Using Deep Mutational Scanning to Understand and Engineer Host Specificity in Bacteriophages

By

Phil Huss

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 06/30/2022

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

Srivatsan Raman, Assistant Professor, Department of Biochemistry

Kristen Bernard, Professor, Department of Pathobiological Sciences

Karthik Anantharaman, Assistant Professor, Department of Bacteriology

Nathan Sherer, Associate Professor, Department of Oncology

Rodney Welch, Professor Emeritus, Department of Medical Microbiology & Immunology

Table of Contents

Abstract

Bacteriophages, or ‘phages’, are bacterial viruses that play a central role in microbial ecosystems. Understanding interactions between phages and their bacterial hosts has contributed to major advances in modern biology and has positioned phages as tools for editing complex microbial communities and as treatments for antibiotic-resistant bacterial infections. A major impediment to advancing phage biology from empirical observations to molecular function has been a lack of high throughput methods to systematically profile gene-function or variant-function relationships. As such, the focus of this dissertation has been developing and implementing a platform for understanding and engineering bacteriophages in a high throughput manner.

To accomplish this goal I first developed ORACLE, a method for creating abundant, unbiased phage libraries that facilitates testing more than 10^5 phage variants at once. We used ORACLE to study the tip domain of the T7 phage receptor binding protein, a key region that determines phage activity. We created a deep mutational scanning (DMS) phage library of this region and functionally profiled the activity of variants on different *E. coli* hosts. We determined where activity-enhancing substitutions were concentrated and revealed host-specific substitutional patterns. Comparison of these results created a functional profile for the phage, establishing a high-resolution map of which genetic changes drive phage activity.

We then leveraged the results of our DMS assay to establish a new process called Motif Curation to find sequence motifs that were responsible for receptor recognition in metagenomic datasets. Motif Curation identified over 10^4 relevant motifs from diverse

metagenomic phages located throughout the T7 tip domain. Our process allowed us to identify motifs that are highly divergent from the wildtype sequence which can be sourced from diverse, non-homologous phage sequences. Screening these phages revealed hundreds of phages with novel host specificity and key combinations of substitutions that drive activity and host range. Overall, the platform technology and approaches established in this dissertation can be used to interrogate sequence-function relationships in phages to create powerful combinatorial libraries to alter phage activity and customize phages for differ roles.

Dedication

For Laura. You inspire me every day and make me want to be the best scientist I can be.

Acknowledgements

Foremost I would like to thank my graduate mentor Srivatsan Raman. Your patience and encouragement has motivated me throughout my PhD work, and I greatly respect your ability to push forward in exciting new directions in the lab. I have never been afraid to try new things or propose new experiments, and for this I am deeply grateful. It has been a pleasure working with you and I look forward to our future ventures.

I would like to thank my committee members Dr. Kristen Bernard, Dr. Karthik Anantharaman, Dr. Nathan Sherer and Dr. Rodney Welch for your continual support and feedback. Your questions have always been astute and I'm grateful that we've always had fun during my committee meetings.

Thank you to members of the Raman lab and our collaborators in the Anantharaman lab, we have all been in this together. Special thanks to Dr. Kyle Nishikawa, Anthony Meger, Dr. Megan Leander, Chutikarn Chitboonthavisuk and Dr. Kris Kieft. Kyle, Tony and Megan you all helped me get off my feet with coding and the basics of creating the plasmid libraries that define our work. Much of this work would have been a great struggle without your help. Chutikarn it has been a pleasure working through phage experiments with you. Kris, it has been great working with you on projects over the years. Thank you all for your support and kindness throughout my time in the lab.

Thank you to the Microbial Doctoral Training Program, especially our coordinators Cathy Davis Grey and Terra Theim who have kept the ship afloat. This is a great community that has created a welcoming and supportive environment.

Thank you to John DiCaro, my former boss at Covidien/Medtronic before I joined graduate school. I learned a lot about science and hard work from you and am grateful for your support of my pursuit for a PhD. I wouldn't be where I am today without your help.

Thank you to my friends and cohort for your friendship and support. Trips to Chicago have been a welcome relief from graduate school and I'm grateful for all the trivia and board games we've been able to play.

Thank you to my parents Phil and Stephanie Huss for putting me through college and always trusting me. Times haven't always been easy, but we've pulled through together.

Finally, thank you to my wonderful fiancée Laura Alexander. You have always been so supportive, so funny, and so smart. Graduate school has been a much better experience with you, and I've always appreciated that we can talk about each other's science. I love you and can't wait to tie the knot.

.

Chapter 1

Introduction

Introduction

Bacteriophages, or ‘phages’, are bacterial viruses that play an essential role in microbial ecosystems. Decades of research exploring how phages interact with their bacterial hosts has enabled development of widely used tools in modern biology including polymerases, recombinases and CRISPR-Cas systems (Salmond and Fineran, 2015). Phages typically have a narrow host range and only infect and kill specific bacteria without affecting other bacteria or mammalian cells. This specificity makes them valuable tools for targeted killing of drug-resistant bacteria, as programmable biocontrol agents (as reviewed in **Chapter 2**), for precise modulation of the microbiome, gene delivery devices, and as diagnostic sensors for pathogens (Canfield and Duerkop, 2020; Chen et al., 2017; Clokie et al., 2011; Dedrick et al., 2019; Eskenazi et al., 2022; Hatfull et al., 2022; Mutalik and Arkin, 2022; Schooley et al., 2017). Understanding how sequence changes drive phage activity is important for tailoring phages for these different roles and allows us to determine how phages interact with their hosts in microbial ecosystems.

Functional studies of phages have been traditionally restricted to evaluating one or only a few gene or genome variants, leaving large swathes of the sequence-function space unexplored. Traditional phage assays, such as passaging and isolating individual plaques, are laborious and simply do not scale for large functional genomics studies. Identifying only a few variants that perform well in a certain condition leaves us without a molecular understanding of how sequence changes drive phage activity. How might those variants fit into the larger picture? How do they compare to other variants in that condition? Are they more or less fit than similar variants, and what does that mean for understanding

the phage? As reviewed in **Chapter 3**, a major impediment to advancing phage biology from empirical observations to molecular function has been a lack of high-throughput methods to systematically and comprehensively profile gene-function or variant-function relationships.

The central focus of this dissertation has been developing and implementing a platform for understanding and engineering bacteriophages in a high throughput manner. Fundamentally our goal has been to find a way to test many phages, all at once, to create a generalized framework to describe sequence-function relationships. This goal had three steps. First, we established a method to create a large pool of phage variants, called a phage library, which can be easily tested in different conditions. Our approach can screen $>10^5$ pre-determined phage variants, a dramatic increase in scale compared to traditional approaches. Phage variants can be scored using deep sequencing to determine which variants do better or worse in different conditions. Second, results can be compared to create a functional profile for the phage, establishing a map of which genetic changes drive changes in phage activity. Finally, these results can be leveraged to create powerful combinatorial libraries to alter phage activity and customize phages for differ roles.

To create the phage library that underpins this analysis we had to overcome two key problems, library bias and library abundance. Library bias is the reduction or elimination of phage variants due to selection pressure during library formation. Creating phage variants necessitates passaging phages on a laboratory bacterial host, which inadvertently biases the library towards variants that are more fit on that host. Variants

that cannot grow or perform poorly on that laboratory host may be active on different hosts or in different conditions. Critically, these variants are likely to be the most relevant to understanding or engineering that phage, as they necessarily can affect the activity and host range of that phage. Loss of these variants would dramatically reduce our ability to understand the sequence-function landscape in a phage.

Library abundance is the proportion of phages in a phage population that are the actual library variants. Many methods for incorporating variant sequences into phages are low frequency events, making library abundance a problem. For example, reported rates of recombination vary wildly, with rates between 10^{-10} and 10^{-3}. Consequently, much of the phage population remains unrecombined or wild type phages. This effect makes screening the phage population substantially more challenging. Dramatically higher sequencing volumes are required to detect the variant population, variants can easily fall below the limit of detection for the assay, and experimental results were expected to become more stochastic. Limiting library bias and ensuring high library abundance was thus essential, and no available approach existed that overcame these problems that we could use for our desired phages.

To create an abundant, unbiased phage library I therefore developed a method called ORACLE (**O**ptimized **R**ecombination, **Ac**cumulation and **L**ibrary **E**xpression). This method is fully described in our publication in **Chapter 4**. ORACLE uses Cre recombinase to insert sequence variants from a plasmid library into phages while they passage through a laboratory host. We established that the rate of this recombination was ~10^{-3} and recombined phages had no plaquing deficiency. To minimize library bias, a separate

helper plasmid constitutively provides the wildtype gene *in trans* such that all progeny phages can amplify comparably regardless of the fitness benefit or deficient of any variant. To increase library abundance, we passage all progeny phages on host with Cas9 targeting unrecombined phages, retaining the helper plasmid to continue minimizing library bias. We showed that recombined phages are successfully enriching >1000 fold using this approach. Finally, we propagate the phages once without any plasmid, allowing the library variant to become expressed on the phage particles. We showed that using ORACLE we successfully retained and expressed 99.7% of desired variants in the final phage library, overcoming library bias. Furthermore, the library was highly abundant, with most of the phage population (~79.6%) consisted of desired library variants while retaining a pre-specified percentage of wildtype and unrecombined phages as controls. ORACLE was thus successfully able to create an unbiased, highly abundant phage library, and was used to create all phage libraries described in **Chapters 4, 5 and 6**.

We used ORACLE to study the tip domain of the T7 phage receptor binding protein (RBP), a key region that determines phage activity. Bacteriophage T7 is a podovirus that infects *E. coli*. As a frequently studied obligate lytic phage (i.e. a phage that does not integrate into its host's genome) T7 is well positioned to be engineered for different roles targeting *E. coli* and related bacteria. T7 phages can also tolerate insertion of additional genetic cassettes and expression of genes can be easily altered, as demonstrated in **Appendix 1** and **Appendix 2**. The RPB of T7 phage consists of six short non-contractile tails that form a homotrimer composed of a rigid shaft ending with a β-sandwich tip domain (Garcia-Doval and Raaij, 2012). The tip domain is likely the very first region of the tail

fiber to interact with host receptor lipopolysaccharide (LPS) and position the phage for successful, irreversible binding with the host (González-García et al., 2015; Molineux, 2001; Qimron et al., 2006). The tip domain is thus a major determinant of host range and activity and is often naturally exchanged between phages to adapt to new hosts (Fraser et al., 2006, 2007; Lin et al., 2012). Substitutions in this region can significantly alter phage activity, and even single amino acid substitutions to this domain are sufficient to alter host range between *E. coli* strains (Heineman et al., 2008). Due to its critical functional role, the tip domain was an excellent choice for comprehensive sequence-function profiling.

Using ORACLE, we created a deep mutational scanning (DMS) phage library of the tip domain of T7 phage to functionally profile this region. These results are documented in our publication in **Chapter 4** (Huss et al., 2021). DMS has been employed to study many proteins including enzymes, transcription factors, signaling domains, and viral surface proteins (Fowler and Fields, 2014; Lee et al., 2018; Raman et al., 2014; Romero et al., 2015). The scale and depth of DMS is used to reveal sites critical for activity, host specificity, and stability in a protein. This phage DMS library consists of a pool of variants, each with a single substitution in the tip domain, encompassing all nineteen non-synonymous and one nonsense substitution at each codon spanning residue positions 472-554 in the RBP (residue numbering based on PDB 4A0T). This library was then selected on three susceptible and two resistant *E. coli* hosts. The relative fitness of each variant was scored by determining the abundance of each variant before and after selection using deep sequencing. Our study identified hundreds of novel function-enhancing substitutions that had not been previously characterized. Using this dataset

for each host we determined what regions were concentrated with activity-enhancing substitutions and host-specific substitutional patterns, many of which were previously unknown.

We then compared results for each host to create a functional profile for the phage, establishing a high-resolution map of which genetic changes drive phage activity. Our unbiased library was essential for this comparison, as many variants highly active on one host performed poorly on others. The functional landscape on each host was unique, reflecting both different molecular preferences of adsorption and the fitness of wildtype T7 on these hosts. Function-enhancing substitutions were densely concentrated in the exterior loops indicating an orientational preference for receptor recognition. However, they were also found on other surface residues, albeit less frequently, suggesting alternative binding modes of the tip domain for host recognition. This functional profile precisely reveals regions that are ideal engineering targets for customizing host range and activity and identifies intolerant residues that should be avoided when engineering synthetic phages.

We next sought to leverage the results of our DMS assay and functional comparisons to create combinatorial libraries to alter phage activity and customize phages for differ roles. We first explored if the functional profile could be leveraged to find sequence motifs that were responsible for receptor recognition in metagenomic datasets. The results of these efforts are documented in **Chapter 5**. Recent years have seen a staggering increase in metagenomic data for bacteriophages. Phage metagenomes has been harvested from diverse sources, including the gut microbiome (Abeles and Pride,

2014; Benler et al., 2021; Shkoporov and Hill, 2019; Shkoporov et al., 2019), from oceans (Brum and Sullivan, 2015; Coutinho et al., 2017; Luo et al., 2022; Winter et al., 2014), and from lakes and soil environments (Bruder et al., 2016; Gu et al., 2021; Pratama and van Elsas, 2018; Wu et al., 2021). These metagenomic datasets represent a treasure trove of sequences that can be mined to engineer new function and alter activity in phages. However, finding sequences that influence phage activity is challenging due to phage diversity, with many sequences in metagenomic databases sharing little to no homology to previously characterized viruses (Coutinho et al., 2017; Fraser et al., 2006; Hatfull, 2008).

To mine the enormous depth of metagenomic databases we thus looked beyond homologous phages and instead to short sequence motifs responsible for critical receptor interactions (Fraser et al., 2006, 2007; Lin et al., 2012). We reasoned that searching for motifs known to be relevant for phage activity was an effective approach to mine metagenomic datasets for such motifs without relying on overall protein homology. To facilitate our search, we established a new process called Motif Curation which leverages functional mapping from our results in **Chapter 4** to find sequence motifs in metagenomic datasets.

Using this dataset, we seeded scores for every possible substitution to determine which theoretical motifs are most relevant for phage activity. We then determined the frequency of the best motifs in NCBI and IMG/VR metagenomic databases to curate a final list of candidate motifs. Motif Curation identified over 10^4 relevant motifs from diverse metagenomic phages located throughout the T7 tip domain. We showed how these motifs

are found in greater abundance in structural proteins and finding motifs requires correctly seeded scores from the DMS set, indicating motifs play a critical functional role. Our process allowed us to identify motifs that are highly divergent from the wildtype sequence which can be sourced from diverse, non-homologous phage sequences.

Screening these phages revealed hundreds of phages with novel host specificity. Furthermore, we saw that metagenomic motifs from distantly related phages primarily drove phage activity, indicating deep mining of metagenomic datasets was essential for tuning host range. We examined position and substitution preference across different hosts to reveal regions in the T7 RBP more tolerant to multiple substitutions and traced key combinations of substitutions that drive activity and host range. We revealed epistatic combinations of substitutions that are individually deleterious but when combined show dramatic increases in activity. Finally, we identified several variants in our library that eliminate pathogenic *E. coli* causing urinary tract infections (UTIs) that are insensitive to wildtype T7, indicating our approach can be used to tune host specificity and increase phage activity on many different hosts.

Results in **Chapter 5** thus lay a foundation for identifying motifs in diverse metagenomic datasets that influence phage activity. This process can be used to mine diverse motifs from non-homologous phage proteins that can tune host range and improve phage activity. These results paint a broader picture of the malleability of the tip domain of the T7 RBP and how changes in the sequence-function landscape facilitate receptor interaction. We envision Motif Curation as a useful tool for tapping into the ever-

increasing wealth of phage metagenomic data to explain sequence-function relationships and engineer phages.

In **Chapter 6** we explored how T7 phage and library variants interacted with *E. coli* BL21 in microgravity aboard the International Space Station. We assessed how wildtype phages were affected by microgravity and found that lysis was delayed in this environment. We incubated phages with hosts over ~23 days and used whole genome sequencing with ~50,000x coverage across the genome to determine which mutations accumulate over time and provide an advantage to the phage. We also evaluated our T7 tip domain DMS library and found a distinct functional profile compared to results seen when incubating in gravity. These results highlight new regions in the phage genome that alter phage activity and new targets for engineering the phage.

In **Chapter 7**, I summarize my findings and propose future directions to improve our platform for engineering and understanding bacteriophages. ORACLE and our DMS approach are well suited to functionally mapping other genes in T7 and in other phages. Natural avenues of research include other structural genes and genes noted to increase activity in microgravity from our results in **Chapter 6**. I also propose evaluating designed combinatorial libraries, combining the most relevant substitutions together directly based on our DMS screen. This approach parallels and complements Motif Curation as a method leveraging our unique understanding of the functional landscape of the tip domain. Several technical aspects of ORACLE can also be improved upon, including recombination efficiency and our approach for minimizing library bias. I describe

experimental approaches that can be used to improve these aspects of ORACLE. Finally, I describe preliminary results for using engineered T7 phages to eliminate two clinically relevant strains, *E. coli* serogroup O121 and multidrug resistant Enterobacter cloacae. Wildtype T7 has limited activity on these strains but our results indicate combinatorial libraries have significantly increased activity on these strains. Using the entirety of our approach to functionally characterize and engineer T7 positions this phage as an excellent chassis phage that can be tailored to eliminate these pathogens.

In conclusion I established ORACLE as a method to create abundant, unbiased libraries of phage variants which can easily be tested in different conditions. We used ORACLE to perform DMS of the tip domain of the RBP of T7 phage and compared these results to create a functional map of the tip domain in different environments. We then used those results to probe metagenomic databases and created combinatorial libraries that drove changes in phage activity and host range. Overall, the platform technology and approaches established in this dissertation can be used to interrogate sequence-function and engineer phages in a high throughput manner.

References

Abeles, S.R., and Pride, D.T. (2014). Molecular Bases and Role of Viruses in the Human Microbiome. J Mol Biol *426*, 3892–3906. https://doi.org/10.1016/j.jmb.2014.07.002.

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A.B., Pevzner, P., and Koonin, E.V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome *9*, 78. https://doi.org/10.1186/s40168-021-01017-w.

Bruder, K., Malki, K., Cooper, A., Sible, E., Shapiro, J.W., Watkins, S.C., and Putonti, C. (2016). Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges. Evol Bioinform Online *12*, 25–33. https://doi.org/10.4137/EBO.S38549.

Brum, J.R., and Sullivan, M.B. (2015). Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat Rev Microbiol *13*, 147–159. https://doi.org/10.1038/nrmicro3404.

Canfield, G.S., and Duerkop, B.A. (2020). Molecular mechanisms of enterococcal-bacteriophage interactions and implications for human health. Curr. Opin. Microbiol. *56*, 38–44. https://doi.org/10.1016/j.mib.2020.06.003.

Chen, M., Zhang, L., Xin, S., Yao, H., Lu, C., and Zhang, W. (2017). Inducible Prophage Mutant of Escherichia coli Can Lyse New Host and the Key Sites of Receptor Recognition Identification. Front. Microbiol. *8*. https://doi.org/10.3389/fmicb.2017.00147.

Clokie, M.R.J., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011). Phages in nature. Bacteriophage *1*, 31–45. https://doi.org/10.4161/bact.1.1.14942.

Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D., Dutilh, B.E., and Thompson, F.L. (2017). Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun *8*, 15955. https://doi.org/10.1038/ncomms15955.

Dedrick, R.M., Guerrero-Bustamante, C.A., Garlena, R.A., Russell, D.A., Ford, K., Harris, K., Gilmour, K.C., Soothill, J., Jacobs-Sera, D., Schooley, R.T., et al. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nature Medicine *25*, 730. https://doi.org/10.1038/s41591-019-0437-z.

Eskenazi, A., Lood, C., Wubbolts, J., Hites, M., Balarjishvili, N., Leshkasheli, L., Askilashvili, L., Kvachadze, L., van Noort, V., Wagemans, J., et al. (2022). Combination

of pre-adapted bacteriophage therapy and antibiotics for treatment of fracture-related infection due to pandrug-resistant Klebsiella pneumoniae. Nat Commun *13*, 302. https://doi.org/10.1038/s41467-021-27656-z.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods *11*, 801–807. https://doi.org/10.1038/nmeth.3027.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507. https://doi.org/10.1016/j.jmb.2006.03.043.

Fraser, J.S., Maxwell, K.L., and Davidson, A.R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr. Opin. Microbiol. *10*, 382–387. https://doi.org/10.1016/j.mib.2007.05.018.

Garcia-Doval, C., and Raaij, M.J. van (2012). Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. PNAS *109*, 9390–9395. https://doi.org/10.1073/pnas.1119719109.

González-García, V.A., Bocanegra, R., Pulido-Cid, M., Martín-Benito, J., Cuervo, A., and Carrascosa, J.L. (2015). Characterization of the initial steps in the T7 DNA ejection process. Bacteriophage *5*. https://doi.org/10.1080/21597081.2015.1056904.

Gu, C., Liang, Y., Li, J., Shao, H., Jiang, Y., Zhou, X., Gao, C., Li, X., Zhang, W., Guo, C., et al. (2021). Saline lakes on the Qinghai-Tibet Plateau harbor unique viral assemblages mediating microbial environmental adaption. IScience *24*, 103439. https://doi.org/10.1016/j.isci.2021.103439.

Hatfull, G.F. (2008). Bacteriophage Genomics. Curr Opin Microbiol *11*, 447–453. https://doi.org/10.1016/j.mib.2008.09.004.

Hatfull, G.F., Dedrick, R.M., and Schooley, R.T. (2022). Phage Therapy for Antibiotic-Resistant Bacterial Infections. Annu Rev Med *73*, 197–211. https://doi.org/10.1146/annurev-med-080219-122208.

Heineman, R.H., Springman, R., and Bull, J.J. (2008). Optimal foraging by bacteriophages through host avoidance. Am. Nat. *171*, E149-157. https://doi.org/10.1086/528962.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. (2021). Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. ELife *10*, e63775. https://doi.org/10.7554/eLife.63775.

Lee, J.M., Huddleston, J., Doud, M.B., Hooper, K.A., Wu, N.C., Bedford, T., and Bloom, J.D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates

of human H3N2 influenza variants. PNAS *115*, E8276–E8285. https://doi.org/10.1073/pnas.1806133115.

Lin, T.-Y., Lo, Y.-H., Tseng, P.-W., Chang, S.-F., Lin, Y.-T., and Chen, T.-S. (2012). A T3 and T7 Recombinant Phage Acquires Efficient Adsorption and a Broader Host Range. PLOS ONE *7*, e30954. https://doi.org/10.1371/journal.pone.0030954.

Luo, E., Leu, A.O., Eppley, J.M., Karl, D.M., and DeLong, E.F. (2022). Diversity and origins of bacterial and archaeal viruses on sinking particles reaching the abyssal ocean. ISME J *16*, 1627–1635. https://doi.org/10.1038/s41396-022-01202-1.

Molineux, I.J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol. Microbiol. *40*, 1–8. https://doi.org/10.1046/j.1365-2958.2001.02357.x.

Mutalik, V.K., and Arkin, A.P. (2022). A Phage Foundry Framework to Systematically Develop Viral Countermeasures to Combat Antibiotic-Resistant Bacterial Pathogens. IScience *25*, 104121. https://doi.org/10.1016/j.isci.2022.104121.

Pratama, A.A., and van Elsas, J.D. (2018). The “Neglected” Soil Virome - Potential Role and Impact. Trends Microbiol *26*, 649–662. https://doi.org/10.1016/j.tim.2017.12.004.

Qimron, U., Marintcheva, B., Tabor, S., and Richardson, C.C. (2006). Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. Proc. Natl. Acad. Sci. U.S.A. *103*, 19039–19044. https://doi.org/10.1073/pnas.0609428103.

Raman, S., Taylor, N., Genuth, N., Fields, S., and Church, G.M. (2014). Engineering Allostery. Trends Genet *30*, 521–528. https://doi.org/10.1016/j.tig.2014.09.004.

Romero, P.A., Tran, T.M., and Abate, A.R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. PNAS *112*, 7159–7164. https://doi.org/10.1073/pnas.1422285112.

Salmond, G.P.C., and Fineran, P.C. (2015). A century of the phage: past, present and future. Nat Rev Microbiol *13*, 777–786. https://doi.org/10.1038/nrmicro3564.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., et al. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrob. Agents Chemother. *61*, e00954-17. https://doi.org/10.1128/AAC.00954-17.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host & Microbe *26*, 527-541.e5. https://doi.org/10.1016/j.chom.2019.09.009.

Winter, C., Garcia, J.A.L., Weinbauer, M.G., DuBow, M.S., and Herndl, G.J. (2014). Comparison of Deep-Water Viromes from the Atlantic Ocean and the Mediterranean Sea. PLOS ONE *9*, e100600. https://doi.org/10.1371/journal.pone.0100600.

Wu, R., Davison, M.R., Nelson, W.C., Graham, E.B., Fansler, S.J., Farris, Y., Bell, S.L., Godinez, I., Mcdermott, J.E., Hofmockel, K.S., et al. (2021). DNA Viral Diversity, Abundance, and Functional Potential Vary across Grassland Soils with a Range of Historical Moisture Regimes. MBio *12*, e0259521. https://doi.org/10.1128/mBio.02595-21.

Chapter 2

Literature Review

High Throughput Approaches to Understand and Engineer Bacteriophages

Authors: Phil Huss[1,2,3,*], Jackie Chen[1,*], and Srivatsan Raman[1,2,4,**]

Contributions:

P.H. and J.C. conceived the review topic, wrote and revised the manuscript

S.R. secured funding, conceived the review topic and revised the manuscript

Affiliations:

[1]Department of Biochemistry, University of Wisconsin-Madison

[2]Department of Bacteriology, University of Wisconsin-Madison

[3]Microbiology Doctoral Training Program, University of Wisconsin-Madison

[4]Department of Chemical and Biological Engineering, University of Wisconsin-Madison

*These authors contributed equally

** Correspondence: sraman4@wisc.edu (Srivatsan Raman)

Abstract

Phages research has been vital to fundamental aspects of modern biology. Advances in metagenomics have revealed treasure troves of new, uncharacterized phages that we have yet to understand. However, our ability to find new phages has outpaced our ability to understand phages. Traditional approaches for characterizing phages are limited in scale and face hurdles determining how changes in sequence drive function. In this review, we describe powerful emerging technologies that can be used to clarify sequence-function relationships in phages through high-throughput genome engineering. Using these approaches up to 10^5 variants can be characterized through pooled selection experiments and deep sequencing. We describe caveats when using these tools and provide examples of basic science and engineering goals pursuable using these approaches.

Keywords

Bacteriophages, Engineered Bacteriophages, Phage-Host Interactions, Deep Sequencing

Highlights

- High throughput approaches can generate large phage libraries that can be screened simultaneously in pooled selection experiments and scored using deep sequencing.
- Approaches for creating phage libraries frequently have a tradeoff between creating mutations throughout larger areas of a phage genome (breadth) versus ability to specify mutations (programmability).
- Untargeted mutagenesis is a simple and effective approach for introducing random mutations and is a useful tool for high-throughput reverse genetics.
- Targeted mutagenesis can characterize phage-host interactions at a high resolution by accurately mapping critical functional regions in a phage genome.
- Phage deletion libraries can be used to study gene essentiality in different environmental and host conditions while insertion libraries can be used to characterize metagenomic gene function.

High Throughput Approaches Enable Interrogating Phage-Host Relationships and Engineering Phages

Understanding the interactions between bacteriophages (or ‘phages’) and bacteria has contributed to major advances in modern biology, including the discovery of polymerases, recombinases, and CRISPR-Cas systems(Salmond and Fineran, 2015). Phages are also increasingly seen as tools for targeted killing of drug-resistant bacteria, precise modulation of the microbiome, gene delivery devices, and diagnostic sensors for pathogens (Eskenazi et al., 2022; Hatfull et al., 2022; Mutalik and Arkin, 2022). Over the decades, phage biologists have painstakingly isolated and characterized thousands of natural phages. Though invaluable in their scope, these studies often lack a molecular understanding of how sequence changes drive phage function. A major impediment to advancing phage biology from empirical observations to molecular function is the lack of high-throughput methods to systematically and comprehensively profile gene-function or variant-function relationships. Traditional phage assays, such as plaque assays, simply do not scale for large functional genomics studies. As a result, functional studies generally are restricted to one or only a few gene or genome variants which leave large swathes of the sequence-function space unexplored.

In this review, we will describe emerging technologies to elucidate sequence-function relationships in phages through high-throughput genome engineering. These approaches enable the functional characterization of up to 10^5 phage variants using pooled selection experiments coupled to deep sequencing and are propelled by a confluence of advances in gene editing, DNA synthesis, and large-scale sequencing.

These functional genomics approaches have been successfully used in bacterial studies and are poised to be implemented in phages. We outline techniques to create diverse **phage libraries** (see Glossary) and explore how these libraries can be selected and scored in different contexts to characterize phage-host interactions, understand phage evolution and uncover gene function. In addition to being a powerful tool to investigate phage biology, these high throughput approaches enable the rational design and engineering of phages with novel properties.

Tools for creating phage libraries frequently have a tradeoff between **breadth** of genome targeting and **programmability** of mutations (Figure 1). For example, chemical mutagenesis and mutagenesis plasmids can introduce mutations across the entire phage genome (higher breadth) but mutations are entirely random (lower programmability). In contrast, specific mutations by oligonucleotide pools (higher programmability) can be inserted into the phage genome but only at a pre-defined locus (lower breadth). Other tools like error-prone PCR or CRISPR **base editors** have greater breadth but lower programmability. The tradeoff between breadth and programmability is thus an important consideration when selecting an approach for creating and scoring phage libraries (Figure 2). In this review, we describe various tools to introduce point mutations, insertions, and deletions on the phage genome and how they relate in breadth and programmability. We motivate the adoption of high-throughput methods by providing examples of basic science questions and engineering goals that can be accomplished using different mutagenesis tools (Figure 3). We conclude with general caveats and considerations when making phage libraries and employing high throughput screens.

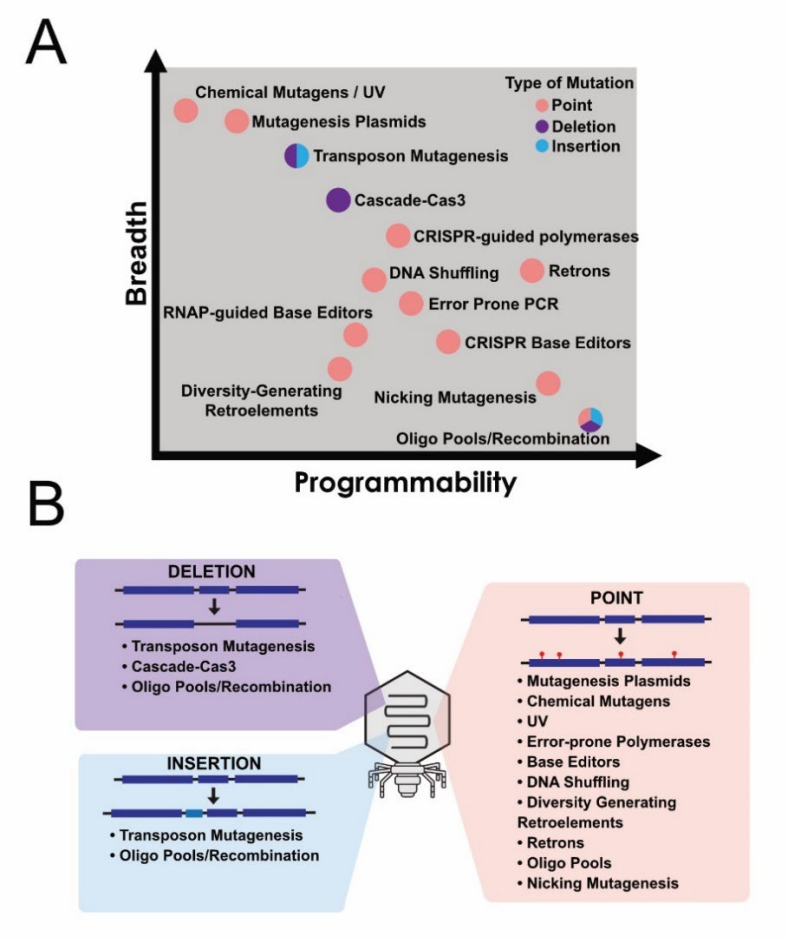


Figure 1. Approaches for creating phage libraries have tradeoffs in breadth and programmability.

(A) Approaches that can create point mutations (pink), deletions (purple) and insertions (blue) in phage libraries and how they compare in ability to create mutations throughout larger areas of a phage genome (breadth) versus the ability to specify mutations (programmability). **(B)** Summary of applicable methods for creating deletions, insertions, and point mutations in phages.

