five reads to call a fragment. For each library, we generated ~50,000 fragments, with a median fragment length of 63 bp (Figure 2.4a). The number of sequenced fragments was several orders of magnitude less than the ~10^8 possible fragments, such that

each unique fragment should represent a single transposition event. For each base in the sequence, we counted the number of fragments containing the mutant base and the wild type base. We calculated an enrichment score for each mutation as the log2 ratio of wildtype-normalized counts between the LB+λ and LB conditions or the M9+maltodextrin and M9+glucose conditions.

2.2.2 lamB missense mutations that confer λ-resistance

We estimated a function score for λ sensitivity (F_λ) by scaling the enrichment scores from the λ selection such that the median nonsense mutation, expected to be λ^r, was assigned a score of 0 and the median synonymous mutation, expected to be λ^s, was assigned a score of 1. Note that this scaling reverses the axis such that enriched variants (with resistance to λ) have lower scores. We ignored mutations represented by fewer than five fragments in the control condition. This fragment cutoff was based on the number of mutations that passed the cutoff and the concordance between replicates (Figure 2.5a). At this threshold, replicates were moderately correlated, with a mean Pearson's r = 0.443 between biological replicates and r = 0.481 between technical replicates. The relatively modest correlation reflects the scores as being averaged over many different variants containing each mutation, and that these variants tended to be different in each replicate due to the large size of the library. To reduce and quantify error, we averaged scores over six replicates, consisting of three biological replicates each performed in two technical replicates.

As expected, nonsense mutations had low F_λ scores compared to synonymous or missense mutations (Figure 2.5b,c). By comparison, synonymous mutations scored relatively well and did not overlap much with the nonsense distribution. Among missense mutations, there was a wide distribution of scores. Over most of the range, the missense mutations were similarly distributed to the synonymous mutations. However, the missense distribution had a longer tail to the left (Figure 2.5c). To classify missense mutations as 'λ-sensitive,' we set a threshold at $F_\lambda = 0.5$, the midpoint between the median nonsense ($F_\lambda = 0$) and median

synonymous scores ($F_\lambda = 1$) (Figure 2.5c). A total of 302 missense mutations (15%) fell below this threshold, along with 79% of nonsense variants and 5% of synonymous mutations.

Spatially, the most λ-resistant and the most λ-sensitive mutations fell throughout the protein structure, including in both the transmembrane β-sheets and the extracellular loops (Figure 2.5d). Stop codons were deleterious (*i.e.* conferring λ-resistance) at any point in the coding sequence, even the last few residues. Most of the trans-membrane strands facing against the lipid bilayer primarily contained λ-resistant mutations, with the exception of the C-terminal strand, which contained many λ-sensitive mutations (Figure 2.6). By contrast, most of the trans-membrane strands facing the other monomers contained an abundance of λ-sensitive mutations, with the exception of the N-terminal strand (Figure 2.6). The relative intolerance to mutation of the N-terminal strand compared to the C-terminal strand, despite their physical proximity, suggests that N-terminal mutations may be disrupting insertion of the protein into the membrane, which occurs in the N-to-C direction, (Zhang and Han, 2016) rather than protein stability per se. However, this hypothesis fails to explain why mutations at the trimer interface were relatively well-tolerated to maintain λ-sensitivity. LamB functions only as a trimer, and we hypothesize that it trimerizes with sufficient affinity that most single mutations would fail to disrupt assembly. This result is consistent with the extreme stability of LamB trimers, which withstand temperatures up to 70ºC in 9M urea (Baldwin et al., 2011).

2.2.3 Most lamB mutations do not disrupt maltodextrin transport

We next investigated the fitness effects of lamB mutations on the transport of maltodextrin. We estimated functional maltodextrin transport (F_{malt}) by scaling the enrichment scores such that the median nonsense mutation, expected to be Mal^-, was assigned a score of 0, and the median synonymous mutation, expected to be Mal^+, was assigned a score of 1. Because of the overall lower mutation rate in the library used for this selection (each lamB variant on average contained 0.84 mutations), we captured fewer mutations overall, with lower fragment counts and higher variance. Replicates correlated only modestly (r=0.26), and thus to

reduce error, we again averaged scores over six replicates, consisting of three biological replicates each performed in two technical replicates (Figure 2.7a). Regardless, we were able to score hundreds of mutations for their effect on the transport of maltodextrin.

We observed a separation of the nonsense and synonymous mutations, with the missense mutations spanning the range of both distributions (Figure 2.7b). We set a threshold for assigning mutations as Mal^+ or Mal^- midway between the median nonsense ($F_{malt} = 0$) and median synonymous mutation ($F_{malt} = 1$), at $F_{malt} = 0.5$. At this threshold, 166 missense mutations (78%) fell above this threshold and were called Mal^+, along with 81% of synonymous mutations and 16% of nonsense mutations (Figure 2.7c). Unlike in the selection for λ resistance, in the selection for growth on maltodextrin, the distributions of missense and synonymous mutations were not statistically significantly different from each other (p=0.65, Kolmogorov-Smirnov test) (Figure 2.7b), implying that most lamB missense mutations are not highly disruptive to maltodextrin transport. The nonsense distribution was significantly different from both the missense and synonymous distributions ($p<10^{-6}$, Kolmogorov-Smirnov test). We conclude that most single amino acid substitutions in lamB do not destroy the ability of the protein to transport maltose derivatives, in contrast to the observation that many mutations in the transmembrane sheets confer at least partial λ-resistance. This finding can be explained if sensitivity to λ and maltose transport differ in the amount of LamB protein needed to confer each activity, or if lamB is sensitized to the effects of mutations by the higher overall mutation rate in the λ selection. High- and low-scoring mutations fell in all regions of the protein, with no region appearing broadly tolerant of mutations (Figure 2.7d).

2.2.4 Comparison of the sequence–function maps for resistance to λ and maltodextrin transport

Scores for λ-sensitivity and maltodextrin transport were correlated ($r = 0.495$, $p<10^{-6}$), and plotting mutations on both axes separates $\lambda^s Mal^+$ synonymous mutations from $\lambda^r Mal^-$ nonsense mutations (Figure 2.8a). Simply asking whether $(F_{malt}+F_\lambda)/2 > 0.5$ is sufficient to

correctly classify 92% of these mutations as nonsense or synonymous mutations, which is higher accuracy than using either selection in isolation. Among missense mutations, the overall correlation of λ-sensitivity and transport of maltodextrin implies that most mutations affect a common protein property, such as stability or trafficking, that is important for both phenotypes. At a finer scale, individual mutations can be identified that appear to specifically interfere with only one activity. However, the identification of these mutations necessarily depends on the score thresholds used to categorize mutations.

In the previous analyses, we used thresholds at F = 0.5 to phenotypically categorize mutations. In the case of the maltodextrin selection, this threshold is probably close to optimal, as fitting Gaussian distributions to the synonymous and nonsense distributions yields an intersection at F = 0.51 that separates the functional and non-functional distributions. In the case of λ-sensitivity, it is more difficult to determine the most appropriate threshold because the score distributions are non-Gaussian and because many missense mutations fall in an intermediate range between the synonymous and nonsense ranges. We therefore decided to take an empirical approach to refine our estimate of the appropriate F_λ threshold. We individually assayed 20 putatively $\lambda^r Mal^+$ mutations with F_λ scores ranging from -0.5 to 0.5. We cloned each of these mutations back into the parent plasmid by site-directed mutagenesis (see Methods), and measured growth curves for each resulting transformed strain in M9+maltodextrin and LB+λ. Of the 20 variants, all of them grew in M9+maltodextrin, although several had either a delay in reaching log-phase, or a decreased maximum growth rate, or both. Nine were fully λ-resistant, 10 were λ-sensitive and one was of intermediate resistance. The mutations that were shown to be individually λ-sensitive were close, in the pooled assay, to the threshold of $F_\lambda = 0.5$, with the lowest $F_\lambda = 0.312$ (Figure 2.8a). By contrast, the λ-resistant mutations had an average $F_\lambda = 0.124$ (Figure 2.8a). From these results, we conclude that decreasing the threshold to $F_\lambda \approx 0.3$-0.35 effectively removes moderately scoring mutations that

do not confer full resistance to λ, although mutations above the threshold may confer partial resistance in the context of a pooled assay and/or other mutations.

2.2.5 Mutations conferring $\lambda^r Mal^+$ phenotype are restricted to Loop L6

Of the nine individually assayed mutations sufficient to confer $\lambda^r Mal^+$, five occur in the extracellular Loop L6, and three others in residues that directly contact this loop (Figure 2.8b). The concentration of $\lambda^r Mal^+$ mutations in this loop strongly implicates Loop L6 as a key structural determinant of λ binding. While λ-resistance mutations in Loop L6 have been previously reported (Gehring et al., 1987; Charbit et al., 1988), the loop structural feature has not been described as a hotspot for λ-resistance and is less surface-exposed than many of the other extracellular loops (Gehring et al., 1987). The ninth individually assayed $\lambda^r Mal^+$ mutation falls in the signal peptide and is not part of the mature protein. We hypothesize that reducing LamB protein levels by altering the signal peptide, without completely abrogating expression, confers some level of λ-resistance. The remaining 11 mutations appeared to confer $\lambda^r Mal^+$ in the pooled selection but were not sufficient to do so in individual assays. These mutations were more spatially scattered but include changes to residues that face the inside of the lipid bilayer, for instance, and might be expected to affect protein stability or abundance, but not λ-binding directly. That these mutations exist but have comparatively moderate F_λ scores (Figure 2.8a) suggests that the stability or abundance effects of mutations are unlikely to act as the primary drivers of λ-resistance in a population with many alleles of lamB.

2.2.6 Among λ^r mutations, Mal^+ and Mal^- phenotypes are approximately equally likely

To predict how the sequence–function map of λ-sensitivity might affect how E. coli acquires resistance to λ, we asked whether λ^r missense mutations are more likely to confer a Mal^+ or Mal^- phenotype. Because our individual assays suggested that the appropriate F_λ threshold for classifying λ^r mutations is likely lower than $F_\lambda = 0.5$, closer to $F_\lambda \approx 0.3$-0.35, we decided to compare Mal^+ and Mal^- mutations across a broad range of F_λ thresholds, rather than picking a single threshold. Below a threshold of $F_\lambda = 0.4$, the ratio of Mal^+ to Mal^- approaches 1.0

(Figure 2.9a), implying that a randomly chosen λ^r mutation is approximately equally likely to be Mal^+ or Mal^-. This effect can also be seen by looking at the distribution of F_{malt} scores. If $F_\lambda < 1$, the median F_{malt} is ~1.0 (comparable to a synonymous mutation), but if $F_\lambda < 0.4$, the median F_{malt} drops to ~0.5 (Figure 2.9b). The median F_{malt} score does not change appreciably between $F_\lambda = 0.0$ and $F_\lambda = 0.4$. We interpret these data as implying that when λ-resistance is mediated by a randomly chosen missense mutation, this mutation is approximately equally likely to disrupt or maintain maltodextrin transport. This result seems surprising given that there are many more ways to disrupt the structure of a protein compared to specifically disrupting a single binding site for a pathogen. However, the extreme stability of folded LamB may reduce the number of mutations that prevent correct folding, thereby resulting in many λ-resistance mutations that still allow maltodextrin transport (Araya et al., 2012; Baldwin et al., 2011; Bloom et al., 2006).

2.3 Discussion

By carrying out assays of LamB variants for their ability to promote two activities – infection by λ and transport of maltodextrin – that inherently conflict, we establish independent sequence-function maps for this single protein. Most missense mutations have similar effects on both activities, pointing to the primacy of fundamental protein properties, like stability, on the effects of mutations. Despite the overall agreement of mutational effects between the two selections, mutations that specifically disrupt a single function, λ-sensitivity, could be isolated and were highly spatially clustered.

These results are subject to certain caveats. First, the mutational complexity of the libraries differed between the two selections; second, the strength of selection may have differed between selections; and third, mutations were scored by comparing heterogeneous variants that do and do not contain each mutation rather than directly assessing a single mutation in a wild type background. These caveats are consistent with some amount of noise in the data, such as the fact that only 92% of synonymous and nonsense mutations can be

correctly classified as such using their F_λ and F_{malt} scores, and that only 9/20 mutations putatively called as λ-resistant were sufficient to confer complete λ-resistance in individual assays. While all technical strategies to measure phenotypes have limitations, we observed a clear separation by function in both selections, demonstrating that the overall distribution of fitness effects is meaningful. Furthermore, by individually assaying mutations, we were able to refine our set of $\lambda^r Mal^+$ mutations, improving our estimates of the appropriate F_λ threshold for determining resistance and helping us narrow in on Loop L6 as the major structural determinant of λ binding.

Nearly all of the missense mutations that conferred complete λ-resistance in individual assays are involved in a single structural motif, Loop L6. Co-crystallization of LamB with maltose suggests that maltose binding is mostly coordinated by Loop L1 and Loop L3, along with the inward faces of some transmembrane beta-strands (Schirmer et al., 1995; Benz et al., 1987). This spatial segregation likely explains why some lamB mutations can disrupt λ binding while maintaining maltodextrin transport. Given that $\lambda^r Mal^+$ mutations are relatively rare and involve only a single loop of the 18-loop LamB, it might be expected that most λ-resistance mutations would be general loss-of-function (*i.e.* conferring the $\lambda^r Mal^-$ phenotype). However, we find that approximately half of the missense mutations that led to λ-resistance did not eliminate maltodextrin transport. We therefore expect that in an environment containing both λ and maltodextrin, E. coli should be able to directly acquire λ-resistance in a single step, without sacrificing maltose transport. An alternative two-step pathway would be for E. coli to first acquire λ-resistance via a general loss-of-function mutation ($\lambda^r Mal^-$) followed by a compensatory mutation that reverts the phenotype to $\lambda^s Mal^+$ once λ has left the environment. Although the single-step pathway seems more advantageous to the bacteria, it can precipitate an evolutionary arms race if λ acquires counter-resistance, driving either E. coli or λ to extinction (Dennehy, 2012; Lenski, 1984; Weitz et al., 2005). In experimental evolution studies that compete λ with E. coli, λ is typically the first to go extinct, yet we know that phage populations

and species can persist for decades in the wild (Koskella and Brockhurst, 2014; Lenski and Levin, 1985; Schrag and Mittler 1996). When species are known to coevolve stably for long periods of time, a different set of dynamics, called Fluctuating Selection Dynamics, are often presumed to dominate (Gómez and Buckling, 2011; Hall et al., 2011). Fluctuating Selection Dynamics would be more consistent with the two-step pathway, with E. coli periodically fluctuating between λ^r and λ^s phenotypes, but such a model would incorrectly predict that $\lambda^r Mal^+$ mutations should be difficult to acquire relative to general loss-of-function mutations.

To square our results with these evolutionary considerations, we hypothesize that selection for maltose transport is close to zero in environments where λ and E. coli interact (Hochberg et al., 2011; Lopez-Pascua and Buckling 2008). This hypothesis predicts that the relative abundance of $\lambda^r Mal^+$ and $\lambda^r Mal^-$ mutations should be equal rather than weighted toward $\lambda^r Mal^-$. Furthermore, phages are thought to cause more selection in nutrient-rich conditions like the human gut, where maltose is not a preferred carbon source (Hochberg et al., 2011). The maintenance of the Mal^+ phenotype could thus be driven by exposure to periodic nutrient-poor conditions when phages are mostly absent and Fluctuating Selection Dynamics are thought to be favored (Lopez-Pascua et al., 2014; Boots, 2011).

A prediction that arises from this work is that a receptor whose loss is tied to a more significant fitness defect in phage-containing environments should have phage-resistance mutations that are more strongly biased towards knocking out the protein's native function. Most phages use receptors with non-essential cellular functions, despite essential receptors being slower to evolve phage resistance (Bertozzi Silva et al., 2016). Future sequence-function maps should assess the role of missense variation in the acquisition or maintenance of phenotypes in different environments.

2.4 Materials and Methods

2.4.1 Strategy

We adopted a strategy known as “deep mutational scanning” (Fowler and Fields 2014), wherein, typically, cells carrying a library of genetic variants in a gene are subjected to a selection condition that requires a function of that gene, and mutational effects are estimated from the frequency changes of each variant as determined by deep sequencing. We generated a library of variants in lamB using error-prone PCR, cloned the library into a lac-inducible expression vector, and expressed the variants in an E. coli strain knocked out for lamB. Cells were then selected separately either for their ability to grow in rich media containing λ, or in minimal media with maltodextrin as a sole carbon source. The lamB gene was amplified from cells after the selection conditions and control conditions, and mutation frequency was estimated by short-read sequencing of random fragments of the lamB amplicon. We calculated functional scores based on differences in the frequency of each mutation in the selection condition compared to the corresponding control condition and used score thresholds to categorize mutational effects. For some mutations categorized as λ-resistant and Mal^+, we assayed the effect of the individual mutation in an otherwise wild type lamB background.

2.4.2 Strains

All assays were done in a DH10B background. DH10B($lamB^{\Delta}$) was generated using the λ-red system. Briefly, pSIM5 (Court et al., 2007) was transformed into DH10B and was induced by heat shock at 42ºC. A linear cassette containing the genes tetA and sacB flanked by homology to the lamB locus was then transformed in, and recombinants were selected by tetracycline. After verifying the lamB deletion, the cassette was removed by transforming an oligonucleotide that annealed flanking the cassette and selecting on fusaric acid and sucrose. The deletion of the lamB coding sequence was verified by Sanger sequencing and the strain was cured of pSIM5 by growing at 37ºC and choosing a chloramphenicol-sensitive colony.

λ_{sam7} was generated by packaging λ DNA from NEB (#N3011, cI857 ind1 Sam7) into λ particles using the MaxPlax λ packaging extract from Lucigen (#MP5120). Prior to the assay, ~10^5 packaged phages were plated on host-strain LE392MP (Lucigen #SS000437-D), grown for

5 hours at 37ºC, and washed into TMS (10mM Tris-HCl+10mM MgSO4+100mM NaCl). The lysate was cleared by centrifugation and filtered through a 0.2µm filter and had a titer of 3.2 x 10^8 pfu/mL. The λ_{sam7} is amber suppressible for lysis, such that it will not lyse cells that recognize an amber codon as a stop codon. Additionally, λ_{sam7} contains the temperature sensitive cI857 allele, such that it is obligate lytic at 37ºC, but can lysogenize at 30ºC. In all cases in this paper, infection with λ_{sam7} was carried out at 37ºC.

λ_{JL801} was a gift from Ben Kerr, and a plate lysate was generated via the same method as for λ_{sam7}, except DH10B was used as the host. The titer was 9.3x10^8 pfu/mL. λ_{JL801} differs from λ_{sam7} in that it is not amber suppressible, and the cI gene contains a large deletion in the 5' end, rendering the phage obligate lytic.

2.4.3 Error-prone PCR and library generation

The lamB gene was amplified from DH10B using PCR and cloned into the lac-inducible high-copy expression vector p44K (addgene #45800) using Gibson assembly (Gibson et al., 17) to produce p44K-lamB. A miniprep of this plasmid was used as the template for error-prone PCR with the Agilent GeneMorph II kit, using 800 ng or 2400 ng of plasmid per reaction. PCR products were treated with DpnI to remove template plasmid, cleaned up on a Zymogen DNA Clean & Concentrator column, and cloned back into the plasmid backbone using Gibson Assembly. The assembled plasmid was transformed into DH10B, obtaining an estimated 1.1x10^6 transformants per library. The plasmid library was miniprepped and transformed back into DH10B($lamB^{\Delta}$).

2.4.4 Media and culture conditions

For selections, 200 µL stock of the library was inoculated into 5 mL LB and shaken for 1 hour at 37ºC to recover. Carbenicillin was then added to 50 µg/mL and IPTG to 100 µM, and the culture was shaken for 1 additional hour, or until the culture reached OD600 = 0.4. For each biological replicate, 1 mL cells were then spun down (30 s x 12,700 rpm) and resuspended in 200 µL TMS.

For the λ-resistance selection, the cells were diluted into 10 mL LB+0.2% maltose+10 mM MgSO4+100 μM IPTG+50 μg/mL carbenicillin for the control condition, or the same buffer with 10^6 pfu/mL λ_{sam7} for the selection condition. The cultures were shaken for 16 hours at 37ºC, then collected.

For the maltodextrin selection, the cells were diluted into M9+0.4% glucose+80 μg/mL leucine+100 μM IPTG+50 μg/mL carbenicillin for the control condition, or the same media with 0.4% maltodextrin and no glucose for the selection condition. The cultures were shaken for 36 hours at 37ºC, then collected.

2.4.5 Sequencing library preparation

After the plasmids were miniprepped from the selections, the lamB coding sequence and flanking ~100 bp were amplified using qPCR for 11 cycles, and the template was digested with DpnI. The amplicon was then subjected to tagmentation using the Illumina Nextera kit. The fragments were amplified and indexed by a further 5 cycles of PCR, then prepared for Illumina sequencing.

2.4.6 Sequencing and sequence processing

The Nextera fragments were sequenced on the Illumina NextSeq, for 2-4 million reads per replicate using 2x38bp reads, or 1-2 million reads per replicate using 2x75bp reads. The raw read pairs were collapsed into unique read pairs, and unique read pairs represented by less than five raw reads were discarded. Trim_galore was used to remove the Nextera adaptor sequences and Pear was used to merge read pairs that overlapped at the 3' end. The unique reads were then aligned to the wild type lamB amplicon using Bowtie2. Samtools and pysamstats were used to count the matches and mismatches between the aligned unique reads and the template amplicon, and the counts were all given an additional pseudocount of 0.1.

2.4.7 Calculating Functional Scores

For each mutation with an average of five or more counts in the libraries from the control conditions, we calculated enrichments, and converted those to functional scores. For the λ selection, we calculated the Enrichments, E_λ as follows:

$$E_\lambda = log_2((C_{\lambda,mut}/C_{\lambda,wt})/(C_{LB,mut}/C_{LB,wt}))$$

Where C is a count of how many times a base appeared at a given position, λ indicates the count from the LB+λ selected population, LB indicates the count from the control population, mut indicates the count refers to a mutant base, and wt indicates the count of the reference base at that position.

Enrichments were converted to functional scores by linearly scaling them such that the median synonymous score was set to 1 and the median nonsense score was set to 0.

$$F_\lambda = \frac{E_\lambda - median(E_{\lambda,stop})}{median(E_{\lambda,syn}) - median(E_{\lambda,stop})}$$

Where E_λ is the enrichment calculated above, mut indicates that E refers to the individual mutation for which F_λ is being calculated, syn refers to enrichments for all mutations that are synonymous to the wild type sequence, and stop indicates that the mutation is a nonsense mutation.

Similarly, for the maltodextrin selection, we calculated:

$$E_{malt} = log_2((C_{M,mut}/C_{M,wt})/(C_{G,mut}/C_{G,wt}))$$

Where M indicates the count from the M9+maltodextrin population and G indicates the count from the M9+glucose population.

Enrichments were converted to functional scores as follows:

$$F_{malt} = \frac{E_{malt,mut} - median(E_{malt,syn})}{median(E_{malt,stop}) - median(E_{malt,syn})}$$

2.4.8 Site-directed mutagenesis for individual mutations

For 20 mutations that were called $\lambda^r Mal^+$, we generated variants containing only the single mutations. Primers containing the mutation with 20 bp of wild type sequence on either end were ordered along with reverse primers abutting the 5' end. These were used to amplify the entire p44k-lamB plasmid, and the PCR product was treated with DpnI to remove template and T4 polynucleotide kinase to phosphorylate the ends, then ligated to form a circular product. The ligation products were transformed into DH5α and the lamB coding sequence was Sanger sequenced to confirm the directed mutation and the absence of other mutations. The sequence-verified plasmids were transformed back into DH10B($lamB^{\Delta}$).