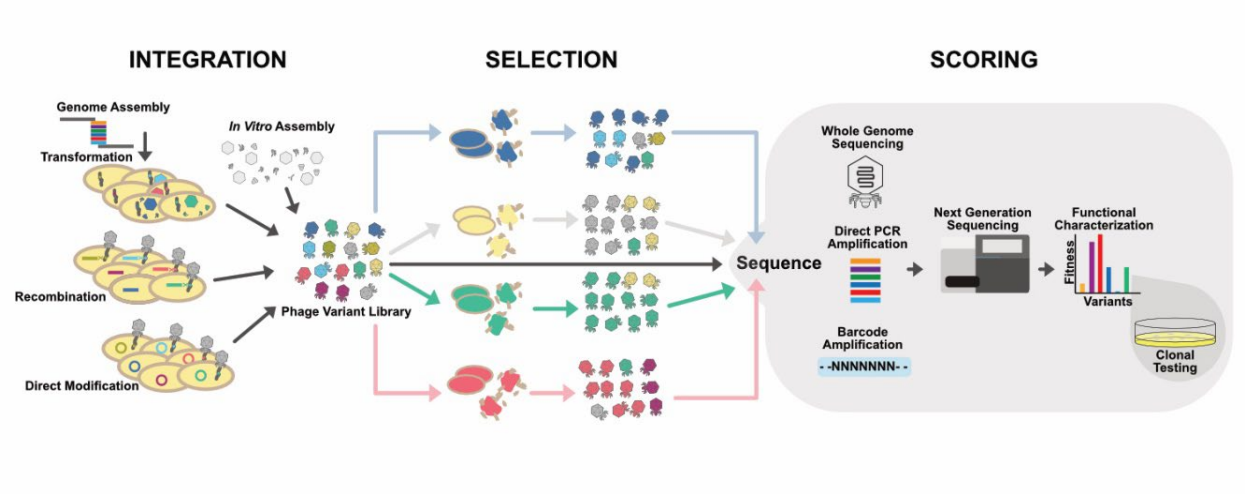


Figure 2. A general outline for integrating, selecting and scoring phage libraries. Various approaches can be used to integrate variants (represented by different color phages) into a pool of phages to create a phage variant library. The phage variant library is selected under different conditions to select for variants with higher fitness in each condition (represented by different color hosts). Selection can be repeated for multiple replication cycles and may include methods for continual diversification of the phage library during selection. All phage populations, including the original phage variant library, are typically deep sequenced to determine the proportion of each variant in each phage pool. This information is used to score variants and determine which variants were more fit in which condition. Sequencing strategies include whole genome sequencing, direct PCR amplification for targeted mutagenesis, and PCR amplification of barcodes previously mapped to phage variants. Deep sequencing ultimately reveals functional characterization that can be further confirmed in clonal testing.

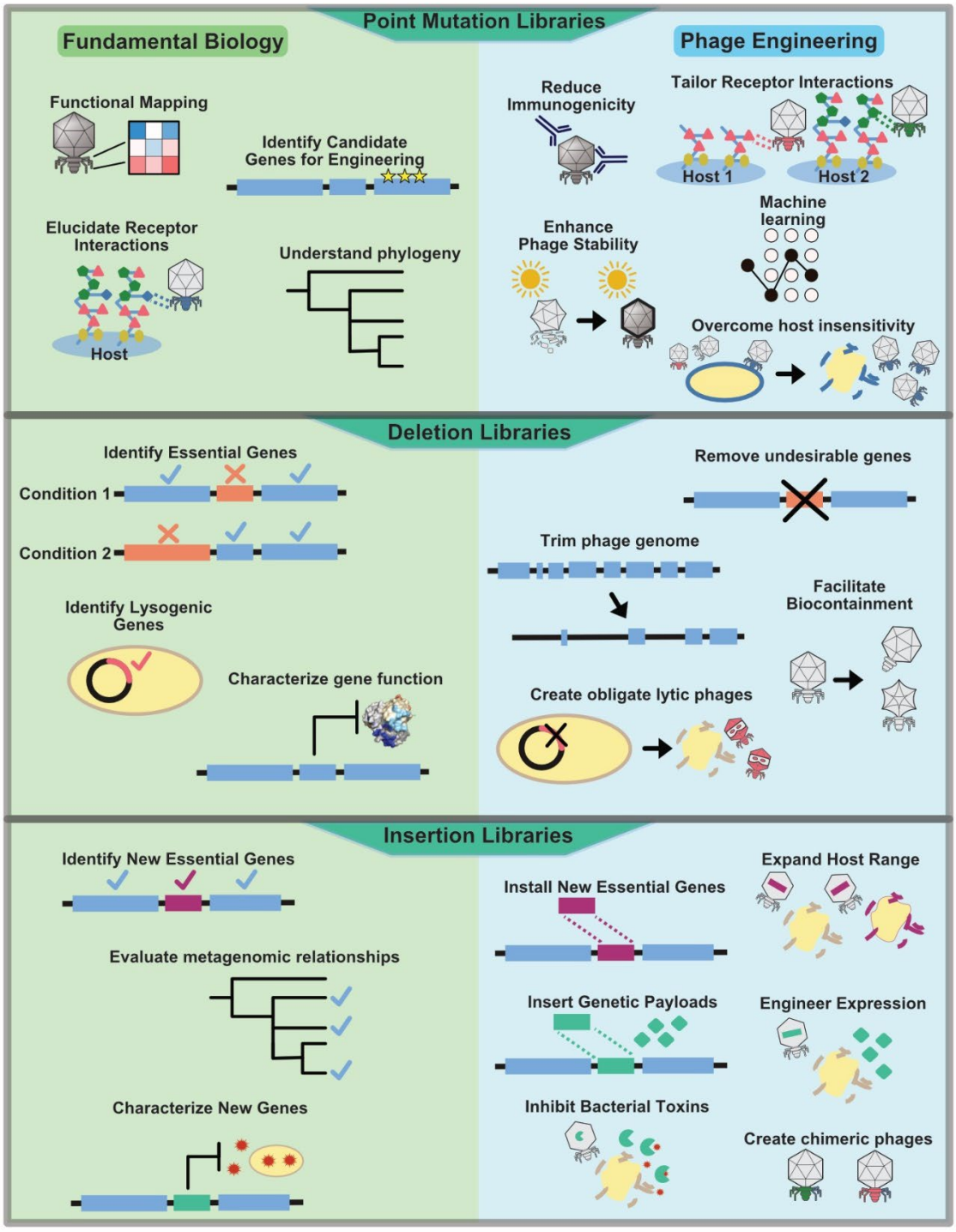

Point Mutation Libraries
Fundamental Biology
Phage Engineering
Functional Mapping
Identify Candidate Genes for Engineering
Elucidate Receptor Interactions
Host
Understand phylogeny
Reduce Immunogenicity
Tailor Receptor Interactions
Host 1
Host 2
Machine learning
Enhance Phage Stability
Overcome host insensitivity
Deletion Libraries
Identify Essential Genes
Condition 1
Condition 2
Identify Lysogenic Genes
Characterize gene function
Remove undesirable genes
Trim phage genome
Facilitate Biocontainment
Create obligate lytic phages
Insertion Libraries
Identify New Essential Genes
Evaluate metagenomic relationships
Characterize New Genes
Install New Essential Genes
Expand Host Range
Insert Genetic Payloads
Engineer Expression
Inhibit Bacterial Toxins
Create chimeric phages

Figure 3. Phage libraries can be used to explore questions in fundamental biology and phage engineering.

Phage libraries can be used to explore diverse applications in fundamental biology (left) and phage engineering (right). Point mutation libraries (top) can be used to map regions relevant for phage function and identify specific residues responsible for host receptor recognition. These approaches can further be leveraged to identify candidate genes for engineering and identify conserved regions to trace phylogeny. This information can be used to engineer phages to reduce host immunogenicity, enhance stability in challenging environments or tailor phages for different receptor interactions. Testing libraries against insensitive hosts may reveal avenues to overcome insensitivity and restore activity. Deletion libraries (middle) can be used to identify essential genes, reveal genes responsible for lysogeny and characterize function by exploring the effects on bacterial hosts. Undesirable genes can be removed, trimming the phage genome to make the phage more efficient, while lysogenic genes can be deleted to create new obligate lytic phages. Alternatively essential genes can be removed to create fully bio-contained phages. Insertion libraries (bottom) can be used to explore roles for metagenomic proteins and identify genes that enable function in new contexts or characterize the role of genes in specific hosts. Insertion of new essential genes can expand host range, or genetic payloads like fluorescent markers, enzymes to degrade bacterial toxins, or other desired genetic cassettes can be added to the phage genome to tailor expression. Alternatively chimeric phages can be constructed by inserting new structural genes into the genome.

Point Mutation Libraries

Methods to introduce point mutations in phages can be broadly grouped into untargeted and targeted approaches.

Untargeted Phage Mutagenesis

Untargeted mutagenesis is a simple and effective approach for introducing random mutations during **continuous evolution** experiments and is particularly useful as a screening tool for high-throughput reverse genetics. Point mutations can be introduced randomly into phages using UV radiation or by chemical mutagens such as ethyl methanesulfonate (EMS) or hydroxylamine (Arimoto-Kobayashi et al., 2000; Favor et al., 2020; Yosef et al., 2017). Radiation and chemical mutagenesis rely on damaging phage DNA which can cause single nucleotide changes during DNA replication. The frequency of mutagenesis can be tuned in a dose-dependent manner. Chemical mutagenesis, in particular, is relatively easy to implement in any laboratory setting but remains an underutilized tool to study laboratory evolution of phages. Radiation and chemical mutagens can generate **unbiased** libraries as they do not require passaging in a replication host. Although the mutations are randomly dispersed genome-wide, the nature of mutations is only semi-random as specific transitions or transversion mutations are favored depending on the mutagen(Li et al., 2016a; Shibai et al., 2017). UV radiation may also impact the survivability of phages by destabilizing structural proteins by altering protein conformations(Santos et al., 2013). An alternative untargeted approach is to use mutagenesis plasmids in the replication host to induce genome-wide mutations through

the expression of proteins that damage DNA or reduce DNA replication fidelity (e.g Phage-assisted continuous evolution) (Badran and Liu, 2015; Esvelt et al., 2011; Miller et al., 2020). However, mutagenesis plasmids require passaging on a replication host, which limits this approach to hosts which can sustain such plasmids and biases phage libraries toward mutations that permit phage variants to grow on that host.

Variant libraries generated by untargeted mutagenesis have high breadth but virtually no programmability. Libraries are highly diverse but quantifying this diversity will require considerable **sequencing depth** as the frequency of any particular mutation may be low in the variant population. Despite low initial frequency, mutations that affect a specific trait are likely to exist in the variant population and can rapidly emerge during laboratory selection, making this a powerful approach for continuous evolution. For example, continuous evolution paired with EMS mutagenesis has been used to identify phage variants with improved thermal stability(Favor et al., 2020). Due to the preponderance of additional mutations, multiple parallel evolution experiments may be needed to identify mutations that cause a specific phenotype.

Targeted Phage Mutagenesis

Targeted mutagenesis allows deeper exploration of sequence space at a specific genetic locus or loci by concentrating mutations in a defined window. While targeted mutagenesis could be used to discover gene function, it is a particularly effective tool for investigating the effects of sequence variants when the functional role of a gene is known *a priori* (e.g: genes encoding phage receptor interaction). Targeted genome variants can

be created by introducing an externally created library into the phage genome at a specific site or by directly editing the phage genome using accessory proteins.

To generate a library of randomized phage gene variants, we can use common techniques such as error-prone PCR, **DNA shuffling**, **nicking mutagenesis**, and degenerate or random primers(Dunne et al., 2019; Stemmer, 1994; Wrenbeck et al., 2016; Yehl et al., 2019). Alternatively, pre-specified variants can be synthesized using commercially available oligonucleotide pools(Huss et al., 2021). These methods are easy-to-implement, guarantee diversity, and generate large variant libraries with high programmability. However, the challenge lies in the efficient integration of the library into the phage genome which can limit breadth. Site-specific recombinases (Huss et al., 2021), lambda Red recombineering (Jensen et al., 2020; Oppenheim et al., 2004) and homologous recombination (Hoshiga et al., 2019; Marinelli et al., 2019; Yehl et al., 2019) have been employed for genome integration.

The phage genome can be directly mutated *in vivo* using CRISPR systems that localize mutagenic proteins to specific sites. For example, cytidine and adenine base editors can be tethered to or recruited by dead Cas9 or Cas9 **nickase** to mutate specific loci(Álvarez et al., 2020; Gaudelli et al., 2017; Hess et al., 2016; Komor et al., 2016; Moore et al., 2018; Zheng et al., 2018). Cas protein-linked error-prone DNA polymerases with nick translation and/or strand displacement capabilities can be used to introduce every type of point mutation in a defined window(Halperin et al., 2018). Base editors attached to RNA polymerase can be used to mutate larger regions of a genome engineered with appropriate promoter and terminator sequences flanking the target region(Cravens et al.,

2021), although the spectrum of induced mutations is currently limited. Recently, bacterial retroelements (retrons) have been used to genetically engineer T5 phage (Ramirez-Chamorro et al., 2021). Retrons are polycistrons consisting of a reverse transcriptase and covalently linked ssRNA and ssDNA which induce site-specific mutagenesis through a mechanism still yet to be fully elucidated. Libraries of these retroelements can be synthesized *in vitro* or mutated *in vivo* for continuous mutagenesis of target sequence(Simon et al., 2018). Likewise lysogenic phage BPP-1 leverages diversity generating retroelements encoded in its genome to introduce targeted hypervariability via **mutagenic homing** to its receptor binding protein(Medhekar and Miller, 2007). Such a system could feasibly be adapted to mutate heterologous genes in different phages(Guo et al., 2011).

Targeted mutagenesis *in vivo* requires a suitable bacterial host to sustain variant libraries or mutator protein systems. Though versatile, CRISPR approaches are constrained by the location of PAM sites for targeting guide RNA (gRNA), and the difficulty in multiplexing gRNAs to simultaneously edit several loci or to expand the window of activity. The diversity generated by CRISPR guided systems and retrons may also vary considerably, with reported rates of 10^{-5}-10^{-6} mutations/bp/generation(Halperin et al., 2018; Kantor et al., 2020; Zheng et al., 2018). Recombination efficiency also varies and may require engineering phages to enable recombination, with reported recombination rates ranging from 10^{-10} to 10^{-3}(Huss et al., 2021; Pires et al., 2016; Yehl et al., 2019). Lower rates of recombination or mutation incorporation may impact **library abundance**, decreasing the total number of variants compared to unrecombined phages. One

approach to increase library abundance is to use counterselection with CRISPR to remove unrecombined or unmutated phages from the phage population (Grigonyte et al., 2020; Huss et al., 2021; Nayeemul Bari et al., 2017). Lytic phages may also require multiple passages on the host used to create the library, biasing the library towards variants capable of productively infecting that host. To minimize bias, a 'helper' plasmid expressing the wildtype gene was used to mask the effect of variants during library generation in the **ORACLE** system(Huss et al., 2021).

Alternatively, some phage genomes can be wholly constructed outside of bacterial hosts. Phage genomes can be assembled in yeast using yeast artificial chromosomes(Ando et al., 2015a; Jaschke et al., 2012), or using various ligation cloning methods *in vitro*(Andrews and Fields, 2021; Dunne et al., 2019; Kilcher et al., 2018a; Pryor et al., 2022; Pulkkinen et al., 2019). Assembled genomes or genome fragments can then be transformed into bacterial hosts(Ando et al., 2015a; Assad-Garcia et al.; Faber et al., 2020; Pulkkinen et al., 2019) to '**reboot**' the phage or packaged into viable phages particles in cell-free systems (Andrews and Fields, 2021; Rustad et al., 2017; Shin et al., 2012). Such *ex vivo* approaches can maintain high variant abundance and limit library bias as they are not reliant on multiple passages in bacteria. However, these methods are limited to phages that can 'reboot' after transformation and by the need for established cell-free systems or bacterial hosts capable of efficient transformation of large DNA constructs, which may limit library size.

Applications of Targeted Phage Mutagenesis Libraries

Targeted mutation libraries can help characterize phage-host interactions and engineer phages at a high resolution by accurately mapping critical functional regions in a phage gene. For example, deep mutational scanning (DMS) of the phage receptor-binding protein and subjecting the variant pool to selection on different hosts(Huss et al., 2021) revealed mutations that play a role in host receptor recognition. Similarly, molecular rules of other phage properties such as capsid assembly, replication fidelity, or host lysis could be probed by DMS of relevant gene(s). This knowledge can be extended to create tailored libraries where variants have been designed for specific purposes, like targeting or avoiding specific hosts, reducing immunogenicity or increasing phage stability in unfavorable physicochemical conditions (e.g: gut environment). Recent advances in machine learning are revolutionizing our ability to decipher the complex relationships between protein sequence, structure, and function(Gelman et al., 2021; Wu et al., 2019). Deep learning models enable us to explore vast sequence spaces to predict sequences that are highly optimized for novel function. Large phage mutation-function datasets are a rich resource to train machine learning models to understand the complex relationship between phage sequences and function and to design novel phage variants with desired properties.

Phage Insertion and Deletion Libraries

Targeted deletions or insertions in phage genomes evaluates the functional impact of larger genomic perturbations ranging from tens of bases to kilobases. Smaller deletions

of tens to hundreds of bases can be created *in vivo* by expressing Cas9 in the replication host and larger deletions up to several kilobases can be generated *in vivo* using Cascade-Cas3 which degrades DNA using dual helicase-nuclease activity (Csörgő et al., 2020). Transposon mutagenesis with site-specific *mariner* class transposons or random insertion transposons such as Tn5, which have been used extensively to create genome-wide loss-of-function libraries in bacteria (Goodman et al., 2009; Langridge et al., 2009; Mutalik et al., 2020; van Opijnen et al., 2009), could be implemented in phages *in vivo* to create gene deletion libraries. Alternatively, transposon mutagenesis can be used to generate more programmable single, double, and triple residue deletions in specific genes. These insertion or deletion libraries can be synthesized in oligonucleotide pools and recombined into the phage genome(Emond et al., 2020) using site-specific recombinases or by homologous recombination(Jones et al., 2008; Mutalik et al., 2020). Cas-based editing requires the user to direct insertions or deletions to defined regions of the genome while transposon-based editing is broader and can cover the entire genome. Barcodes may be inserted into the donor or transposon sequence to score the genome perturbations using short read sequencing(Wetmore et al., 2015). Insertion of sequences encoding sequence motifs to small protein domains can be created but currently suffer tradeoffs between library size, genome insertion efficiency, and cost. Large, diverse libraries can be created from oligonucleotide pools at relatively low cost, but the efficiency of genome insertion into the phage genome is the limiting factor.

Applications of Phage Deletion and Insertion Libraries

Phage deletion libraries are helpful to study gene essentiality in different environmental and host conditions. When the phage deletion library is subjected to selection under a certain condition, variants with deletions in genes essential in that condition will deplete. Since phage genomes have evolved to maintain overlapping regulation, this approach can be further extended to study regulatory networks in the phage genome. Lysogenic phages which are disfavored for therapeutic use could be converted into obligate lytic phages by the systematic removal of recombinases, transcription factors or other unknown mechanisms that regulate lysogeny (Kilcher et al., 2018a). Deletions may also improve the activity of phage in different conditions, as some genetic elements may only be beneficial in specific contexts and are otherwise disadvantageous to the phage, such as gene 1.7 in T7 phage which confers sensitivity to dideoxythymidine(Tran et al., 2008).

Deletion libraries are helpful for phage engineering. Mapping essential genes can be used to create synthetic phages with a minimal genome(Pires et al., 2021), whose activity may be higher than wildtype. Host range can be customized in engineered phages by removing essential genes for one host but not another. Targeted deletions of key genes can be used as a strategy to improve the biocontainment of phages, as deletion of essential genes restricts phage replication.

Larger-sized insertions can be used to characterize metagenomic gene function. While phage genome databases are rapidly growing, tools to characterize functions of these sequences has lagged far behind. Libraries of metagenomic sequences can be

rapidly tested by inserting them into the phage genome. For example, exchanging receptor binding proteins between different phages can show evolutionary relationships between phages or identify new phage chimeras that are able to infect novel hosts. Exchanging or inserting phage genes predicted to deter host defenses, like Cas inhibitors, would allow us to study which hosts harbor which defense mechanisms(Landsberger et al., 2018).

Phage genomes could be augmented with new capabilities (e.g: new lysins) by inserting characterized genes into non-essential regions. Addition of host genes that are essential to the phage, like trxA for T7 phage, removes reliance on the host, while insertion of genes that deter host defenses like Cas inhibitors can improve the ability of engineered phages to target and/or eliminate new hosts (Grigonyte et al., 2020; Landsberger et al., 2018; Qimron et al., 2006). The insertion of larger sequences can enable tailored protein production during the phage lifecycle, such as fluorescent markers, or can be used to inactivate undesirable proteins made by the bacterial host, such as toxins. Gene cassettes that facilitate host killing may be incorporated into phages, such as Crispr-Cas elements that target the host genome, ensuring host cell destruction(Selle et al.).

Considerations for Creating and Screening Phage Libraries

Any approach used to create phage libraries should account for library bias. Bias is the reduction or elimination of phage variants due to selection pressure during library formation. Bias frequently occurs during passage on a replication host. Variants lost to

library bias may be active on different hosts or in different conditions. These variants are likely to be the most relevant to understanding or engineering that phage, as they necessarily can affect the activity and host range of that phage. Bias can be attenuated by limiting rounds of replication or by using a helper plasmid for libraries of specific genes. *In vitro* assembly of phage libraries avoids selection bias from replication hosts but could introduce bias if altered structural proteins assemble differently *in vivo* versus *in vitro*. Methods for scoring the phage library should be able to take the extent of bias into account. If the library is sequenced prior to selection, it is critical that the library be assessed after any bias may have altered library distribution. For example, sequencing a plasmid library before integration into a phage genome may produce dramatically different results than sequencing the phage library after integration. If the library is not assessed prior to selection, possible uncharacterized bias should be considered when analyzing results.

Library abundance should be evaluated to ensure that an experimentally appropriate number of unique phage variants are used during selection. Experimental consistency is likely to be improved if there are hundreds or thousands of a particular variant during selection. Reduced phage-host ratios during selection may result in complete attenuation of some variants in the phage pool. If a library has been biased during creation, some members may have much lower proportional representation than other members. Some methods such as homologous recombination result in an extreme overabundance of wild type phages. This can dramatically reduce the number of relevant variants in the phage pool and add significant noise to the experiment. Wild type can be

reduced using counterselection using CRISPR during passage (Grigonyte et al., 2020; Huss et al., 2021; Nayeemul Bari et al., 2017) or by using techniques that avoid wild type overrepresentation.

Variants that dramatically outperform other library members may overtake the phage population over a few rounds of replication. This effect can semi-randomly mask variants with less pronounced effects that are still be valuable for understanding phage host interactions or engineering the phage. Comparison of biological replicates can reveal this effect, which can be reduced by lowering the number of passages used for selection.

Libraries are sequenced to score phage variants, ideally before and after selection. The sequencing depth necessary should be based on the ability to detect the lowest abundant variants. Low sequencing coverage or low initial abundance can dramatically alter the limit of detection for the assay and this effect must be kept in mind when analyzing results. Sequencing and scoring variants can rely on whole genome sequencing or sequencing of specific target regions. Long read sequencing such as PacBio or Oxford Nanopore and short read sequencing such as Illumina platforms are viable options. Longer reads come at the cost of depth, as short read sequencing can sequence millions of times (Hu et al., 2021). Targeted mutation libraries or short, ~500 bp windows are easiest to sequence and score as they can be directly sequenced using abundant short reads. For larger windows and changes which surpass ~500 bp, more costly long read sequencing is generally necessary. Alternatively, short nucleotide barcodes can be integrated in tandem with the variant sequence. Strategies include integrating barcodes in a known region alongside the variant(Andrews and Fields, 2021),

or barcodes could replace deleted regions(Wetmore et al., 2015). This barcode can be sequenced as a proxy readout for the mutated region after mapping unique barcodes to individual variants using long read sequencing, after which mutants can be scored using less expensive short read sequencing. These strategies which use targeted deep sequencing of variants enable researchers to obtain quantitative fitness data for thousands, if not millions, of phage variants, exceeding the throughput of clonal testing by several orders of magnitude.

Concluding Remarks

Over the decades our understanding of how phages and bacteria interact has continued to evolve. These studies have already revealed a wealth of novel biologic functions that are fundamental to many fields like polymerases, recombinases, and Crispr-Cas systems. Our increasing understanding of phages has positioned phages as potent tools for precision editing of complex microbial communities and as treatments for drug-resistant bacterial infections. Still, advances in metagenomics have revealed many new phages that remain uncharacterized, and many fundamental rules of phage-host interactions remain to be investigated. High throughput approaches described in this review enable systematic characterization of phages to establish sequence-function relationships. While many questions remain as to the best approach to implement these techniques in phages (see Outstanding Questions), these approaches present exciting new avenues for research to enhance our understanding of these viruses so essential to our ecosystem and our study of biology.

Acknowledgements

This work was supported by NIAID grant R21AI156785 (to S.R.) and the National Institute of General Medical Sciences of the National Institute of Health under Award Number T32GM135066 (to J.C.).

Declarations of Interest

No interests are declared.

Outstanding Questions Box

- Continuous evolution experiments can frequently introduce variants that are immediately selected against. These variants are difficult to detect and score. How can those variants that do poorly be identified in these experiments?
- How can we link activity of specific phage variants to different members of complex microbial communities?
- How can information from screens be used to create machine learning models to tailor phage function?
- Can deep sequencing be used to mine for related phages in complex metagenomic databases?
- What generalizable approaches can be developed to limit bias when creating phage libraries?
- How do rates of mutagenesis and recombination rates differ for phage genomes compared to bacterial genomes?
- How can mutagenesis and recombination rates be optimized within the setting of an active phage infection?

Glossary

Base Editors: proteins which introduce single nucleotide alteration without forming double-stranded breaks, typically by damaging DNA through deamination.

Breadth: The ability of an approach to create mutations throughout larger areas of a phage genome. High breadth indicates mutations can be incorporated over much of the phage genome, while low breadth indicates an approach is restricted to smaller regions.

Continuous Evolution: An experimental approach whereby phage diversification and selection occur uninterrupted over multiple rounds of replication. The approach for mutating phages may be included in every round of replication to allow for increased diversification.

DNA Shuffling: Approach for generated diversity in a gene by treatment with restriction enzymes and reassembly by PCR without primers or through nonhomologues random recombination.

Library Abundance: The proportion of phages in a phage population that are the actual library variants. For example, a library is 1% abundant if 99% of a phage population is wild type and 1% of phages are the library phage variants.

Library Bias: The degree to which phage variants have been proportionally reduced or eliminated in a phage library due to selection pressure during library formation, skewing the phage library towards variants more fit to the condition in which the library was created. Bias can eliminate variants that are viable in other conditions and alter interpretation of results if unaccounted for.

Mutagenic homing: Transposition of a mutated genetic element into a targeted region, where the mutations found in the transposed element are inherited by the target sequence.

Nickase: A Cas9 variant with one of the two nuclease domains inactivated. Instead of generating double-stranded breaks at the target site, these enzymes introduce a single-stranded nick.

Nicking mutagenesis: PCR-based method which generates diversity at target loci on a dsDNA plasmid using mutagenic oligos where wildtype DNA strands are successively degraded with nicking enzymes and exonuclease treatment.

ORACLE: A method (Optimized Recombination, Accumulation and Library Expression) designed to create unbiased phage libraries using a combination of site-specific recombination, CRISPR counterselection and helper plasmids.

Phage Library: A pool of phages which contains all the phage variants to be evaluated in a mixed population.

Programmability: The ability of an approach to create specific, desired variants in a phage library. High programmability indicates precise control over which mutations are incorporated into a library, while low programmability indicates no control over which mutations are incorporated.

Rebooting: Process by which phage genomes are directly transformed into compatible hosts to initiate infection and create viable phage particles.

Sequencing Depth: The number of times a given nucleotide position is read during deep sequencing when scoring libraries. Every different read is presumed to be a different sequence from a different phage. For example, if a position has a sequencing depth of 1000 it is presumed 1000 different phages have been sequenced at that position.

Bibliography

Abedon, S.T. (2011). Lysis from without. Bacteriophage *1*, 46–49. https://doi.org/10.4161/bact.1.1.13980.

Álvarez, B., Mencía, M., de Lorenzo, V., and Fernández, L.Á. (2020). In vivo diversification of target genomic sites using processive base deaminase fusions blocked by dCas9. Nat Commun *11*, 6436. https://doi.org/10.1038/s41467-020-20230-z.

Ando, H., Lemire, S., Pires, D.P., and Lu, T.K. (2015a). Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. Cell Systems *1*, 187–196. https://doi.org/10.1016/j.cels.2015.08.013.

Ando, H., Lemire, S., Pires, D.P., and Lu, T.K. (2015b). Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. Cell Syst *1*, 187–196. https://doi.org/10.1016/j.cels.2015.08.013.

Andrews, B., and Fields, S. (2021). Balance between promiscuity and specificity in phage λ host range. ISME J *15*, 2195–2205. https://doi.org/10.1038/s41396-021-00912-2.

Arimoto-Kobayashi, S., Anma, N., Yoshinaga, Y., Douki, T., Cadet, J., and Hayatsu, H. (2000). Oxidative damage and induced mutations in M13mp2 phage DNA exposed to N-nitrosopyrrolidine with UVA radiation. Mutagenesis *15*, 473–477. https://doi.org/10.1093/mutage/15.6.473.

Arthur, M., Arbeit, R.D., Kim, C., Beltran, P., Crowe, H., Steinbach, S., Campanelli, C., Wilson, R.A., Selander, R.K., and Goldstein, R. (1990). Restriction fragment length polymorphisms among uropathogenic Escherichia coli isolates: pap-related sequences compared with rrn operons. Infect Immun *58*, 471–479. .

Assad-Garcia, N., D'Souza, R., Buzzeo, R., Tripathi, A., Oldfield, L.M., Vashee, S., and Fouts, D.E. Cross-Genus "Boot-Up" of Synthetic Bacteriophage in Staphylococcus aureus by Using a New and Efficient DNA Transformation Method. Applied and Environmental Microbiology *88*, e01486-21. https://doi.org/10.1128/AEM.01486-21.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol *2*, 2006.0008. https://doi.org/10.1038/msb4100050.

Badran, A.H., and Liu, D.R. (2015). Development of potent in vivo mutagenesis plasmids with broad mutational spectra. Nat Commun *6*, 8425. https://doi.org/10.1038/ncomms9425.

Bauer, D.F. (1972). Constructing Confidence Sets Using Rank Statistics. Journal of the American Statistical Association *67*, 687–690. https://doi.org/10.1080/01621459.1972.10481279.

Bertozzi Silva, J., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. FEMS Microbiol Lett *363*. https://doi.org/10.1093/femsle/fnw002.

Canfield, G.S., and Duerkop, B.A. (2020). Molecular mechanisms of enterococcal-bacteriophage interactions and implications for human health. Curr. Opin. Microbiol. *56*, 38–44. https://doi.org/10.1016/j.mib.2020.06.003.

Chen, M., Zhang, L., Xin, S., Yao, H., Lu, C., and Zhang, W. (2017). Inducible Prophage Mutant of Escherichia coli Can Lyse New Host and the Key Sites of Receptor Recognition Identification. Front. Microbiol. *8*. https://doi.org/10.3389/fmicb.2017.00147.

Chen, Z., Guo, L., Zhang, Y., Walzem, R.L., Pendergast, J.S., Printz, R.L., Morris, L.C., Matafonova, E., Stien, X., Kang, L., et al. (2014). Incorporation of therapeutically modified bacteria into gut microbiota inhibits obesity. J. Clin. Invest. *124*, 3391–3406. https://doi.org/10.1172/JCI72517.

Clokie, M.R.J., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011). Phages in nature. Bacteriophage *1*, 31–45. https://doi.org/10.4161/bact.1.1.14942.

Cravens, A., Jamil, O.K., Kong, D., Sockolosky, J.T., and Smolke, C.D. (2021). Polymerase-guided base editing enables in vivo mutagenesis and rapid protein engineering. Nat Commun *12*, 1579. https://doi.org/10.1038/s41467-021-21876-z.

Csörgő, B., León, L.M., Chau-Ly, I.J., Vasquez-Rifo, A., Berry, J.D., Mahendra, C., Crawford, E.D., Lewis, J.D., and Bondy-Denomy, J. (2020). A compact Cascade–Cas3 system for targeted genome engineering. Nat Methods *17*, 1183–1190. https://doi.org/10.1038/s41592-020-00980-w.

Dedrick, R.M., Guerrero-Bustamante, C.A., Garlena, R.A., Russell, D.A., Ford, K., Harris, K., Gilmour, K.C., Soothill, J., Jacobs-Sera, D., Schooley, R.T., et al. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nature Medicine *25*, 730. https://doi.org/10.1038/s41591-019-0437-z.

Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M.J., et al. (2019). Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Reports *29*, 1336-1350.e4. https://doi.org/10.1016/j.celrep.2019.09.062.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., et al. (2008). The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. J Bacteriol *190*, 2597–2606. https://doi.org/10.1128/JB.01695-07.

Edgar, R.C., and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics *31*, 3476–3482. https://doi.org/10.1093/bioinformatics/btv401.

Elena, S.F., Agudelo-Romero, P., and Lalić, J. (2009). The Evolution of Viruses in Multi-Host Fitness Landscapes. Open Virol J *3*, 1–6. https://doi.org/10.2174/1874357900903010001.

Emond, S., Petek, M., Kay, E.J., Heames, B., Devenish, S.R.A., Tokuriki, N., and Hollfelder, F. (2020). Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. Nat Commun *11*, 3469. https://doi.org/10.1038/s41467-020-17061-3.

Eskenazi, A., Lood, C., Wubbolts, J., Hites, M., Balarjishvili, N., Leshkasheli, L., Askilashvili, L., Kvachadze, L., van Noort, V., Wagemans, J., et al. (2022). Combination of pre-adapted bacteriophage therapy and antibiotics for treatment of fracture-related infection due to pandrug-resistant Klebsiella pneumoniae. Nat Commun *13*, 302. https://doi.org/10.1038/s41467-021-27656-z.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011). A System for the Continuous Directed Evolution of Biomolecules. Nature *472*, 499–503. https://doi.org/10.1038/nature09929.

Faber, M.S., Van Leuven, J.T., Ederer, M.M., Sapozhnikov, Y., Wilson, Z.L., Wichman, H.A., Whitehead, T.A., and Miller, C.R. (2020). Saturation Mutagenesis Genome Engineering of Infective ΦX174 Bacteriophage via Unamplified Oligo Pools and Golden Gate Assembly. ACS Synth Biol *9*, 125–131. https://doi.org/10.1021/acssynbio.9b00411.

Favor, A.H., Llanos, C.D., Youngblut, M.D., and Bardales, J.A. (2020). Optimizing bacteriophage engineering through an accelerated evolution platform. Sci Rep *10*, 13981. https://doi.org/10.1038/s41598-020-70841-1.

Fister, S., Robben, C., Witte, A.K., Schoder, D., Wagner, M., and Rossmanith, P. (2016). Influence of Environmental Factors on Phage–Bacteria Interaction and on the Efficacy and Infectivity of Phage P100. Front Microbiol *7*. https://doi.org/10.3389/fmicb.2016.01152.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods *11*, 801–807. https://doi.org/10.1038/nmeth.3027.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507. https://doi.org/10.1016/j.jmb.2006.03.043.

Fraser, J.S., Maxwell, K.L., and Davidson, A.R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr. Opin. Microbiol. *10*, 382–387. https://doi.org/10.1016/j.mib.2007.05.018.

Garcia, E., Elliott, J.M., Ramanculov, E., Chain, P.S.G., Chu, M.C., and Molineux, I.J. (2003). The genome sequence of Yersinia pestis bacteriophage phiA1122 reveals an intimate history with the coliphage T3 and T7 genomes. J. Bacteriol. *185*, 5248–5262. .

Garcia-Doval, C., and Raaij, M.J. van (2012). Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. PNAS *109*, 9390–9395. https://doi.org/10.1073/pnas.1119719109.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464–471. https://doi.org/10.1038/nature24644.

Gebhart, D., Williams, S.R., and Scholl, D. (2017). Bacteriophage SP6 encodes a second tailspike protein that recognizes Salmonella enterica serogroups C2 and C3. Virology *507*, 263–266. https://doi.org/10.1016/j.virol.2017.02.025.

Gelman, S., Fahlberg, S.A., Heinzelman, P., Romero, P.A., and Gitter, A. (2021). Neural networks to learn protein sequence–function relationships from deep mutational scanning data. Proceedings of the National Academy of Sciences *118*, e2104878118. https://doi.org/10.1073/pnas.2104878118.

Gietz, R.D., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Meth. Enzymol. *350*, 87–96. .

González-García, V.A., Bocanegra, R., Pulido-Cid, M., Martín-Benito, J., Cuervo, A., and Carrascosa, J.L. (2015). Characterization of the initial steps in the T7 DNA ejection process. Bacteriophage *5*. https://doi.org/10.1080/21597081.2015.1056904.

Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R., and Gordon, J.I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe *6*, 279–289. https://doi.org/10.1016/j.chom.2009.08.003.

Grigonyte, A.M., Harrison, C., MacDonald, P.R., Montero-Blay, A., Tridgett, M., Duncan, J., Sagona, A.P., Constantinidou, C., Jaramillo, A., and Millard, A. (2020). Comparison of CRISPR and Marker-Based Methods for the Engineering of Phage T7. Viruses *12*, 193. https://doi.org/10.3390/v12020193.

Guo, H., Tse, L.V., Nieh, A.W., Czornyj, E., Williams, S., Oukil, S., Liu, V.B., and Miller, J.F. (2011). Target Site Recognition by a Diversity-Generating Retroelement. PLOS Genetics *7*, e1002414. https://doi.org/10.1371/journal.pgen.1002414.

Halperin, S.O., Tou, C.J., Wong, E.B., Modavi, C., Schaffer, D.V., and Dueber, J.E. (2018). CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. Nature *560*, 248–252. https://doi.org/10.1038/s41586-018-0384-8.

Hatfull, G.F., Dedrick, R.M., and Schooley, R.T. (2022). Phage Therapy for Antibiotic-Resistant Bacterial Infections. Annu Rev Med *73*, 197–211. https://doi.org/10.1146/annurev-med-080219-122208.

Heineman, R.H., Springman, R., and Bull, J.J. (2008). Optimal foraging by bacteriophages through host avoidance. Am. Nat. *171*, E149-157. https://doi.org/10.1086/528962.

Hess, G.T., Frésard, L., Han, K., Lee, C.H., Li, A., Cimprich, K.A., Montgomery, S.B., and Bassik, M.C. (2016). Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat Methods *13*, 1036–1042. https://doi.org/10.1038/nmeth.4038.

Hollander, M., and Wolfe, D.A. (1973). Nonparametric Statistical Methods (Wiley).

Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (2020). A continuous evolution system for contracting the host range of bacteriophage T7. Scientific Reports *10*, 1–8. https://doi.org/10.1038/s41598-019-57221-0.

Hoshiga, F., Yoshizaki, K., Takao, N., Miyanaga, K., and Tanji, Y. (2019). Modification of T2 phage infectivity toward Escherichia coli O157:H7 via using CRISPR/Cas9. FEMS Microbiology Letters *366*, fnz041. https://doi.org/10.1093/femsle/fnz041.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. Human Immunology *82*, 801–811. https://doi.org/10.1016/j.humimm.2021.02.012.

Huss, P., and Raman, S. (2020). Engineered bacteriophages as programmable biocontrol agents. Curr Opin Biotechnol *61*, 116–121. https://doi.org/10.1016/j.copbio.2019.11.013.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. (2021). Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. ELife *10*, e63775. https://doi.org/10.7554/eLife.63775.

Hyman, P., and Abedon, S.T. (2010). Bacteriophage Host Range and Bacterial Resistance. Advances in Applied Microbiology *70*, 217–248. https://doi.org/10.1016/S0065-2164(10)70007-1.

Jaschke, P.R., Lieberman, E.K., Rodriguez, J., Sierra, A., and Endy, D. (2012). A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. Virology *434*, 278–284. https://doi.org/10.1016/j.virol.2012.09.020.

Jensen, J.D., Parks, A.R., Adhya, S., Rattray, A.J., and Court, D.L. (2020). λ Recombineering Used to Engineer the Genome of Phage T7. Antibiotics (Basel) *9*, 805. https://doi.org/10.3390/antibiotics9110805.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotechnology *31*, 233–239. https://doi.org/10.1038/nbt.2508.

Jones, G.M., Stalker, J., Humphray, S., West, A., Cox, T., Rogers, J., Dunham, I., and Prelich, G. (2008). A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae. Nat Methods *5*, 239–241. https://doi.org/10.1038/nmeth.1181.

de Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., and Dutilh, B.E. (2019). Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends in Microbiology *27*, 51–63. https://doi.org/10.1016/j.tim.2018.08.006.

Kantor, A., McClements, M.E., and MacLaren, R.E. (2020). CRISPR-Cas9 DNA Base-Editing and Prime-Editing. International Journal of Molecular Sciences *21*, 6240. https://doi.org/10.3390/ijms21176240.

Kellogg, E.H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins *79*, 830–838. https://doi.org/10.1002/prot.22921.

Khan Mirzaei, M., and Nilsson, A.S. (2015). Isolation of Phages for Phage Therapy: A Comparison of Spot Tests and Efficiency of Plating Analyses for Determination of Host Range and Efficacy. PLoS One *10*. https://doi.org/10.1371/journal.pone.0118557.

Kilcher, S., and Loessner, M.J. (2019). Engineering Bacteriophages as Versatile Biologics. Trends in Microbiology *27*, 355–367. https://doi.org/10.1016/j.tim.2018.09.006.

Kilcher, S., Studer, P., Muessner, C., Klumpp, J., and Loessner, M.J. (2018a). Cross-genus rebooting of custom-made, synthetic bacteriophage genomes in L-form bacteria. PNAS *115*, 567–572. https://doi.org/10.1073/pnas.1714658115.

Kilcher, S., Studer, P., Muessner, C., Klumpp, J., and Loessner, M.J. (2018b). Cross-genus rebooting of custom-made, synthetic bacteriophage genomes in L-form bacteria. PNAS *115*, 567–572. https://doi.org/10.1073/pnas.1714658115.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424. https://doi.org/10.1038/nature17946.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. *110*, 14024–14029. https://doi.org/10.1073/pnas.1301301110.

Kuijpers, N.G., Solis-Escalante, D., Bosman, L., van den Broek, M., Pronk, J.T., Daran, J.-M., and Daran-Lapujade, P. (2013). A versatile, efficient strategy for assembly of multi-fragment expression vectors in Saccharomyces cerevisiae using 60 bp synthetic recombination sequences. Microb Cell Fact *12*, 47. https://doi.org/10.1186/1475-2859-12-47.

Kutter, E.M., Kuhl, S.J., and Abedon, S.T. (2015). Re-establishing a place for phage therapy in western medicine. Future Microbiol *10*, 685–688. https://doi.org/10.2217/fmb.15.28.

Landsberger, M., Gandon, S., Meaden, S., Rollie, C., Chevallereau, A., Chabas, H., Buckling, A., Westra, E.R., and van Houte, S. (2018). Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. Cell *174*, 908-916.e12. https://doi.org/10.1016/j.cell.2018.05.058.

Langer, S.J., Ghafoori, A.P., Byrd, M., and Leinwand, L. (2002). A genetic screen identifies novel non-compatible loxP sites. Nucleic Acids Res *30*, 3067–3077. .

Langridge, G.C., Phan, M.-D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., et al. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res *19*, 2308–2316. https://doi.org/10.1101/gr.097097.109.

Lee, J.M., Huddleston, J., Doud, M.B., Hooper, K.A., Wu, N.C., Bedford, T., and Bloom, J.D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. PNAS *115*, E8276–E8285. https://doi.org/10.1073/pnas.1806133115.

Li, C.-L.F., Santhanam, B., Webb, A.N., Zupan, B., and Shaulsky, G. (2016a). Gene discovery by chemical mutagenesis and whole-genome sequencing in Dictyostelium. Genome Res. *26*, 1268–1276. https://doi.org/10.1101/gr.205682.116.

Li, Z., Tang, J., and Guo, F. (2016b). Identification of 14-3-3 Proteins Phosphopeptide-Binding Specificity Using an Affinity-Based Computational Approach. PLOS ONE *11*, e0147467. https://doi.org/10.1371/journal.pone.0147467.

Lin, T.-Y., Lo, Y.-H., Tseng, P.-W., Chang, S.-F., Lin, Y.-T., and Chen, T.-S. (2012). A T3 and T7 Recombinant Phage Acquires Efficient Adsorption and a Broader Host Range. PLOS ONE *7*, e30954. https://doi.org/10.1371/journal.pone.0030954.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507.

Marinelli, L.J., Piuri, M., Swigoňová, Z., Balachandran, A., Oldfield, L.M., Kessel, J.C. van, and Hatfull, G.F. (2008). BRED: A Simple and Powerful Tool for Constructing Mutant and Recombinant Bacteriophage Genomes. PLOS ONE *3*, e3957. https://doi.org/10.1371/journal.pone.0003957.

Marinelli, L.J., Piuri, M., and Hatfull, G.F. (2019). Genetic Manipulation of Lytic Bacteriophages with BRED: Bacteriophage Recombineering of Electroporated DNA. Methods Mol. Biol. *1898*, 69–80. https://doi.org/10.1007/978-1-4939-8940-9_6.

Medhekar, B., and Miller, J.F. (2007). Diversity-generating retroelements. Current Opinion in Microbiology *10*, 388–395. https://doi.org/10.1016/j.mib.2007.06.004.

Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., and Lenski, R.E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. Science *335*, 428–432. https://doi.org/10.1126/science.1214449.

Miller, S.M., Wang, T., and Liu, D.R. (2020). Phage-assisted continuous and non-continuous evolution. Nat Protoc *15*, 4101–4127. https://doi.org/10.1038/s41596-020-00410-3.

Mizuno, C.M., Luong, T., Cederstrom, R., Krupovic, M., Debarbieux, L., and Roach, D.R. (2020). Isolation and Characterization of Bacteriophages That Infect Citrobacter rodentium, a Model Pathogen for Intestinal Diseases. Viruses *12*. https://doi.org/10.3390/v12070737.

Molineux, I.J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol. Microbiol. *40*, 1–8. https://doi.org/10.1046/j.1365-2958.2001.02357.x.

Moore, C.L., Papa, L.J., and Shoulders, M.D. (2018). A Processive Protein Chimera Introduces Mutations across Defined DNA Regions In Vivo. J. Am. Chem. Soc. *140*, 11560–11564. https://doi.org/10.1021/jacs.8b04001.

Mutalik, V.K., and Arkin, A.P. (2022). A Phage Foundry Framework to Systematically Develop Viral Countermeasures to Combat Antibiotic-Resistant Bacterial Pathogens. IScience *25*, 104121. https://doi.org/10.1016/j.isci.2022.104121.

Mutalik, V.K., Adler, B.A., Rishi, H.S., Piya, D., Zhong, C., Koskella, B., Kutter, E.M., Calendar, R., Novichkov, P.S., Price, M.N., et al. (2020). High-throughput mapping of the phage resistance landscape in E. coli. PLOS Biology *18*, e3000877. https://doi.org/10.1371/journal.pbio.3000877.

Nayeemul Bari, S.M., Walker, F.C., Cater, K., Aslan, B., and Hatoum-Aslan, A. (2017). Strategies for Editing Virulent Staphylococcal Phages Using CRISPR-Cas10. ACS Synth Biol *6*, 2316–2325. https://doi.org/10.1021/acssynbio.7b00240.

Nobrega, F.L., Vlot, M., de Jonge, P.A., Dreesens, L.L., Beaumont, H.J.E., Lavigne, R., Dutilh, B.E., and Brouns, S.J.J. (2018). Targeting mechanisms of tailed bacteriophages. Nature Reviews Microbiology *16*, 760–773. https://doi.org/10.1038/s41579-018-0070-8.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq; high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods *6*, 767–772. https://doi.org/10.1038/nmeth.1377.

Oppenheim, A.B., Rattray, A.J., Bubunenko, M., Thomason, L.C., and Court, D.L. (2004). In vivo recombineering of bacteriophage lambda by PCR fragments and single-strand oligonucleotides. Virology *319*, 185–189. https://doi.org/10.1016/j.virol.2003.11.007.

Pagnout, C., Sohm, B., Razafitianamaharavo, A., Caillet, C., Offroy, M., Leduc, M., Gendre, H., Jomini, S., Beaussart, A., Bauda, P., et al. (2019). Pleiotropic effects of rfa-gene mutations on Escherichia coli envelope properties. Sci Rep *9*. https://doi.org/10.1038/s41598-019-46100-3.

Pires, D.P., Cleto, S., Sillankorva, S., Azeredo, J., and Lu, T.K. (2016). Genetically Engineered Phages: a Review of Advances over the Last Decade. Microbiol. Mol. Biol. Rev. *80*, 523–543. https://doi.org/10.1128/MMBR.00069-15.

Pires, D.P., Monteiro, R., Mil-Homens, D., Fialho, A., Lu, T.K., and Azeredo, J. (2021). Designing P. aeruginosa synthetic phages with reduced genomes. Sci Rep *11*, 2164. https://doi.org/10.1038/s41598-021-81580-2.

Principi, N., Silvestri, E., and Esposito, S. (2019). Advantages and Limitations of Bacteriophages for the Treatment of Bacterial Infections. Front. Pharmacol. *10*. https://doi.org/10.3389/fphar.2019.00513.

Pryor, J.M., Potapov, V., Bilotti, K., Pokhrel, N., and Lohman, G.J.S. (2022). Rapid 40 kb Genome Construction from 52 Parts through Data-optimized Assembly Design. ACS Synth. Biol. https://doi.org/10.1021/acssynbio.1c00525.

Pulkkinen, E.M., Hinkley, T.C., and Nugen, S.R. (2019). Utilizing in vitro DNA assembly to engineer a synthetic T7 Nanoluc reporter phage for Escherichia coli detection. Integrative Biology *11*, 63–68. https://doi.org/10.1093/intbio/zyz005.

Qimron, U., Marintcheva, B., Tabor, S., and Richardson, C.C. (2006). Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. Proc. Natl. Acad. Sci. U.S.A. *103*, 19039–19044. https://doi.org/10.1073/pnas.0609428103.

Raman, S., Taylor, N., Genuth, N., Fields, S., and Church, G.M. (2014). Engineering Allostery. Trends Genet *30*, 521–528. https://doi.org/10.1016/j.tig.2014.09.004.

Ramirez-Chamorro, L., Boulanger, P., and Rossier, O. (2021). Strategies for Bacteriophage T5 Mutagenesis: Expanding the Toolbox for Phage Genome Engineering. Frontiers in Microbiology *12*. .

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012). Going viral: next generation sequencing applied to human gut phage populations. Nat Rev Microbiol *10*, 607–617. https://doi.org/10.1038/nrmicro2853.

Romero, P.A., Tran, T.M., and Abate, A.R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. PNAS *112*, 7159–7164. https://doi.org/10.1073/pnas.1422285112.

Rousset, F., Cui, L., Siouve, E., Becavin, C., Depardieu, F., and Bikard, D. (2018). Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. PLOS Genetics *14*, e1007749. https://doi.org/10.1371/journal.pgen.1007749.

Rustad, M., Eastlund, A., Marshall, R., Jardine, P., and Noireaux, V. (2017). Synthesis of Infectious Bacteriophages in an E. coli-based Cell-free Expression System. JoVE (Journal of Visualized Experiments) e56144. https://doi.org/10.3791/56144.

Salmond, G.P.C., and Fineran, P.C. (2015). A century of the phage: past, present and future. Nat Rev Microbiol *13*, 777–786. https://doi.org/10.1038/nrmicro3564.

Santos, A.L., Moreirinha, C., Lopes, D., Esteves, A.C., Henriques, I., Almeida, A., Domingues, M.R.M., Delgadillo, I., Correia, A., and Cunha, Â. (2013). Effects of UV Radiation on the Lipids and Proteins of Bacteria Studied by Mid-Infrared Spectroscopy. Environ. Sci. Technol. *47*, 6306–6315. https://doi.org/10.1021/es400660g.

Sausset, R., Petit, M.A., Gaboriau-Routhiau, V., and De Paepe, M. (2020). New insights into intestinal phages. Mucosal Immunology *13*, 205–215. https://doi.org/10.1038/s41385-019-0250-5.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., et al. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrob. Agents Chemother. *61*, e00954-17. https://doi.org/10.1128/AAC.00954-17.

Selle, K., Fletcher, J.R., Tuson, H., Schmitt, D.S., McMillan, L., Vridhambal, G.S., Rivera, A.J., Montgomery, S.A., Fortier, L.-C., Barrangou, R., et al. In Vivo Targeting of Clostridioides difficile Using Phage-Delivered CRISPR-Cas3 Antimicrobials. MBio *11*, e00019-20. https://doi.org/10.1128/mBio.00019-20.

Shen, T.-C.D., Albenberg, L., Bittinger, K., Chehoud, C., Chen, Y.-Y., Judge, C.A., Chau, L., Ni, J., Sheng, M., Lin, A., et al. (2015). Engineering the gut microbiota to treat hyperammonemia. J Clin Invest *125*, 2841–2850. https://doi.org/10.1172/JCI79214.

Shibai, A., Takahashi, Y., Ishizawa, Y., Motooka, D., Nakamura, S., Ying, B.-W., and Tsuru, S. (2017). Mutation accumulation under UV radiation in Escherichia coli. Sci Rep *7*, 14531. https://doi.org/10.1038/s41598-017-15008-1.

Shin, J., Jardine, P., and Noireaux, V. (2012). Genome replication, synthesis, and assembly of the bacteriophage T7 in a single cell-free reaction. ACS Synth Biol *1*, 408–413. https://doi.org/10.1021/sb300049p.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

Silva, Y.J., Costa, L., Pereira, C., Mateus, C., Cunha, A., Calado, R., Gomes, N.C.M., Pardo, M.A., Hernandez, I., and Almeida, A. (2014). Phage therapy as an approach to prevent Vibrio anguillarum infections in fish larvae production. PLoS ONE *9*, e114197. https://doi.org/10.1371/journal.pone.0114197.

Simon, A.J., Morrow, B.R., and Ellington, A.D. (2018). Retroelement-Based Genome Editing and Evolution. ACS Synth. Biol. *7*, 2600–2611. https://doi.org/10.1021/acssynbio.8b00273.

Sordi, L.D., Lourenço, M., and Debarbieux, L. (2019). The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract. Cell Host & Microbe *25*, 210–218. https://doi.org/10.1016/j.chom.2019.01.018.

Stemmer, W.P.C. (1994). Rapid evolution of a protein in vitro by DNA shuffling. Nature *370*, 389–391. https://doi.org/10.1038/370389a0.

Tran, N.Q., Rezende, L.F., Qimron, U., Richardson, C.C., and Tabor, S. (2008). Gene 1.7 of bacteriophage T7 confers sensitivity of phage growth to dideoxythymidine. Proc Natl Acad Sci U S A *105*, 9373–9378. https://doi.org/10.1073/pnas.0804164105.

Valvano, M.A., Messner, P., and Kosma, P. (2002). Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. Microbiology, *148*, 1979–1989. https://doi.org/10.1099/00221287-148-7-1979.

Wetmore, K.M., Price, M.N., Waters, R.J., Lamson, J.S., He, J., Hoover, C.A., Blow, M.J., Bristow, J., Butland, G., Arkin, A.P., et al. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. MBio *6*, e00306-15. https://doi.org/10.1128/mBio.00306-15.

Wrenbeck, E.E., Klesmith, J.R., Stapleton, J.A., Adeniran, A., Tyo, K.E.J., and Whitehead, T.A. (2016). Plasmid-based one-pot saturation mutagenesis. Nat Methods *13*, 928–930. https://doi.org/10.1038/nmeth.4029.

Wu, Z., Kan, S.B.J., Lewis, R.D., Wittmann, B.J., and Arnold, F.H. (2019). Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences *116*, 8852–8858. https://doi.org/10.1073/pnas.1901979116.

Yehl, K., Lemire, S., Yang, A.C., Ando, H., Mimee, M., Torres, M.D.T., de la Fuente-Nunez, C., and Lu, T.K. (2019). Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. Cell *179*, 459-469.e9. https://doi.org/10.1016/j.cell.2019.09.015.

Yosef, I., Goren, M.G., Globus, R., Molshanski-Mor, S., and Qimron, U. (2017). Extending the Host Range of Bacteriophage Particles for DNA Transduction. Molecular Cell *66*, 721-728.e3. https://doi.org/10.1016/j.molcel.2017.04.025.

Zheng, K., Wang, Y., Li, N., Jiang, F.-F., Wu, C.-X., Liu, F., Chen, H.-C., and Liu, Z.-F. (2018). Highly efficient base editing in bacteria using a Cas9-cytidine deaminase fusion. Commun Biol *1*, 1–6. https://doi.org/10.1038/s42003-018-0035-5.

Chapter 3

Literature Review

Engineered bacteriophages as programmable biocontrol agents

Authors: Phil Huss[1,2,3,] and Srivatsan Raman[1,2,4,*]

Contributions:

P.H. conceived the review topic, wrote and revised the manuscript

S.R. secured funding, conceived the review topic and revised the manuscript

Affiliations:

[1]Department of Biochemistry, University of Wisconsin-Madison

[2]Department of Bacteriology, University of Wisconsin-Madison

[3]Microbiology Doctoral Training Program, University of Wisconsin-Madison

[4]Department of Chemical and Biological Engineering, University of Wisconsin-Madison

[**] Correspondence: sraman4@wisc.edu (Srivatsan Raman)

Published in Current Opinion in Biotechnology 2020, 61:116–121 and included here under creative commons (CC BY 4.0) allowing unrestricted used and redistribution provided original authors and sources are credited. Article Link: https://doi.org/10.1016/j.copbio.2019.11.013.

Abstract

Bacteriophages (or ‘phages’) can be potent biocontrol agents but their potential has not been fully realized due to inherent limitations of natural phages. By leveraging new tools in synthetic biology, natural phages can be engineered to overcome these limitations to markedly improve their efficacy and programmability, making them ideal biocontrol agents. Engineered phages can be used for targeted detection and removal of pathogens, *in situ* microbiome editing, gene delivery and programable control of phage-bacterial interactions. In this mini review we examine different ways natural phages can be engineered as effective biocontrol agents and identify novel applications of engineered phages in food biotechnology.

Highlights

- Engineered bacteriophages are promising tools for use in food biotechnology
- Diverse natural bacteriophages can be leveraged by engineering for specificity and infectivity
- Engineered bacteriophages are potent tools for pathogen biocontrol
- Engineered bacteriophages can be used for targeted delivery vectors and pathogen detection.

Introduction

Foodborne infections caused by bacterial pathogens are a serious threat with hundreds of thousands of deaths every year globally [1]. Healthcare costs of foodborne illness are estimated at a staggering $75 billion/year in the United States [2], with cascading economic losses from discarded food, culled farm animals, and food recalls. Traditional biocontrol of bacterial pathogens has relied on broad-spectrum approaches such as antibiotics or pasteurization that vary in effectiveness, impact natural microflora of food, and can negatively affect food quality [3], [4]. Bacteriophages, or 'phages', are viruses that kill and are a promising alternative for bacterial biocontrol. They are ubiquitous natural predators of bacteria that are cheap-to-produce and can precisely target and kill pathogens without affecting food quality [5], [6]. Although products based on natural phages have been in the market for decades, their adoption by industry is low and their game-changing potential remains unfulfilled. [7]. Key factors impeding broader use of natural phages are poor efficacy compared to traditional biocontrol methods and poor scalability for high-volume production due to the narrow specificity profile of phages. However, synthetic biology offers exciting new tools to build engineered phages through a variety of recombineering approaches and *in vitro* genome assembly [8]. Engineering phages without natural limitations could lead to a design-build-test-learn platform to rapidly prototype new phages with user-defined properties. In this mini review, we will examine different ways natural phages can be engineered as more effective biocontrol agents and identify novel applications for engineered phages in food biotechnology.

Engineering phages for higher efficacy

A major hurdle in the use of natural phages for biocontrol is their low efficacy. Although initial application of natural phages logarithmically reduces target bacterial levels, the residual bacterial load remains high. Even this limited efficacy is achieved with high phage to bacteria ratios which may not be feasible in applications outside a laboratory setting [9]–[14]. Bacteria often continue to grow and/or quickly recover after phage application [9]–[11], [13]–[15], indicating low phage susceptibility and swift emergence of bacterial resistance. In this section, we evaluate factors that can limit the efficacy of natural phages and examine how engineering approaches can overcome these shortcomings (Figure 1).

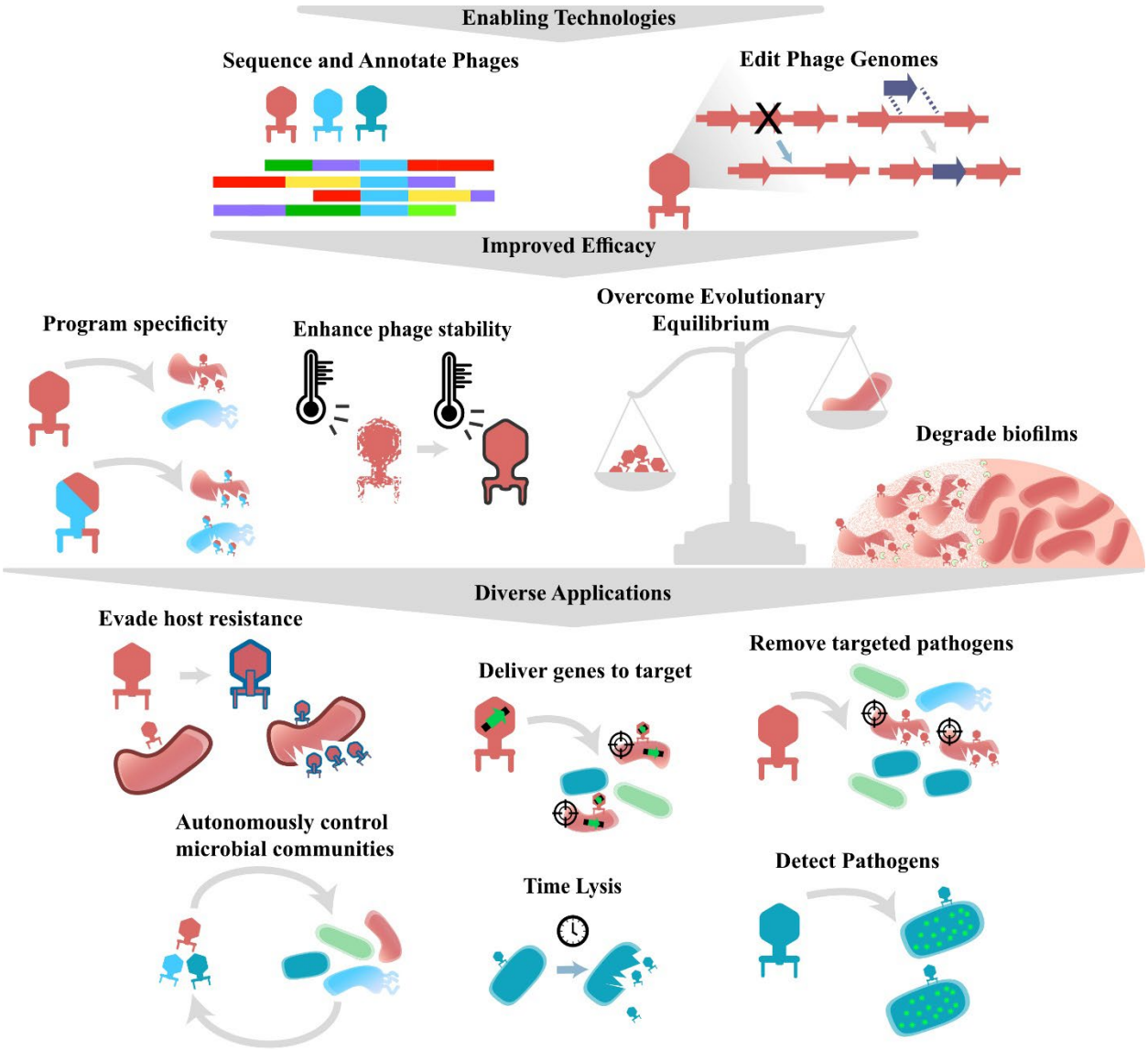


Figure 1 – Strategies for engineering phages and their applications

The confluence of synthetic biology, genome engineering, viral metagenomics and deep sequencing has provided tools for rapid, evolution-guided and rational design of phages with tailored properties. Engineering approaches and applications in food biotechnology have been graphically summarized below.

Overcoming evolutionary equilibrium in natural phages

Limited efficacy of natural phages is not surprising from an evolutionary perspective because a phage that fully eliminates its bacterial host will itself perish too. Therefore, as predators, phages must co-exist in equilibrium with their bacterial prey. In fact, evolutionary models show that natural selection favors mediocre killers over highly efficient phages [16]. Cocktails of several natural phages can be more efficacious but still face this evolutionary pressure to equilibrate with the bacterial host. To achieve a high pathogen clearance, we need to engineer phages which are not subject to evolutionary constraints imposed on natural phages. Every stage of a phage life cycle can be engineered to counteract various modes of resistance. For instance, phages with mutated tail fibers outcompete natural phages against *Pseudomonas aeruginosa* [17], and a single point mutation can confer a 1000-fold increase in efficacy for phages of *Mycobacterium abscessus* [18]**. Phage genome editing can remove built-in mechanisms in phages that reduce phage efficacy such as self-inhibition when their bacterial host experiences starvation [19], [20]. Self-inhibition is beneficial for natural phages because it gives bacterial populations time to recover before phages prey again, but is undesirable for biocontrol. Engineered phages devoid of starvation-induced regulatory genes continued to grow on bacteria in starvation conditions [21]. Lytic capabilities of engineered phages have also been enhanced by overexpressing genes such as phage holins, which are proteins that permeabilize cell membranes [22]. Bacterial resistance can be overcome by incorporating new genes into the engineered phage genome, including phage defense mechanisms such as anti-CRISPR genes or additional genes lethal to the target bacteria

[23]–[25]. A broad strategy to overcome bacterial resistance is to employ genome-scale screens such as CRISPRi or transposon insertion to reveal mechanisms of resistance [26], [27] and guide efforts to overcome evolutionary equilibrium. Expanding these screens to a larger set of bacteria could lead to understanding common resistance patterns and establishing engineering targets for phages.

Engineering lysogenic to lytic conversion

Due to their high abundance, lysogenic phages represent a treasure trove of natural phages that can be used for biocontrol applications. During their lifecycle, lysogens become dormant after integrating into the host genome, only to become activated later to kill the cell. Therefore, unlike lytic phages, lysogenic phages are not considered suitable for biocontrol and therapeutic applications due to the risk of horizontal gene transfer. Natural phages must be thoroughly screened for lysogeny, significantly increasing time, cost, and uncertainty in product development cycles. To tap the vast natural phage reservoir, lysogenic phages could be converted to obligate lytic phages by removing factors that allow for lysogeny such as recombinases and repressors [28]. These converted phages have increased lethality and host range [24]**, [29], as evidenced by those used to treat *M. abscessus* showing a 10,000-fold increase in efficacy after removing lysogeny genes [18]**. New bioinformatics tools tailored to identify these lysogenic phages and the factors that make them lysogenic can quickly screen for candidate phages and genes for this conversion [30], [31]. Lysogenic to lytic conversion greatly enhances the diversity and effectiveness of phages for biocontrol applications.