2.4.8 Growth curves for individual mutations

For each of the individual mutations, cultures were grown overnight by inoculating a single colony into 2 mL LB ampicillin and shaking at 37ºC. In the morning, the cultures were back-diluted 1:50 into LB+ampicillin+100 µM IPTG and grown for 2 hours. The cells were harvested and resuspended into an equal volume of TMS. The cells were then diluted in duplicate (10µL cells+190µL media) onto a 96-well plate, where the media was LB+0.2% maltose+10mM MgSO4+100µM IPTG+10^4 pfu/mL λ_{JL801} when measuring λ resistance, or M9+100 µM IPTG+0.4% maltodextrin+80 µg/mL leucine when measuring maltodextrin transport. The plate was incubated at 37ºC with shaking, and OD600 was measured every 10 minutes for 48 hours.

2.5 Figures and Tables

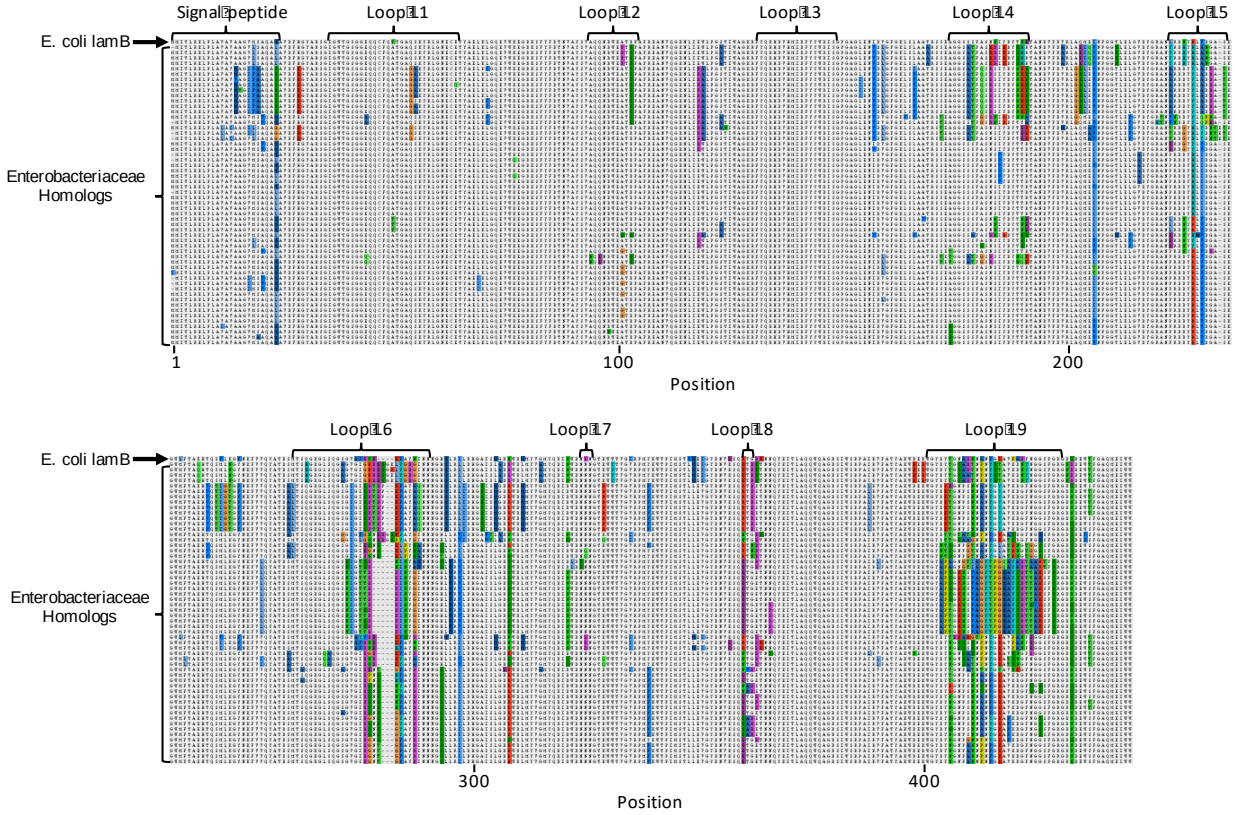

Figure 2.1: Multiple sequence alignment of LamB Homologs.
A protein alignment of homologs of LamB from the family Enterobacteriaceae, ranging from 82-97% identical to the query, is shown, with residues that deviate from the consensus residue colored by residue property. The extracellular loops and signal peptide are each annotated, showing a high rate of missense variation in some, but not all, of the extracellular loops, especially Loops L4, L5, L6, and L9. These loops are likely diversified by selection against predation, as they are easily accessible to predators on the extracellular face of the protein.

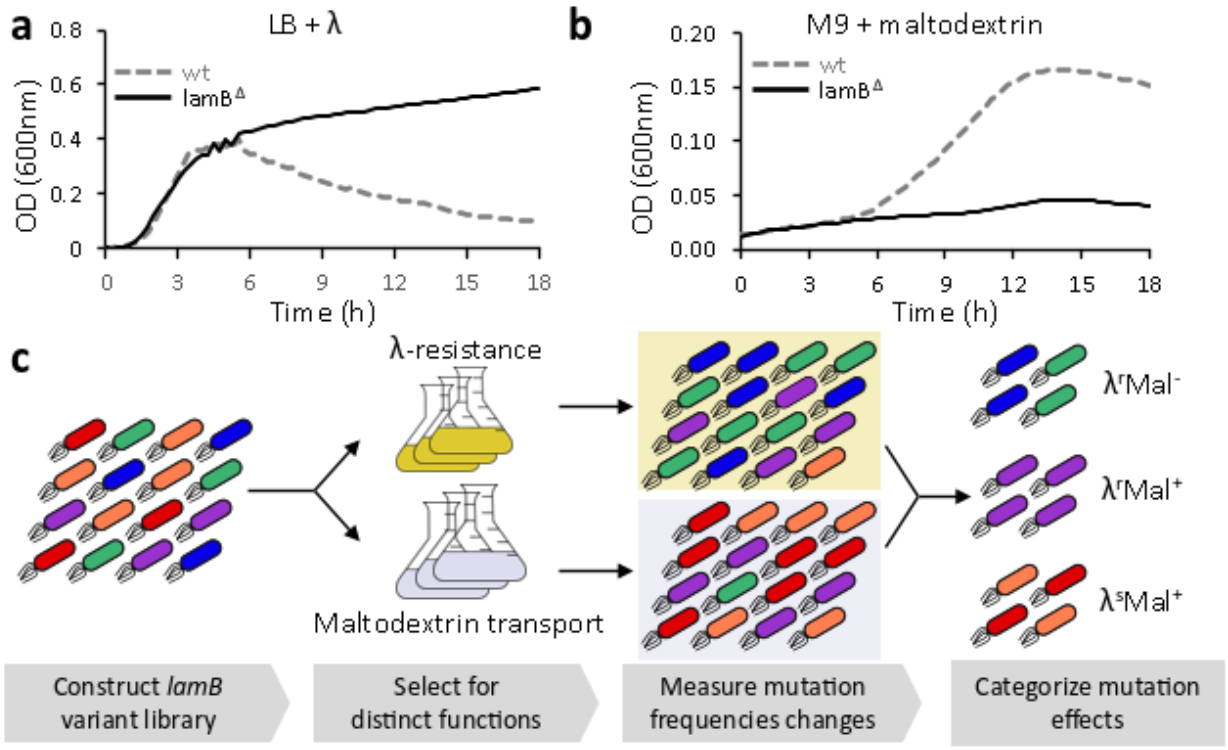

Figure 2.2: Separation of the functions of LamB via growth in different selective conditions.
(a) E. coli (strain DH10B) containing wild type lamB or a deletion of the entire lamB coding sequence were grown for 18 hours in LB+λJL801 at 37ºC, and OD600 was measured every 10 minutes. The average of three replicate wells are plotted over time. **(b)** E. coli (strain DH10B) containing wild type lamB or a deletion of the entire lamB coding sequence were grown for 18 hours in M9+maltodextrin at 37ºC, and OD600 was measured every 10 minutes. The average of three replicate wells are plotted over time. **(c)** The experimental strategy for determining effects of missense mutations involves subjecting a mutational library of lamB variants, each expressed from a plasmid in an E. coli cell, to distinct selection pressures of λ resistance or maltodextrin transport. This general strategy is sometimes called deep mutational scanning (Fowler and Fields, 2014).

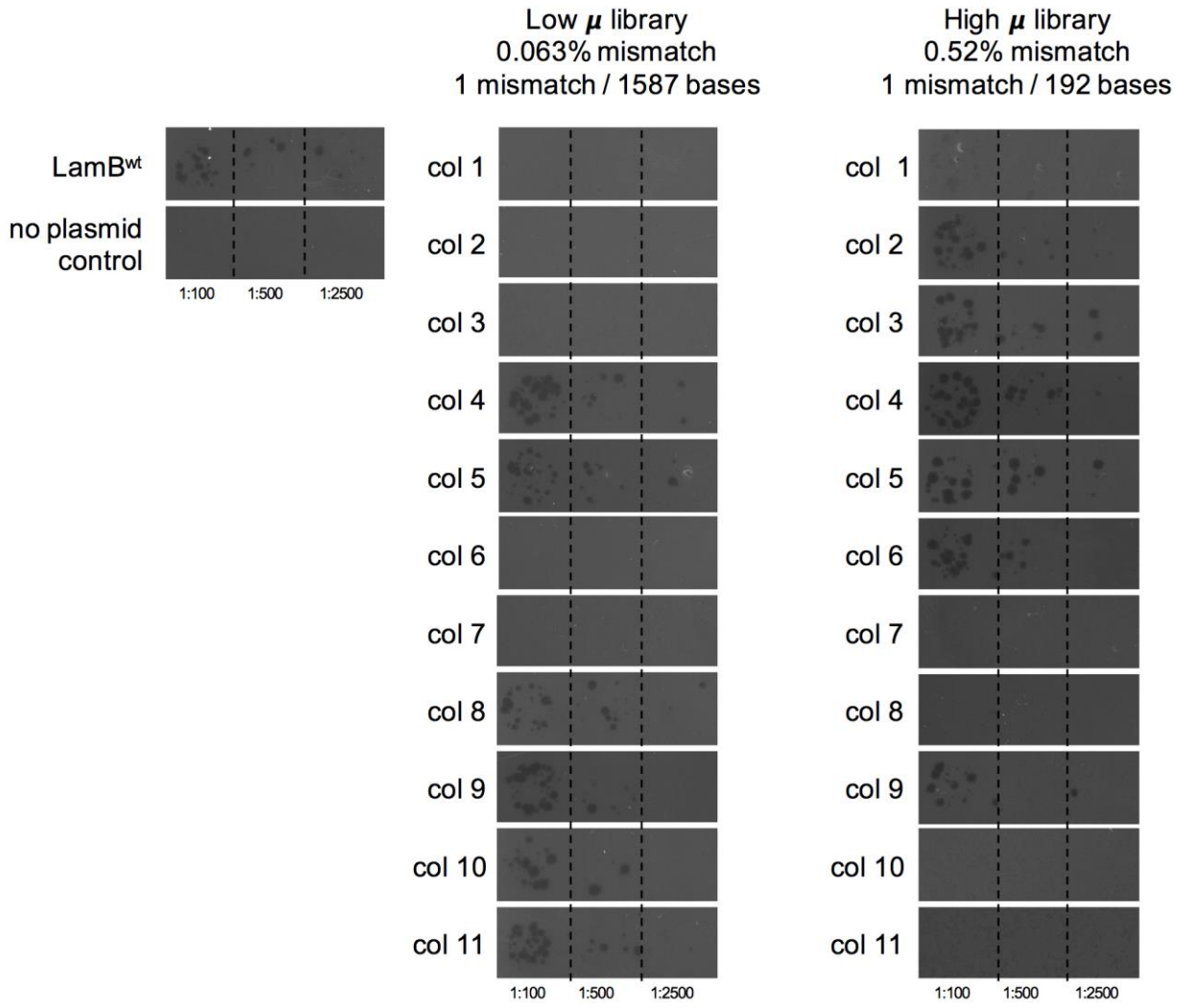

Figure 2.3: Individual phenotyping of randomly chosen variants from the error-prone PCR libraries.
Eleven variants from each library were chosen at random by plating on non-selective media and picking colonies for individual assays of λ-sensitivity or maltodextrin transport. Each colony was grown up overnight, back-diluted 1:100 into induction media containing IPTG, then plated in top agar on LB. The indicated dilutions of λ were then spotted onto each plate, using 3 µL of diluted lysate per spot, and the plates were incubated at 37°C overnight before imaging.

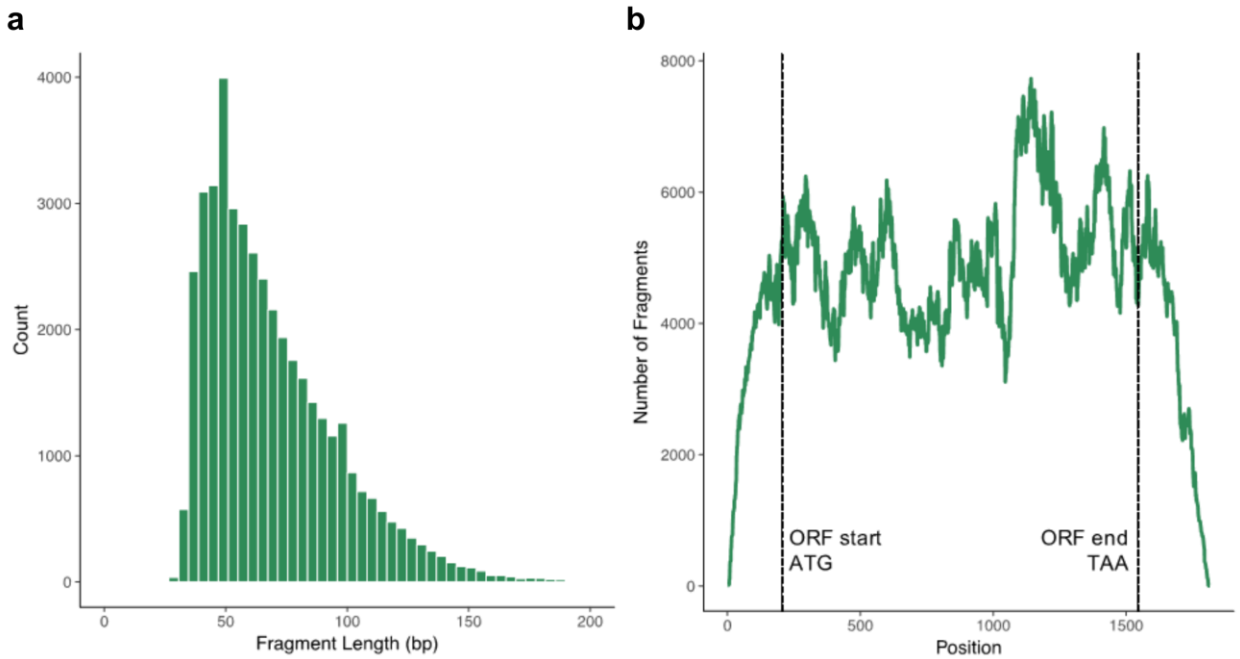

Figure 2.4: Sequencing output from a representative sample.
(a) The length distribution of Tn5-generated fragments sequenced from the first replicate of the control population for the λ selection. Lengths shown have had adapter and index sequence trimmed off. **(b)** Coverage depth, in terms of number of unique fragments, over the length of the amplicon that was sequenced. The *lamB* ORF is indicated, as well as flanking regions that were included due to expected coverage falloffs at the DNA termini.

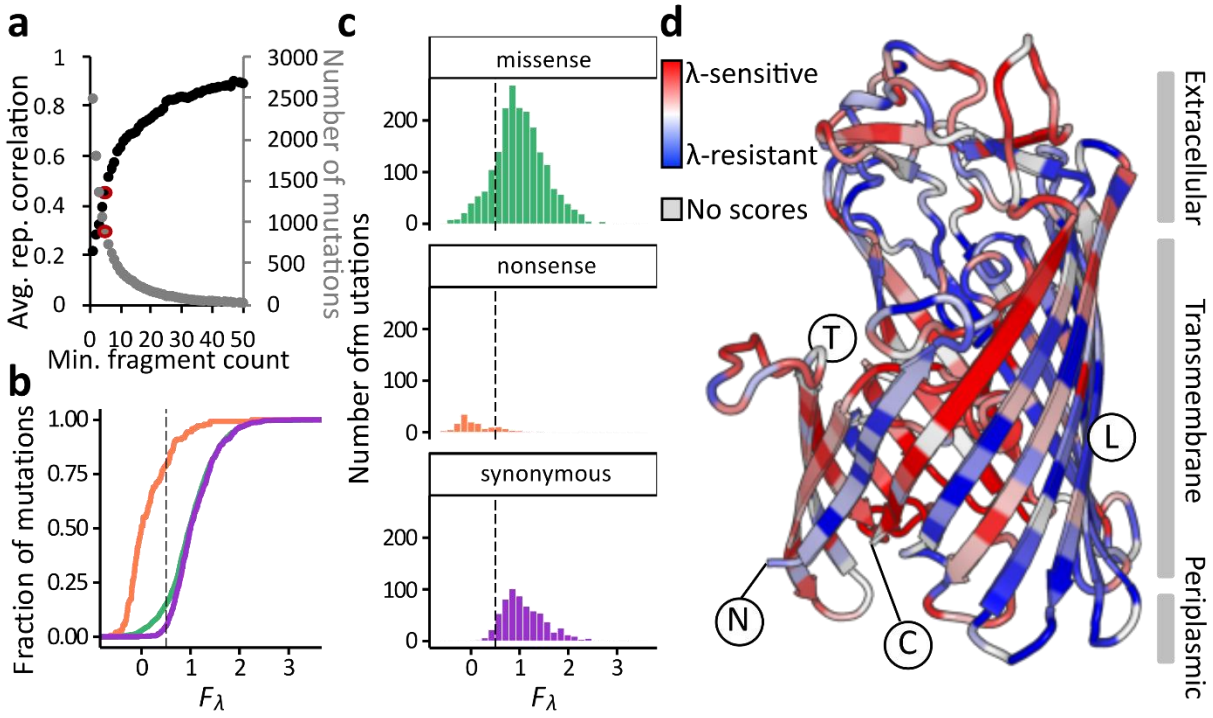

Figure 2.5: Missense mutations can drive large changes in λ infectivity.
(a) Mutations in lamB were scored in six replicates for their effects on λ sensitivity by comparing mutation frequencies before and after selection in LB+λ, and the scores were compared between replicates. The correlation between replicates is a function of the fragment input count threshold and increases as the threshold is raised. However, raising the fragment input count threshold decreases the number of mutations that can be analyzed because fewer meet the threshold. The fragment input count threshold used throughout the manuscript, five, is highlighted in red. 'rep': replicate. **(b)** The empirical Cumulative Distribution Function of F_λ scores, which estimate mutational effects on λ-sensitivity from zero (nonsense-like) to one (synonymous-like). The missense distribution mostly tracks the synonymous distribution, with a slightly longer tail to the left. Green: missense mutations. Orange: nonsense mutations. Purple: synonymous mutations. The dotted line represents the threshold, F_λ=0.5, above which mutations are initially called λ-sensitive. The placement of this line is reconsidered on the basis of individual assays for Figures 2.4 and 2.5. **(c)** The distribution of fitness effects for the λ selection, for all scored single mutations, averaged over six replicates. Colors and dotted line are as in B. **(d)** The mean F_λ score of mutations at each residue in LamB, colored relative to other residues from λ-resistant (red) to λ-sensitive (blue). Structural features described in the text are annotated. N: the amino-terminus of the mature protein. C: the carboxyl-terminus of the mature protein. T: the face of the protein that binds other monomers to form the functional trimer. L: the face of the protein exposed to the lipid bilayer.

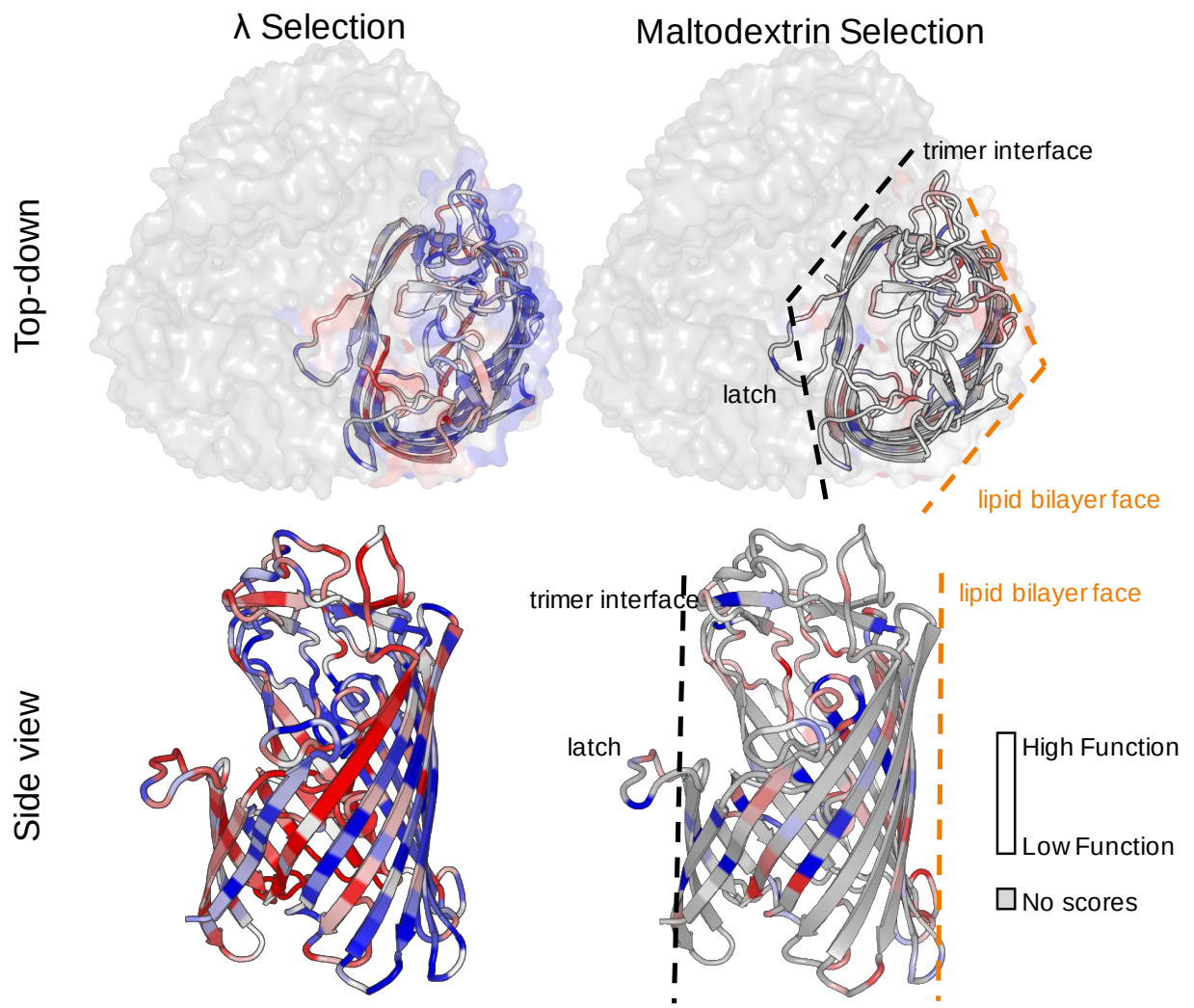

Figure 2.6: Structural components of LamB with respect to assayed phenotypes.
The structure of LamB, as determined by x-ray crystallography (Schirmer et al., 1995), is shown as viewed from outside the cell (top-down) vs. laterally from within the membrane (side view). In the top-down view, the other monomers of the homotrimer are shown in light grey. In the λ selection, residues that are against the trimer interface generally contain λ-sensitive mutations, while residues facing the lipid bilayer generally contain λ-resistant mutations.