Programming host specificity

The bacterial host range of different natural phages varies significantly and finding the right combination of natural phages can be arduous. Engineering phages to reprogram host range removes this constraint. Specificity can be tailored by swapping or complementing host-binding proteins [32]–[34]*, or by removal of lysogeny genes or other genes that dictate specificity [29], [35]. Programmable engineered phages would provide a platform for phage treatments for any of a set of closely related bacteria. Programming narrow host ranges is ideal for food reliant on specific microbiota compositions like cheeses where contaminating bacteria need to be removed without disturbing other microflora [36], whereas a programmable broad host range is more ideal for phages designed for biocontrol.

Engineering stable phages

Phage stability is essential for biocontrol as phages are exposed to harsh environments including continual UV irradiation, low pH, and high temperature [9], [37], [38]. For example, pH and temperature in the gastrointestinal (GI) tract can vary between 2.5-5.7 and 38°C to 42°C, respectively [39]*. Phages can degrade under these conditions, which can affect their proliferation and efficacy. Engineered phages with protective surface phospholipids have increased survival in the hostile GI tract environment [39]*. Some natural phage have also exhibited stability under harsh environmental conditions [40]. Genes responsible for these attributes can be identified and integrated into engineered phages to improve stability. Alternatively, approaches such as computational

protein design, which has met with remarkable success in improving protein stability, could be employed to stabilize phage coat proteins without losing the flexibility required for function [41]. Stable phages could be used for crop protection and can also act as biocontrol ‘sentinels’ that can protect against bacterial contamination in the future.

Degrading biofilms using engineered phages

Bacteria that form biofilms are an enormous challenge in food safety, as the biofilm provides a protective cover against traditional biocontrol agents including phages [42], [43]. Phage enzymes that can degrade biofilms have been characterized [44] and phages engineered to include biofilm degrading peptides and enzymes have effectively dispersed biofilm [45]**, [46]. Phages targeting biofilm-creating organisms such as *Listeria monocytogenes, Staphylococcus aureus,* and *Escherichia coli* could be enhanced with these genes, establishing a unique advantage over traditional methods of biocontrol for these challenging pathogens.

Additional applications in food biotechnology

Using engineered phages as targeted delivery vectors

Currently, we lack tools to deliver genes to specific targets in a mixed microbial community *in situ*. Engineered phages are ideally suited as targeted delivery vectors due to their tailored host specificity [34]*. Gene delivery is useful to modulate the composition of a community by enhancing or reducing the fitness of a target species. This approach

can also deliver enzymes or metabolic operons to produce nutrients or signaling molecules. Engineered phages have been used as delivery vectors to resensitize bacteria to antibiotics by providing a drug-sensitizing DNA cassette [47]. CRISPR-Cas systems engineered into phages have also successfully disrupted virulence genes in bacteria [33]. These powerful tools could be leveraged to deliver and incorporate advantageous genes into the genome, such as proteases needed for food flavoring or acid production genes for fermentation.

Engineered phages can compete with undesirable phages

Complex mixed communities of starter cultures used to make cheese, yogurt, and other fermented milk products are highly susceptible to phage infection. Economic losses from discarded production batches and sanitizing equipment can be significant. Current methods such as air flow control and strict sanitary conditions for controlling phages are expensive and not particularly effective [48]. Counterintuitively, phages themselves may be the solution to this issue. Phages naturally compete with other phages and have anti-phage genes that can block expression of phage genes, prevent infection, or compete for insertion sites in the bacterial genome [49]. Engineering lysogenic phages to encode these anti-phage genes while removing their own ability to propagate could protect starter cultures from unwanted phages.

Using engineered phages as rapid detection tools

Rapid, low-cost detection of bacterial pathogens is critical for food safety. Methods such as antibody tagging are effective but are not cost effective, and culture-based mechanisms are laborious and time consuming. Engineered phages can deliver a bioluminescent reporter enzyme to readily detect pathogens [45]**, [50], [51] as a rapid and accurate tool during food processing or in final products.

Using engineered phages for controlled lysis

Timely lysis of starter cultures can be beneficial. For example, starter culture lysis is an important consideration in cheese maturation [52]. Lysogenic phages could be engineered to lyse the cell under an inducible condition such as access to a small molecule. Engineered lysogenic phages with this ability would trigger controlled, reproducible, and exponential lysis of the culture at an ideal time for cheese maturation.

Platform for rapid phage engineering

When creating phage-based products, engineered phages have several key advantages over natural phages. Natural phage discovery is a serial, time consuming, and laborious process. The physiology of newly discovered phages is often poorly understood, which may lead to batch-to-batch inconsistencies during manufacturing and making mass production unsustainable. Natural phage production pipelines can also be interrupted by a frequent need to discover new natural phages to combat emerging bacterial resistance or to create cocktails to cover strain variations in pathogens.

Engineered phages could provide a flexible product development platform and a scalable and customizable workflow (Figure 2). We envision developing well-characterized chassis phages against different key bacterial clades and engineering them using natural and synthetic parts to achieve the desired bacterial host range and effect. Mixing and matching these chassis phages may require minor adjustments but not a complete overhaul of a production pipeline, minimizing batch-to-batch variability and ensuring product quality.

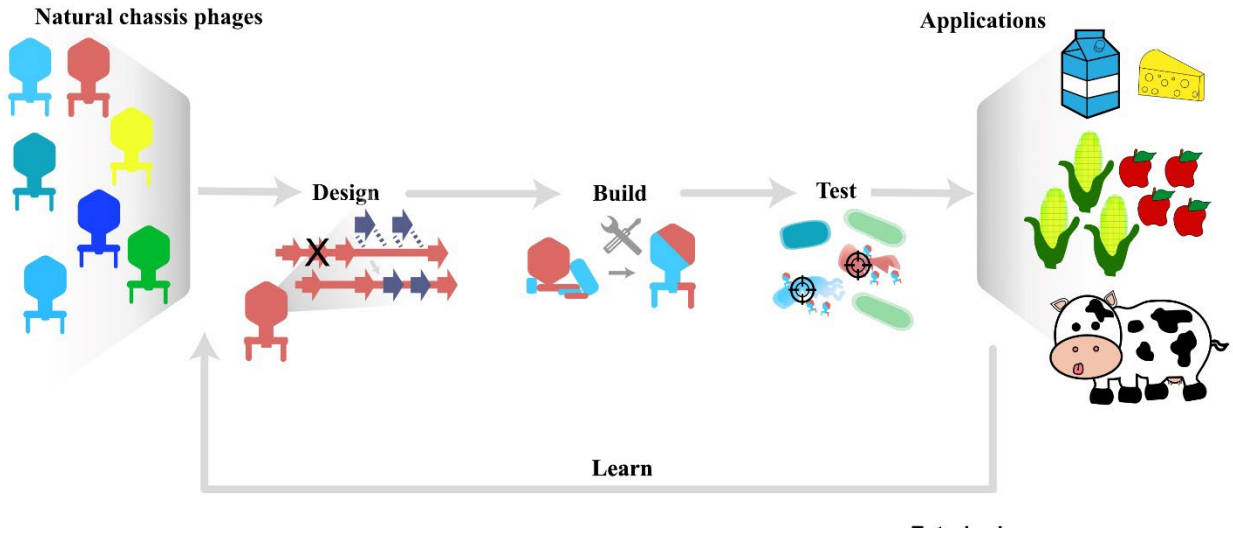


Figure 2 – A platform for rapid phage engineering

A generalized platform for rapid engineering of phages using a design-test-build-learn approach.

Conclusions

In this review we have outlined methods for improving natural phages for use in many applications in food biotechnology. Phages have enormous potential and are diverse tools that we have barely begun to explore. Phages are tractable and can be modified in many ways to improve effectiveness with even minimal engineering, and numerous methods now exist for modifying phage genomes to produce engineered phages. We envision that engineered phages will serve as a platform for developing biocontrol and delivery mechanisms to a broad range of bacteria to solve a variety of current problems in food biotechnology.

Acknowledgements

This work is partially supported by Gates Grand Challenges grant (OPP1150209) and US Department of Agriculture Hatch award (WIS02066). We thank Laura Alexander, Megan Leander, Chutikarn Chitboonthavisuk and Kyle Nishikawa for comments on the review.

Conflict of Interest

The authors declare no conflict of interest

References

[1] A. H. Havelaar *et al.*, “World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010,” *PLoS Med.*, vol. 12, no. 12, p. e1001923, Dec. 2015.

[2] R. L. Scharff, “Economic burden from health losses due to foodborne illness in the United States,” *J. Food Prot.*, vol. 75, no. 1, pp. 123–131, Jan. 2012.

[3] F. Al-juhaimi *et al.*, “Effect of various food processing and handling methods on preservation of natural antioxidants in fruits and vegetables,” *J. Food Sci. Technol.*, vol. 55, no. 10, pp. 3872–3880, Oct. 2018.

[4] T. F. Landers, B. Cohen, T. E. Wittum, and E. L. Larson, “A Review of Antibiotic Use in Food Animals: Perspective, Policy, and Potential,” *Public Health Rep.*, vol. 127, no. 1, pp. 4–22, 2012.

[5] E. C. Keen and G. Dantas, “Close Encounters of Three Kinds: Bacteriophages, Commensal Bacteria, and Host Immunity,” *Trends Microbiol.*, vol. 26, no. 11, pp. 943–954, Nov. 2018.

[6] M. N. Perera, T. Abuladze, M. Li, J. Woolston, and A. Sulakvelidze, “Bacteriophage cocktail significantly reduces or eliminates Listeria monocytogenes contamination on lettuce, apples, cheese, smoked salmon and frozen foods,” *Food Microbiol.*, vol. 52, pp. 42–48, Dec. 2015.

[7] Z. D. Moye, J. Woolston, and A. Sulakvelidze, “Bacteriophage Applications for Food Production and Processing,” *Viruses*, vol. 10, no. 4, Apr. 2018.

[8] S. Kilcher and M. J. Loessner, “Engineering Bacteriophages as Versatile Biologics,” *Trends Microbiol.*, vol. 27, no. 4, pp. 355–367, Apr. 2019.

[9] S. Fister, C. Robben, A. K. Witte, D. Schoder, M. Wagner, and P. Rossmanith, “Influence of Environmental Factors on Phage–Bacteria Interaction and on the Efficacy and Infectivity of Phage P100,” *Front. Microbiol.*, vol. 7, Jul. 2016.

[10] S. Guenther, D. Huwyler, S. Richard, and M. J. Loessner, “Virulent Bacteriophage for Efficient Biocontrol of Listeria monocytogenes in Ready-To-Eat Foods,” *Appl. Environ. Microbiol.*, vol. 75, no. 1, pp. 93–100, Jan. 2009.

[11] S. Fister, S. Fuchs, B. Stessl, D. Schoder, M. Wagner, and P. Rossmanith, “Screening and characterisation of bacteriophage P100 insensitive Listeria monocytogenes isolates in Austrian dairy plants,” *Food Control*, vol. 59, pp. 108–117, Jan. 2016.

[12] H.-W. Kang, J.-W. Kim, T.-S. Jung, and G.-J. Woo, “wksl3, a New Biocontrol Agent for Salmonella enterica Serovars Enteritidis and Typhimurium in Foods: Characterization, Application, Sequence Analysis, and Oral Acute Toxicity Study,” *Appl. Environ. Microbiol.*, vol. 79, no. 6, pp. 1956–1968, Mar. 2013.

[13] S. Guenther, O. Herzig, L. Fieseler, J. Klumpp, and M. J. Loessner, “Biocontrol of Salmonella Typhimurium in RTE foods with the virulent bacteriophage FO1-E2,” *Int. J. Food Microbiol.*, vol. 154, no. 1, pp. 66–72, Mar. 2012.

[14] Y. J. Silva *et al.*, “Phage therapy as an approach to prevent Vibrio anguillarum infections in fish larvae production,” *PloS One*, vol. 9, no. 12, p. e114197, 2014.

[15] L. Rodríguez-Rubio, P. García, A. Rodríguez, C. Billington, J. A. Hudson, and B. Martínez, “Listeriaphages and coagulin C23 act synergistically to kill Listeria monocytogenes in milk under refrigeration conditions,” *Int. J. Food Microbiol.*, vol. 205, pp. 68–72, Jul. 2015.

[16] S. Heilmann, K. Sneppen, and S. Krishna, “Sustainability of Virulence in a Phage-Bacterial Ecosystem,” *J. Virol.*, vol. 84, no. 6, pp. 3016–3022, Mar. 2010.

[17] L. Latino, C. Midoux, G. Vergnaud, and C. Pourcel, “Investigation of Pseudomonas aeruginosa strain PcyII-10 variants resisting infection by N4-like phage Ab09 in search for genes involved in phage adsorption,” *PLoS ONE*, vol. 14, no. 4, Apr. 2019.

[18] R. M. Dedrick *et al.*, “**Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus**,” *Nat. Med.*, vol. 25, no. 5, p. 730, May 2019.

[19] P. Golec, A. Wiczk, J. M. Łoś, G. Konopa, G. Węgrzyn, and M. Łoś, “Persistence of bacteriophage T4 in a starved Escherichia coli culture: evidence for the presence of phage subpopulations,” *J. Gen. Virol.*, vol. 92, no. 4, pp. 997–1003, 2011.

[20] M. Briggiler Marcó, J. Reinheimer, and A. Quiberoni, “Phage adsorption and lytic propagation in Lactobacillus plantarum: Could host cell starvation affect them?,” *BMC Microbiol.*, vol. 15, Dec. 2015.

[21] P. Golec *et al.*, “Proteomic profiles and kinetics of development of bacteriophage T4 and its rI and rIII mutants in slowly growing Escherichia coli,” *J. Gen. Virol.*, vol. 94, no. 4, pp. 896–905, 2013.

[22] T. A. T. Tran, D. K. Struck, and R. Young, “Periplasmic domains define holin-antiholin interactions in t4 lysis inhibition,” *J. Bacteriol.*, vol. 187, no. 19, pp. 6631–6640, Oct. 2005.

[23] M. Landsberger *et al.*, “Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity,” *Cell*, vol. 174, no. 4, pp. 908-916.e12, Aug. 2018.

[24] S. Kilcher, P. Studer, C. Muessner, J. Klumpp, and M. J. Loessner, “**Cross-genus rebooting of custom-made, synthetic bacteriophage genomes in L-form bacteria**,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 3, pp. 567–572, Jan. 2018.

[25] M. Hupfeld, D. Trasanidou, L. Ramazzini, J. Klumpp, M. J. Loessner, and S. Kilcher, “A functional type II-A CRISPR–Cas system from Listeria enables efficient genome editing of large non-integrating bacteriophage,” *Nucleic Acids Res.*, vol. 46, no. 13, pp. 6920–6933, Jul. 2018.

[26] F. Rousset, L. Cui, E. Siouve, C. Becavin, F. Depardieu, and D. Bikard, “Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors,” *PLOS Genet.*, vol. 14, no. 11, p. e1007749, Nov. 2018.

[27] L. A. Cowley *et al.*, “Transposon Insertion Sequencing Elucidates Novel Gene Involvement in Susceptibility and Resistance to Phages T4 and T7 in Escherichia coli O157,” *mBio*, vol. 9, no. 4, pp. e00705-18, Sep. 2018.

[28] C. Howard-Varona, K. R. Hargreaves, S. T. Abedon, and M. B. Sullivan, “Lysogeny in nature: mechanisms, impact and ecology of temperate phages,” *ISME J.*, vol. 11, no. 7, pp. 1511–1520, Jul. 2017.

[29] H. Zhang, D. E. Fouts, J. DePew, and R. H. Stevens, “Genetic modifications to temperate Enterococcus faecalis phage ϕEf11 that abolish the establishment of

lysogeny and sensitivity to repressor, and increase host range and productivity of lytic infection," *Microbiology*, vol. 159, no. Pt 6, pp. 1023–1035, Jun. 2013.

[30] A. L. de Sousa *et al.*, "PhageWeb – Web Interface for Rapid Identification and Characterization of Prophages in Bacterial Genomes," *Front. Genet.*, vol. 9, 2018.

[31] D. Arndt, A. Marcu, Y. Liang, and D. S. Wishart, "PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes," *Brief. Bioinform.*, Sep. 2017.

[32] H. Ando, S. Lemire, D. P. Pires, and T. K. Lu, "Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing," *Cell Syst.*, vol. 1, no. 3, pp. 187–196, Sep. 2015.

[33] J. Y. Park, B. Y. Moon, J. W. Park, J. A. Thornton, Y. H. Park, and K. S. Seo, "Genetic engineering of a temperate phage-based delivery system for CRISPR/Cas9 antimicrobials against *Staphylococcus aureus*," *Sci. Rep.*, vol. 7, p. 44929, Mar. 2017.

[34] I. Yosef, M. G. Goren, R. Globus, S. Molshanski-Mor, and U. Qimron, "**Extending the Host Range of Bacteriophage Particles for DNA Transduction**," *Mol. Cell*, vol. 66, no. 5, pp. 721-728.e3, Jun. 2017.

[35] F. Pouillot, H. Blois, and F. Iris, "Genetically engineered virulent phage banks in the detection and control of emergent pathogenic bacteria," *Biosecurity Bioterrorism Biodefense Strategy Pract. Sci.*, vol. 8, no. 2, pp. 155–169, Jun. 2010.

[36] M. Gobbetti, R. Di Cagno, M. Calasso, E. Neviani, P. F. Fox, and M. De Angelis, "Drivers that establish and assembly the lactic acid bacteria biota in cheeses," *Trends Food Sci. Technol.*, vol. 78, pp. 244–254, Aug. 2018.

[37] F. B. Iriarte, B. Balogh, M. T. Momol, L. M. Smith, M. Wilson, and J. B. Jones, "Factors affecting survival of bacteriophage on tomato leaf surfaces," *Appl. Environ. Microbiol.*, vol. 73, no. 6, pp. 1704–1711, Mar. 2007.

[38] E. Jończyk, M. Kłak, R. Międzybrodzki, and A. Górski, "The influence of external factors on bacteriophages--review," *Folia Microbiol. (Praha)*, vol. 56, no. 3, pp. 191–200, May 2011.

[39] F. L. Nobrega *et al.*, “**Genetically manipulated phages with improved pH resistance for oral administration in veterinary medicine**,” *Sci. Rep.*, vol. 6, p. 39235, Dec. 2016.

[40] J.-G. Yu *et al.*, “Isolation and Characterization of Bacteriophages Against Pseudomonas syringae pv. actinidiae Causing Bacterial Canker Disease in Kiwifruit,” *J. Microbiol. Biotechnol.*, vol. 26, no. 2, pp. 385–393, Feb. 2016.

[41] N. Koga *et al.*, “Principles for designing ideal protein structures,” *Nature*, vol. 491, no. 7423, pp. 222–227, Nov. 2012.

[42] L. Vidakovic, P. K. Singh, R. Hartmann, C. D. Nadell, and K. Drescher, “Dynamic biofilm architecture confers individual and collective mechanisms of viral protection,” *Nat. Microbiol.*, vol. 3, no. 1, pp. 26–31, Jan. 2018.

[43] S. Galié, C. García-Gutiérrez, E. M. Miguélez, C. J. Villar, and F. Lombó, “Biofilms in the Food Industry: Health Aspects and Control Methods,” *Front. Microbiol.*, vol. 9, May 2018.

[44] D. P. Pires, H. Oliveira, L. D. R. Melo, S. Sillankorva, and J. Azeredo, “Bacteriophage-encoded depolymerases: their diversity and biotechnological applications,” *Appl. Microbiol. Biotechnol.*, vol. 100, no. 5, pp. 2141–2151, Mar. 2016.

[45] Y. Born, L. Fieseler, V. Thöny, N. Leimer, B. Duffy, and M. J. Loessner, “**Engineering of Bacteriophages Y2::dpoL1-C and Y2::luxAB for Efficient Control and Rapid Detection of the Fire Blight Pathogen, Erwinia amylovora**,” *Appl. Environ. Microbiol.*, vol. 83, no. 12, pp. e00341-17, Jun. 2017.

[46] R. Pei and G. R. Lamas-Samanamud, “Inhibition of Biofilm Formation by T7 Bacteriophages Producing Quorum-Quenching Enzymes,” *Appl. Environ. Microbiol.*, vol. 80, no. 17, pp. 5340–5348, Sep. 2014.

[47] R. Edgar, N. Friedman, S. Molshanski-Mor, and U. Qimron, “Reversing Bacterial Resistance to Antibiotics by Phage-Mediated Delivery of Dominant Sensitive Genes,” *Appl. Environ. Microbiol.*, vol. 78, no. 3, pp. 744–751, Feb. 2012.

[48] M. B. Marcó, S. Moineau, and A. Quiberoni, “Bacteriophages and dairy fermentations,” *Bacteriophage*, vol. 2, no. 3, pp. 149–158, Jul. 2012.

[49] J. Mahony, S. McGrath, G. F. Fitzgerald, and D. van Sinderen, “Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes,” *Appl. Environ. Microbiol.*, vol. 74, no. 20, pp. 6206–6215, Oct. 2008.
[50] J. Kim, M. Kim, S. Kim, and S. Ryu, “Sensitive detection of viable Escherichia coli O157:H7 from foods using a luciferase-reporter phage phiV10lux,” *Int. J. Food Microbiol.*, vol. 254, pp. 11–17, Aug. 2017.
[51] Ł. Richter, M. Janczuk-Richter, J. Niedziółka-Jönsson, J. Paczesny, and R. Hołyst, “Recent advances in bacteriophage-based methods for bacteria detection,” *Drug Discov. Today*, vol. 23, no. 2, pp. 448–455, Feb. 2018.
[52] S. Lortal and M.-P. Chapot-Chartier, “Role, mechanisms and control of lactic acid bacteria lysis in cheese,” *Int. Dairy J.*, vol. 15, no. 6, pp. 857–871, Jun. 2005.

Reference Annotations

18** Authors used several engineered phages to successfully treat *Mycobacterium abscessus*. Phage efficacy was greatly improved by lysogenic to lytic conversion and through single mutations. This work is emblematic of the power that even minor engineering can have on phage efficacy.

24**Authors created a synthetic phage genome *in vitro* and rebooted it in host bacteria, increased efficacy through lysogenic to lytic conversion, and loaded the phage with enzybiotics to target nearby phage-resistant bacteria. This work is an excellent example of the flexibility engineered phages provide as a chassis to improve efficacy and overcome bacterial resistance.

34* Authors developed hybrid phages with different tails/tail fibers combinations designed to efficiently transduce DNA, extending the host range of these engineered particles and creating a platform for potentially tailoring host range of any phage.

39*Authors engineered a phage that expressed phospholipids on the surface of the phage to dramatically increase phage stability in the gastrointestinal tract.

45**Authors engineered phages for the fire blight pathogen *Erwinia amylovora* to contain a depolymerase to eliminate biofilm and increase phage efficacy. Authors also engineered a phage to contain a luciferase reporter for rapid fluorescent detection of this pathogen, demonstrating the multiple ways a single engineered phage can be used.

Chapter 4

Publication

Mapping the Functional Landscape of the Receptor Binding Domain of T7 Bacteriophage by Deep Mutational Scanning

Authors: Phil Huss[1,2], Anthony Meger[1], Megan Leander[1], Kyle Nishikawa[1] and Srivatsan Raman[1,2,3,§]

Contributions:

P.H.: Conceptualization, Data curation, Software, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing

A.M.: Software

M.L.: Methodology

K.N.: Methodology

S.R.: Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing - review and editing

Affiliations:

1 Department of Biochemistry, University of Wisconsin-Madison, Madison, WI

2 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI

3 Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI

§ Corresponding author: sraman4@wisc.edu

Published in eLife 10:e63775 and included here under creative commons (CC BY 4.0) allowing unrestricted used and redistribution provided original authors and sources are credited. Article Link: https://doi.org/10.7554/eLife.63775.

Abstract

The interaction between a bacteriophage and its host is mediated by the phage's receptor binding protein (RBP). Despite its fundamental role in governing phage activity and host range, molecular rules of RBP function remain a mystery. Here, we systematically dissect the functional role of every residue in the tip domain of T7 phage RBP (1660 variants) by developing a high-throughput, locus-specific, phage engineering method. This rich dataset allowed us to cross compare functional profiles across hosts to precisely identify regions of functional importance, many which were previously unknown. Substitution patterns showed host-specific differences in position and physicochemical properties of mutations, revealing molecular adaptation to individual hosts. We discovered gain-of-function variants against resistant hosts and host-constricting variants that eliminated certain hosts. To demonstrate therapeutic utility, we engineered highly active T7 variants against urinary tract pathogen. Our approach presents a generalized framework for characterizing sequence-function relationships in many phage-bacterial systems.

Introduction

Bacteriophages (or “phages”) shape microbial ecosystems by infecting and killing targeted bacterial species. As a result, they are promising tools for treatment of antibiotic resistant bacterial infections and microbiome manipulation (Canfield and Duerkop, 2020; Chen et al., 2014; Clokie et al., 2011; Dedrick et al., 2019; Kilcher and Loessner, 2019; Kutter et al., 2015; Mizuno et al., 2020; Sausset et al., 2020; Schooley et al., 2017; Shen et al., 2015; Shkoporov and Hill, 2019; Sordi et al., 2019). Interaction of phages with their bacterial receptors is a key determinant of their host range and virulence (Bertozzi Silva et al., 2016; de Jonge et al., 2019; Rousset et al., 2018). This interaction is primarily mediated by the receptor binding proteins (RBPs) of the phage (Nobrega et al., 2018). RBPs enable phages to adsorb to diverse cell surface molecules including proteins, polysaccharides, lipopolysaccharides and carbohydrate-binding moieties. Phages exhibit high functional plasticity through genetic alterations to RBPs and by natural and laboratory-guided evolution which can modulate activity and host range to different hosts and environments (Ando et al., 2015b; Chen et al., 2017; Dedrick et al., 2019; Dunne et al., 2019; Garcia et al., 2003; Gebhart et al., 2017; Holtzman et al., 2020; Lin et al., 2012; Meyer et al., 2012; Yehl et al., 2019; Yosef et al., 2017). In essence, survivability of a phage is intimately linked to the adaptability of its RBP. The challenge now is to understand the molecular code of RBPs in sufficient depth to enable predictable manipulation of host range and virulence. We sought to do so by combining deep mutational scanning of the RBP with powerful selections on multiple hosts.

Although RBPs remain the focus of many mechanistic, structural and evolutionary studies and are a prime target for engineering, we currently lack a systematic and comprehensive understanding of how RBP mutations influence phage activity and host range. Though insightful, directed evolution enriches only a small group of 'winners' which makes it difficult to glean a comprehensive mutational landscape of the RBP (Holtzman et al., 2020). Random mutagenesis-based screens generate multi-mutant variants whose individual effects cannot be easily deconvolved (Dunne et al., 2019; Yehl et al., 2019). Other approaches including swapping homologous RBPs lead to gain of function, however the underlying molecular determinants of function can be difficult to explain (Ando et al., 2015b; Chen et al., 2017; Gebhart et al., 2017; Yosef et al., 2017). In summary, despite the extraordinary functional potential of phage RBPs, how systematic changes to their sequence shape the overall functional landscape of a phage remains unknown.

Here we carried out deep mutational scanning (DMS), a high-throughput experimental technique, of the tip domain of the T7 phage RBP (tail fiber) to uncover molecular determinants of activity and host range. The tip domain is the distal region of the tail fiber that makes primary contact with the host receptor (González-García et al., 2015; Molineux, 2001; Qimron et al., 2006). We developed ORACLE (**O**ptimized **R**ecombination, **Ac**cumulation and **L**ibrary **E**xpression), a high-throughput, locus-specific, phage genome engineering method to create a large, unbiased library of phage variants at a targeted gene locus. Using ORACLE, we systematically and comprehensively mutated the tip domain by making all single amino acid substitutions at every site (1660

variants) and quantified the functional role of all variants on multiple bacterial hosts. We generated high resolution functional maps delineating regions concentrated with function-enhancing substitutions and host-specific substitutional patterns, many of which were previously unknown. We discovered T7 variants with far greater virulence than wildtype T7, demonstrating that even those natural phages that are well adapted to a host can be engineered for higher efficacy.

However, many variants highly adapted to one host performed poorly on others, underscoring a tradeoff between activity and host range. This functional screening highlights ideal regions of the tip domain for engineering host range. Furthermore, we demonstrated the functional potential of RBPs by discovering gain-of-function variants against resistant hosts and host-constriction variants that selectively eliminate specific hosts. To demonstrate the therapeutic value of ORACLE, we engineered T7 variants that avert emergence of spontaneous resistance in pathogenic *E. coli* causing urinary tract infections.

Our study explains the molecular drivers of adaptability of the tip domain and identifies key functional regions determining activity and host range. ORACLE provides a generalized framework to describe sequence-function relationships in phages to elucidate the molecular basis of phages, the most abundant life form on earth.

Results

Creating an unbiased library of phage variants using ORACLE

ORACLE is a high-throughput precision phage genome engineering technology designed to create a large, unbiased library of phage variants to investigate sequence-function relationships in phages. ORACLE overcomes three major hurdles. First, phage variants are created during the natural infection cycle of the phage which eliminates a common bottleneck from transforming DNA libraries. By recombining a donor cassette containing prespecified variants to a targeted site on the phage genome, ORACLE allows sequence programmability and generalizability to phages with transformable bacterial hosts capable of maintaining a plasmid library. Second, ORACLE minimizes library bias that can rapidly arise due to fitness advantage or deficiency of any variant on the propagating host that may then be amplified due to exponential phage growth. Minimizing bias is critical because variants that perform poorly on a propagating host but well on targeted hosts may disappear during propagation. Third, ORACLE prevents extreme abundance of wildtype over variants, which allows for resolving and scoring even small functional differences between variants. The development of this technology was necessary to overcome challenges with existing engineering approaches for creating a large, unbiased phage library. Direct transformation of phage libraries, while ideal for creating one or small groups of synthetic phages, will not work because phage genomes are typically too large for library transformation (Ando et al., 2015b; Kilcher et al., 2018b; Marinelli et al., 2008, 2019). Homologous recombination has low, variable recombination

rates and high levels of wildtype phage are retained, which mask library members (Pires et al., 2016; Yehl et al., 2019). Libraries of lysogenic phages could potentially be made using conventional bacterial genome engineering tools as the phage integrates into the host genome. However, this approach is not applicable to obligate lytic phages. Our desire to develop ORACLE for obligate lytic phages is motivated by their mandated use for phage therapy. Any phage, including lysogenic phages, with a sequenced genome and a transformable host that can maintain a plasmid library should be amenable to ORACLE.

ORACLE is carried out in four steps: (a) making acceptor phage (b) inserting gene variants through recombination (c) accumulating recombined phages (d) expressing the library for selection (Figure 1A). An ‘acceptor phage’ is a synthetic phage genome where the gene of interest (i.e., tail fiber) is replaced with a fixed sequence flanked by Cre recombinase sites to serve as a landing site for inserting variants (Figure 1 - Figure Supplement 1). We created T7 acceptor phages by assembling PCR fragments of the phage genome in yeast (Ando et al., 2015; Jaschke et al., 2012) (see methods). T7 acceptor phages lacking a wildtype tail fiber gene cannot plaque on *E. coli* and do not spontaneously re-acquire the tail fiber during propagation (Figure 1B and Figure 1 - Figure Supplement 2A). Furthermore, the T7 acceptor phages have no plaquing deficiency relative to wildtype when the tail fiber gene is provided from a donor plasmid (Figure 1 - Figure Supplement 2A). Thus, the tail fiber gene is decoupled from the rest of the phage genome for interrogation of function. Next, phage variants are generated within the host during the infection cycle by **O**ptimized **R**ecombination by inserting tail fiber variants from

a donor plasmid into the landing sites in the acceptor phage using site specific recombination. To minimize biasing of variants during propagation, a helper plasmid constitutively provides the wildtype tail fiber *in trans* such that all progeny phages can amplify comparably regardless of the fitness benefit or deficient of any variant. At this stage, we typically have ~1 recombined phage among 1000 acceptor phages (Figure 1C). To enrich recombined phages in this pool, we passage all progeny phages on *E. coli* expressing Cas9 and a gRNA targeting the fixed sequence flanked by recombinase sites we introduced into the acceptor phage. The helper plasmid is retained during this stage to continue minimizing bias by providing the wildtype tail fiber *in trans*. As a result, only unrecombined phages will be inhibited while recombined phages with tail fiber variants are **Ac**cumulated without bias. The Cas9-gRNA system successfully inhibits acceptor phages but has no effect on plaquing of untargeted phages (Figure 1 - Figure Supplement 2A-D). Recombined phages were highly enriched by over 1000-fold in the phage population when an optimized gRNA targeting the fixed sequence was used, whereas a randomized control gRNA yielded no enrichment of recombined phages (Figure 1D and Figure 1 - Figure Supplement 2E-F). In the final step, phages are propagated on *E. coli* which lack the helper plasmid that previously provided the wildtype tail fiber *in trans* to prevent bias. In this **L**ibrary **E**xpression, propagation on this host allows for full expression of the library variant – this is the first time during library creation that the variant is fully expressed on the phage particle. We sequenced the distribution of the library of tail fiber variants integrated on the phage genome after ORACLE. We compared this distribution to the distribution of variants on the recombination plasmid library to evaluate how

effective ORACLE was at integrating variants and preventing bias during library creation. The post-ORACLE phages were mildly skewed towards more abundant members but remained generally evenly distributed and comparable to the distribution of variants in the input donor plasmid library, retaining 99.8% coverage (Figure 1E). Comparison of variant libraries with and without DNAse treatment was well correlated (R=0.994), indicating no unencapsidated phage genomes influenced library distribution (Figure 1 - Figure Supplement 3). In summary, ORACLE is a generalizable tool for creating large, unbiased variant libraries of obligate lytic phages. These phage variants, including those that have a fitness deficiency on the host used to create the library, can all be characterized in a single selection experiment by deep sequencing phage populations before and after selection in a host. Compared to traditional plaque assays this represents increased throughput by nearly 3-4 orders of magnitude.