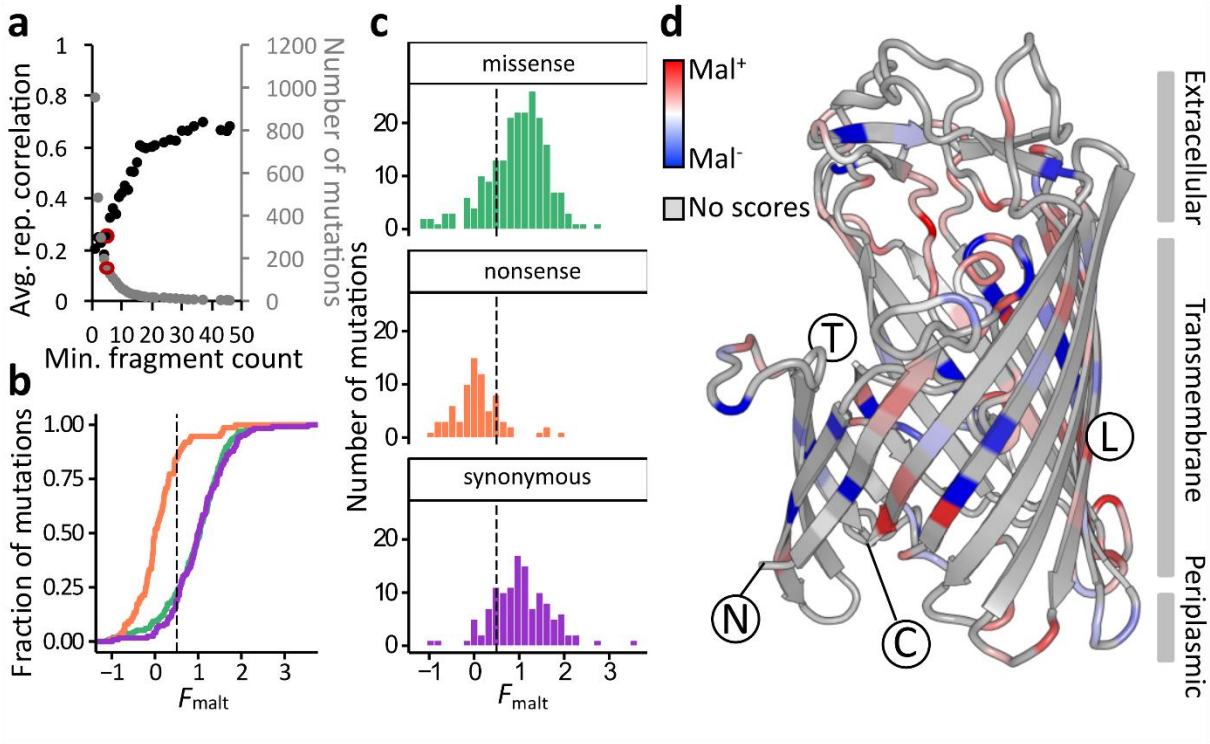

Figure 2.7: Most missense mutations are unable to strongly disrupt maltodextrin transport.

(a) Mutations in lamB were scored in six replicates for their effects on maltodextrin transport by comparing mutation frequencies before and after selection in M9+maltodextrin, and the scores were compared between replicates. The relationship between the replicate correlation and fragment input count threshold is as for Figure 2.2. The fragment input count threshold of five is highlighted in red. 'rep': replicate. **(b)** The empirical Cumulative Distribution Function of F_{malt} scores, which estimate mutational effects on maltodextrin transport from zero (nonsense-like) to one (synonymous-like). The missense distribution is not significantly different from the synonymous distribution (p=0.65, Kolmogorov-Smirnov test). Green: missense mutations. Orange: nonsense mutations. Purple: synonymous mutations. The dotted line represents the threshold, F_{malt}=0.5, above which mutations are called Mal^+. **(c)** The distribution of fitness effects for the maltodextrin selection, for all scored single nucleotide mutations, averaged over six replicates. Colors and dotted line are as in B. (D) The mean F_{malt} score of mutations at each residue in LamB, colored relative to other residues from Mal^+ (red) to Mal^- (blue). Structural features described in the text are annotated. N: the amino-terminus of the mature protein. C: the carboxyl-terminus of the mature protein. T: the face of the protein that binds other monomers to form the functional trimer. L: the face of the protein exposed to the lipid bilayer.

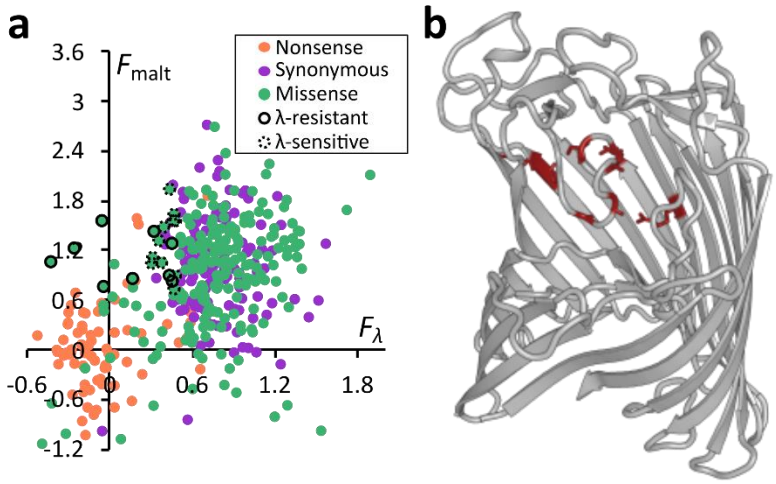

Figure 2.8: λ^{r}Mal$^+$ mutations are relatively rare and constrained to Loop L6.
(a) The functional scores, F_{malt} and F_λ, for each variant are significantly correlated between selections, with Pearson's r = 0.495, $p < 10^{-6}$. Mutations that were chosen for individual assays are outlined, with solid outlines indicating agreement between the individual and the pooled assay, while dotted outlines indicate the mutations that are at least partially sensitive to λ in isolation. Green: missense mutations. Orange: nonsense mutations. Purple: synonymous mutations. **(b)** The mutations that were determined in individual assays to be λ^{r}Mal$^+$ are shown in red on the structure of LamB. Eight of these are either contained on or directly contact Loop L6, with the ninth occurring on the signal peptide (not shown).

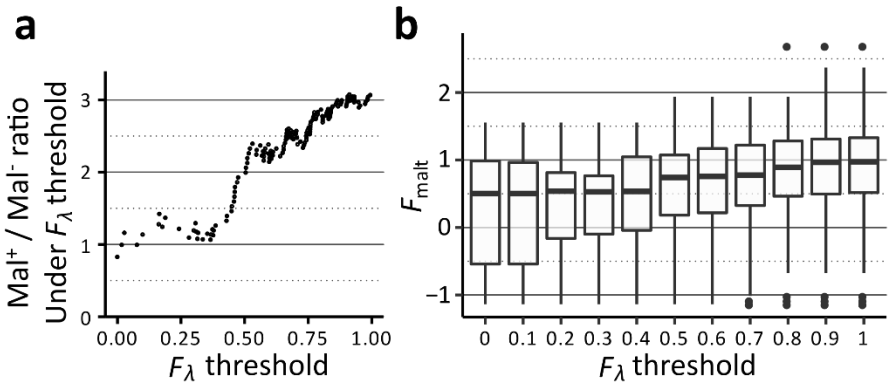

Figure 2.9: Comparison of Mal$^+$ and Mal$^-$ mutations across a range of F_λ thresholds. **(a)** For λ-resistant variants, as determined by F_λ below a defined threshold, we calculate the ratio of mutations that are Mal$^+$ or Mal$^-$. As the threshold is decreased below F_λ=0.35, this ratio approaches one, although noise increases due to the lower number of mutations under the more stringent threshold. The F_{malt} threshold used for this and all other analyses is 0.5. **(b)** The distributions of F_{malt} scores below a given threshold of F_λ are shown. The median F_{malt} decreases from ~1.0 (equivalent to a synonymous codon) with a high threshold to ~0.5 with a lower threshold at F_λ=0.4. However, the median F_{malt} score does not appreciably change if the threshold is decreased further.

Chapter 3: Balance between promiscuity and specificity in phage λ host range

As hosts acquire resistance to viruses, viruses must overcome that resistance to re-establish infectivity, or go extinct. Despite the significant hurdles associated with adapting to a resistant host, viruses are evolutionarily successful and maintain stable coevolutionary relationships with their hosts. To investigate the factors underlying how pathogens adapt to their hosts, we performed a deep mutational scan of the region of the λ tail fiber tip protein that mediates contact with the λ host, E. coli. Phages harboring amino acid substitutions were subjected to selection for infectivity on wild type E. coli, revealing a highly restrictive fitness landscape, in which most substitutions completely abrogate function. By comparing this lack of mutational tolerance to evolutionary diversity, we highlight a set of mutationally intolerant and diverse positions associated with host range expansion. Imposing selection for infectivity on three λ-resistant hosts, each harboring a different missense mutation in the λ receptor, reveals hundreds of adaptive variants in λ. We distinguish λ variants that confer promiscuity, a general ability to overcome host resistance, from those that drive host-specific infectivity. Both processes may be important in driving adaptation to a novel host.

A version of the chapter has previously been published as:

Andrews, Bryan, and Stanley Fields. "Balance between promiscuity and specificity in phage λ host range." bioRxiv (2020).

3.1 Introduction

Viruses and their hosts engage in an evolutionary battle: mutations in the virus that increase infectivity come at a cost to bacterial growth, and mutations in the host that confer resistance to viruses come at a cost to the virus. In phages, as in other viruses, much of this

battle is centered on the binding relationship between the host receptor and the viral protein that contacts it. Despite the adaptability of viruses to resistant hosts, viruses face significant evolutionary hurdles that their hosts do not (Figure 3.1). First, random mutations are much more likely to disrupt an existing binding relationship than to generate a new one. Thus, viruses must survey a much broader sequence space than their hosts do to compete in an evolutionary arms race. Second, many receptors are readily lost in the host, but are essential to the virus. Some viruses, including λ, switch receptors following loss of their canonical receptor, although it is unclear if this is a common response (Morona and Henning, 1984; Doulatov et al., 2004; Meyer et al., 2010). Third, host populations are polymorphic, such that selection for viral resistance may give rise to multiple sub-clades with distinct resistance mechanisms. Overcoming any single resistance mechanism may not be sufficient for a virus to reestablish infectivity at the population level. Given these evolutionary hurdles, there remains the need to explain why viruses are so successful over evolutionary time scales (Lenski, 1984; Lenski and Levin, 1985; Koskella and Brockhurst, 2014). In humans and other long-lived organisms, differences in relative generation times and mutation rates likely play a role (Hall et al., 2013). However, phages are also extremely successful predators despite the short generation times and rapid adaptation of their bacterial hosts, which frequently outcompete phages and drive them to extinction in co-culture (Koskella and Brockhurst, 2014; Dennehy, 2012).

Phages such as λ serve as a useful model for analyzing host-pathogen coevolution (Koskella and Brockhurst, 2014; Buckling and Rainey, 2002). Co-culture of λ with its E. coli host has resulted in the isolation of λ-resistant E. coli strains, which typically harbor null mutations of a maltose porin, LamB, that serves as the λ receptor (Lederberg, 1955; Thirion and Hofnung 1972; Gehring et al., 2987). Selection for both λ-resistance and maltose uptake has revealed LamB missense variants that disrupt λ binding (Shaw et al., 1977). λ variants can then arise that can re-establish infection on these previously resistant strains (Werts et al., 1994). When co-

cultured with E. coli not expressing LamB, λ strains have been isolated that switch to use the non-canonical receptor OmpF, a LamB paralog (Meyer et al., 2010).

Phage tail fibers, which bind to the host, evolve extremely quickly due to strong selective pressures and dedicated diversification mechanisms (Doulatov et al., 2004; Tetart et al., 1998; Paterson et al., 2010). The λ tail fiber J protein consists of 1,132 amino acids, with high conservation across its N-terminal ~980 amino acids and extreme diversification across its ~150 amino acid C-terminal domain, which contacts the receptor (Werts et al., 1994). As examples, the J proteins from two recently isolated phages, lambda_2H10 and lambda_2G7b, are each >97% identical to λ across the N-terminal 982 amino acids, but only 40-60% identical across the C-terminal ~150 residues (Mathieu et al., 2020). The isolates are similarly diverged from each other and have unique host ranges. Although the structure of J or homologous tail fiber proteins has not been solved, tail fibers such as that of T4 form highly intertwined trimers (Bartual et al., 2010; Hyman and van Raaij, 2018). The trimer is composed almost entirely of β-sheets arranged in helical barrels, with the distal tip forming a more globular structure that directly contacts the receptor. Consistent with this structure, computational analysis of the secondary structure of J predicts β-sheets across the C-terminal domain, with a single α-helix ~65 residues from the C-terminus. In T4, the tail fiber turns back on itself to form a six-sheeted, rather than three-sheeted, barrel, suggesting that the α-helix may be part of the distal tip of J.

We decided to interrogate λ host range by generating a library of thousands of genetic variants in the C-terminal domain of J. We first imposed selection for infectivity on wild type (wt) E. coli, comparing the patterns of infectivity with the evolution and possible structure of the J protein. We then imposed selection on the same library for infectivity on three λ-resistant hosts, uncovering shared and unique paths to adaptation. By comparing variants within and between conditions, we consider the likely routes by which λ may overcome host resistance, and we ascribe distinct roles for variants that increase promiscuity from those that drive changes to specificity.

3.2 Results

3.2.1 Mutational scanning strategy yields infectivity measurements for thousands of J variants

We performed a 'deep mutational scan' (Fowler and Fields 2014) of phage infectivity in λ by generating a library of phage variants with single amino acid changes in the tail fiber and imposing selection for infectivity on that library (Figure 3.2a). The library consisted of nearly all single amino acid substitutions and several thousand double amino acid substitutions across the C-terminal 150 amino acids of J (Figure 3.2b,c). For readability, we numbered the positions starting at the first residue of the mutagenized region. We imposed selection on this library by mixing the phage variants with wild type *E. coli* and isolating intracellular phages after a single round of growth. This selection was repeated for four infectious cycles, and we estimated the abundance of each barcode in the expanding population over time using a novel approach we call 'Model-Bounded Scoring' that directly estimates the progeny produced per generation. Barcode abundances between replicates showed strong agreement, as did different barcodes representing the same variant (Figure 3.2d, Figure 3.3). We used the distribution of nonsense variants and synonymous variants to categorize variant effects. We considered variants with growth rate within two standard deviations of the mean nonsense variant to be 'null-like' and those within two standard deviations of the mean synonymous variant to be 'wt-like' (Figure 3.2e). Variants falling between these two distributions were considered 'deleterious,' while variants above the synonymous distribution were considered 'hyper-infective.'

A large fraction (69%) of single missense variants conferred a null-like phenotype (Figure 3.4a). Although phage tails rapidly diversify to expand their host range, the fitness landscape of J with respect to its wild type host is much more restrictive than has been seen for other, much more conserved, proteins (Roscoe and Bolon, 2014; Mishra et al., 2016; Bandaru et al., 2017; Weile et al., 2017). The pattern of mutational tolerance in J is broadly consistent with a T4-like

barrel structure: extreme intolerance to substitutions to glycine and proline and three regions of high periodicity in mutational tolerance, both of which suggest β-sheet richness (Figure 3.4a). We can use this pattern of mutational tolerance to construct a coarse model of J's structure (Figure 3.5), with the receptor-binding portion of the protein encompassing positions ~50-100. This region of the protein is not, however, enriched for hyper-infective variants. We identified 12 such hyper-infective variants, two of which are accessible through point mutations. That these hyper-infective mutations are found throughout the mutagenized region suggests that residues that do not directly contact the receptor can nevertheless strongly influence receptor binding.

3.2.2 J fitness landscape is more restrictive than predicted from evolutionary diversity

The extreme intolerance to mutation of J relative to other proteins conflicts with the observation that phage tail fibers, including J homologs, rapidly diversify over evolutionary time. We therefore decided to directly compare the patterns of mutational tolerance from our assay against patterns of mutational tolerance from an alignment of 910 orthologs of J. Positions that harbored hyper-infective variants do not correspond to positions of high amino acid diversity across J orthologs (Figure 3.4b). More broadly, the average progeny of variants at a position, a proxy of mutational tolerance, correlated only weakly ($r^2 = 0.06$) with the evolutionary diversity of those positions (Figure 3.4c) (Meyer et al., 2010; Werts et al., 1994). This observation contrasts with deep mutational scans of conserved cellular proteins, for which much stronger associations are observed between diversity and mutational tolerance (Roscoe and Bolon, 2014; Mishra et al., 2016; Boucher et al., 2014). Additionally, the distribution of variant effects argues that J is not trapped on a fitness peak with respect to rapidly binding to its wild type host. Point mutations were slightly less likely to be strongly deleterious than all possible amino acid substitutions, even when synonymous mutations are excluded. Moreover, the best-scoring point mutation, I15F, was 4.2 standard deviations above the distribution of synonymous variants, producing 35% more progeny per generation than the wild type. Thus, the selection pressures that have shaped the evolutionary

history and the wild type sequence of λ may have acted on a different property than is being selected for in our assays.

In J, peaks of diversity (Figure 3.4b) that corresponded to many wild type-like variants (*e.g.* positions 67 or 108) can be separated from those that corresponded to many deleterious or null-like variants (*e.g.* positions 30 and 97). Furthermore, λ host range mutations tended to fall at these diverse, mutationally intolerant positions (gold points, Figure 3.4c). We posit that by imposing selection for infectivity on a single host, we measured a selective pressure that is too narrow to adequately reflect λ's entire evolutionary history on multiple hosts. Positions important for mediating infection of a single host would be mutationally intolerant in our assay, as the amino acid optimal for binding wild type LamB would likely have already been fixed in λ. However, these same positions could be under diversifying selection over evolutionary time during which the host has varied. Therefore, a complete understanding of λ evolution requires comparing the effects of J variants across many hosts.

3.2.3 Adaptation to a set of resistant hosts

To investigate mechanisms of adaptation to a resistant host, we challenged the library of J variants with a set of *E. coli* hosts bearing novel λ resistance mutations. A deep mutational scan of LamB, the λ receptor, showed that many point mutations specific for λ resistance are in or proximal to loop L6, the presumptive λ binding site (Andrews and Fields, 2020). We chose three receptor mutations, LamB-R219H, LamB-T264I, and LamB-G267D, that confer λ resistance but allow maltodextrin transport, on the basis that these mutations specifically disrupt the λ binding site rather than disrupting stability of the receptor. In a *lamB*-deletion background, we generated a *lac*-inducible expression vector for the wild type and the three *lamB* alleles, which were expressed using 0.1 mM IPTG. We imposed selection for infectivity on the λ library on each of the four *E. coli* strains, and we estimated progeny using Model-Bounded Scoring.

Similar to selection on wild type *E. coli*, synonymous J variants were more infective than nonsense variants on each of the three resistant *E. coli* strains (Figure 3.6a). Because Model-

Bounded Scoring estimates real growth rate, rather than relative fitness, we can directly compare the growth of λ bearing the same J variants on different hosts without normalizing to a reference allele. λ was slightly more infective on the IPTG-expressed wild type receptor than when this receptor was expressed from the endogenous *lamB* locus (31 progeny vs. 27 progeny for wild type J), but infectivity of the variants was highly correlated between the selections (r^2 = 0.958, Figure 3.6c). The resistant *lamB* alleles did not confer absolute resistance, but decreased the progeny produced by wild type λ from 31 (LamB-wt) to 22 (LamB-G267D), 2.7 (LamB-T264I), and 2.0 (LamB-R219H). For all three *lamB* mutants, we could identify single missense variants in J that restored progeny to 40-100% of what was produced on the wild type host (Figure 3.6a). Moreover, a much larger fraction of variants significantly outperformed the synonymous distribution on each resistant host. Adaptive variants in each case were broadly distributed over the sequence, not highlighting a domain or structural feature uniquely necessary for adaptation (Figure 3.6b). With some exceptions, adaptive variants were neutral-to-beneficial at infecting LamB-wt (Figure 3.6c). However, the reverse was not true; variants that were highly infective on the wild type receptor were frequently less infective on the novel hosts.

3.2.4 Specific and general mechanisms of adaptation

A comparison of the infectivity of J variants between two hosts shows that the most infective J variant with respect to one host was frequently highly infective on the other host as well (Figure 3.6c, 3.7). In cases for which a J variant conferred infectivity on one host but not the other, it nearly always conferred infectivity on the host that is less resistant to wild type λ. Reciprocally, for hosts with greater resistance to wild type λ, a smaller fraction of J variants conferred infectivity (*i.e.*, were not null-like). This patten is consistent with a 'nested' model in which each host and pathogen has a set amount of resistance or counter-resistance that is not dependent on the other player (Figure 3.8a). In this model, infectivity is determined by the relative strength of resistance and counter-resistance. This model would contrast with a 'lock-key' model (Weitz et al., 2005), in which infectivity is determined by how a well a given λ variant matches a

given receptor (Figure 3.8b). The relevant distinction is that under a nested model, but not a lock-key model, a J variant may gain generic counter-resistance that improves its infectivity on many hosts.

The property of generic counter-resistance can be compared to the property of enzyme 'promiscuity,' in which enzymes weakly catalyze non-canonical reactions in addition to their normal biological activities (Arnold, 2009). Under the right selection pressure, mutations can expand or improve these promiscuous activities, and evolution may eventually lead to specialization in the new activities (Arnold, 2009). By analogy, J might be said to have a baseline promiscuity in that it can weakly bind non-LamB-wt receptors, and that variants have increased promiscuity if they improve infectivity broadly over the space of potential hosts, rather than only on LamB-wt (Figure 3.8c). To estimate the promiscuity of J variants, we plotted in log-space the progeny conferred by a variant on each host, compared to the susceptibility of these hosts to wild type λ. For synonymous variants (Figure 3.8c, black line), this relationship is 1:1 by definition. However, some variants, like A84M, have a shallower slope indicating that they were less affected by the resistance of receptor mutations; we call variants with significantly greater area under the curve than synonymous variants 'promiscuous.' Most promiscuous variants grew well on the wild type host (Figure 3.9), implying that these promiscuous variants could persist in a λ population prior to encountering a resistant host. However, that these variants have not already been fixed in the population may imply subtle costs to promiscuity, such as thermodynamic instability (Arnold, 2009; Bloom 2006).