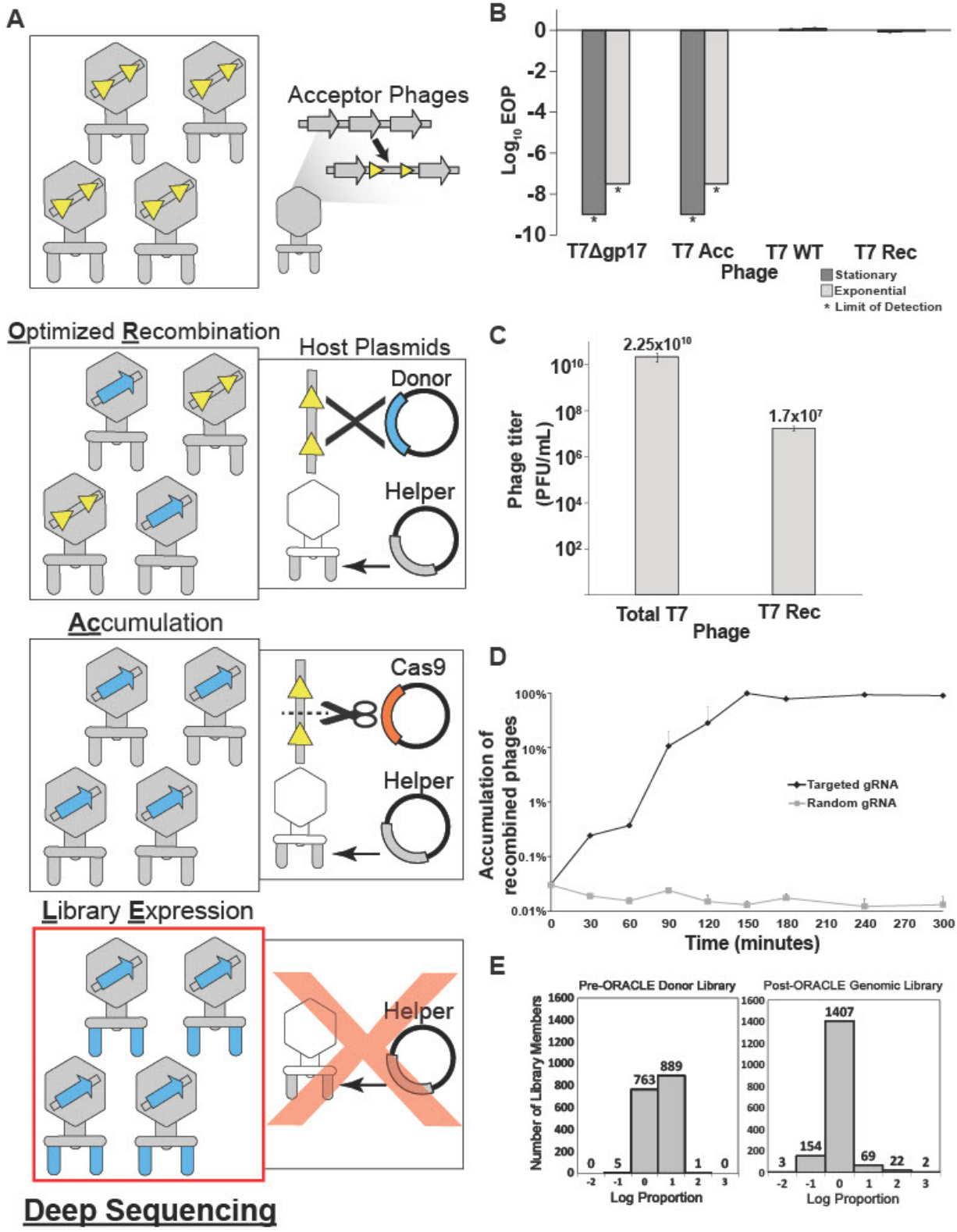

A
Acceptor Phages
Optimized Recombination
Host Plasmids
Donor
Helper
Accumulation
Cas9
Helper
Library Expression
Helper
Deep Sequencing
B
Log10 EOP
0
-2
-4
-6
-8
-10
T7Δgp17
T7 Acc
T7 WT
T7 Rec
Phage
Stationary
Exponential
* Limit of Detection
C
Phage titer (PFU/mL)
2.25x10^10
1.7x10^7
Total T7
T7 Rec
Phage
D
Accumulation of recombined phages
100%
10%
1%
0.1%
0.01%
Targeted gRNA
Random gRNA
0
30
60
90
120
150
180
210
240
270
300
Time (minutes)
E
Pre-ORACLE Donor Library
Post-ORACLE Genomic Library
Number of Library Members
1600
1400
1200
1000
800
600
400
200
0
0
5
763
889
1
0
3
154
1407
69
22
2
-2
-1
0
1
2
3
Log Proportion

Figure 1. Optimized Recombination, Accumulation and Library Expression (ORACLE) workflow for creating phage variant libraries

(A) Schematic illustration of the four steps of ORACLE: creation of acceptor phage, inserting gene variants (**O**ptimized **R**ecombination), enriching recombined phages (**Ac**cumulation) and expressing library for selection (**L**ibrary **E**xpression). Color notations are as follows: yellow triangles – Cre recombinase sites, blue colored segments – gene variants, orange colored segment – Cas9, grey colored segments – wild type phage parts including the wildtype tail fiber from the helper plasmid **(B)** Ability of different versions of T7 to infect *E. coli* 10G in stationary (dark gray bar) and exponential (light gray bar) phases by Efficiency of Plating (EOP) using exponential 10G with *gp17* tail fiber helper plasmid as reference host. T7 without tail fiber (T7Δgp17) and T7 Acceptor phages (T7 Acc) cannot visibly plaque, but wildtype T7 (T7 WT), and T7 with *gp17* recombined into the acceptor locus (T7 Rec) plaque efficiently. **(C)** Concentration of total (Total T7) and recombined (T7 Rec) phages after a single passage on host containing Cre recombinase system. Recombination rate is estimated to be ~7.19x10-4. **(D)** Percentage of recombined phages in total phages when using gRNA targeting fixed sequence at acceptor site T7 Acc (Targeted) or randomized gRNA (Random). **(E)** Histogram of abundance of variants in the input plasmid library (left) and on the phage genome after ORACLE (right) binned using log proportion centered on equal representation. All data represented as mean ± SD of biological triplicate.

Figure 1 – Figure Supplement 1. Sequence rearrangements before and after recombinase-mediated cassette exchange.

Figure 1 – Figure Supplement 2. Effect of Cas9-gRNA system on acceptor and control phages.

Figure 1 – Figure Supplement 3. Correlation between expression library with and without DNAse treatment.

Figure 1 - Source Data 1. Deep sequencing summary for phage variant expression library with and without DNAse treatment after ORACLE. Related to Figure 1E and Figure 1 – Figure Supplement 3.

Figure 1 - Source Data 2. Percentage distribution of each variant in the expression library. Related to Figure 1E and Figure 1 – Figure Supplement 3.

Deep mutational scanning of tip domain shows phage adaptation at molecular resolution

Deep mutational scanning (DMS) is a high-throughput experimental technique to characterize sequence-function relationships through large-scale mutagenesis coupled to selection and deep sequencing. The scale and depth of DMS is used to reveal sites critical for activity, host specificity, and stability in a protein. DMS has been employed to study many proteins including enzymes, transcription factors, signaling domains, and viral surface proteins (Fowler and Fields, 2014; Lee et al., 2018; Raman et al., 2014; Romero et al., 2015).

Bacteriophage T7 is a podovirus that infects *E. coli*. T7 has a short non-contractile tail made up of three proteins including the tail fiber encoded by *gp17*. Each of the six tail fibers is a homotrimer composed of a relatively rigid shaft ending with a β-sandwich tip domain connected by a short loop (Garcia-Doval and Raaij, 2012). The tip domain is likely the very first region of the tail fiber to interact with host lipopolysaccharide (LPS) and position the phage for successful, irreversible binding with the host (González-García et al., 2015; Molineux, 2001; Qimron et al., 2006). The tip domain is a major determinant of host range and activity and is often naturally exchanged between phages to readily adapt to new hosts (Fraser et al., 2006, 2007; Lin et al., 2012). Even single amino acid substitutions to this domain are sufficient to alter host range between *E. coli* strains (Heineman et al., 2008). Due to its critical functional role, we chose the tip domain to comprehensively characterize phage activity and host range by DMS.

We generated a library of 1660 single mutation variants of the tip domain, pre-specified as chip-based oligonucleotides, where all nineteen non-synonymous and one nonsense substitution were made at each codon spanning residue positions 472-554 (Figure 2A, residue numbering based on PDB 4A0T). Using ORACLE, the library was inserted into T7 to generate variants to be selected and deep sequenced (Figure 2B) on three laboratory *E. coli* hosts: B strain derivative BL21, K-12 derivative BW25113 and DH10B derivative 10G. Each variant was given a functional score, F, based on the ratio of their relative abundance before and after selection consisting of an estimated four infection cycles, which was then normalized to wildtype to yield F_N where wildtype $F_N = 1$ (Figure 2C-E, see methods). Selection on each host gave excellent correlation across biological triplicates (Figure 2 - Figure Supplement 1). To validate the functional relevance of the screen, we hypothesized that the flexible C-terminal end (residue positions 552-554 and a three-residue extension if the stop codon is substituted) is unlikely to have any structural or host recognition role. As expected, these positions broadly tolerated nearly all substitutions across all three hosts indicating that the functional scores likely reflect true biological effects (Figure 2C-E).

We compared the activities of phage variants across hosts to assess their fitness and evolutionary adaptation to each host. Between the three hosts, T7 variants appeared most and least adapted to BW25113 and 10G, respectively, as evidenced by the fraction of depleted variants ($F_N<0.1$) after selection on each host (10G: 0.66±0.03, BL21: 0.59±0.01, and BW25113: 0.51±0.01, all significantly different from each other with $p<0.05$) (Figure 2 – Figure Supplement 2A-C). Furthermore, wildtype T7 fared relatively

poorly on 10G (F=0.77±0.05), indicating a fitness impediment, but performed significantly better on BL21 (F=2.92±0.2, p<0.01) and BW25113 (F=2.26±0.1, p<0.01) (Figure 2 – Source Data 1). The fitness impediment gave many more variants competitive advantage resulting in greater enrichment ($F_N \geq 2$) over wildtype on 10G (48 variants) compared to BL21 (2 variants) and BW25113 (16 variants) (Figure 2 – Figure Supplement 2A-C). In fact, the best performing variants on 10G were 10 times more enriched than wildtype suggesting substantially higher activity (Figure 2F and Figure 2 – Figure Supplement 2D). Examining enriched variants on each host ($F_N \geq 2$) provides compelling evidence of the tradeoff between activity and host range (Figure 2F and Figure 2 – Figure Supplement 2E). The top ranked variants on each host were remarkably distinct from those on other hosts (except G479Q shared between 10G and BL21). Hierarchical clustering of F_N across all three hosts revealed grouping of similar variants that performed better selectively on some hosts but not others (Figure 2 – Figure Supplement 3). No variant performed exceptionally well on all hosts (F_N>2, Figure 2F), however 406 variants were tolerated on all three hosts (Figure 2 – Source Data 2). Thus, specialization toward a host comes at the cost of sacrificing breadth, mirroring observations made of natural phage populations (Elena et al., 2009).

We investigated the global physicochemical properties and topological preferences of substitutions after selection on each host (Figure 2 - Figure Supplement 2F-H). On 10G, there was enrichment of larger and more hydrophilic amino acids and depletion of hydrophobic amino acids (all p<0.001, r>0.12), which is visually striking on the heatmap (see R, K and H on Figure 2C). In contrast, no significant enrichment or

depletion was observed on BL21 (Figure 2 – Figure Supplement 2F-H). This is consistent with our earlier observation that wildtype T7 is generally well adapted to BL21 since it had the fewest variants outperforming wildtype. We reasoned that since BL21 has been used to propagate T7, it may have already adapted well to this host over time. On BW25113, hydrophobic residues were modestly enriched (all $p<0.034$, $r>0.07$) Figure 2 - Figure Supplement 2H), a trend opposite to 10G. This provides a molecular explanation as to why high scoring substitutions on one host fare poorly on others (Figure 2F). We mapped positions of enriched substitutions ($F_N \geq 2$) on each host onto the structure to determine topologically distinct patterns of substitution that may be masked in global comparisons of the entire tip domain (Figure 2 - Figure Supplement 2E). These fall predominantly on four exterior loops (BC, DE, FG, and HI), the adjoining region (β-strand H) close to exterior loop HI, and less frequently on the 'side' of the tip domain. This suggests directionality to phage-bacterial interactions and orientational bias of the tip domain with respect to the bacterial surface. Directionality and orientational bias is particularly valuable information since no high-resolution structure of this phage bound to receptor exists.

Several key lessons emerged from these host screens. First, single amino acid substitutions alone can generate broad functional diversity, highlighting the evolutionary adaptability of the RBP. Second, T7 can be optimized and activity can be increased, even on hosts that T7 is already considered to grow well on. Third, enrichment patterns on each host follow broad trends but have nuance at each position.

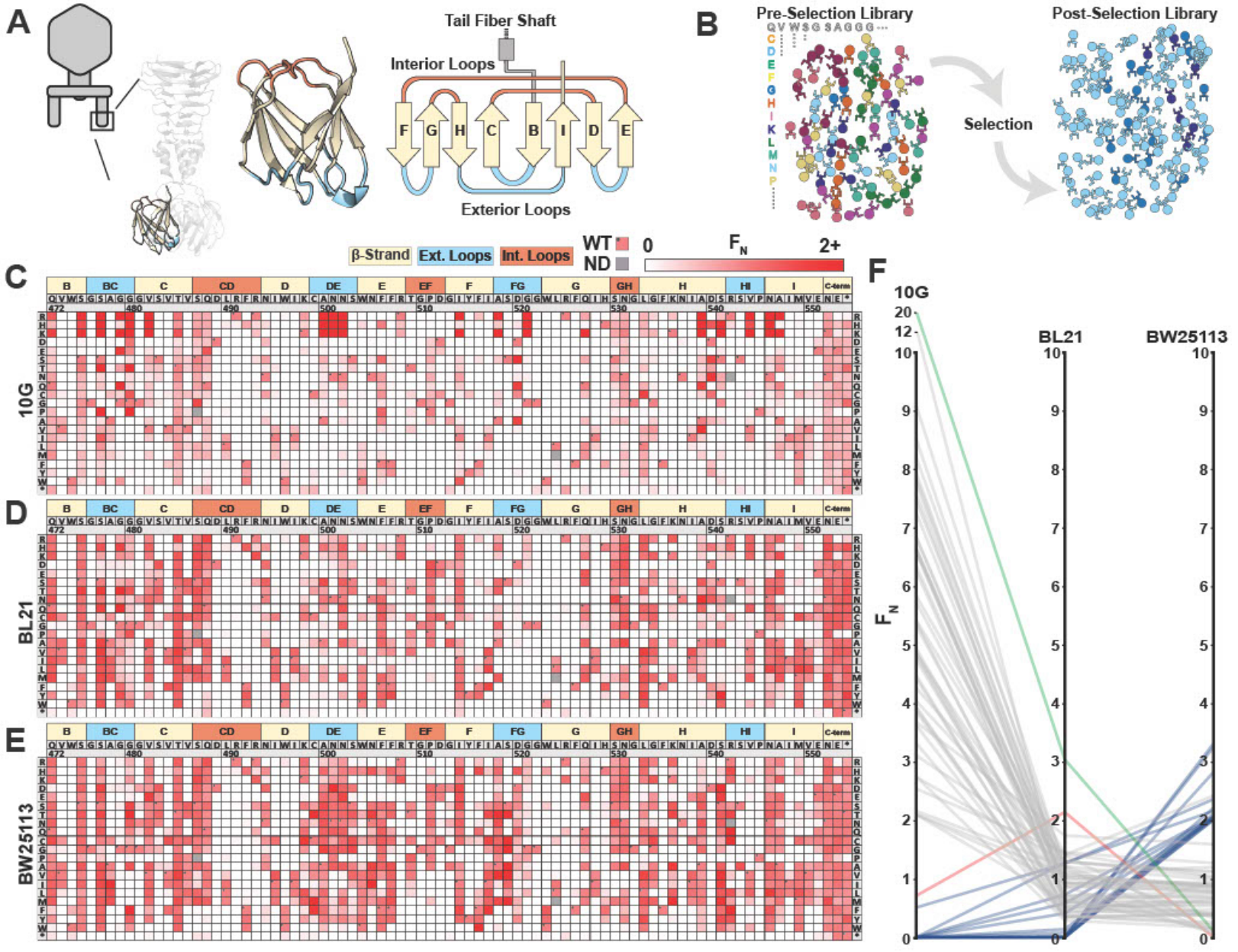

A
Tail Fiber Shaft
Interior Loops
F G H C B I D E
Exterior Loops
B
Pre-Selection Library
Post-Selection Library
Selection
β-Strand
Ext. Loops
Int. Loops
WT
ND
0
F_N
2+
C
10G
D
BL21
E
BW25113
F
10G
BL21
BW25113
F_N

Figure 2. Deep mutational scanning of tip domain shows phage adaptation at molecular resolution

(A) Crystal structure and secondary structure topology of the tip domain color coded as interior loops (red), β-sheets (beige) and exterior loops (blue) **(B)** Functional analysis of variants by comparing their abundances pre- and post-selection on a host. **(C-E)** Heat maps showing normalized functional scores (F_N) of all substitutions (red gradient) and wildtype amino acid (F_N=1 and black dot upper left) at every position for *E. coli* 10G (**C**), BL21 (**D**) and BW25113 (**E**). Residue numbering (based on PDB 4A0T), wildtype amino acid and secondary structure topology are shown above left to right, substitutions listed top to bottom. **(F)** Parallel plot showing FN for enriched (F_N≥2) variants on 10G, BL21, and BW25113. Coloring indicates enrichment only on 10G (grey), only on BL21 (red), only on BW25113 (blue) enriched on 10G and BL21 (green). Connecting lines indicate FN of the same variant on other hosts.

Figure 2 – Figure Supplement 1. Correlation between biological replicates after selection of phage variant library on *E. coli* 10G, BL21 and BW25113.

Figure 2 – Figure Supplement 2. Distribution, enrichment profile and physicochemical properties of variants after selection on *E. coli* 10G, BL21 and BW25113.

Figure 2 – Figure Supplement 3. Hierarchical clustering of phage variants by F_N score.

Figure 2 – Source Data 1. Deep sequencing summary for phage variant library after selection on different hosts. Related to Figure 2C-E.

Figure 2 – Source Data 2. Variant-specific FN for phage variants after selection on *E. coli* 10G, BL21 and BW25113 and physicochemical statistics. Related to Figure 2C-F, Figure 2 – Figure Supplement 2, and Figure 3.

Comparison across hosts reveals regions of functional importance

Next, we sought to elucidate features of each residue unique to each host or common across all hosts. There were over 30 residues with contrasting substitution patterns between different hosts, revealing fascinating features of receptor recognition for T7 (Figure 2 – Source Data 2). Here we focus on five of these residues, N501, R542, G479, D540, and D520, which showed starkly contrasting patterns of selection (Figure 3A). N501 and R542 are located on exterior loops oriented away from the phage and toward the receptor (Figure 3C). In fact, R542 forms a literal 'hook' to interact with the receptor (Garcia-Doval & Raaij, 2012). On 10G and BL21, only positively charged residues (R, K and H) were tolerated at residues 501 and 542, while in contrast many more substitutions were tolerated at both residues on BW25113. One such substitution, R542Q, is the best performing variant on BW25113 (F_N = 3.31) but is conspicuously depleted on 10G and BL21 suggesting that even subtle molecular disparities can lead to large biases in activity. The substitution profiles of G479 and D540 are loosely the inverse of N501 and R542 as many substitutions are tolerated on BL21 and 10G, but very few are tolerated on BW25113 (Figure 3A). We hypothesize that D540 is critical for host-recognition on BW25113. Since D540, a receptor-facing position on an exterior loop, is only 6 Å from G479, it is likely that any substitution at G479 may sterically hinder D540, resulting in the noted depletion of G479 substitutions on BW25113. This hypothesis is further supported by enrichment of adjacent S541D on BW25113 (F_N = 2.82, the third highest scoring substitution), while this substitution is depleted on 10G and BL21 (Figure 2 – Source Data 2). D520 displays a third variation in substitution patterns where substitutions are generally tolerated on 10G and BW25113, but not tolerated on BL21

(Figure 3A). This loop is also oriented downwards towards the receptor and we hypothesize that D520 or the local region around this exterior loop is more important for receptor recognition in BL21 than it is for the other two hosts, mirroring the result for D540 for BW25113. Another stark contrast can be drawn at adjacent S519, where no substitutions are tolerated in BL21 or 10G but several substitutions are enriched on BW25113, indicating substitutions can improve receptor binding on one host while reducing function on another host. Overall, these host-specific substitution patterns reveal a nuanced relationship between the tip domain composition and receptor preferences.

We quantitatively characterized the role of every residue by integrating selection data across all hosts to reveal a functional map of the tip domain at granular resolution (Figure 3B and 3C, Figure 3 – Source Data 1). We classified every residue as ‘intolerant’, ‘tolerant’ or ‘functional’ based on aggregated F_N scores of all substitutions across all three hosts at every residue. Our method of classifying functional regions was robust to adjusting the F_N threshold used to identify functional variants (Figure 3 – Figure Supplement 2). Residues where the majority of substitutions were depleted were considered intolerant to substitution, while residues where at least a third of substitutions were depleted in one host and tolerated or enriched in another host were considered functional; the remaining positions were considered tolerant (see methods). The hydrophobic core comprising W474, I495, W496, I497, Y515, W523, L524, F526, I528, F535 and I548 is essential for stability and therefore is highly intolerant to substitutions (Figure 3B). Other intolerant positions include an elaborate network of salt bridge interactions involving D489, R491, R493, R508 and D512 in the interior loops, which likely constrain the orientation of the tip domain relative to the shaft (Figure 3C). Glycines

generally provide conformational flexibility between secondary structure elements and normally tend to be mutable. Interestingly, several glycines (G476, G510, G522 and G532) are highly intolerant to substitutions. These glycines may be essential to minimize steric obstruction to adjacent larger residues, similar to G479 and D540 on BW25113 (Figure 3C). For example, G510 and G532 may facilitate formation of salt bridges in the interior loop, while G476 and G522 may facilitate a required receptor interaction in exterior loops for all three hosts.

It has been previously assumed that exterior loops are the primary functional region of the tip domain (Garcia-Doval and Raaij, 2012; Yehl et al., 2019). We found that functional positions did typically point outward and are densely concentrated along exterior loops BC, DE, FG, and HI, as well as adjacent β-sheet residues. This is consistent with two specificity switching substitutions found in a previous study, D520Q and V544A, which are both located in exterior loops (Heineman et al., 2008). However, several residues in exterior loops, such as G476 and S543, were notably intolerant, indicating these residues may be poor targets for engineering or future combinatorial studies. Functional positions were also found in regions other than exterior loops, such as I514, Q527 and K536 which are β-sheet residues located along one side of the tip domain (Figure 3C). This suggests the phage can use the ‘side’ of the tip domain to engage the receptor, increasing the apparent functional area of the tip domain and highlighting several new regions as valuable engineering targets.

We also determined if the functionally important regions could be predicted computationally, as the ability to predict functionally important regions without DMS could rapidly accelerate engineering efforts. We used Rosetta, a state-of-the-art protein

modeling software, to calculate the change in Gibbs free energy (ΔΔG) for each of the 1660 mutations and compared this distribution to our DMS results (Figure 3 – Figure Supplement 1, also see methods) and generated a truth table to summarize results compared to our functional data (Figure 3- Figure Supplement 3). Predicted thermodynamic changes in stability mapped very well with over 93% of tolerated or functional positions having a substitution that was predicted to be stabilizing. The remaining 7% of tolerant or enriched substitutions were predicted to be destabilizing, and we hypothesis that this may indicate these substitutions result in improved dynamic or induced fit positioning of the tip domain for productive infection. Incorporating stability estimations could further improve the engineering power of the assay. For example, substitutions predicted to be stable but that are intolerant in the DMS assay may indicate that the substituted residue is necessary for all three hosts.

Overall, these results paint a complex enrichment profile for each host with some broad trends but subtle host-specific effects. These results suggest that exterior loops and some outward facing positions in β-sheets act as a reservoir of function-switching and function-enhancing mutations, likely promoting host-specific and orientation-dependent interactions between phage and bacterial receptors. Functional positions identified by this comparison are ideal engineering targets to customize host range and activity.

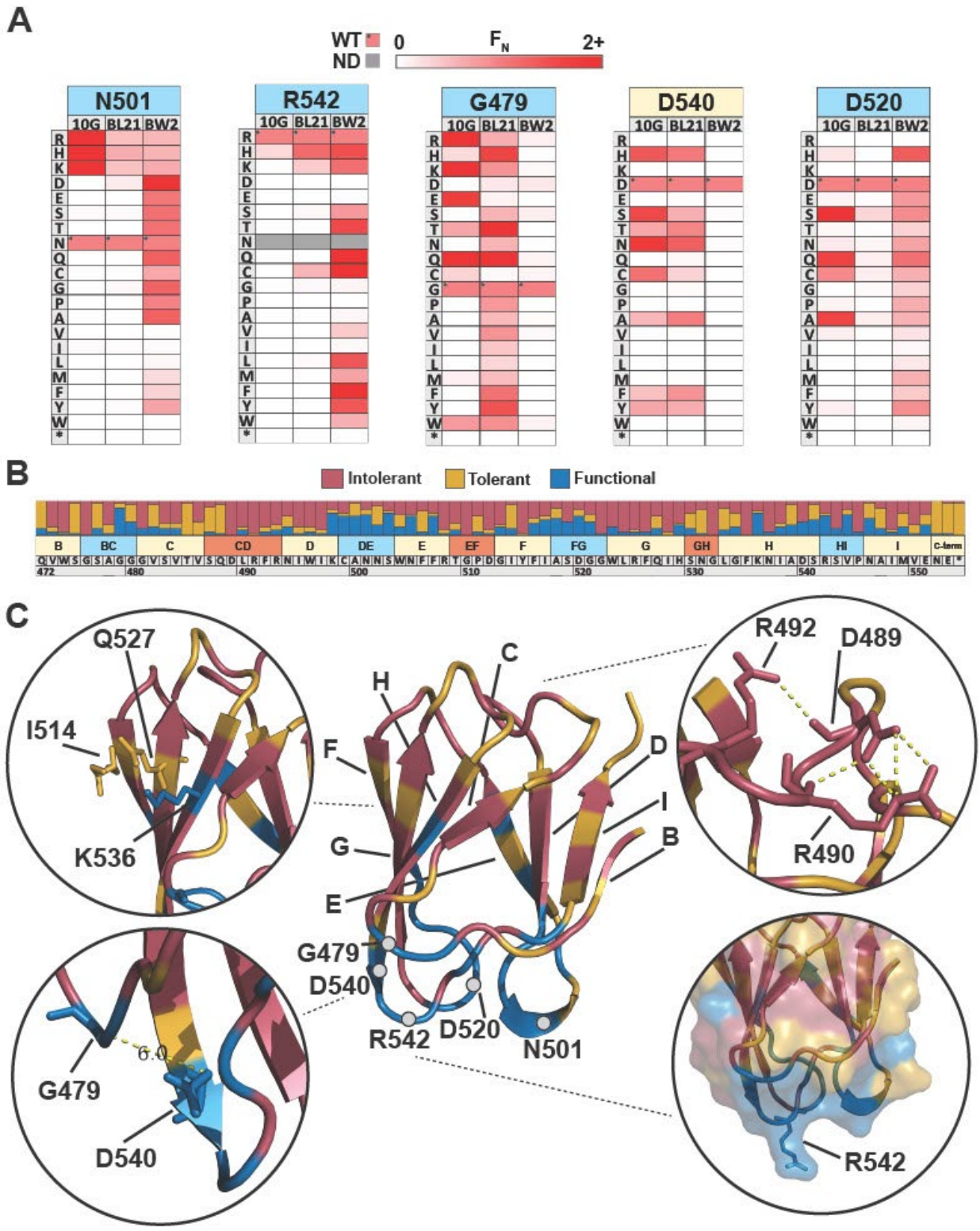

A
WT
ND
0
F_N
2+
N501
R542
G479
D540
D520
10G BL21 BW2
B
Intolerant
Tolerant
Functional
B BC C CD D DE E EF F FG G GH H HI I C-term
C
Q527
I514
K536
H
C
F
D
I
G
B
E
G479
D540
R542
D520
N501
6.0
R492
D489
R490
G479
D540
R542

Figure 3. Comparison across hosts reveals regions of functional importance

(A) Host-specific differences in substitution patterns at five positions in the tip domain recapitulated from Figure 2. **(B)** Role of each position determined by aggregating scores of all substitutions in all hosts at that position. Substitutions are classified as intolerant ($F_N < 0.1$ in all hosts), tolerant ($F_N \geq 0.1$ in all hosts), or functional ($F_N < 0.1$ in one host, $F_N \geq 0.1$ in another host) and bar plots are shown as proportion of classified variants at that position. **(C)** Crystal structure of the tip domain (center) with each residue colored as intolerant, tolerant, or functional based on the dominant effect at that position, β-sheets and residues listed in **(A)** are labeled. Key interactions defining function and orientation are highlighted in peripheral panels.

Figure 3 – Figure Supplement 1. Comparing FD to computationally predicted stability of variant ΔΔG.

Figure 3 – Figure Supplement 2. Classification of tolerant, intolerant and functional regions based on different cutoff conditions.

Figure 3 – Figure Supplement 3. Truth table comparing functional results to predicted stability.

Figure 3 – Source Data 1. Functional comparison for each variant on susceptible hosts. Related to Figure 3B-C.

Figure 3 – Source Data 2. ΔΔG and FD Conversion for all variants. Related to Figure 3 – Figure Supplement 1

Discovery of gain-of-function variants against resistant hosts

The tail fiber is considered a reservoir of gain-of-function variants due to its principal role in determining fitness of a phage through host adsorption (Holtzman et al., 2020; Yehl et al., 2019). We hypothesized that novel gain of function variants against a resistant host could be discovered by subjecting our tail fiber variant library to selection on a resistant host. To identify a resistant host, we focused on host genes *rfaG* and *rfaD* involved in the biosynthesis of surface lipopolysaccharide (LPS) which is a known receptor for T7 in *E. coli* (González-García et al., 2015, 2015; Molineux, 2001; Qimron et al., 2006). Gene *rfaG* (synonyms *WaaG* or *pcsA*) transfers glucose to the outer core of LPS and deletion strains lack the outer core of LPS (Pagnout et al., 2019), while *rfaD* (synonyms *gmhD* or *WaaD*) encodes a critical epimerase required for building the inner core of LPS (Valvano et al., 2002) (Figure 4A). Deletion of either gene reduces the ability of T7 to infect *E. coli* by several orders of magnitude (Figure 4F). We challenged the library of T7 variants against *E. coli* deletion strains BW25113Δ*rfaG* and BW25113Δ*rfaD* through pooled selection and deep sequencing as before (Figure 2) and determined a F_N score for each substitution on both strains (Figure 4B-C). Independent replicates showed good correlation for BW25113Δ*rfaG* (R=0.99, 0.93, 0.93) but only adequate correlation for BW25113Δ*rfaD* (R=0.51, 0.68, 0.39) (Figure 4 - Figure Supplement 1). Although the scale of F_N was inconsistent across replicates on BW25113Δ*rfaD*, the same substitutions were largely enriched in all three replicates, suggesting reproducibility of results (Figure 4 – Figure Supplement 3). Inconsistencies in F_N scores may arise due to severe loss of diversity causing stochastic differences in enrichment to become magnified across independent experiments and the four infection cycles used for selection. Separately we

examined correlation after selection using only a single infection cycle, which produced more highly correlated results for BW25113Δ*rfaD* (R=0.89, 0.90, 0.89) (Figure 4 - Figure Supplement 4), indicating fewer infection cycles may be ideal for future work with highly resistant hosts.