Additionally, 49 single missense variants conferred growth on only a single host, most of which conferred growth on LamB-G267D (such as Q96E in Figure 3.8d). With a few exceptions, like K141C (Figure 3.8d), these host-specific variants did not confer greater infectivity than wild type J on any host we tested; they were merely dramatically more infective on a particular host compared to the other hosts. Thus, while these host-specific variants may be an important part of adapting to a host, they are generally insufficient to directly overcome resistance. Rather, we posit

that the acquisition of host-specific variants may follow after the acquisition of promiscuous variants and either ameliorate thermodynamic costs or prevent off-target binding. This process could explain why very few LamB-wt-specific variants were observed, and why they were less infective than variants specific to other hosts (Figure 3.10), as if the wild type J sequence has already been selected for near-maximal specificity to its host.

Based on our analysis of J variants growing on a wild type host, positions with low mutational tolerance but high diversity over evolutionary time are predicted to be enriched within variants that drive adaptation to a novel receptor (Figure 3.4c). Contrary to expectations, promiscuous variants tended to occur at positions with higher-than-average progeny across variants (Figure 3.8e). By contrast, host-specific variants tended to occur at positions that are intolerant to mutation and diversified over evolutionary time (Figure 3.8f). Thus, while our hypothesis that these mutationally intolerant, evolutionarily diverse positions are driving host-specificity is largely supported, host-specificity is not equivalent to overcoming resistance, which can happen through host-nonspecific mechanisms (*i.e.*, promiscuity). This distinction implies that host range mutations arising in experimental evolution studies have not generally been promiscuous variants, as these host range mutations have mostly fallen at positions both mutationally intolerant and diverse (Figure 3.4c). In our analysis, many of these host range mutations were null-like on all the hosts tested, including LamB-wt (Table 3.1), suggesting that the effects of these variants may be specific to the hosts used in those studies and/or co-occurring mutations in J.

3.2.5 Positive epistasis potentiates adaptation to a new host

In addition to single missense variants, the J library contained approximately 7,500 variants with two missense mutations, sparsely surveying the space of >4 million possible double missense variants. We wondered whether combinations of promiscuous and host-specific mutations could help mediate adaptation to a novel receptor beyond what either class of mutations would confer in isolation. For example, A94S is a promiscuous variant, but was mildly

impaired for infectivity on LamB-G267D, and S30W is a LamB-G267D-specific variant that was mildly impaired on LamB-G267D but null-like on all others (Figure 3.11a). However, the double missense variant A94S+S30W had both high growth and moderate specificity on LamB-G267D. This double missense variant therefore exhibits positive epistasis on one host, though it is poorly infective on the other hosts.

On LamB-wt, only 6.2% of double missense variants were infective (*i.e.*, not null-like), compared to 26% of single missense variants that were infective. For each double missense variant, we calculated the expected progeny given the progeny from each single variant, using a simple multiplicative model. Most variants yielded similar values for measured and expected progeny, but ~45% of infective variants (2.7% of total variants) exhibited significant epistasis (Figure 3.11b), including 22 infective variants for which both mutations were null-like on their own. These variants with dramatic reciprocal sign epistasis were enriched for positions with host range mutations (Meyer et al., 2010; Werts et al., 1994), which appeared 20 times among these 22 variants. As a control, we also considered variants in which a single missense mutation was paired with a synonymous mutation. Such variants exhibited better agreement between expected and empirical scores (r^2 = 0.92) than double missense variants (r^2 = 0.64), suggesting that the prevalence of epistasis in double missense variants is not an artifact of the fact that most double missense variants were less abundant in the library than single missense variants.

Similar to the selection on the wild type host, strong positive epistasis was prevalent on each resistant host. Single missense variants that conferred promiscuity frequently interacted epistatically with other variants (Figure 3.11c). Of double missense variants that contained a promiscuous single missense variant, 73/550 (13.3%) exhibited significant epistasis compared to only 2.2% of all double missense variants. The rare cases of double missense variants consisting of one promiscuous variant and one host-specific variant were even more likely to exhibit epistasis (5/21, 23.8%). In these five examples, epistasis was always in the positive direction.

Additionally, the effects of positive epistasis became more salient when λ was challenged with a resistant host. Double missense variants that were infective on LamB-wt had varying levels of promiscuity and positive epistasis (Figure 3.11d). By contrast, double missense variants that were infective on the most resistant host, LamB-R219H, nearly exclusively had both high promiscuity and positive epistasis. The other two hosts revealed intermediate effects.

3.3 Discussion

Although phages and bacteria are often presumed to coevolve stably and indefinitely, an "asymmetry in coevolutionary potential of these hosts and parasites" exists (Lenski and Levin, 1985). To investigate how λ overcomes host resistance, we analyzed thousands of variants of its tail fiber protein, J, on a small set of resistant *E. coli* hosts. We find that promiscuous J variants, which increase infectivity on a broad range of hosts, underlie the re-establishment of infection on resistant hosts. These variants co-exist with other, host-specific, variants that generally do not increase infectivity on any host but have smaller losses to infectivity on a single host. We posit that both types of variants are important for a phage to adapt to a new host, with host-specific variants likely ameliorating costs associated with promiscuity. This framing has implications for experimental evolution studies, protein adaptation more broadly, and natural phage-bacteria communities.

When phages and bacteria cyclically develop resistance and counter-resistance in experimental evolution studies, this coevolution is frequently characterized by an initial escalation of both host resistance and phage counter-resistance. This escalation eventually reaches an asymptote and is followed by negative frequency-dependent selection ("Kill the winner" dynamics) in which the dominant phages are most infective on the most common hosts (Koskella and Brockhurst, 2014; Hall et al., 2013). This pattern is well explained by a model in which promiscuous variants drive broadened host range but come at a cumulative cost, manifesting as lower growth rate relative to their host-specific counterparts (Poullain et al., 2007). The sequential

acquisition of promiscuous variants could also open up pathways for a phage to infect a highly resistant host by first adapting to a less resistant host in the same environment. For example, Werts *et al.* (1994) could not directly isolate λ that overcame the resistance allele LamB-G151D, but by pre-adapting the phage to other hosts with weaker resistance, they found double mutants able to grow on LamB-G151D. However, thermodynamic costs associated with promiscuity may also constrain paths to adaptation. In their isolation of LamB-independent λ strains that use OmpF as receptor, Meyer *et al.* (2012) repeatedly identified the same set of 4-5 adaptive variants across dozens of independent cultures, suggesting considerable constraint on the path to counter-resistance (Franke et al., 2011). In a follow-up study, the bi-specific intermediate, which binds both LamB and OmpF, was less stable than the LamB-specific parent, and selection for OmpF specificity was sufficient to restore stability (Petrie et al., 2018).

In the context of enzymes, promiscuous activities provide convenient starting points for adaptation to novel protein functions (Arnold, 2009). Increasing the stability of the parent enzyme can potentiate greater adaptation by compensating for the mild destabilization associated with some adaptive variants (Bloom et al., 2006; Tokuriki and Tawfik, 2009a). Similarly, the low level of infectivity mediated by wild type J on resistant hosts serves as a starting point for adaptation to those hosts. Second mutations can compensate for the initial mutations that confer promiscuity, offsetting the potential costs of these initial mutations and resulting in highly infective and/or promiscuous double variants. This effect can be seen in the strong association between positive epistasis and promiscuity, whereby promiscuous single variants were more likely to have positive epistatic interactions than non-promiscuous variants (Figure 3.11c) and rare promiscuous variants that drive adaptation to LamB-R219H were always positively epistatic (Figure 3.11d). At a mechanistic level, promiscuous J variants may shift between multiple semi-stable protein conformations, or they may heterogeneously fold into one of multiple stable conformations. The latter mechanism was found to underlie the LamB-OmpF bi-specific intermediates characterized by Petrie *et al.* (2018). Under a model requiring multiple protein conformations, destabilization

may be fundamentally linked to promiscuity rather than incidental to it. Therefore, counter to observations with promiscuous enzymes (Bloom et al., 2006; Tokuriki and Tawfik, 2009a), stabilizing mutations in phages are unlikely to precede adaptive ones. Instead, a destabilizing mutation that increases promiscuity must come first, followed by a compensatory mutation that re-stabilizes the protein into an optimal conformation for infection of the most abundant host. This prediction would also explain why J is so broadly intolerant of mutation: repeated evolutionary transitions between stable and unstable sequences leave J close to a threshold of severe destabilization, compared to proteins whose evolutionary histories are dominated by stable sequences (Gong et al., 2013; Tokuriki and Tawfik, 2009b).

Naturally occurring phage–bacteria interactions also show patterns consistent with a balance between host-specific and promiscuous variation. Phages and bacteria isolated from the same environment form infection networks exhibiting both nestedness and 'modularity,' a property of lock-key models, with nestedness dominating at small scales involving highly related strains (Flores et al., 2011; Weitz et al., 2013). This pattern is consistent with promiscuity driving counter-resistance to a newly resistant host, and modularity arising between more diverged hosts. The extent to which the evolution of phages involves adaptation to diverged hosts remains unclear. Orthologs of both J and LamB are broadly distributed among enterobacteria, and even appear in distant ε-proteobacteria species, suggesting either an ancient origin of this host–pathogen relationship or frequent cross-taxa jumps in host range. However, the limited breadth of hosts in which nestedness is observed suggests practical constraints to the promiscuity of a phage tail: a single promiscuous phage variant is more likely to be infective on multiple receptor variants within a single host species than on receptors of multiple related host species. In our hands, an *E. coli* host expressing a LamB ortholog from another enterobacteria (*C. freundii*, *Y. pestis*, or *S. marcescens*) did not support growth of our library, which mostly contained single missense mutations, suggesting that multiple mutations may be required for jumps between species.

We conclude that although λ faces significant evolutionary hurdles not faced by its host, it can establish common paths to adaptation on multiple potential hosts by maintaining a balance between promiscuity and host-specificity. This balance may be mediated by mild destabilization of the protein, allowing it to sample multiple conformations, although further work is needed to directly test this hypothesis. Although we surveyed adaptation to only a small set of resistant hosts, this general framework is consistent with prior observations of how λ evolves to switch to a novel receptor. This framework may apply broadly to other phages and viruses for which mutations are more difficult to assay *en masse* (Hall et al., 2013; Carat et al., 2007; Liao et al., 2013).

3.4 Materials and Methods

3.4.1 Library generation

The library of λ variants was generated by first digesting the λ genome with PasI, then cloning the small fragment that contains all of J plus some flanking DNA, onto the high-copy *E. coli* vector p44K. This vector contains three BbvCI sites, with two in the forward and one in the reverse orientation. Since all BbvCI sites must be in the same orientation for Nicking Mutagenesis, the reverse site was altered at two positions using site-directed mutagenesis. This BbvCI site was in the J coding sequence, and both edits were synonymous changes. The entire J coding sequence was subsequently confirmed by Sanger sequencing. We performed Nicking mutagenesis as described by Wrenbeck *et al.* (2016), using 150 primers, each containing a random (NNN) codon. We performed one round of mutagenesis, in three pools of 50 primers each, and Sanger sequenced colonies to check the level of mutagenesis. There was a high carryover of the wild type sequence (~30%), so we pooled the libraries and performed an additional round of mutagenesis. After the second round, we determined that the library contained mostly single NNN replacements, with some double mutations and ~12% wild type. This library was used for all experiments.

To barcode the library, we amplified the Kan^r gene from pET-9a using primers that added BamHI sites at each end, a 15-base barcode at the 5' end, and homology to the vector on both ends. We digested the p44k-borne library with HinDIII and used Gibson assembly (Gibson et al., 2009) to insert the Kan^r amplicon, selecting on 50 mg/mL kanamycin to remove the parent plasmid. The kanamycin cassette was removed by digesting the plasmid with BamHI, cutting the fully cut product out of an agarose gel, and performing a self-ligation with T4 DNA ligase. The ligation product did not produce any Kan^r colonies.

To computationally link barcodes and variants, we amplified from immediately upstream of the mutagenized region to immediately downstream of the barcode and sequenced the amplicon on the Illumina MiSeq platform. Consensus reads of at least five barcodes were constructed, conditional on >80% agreement between reads, and assigned to barcodes. These reads were also used to calculate input frequencies for each barcode for the purpose of determining input read thresholding and expected barcode counts. We measured input frequency (infectious cycle = 0, the phage population prior to selection) separately for each replicate by amplifying barcodes from the input phage population, and these replicate-specific frequencies were used for determining scores for variants.

Prior to each selection, a midiprep of the plasmid borne library was digested with PasI and ligated with PasI-digested de-phosphorylated λ genome. The ligation product was packaged into λ using the MaxPlax λ packaging extract. The variants prior to the first infectious cycle, while containing the correct DNA variants, therefore have a wild type J protein. The wild type J protein is replaced with the variant protein at the first infectious cycle.

3.4.2 Strains and plasmids

Selections were performed in the background of DH10B. For the first selection, the DH10B cells were grown from ElectroMax cells (Thermo-Fisher). For the second set of selections, we generated DH10B-lamB$^{\Delta}$, which harbors an in-frame scarless deletion of the lamB gene. Briefly, pSIM5 was transformed into DH10B and the transformed strain was heat shocked at 42^{O}C to

induce expression of the λ-red recombination system, then made electrocompetent. A linear cassette containing tetA and sacB flanked by homology to the lamB locus was then transformed in, and cells were selected on tetracycline. pSIM5, which was lost during heat shock, was retransformed, and the cells were heat shocked and made electrocompetent. Finally, an oligo flanking both sides of the lamB locus was transformed in and cells were selected on sucrose + fusaric acid, which selects against the tetA-sacB cassette. Clean deletion of the lamB gene was confirmed by Sanger sequencing and pSIM5 was cured by growing cells at 37^{O}C overnight and choosing a chloramphenicol-sensitive colony by replicate streaking on LB and LB+chloramphenicol.

For the second set of selections, we expressed lamB variants from the high-copy expression vector p44k. We used PCR to amplify lamB from the DH10B genome and cloned it into p44K such that the gene is *lac*-inducible, producing p44K-lamB. We performed site-directed mutagenesis to separately introduce the alleles g659a(R219H), c794t(T264I), and g803a(G267D), and the mutations were confirmed by Sanger sequencing the entire lamB coding sequence. We transformed each lamB variant into DH10B-lamB$^{\Delta}$.

We used λ DNA from NEB, which contains the cI857 and Sam7 alleles. cI857 confers temperature sensitive lysogeny and is obligately lytic at 37^{O}C, at which experiments were performed. Sam7 confers amber-suppressible lysis. In DH10B, λ_{Sam7} produces intracellular progeny but does not lyse cells following infection. However, progeny can be released if cells are lysed exogenously. We used the amber-suppressor strain LE392MP (Lucigen) for measuring phage stocks.

3.4.3 Media and expression

Prior to each selection, we prepared cells for infection. Cells were grown overnight at 37^{O}C and 250 rpm in LB + 10 mM $MgSO_4$ + 0.2% maltose for DH10B or DH10B-lamB$^{\Delta}$, or the same media with 100 mg/mL ampicillin and 0.1 mM IPTG for DH10B-lamB$^{\Delta}$(p44K-lamB). Cells were back-diluted 1:100 into the same media in the morning, grown for 4 hours, pelleted and resuspended

into 10 mM $MgSO_4$ to a volume such that OD_{600nm} = 0.5. Cells were stored this way at 4^OC for up to 72 hours and plated at a 10^{-6} dilution on LB immediately prior to selection to determine viability.

3.4.4 Selection conditions

For selections, 1 mL of cells were pelleted and resuspended in 1 mL LB + 10mM $MgSO_4$ + 0.2% maltose. As a negative control, we included a tube containing DH10B-lamB$^{\Delta}$, which produce very few (though non-zero) phage progeny. The phage library was added, using >10^6 pfu (plaque-forming units), and the culture was shaken at 37^OC and 600 rpm for the prescribed binding time (10 minutes for the first selection or 60 minutes for the second set of selections). Cells were pelleted, washed twice in 1 mL LB, then grown at 37^OC and 600 rpm for 90 minutes.

Following this growth, cells were manually lysed as follows. Cells were pelleted and resuspended in 100 mL STE (100 mM NaCl + 10 mM Tris + 0.5 mM EDTA). 0.1 µL ReadyLyse lysozyme (Lucigen) was added and cells were incubated for 10 minutes at room temperature. 250 mg of fine glass beads were added and the mixture was shaken in a bead beater for 30 seconds at 4^OC. Lysate was recovered by adding 500 mL TMS (100 mM NaCl + 10 mM Tris + 10 mM $MgSO_4$), vortexing, letting the beads settle, and pipetting off the supernatant. The lysate was cleared by centrifugation (12,700 rpm x 2 minutes), and the cleared lysate was driven through a 0.2 mm filter to remove any remaining cells.

3.4.5 Sequencing library preparation

At each replicate and for each timepoint, 5 mL lysate was used as the template for SYBR-based qPCR using primers that add sequencing adapters and custom indices. Because of the high magnesium content of the lysate, we used Phusion with the 5x GC buffer, and added DMSO to 3% final concentration. The barcodes were amplified and monitored manually, with tubes removed as they entered exponential phase. All tubes were removed by cycle 15. DNA was cleaned up from each reaction on a Zymo clean & concentrator column, the product concentration was determined by Qubit, and libraries were mixed at equimolar ratio and prepared for Illumina NextSeq sequencing.

3.4.6 Sequencing and barcode counting

Amplicons containing the indexed barcodes were deeply sequenced by the Illumina NextSeq, yielding at least 10x average coverage of the library for each replicate and timepoint. The sequences were trimmed down to just the barcode, and they were counted using Enrich2 with no filters. The counts files were then used for Model-Bounded Scoring.

3.4.7 Model-Bounded Scoring

We developed a custom scoring pipeline, which we call “Model-Bounded Scoring.” The name refers to the fact that each barcode is initially modeled as a zero-fitness variant, with the model placing an effective lower bound on the score a variant can receive. For variants that have non-zero counts at later timepoints, the model does not substantially alter the score.

For each barcode, we want to estimate the growth rate, r_v. Typically, we would attempt to estimate the number of viruses containing that barcode at each timepoint, $n_{v,t}$, compared to the number of viruses containing that barcode at t=0, $n_{v,0}$. We could then solve for r_v using the equation:

$$e^{r_v * t} = \frac{n_{v,t}}{n_{v,0}}$$

Assuming each barcode has a minimum growth rate of zero, $n_{v,t}$ should always be positive, and this equation will hold. However, when $n_{v,t}$ is estimated from counting barcodes, it is common for many barcodes to have zero counts, which will generally cause $n_{v,t}$ to be estimated as zero. Since log(0) is undefined, this estimate makes it impossible to come up for a useful estimate of r_v. A common way to address this problem is to add a constant positive value, termed a ‘pseudocount’ to each barcode count, therefore causing $n_{v,t}$ to always be positive. However, this can have distortionary effects on the scores.

Instead of trying to solve equation 1 as written, we assign each barcode a modelled growth rate, r^{Ψ}. We assign $r^{\Psi} = 0$ for all barcodes, but it is not strictly necessary to do so. We then try to solve the equation:

$$e^{(r_v + r^{\Psi}) * t} = \frac{n_{v,t} + n^{\Psi}_{v,t}}{n_{v,0} + n^{\Psi}_{v,0}}$$

Since we assign $r^{\Psi} = 0$, the left half of this equation is equivalent to the left half of the previous equation. Values for n can be estimated from sequencing counts as follows:

$$n_{v,t} = \frac{c_{v,t} * N_t}{C_t}$$

where $c_{v,t}$ is the counts of the barcode, v, at a given timepoint, t; C_t is the total counts of all barcodes at that timepoint; and N_t is the total population size at that timepoint. For the modelled counts, we estimate as follows:

$$n^{\Psi}_{v,t} = n^{\Psi}_{v,0} e^{r^{\Psi} * t}$$

Combining these equations, we can solve for:

$$r_v * t = ln\left(\frac{c_{v,t} * N_t * C_0}{c_{v,0} * N_0 * C_t} + e^{r^{\Psi} * t}\right) + ln(1/2)$$

In practice, rather than solving this equation separately for each timepoint, we regress the right half of the equation over time and estimate r_v as the slope of the regression. Since ln(1/2) is constant at every timepoint, it does not contribute to the slope. Thus, the term we actually compute is

$$ln\left(\frac{c_{v,t} * N_t * C_0}{c_{v,0} * N_0 * C_t} + 1\right)$$

which is regressed over time with an intercept at the origin.

We calculate slope as described separately for each barcode and each replicate. We also record the standard error of the estimate from the regression, which is propagated to form the error term.

This calculation requires an estimation of the population size, N_t, which we measured by plating the library at each timepoint. In some experimental designs, however, library sizes are not estimated. Unlike enrichment scoring, Model-Bounded Scoring does not normalize all scores to a reference allele and attempting to do so is not recommended. If the growth rate of wild type (or

another reference allele) is known, the population size can be estimated from the growth rate, and the appropriate term to be regressed over time is:

$$ln\left((2e^{r_{wt}*t}-1)\frac{c_{wt,0}*c_{v,t}}{c_{wt,t}*c_{v,0}}+1\right)$$

In our dataset, using this approach slightly improved replicability between different barcodes, but substantially decreased the linearity of each barcode over time (the improvement in replicability may be a result of more stringent filtering of marginally linear barcodes). This problem is expected to be worse on datasets for which exponential growth over the whole length of the experiment cannot be assumed. Therefore, we do not recommend this method when timepoint population sizes are available.

3.4.8 Score aggregation and data filtering

We calculate the slope separately for each barcode and timepoint, taking the standard error of the estimate as a starting point for error. Typically, we want to assess scores for protein-level variants, which may be represented by multiple DNA-level variants and/or multiple barcodes. First, we calculate a score for each barcode averaged across replicates. The average score is the arithmetic mean of the slopes, and the error is:

$$\frac{\sqrt{\epsilon_a^2+\epsilon_b^2+\ldots\epsilon_n^2}}{n}$$

where ε_n represents the standard error from a single replicate. We then aggregate barcodes that represent the same protein variant. For variants with missense or nonsense mutations, co-occurring synonymous mutations are ignored and all variants that produce the same protein sequence are averaged. Error is propagated over n barcodes. We separately calculate the standard error between barcodes that represent the same protein variant. The number of barcodes is also reported for each variant. For synonymous variants, barcodes are aggregated across all variants with synonymous mutations in the same codon and no co-occurring mutations. Variants with synonymous mutations in multiple codons and no missense mutations are

discarded. Barcodes are aggregated under the label ‘Wild type’ if the variant they represent exactly matches the wild type DNA sequence.

We impose data quality filters at four levels. First, barcodes are discarded if they do not meet a minimum threshold for input reads, which we set at 20. We observe that barcodes with fewer than 20 input reads have higher estimated error, on average. Second, if barcodes are significantly non-linear when regressing estimated log abundance over time, we discard it. We set a static threshold at S.E. = 0.2, discarding barcodes above this threshold. Third, we perform a repeated-measures ANOVA for each barcode to detect whether different replicates produce significantly different estimates across timepoints. Barcodes falling below a significance threshold, which we set at p = 0.05, are discarded. Fourth, we discard variants with significant disagreement between synonymous barcodes. If the standard error across all barcodes for a variant exceeds a static threshold, which we set at S.E. = 0.2, we discard the variant.

3.4.9 Modelling epistasis

For each double missense variant for which we scored each single missense variant composing it, we estimated the progeny expected for the double missense variant as follows:

$$r_{exp} = \frac{r_1 * r_2}{r_{wt}}$$

where r_{exp} is the expected growth rate; r_1 and r_2 are the measured growth rates of the single missense variants; and r_{wt} is the mean growth rate of variants synonymous to the wild type sequence. We estimated standard error for the expectation as follows:

$$SE_{exp} = \sqrt{SE_1^2 + SE_2^2 + 2SE_{wt}^2}$$

We calculated growth rates and error for the double missense variant using the same procedure as for single missense variants. We then compared the empirical growth rate to the expected growth rate for each variant. Variants were significantly epistatic if:

$$t = \frac{r - r_{exp}}{\sqrt{SE^2 + SE_{exp}^2}} > 1.96$$

3.4.10 Determining promiscuity

For each single missense variant that we scored on all four receptors, we estimated promiscuity as a rough estimate of the average infectivity over the space of possible receptor, relative to wild type. We calculated an ordinary least squares regression line, y = mx +b, of the variant-specific growth rate on each receptor over the mean growth rate of synonymous variants on each receptor. We calculated the area under the curve for each variant divided by the area under the curve for synonymous variants. For variants with a positive intercept, this was:

$$AUC = \frac{r_{wt}(r_{wt}m + 2b)}{2}$$

While for variants with a negative intercept, this was:

$$AUC = \frac{(r_{wt}m + b)(r_{wt} + b/m)}{2}$$

3.5 Figures and Tables

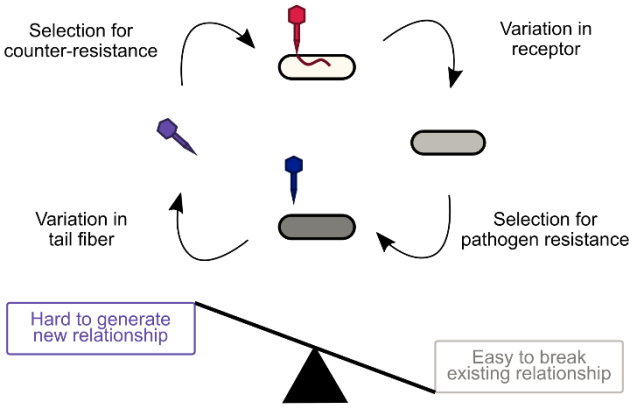

Figure 3.1: Phage–bacteria coevolution is fundamentally weighted in favor of bacteria. Phages and bacteria are presumed to coevolve stably and indefinitely in the wild, but theoretical considerations and experimental co-culture systems mostly predict long-term dominance of bacteria. Explaining the success of phages over evolutionary timescales is further hampered the limited mechanistic understanding of how phages adapt to a new and/or resistant host.

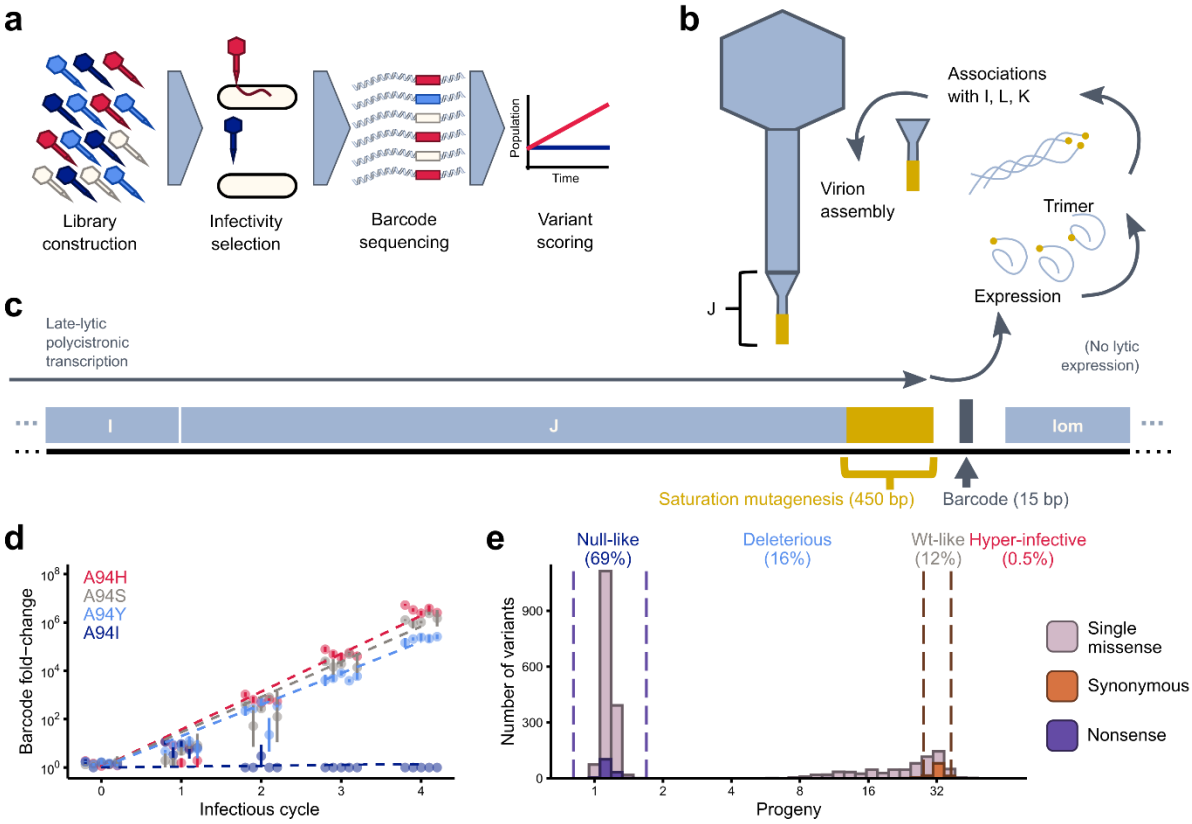

Figure 3.2: A sequencing-based method to assess phage infectivity across thousands of variants.

(a) The experimental strategy. A library of phage variants was constructed and subjected to selection for the ability to infect *E. coli*. After each infectious cycle, barcodes corresponding to each variant were deeply sequenced, and their frequencies were used to score each variant using Model-Bounded Scoring. **(b)** Binding of λ to the host is mediated by the J protein, which contacts the receptor, LamB. J spontaneously forms a homotrimer and then associates with accessory proteins that fold J into a mature conformation (Rajagopala et al., 2011). **(c)** J is the 3'-most gene on a long polycistronic transcript expressed in late lytic phase that contains most of the capsid genes. We mutagenized the 3'-most 450 bp of J (excluding the stop codon) using NNN codon replacement and inserted a 15 bp barcode downstream of the gene. **(d)** For each of four missense variants at A94, five randomly selected barcodes are plotted by their abundance in the phage population after each infectious cycle relative to the pre-packaging DNA pool. Infectious cycle = 0 corresponds to the packaged but not yet selected phage population. Error bars represent the standard error between 3 replicate selections. For each variant, the growth rate, r, is the slope of an ordinary least squares regression line calculated separately for each barcode and averaged across all barcodes representing the same protein-level variant. The average growth rates for the selected variants are shown by slopes of the dashed lines. **(e)** Progeny per infectious cycle, equal to the exponential of the growth rate (e^r), is shown for λ bearing synonymous (to wild type) variants in orange, nonsense variants in purple, and single missense variants in mauve. Dashed lines indicate score boundaries used to categorize variants as null-like, deleterious, wt-like, or hyper-infective.

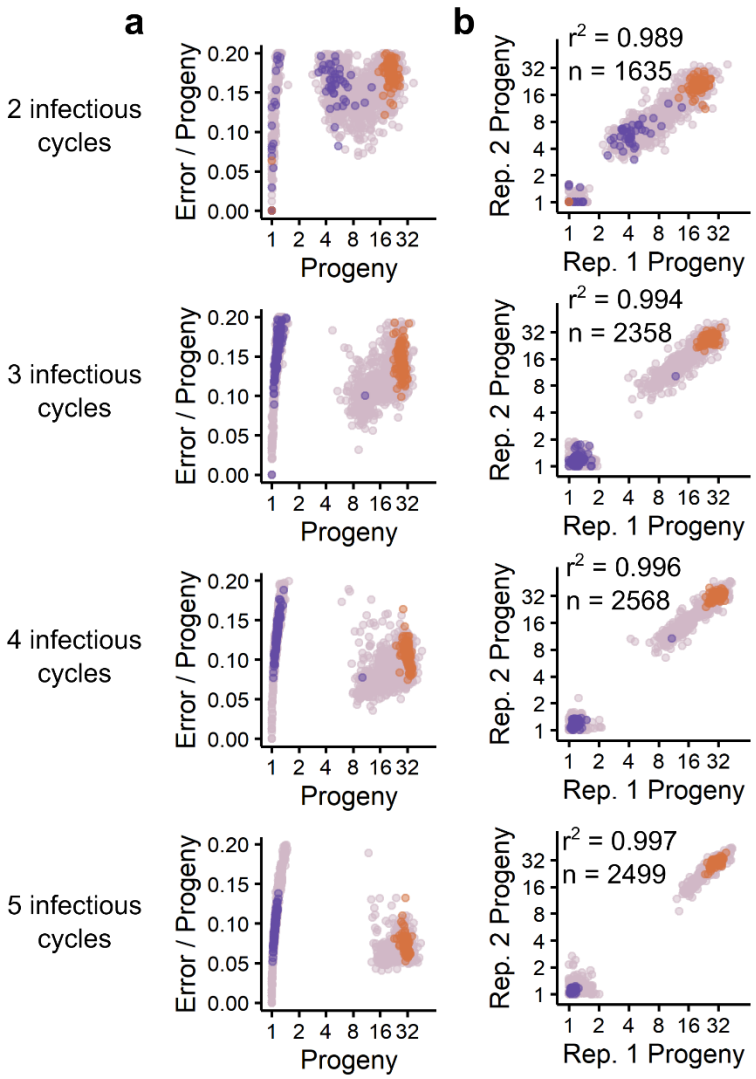

Figure 3.3: Replicability of selection for λ infectivity, as assessed by Model-Bounded Scoring.

(a) As more infectious cycles are added, the estimated error per variant decreases and the scoring better separates the nonsense and synonymous distributions. However, we start to lose variants that are strongly deleterious but not null-like as they get swept from the population, which is increasing in average fitness over time. We used scores from four infectious cycles in order to balance precision and coverage. Mauve: single missense variants. Orange: synonymous variants. Purple: nonsense variants. **(b)** Adding infectious cycles improve between-replicate correlation, but only modest improvements are possible.

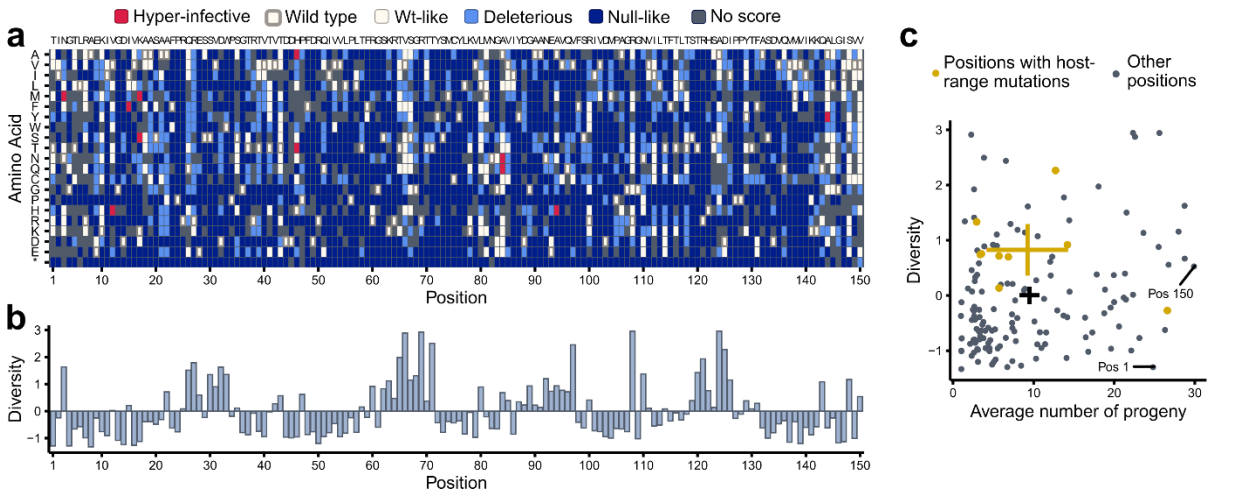

Figure 3.4: Comparison of empirical mutational tolerance and evolutionary diversity. **(a)** Categorical effects of all amino acid substitutions. Categories are defined in Figure 3.2e. The wt sequence of the J region is shown across the top and the amino acid substitutions on the Y axis. **(b)** Amino acid diversity, calculated from ConSurf (Ashkenazy et al., 2016), across 910 orthologs of J. Zero represents the mean diversity across positions. **(c)** For each amino acid position, the evolutionary diversity (y-axis) is compared to the average progeny produced by amino acid substitutions at that position (x-axis). Diversity of each position correlates only weakly with the average number of progeny ($r^2 = 0.06$, $p<0.01$), in contrast to cellular proteins for which mutational tolerance and evolutionary diversity are more strongly related (Roscoe and Bolon, 2014; Mishra et al., 2016; Boucher et al., 2014). Gold points indicate positions where mutations have been reported that expand host range (Meyer et al., 2010; Werts et al., 1994); these positions are more diverse but no more mutationally tolerant than the average position. Crosses represent 95% confidence intervals of the mean diversity and progeny of variants for positions with host range mutations (gold) or all positions (black). 'Pos': position.

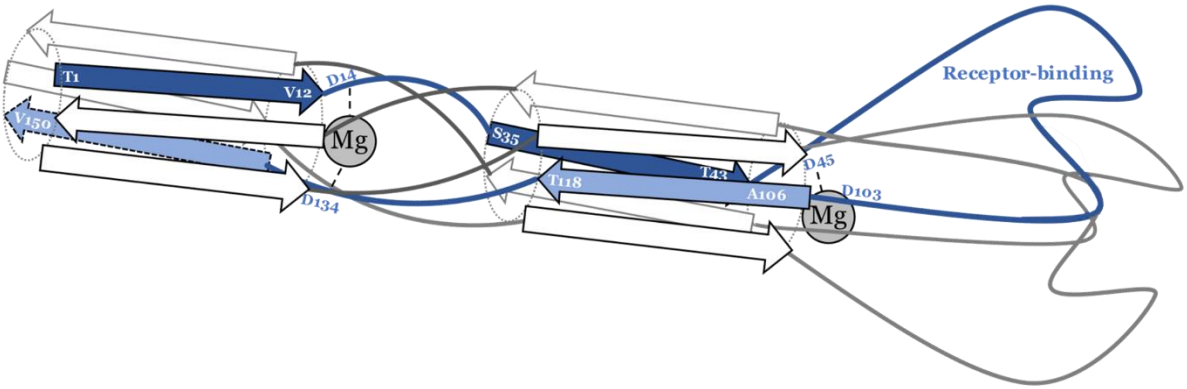

Figure 3.5: Coarse structural model for consideration of J structural features.
Based on the patterns of mutational tolerance and using the T4 tail fiber structure (Bartual et al., 2010) as a general guide, we propose this coarse model of J's structure. Antiparallel β-sheets form a helical barrel in the 'stalk' of the protein, with junctions between these sheets and less-structured regions defined by Mg-coordinating aspartate residues.

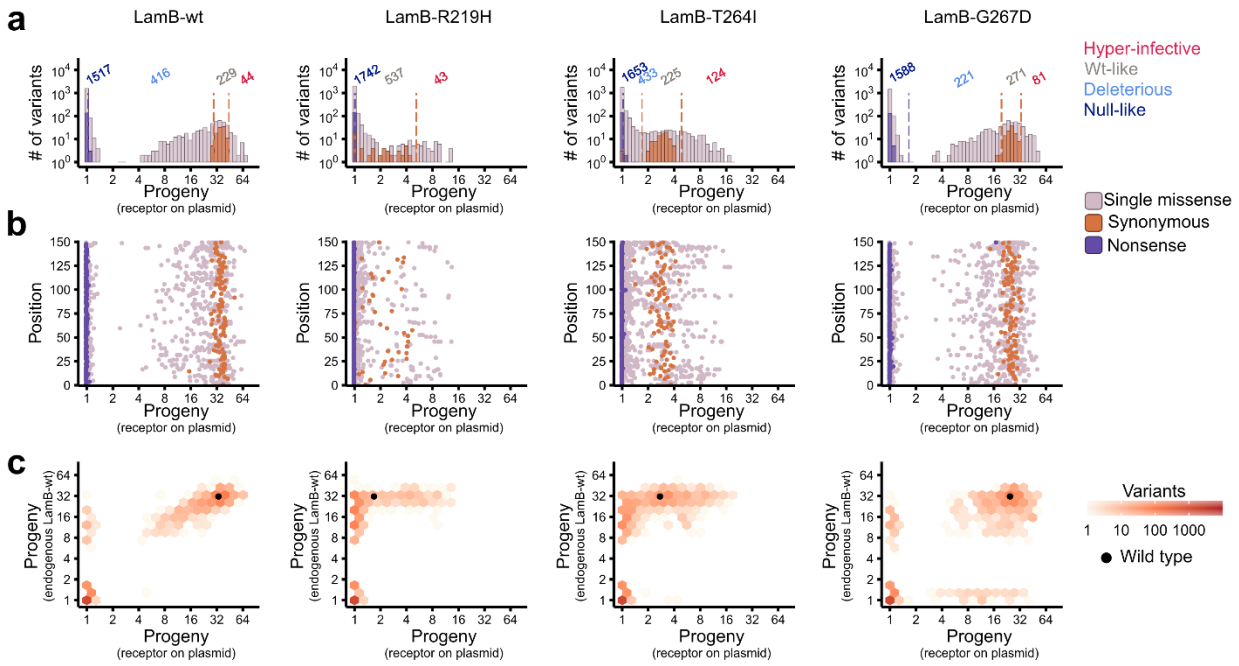

Figure 3.6: Selection of λ bearing J variants on λ-resistant hosts.
(a) Distribution of fitness effects on each host. In each case, the synonymous and nonsense distributions can be separated, despite the nominal λ-resistance of each lamB allele. Variants that outperform wild type λ do so by a larger margin on hosts that are more resistant. **(b)** Progeny for each J variant shown by the position of the mutation in the sequence. Adaptive mutations occur frequently at a subset of positions, but these positions are spread over the entire mutagenized region. Positions with many hyper-infective mutations are more apparent on the more resistant hosts LamB-R219H and LamB-T264I than on LamB-wt or LamB-G267D. **(c)** Correlation between progeny produced by λ bearing each J variant on its wild type host (y-axis), or on a host bearing a plasmid-borne lamB allele (x-axis). Variants that are highly infective on a non-wt host tend to also be infective on the wild type host, with some exceptions. However, many variants that are highly infective on the wild type host are poorly infective on non-wt hosts.

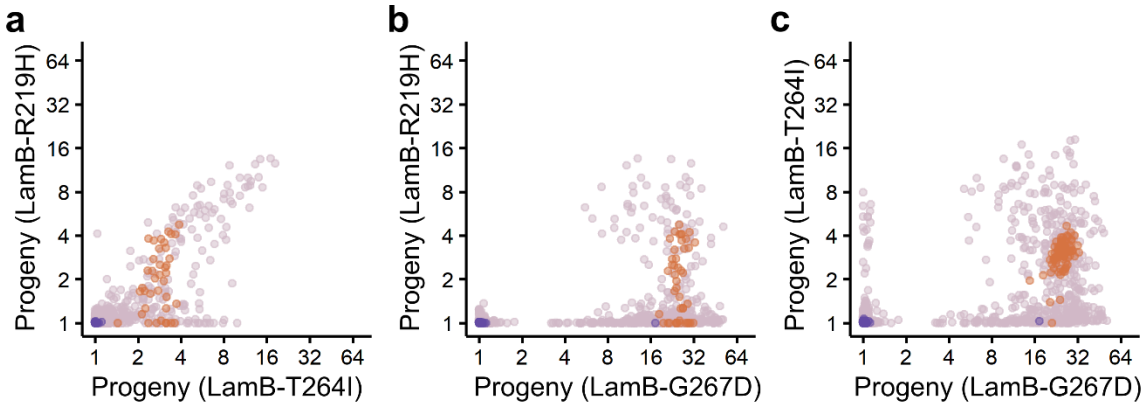

Figure 3.7: Comparison of J variant infectivity between multiple novel hosts.
For each pairwise comparison of hosts, excluding LamB^{wt}, the infectivity of all J variants measured on both hosts is shown. **(a)** LamB-R219H by LamB-T264I. **(b)** LamB-R219H by LamB-G267D. **(c)** LamB-T264I by LamB-G267D. For all hosts, the most infective viral variants are likely to be highly infective on the other host as well.

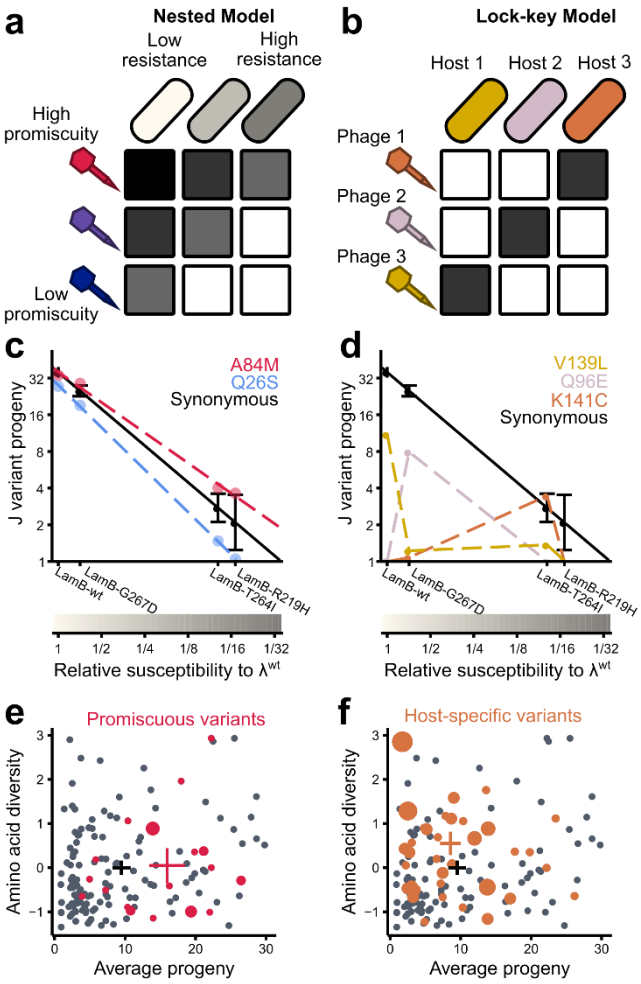

Figure 3.8: Discrimination of promiscuity from specificity by comparison of variant effects across different hosts.