We engineered several gain-of-function T7 variants that could infect both deletion strains with activity comparable to wildtype T7 infecting susceptible BW25113 (Figure 4F). Low sequence diversity and high enrichment scores of T7 variants indicates a strong selection bottleneck which is consistent with diminished activity of wildtype T7 on the deletion strains. This is reflected in the significantly lower functional score of wildtype T7 on BW25113Δ*rfaG* and BW25113Δ*rfaD* (F=0.09±0.3 and F=0.03±0.2, respectively) in comparison to BW25113 (F=2.26±0.1, $p<0.001$) (Figure 2 – Source Data 1). The number of enriched variants outperforming wildtype T7 ($F_N \geq 2$) on the deletion strains (BW25113Δ*rfaG*: 55 variants, 3.3% and BW25113Δ*rfaD*: 68 variants, 4.1%) was over 3 times higher than BW25113 (16 variants, 1%) but comparable to 10G (48 variants, 2.9%) (Figure 4 – Figure Supplement 2A-B). However, the enrichment scores of top performing variants such as G521H and G521R on BW25113Δ*rfaG* and S541K and N501H on BW25113Δ*rfaD* was over 100 times greater than wildtype T7, suggesting strong gain-of-function on the deletion strains (Figure 4 – Figure Supplement 2C). Of the 78 variants with $F_N \geq 2$ on either deletion strain, 45 variants had $F_N \geq 2$ on both strains indicating that variants that performed well on one strain typically performed well on the other strain. This implies the enriched variants may have broad affinity for truncated LPS but cannot discriminate based on the length of the LPS. Nonetheless, hydrophilic substitutions were more strongly enriched on *BW25113ΔrfaG* ($p<0.001$, $r>0.11$), but not as significantly on

BW25113ΔrfaD ($p<0.033$, $r<0.10$), suggesting subtle differences in surface chemical properties of deletion strains leading to host-specific enrichments (Figure 4 – Figure Supplement 2D-F). Indeed, there were several variants with contrasting F scores on both strains such as S541T (*BW25113ΔrfaD* $F_N = 44.8$, *BW25113ΔrfaG* $F_N = 0.6$) and G521E (*BW25113ΔrfaD* $F_N = 0$, *BW25113ΔrfaG* $F_N = 17.4$) suggest potential host preference. Most substitutions were concentrated in the exterior loops BG, FG, HI, and β-strand H, all pointing downwards towards the bacterial surface, reinforcing the functional importance of these regions of the tip domain (Figure 4D and 4E). Notably, the most enriched variants had large positively charged substitutions (K, R, and H) akin to the enrichment pattern on 10G, suggesting the bacterial surface of these truncated mutants likely resembles that of 10G. Our results are consistent with a recent continuous evolution study, which identified G480E and G521R as possible gain-of-function variants on a strain similar to *BW25113ΔrfaD* and G479R and G521S as possible gain of function variants on *BW25113ΔrfaG* (Holtzman et al., 2020), although these variants only represent a small fraction of the gain of function variants discovered in our study.

We validated the results of the pooled selection experiment by clonally testing the ability of phage variants with high F_N (A539R, G521H, and D540S) to plaque on the deletion strains based on a standard efficiency of plating assay (EOP). Indeed, EOP results showed significant gain of function in these variants on the deletion strains (Figure 4F). D540S was particularly noteworthy as it performed better on the deletion strain *BW25113ΔrfaG* over wildtype BW25113 by 1-2 orders of magnitude. Based on these results, we conclude that D540 is critical for infecting wildtype BW25113 (Figure 3) likely by interacting with the outer core of LPS. When the outer core of the LPS is missing

(*BW25113ΔrfaG*), a substitution at this position becomes necessary for adsorption either to a different LPS moiety or to an alternative receptor.

We introduced stop codon at every position to systematically evaluate the function of tip domains truncated to different lengths. Many truncated variants performed well, especially on BW25113Δ*rfaG* which included some with $F_N \geq 10$ (Figure 4 – Source Data 2). Truncated variants that performed well are distributed throughout the tip domain and are not localized to any one region (Figure 4 – Figure Supplement 5). We clonally tested variant R525*, the best performing truncated library member (BW25113Δ*rfaG* $F_N = 9.55$, BW25113Δ*rfaD* $F_N = 75.7$), and found that this mutant showed no ability to plaque on any host unless provided the tail fiber *in trans*. These truncated phages, detectable here only using deep sequencing, may demonstrate how obligate lytic phages could become less active in a bacterial population, slowly replicating alongside their bacterial hosts, requiring only a single mutation to become fully active again. In fact, acceptor phages altogether lacking a tail fiber were present at extremely low abundance (Figure 4 – Source Data 1). These phages are not artifacts from library creation as some ability to replicate is required to produce detectable concentrations of each phage. We concluded that these are viable phage variants albeit with a much slower infection cycle resulting in their inability to form visible plaques.

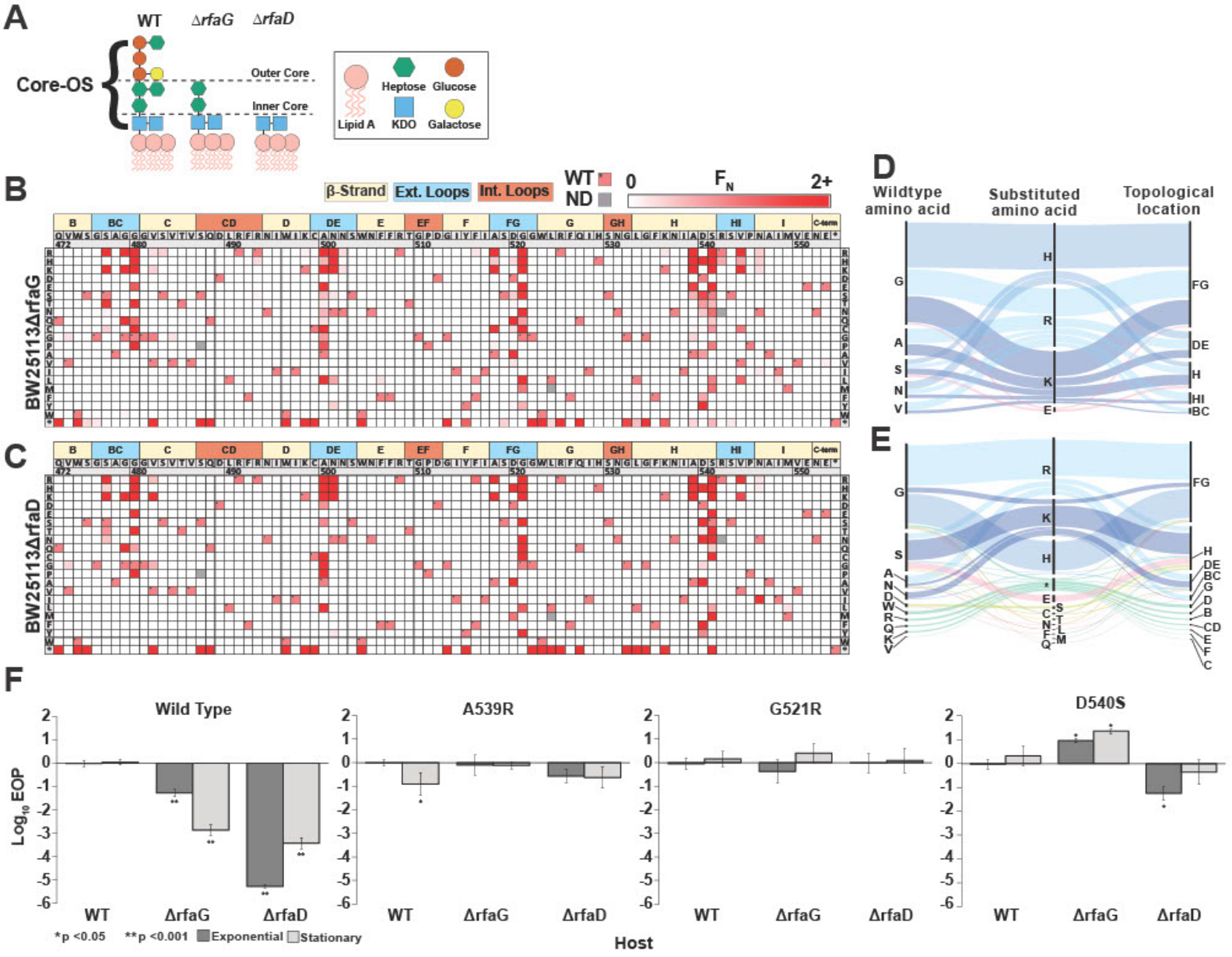

A
WT
ΔrfaG
ΔrfaD
Core-OS
Outer Core
Inner Core
Heptose
Glucose
Lipid A
KDO
Galactose
B
β-Strand
Ext. Loops
Int. Loops
WT
ND
0
F_N
2+
BW25113ΔrfaG
C
BW25113ΔrfaD
D
Wildtype amino acid
Substituted amino acid
Topological location
E
F
Wild Type
A539R
G521R
D540S
Log10 EOP
WT
ΔrfaG
ΔrfaD
*p <0.05
**p <0.001
Exponential
Stationary
Host

Figure 4. Discovery of gain-of-function variants against resistant hosts

(A) Schematic view of the LPS on wildtype BW25113, BW25113Δ*rfaG* and BW25113Δ*rfaD*. **(B-C)** Heat maps showing normalized functional scores (F_N) of all substitutions (red gradient) and wildtype amino acid (F_N=1 and black dot upper left) at every position for BW25113Δ*rfaG* **(B)** and BW25113Δ*rfaD* **(C) (D-E)** Among highly enriched variants ($F_N \geq 10$), targeted amino acids (left), their substitutions (middle) and topological location on the structure (right) on BW25113Δ*rfaG* **(D)** and BW25113Δ*rfaD* **(E)**, with each alluvial colored based on the substituted amino acid and scaled by FN. **(F)** EOP (mean ± SD, biological triplicates) for wildtype phage and select variants on BW25113 (Wild Type), BW25113Δ*rfaG* and BW25113Δ*rfaD* in exponential (dark gray) and stationary phases (light gray) using BW25113 as a reference host.

Figure 4 – Figure Supplement 1. Correlation between biological replicates after selection of phage variant library on *E. coli* BW25113*ΔrfaG* and BW25113*ΔrfaD*.

Figure 4 – Figure Supplement 2. Distribution, enrichment profile and physicochemical properties of variants after selection on *E. coli* BW25113*ΔrfaG* and BW25113*ΔrfaD*.

Figure 4 – Figure Supplement 3. Ranking FN of the ten most enriched variants in each biological replicate for (A) *E. coli* BW25113ΔrfaG and (B) BW25113ΔrfaD.

Figure 4 – Figure Supplement 4. Correlation between biological replicates for selection of phage variant library after one infection cycle on (A) *E. coli* BW25113ΔrfaG and (B) BW25113ΔrfaD.

Figure 4 - Figure Supplement 5. Location of truncations in the tip domain enriched after selection.

Figure 4 – Source Data 1. Deep sequencing summary for phage variant library after

selection on different hosts. Related to Figure 4B-C.

Figure 4 – Source Data 2. Variant-specific FN for phage variants after selection on *E. coli* BW25113*ΔrfaG* and BW25113*ΔrfaD* and physicochemical statistics. Related to Figure 4B-C and Figure 4 – Figure Supplement 2.

Targeting pathogenic *E. coli* causing urinary tract infection using T7 variants

Phage therapy is emerging as a promising solution to the antibiotic resistance crisis. Recent clinical success stories against multidrug resistant *Acinetobacter* and *Mycobacterium* showcase the enormous potential of phage therapy (Dedrick et al., 2019; Schooley et al., 2017). Despite notable exceptions, in general development of effective phage-based therapeutics is hindered by onset of bacterial resistance resulting in low phage susceptibility. Although initial application of phages in a laboratory setting may reduce bacterial levels, the residual bacterial load remains high, causing bacteria to quickly recover after phage application (Fister et al., 2016; Huss and Raman, 2020; Silva et al., 2014). A high ratio of phage to bacteria (multiplicity of infection or MOI) may productively kill bacteria in a laboratory setting by overwhelming a host with many phages (Abedon, 2011). However, ensuring an overwhelming amount of phages in a clinical setting is not always feasible (Principi et al., 2019). Engineering highly active phages that overcome bacterial insensitivity and can therefore productively eliminate bacterial populations at low MOI in a laboratory setting would greatly enhance phage-based therapeutics. We hypothesized that engineered tip domain variants may abate bacterial insensitivity and be active even at low MOI by better adsorbing to the native receptor or recognizing a new bacterial receptor altogether.

To test this hypothesis, we chose pathogenic *E. coli* strain isolate 473 isolated from a patient with urinary tract infection (UTI) (Arthur et al., 1990). Although T7 can infect this UTI strain, insensitivity arises rapidly, a phenomenon all too common with the use of natural phages. EOP assays for wildtype T7 showed insensitive plaque morphology consisting of small, slow-growing plaques. No visible lysis was detected after overnight incubation when wildtype T7 was applied in liquid culture (MOI=1) indicating onset of insensitivity. However, the variant library applied on the UTI strain cleared the culture (MOI=1) after overnight incubation, suggesting T7 variants are capable of lysing and attenuating insensitivity existing in the pool. We clonally characterized three variants (N501H, D520A and G521R) isolated from plaques. All three variants vastly outperformed wildtype T7 in terms of onset of insensitivity. Insensitivity emerged approximately 11-13 hours after initial lysis for the three variants whereas it took merely 1-2 hours after initial lysis for wildtype (Figure 5A). In particular, the N501H variant lysed cells faster and produced a lower bacterial load post lysis, suggesting far greater activity compared to wildtype T7. Next, we compared the effect of phage MOI (MOI=10^{2}-10^{-5}) on the lysing activity of N501H and wildtype T7 (Figure 5B). At all MOIs wildtype phages lysed UTI473 significantly more slowly compared to N501H phages (all $p<0.05$). At lower MOI, time to lysis of N501H was half that of wildtype T7, though they were more comparable at higher MOI.

A striking contrast between N501H and wildtype T7 is evident in reduced bacterial insensitivity at progressively lower MOI (Figure 5C). Between a MOI of 100 to 1, application of both N501H and wildtype phage resulted in similar bacterial insensitivity. However, between an MOI of 10^{-1} and 10^{-5}, application of N501H phage reduced

insensitivity over a 10-hour window, while application of wildtype phages resulted in rapid onset of insensitive bacteria. We postulate that at high MOI, wildtype T7 simply overwhelms the host before insensitivity arises, while at lower MOI, insensitivity can emerge, and only variants adapted to the host can effectively kill the host. These results indicate that ORACLE can generate phage variants superior to wildtype phage that could then become starting points for further engineering therapeutic phages. Further experiments will be required to assess the *in vivo* efficacy of the T7 variants.

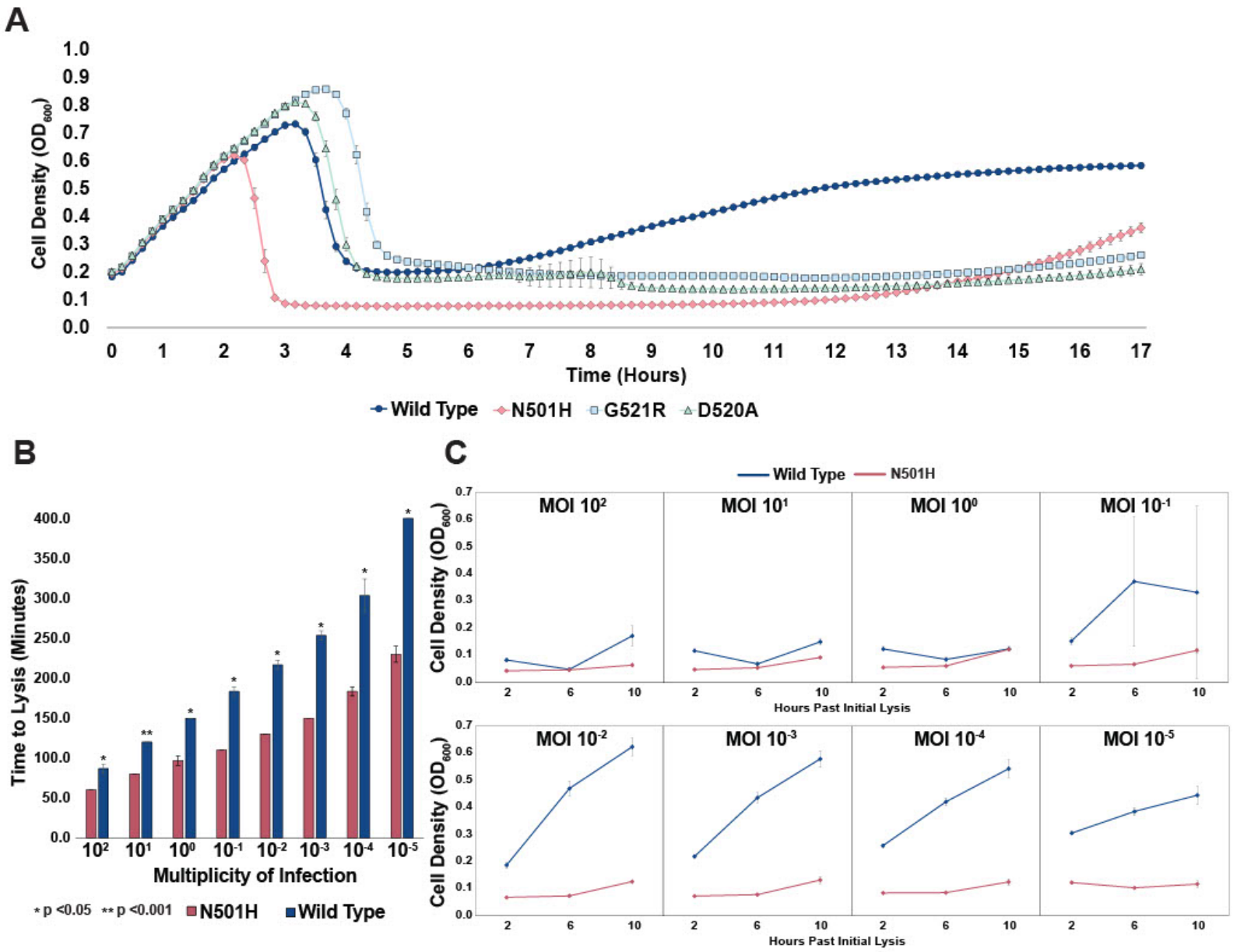

A
Cell Density (OD_{600})
Time (Hours)
Wild Type
N501H
G521R
D520A
B
Time to Lysis (Minutes)
Multiplicity of Infection
* p <0.05 ** p <0.001
N501H
Wild Type
C
Wild Type
N501H
MOI 10^{2}
MOI 10^{1}
MOI 10^{0}
MOI 10^{-1}
MOI 10^{-2}
MOI 10^{-3}
MOI 10^{-4}
MOI 10^{-5}
Cell Density (OD_{600})
Hours Past Initial Lysis

Figure 5. Targeting pathogenic *E. coli* causing urinary tract infection using T7 variants

(A) Growth time course of UTI473 strain subject to wildtype T7 and select variants. Phages were applied after an hour at an MOI of ~10-2 **(B)** Estimated time to lysis of UTI473 strain incubated with wildtype T7 and N501H variant over a range of MOIs, derived from time course experiments. **(C)** Cell density (OD600) of UTI473 strain when incubated with wildtype T7 and N501H variant at select timepoints after initial lysis. All data represented as mean ± SD of biological triplicate.

Host range constriction emerges from global comparison across variants

Most phages are specialists that selectively target a narrow range of hosts but are unable to productively infect other closely related hosts (Hyman and Abedon, 2010). We wanted to assess differences in the host range of individual variants on 10G, BL21 and BW25113 and identify variants with constricted host ranges. Ideally, host specificities can be determined by subjecting a co-culture of all three hosts to the phage library. However, deconvolving specificities of thousands of variants from a pooled co-culture experiment can be technically challenging. Instead, we sought to estimate specificities by comparing F_N of a phage variant on all three hosts. Although F_N compares activity of variants within a host, it could nonetheless be a useful proxy for estimating specificities across hosts. For instance, a phage variant with high F_N on BL21 but completely depleted on BW25113 is more likely to specifically lyse BL21 than BW25113 in a co-culture experiment. Based on this rationale, we considered different metrics of comparison of F_N, and settled on difference in F_N of a variant with reduced weight for enrichment (or F_D, see methods)

between any two hosts as an approximate measure of host preference. This metric is not an absolute measure of host specificity, but one devised to reveal broad trends in specificity to prioritize variants for downstream validation.

To assess if variants preferred one host over another, we computed F_D for all three pairwise combinations and plotted functional substitutions as points on or above/below a 'neutral' line (Figure 6A-C). Variants above the line favor lysis of the noted host, and vice versa for variants below the line. To check if this F_D-based approach is suitable for assessing host specificity, we compared our results with previously published data. Two substitutions, D520Q and V544A, that were reported to have a preference for BW25113 and BL21, respectively, in head to head comparisons (Heineman et al., 2008) were placed correctly in our plots, confirming the validity of our F_D-based classification scheme. We identified 118 out of 1660 variants as good candidates for constricting host range ($|F_D| \geq 1$, see Figure 6 – Source Data 1). Of the 118 variants, 53 variants favor BW25113 over BL21 and 98 variants favor BW25113 over 10G in pairwise comparisons (Figure 6A and 6C). Between BL21 and 10G, there are 15 variants that favor BL21 but none that favor 10G (Figure 6B).

Certain key positions including G479, D540, R542 and D520 which we previously identified as functionally important (Figure 3A) are the molecular drivers of specificity between hosts (Figure 6A-C). Taken together, our data suggests that it would be easier to find a variant capable of specifically lysing BW25113, less so for BL21, and most challenging for 10G.

To validate our analysis, we clonally tested variant R542Q which had a greater preference for BW25113 than BL21 or 10G (BW25113 F_D =2.0, BL21 F_D =0, 10G F_D=0)

and variant D540S which had a greater preference for BL21 and 10G than BW25113 (10G F_D = 1.03, BL21 F_D= 0.62, BW25113 F_D = 0.03). Indeed, R542Q showed a significant ~100-fold decrease in the ability to plaque on BL21 compared to BW25113 while 10G plaques were atypically small, indicating a severe growth defect (Figure 6D). In contrast, D540S showed a significant ~100-fold decrease in the ability to plaque on BW25113 compared to BL21 and 10G (Figure 6D), confirming the host constriction properties of these variants. In summary, pairwise comparison is a powerful tool to map substitutions that constrict host range and can be leveraged to tailor engineered phages for targeted hosts.

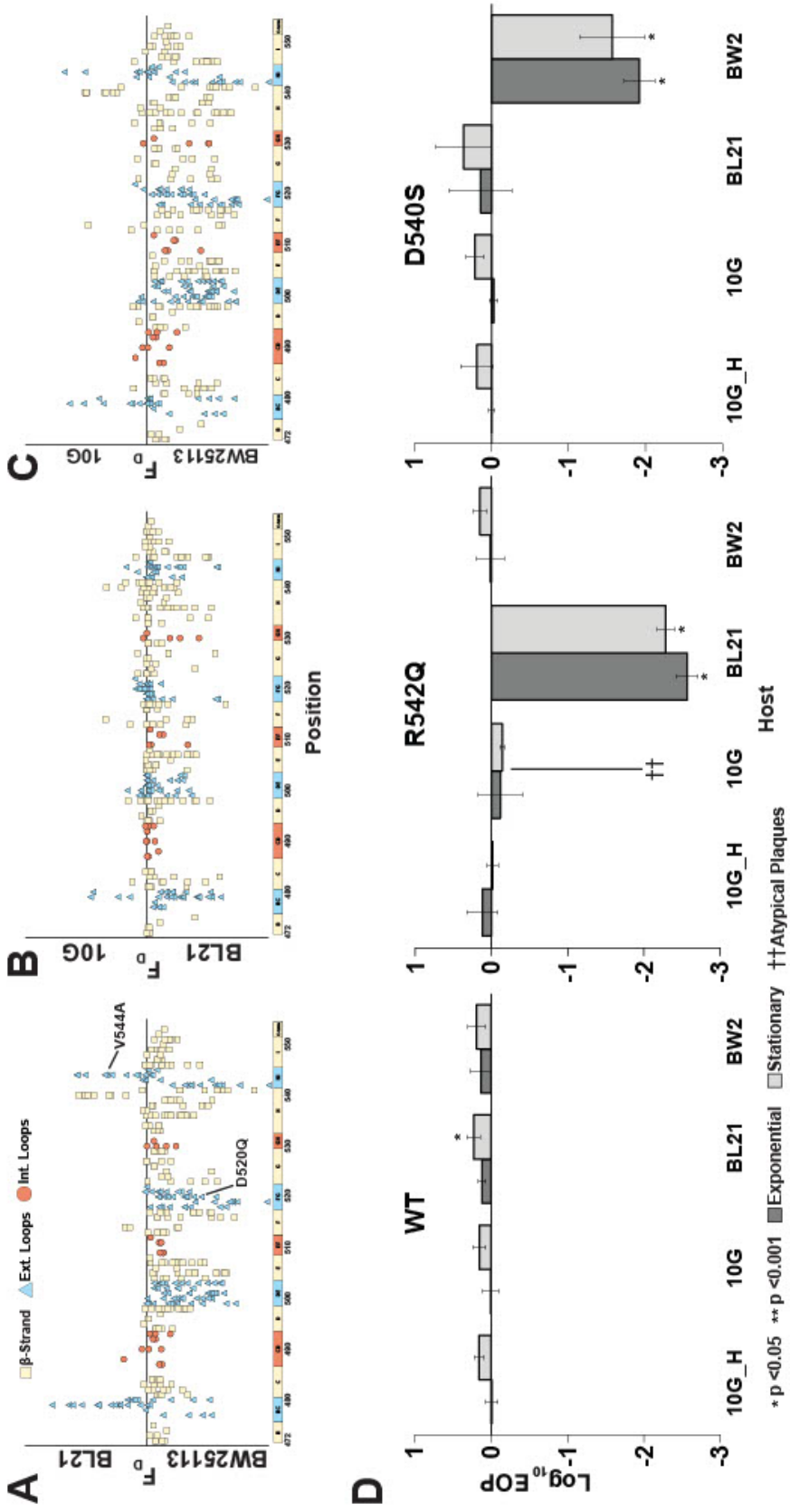

A
B
C
D
β-Strand
Ext. Loops
Int. Loops
V544A
D520Q
BW25113
BL21
10G
Position
WT
R542Q
D540S
10G_H
10G
BL21
BW2
Log10EOP
Host
* p <0.05 ** p <0.001 Exponential Stationary ††Atypical Plaques

Figure 6. Host range constriction emerges from global comparison across variants

(A-C) Pairwise comparison of differences in functional scores of variants between hosts (see Methods). Variants above the line favor lysis of host noted above the line, and vice versa for variants below the line. **(D)** EOP (mean ± SD, biological triplicates) for wildtype T7 and select variants on BW25113, 10G and BL21 in exponential (dark gray) and stationary phases (light gray) using exponential 10G with gp17 tail fiber helper plasmid (10G_H) as a reference host. R542Q plaques are atypically small until EOP ~10-2.

Figure 6 – Source Data 1. ΔΔG and FD Conversion for all variants.

Discussion

In this study, we used ORACLE to create a large, unbiased library of T7 phage variants to comprehensively characterize the mutational landscape of the tip domain of the tail fiber. Our study identified hundreds of novel function-enhancing substitutions that had not been previously characterized. We mapped regions of function-enhancing substitutions on to the crystal structure to rationalize how sequence and structure influence activity and host range. Several important insights emerged from these results. Cross-comparison between different hosts and selection on resistant hosts allowed us to map key substitutions leading to host discrimination and gain of function. Single amino acid substitutions are sufficient to enhance activity and host range including some that confer dramatic increases in activity or specificity. The functional landscape on each host is unique, reflecting both different molecular preferences of adsorption and the fitness of wildtype T7 on these hosts. For instance, hydrophilic substitutions were enriched in 10G while hydrophobic substitutions were enriched in BW25113. Notably, substitutions on 10G (an *E. coli* K-12 derivative lacking LPS components) mirrored substitutions that recovered function on BW25113 mutants with truncated LPS which shows convergence of selection. Function-enhancing substitutions were densely concentrated in the exterior loops indicating an orientational preference for receptor recognition. However, they were also found on other surface residues, albeit less frequently, suggesting alternative binding modes of the tip domain for host recognition, and several intolerant residues were located in exterior loops. Taken together, these results highlight the extraordinary functional potential of the tip domain and rationalizes the pervasive use of this structural fold in nature for molecular recognition. Comparison of these functional profiles precisely reveals

regions that are ideal engineering targets for customizing host range and activity and identifies intolerant residues that should be avoided when engineering synthetic phages.

These results also highlight the power of deep sequencing to detect and resolve small functional effects over traditional low-throughput plaque assays. This is best illustrated in the case of truncated variants visible only to deep sequencing, but incapable of plaque formation without a helper plasmid. The truncated variants are likely not experimental artifacts as some ability to replicate is required to survive multiple rounds of selection on the host. Truncation of the tip domain may misorient the phage relative to the receptor, likely resulting in slower growth and deficiency in plaquing, while still capable of replicating. Since plaque formation is a complex process, inability to plaque may not imply a functionally incompetent phage.

ORACLE is designed as a foundational technology to elucidate sequence-function relationships in phage genes. On T7, ORACLE can be used to investigate the function of several important genes including the remainder of the tail fiber and tail structure, capsid components, or lysins and holins. Together, these will provide a comprehensive view of the molecular determinants of the structure, function, and evolution of a phage. Once the phage variants are created, scaling up ORACLE to investigate potentially tens of hosts merely scales up sequencing volume, not experimental complexity. Such a large-scale study will lead to a detailed molecular understanding and adaptability of phage bacterial interactions. Any phage with a sequenced genome and a transformable host capable of maintaining a plasmid library should be amenable to ORACLE because the phage variants are created during the natural infection cycle. This approach can be leveraged

to tune activity for known phages with high activity, such as T7, or to identify engineering targets that dramatically increase activity for newly isolated natural phages.

The confluence of genome engineering, high-throughput DNA synthesis, and sequencing enabled by ORACLE together with viral metagenomics could transform phage biology. Phages constitute unparalleled biological variation found in nature and are aptly called the “dark matter” of the biosphere. Their sequence diversity and richness are coming to light in the growing volume of viral metagenome databases. However, what functions these sequences encode remains largely unknown. For instance, fecal viromes estimate 10^8-10^9 virus-like particles per gram of feces, but less than a quarter of sequence reads align to existing databases (Reyes et al., 2012). While this knowledge gap is daunting, it also presents an opportunity to mine metagenomic sequences to characterize their function and engineer programmable phages. By enabling sequence programmability, we envision ORACLE as a powerful tool to discover new phage ‘parts’ from metagenomic sequences.

Acknowledgements

We would like to thank Dr. Rodney Welch for UTI473 strain and Dr. Douglas Weibel for BW25113 deletion mutant strains. We thank Dr. Karthik Anantharaman, Dr. John Yin, Laura Alexander and Chutikarn Chitboonthavisuk for critical review of the manuscript. This work is partially supported by the US Department of Agriculture Hatch award (WIS02066) and the Gates Grand Challenges grant (OPP1150209). A.M is supported by the Great Lakes Bioenergy Research Center (U. S. Department of Energy Award Number

DE-SC0018409). M.L is supported by NIH Molecular Biophysics Training Program T32 GM08293 and William H. Peterson Fellowship in Biochemistry. K.N is supported by NIH National Research Service Award T32 GM07215 and the Robert and Katherine Burris Biochemistry Fund.

Competing Interests

P.H and S.R have filed a provisional patent application on this technology. S.R is on the scientific advisory board of MAP/PATH LLC.

Figures, Figure Supplements and Supplementary Files

Figure 1. ORACLE workflow for creating phage variant libraries

- **Figure 1 – Figure Supplement 1.** Sequence rearrangements before and after recombinase-mediated cassette exchange.
- **Figure 1 – Figure Supplement 2.** Effect of Cas9-gRNA system on acceptor and control phages.
- **Figure 1 – Figure Supplement 3.** Correlation between expression library with and without DNAse treatment.
- **Figure 1 - Source Data 1.** Deep sequencing summary for phage variant expression library with and without DNAse treatment after ORACLE. Related to Figure 1E and Figure 1 - Figure Supplement 3.
- **Figure 1 - Source Data 2.** Percentage distribution of each variant in the expression library. Related to Figure 1E and Figure 1 - Figure Supplement 3.

Figure 2. Deep mutational scanning of tip domain shows phage adaptation at molecular resolution

- **Figure 2 – Figure Supplement 1.** Correlation between biological replicates after selection of phage variant library on *E. coli* 10G, BL21 and BW25113
- **Figure 2 – Figure Supplement 2.** Distribution, enrichment profile and physicochemical properties of variants after selection on *E. coli* 10G, BL21 and BW25113.
- **Figure 2 – Figure Supplement 3.** Hierarchical clustering of phage variants based on F_N score.

- **Figure 2 – Source Data 1.** Deep sequencing summary for phage variant library after selection on different hosts. Related to Figure 2C-E.
- **Figure 2 – Source Data 2.** Variant-specific F_N for phage variants after selection on *E. coli* 10G, BL21 and BW25113 and physicochemical statistics. Related to Figure 2C-F, Figure 2 - Figure Supplement 2 and Figure 3.