(a) Under a nested model, phages have more or less ability to generally infect hosts, which we describe as 'promiscuity,' which contends with the level of resistance of potential hosts. **(b)** Under a lock-key model, the specific relationship between a host and phage determines infectivity, rather than host-independent properties of the phage or phage-independent properties of the host. **(c)** We can estimate the level of resistance of a given host by asking how well λ^{wt} produces progeny on it compared to LamB-wt (x-axis). Some J variants, like A84M are less affected by resistance than wild type λ, whereas others, like Q26S, are more affected. We calculate promiscuity as the area under the curve for a regression line comparing the variant infectivity to host susceptibility, relative to wild type λ. The black line represents synonymous variants, with error bars equal to +/- 1 standard deviation. **(d)** Some J variants confer infectivity on only a single host and are null-like on all others. Most of these variants, like Q96E, are deleterious, even on the host for which they are specific. Therefore, most of these variants cannot drive adaptation to a novel host by themselves, though they may work in concert with other variants. **(e)** Positions with promiscuous variants are shown in red with respect to their tolerance to mutation and amino acid diversity, with the size of the circle representing the number of unique promiscuous variants. The weighted average of these positions (cross, red) is more tolerant to mutation but not more diverse than the average of all positions (cross, black). Crosses represent 95% confidence intervals. **(f)** Variants that display specific infectivity on a single LamB variant fall at positions shown in orange. The weighted average of these positions has higher amino acid diversity, but is not more mutationally tolerant, than the average position.

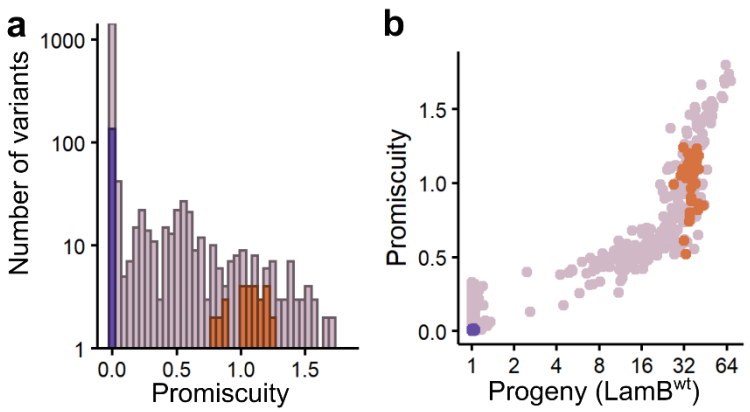

Figure 3.9: Distribution of promiscuity.
Promiscuity, as described in Figure 5, is shown for all synonymous (orange), nonsense (purple) and single missense (mauve) variants. **(a)** The distribution of promiscuity scores. Scores are calculated relative to wild type, such that the mean promiscuity of synonymous variants is 1. **(b)** Promiscuity is positively, though not linearly, associated with infectivity. The most promiscuous variants are therefore neutral-to-beneficial on a wild type host.

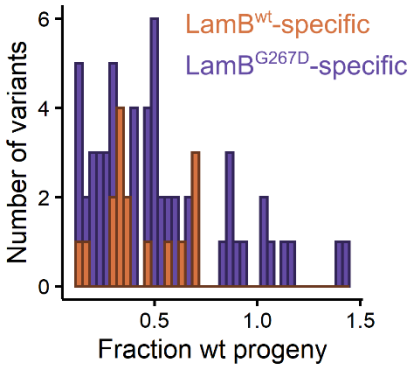

Figure 3.10: Distribution of scores for LamB^{wt}- and LamB^{G267D}-specific variants, on the host they are specific for.

For J variants that are specific to LamB^{G267D} or LamB^{wt}, the deleteriousness of those variants is shown as their progeny relative to wild type J. LamB^{wt}-specific variants are much less likely to be neutral or beneficial. We posit that this bias arises because wild type J has already been selected for specificity to LamB^{wt}, removing those variants except when they have a significant cost.

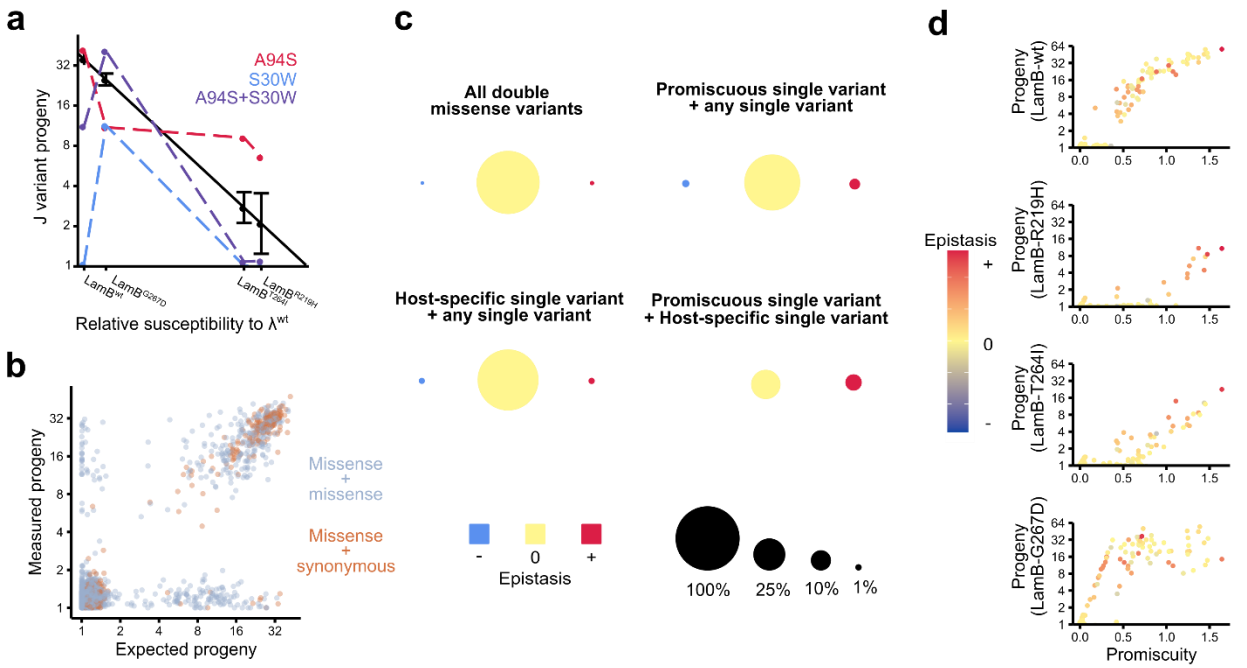

Figure 3.11: Double missense variants in J can mediate adaptation to new hosts. **(a)** The double missense variant S30W, A94S is a combination of a promiscuous variant (A94S) and a LamB-G267D-specific variant (S30W). The double missense variant exhibits sign epistasis, improving infectivity on LamB-G267D despite the deleteriousness of each single missense variant. **(b)** For double missense variants (silver), we calculated the expected progeny from the progeny of each of the single missense variants using a simple multiplicative model. A subset of double missense variants strongly deviates from the multiplicative model, in contrast to variants in which a single missense mutation is paired with a single synonymous mutation (orange), which are more likely to agree with the multiplicative model (r^2 = 0.92 missense x synonymous vs. r^2 = 0.64 missense x missense). **(c)** Across the four hosts, most double missense variants in J do not exhibit significant epistasis (top left panel). However, double missense variants that contain a promiscuous variant, a host-specific variant, or both, are more likely to exhibit significant epistasis. We measured significant epistasis in 13.2% of double missense variants containing promiscuous variants, 5.4% containing host-specific variants, and 23.8% containing both, compared to 1.8% containing neither. **(d)** For each host, progeny is positively associated with promiscuity and with positive epistasis. However, these associations become more salient on resistant hosts, with all infective variants on LamB-R219H being promiscuous and positively epistatic, compared to a minority of variants on LamB-wt.

		selection 1			selection 2				
J variant	*E. coli* allele(s) isolated from	Progeny	Error / estimate	barcodes	Progeny (wt)	Progeny (R219H)	Progeny (T264I)	Progeny (G267D)	Promiscuity
S29G	$LamB^{G151D}$	1.17	0.14	6	not measured	not measured	1.27	1.05	not measured
S30G	$LamB^{-}$	1.08	0.11	20	1.00	1.00	1.00	1.02	0.00
T58M	$LamB^{S247L}$,$LamB^{G246R}$,$LamB^{-}$	1.08	0.10	11	1.02	1.03	1.00	1.00	0.01
E93V	$LamB^{G151D}$	1.15	0.14	15	1.00	1.00	1.02	1.00	0.00
A94S	$LamB^{G151D}$	28.73	0.10	23	42.00	6.60	9.20	11.03	1.40
V95A	$LamB^{S247L}$	1.10	0.11	29	1.01	1.00	1.00	1.01	0.00
Q96R	$LamB^{G151D}$	1.09	0.11	58	1.01	1.00	1.00	1.01	0.00
H122L	$LamB^{-}$	1.18	0.15	12	1.06	1.00	1.01	1.00	0.01
D125K	$LamB^{-}$	not measured	not measured	not measured	not measured	not measured	2.62	1.00	not measured
L145P	$LamB^{E148K}$, $LamB^{G245R}$	1.10	0.12	4	1.03	1.00	1.00	1.00	0.00

Table 3.1: Previously published host range mutations detected in our assays.
For host range mutations reported in Meyer et al. (2012) or Werts et al. (1994), we report the estimated progeny in our selections. Most variants were strongly deleterious in our assays. Most of these host range mutations were identified in the context of other mutations, and we did not capture most of the double missense combinations. Therefore, these mutations could be advantageous in the context of positive epistasis in other assays.

Chapter 4: Deep mutational scanning data reveals evidence of genetic robustness at the level of synonymous codons

Synonymous codons vary in the mutations that are available to them, and DNA sequences encoding a specific protein can therefore have favorable or unfavorable mutational neighborhoods. Robustness, or the tendency of genes to maintain more positive mutational neighborhoods than expected by chance, has been postulated to result from population-level selection against deleterious mutations and/or individual-level selection against deleterious mistranslation events. Using data from published Deep Mutational Scans (DMS), we find pervasive evidence of robustness across 20 proteins, or 4,526 codons. Robustness is a feature of natural sequences, not a technical artifact of performing bulk assays, and cannot be fully explained by codon bias. In fact, robustness results primarily from position-specific mutational effects being well-matched to the codon used at that position, rather than overrepresentation of any particular codon. We find that robustness correlates with other sequence features, namely evolutionary conservation and amino acid buriedness, as has been previously reported. However, the strongest predictor of robustness is the empirical sensitivity of the position. Although we cannot rule out the hypothesis that selection for robustness is mediated by mistranslation, we do not observe signatures that might be expected given that mistranslation events are highly non-random. We suggest that further empirical studies are needed to investigate the possibility that the robustness of natural sequences results from population-level selection against deleterious mutations.

This work was performed in collaboration with Peter Conlin Ben Kerr. Peter and Ben conceived the initial project. Peter and I curated the data and performed the analyses. I wrote this manuscript, presently unpublished, in consultation with Peter, Ben, and Stan.

4.1 Introduction

Robustness, the ability of a system to maintain consistency in the face of errors, is present at many levels of biology. From signaling pathways with built-in redundancies to tumor suppressor genes that prevent runaway growth, cells have tools to handle exogenous and endogenous perturbations. At the molecular level, a random missense mutation may or may not disrupt a protein's function, and we might consider a protein more robust if mutations are less likely to disrupt function. Because the genetic code is redundant, there are many ways to encode the same protein sequence in DNA, with different encodings corresponding to different potential mutations. For example, it is possible for CGG(Arg) to change to CTG(Leu) via a single base substitution, but it is not possible for AGG(Arg) to change to any Leu codon via a single base substitution. It is therefore theoretically possible to optimize a DNA sequence to mitigate the effects of common errors, without altering the protein sequence. Whether natural evolutionary processes effectively optimize native gene sequence for robustness has not been conclusively demonstrated, except in 'digital organisms' (Wilke et al., 2001) and species with very high mutation rates (Lauring et al., 2012).

With any non-zero mutation rate, a large population with a single genotype at time zero will inevitably form a 'cloud' of related genotypes, sometimes called a quasispecies, with an average fitness that depends on the mutation rate and the local fitness landscape (Eigen, 1971; Wilke, 2005). Selection for favorable local fitness landscapes early in the history of life may explain why the genetic code is structured such that similar single base changes are frequently synonymous or lead to chemically similar amino acid substitutions (Freeland and Hurst, 1998). One could imagine similar selective pressures acting on synonymous codon choice in a quasispecies, and this has been experimentally demonstrated in certain cases (Lauring et al., 2012), but this quasispecies framework is generally thought to only apply to populations with very high error rates, such as RNA viruses or pre-biotic self-replicating sequences. More recently, it has been proposed that robustness could be mediated by mistranslation events,

which occur at ~10^{-3}-10^{-4}/position and thus affect ~20% of full-length proteins (Akashi, 1994; Shah and Gilchrist, 2010; Drummond et al., 2005). Since mistranslation is strongly biased towards missense variants that represent single-base substitutions (Mordret et al., 2019), this could plausibly select for codons with favorable local fitness landscapes because the landscape would be explored by the population of proteins inside a cell even absent DNA-level mutations.

Attempts to quantify robustness in real genes have been hampered by lack of accurate high-resolution maps of local fitness landscapes. Since arginine codons differ in their access to nonsense mutations, which are more likely to be deleterious than missense mutations, it is possible to detect robustness using arginine alone, but this approach has limited resolution (Plotkin et al., 2004; Plotkin et al., 2006). Another approach is to incorporate chemical similarity scores between different amino acids (Archetti, 2006). However, different protein positions often have unique amino acid preferences that differ in both magnitude and rank-order of missense effects, limiting the usefulness of chemical similarity scores.

With the advent of Deep Mutational Scanning (DMS), it is possible to empirically determine these amino acid preferences over thousands of positions, allowing one to assess robustness of a long protein-coding sequence. As an example, the Influenza A protein PB2 has arginine residues that are encoded by AGG at position 597 and CGG at position 604, but the amino acid preference profiles at these two nearby positions are dramatically different (Figure 4.1A). At position 597, serine substitutions are much more tolerated than leucine substitutions, while the inverse is true at position 604. In fact, at both sites, the mean variant effect of single-base substitutions from the wild type codon, marked in green, is less deleterious than the mean variant effect of single-base substitutions from the codon used at the other position, marked in gold (Figure 4.1b).

4.2 Results

4.2.1 Combining DMS datasets allows estimation of robustness over many positions

We estimate robustness as the difference between the mean variant effect of single-base substitutions from a codon in question, and the mean variant effect of single-base substitutions from all codons that encode the same amino acid (Equation 1).

Equation 1:

$$R_{c,p} = \frac{1}{9}\sum_{i=1}^{64}(t_{c\to i} * F_{i,p}) - \frac{1}{n}\sum_{j\in S_c}\frac{1}{9}\sum_{j=1}^{64}(t_{c\to j} * F_{j,p})$$

We gathered data from 20 previously published DMS experiments, normalized the scores such that wild type-like variant are centered around one and null-like variants are centered around zero, and calculated robustness for 4,526 positions, excluding any position where calculation of robustness required data that were missing. We grouped all synonymous variants into a single protein variant scores, with all wild type-synonymous variants scored '1' and all nonsense variants scored '0.' All positions encoding one of the 10 amino acids, C, D, E, F, H, M, N, Q, W, or Y, have robustness = 0, either because only a single codon encodes that amino acid, or because the synonymous codons have identical mutational neighborhoods. Of the remainder, 60.7% of positions have positive robustness, compared to only 39.3% of positions with negative robustness (Figure 4.1c). When we calculate average robustness across the 12 amino acid portion of the PB2 protein shown in Figure 4.1a, the wild type sequence is close to the optimal sequence (Figure 4.1d). Of 442,368 possible ways to encode this 12 amino acid sequence in DNA, only 2624 (0.6%) have average robustness greater than or equal to the wild type sequence.

4.2.2 Most genes in our dataset are more robust than expected by chance

Because the number of possible ways to encode a protein grows approximately exponentially with length, it is not computationally feasible to calculate the robustness of all possible encodings of a protein with hundreds of positions. Instead, we calculated robustness for 100,000 possible encodings and compared wild type to the distribution of possible encodings as a Z-score. The possible encodings are in all cases normally distributed – as expected from

the Lyapunov variant of the central limit theorem – centered around zero, and with variance negatively correlated with the length of the sequence (Figure 4.2a). In contrast, the mean wild type sequence has positive robustness and a Z-score of approximately 2 (95% C.I. [1.30,2.73]).

4.2.3 Gene-level robustness signatures remain after controlling for codon use

Codon use is not random because some codons are translated more efficiently than others, thus most genes have a codon bias that restricts the number of plausible encodings of a given protein (Grantham et al., 1980). It is plausible that codons that are translated optimally are also generally more robust, and this could lead to wild type sequences appearing robust when selection acts on translational speed. We therefore reconstructed the distribution of possible encodings for each gene in accordance with the codon bias of that organism at the genome or proteome (Figure 4.2b). In either case, most of the sequence maintain a positive Z-score, suggesting that codon bias alone does not lead to the levels of robustness we observe in wild type sequences. Since codon bias can vary between genes in a given organism, we also tried generating possible encodings by shuffling synonymous codons within the sequences that we analyzed (Figure 4.2c). The shuffled sequences in general had positive robustness, in some cases causing the difference between wild type and the mean shuffled sequence to be small. However, this effect was strongly negatively correlated with the length of the gene being shuffled, suggesting that the positive robustness is a result of similarity to the wild type sequence rather than the overall codon content of the gene.

4.2.4 Robustness is not an artifact of performing pooled assays

Because DMS experiments operate on the wild type sequence but not the hypothetical alternative sequences, an undetected bias for detecting variants with single base mutations compared to multiple base mutations could result in a spurious signal of robustness. Given that we have gathered data from experiments with different library generation, assay, and analysis formats, this possibility seems unlikely. Two further lines of evidence suggest that the robustness we measure is a real feature of the sequences in question. First, fitness scores for

natural sequences that are predicted from machine learning platform Envision, using evolutionary, physiochemical, and structural features show evidence of robustness (Figure 4.2d, 95% C.I. [0.15,1.59]). Although Envision was trained on DMS data, no actual experimental data was used to generate these scores. Second, computationally designed proteins, which have no natural evolutionary history, do not show strong evidence of robustness when subjected to DMS, with the caveat that designed proteins are generally much shorter than natural proteins and thus it is harder to detect robustness (Figure 1E, 95% C.I. [-0.67,0.23]). Taken together, these analyses imply that robustness is a result of the evolutionary history of the protein and not from technical biases inherent in collecting the scores.

4.2.5 Most of robustness signature comes from position-specific effects

The key advantage of using DMS data is that the most robust codon can vary between different synonymous positions. Thus, robustness could arise from using typically robust codons more frequently, or from using a given collection of synonymous codons in the optimal positions within a gene. As an example, consider glycine, where it is easy to predict which codons should be more robust. Namely, GGT and GGC both have single nucleotide variation that leads to missense substitutions S, C, and D that are, like glycine, relatively small in molecular weight and are less deleterious on average (Figure 4.3a). By contrast, GGA and GGG have single nucleotide variation that leads to nonsense variants or the comparatively large amino acids R, E, and W. We would therefore predict that GGT and GGC should have positive robustness while GGA and GGG should have negative robustness, and a long sequence could have positive robustness by having an overabundance or GGC/GGT compared to GGA/GGG, by using GGC/GGT disproportionately at sites that have a strong preference for small amino acids and GGA/GGG at sites that tolerate larger amino acids. Across all glycine positions in our dataset, every codon has positions where that codon has positive or negative robustness, indicating significant position-dependency (Figure 4.3b). Additionally, we can decompose the robustness of every wild type position into a predicted component (d_p) representing the median

robustness of that codon over all positions and a contextual component (d_c) represent the difference between the actual robustness and the predicted component. When averaged over all glycine positions, d_p represents the effect of using codons that are generally more robust more frequently, while d_c represents the effect of using codons at positions where they are more robust than they would be at the median position. For every amino acid with non-zero robustness, mean d_c > mean d_p (Figure 4.3c). This implies that the position-dependent component of robustness, which requires DMS data to observe, is generally greater than the any part of robustness arising purely from codon bias.

4.2.6 Robustness is associated with mutationally sensitive sites

Selection for robustness should be strongest at positions where errors are more deleterious. For example, in the Gal4 DNA-binding domain (Figure 4.4a), residues involved in contacting DNA or the other monomer are more robust than other residues. Over all of the positions in our dataset, we estimated evolutionary conservation and empirical mutational sensitivity, and we used crystal structures where available to estimate buriedness. We saw a significant correlation between robustness and buriedness, with the most buried decile of residues having ~11-fold higher robustness on average compared to least buried decile (Figure 4.4b). As has been previously observed (Mishra et. al., 2016), we saw a strong correspondence between evolutionary conservation and empirical mutational sensitivity, and both also correlate with robustness (Figure 4.4c). Since all of these features are co-correlated, we constructed a linear model to predict robustness from the other three features. In this linear model, robustness is significantly, though weakly, predicted by sensitivity ($p = 7.8 \times 10^{-6}$) and weakly predicted by buriedness ($p = 0.032$), but is not well predicted by conservation ($p = 0.454$), suggesting that conservation is only associated with robustness to the extent that it is a proxy for sensitive and/or buried positions. One explanation for this result is that selection for robustness may require positions to change over time.

4.3 Discussion

Our analysis was designed to detect signatures of genetic robustness, but not to test for any particular mechanism for the evolution of robustness. However, we might expect that different mechanisms would lead to different signatures of robustness. First, genes could be optimized to minimize protein misfolding events that result from mistranslation and cause toxic protein aggregation (Drummond and Wilke, 2008; Drummond et al., 2005; Akashi, 1994). However, calculating robustness scores weighted by mistranslation probabilities does not make sequences appear more robust. Similarly, the types of substitutions that contribute the most to robustness of wild type sequences do not correspond to common mistranslation events. Averaging over all 20 proteins, the most buffered mutations were c2a, t2g, t2a, and c2g, whereas mistranslation is driven primarily by G-U mispairing, or g1a and g2a in this notation (Mordret et al. 2019). Additionally, one might predict that viral proteins would be under less pressure to avoid causing proteotoxicity given that they often kill their host cells. In fact, we see that viral proteins are more robust than many cellular proteins. Last, we note that buriedness, which is thought to be associated with mutations leading to aggregation, predicts robustness much more weakly than mutational sensitivity.

Second, consider the framework, sometimes called ‘survival of the flattest,’ where a quasispecies has an average fitness that depends on the mutation rate and the local fitness landscape (Eigen, 1971; van Nimwegen et al., 1999; Wilke, 2001; Plotkin et al., 2006). Sequences randomly walking through a neutral or near-neutral network will be biased towards parts of the network with many neutral neighbors and few deleterious neighbors. Moreover, a population on a relatively ‘flat’ fitness peak (*i.e.*, where most mutations are near-neutral) can outcompete a population on a steep fitness peak, even if the maximum fitness of the steep fitness peak is higher (Wilke et al., 2001). However, this type of selection is dependent on effective population size and mutation rate, and specifically are only thought to be important

when N*μ >> 1 (2N*μ >> 1 for diploids) (van Nimwegen et al., 1999, Plotkin et al., 2006). While this condition may be met for some microbes, especially considering transient hypermutator phenotypes, (Giraud et al., 2001; Plotkin et al., 2006), it is probably not true for humans. We note that the robustness we measure is stronger for viral and bacterial proteins compared to human proteins, but there is a clear signal in human proteins as well. Recalculating robustness scores with weights derived from DNA mutation spectra does not improve them, and the most substitutions contributing most to robustness include several less-frequent mutations, such as transversions.