Figure 3. Comparison across hosts reveals regions of functional importance

- **Figure 3 – Figure Supplement 1.** Comparing F_D to computationally predicted stability of variant ΔΔG.
- **Figure 3 – Figure Supplement 2.** Classification of tolerant, intolerant and functional regions based on different cutoff conditions.
- **Figure 3 – Figure Supplement 3.** Truth table comparing functional results to predicted stability.
- **Figure 3 – Source Data 1.** Functional comparison for each variant on susceptible hosts. Related to Figure 3B-C.
- **Figure 3 – Source Data 2.** ΔΔG and F_D Conversion for all variants. Related to Figure 3 - Figure Supplement 1

Figure 4. Discovery of gain-of-function variants against resistant hosts

- **Figure 4 – Figure Supplement 1.** Correlation between biological replicates after selection of phage variant library on *E. coli* BW25113*ΔrfaG* and BW25113*ΔrfaD*.

- **Figure 4 – Figure Supplement 2.** Distribution, enrichment profile and physicochemical properties of variants after selection on *E. coli* BW25113*ΔrfaG* and BW25113*ΔrfaD*.
- **Figure 4 – Figure Supplement 3.** Ranking FN of the ten most enriched variants in each biological replicate for (A) E. coli BW25113ΔrfaG and (B) BW25113ΔrfaD.
- **Figure 4 – Figure Supplement 4.** Correlation between biological replicates for selection of phage variant library after one infection cycle on (A) E. coli BW25113ΔrfaG and (B) BW25113ΔrfaD.
- **Figure 4 - Figure Supplement 5.** Location of truncations in the tip domain enriched after selection.
- **Figure 4 – Source Data 1.** Deep sequencing summary for phage variant library after selection on different hosts. Related to Figure 4B-C.
- **Figure 4 – Source Data 2.** Variant-specific F_N for phage variants after selection on *E. coli* BW25113*ΔrfaG* and BW25113*ΔrfaD* and physicochemical statistics. Related to Figure 4B-C and Figure 4 – Figure Supplement

Figure 5. Targeting pathogenic *E. coli* causing urinary tract infection using T7 variants

Figure 6. Host range constriction emerges from global comparison across variants

- **Figure 6 – Source Data 1.** ΔΔG and F_D Conversion for all variants.

Supplementary File 1. List of primers used for experimentation

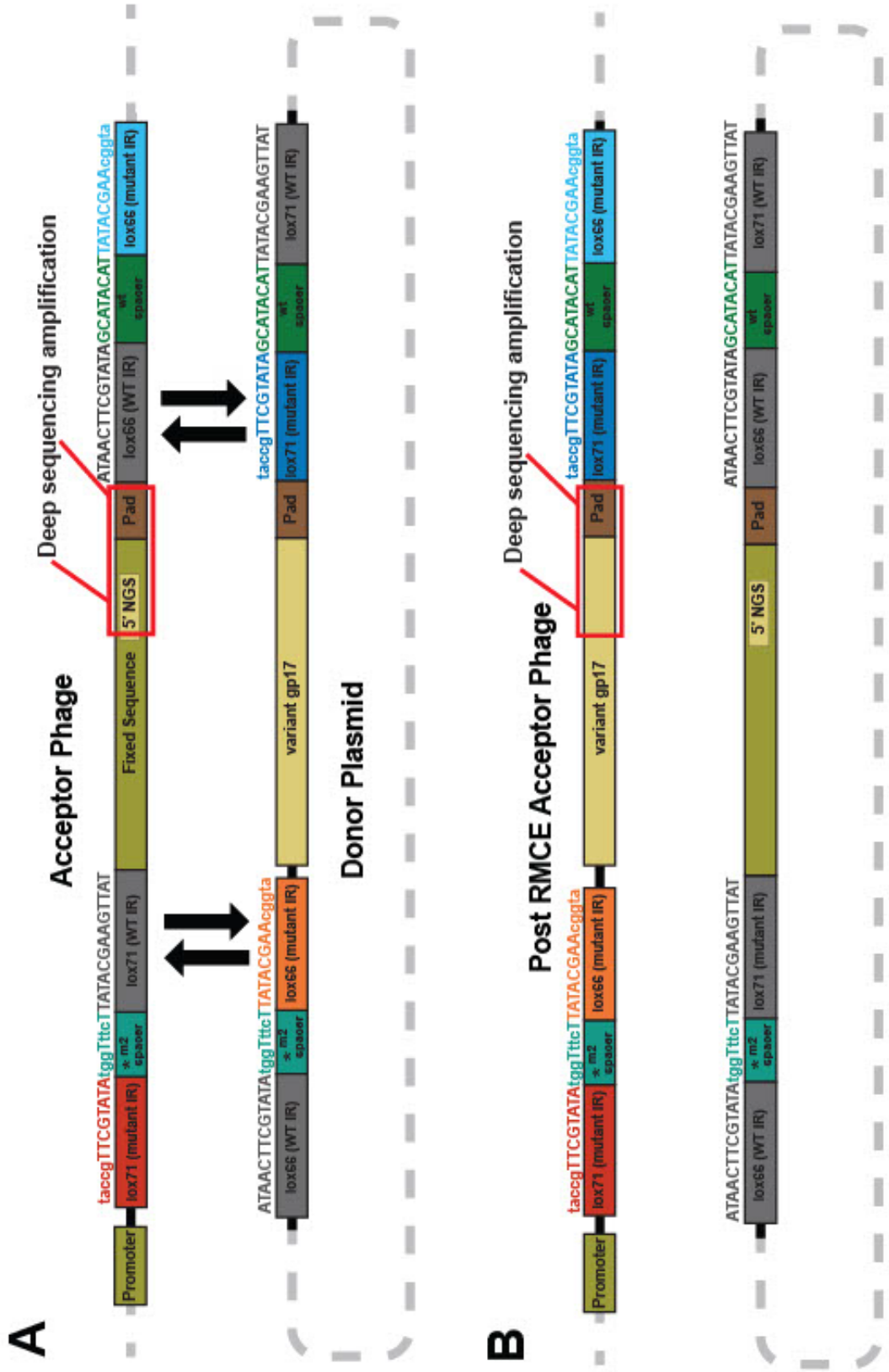

A
Acceptor Phage
Deep sequencing amplification
taccgTTCGTATAtggTttcTTATACGAAGTTAT
ATAACTTCGTATAGCATACATTATACGAAcggta
Promoter
lox71 (mutant IR)
* m2 spacer
lox71 (WT IR)
Fixed Sequence
5' NGS
Pad
lox66 (WT IR)
wt spacer
lox66 (mutant IR)
Donor Plasmid
ATAACTTCGTATAtggTttcTTATACGAAcggta
taccgTTCGTATAGCATACATTATACGAAGTTAT
lox66 (WT IR)
* m2 spacer
lox66 (mutant IR)
variant gp17
Pad
lox71 (mutant IR)
wt spacer
lox71 (WT IR)
B
Post RMCE Acceptor Phage
Deep sequencing amplification
taccgTTCGTATAtggTttcTTATACGAAcggta
ATAACTTCGTATAGCATACATTATACGAAcggta
Promoter
lox71 (mutant IR)
* m2 spacer
lox66 (mutant IR)
variant gp17
Pad
lox71 (mutant IR)
wt spacer
lox66 (mutant IR)
Post RMCE Donor Plasmid
ATAACTTCGTATAtggTttcTTATACGAAGTTAT
ATAACTTCGTATAGCATACATTATACGAAGTTAT
lox66 (WT IR)
* m2 spacer
lox71 (mutant IR)
5' NGS
Pad
lox66 (WT IR)
wt spacer
lox71 (WT IR)

Figure 1 – Figure Supplement 1. Sequence rearrangements before and after recombinase-mediated cassette exchange. Schematic illustration of sequence rearrangements in acceptor phage and donor plasmid (A) before and (B) after recombinase-mediated cassette exchange (RMCE). Specific lox recombinase sites required for exchanging sequence cassettes (variant and fixed sequence) are shown. Lox sites have wild type (WT) or mutated inverted repeats (IR) and one-way RMCE can only occur if one IR is wild type, while the m2 spacer forces recombination in the correct orientation and prevents adverse recombination events [53]. Deep sequencing targets the area boxed in red between the 5’ NGS region and 3’ pad on both acceptor phages and the variant library.

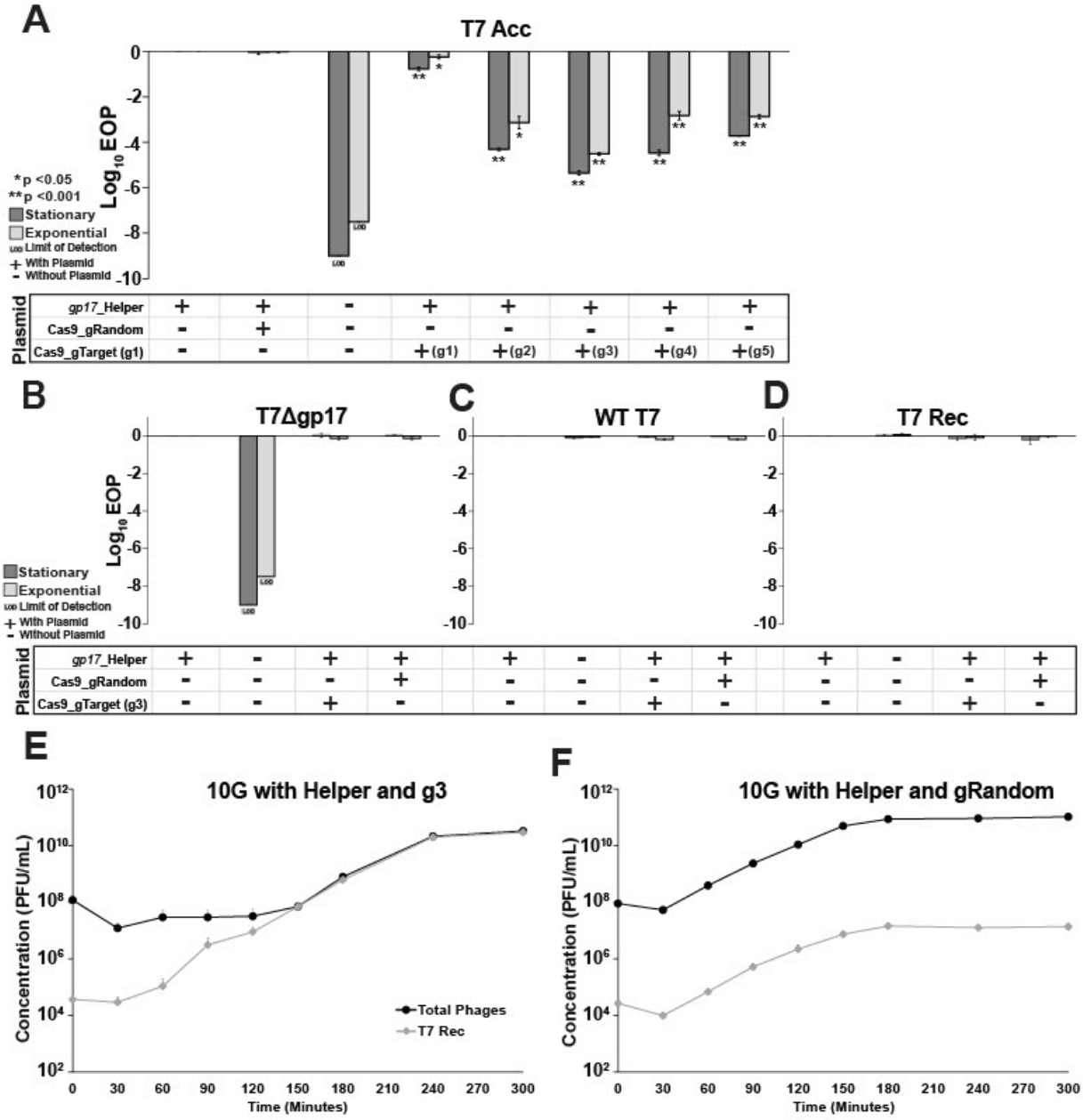

A
T7 Acc
Log10 EOP
*p <0.05
**p <0.001
Stationary
Exponential
LOD Limit of Detection
+ With Plasmid
- Without Plasmid
Plasmid
gp17_Helper
Cas9_gRandom
Cas9_gTarget (g1)
+(g1)
+(g2)
+(g3)
+(g4)
+(g5)
B
T7Δgp17
C
WT T7
D
T7 Rec
Cas9_gTarget (g3)
E
10G with Helper and g3
Concentration (PFU/mL)
Time (Minutes)
Total Phages
T7 Rec
F
10G with Helper and gRandom

Figure 1 – Figure Supplement 2. Effect of Cas9-gRNA system on acceptor and control phages.

(A) Efficiency of plating (EOP) measurements on *E. coli* 10G in stationary (dark grey bar) and exponential (light gray bar) phases for T7 acceptor phage (T7 Acc) with combinations of gp17 helper plasmid, Cas9-gRNA plasmid with a random guide (gRandom) or targeting guides 1 through 5 (gTarget, g1-g5). T7 acceptor phages cannot plaque without helper plasmid. Their plaquing is unaffected by Cas9-gRNA with a random guide (gRandom). Among five different gRNAs targeting fixed sequence, g3 shows highest targeting efficiency and was used for phage library construction with ORACLE. **(B-D)** The Cas9-gRNA system does not adversely affect the plaquing activity of untargeted phages. Efficiency of plating (EOP) measurements on E. coli 10G with combinations of gp17 helper plasmid, Cas9-gRNA plasmid with a random guide or targeting guide 3 (gTarget, g3) for **(B)** T7 phage without gp17 (T7Δgp17), **(C)** wildtype T7 (WT T7) and **(D)** acceptor T7 phage recombined with wildtype gp17 (T7 Rec). **(E and F)** Comparison of accumulation of recombined phages (T7 Rec) with respect to total phages using **(E)** 10G with gp17 helper plasmid and Cas9-gRNA (g3) and **(F)** 10G with gp17 helper plasmid and Cas9-gRNA (gRandom). All data shown is biological triplicates (mean + SD), all EOP data uses 10G with gp17 helper as a reference host.

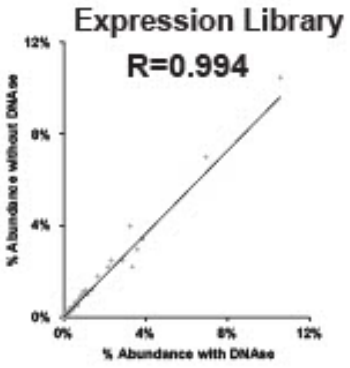


Figure 1 – Figure Supplement 3. Correlation between expression library with and without DNAse treatment.

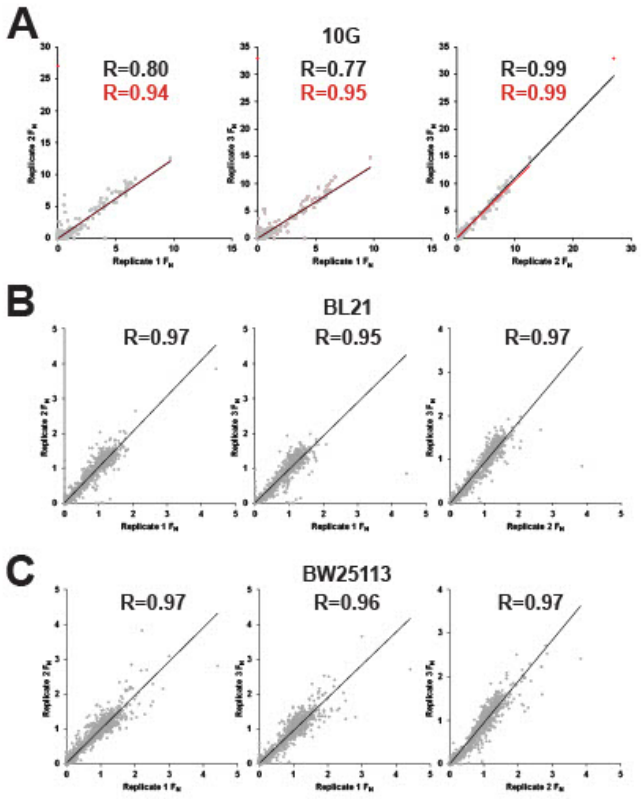


Figure 2 – Figure Supplement 1. Correlation between biological replicates for selection of phage variant library under different conditions. Correlation of FN scores between biological replicates of phage variant library on multiple hosts including **(A)** *E. coli* 10G **(B)** BL21 and **(C)** BW25113. R values and trendlines are displayed for all variants (black line) and with outliers excluded (red line) for 10G with outlier points in red and all other points in grey.

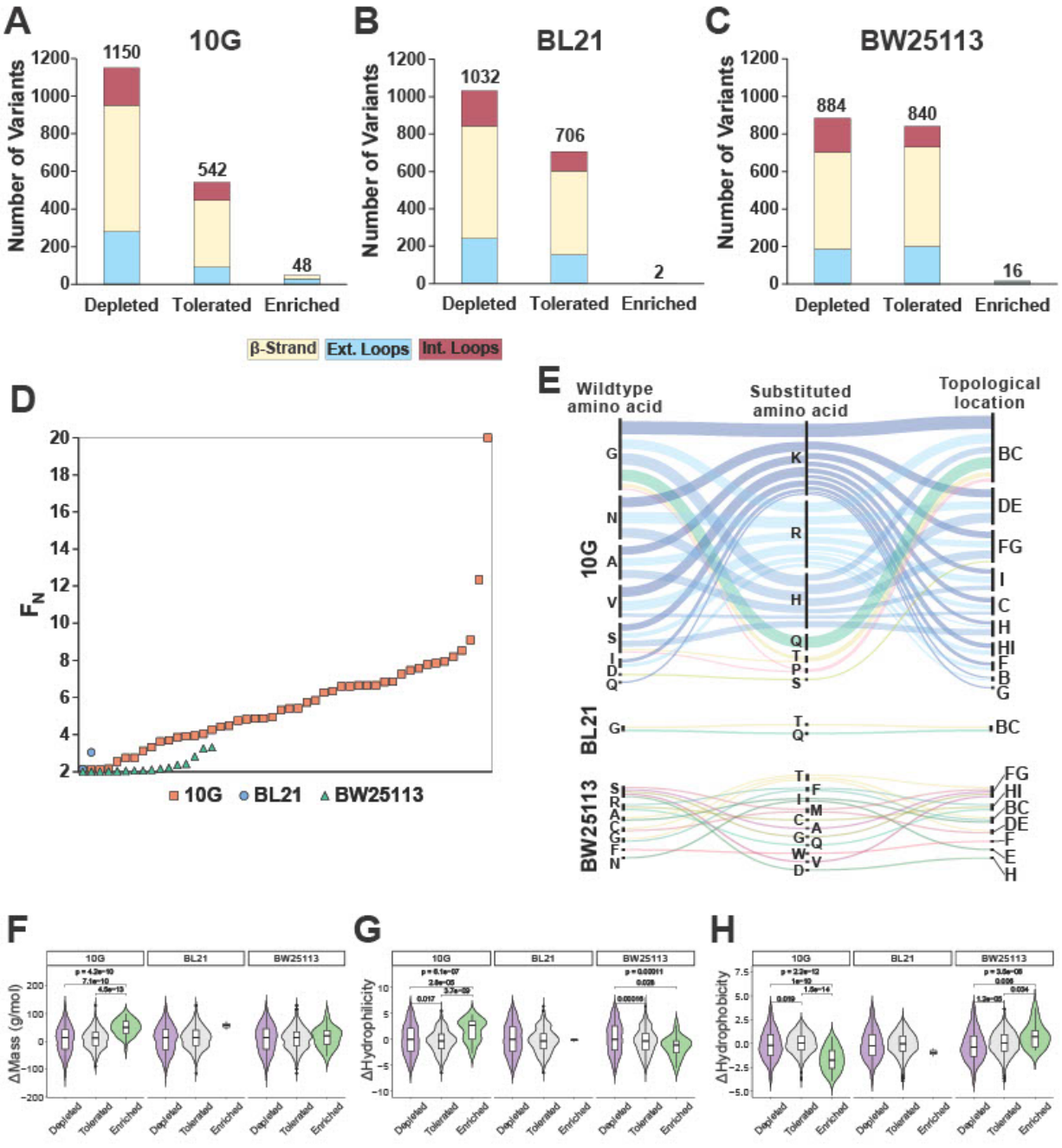

A
10G
Number of Variants
1150
542
48
Depleted
Tolerated
Enriched
B
BL21
1032
706
2
C
BW25113
884
840
16
β-Strand
Ext. Loops
Int. Loops
D
F_N
10G
BL21
BW25113
E
Wildtype amino acid
Substituted amino acid
Topological location
10G
BL21
BW25113
F
ΔMass (g/mol)
G
ΔHydrophilicity
H
ΔHydrophobicity

Figure 2 - Figure Supplement 2. Distribution, enrichment profile and physicochemical properties of variants after selection on 10G, BL21 and BW25113.

Number of variants that were depleted (FN ≤ 0.1) tolerated (FN >0.1 and <2) or enriched (FN ≥ 2) after selection on **(A)** E. coli 10G, **(B)**, BL21, and **(C)** BW25113, separated by topology of the tip domain color coded as interior loops (red), β-sheets (beige) and exterior loops (blue). **(D)** Average FN of enriched variants (FN ≥ 2) for 10G (orange squares), BL21 (blue circles), and BW25113 (teal triangles) ordered left to right from lowest to highest FN. **(E)** Alluvial distribution of enriched variants (FN ≥ 2) on 10G (upper), BL21 (middle) and BW25113 (bottom), showing wild type amino acids (left), their substitution (middle) and topological location on the structure (right). Each alluvial is colored based on the substituted amino acid and scaled by FN across hosts. Violin plots comparing **(F)** change in mass, **(G)** change in hydrophilicity, and **(H)** change in hydrophobicity for grouped depleted (FN ≤ 0.1) tolerated (FN >0.1 and <2) or enriched (FN ≥ 2) substitutions on *E. coli* 10G, BL21 and BW25113. p-values are shown if only if <0.05, the upper p-value is the result of a Kruskal-Wallis test among all three groups while pairwise p-values from a Wilcoxon test are shown linking each group.

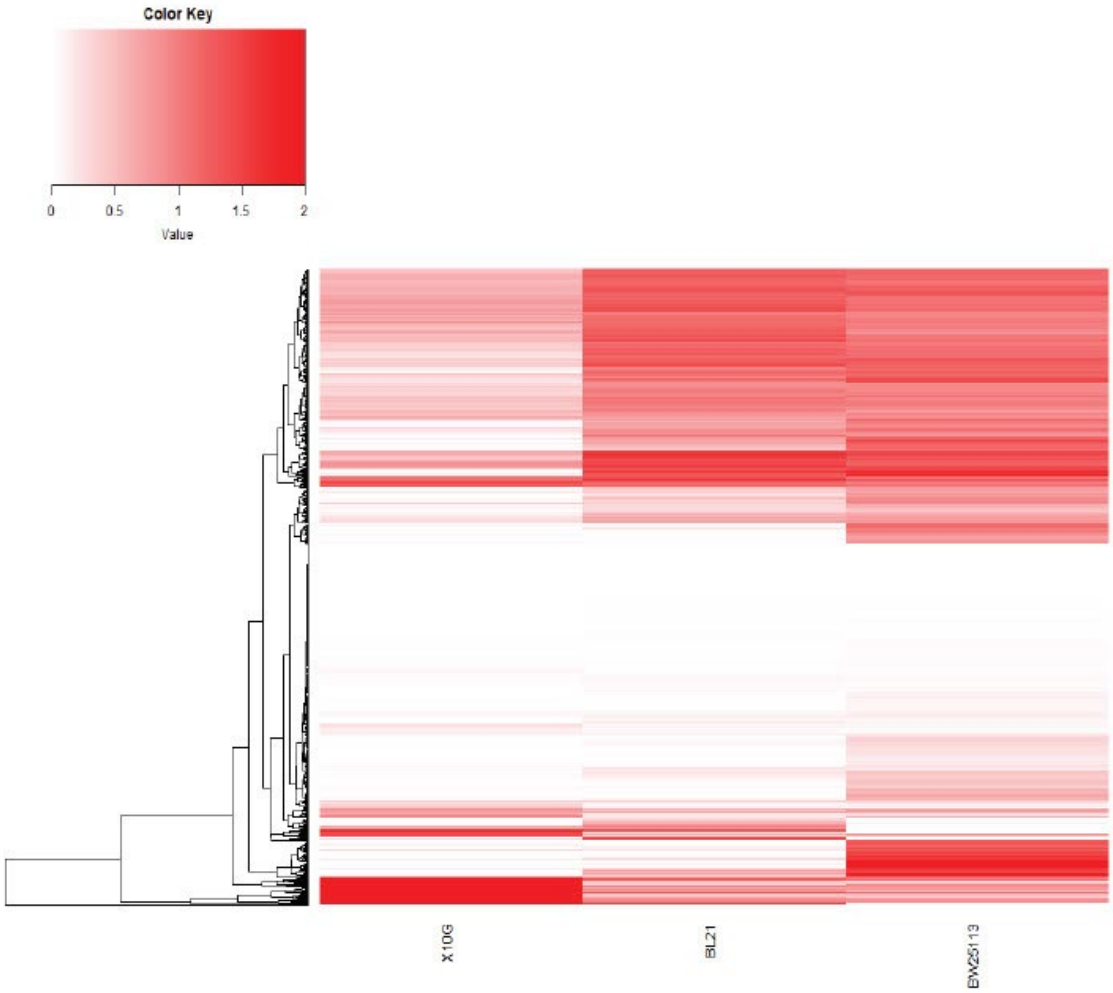


Figure 2 - Figure Supplement 3. Hierarchical clustering of phage variant based on FN score

Hierarchical clustering of phage variants based on their FN score across *E. coli* 10G, BL21 and BW25113. Variants are shaded in white (FN = 0) to red (FN = 2+). Only variants with an FN above the limit of detection on at least one host are included.

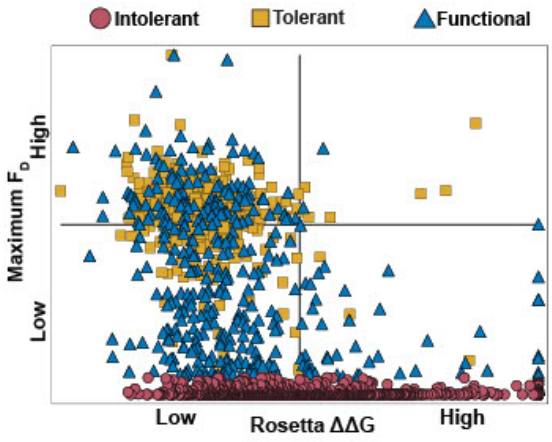


Figure 3 - Figure Supplement 1. Comparing FD to computationally predicted stability of variant ΔΔG.

Comparison of maximum FD values between *E. coli* 10G, BL21 and BW25113 to computationally predicted change in stability (ΔΔG, see methods) for each intolerant (red circles), tolerant (yellow squares) and functional (blue triangles) in the variant library.

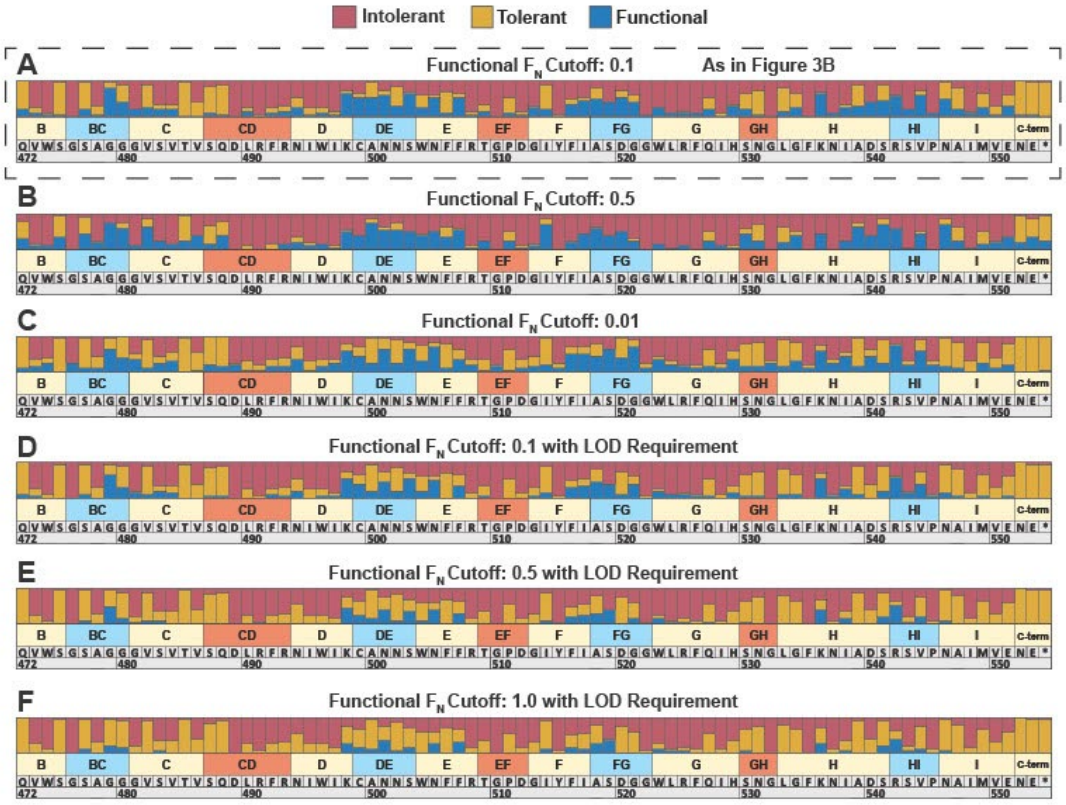


Figure 3 - Figure Supplement 2. Classification of tolerant, intolerant and functional regions based on different cutoff conditions.

Phenotypic classification of each position is determined by aggregating scores of all substitutions in all hosts at that position. **(A)** Substitutions are classified using an FN cutoff of 0.1 as intolerant (FN < 0.1 in all hosts), tolerant (FN ≥ 0.1 in all hosts), or functional (FN < 0.1 in one host, FN ≥ 0.1 in another host) as seen in Figure 3B. Variations of this cutoff **(B)** using an FN cutoff of 0.5 or **(C)** using an FN cutoff of 0.01. An alternative, more conservative approach to classifying a functional substitution requires an FN below the limit of detection (LOD) in one host and a higher FN of **(D)** 0.1, **(E)** 0.5, or **(F)** 1.0 in another host. All cutoff conditions broadly result in similar overall trends.

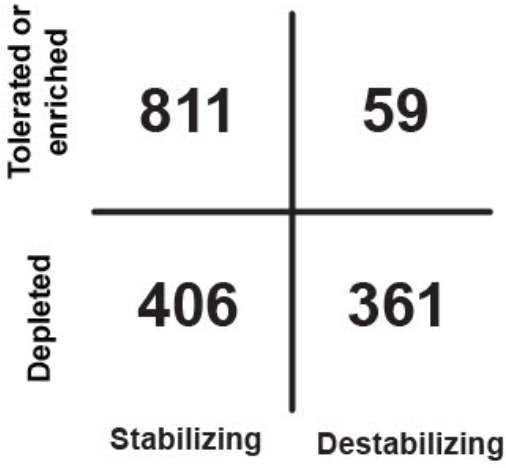


Figure 3 - Figure Supplement 3. Truth table comparing functional results to predicted stability.

Comparison of functional results to computationally predicted change in stability (ΔΔG, see methods). Substitutions are binned as tolerated or enriched if maximum FD > 0.1 and depleted if FD < 0.1 across any strain. Substitutions are considered stabilizing if predicted ΔΔG was <10 and destabilizing if >10.

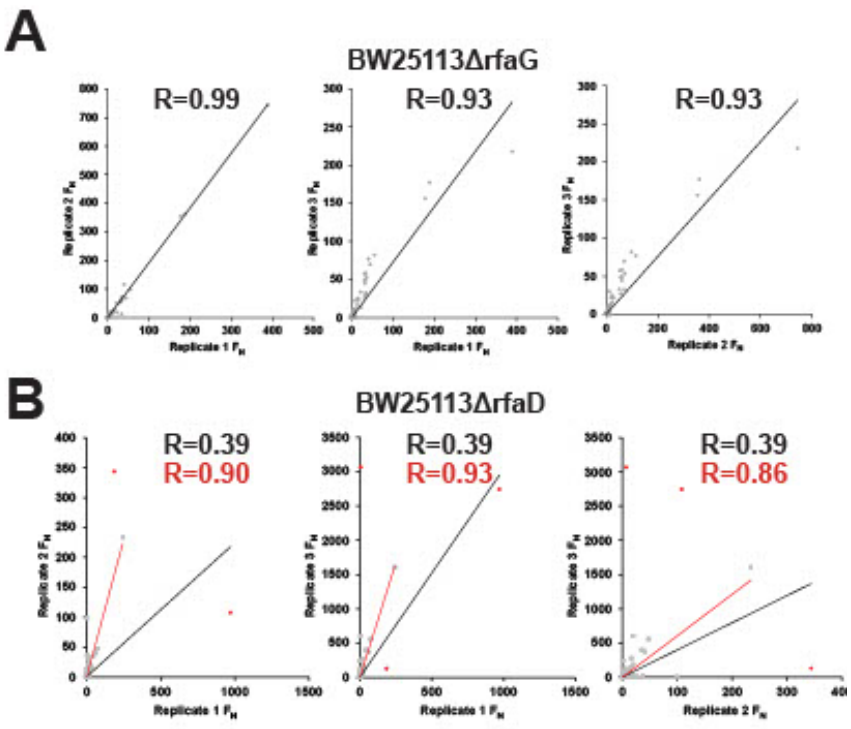


Figure 4 – Figure Supplement 1. Correlation between biological replicates for selection of phage variant library under different conditions. Correlation of FN scores between biological replicates of phage variant library on **(A)** *E. coli* BW25113Δ*rfaG* and **(B)** BW25113Δ*rfaD*. R values and trendlines are displayed for all variants (black line) and with outliers excluded (red line) for BW25113Δ*rfaD* with outlier points in red and all other points in grey.