Neither mistranslation nor direct selection on quasispecies, then, can readily explain the signature of genetic robustness we observe. Nevertheless, sequences are in a more favorable mutational neighborhood than we would expect by chance. We note that every time a near-neutral missense substitution occurs in a protein's evolution, the derived codon is guaranteed at least one neutral substitution in its neighborhood (to the ancestral codon), while other codons synonymous to the derived codon do not have this guarantee. Thus, given a large number of near-neutral missense substitutions, we might expect signatures of robustness, even without selection directly for it. Further theoretical work is needed to estimate the strength of this effect and determine if it matches our empirical observations.

4.4 Materials and Methods

4.4.1 Data collection and normalization

Data were aggregated for 20 protein-coding sequences for which the functional effects of all or nearly all missense mutations have been empirically measured. The data originate from 17 unique publications and include genes from across the tree of life, mostly originating from *E. coli*, *S. cerevisiae*, or *H. sapiens*. Each of the selected sequence include data for >80% of possible missense mutations over a span of >60 amino acids. In cases where multiple assays

were done on the same library a single assay was selected that best represents the native context for the protein and shows evidence of moderate levels of selection. If multiple replicates of a single assay were performed, we averaged across all replicates.

For each of the selected datasets, we collected the logged enrichment values for single missense variants, and linearly rescaled them according to *Formula 1*, so that the mode of values for wildtype-like variants is centered around 1 and the mode of values for null-like variants is centered around 0.

$$E = \frac{E_m - M_0}{M_{wt} - M_0}$$

Where E_m is measured logged enrichment value for a variant, M_0 is the mode of logged enrichment values for null-like variants, and M_{wt} is the mode of logged enrichment values for wildtype-like variants. Modes were determined using the amps function in the modes package for R.

4.4.2 Determining the codon-level mean effect of mutation, K

Initially, we worked under the assumption that all mutations representing substitution of a single base are equally likely and that mutations representing multiple substitutions or indels have negligible likelihood. For each codon in the wild type DNA sequences, we calculated the average effect of base substitutions, K, by summing the variant scores, E, for all nine single base substitutions for that codon and dividing by nine.

$$\kappa_{c,p} = \sum_{j=1}^{64} t_{c \to j} E_{j,p} \qquad t_{c \to j} = \begin{cases} 1 & if \quad dist(c,j) = 1 \\ 0 & if \quad dist(c,j) \neq 1 \end{cases}$$

Synonymous mutations were assigned E = 1, and nonsense mutations were assigned E = 0. Where data was missing for a missense mutation needed for the calculation, we did not

analyze the position. We additionally filtered out sites where data was missing for a missense mutation from a synonymous codon, even if all data were present for the wild type codon.

4.4.3 Determining the weighted codon-level mean effect of mutation, K*

The possibility that errors can occur at different rates depending on the parent and child codons was considered by estimating $t_{c->j}$ for all codons c and j, under various transformative models. For all models, we approximated $t_{c->j}=0$ for cases where dist(c,j) > 1. Under the mutational model, when dist(c,j) = 1, we estimated $t_{c->j} = \mu_{x->y}$, the mutation rate between bases x and y that vary between codons c and j, as determined empirically for the species of origin. Under the transcription error model, we estimated $t_{c->j} = \theta_{v->w}$, the transcription error rate from base v to w that vary between codons c and j, as determined empirically for the species of origin, or the most closely related species for which there exist estimates of transcription error rates. Under the mistranslation model, we estimated $t_{c->j} = [\tau_j]w_{c->j} / \sum_{k=1}^{K}[\tau_k]w_{k\rightarrow j}$, where $[\tau_i]$ is the gene copy number of tRNAs with exact anticodon matches to codon i, $w_{i->j}$ is a weighting term describing the efficiency of a tRNA with an exact anticodon match to codon i for decoding codon j, and $K = \{k_1, k_2, \ldots k_K\}$ is the set of codons encoding the same amino acid as j (including j). The weighting terms w, were optimized via linear modeling by training on a dataset of mass spectrometry-derived natural mistranslation events in *S. cerevisiae* and *E. coli*, using the base position and identity as free terms.

4.4.4 Determining the sequence robustness score, Ψ_s

To generate robustness scores for a sequence as a whole, we averaged **K** or **K*** values for every position in the sequence that was not excluded due to missing data.

$$\psi = \frac{1}{P}\sum_{p=1}^{P}\kappa_{c,p}$$

The sequence robustness score of the wild type sequence, Ψ_{wt}, was compared to many DNA sequences that are synonymous to the wild type sequence. The synonymous DNA sequences were generated by randomly drawing a codon for each amino aicd position from the set of codons that encode the appropriate amino acid, either with equal probability, or with probability proportional to that codon's frequency in the genome, or with probability proportioinal to that codon's normalized translational efficiency (nTE). In the codon-shuffle case, synonymous sequences were generated by randomly picking correctly coding codons without replacement for each site from the pool of codons used by the wild type sequence. For each of the synonymous sequences generated, typically 1000 per gene, we calculated Ψ_{syn} by method described above, and compared Ψ_{wt} to the distribution of Ψ_{syn}. We calculated a Z-score as follows:

$$Z = \frac{\psi_{wt} - M_{\psi_{syn}}}{\sigma_{\psi_{syn}}}$$

4.4.5 Calculating the normalized codon robustness score effect, $\Delta_{c,p}$

To compare K values across datasets, we normalized K to the average K of possible synonymous codons at that site, and to the average K value for all possible encodings of the full-length sequence as follows:

$$\Delta_{c,p} = \frac{\kappa_{c,p} - \frac{1}{S}\sum_{s=1}^{S}\kappa_{s,p}}{\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{S}\sum_{s=1}^{S}\kappa_{s,q}}$$

Where S is the set of codons synonymous to c, including c.

4.4.6 Calculating the mean effect and contextual contributions to robustness, d_p and d_c

We decomposed $\Delta_{c,p}$ into components that are attributable to the mean effect of using a particular codon (*i.e.*, the predicted effect of using a particular codon), and the context-specific

difference between the mean effect and the actual measured effect of using that codon, as follows:

$$\delta_{predicted} = \frac{1}{R_c} \sum_{r_c=1}^{R_c} \Delta_{c,r_c}$$

$$\delta_{context} = \Delta_{c,p} - \frac{1}{R_c} \sum_{r_c=1}^{R_c} \Delta_{c,r_c}$$

4.5 Figures and Tables

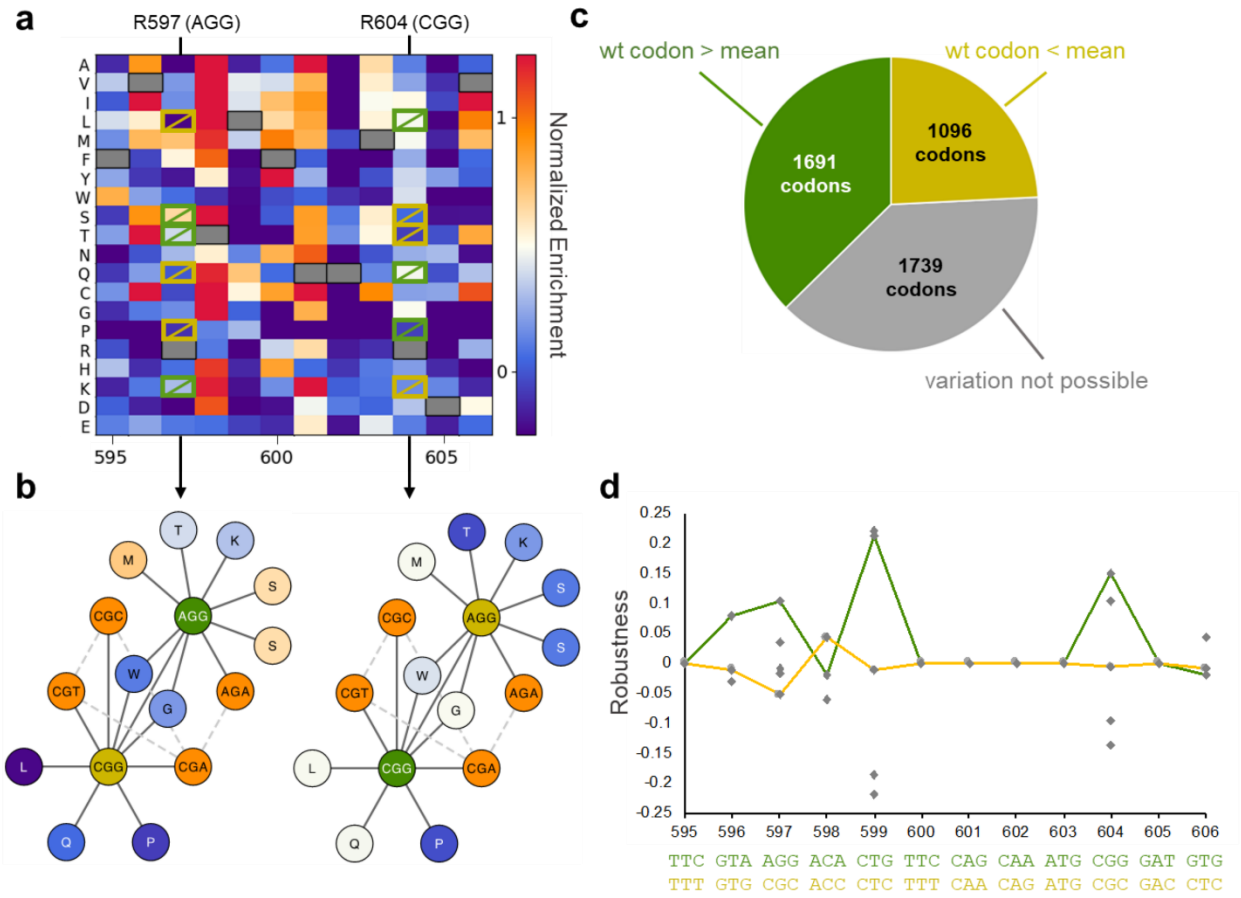

Figure 4.1: Different synonymous encodings of proteins potentiate different single nucleotide variation.
(a) A subset of the sequence of PB2 is shown with enrichment scores for each missense mutation. Two arginine residues encoded by different codons at positions 597 and 604 have substantially different mutagenic effects. **(b)** The protein-level variants that are possible via single base pair changes to the arginine codons AGG and CGG are displayed and colored according to the enrichment of that variant at position 597 or 604, respectively. At both positions, the wild type codon (green) has a more favorable neighborhood than the alternative codon (gold) **(c)** Across all the analyzed proteins, at 1691 positions (37.4%), the wild type codon has positive robustness, compared to 1096 positions (24.2%) where the wild type codon has negative robustness. **(d)** Across the positions shown in (a), the robustness of the wild type sequence (green) and a hypothetical alternative encoding (gold) are shown. Of the 442,368 possible ways to encode these 12 residues, only 2624 (0.6%) have equal or higher average robustness than wild type.

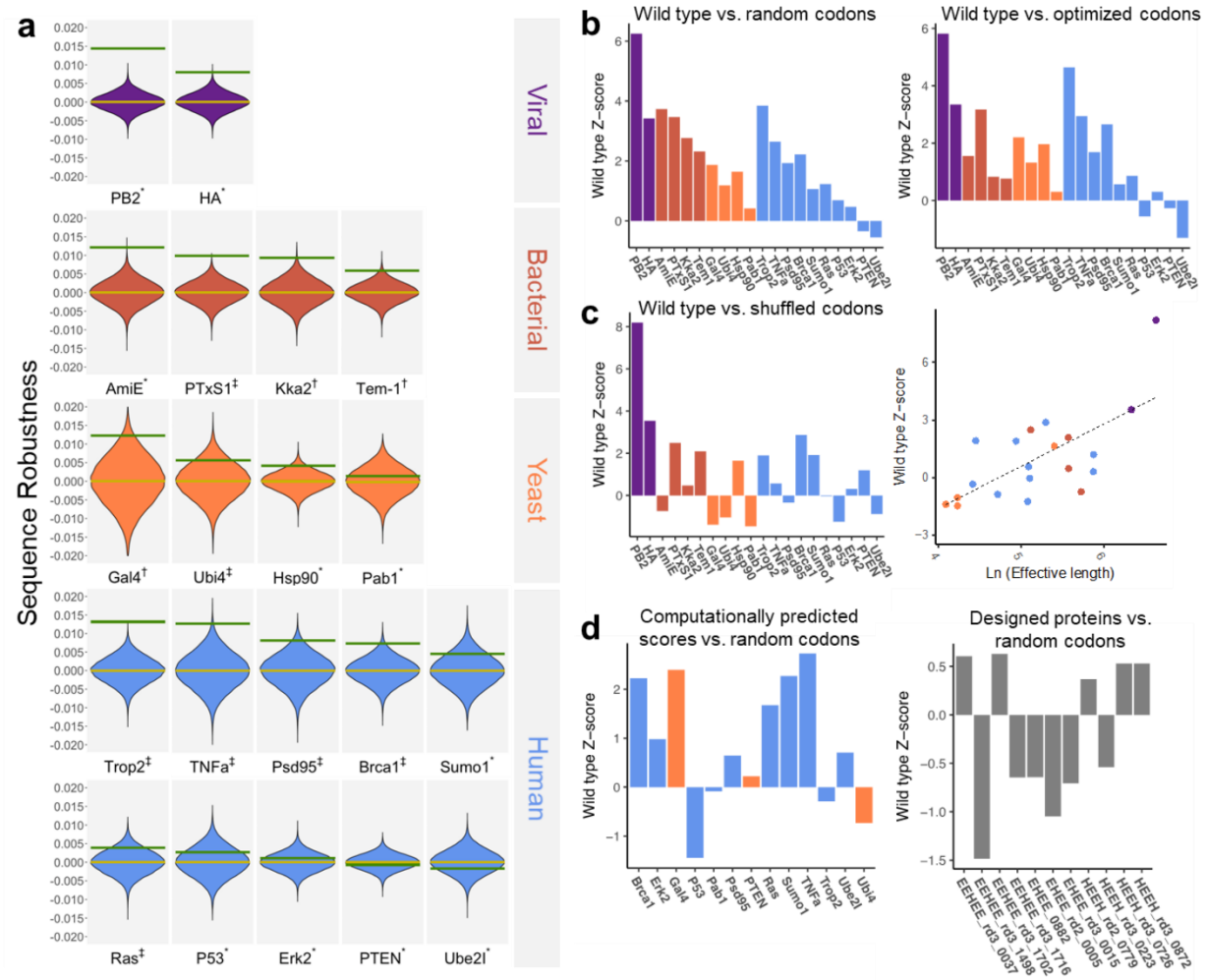

Figure 4.2: Most gene sequences are more robust than synonymous encodings of the same protein.

(a) The robustness of the wild type sequence is shown relative to the distribution of robustness of synonymous encodings of the same protein, generated by randomly drawing codon for each position with equal probability. *Proliferation or growth rate assay, †Drug- or reporter-based assay, ‡Binding-based assay (surface display or two-hybrid). **(b)** The Z-score of the wild-type sequence compared to the distribution of synonymous sequences generated by randomly drawing codons for each position with either equal probability, or probability corresponding to the frequency of codons in highly-expressed genes in the species of origin (E. coli used for all bacterial species). **(c)** The Z-score of the wild type sequence compared to the distribution of synonymous sequences generated by shuffling the positions of codons within a gene. The wild type Z-score is positively correlated with the natural-logged length of the sequence being shuffled (R^2 = 0.462, p = 0.00097). **(d)** Scores from Envision were used to calculate robustness in place of empirical scores. The wild type Z-scores are generally positive (95% C.I. [0.15,1.59]). **(e)** Deep mutational scans were performed on 11 proteins designed with Rosetta. The Z-scores of the reported parental DNA sequences are not significantly different from zero (95% C.I. [-0.67,0.23])

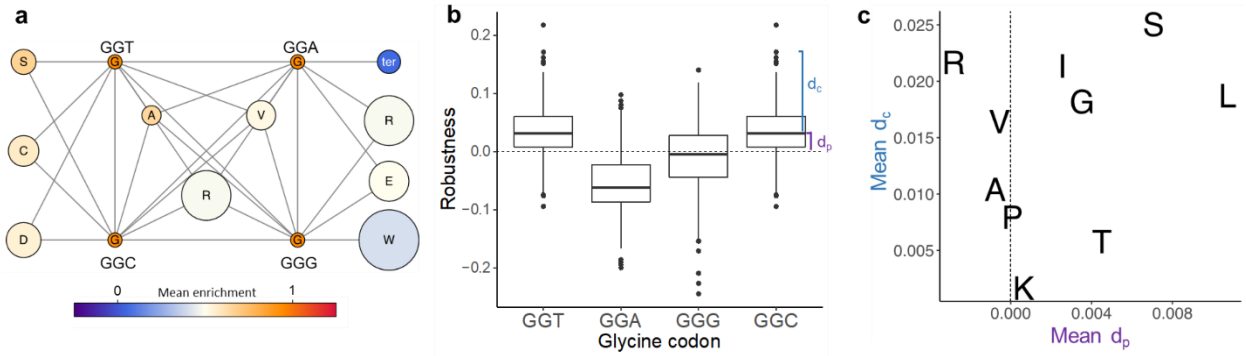

Figure 4.3: Robustness is highly dependent on the positional context within the protein. **(a)** The mutational neighborhood for glycine codons is shown with each node colored by the mean effect of substitutions from glycine across all positions, and sizes corresponding to the mass of each amino acid. **(b)** For each glycine codon, the robustness is plotted. Robustness can be decomposed into d_p and d_c, based on comparing the actual robustness of each point to the median robustness for that codon. **(c)** For each amino acid where robustness is possible, the mean d_p and d_c of wild type codons are compared. In all cases, mean d_c > mean d_p.

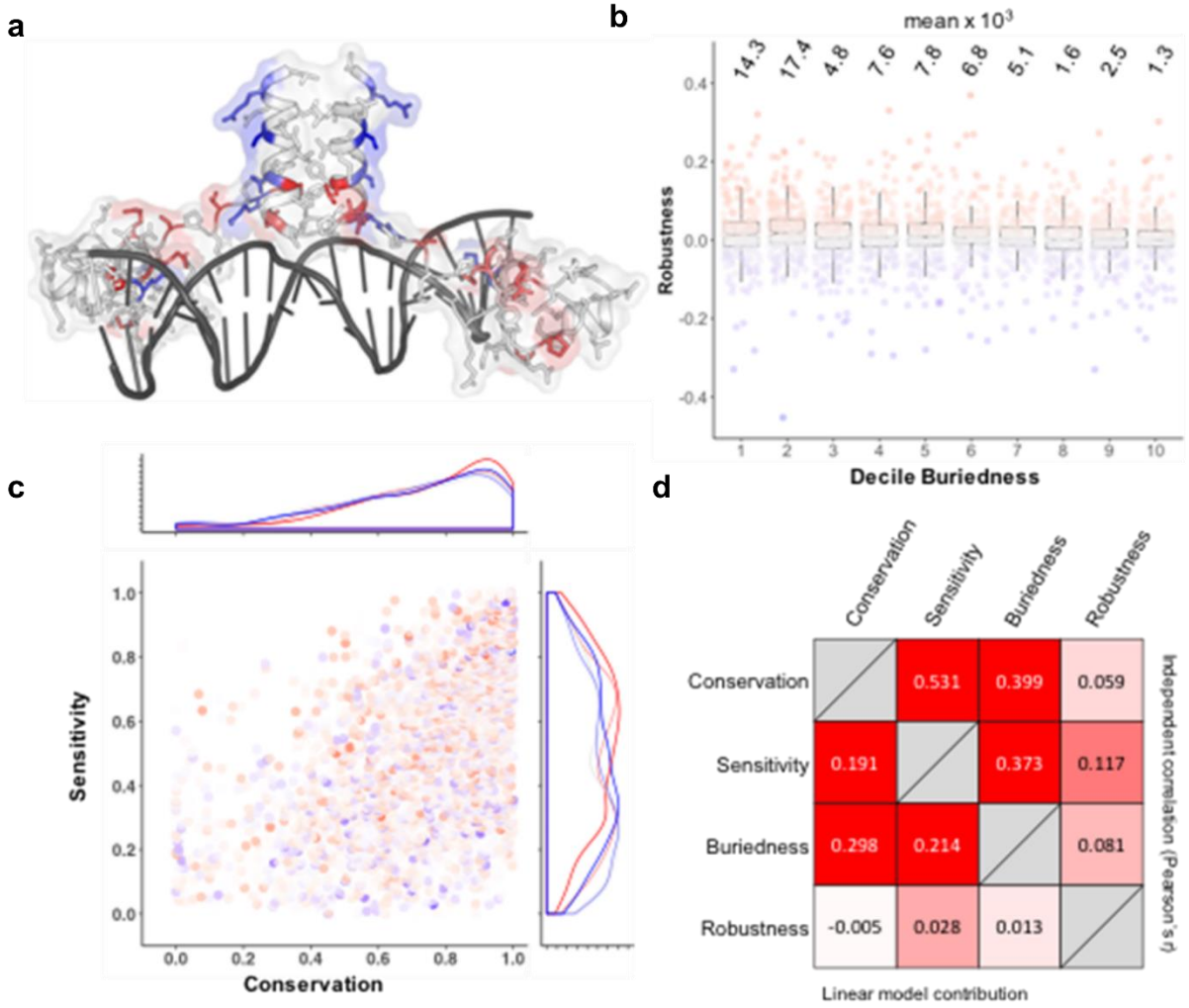

Figure 4.4: Highly robust codons are associated with evolutionarily and structurally sensitive positions.
(a) The structure of the dimeric Gal4 DNA-binding domain, from PDB 1D66, with amino acids colored according to robustness. Red: positive robustness. Blue: negative robustness. White: robustness not possible. **(b)** Robustness of each position that has structural information in PDB is negatively correlated with the solvent-accessible surface area of that residue. Pearson's r = 0.083, p = 1.7×10^{-7} **(c)** Robustness at each position varies with both conservation, determined from the frequency of mutations across taxa, and sensitivity, determined from the mean variant effect in the deep mutational scan. **(d)** Correlation analysis between the different factors analyzed reveals high correspondence between conservation, sensitivity, and buriedness. Each of these factors also correlates significantly with robustness, but in a linear model, sensitivity is by far the strongest predictor of robustness, with buriedness also providing a minor contribution. Red color indicates the magnitude of the correlation coefficient

Chapter 5: Conclusion

In this dissertation, I've discussed three related processes in protein evolution. First, I described the process by which a host receptor, LamB, evades a pathogenic interaction while maintaining its core function. Second, I described a viral protein, J, that must constantly expand its set of hosts and overcome host resistance. Third, I described a signature of robustness in many proteins, in which the mutational neighborhoods of these proteins are more favorable than would be expected by chance. Each of these processes exemplifies a different aspect of how proteins change over time, and it is worth making some direct comparisons between them.

LamB and J, despite both being rapidly co-evolving proteins, reside in remarkably different mutational neighborhoods. LamB is very difficult to disrupt, with most missense mutations behaving similarly to synonymous mutations. By contrast, J is very easy to disrupt, with most missense mutations conferring a null-like phenotype. This difference in mutational sensitivity corresponds to a difference in thermal stability – LamB is extremely difficult to denature, while J is only marginally stable. Moreover, the difference in mutational neighborhoods may affect the evolutionary outcomes of the arms race between them.

5.1 High stability of LamB promotes specific resistance mechanisms

If LamB were marginally stable, we would expect that most λ resistance mutations would come from the protein being destabilized, as opposed to specific missense mutations disrupting the λ binding site. If *E. coli* were under constant selection for LamB's function, maltose transport, this might not be a problem – evolution could still find those rare mutations that confer specific resistance. However, we know that *E. coli* is often not under selection for maltose transport, as maltose is not a preferred carbon source and is rapidly broken down into glucose in the human gut (Quezada-Calvillo et al., 2008). However, the ability to transport maltose may be periodically useful to the cell, if it happens to be in a low-glucose, high-maltose environment.

Therefore, maintaining LamB function even in the absence of selection is likely important for a lineage's long-term viability. It is precisely the stability of LamB that allows this to happen, by biasing the space of possible λ resistance mutations toward specific resistance mutations and away from destabilizing mutations.

In addition to stability, the fact that the maltose pore and the λ binding site are spatially segregated makes specific resistance mutations more likely. In the LamB structure, the maltose pore is occluded by several loosely structured loops, which evolve rapidly and are the targets of diverse phages. These loops have been proposed to have weak affinity for maltose, increasing its local concentration around the pore (Schirmer et al., 1995). I propose they may play an additional role in providing phages a non-essential target to latch on to, which can then tolerate mutations without consequence to maltose transport. This role would be analogous to hemagglutinin protein of influenza, which occludes its functional domain with hypervariable loops that stimulate an antibody response while also tolerating mutations to escape those antibodies (Xu et al., 2013). That this analogy draws a connection between a host protein to a viral protein suggests that this is not a property of a particular type of organism, but of the process of escaping from a deleterious interaction while maintaining a central function.

5.2 Sensitivity of J to mutation affords access to promiscuous variants

Although J has a very different mutational neighborhood to LamB, J's neighborhood is similarly well suited for effective coevolution. In chapter 3, I showed that a subset of mutations confers 'promiscuity,' causing the tail fiber to be less affected by various resistance mutations. On a mechanistic level, these promiscuous variants may represent increased flexibility, with the protein sampling multiple semi-stable conformations that allow binding to a broader set of receptors. This is a testable hypothesis, and while it is in consistent with prior literature (Petrie et al., 2018), more work is needed to elucidate this mechanism. The fact that J is marginally stable, with many deleterious mutations in its immediate neighborhood, affords J access to

destabilized variants that are important for overcoming host resistance. That a favorable mutational neighborhood for J means precisely the opposite as that for LamB may reflect that J is trying to maximize binding interactions while LamB is trying to minimize binding interactions.

5.3 Selection for favorable mutational neighborhoods

Evolution is not goal oriented – it only acts on the local population of variants with respect to their present environment and cannot 'plan' for some future sequence or environment. On this basis, it is worth being skeptical of the idea that proteins with favorable mutational neighborhoods could be selected in favor of equivalent proteins with less favorable neighborhoods. That said, if two populations, one with more favorable mutational neighborhood than the other, were mixed in an environment that required some adaptation, it is plausible that clones isolated at the end might originate mostly from the population with the more favorable neighborhood. This form of 'second-order selection' was observed by Woods et al. (2011) in an experimentally evolved *E. coli* lineage. In the context of virus-host coevolution, it is surprising that so many systems of antagonistic coevolution are stable throughout many cycles of resistance and counter-resistance. One explanation for this stability is that most systems die out within a few cycles, and the remaining systems are those that are capable of producing many rounds of resistance and counter-resistance. That is, there may be second-order selection for hosts and pathogens that are able to adapt across many cycles of resistance and counter-resistance.

Second-order selection is thought to depend heavily on the effective population size and mutation rate of the organism(s) in question (Wilke, 2001; Plotkin et al., 2006). However, we observe a signature of genetic robustness in human proteins (see chapter 4), indicating that mutational neighborhoods are unexpectedly favorable even when population size and mutation rate are small. In these cases, selection for some other property, such as translational fidelity, may underlie favorable mutational neighborhoods. While we did not observe specific evidence

that translational fidelity is being selected for, it remains possible that robustness results from selection for other qualities that coincide with a favorable mutational landscape. Similarly, while the mutational neighborhoods of J and LamB appear favorable for their potential to evolve resistance or counter-resistance, we cannot discount the possibility that these neighborhoods result from selection for other protein properties. For example, LamB is an outer membrane protein, and it may be exposed to harsh conditions, such as high or low pH, changing salinity, or denaturants like bile salts. Selection for tolerance to these conditions could lead to a hyper-stable protein structure, which would in turn lead to high tolerance to mutation.

5.4 Selection on general protein properties as a driver of evolutionary outcomes

Deep Mutational Scanning (DMS) studies that assess multiple functions frequently focus on the rare mutations that affect functions differently. Mutations that affect general properties like stability are often considered to be less relevant to the protein's evolution, especially in the case of destabilizing mutations. With LamB and J however, stability seems to be crucially important in determining evolutionary outcomes. To what extent does selection for general protein properties drive evolutionary outcomes for specific functions of those proteins? I think this question remains understudied, in part because of the technical limitations of current DMS experiments. Most experiments are set up to link variants of interest to cellular fitness or a cell phenotype that can be sorted on, and how these cell-level properties relate to protein-level properties like stability may be complex and is rarely explicitly modeled. As technologies improve, especially technologies that directly estimate protein stability (Matreyak et al., 2018; Rocklin et al., 2018), improvements are needed in out theoretical understanding of how mutations affect first proteins and then cell-level phenotypes.

5.5 Concluding remarks

Proteins are remarkable in their sophistication and efficiency, but what I find most remarkable is how they arise spontaneously from the simple process of mutation and selection. Genetics has a long history as a tool for linking DNA variants to organismal phenotype. As technology has improved, biologists have begun to dissect that relationship, first with microscopy uncovering cell-to-organism connections and later protein-to-cell connections, and now with sequencing elucidating the link between DNA variants and protein variants. Improvements to technology and theory are both needed to help fully resolve this link. The motivations for doing so are many-fold, encompassing improvements to medicine and a basic desire to understand how we, as living things, came to be. Understanding proteins as they are today is not sufficient for answering either question. We must also understand the forces that shape these proteins, both in the past and future.

References

Akashi H. Synonymous Codon Usage in Drosophila melanogaster: Natural Selection and Translational accuracy. Genetics. 1994;136:927–35.

Andrews B, Fields S. Distinct patterns of mutational sensitivity for λ resistance and maltodextrin transport in Escherichia coli LamB. Microb Genomics. 2020;6(4).

Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. PNAS. 2012;109(42):16858–63.

Archetti M. Genetic robustness and selection at the protein level for synonymous codons. J Evol Biol. 2006;19:353–65.

Arnold FH. How proteins adapt: Lessons from directed evolution. Cold Spring Harb Symp Quant Biol. 2009;74:41–6.

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44:W344–50.

Bandaru P, Shah NH, Bhattacharyya M, Barton JP, Kondo Y, Cofsky JC, et al. Deconstruction of the ras switching cycle through saturation mutagenesis. Elife. 2017;6.

Bartual SG, Otero JM, Garcia-Doval C, Llamas-Saiz AL, Kahn R, Fox GC, et al. Structure of the bacteriophage T4 long tail fiber receptor-binding tip. PNAS. 2010;107(47):20287–92.

Benz R, Schmid A, Vos-Scheperkeuter G. Mechanism of Sugar Transport through the Sugar-Specific LamB Channel of Escherichia coli Outer Membrane. J Membr Biol. 1987;100:21–9.

Berkane E, Orlik F, Stegmeier JF, Charbit A, Winterhalter M, Benz R. Interaction of bacteriophage lambda with its cell surface receptor: An in vitro study of binding of the viral tail protein gpJ to LamB (maltoporin). Biochemistry. 2006;45(8):2708–20.

Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. PNAS. 2006;103(15):5869–74.

Boucher JI, Cote P, Flynn J, Jiang L, Laban A, Mishra P, et al. Viewing protein fitness landscapes through a next-gen lens. Genetics. 2014;198(2):461–71.

Buckling A, Rainey PB. Antagonistic coevolution between a bacterium and a bacteriophage. Proc R Soc B Biol Sci. 2002;269(1494):931–6.

Burch CL, Chao L. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature. 2000;406(August):625–8.

Carrat F, Flahault A. Influenza vaccine: The challenge of antigenic drift. Vaccine. 2007;25:6852–62.

Chai T, Foulds J. Inactivation of Bacteriophages by Protein E, a New Major Membrane Protein Isolated from an Escherichia coli Mutant. J Bacteriol. 1979;137(1):226–33.

Charbit A, Gehring K, Nikaido H, Ferenci T, Hofnung M. Maltose Transport and Starch Binding in Phage-resistant Point Mutants of Maltoporin: Functional and Topological Implications. J Mol Biol. 1988;201:487–96.

Dai W, Hodes A, Hui WH, Gingery M, Miller JF, Zhou ZH. Three-dimensional structure of tropism-switching Bordetella bacteriophage. PNAS. 2010;107(9):4347–52.

Dennehy JJ. What Can Phages Tell Us about Host-Pathogen Coevolution? Int J Evol Biol. 2012;2012.

Dingens AS, Arenz D, Weight H, Overbaugh J, Bloom JD, Dingens AS, et al. An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes. Immunity. 2019;50(2):520–32.

Doud MB, Bloom JD. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. Viruses. 2016;8(155).

Doulatov S, Hodes A, Dal L, Mandhana N, Liu M, Deora R, et al. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature. 2004;431(7007):476–81.

Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43:389–94.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. PNAS. 2005;102(40):14338–43.

Drummond DA, Wilke CO. Theory Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. 2008;341–52.

Durfee T, Nelson R, Baldwin S, Plunkett G, Burland V, Mau B, et al. The complete genome sequence of Escherichia coli DH10B: Insights into the biology of a laboratory workhorse. J Bacteriol. 2008;190(7):2597–606.

Egloff P, Zimmermann I, Arnold FM, Hutter CAJ, Morger D, Opitz L, et al. Engineered peptide barcodes for in-depth analyses of binding protein libraries. Nat Methods. 2019;16(5):421–8.

Eigen M. Selforganization of Matter and the Evolution of Biological Macromolecules. Naturwissenschaften. 1971;58(10):465–523.

Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions. PNAS. 2011;108(28):E288–97.

Fowler DM, Fields S. Deep mutational scanning: A new style of protein science. Nat Methods. 2014;11(8):801–7.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence- function relationships. Nat Methods. 2010;7(9):741–6.

Franke J, Kloser A, Visser JAGM De, Krug J. Evolutionary Accessibility of Mutational Pathways. PLoS Comput Biol. 2011;7(8).

Freeland SJ, Hurst LD. The Genetic Code Is One in a Million. J Mol Evol. 1998;47:238–48.

Gehring K, Charbit A, Brissaud E, Hofnung M. Bacteriophage lambda receptor site on the Escherichia coli K-12 LamB protein. J Bacteriol. 1987;169(5):2103–6.

Gibson B, Wilson DJ, Feil E, Eyre-walker A, Eyre-walker A. The distribution of bacterial doubling times in the wild. RSPB. 2018;285.

Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6(5):343–5.

Giraud A, Radman M, Matic I, Taddei F. The rise and fall of mutator bacteria. Curr Opin Microbiol. 2001;4:582–5.

Gómez P, Buckling A. Bacteria-Phage Antagonistic Coevolution in Soil. Science. 2011;332(April):106–10.

Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife. 2013;(2).

Grantham R, Gautier C, Gouy M. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 1980;8(9):1893–912.

Guo H, Arambula L, Ghosh P, Miller JF. Diversity-generating Retroelements in Phage and Bacterial Genomes. Mob DNA III. 2015;1237–52.

Gurnev PA, Oppenheim AB, Winterhalter M, Bezrukov SM. Docking of a Single Phage Lambda to its Membrane Receptor Maltoporin as a Time-resolved Event. J Mol Biol. 2006;359(5):1447–55.

Hall A, Scanlan P, Morgan A, Buckling A. Host – parasite coevolutionary arms races give way to fluctuating selection. Ecol Lett. 2011;14:635–42.

Hall JPJ, Harrison E, Brockhurst MA. Viral host-adaptation: Insights from evolution experiments with phages. Curr Opin Virol. 2013;3(5):572–7.

Harrison E, Laine A, Hietala M, Brockhurst MA. Rapidly fluctuating environments constrain coevolutionary arms races by impeding selective sweeps. RSBP. 2013;280.

Heine H, Francis G, Lee KINS, Ferenci T. Genetic Analysis of Sequences in Maltoporin That Contribute to Binding Domains and Pore Structure. J Bacteriol. 1988;170(4):1730–8.

Hendrix RW, Casjens S. Bacteriophage λ and its Genetic Neighborhood. In: Calendar R, editor. The Bacteriophages. 2nd ed. 2006. p. 409–47.

Hochberg ME, Baalen M Van. Antagonistic Coevolution over Productivity Gradients. Am Nat. 1998;152(4):620–34.

Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. Bacterial Adaptation through Loss of Function. PLoS Genet. 2013;9(7).

Husain K, Murugan A. Physical Constraints on Epistasis. Mol Biol Evol. 2020;

Hyman P, van Raaij M. Bacteriophage T4 long tail fiber domains. Biophys Rev. 2018;10(2):463–71.

Kaiser D, Masuda T. In vitro assembly of bacteriophage Lambda heads. PNAS. 1973;70(1):260–4.

Kaltenbach M, Emond S, Hollfelder F, Tokuriki N. Functional Trade-Offs in Promiscuous Enzymes Cannot Be Explained by Intrinsic Mutational Robustness of the Native Activity. 2016;1–18.

Koskella B, Brockhurst MA. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiol Rev. 2014;38(5):916–31.

Lauring AS, Acevedo A, Cooper SB, Andino R. Codon Usage Determines the Mutational Robustness, Evolutionary Capacity, and Virulence of an RNA Virus. Cell Host Microbe [Internet]. 2012;12(5):623–32. Available from: http://dx.doi.org/10.1016/j.chom.2012.10.008

Lauring AS, Frydman J, Andino R. The role of mutational robustness in RNA virus evolution. Nat Rev Microbiol. 2013;11(May):327–36.

Lederberg E. Pleiotropy for maltose fermentation and phage resistance in Escherichia coli K-12. Genetics. 1955;40(5):580–1.

Lenski RE. Coevolution of bacteria and phage: Are there endless cycles of bacterial defenses and phage counterdefenses? J Theor Biol. 1984;108(3):319–25.

Lenski RE, Levin BR. Constraints on the Coevolution of Bacteria and Virulent Phage: A Model, Some Experiments, and Predictions for Natural Communities. Am Nat. 1985;125(4):585–602.

Liao HX, Lynch R, Zhou T, Gao F, Munir Alam S, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature. 2013;496(7446):469–76.

Lopez-Pascua L, Buckling A. Increasing productivity accelerates host – parasite coevolution. J Evol Biol. 2008;21:853–60.

Lopez-Pascua L, Alex R, Boots M. Higher resources decrease fluctuating selection during host – parasite coevolution. Ecol Lett. 2014;17:1380–8.

Mateus A, Määttä TA, Savitski MM. Thermal proteome profiling: unbiased assessment of protein state through heat-induced stability changes. Proteome Sci. 2017;15(13).

Mathieu A, Dion M, Deng L, Tremblay D, Moncaut E, Shah SA, et al. Virulent coliphages in 1-year-old children fecal samples are fewer, but more infectious than temperate coliphages. Nat Commun. 2020;11(1).

Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 2018;50(June):874–82.

McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012;491(7422):138–42.

Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 2014;42(14):1–8.

Merrikh CN, Merrikh H. Gene inversion potentiates bacterial evolvability and virulence. Nat Commun. 2018;9(1).

Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. Repeatability and contingency in the evolution of a key innovation in phage lambda. Science. 2012;335(6067):428–32.

Mighell TL, Evans-Dutson S, Roak BJO. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. Am J Hum Genet. 2018;102:943–55.

Mishra P, Flynn J, Starr TN, Bolon DNA. Systematic Mutant Analyses Elucidate General and Client-specific Aspects of Hsp90 Function. Cell Rep. 2016;15:588–98.

Mordret E, Dahan O, Asraf O, Geiger T, Lindner AB, Pilpel Y. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. Mol Cell. 2019;75:427–41.

Morona R, Henning U. Host Range Mutants of Bacteriophage Ox2 Can Use Two Different Outer Membrane Proteins of Escherichia coli K-12 as Receptors. J Bacteriol. 1984;159(2):579–82.

Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, et al. Antagonistic coevolution accelerates molecular evolution. Nature. 2010;464(7286):275–8.

Petrie KL, Palmer ND, Johnson DT, Medina SJ, Yan SJ, Li V, et al. Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. Science. 2018;359:1542–5.

Plotkin JB, Dushoff J, Desai MM, Fraser HB. Codon Usage and Selection on Proteins. J Mol Evol. 2006;63:635–53.

Plotkin JB, Dushoff J, Fraser HB. Detecting selection using a single genome sequence of M . tuberculosis and P . falciparum. Nature. 2004;428(April):942–5.

Pokusaeva VO, Usmanova DR, Putintseva E V., Espinar L, Sarkisyan KS, Mishin AS, et al. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. PLoS Genet. 2019;15(4):1–30.

Poullain V, Gandon S, Brockhurst MA, Buckling A, Hochberg ME. The Evolution of Specificity in Evolving and Coevolving Antagonistic Interactions between a Bacteria and its Phage. Evolution. 2007;62(1).

Quezada-Calvillo R, Sim L, Ao Z, Hamaker BR, Quaroni A, Brayer GD, et al. Luminal Starch Substrate ‘“ Brake ”’ on Maltase-Glucoamylase Activity Is Located within the Glucoamylase Subunit 1 – 3. J Nutr. 2008;138:685–92.

Rajagopala S V, Casjens S, Uetz P. The protein interaction map of bacteriophage lambda. BMC Microbiol. 2011;11(213).

Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017;357(6347):168–75.

Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, et al. Inferring protein 3D structure from deep mutation scans. Nat Genet. 2019;51(July).

Roscoe BP, Bolon DNA. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. J Mol Biol. 2014;426(15):2854–70.

Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, et al. A statistical framework for analyzing deep mutational scanning data. Genome Biol. 2017;18(150).

Salverda MLM, de Visser JAGM, Barlow M. Natural evolution of TEM-1 β-lactamase: Experimental reconstruction and clinical relevance. FEMS Microbiol Rev. 2010;34(6):1015–36.

Schirmer T, Keller TA, Wang Y, Rosenbusch JP. Structural Basis for Sugar Translocation Through Maltoporin Channels at 3.1 Å Resolution. Science. 1995;267(5197):512–4.

Schmiedel JM, Lehner B. Determining protein structures using deep mutagenesis. Nat Genet. 2019;51(7):1177–86.

Schrag SJ, Mittler JE. Host-Parasite Coexistence : The Role of Spatial Refuges in Stabilizing Bacteria-Phage Interactions. Am Nat. 1996;148(2):348–77.

Shah P, Gilchrist MA. Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. PLoS Genet. 2010;6(9).

Shaw JE, Bingham H, Fuerst CR, Pearson ML. The Multisite Character of Host-Range Mutations in Bacteriophage λ. Virology. 1977;83:180–94.

Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, et al. Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. Science. 2012;336(June):1157–61.

Silva JB, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 2016;363(4):1–11.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively parallel functional analysis of BRCA1 RING domain variants. Genetics. 2015;200(2):413–22.

Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016;25:1204–18.

Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell. 2015;160(5):882–92.

Szmelcman S, Schwartz M. Maltose Transport in Escherichia coli K12. Eur J Biochem. 1976;65:13–9.

Tetart F, Desplats C, Krisch HM. Genome Plasticity in the Distal Tail Fiber Locus of the T-even Bacteriophage: Recombination between Conserved Motifs Swaps Adhesin Specificity. J Mol Biol. 1998;282:543–56.

Thirion JP, Hofnung M. On some genetic aspects of phage λ resistance. Genetics. 1972;71:207–16.

Thomason L, Court D, Bubunenko M, Constantino N, Wilson H, Datta S, et al. Recombineering : Genetic Engineering in Bacteria Using Homologous Recombination. Curr Protoc Mol Biol. 2007;1.16(1):1–24.

Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol. 2009;19:596–604.

Tokuriki N, Tawfik DS. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature. 2009;459(7247):668–73.

van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. PNAS. 1999;96(August):9716–20.

Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, et al. A framework for exhaustively mapping functional missense variants. Mol Syst Biol. 2017;13(957).

Weitz JS, Hartman H, Levin SA. Coevolutionary arms races between bacteria and Bacteriophage. PNAS. 2005;102(27):9535–40.

Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, et al. Phage-bacteria infection networks. Trends Microbiol. 2013;21(2):82–91.

Werts C, Michel V, Hofnung M, Charbit A. Adsorption of bacteriophage lambda on the LamB protein of Escherichia coli K-12: Point mutations in gene J of lambda responsible for extended host range. J Bacteriol. 1994;176(4):941–7.

Wilke CO. Quasispecies theory in the context of population genetics. BMC Evol Biol. 2005;8.

Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature. 2001;412(6844):331–3.

Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. Second-order selection for evolvability in a large Escherichia coli population. Science. 2011;331(6023):1433–6.

Wrenbeck EE, Azouz LR, Whitehead TA. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nat Commun. 2017;8:1–10.

Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA. Plasmid-based one-pot saturation mutagenesis. Nat Methods. 2016;13(11):928–32.

Wroblewska A, Dhainaut M, Ben-Zvi B, Rose SA, Park ES, Amir EAD, et al. Protein Barcodes Enable High-Dimensional Single-Cell CRISPR Screens. Cell. 2018;175(4):1141-1155.e16.

Xu R, Krause JC, Mcbride R, Paulson JC, Crowe JE, Wilson IA. A recurring motif for antibody recognition of the receptor-binding site of influenza hemagglutinin. Nat Struct Mol Biol. 2013;20(3):363–70.

Zhang XC, Han L. How does a β-barrel integral membrane protein insert into the membrane? Protein Cell. 2016;7(7):471–7.

Zheng Y, Roberts RJ, Kasif S. Identification of genes with fast-evolving regions in microbial genomes. Nucleic Acids Res. 2004;32(21):6347–57.

Vita

Bryan Andrews was raised in Davis, California, where he had his first contact with experimental science through an internship in Pamela Ronald's lab at UC Davis. He later completed his B.S. in Biological Sciences and minor in Philosophy at Cal Poly San Luis Obispo, studying the regenerative processes of colonial ascidians with Elena Keeling. While an undergrad, Bryan also worked at the Applied Biotechnology Institute optimizing protein expression in maize, focusing on an oral vaccine for Hepatitis B. After graduating from Cal Poly, Bryan joined the Molecular and Cellular Biology program at the University of Washington. He joined the Fields lab in 2015. Bryan plans to continue investing mechanisms of molecular evolution as a postdoc with Rama Ranganathan.