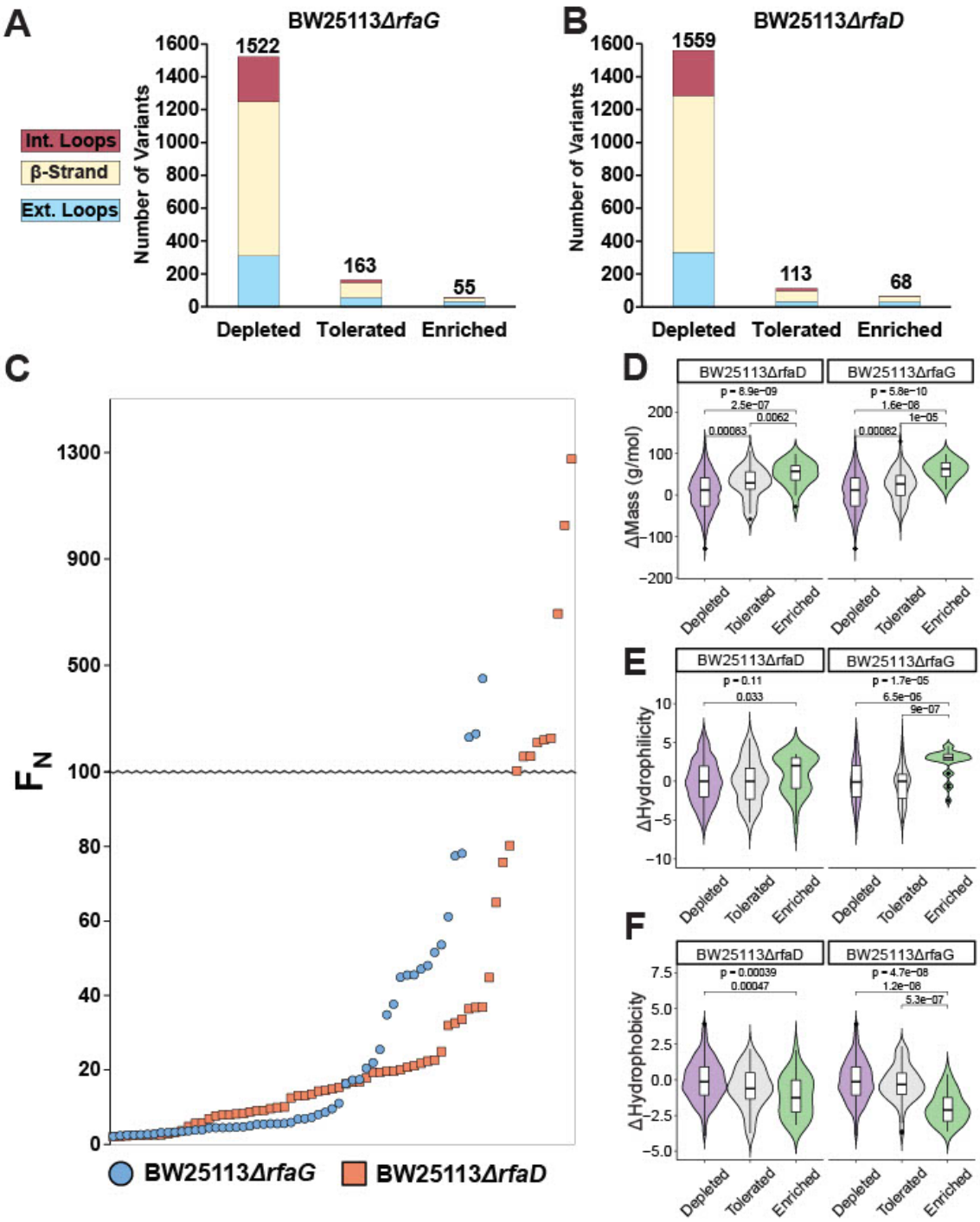

A
BW25113ΔrfaG
Int. Loops
β-Strand
Ext. Loops
Number of Variants
1600
1400
1200
1000
800
600
400
200
0
1522
163
55
Depleted
Tolerated
Enriched
B
BW25113ΔrfaD
1559
113
68
C
F_N
1300
900
500
100
80
60
40
20
0
BW25113ΔrfaG
BW25113ΔrfaD
D
BW25113ΔrfaD
BW25113ΔrfaG
p = 8.9e−09
2.5e−07
0.0062
0.00063
p = 5.8e−10
1.6e−08
1e−05
0.00062
ΔMass (g/mol)
200
100
0
−100
−200
E
p = 0.11
0.033
p = 1.7e−05
6.5e−06
9e−07
ΔHydrophilicity
10
5
0
−5
−10
F
p = 0.00039
0.00047
p = 4.7e−08
1.2e−08
5.3e−07
ΔHydrophobicity
7.5
5.0
2.5
0.0
−2.5
−5.0

Figure 4 – Figure Supplement 2. Distribution, enrichment profile and physicochemical properties of variants after selection on BW25113Δ*rfaG* and BW25113Δ*rfaD*.

Number of variants that were depleted (FN ≤ 0.1) tolerated (FN >0.1 and <2) or enriched (FN ≥ 2) after selection on **(A)** BW25113Δ*rfaG* or **(B)** BW25113Δ*rfaD*, separated by topology of the tip domain color coded as interior loops (red), β-sheets (beige) and exterior loops (blue). **(C)** Average FN of enriched variants (FN ≥ 2) for BW25113Δ*rfaD* (orange squares) and BW25113Δ*rfaG* (blue circles) ordered left to right from lowest to highest FN. Violin plots comparing **(D)** change in mass, **(E)** change in hydrophilicity, and **(F)** change in hydrophobicity for grouped depleted (FN ≤ 0.1) generally tolerated (FN >0.1 and <10) or well enriched (FN ≥ 10) substitutions on *E. coli* BW25113Δ*rfaD* and BW25113Δ*rfaG*. The upper p-value is the result of a Kruskal-Wallis test among all three groups while pairwise p-values from a Wilcoxon test are shown linking each group; p-values are shown if only if <0.05.

A

BW25113ΔrfaG Ranking

Variant	Replicate 1 Rank	Replicate 2 Rank	Replicate 3 Rank
V544R	6	4	5
S541K	5	8	6
S541H			8
N501R	8	7	
N501H	4	5	4
G521R	3	3	3
G521K	2	2	2
G521H	1	1	1
A539K	10	9	7
A518R	7	6	9
A500R	9	10	10

B

BW25113ΔrfaD Ranking

Variant	Replicate 1 Rank	Replicate 2 Rank	Replicate 3 Rank
S541R	5	6	7
S541K	2	2	3
S541H		7	
S541E	4	5	5
R525*	8		10
N501K			4
N501H		4	
G521R	1	3	2
G521K	6	8	6
G521H			1
G521C	9		
G480R	10		8
G479Q		9	
D540S			9
A539K	7		
A500R	3	1	

Figure 4 – Figure Supplement 3. Ranking FN of the ten most enriched variants in each biological replicate for **(A)** *E. coli* BW25113Δ*rfaG* and **(B)** BW25113Δ*rfaD.*

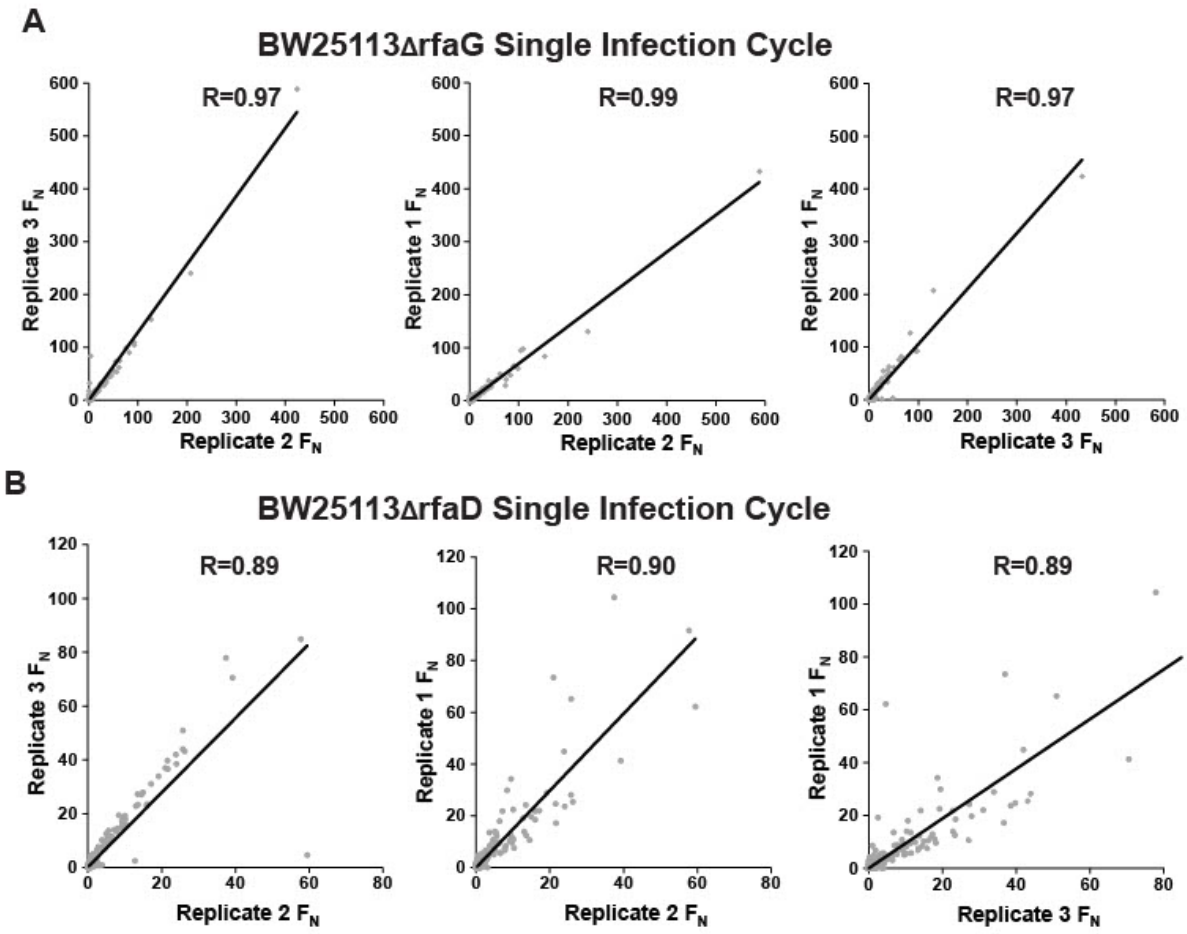


Figure 4 – Figure Supplement 4. Correlation between biological replicates for selection of phage variant library after one infection cycle on **(A)** *E. coli* BW25113Δ*rfaG* and **(B)** BW25113Δ*rfaD*.

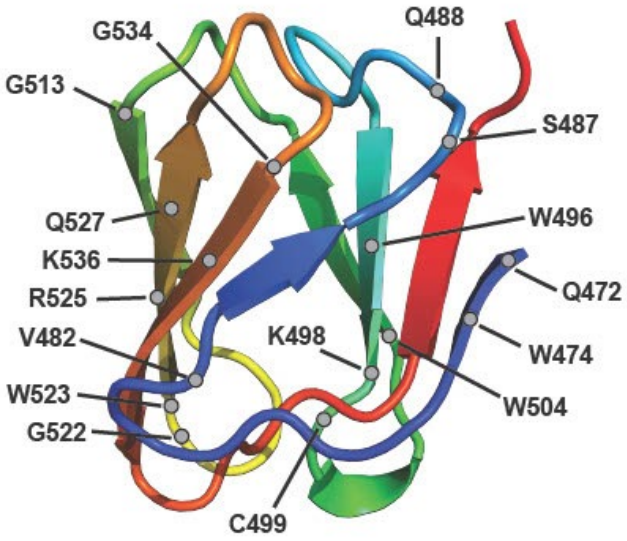


Figure 4 - Figure Supplement 5. Location of truncations in the tip domain enriched after selection

Location (grey circles) of truncations in the tip domain with an FN>10 on BW25113Δ*rfaD*. The tip domain is colored in a rainbow gradient from N (blue) to C (red) terminus.

Methods

Key Resources Table				
Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
strain, strain background (*Escherichia coli*)	*E. coli* 10G	Lucigen	Lucigen:60107-1	
strain, strain background (*Escherichia coli*)	*E. coli* BL21	ATCC	ATCC:BAA-1025	
strain, strain background (*Escherichia coli*)	*E. coli* 10-beta	NEB	NEB:C3020	
strain, strain background (*Escherichia coli*)	*E. coli* BW25113	(Baba et al., 2006)	BW25113	
strain, strain background (*Escherichia coli*)	*E. coli* BW25113Δ*rfaG*	(Baba et al., 2006)	BW25113Δ*rfaG*	
strain, strain background (*Escherichia coli*)	*E. coli* BW25113 Δ*rfaD*	(Baba et al., 2006)	BW25113 Δ*rfaD*	
strain, strain background (*Escherichia coli*)	*E. coli* UTI473	(Arthur et al., 1990)	UTI473	

strain, strain background (T7 Bacteriophage)	T7 Bacteriophage	ATCC	ATCC:BAA-1025-B2	
strain, strain background (T7 Bacteriophage)	T7 Bacteriophage variants	This paper	Available on request	DMS variants, available from the Raman lab
commercial assay or kit	KAPA HiFi PCR Kit	Roche	Roche:KK2101	
commercial assay or kit	KAPA2G Robust PCR Kit with dNTPS	Roche	Roche:KK5005	
commercial assay or kit	Golden Gate Assembly Kit (BsaI-HFv2)	NEB	NEB:E1601L	
recombinant DNA reagent	pHT7Helper1 (plasmid)	This paper		Helper with T7 gp17. See methods for full details.
recombinant DNA reagent	pHRec1 and derivatives (plasmids)	This paper		Recombination plasmid. See methods for details
recombinant DNA reagent	pHCas9 and derivatives (plasmids)	This paper		Plasmid with Cas9 targeting acceptor phage. See methods for full details.
software, algorithm	R scripts for DMS Analysis	This paper	N/A	Available here https://github.com/raman-lab/oracle
software, algorithm	R scripts for Physicochemical Comparisons	This paper	N/A	Available here https://github.com/raman-lab/oracle

software, algorithm	R scripts for Rosetta ΔΔG calculations	This paper	N/A	Available here https://github.com/raman-lab/oracle

Microbes and Culture Conditions

T7 bacteriophage was obtained from ATCC (ATCC® BAA-1025-B2). *Saccharomyces cerevisiae* BY4741, *Escherichia coli* BL21 is a lab stock, *E. coli* 10G is a highly competent DH10B derivative (Durfee et al., 2008) originally obtained from Lucigen (60107-1). *E. coli* 10-beta was purchased from NEB (C3020). *E. coli* BW25113, BW25113Δ*rfaD* and BW25113Δ*rfaG* were obtained from Doug Weibel (University of Wisconsin, Madison) and are derived from the Keio collection (Baba et al., 2006). UTI473 was obtained from Rod Welch (University of Wisconsin, Madison) and originates from a UTI collection (Arthur et al., 1990).

All bacterial hosts are grown in and plated on Lb media (1% Tryptone, 0.5% Yeast Extract, 1% NaCl in dH2O, plates additionally contain 1.5% agar, while top agar contains only 0.5% agar) and Lb media was used for all experimentation. Kanamycin (50 µg/ml final concentration, marker for pHT7Helper1) and spectinomycin (115 µg/ml final concentration, marker for pHRec1, pHRec1-Lib and pHCas9 and derivatives) was added as needed. All incubations of bacterial cultures were performed at 37°C, with liquid cultures shaking at 200-250 rpm unless otherwise specified. Bacterial hosts were streaked on appropriate Lb plates and stored at 4°C.

S. cerevisiae BY4741 was grown on YPD (2% peptone, 1% yeast extract, 2% glucose in dH2O, plates additionally contain 2.4% agar), after Yeast Artificial

Chromosomes (YAC) transformation *S. cerevisiae* BY4741 was grown on SD-Leu (0.17% yeast nitrogen base, 0.5% ammonium sulfate, 0.162% amino acids – Leucine [Sigma Y1376], 2% glucose in dH20, plates additionally contain 2% agar). All incubations of *S. cerevisiae* were performed at 30°C, with liquid cultures shaking at 200-250 rpm. *S. cerevisiae* BY4741 was streaked on YPD or SD-Leu plates as appropriate and stored at 4°C.

T7 bacteriophage was propagated using *E. coli* BL21 after initial receipt from ATCC and then as described on various hosts in methods. All phage experiments were performing using Lb and culture conditions as described for bacterial hosts. Phages were stored in Lb at 4°C.

For long term storage all microbes were stored as liquid samples at -80°C in 10% glycerol, 90% relevant media.

SOC (2% tryptone, 0.5% yeast extract, 0.2% 5M NaCl, 0.25% 1M KCL, 1% 1M $MgCl_2$, 1% 1M $MgSO_4$, 2% 1M glucose in dH2O) was used to recover host and phages after transformation.

General Cloning Methods

PCR was performed using KAPA HiFi (Roche KK2101) for all experiments with the exception of multiplex PCR for screening Yeast Artificial Chromosomes (YACs), which was performed using KAPA2G Robust PCR kits (Roche KK5005). Golden Gate assembly was performed using New England Biosciences (NEB) Golden Gate Assembly Kit (BsaI-HFv2, E1601L). Restriction enzymes were purchased from NEB with the exception of

DNAse I (Roche 4716728001). DNA purification was performed using EZNA Cycle Pure Kits (Omega Bio-tek D6492-01) using the centrifugation protocol. YAC extraction was performed using YeaStar Genomic DNA Extraction kits (Zymo Research D2002). Gibson assembly was performed according to the Gibson Assembly Protocol (NEB E5510) but Gibson Assembly Master Mix was made in lab (final concentration 100 mM Tris-HCL pH 7.5, 20 mM $MgCl_2$, 0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 10 mM dTT, 5% PEG-8000, 1 mM NAD^+, 4 U/ml T5 exonuclease, 4 U/µl Taq DNA Ligase, 25 U/mL Phusion polymerase). All cloning was performed according to manufacturer documentation except where noted in methods. If instructions were variable and/or specific conditions are relevant for reproducing results, those conditions are also noted in the relevant methods section.

PCR reactions use 1 µl of ~1 ng/µl plasmid or ~0.1 ng/µl DNA fragment as template for relevant reactions. PCR reactions using phage as template use 1 µl of undiluted or 1:10 diluted phage stock, genomic extraction was unnecessary. Phage template was initially treated at 65°C for 10 minutes (for YAC cloning), but we later simply extended the 95°C denaturation step to 5 minutes (for deep sequencing).

DpnI digest was performed on all PCR that used plasmid as template. Digestion was performed directly on PCR product immediately before purification by combining 1-2 µl DpnI (20-40 units), 5 µl 10x CutSmart Buffer, PCR product, and dH2O to 50 ul, incubating at 37°C for 2 hours then heat inactivating at 80C for 20 minutes.

DNAse treatment of phages was performed by adding 5 µl undiluted phages, 2 µl 10x DNAse I buffer, 1 µl of 2 U/µl DNAse I, dH2O to 20 ul, then incubating for 20 minutes

at 37°C, followed by heat inactivation at 75C for 10 minutes. 1 µl of this reaction was used as template for relevant PCR.

Electroporation of plasmids and YACs was performed using a Bio-rad MicroPulser (165-2100), Ec2 setting (2 mm cuvette, 2.5 kV, 1 pulse) using 25-50 µl competent cells and 1-2 µl DNA for transformation. Electroporated cells were immediately recovered with 950 µl SOC, then incubated at 37°C for 1 to 1.5 hours and plated or grown in relevant media.

E. coli 10G competent cells were made by adding 8 mL overnight 10G cells to 192 mL SOC (with antibiotics as necessary) and incubating at 21°C and 200 rpm until ~OD_{600} of 0.4 as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation (actual incubation time varies based on antibiotic, typically overnight). Cells are centrifuged at 4°C, 800g-1000g for 20 minutes, the supernatant is discarded, and cells are resuspended in 50 mL 10% glycerol. Centrifugation and washing are repeated three times, then cells are resuspended in a final volume of ~1 mL 10% glycerol and are aliquoted and stored at -80°C. Cells are competent for plasmid and YACs.

Site Directed Mutagenesis (SDM) was performed, in brief, using complementary primers with the desired mutation in the middle of the primer, using 16x cycles of PCR, followed by DpnI digestion and electroporation into competent *E. coli* 10G. Splicing by Overlap Extension (SOE, also known as PCR overlap extension) was performed, in brief, using equimolar ratios of fragments, 16x cycles of PCR using extension based on the combined length of fragments, then a second PCR reaction using 1/100 or 1/1000 diluted product of the first reaction in a typical PCR reaction using 5’ and 3’ end primers for each fragment.

Detailed protocols for cloning are available on request. All primers used in experiments in this publication are listed in supplementary file 1.

Plasmid Cloning and Descriptions

pHT7Helper1 contains a pBR backbone, kanamycin resistance cassette, mCherry, and the T7 tail fiber *gp17*. Both mCherry and *gp17* are under constitutive expression. *Gp17* was combined with promoter apFAB47 (Kosuri et al., 2013) using SOE and the plasmid assembled by Gibson assembly. There is a single nucleotide deletion in the promoter that has no effect on plaque recovery for phages that require *gp17* to plaque. This plasmid is used during optimized recombination and during accumulation in ORACLE to prevent library bias and depletion of variants that grow poorly on *E. coli* 10G.

pHRec1 contains an SC101 backbone, Cre recombinase, a spectinomycin resistance cassette, and the T7 tail fiber *gp17* flanked by Cre lox66 sites with an m2 spacer, a 3' pad region and lox71 sites with a wt spacer (Langer et al., 2002) (Figure 1 – Figure Supplement 1). Cre recombinase is under constitutive expression. This plasmid was assembled with sequential PCR and Gibson assembly. During assembly we used PCR overhangs and SDM to create two synonymous substitutions in *gp17* to remove two BsaI restriction sites, facilitating downstream golden gate assembly. This plasmid was used in recombination assays, as it allows for recombination of wildtype *gp17*, and is used as template to generate the DMS variant library. The DMS variant library is referred to as pHRec1-Lib and is used during optimized recombination in ORACLE. Note this assembly was not tolerated in higher copy number plasmids.

pHCas9 contains an SC101 backbone, a spectinomycin resistance cassette and cas9 cassette capable of ready BsaI cloning of gRNA (Jiang et al., 2013). This plasmid is used directly as part of the negative control for the accumulation assay, and has five derivatives, pHCas9-1 through -5, each with a different gRNA targeting the fixed region in the T7 acceptor phage. pHCas9 was created with Gibson assembly, while derivatives were assembled by phosphorylation and annealing gRNA oligos (100 uM forward and reverse oligo, 5 µl T4 Ligase buffer, 1 µl T4 PNK, to 50 µl dH2O, incubate at 37°C for 1 hour, 96C for 6 minutes then 0.1C/s temperature reduction to 23C), then Golden Gate cloning (1 µl annealed oligo, 75 ng pHCas9, 2 µl T4 DNA Ligase Buffer, 1 µl Golden Gate Enzyme Mix, dH20 to 20 ul. Incubation at 37°C for 1 hour then 60C for 5 minutes, followed by direct transformation of 1 ul, plated on Lb with spectinomycin). Note pHCas9-3 was the most inhibitory (Figure 1 - Figure Supplement 2A) and was the only plasmid used in accumulation during ORACLE. This assembly was also not tolerated in higher copy number plasmids. All plasmid backbones and gene fragments are lab stocks.

General Bacteria and Phage Methods

Bacterial concentrations were determined by serial dilution of bacterial culture (1:10 or 1:100 dilutions made to 1 mL in 1.5 microcentrifuge tubes in Lb) and subsequent plating and bead spreading of 100 µl of a countable dilution (targeting 50 colony forming units) on Lb plates. Plates were incubated overnight and counted the next morning. Typically, two to three dilution series were performed for each host to initially establish concentration at different OD_{600} and subsequent concentrations were confirmed with a single dilution series for later experiments.

Stationary phase cultures are created by growing bacteria overnight (totaling ~20-30 hours of incubation) at 37°C. Cultures are briefly vortexed then used directly. Exponential phase culture consists of stationary culture diluted 1:20 in Lb then incubated at 37°C until an OD_{600} of ~0.4-0.8 is reached (as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation), typically taking 40 minutes to 1 hour and 20 minutes depending on the strain and antibiotic, after which cultures are briefly vortexed and used directly.

Phage lysate was purified by centrifuging phage lysate at 16g, then filtering supernatant through a 0.22 uM filter. Chloroform was not routinely used unless destruction of any remaining host was considered necessary and is mentioned in such cases.

To establish titer, phage samples were typically serially diluted (1:10 or 1:100 dilutions made to 1 mL in 1.5 microcentrifuge tubes) in Lb to a 10^{-8} dilution for preliminary titering by spot assay. Spot assays were performed by mixing 250 µl of relevant bacterial host in stationary phase with 3.5 mL of 0.5% top agar, briefly vortexing, then plating on Lb plates warmed to 37°C. After plates solidified (typically ~5 minutes), 1.5 µl of each dilution of phage sample was spotted in series on the plate. Plates were incubated and checked at 2-4 hours and in some cases overnight (~20-30 hours) to establish a preliminary titer. After a preliminary titer was established, phage samples were serially diluted in triplicate for efficiency of plating (EOP) assays. EOP assays were performed using whole plates instead of spot plates to avoid inaccurate interpretation of results due to spotting error (Khan Mirzaei and Nilsson, 2015). To perform the whole plate EOP assay, 250 µl of bacterial host in stationary or exponential phase was mixed with between 5 to

50 µl of phages from a relevant dilution targeted to obtain 50 plaque forming units (PFUs) after overnight incubation. The phage and host mixture was briefly vortexed, briefly centrifuged, then added to 3.5 mL of 0.5% top agar, which was again briefly vortexed and immediately plated on Lb plates warmed to 37°C. After plates solidified (typically ~5 minutes), plates were inverted and incubated overnight. PFUs were typically counted at 4-6 hours and after overnight incubation (~20-30 hours) and the total overnight PFU count used to establish titer of the phage sample. PFU totals between 10 and 300 PFU were typically considered acceptable, otherwise plating was repeated for the same dilution series. This was repeated in triplicate for each phage sample on each relevant host to establish phage titer.

EOP was determined using a reference host, typically *E. coli* 10G with pHT7Helper1 but stated if otherwise. EOP values were generated for each of the three dilutions by taking the phage titer on the test host divided by the phage titer on the reference host, and this value was subsequently $\log_{10}$ transformed. Values are reported as mean ± SD.

MOI was calculated by dividing phage titer by bacterial concentration. MOI for the T7 variant library after the variant gene is expressed was estimated by titering on 10G with pHT7Helper1.

Limit of Detection (LOD) for T7 acceptor phages (T7 Acc) and T7 lacking a tail fiber (T7Δ*gp17*) was established based on the ability of these phages to clear a bacterial lawn. These phages are unable to plaque on host lacking pHT7Helper1, but phages do express a tail fiber due to being propagated on host with pHT7Helper1. Functionally this allows these phages to complete one infection cycle and kill one host but does not allow the

creation of plaque-viable progeny phages if that host does not also contain pHT7Helper1. At an MOI of greater than ~2 we noted plates no longer form lawns of bacteria but instead contain individual colonies or are clear, reflective of these singular assassinations. As expected, this effect occurs at different concentrations of phages for exponential or stationary host due to different host concentration at those stages of growth. As plaques cannot form under these conditions, and these infections are not productive beyond a single infection, we simply used this cut off as the limit of detection for this assay.

Growth time courses for UTI473 (Figure 5) and OD_{600} was performed using a Synergy HTX Multi-Mode 96-well plate reader, using 140 µl of host and 10 µl of relevant phage titers. Phages were applied after an hour of incubation in the plate reader.

Recombination Rate and Accumulation Assays

To establish recombination rate (Figure 1C), we passaged T7 acceptor phages on 5 mL exponential phase *E. coli* 10G containing pHT7Helper1 and pHRec1. pHRec1 was used because the recombined *gp17* tail fiber is wildtype, ensuring every recombined phage is plaque-capable (derived from Figure 1 - Figure Supplement 2B results). We sought to evaluate recombination rate after only 1 passage through the host to avoid misinterpretation of results in case recombined phages had different fitness than unrecombined phages. We used an MOI of 10 and allowed passage for 30 minutes, sufficient time for one wildtype phage passage, after which we halted any remaining reactions by adding 200 µl of chloroform and lysing the remaining bacterial host. Phages were then purified to acquire the final phage population. We established the phage

population titer on 10G and 10G with pHT7Helper1. Both acceptor phages and recombined phages are capable of plaquing on 10G with pHT7Helper1 and this phage titer is used to count the total phage population. Only recombined phages are capable of plaquing on 10G and this titer is used to count recombined phages. Recombination rate was established as the fraction titer of recombined phages divided by recombined phages. This was repeated in triplicate and reported as mean ± SD.

> Method Note: It should be noted that this assay does not delineate for when recombination occurs in the host or how frequently recombination occurs in any one host. For example, if recombination were to occur on the original phage genome, all subsequent progeny phages could contain the recombined gene. In contrast, if recombination were to occur on the phage genome while it is being replicated, anywhere from one individual progeny to all progeny could contain the recombined gene.

To validate accumulation of recombined phages over acceptor phages (Figure 1D and Figure 1 - Figure Supplement 2E-F) we first generated a population of recombined phages using the same scheme as outlined for the recombination rate assay. After recombination this phage population contained primarily T7 acceptor phages with a small percentage of recombined phage containing a wildtype *gp17* tail fiber. This phage population was passaged on 10G containing pHT7Helper1 and either pHCas9-3 (targeting the fixed region in the acceptor phage using g3, the most effective guide by EOP, Figure 1 - Figure Supplement 2A) or pHCas9 (randomized control). Phages were incubated with host in 5 mL total at an initial MOI of 1 based on the titer of the whole

phage population. Every 30 minutes until 180 minutes, and thereafter every 60 minutes until 300 minutes, ~250 µl of culture was removed, infection was stopped by adding 100 µl of chloroform, and phage samples were purified to establish the phage population at that timepoint. Titer at each timepoint was determined on both 10G and 10G with pHT7Helper1 with a single dilution series using whole plate plaque assay. Percent accumulation was derived by dividing titer on 10G by titer on 10G with pHT7Helper1. Accumulation on both hosts was repeated in triplicate and reported as mean ± SD.

DMS Plasmid Library Preparation

To create the DMS variant plasmid library, oligos were first designed and ordered from Agilent as a SurePrint Oligonucleotide Library (Product G7220A, OLS 131-150mers). Every oligo contained a single substitution at a single position in the tip domain, overall including all non-synonymous substitutions, a single synonymous substitution, and a stop codon from position 472-554. Note we did mutate the stop codon, which is position 554, which when substituted results in a 3 amino acid extension (-DAR) of *gp17*. We used the most frequently found codon for each amino acid in the *gp17* tail fiber to define the codon for each substitution. Oligos contained BsaI sites at each end to facilitate Golden Gate cloning. To accommodate a shorter oligo length the library was split into three pools covering the whole tip domain. Oligo pools were amplified by PCR using 0.005 pmol total oligo pool as template and 15 total cycles to prevent PCR bias, then pools were purified. pHRec1 was used as template in a PCR reaction to create three backbones for each of the three pools. Backbones were treated with BsaI and Antartic Phosphatase as follows. 5 µl 10x CutSmart, 2 µl BsaI, ~1177 ng backbone, dH2O to 50 µl was mixed and incubated

at 37°C for 2 hours, after which 1 µl additional BsaI, 2 µl Antartic Phosphatase, 5.89 µl 10x Antartic Phosphatase buffer was spiked into reaction. Reaction was incubated for 1 more hour at 37°C, then enzymes were heat inactivated at 65°C for 20 minutes (concentration ~20ng/µl at this point) and used directly (no purification) in Golden Gate Assembly. Golden gate assembly was performed using ~100 ng of relevant pool backbone and a 2x molar ratio for oligos (~10 ng), combined with 2 µl 10x T4 DNA ligase buffer, 1 µl NEB Golden Gate Enzyme mix and dH2O to 20 ul. These reactions were cycled from 37°C to 16°C over 5 minutes, 30x, then held at 60°C for 5 minutes to complete Golden Gate assembly. Membrane drop dialysis was then performed on each library pool for 75 minutes to enhance transformation efficiency. 2 µl of each pool was transformed into 33 µl competent *E. coli* 10-beta (NEB C3020) cells. Drop plates were made at this point (spotting 2.5 µl of dilutions of each library on Lb plates with spectinomycin) and total actual transformed cells were estimated at ~$2x10^5$ CFU/mL. Each 1 mL pool was added to 4 mL Lb with spectinomycin and incubated overnight, then plasmids were purified. Plasmids concentration was determined by nanodrop and pools were then combined at an equimolar ratio to create the final phage variant pool, denoted as pHRec1-Lib. pHRec1-Lib was transformed into *E. coli* 10G with pHT7Helper1. Drop plates were made (spotting 2.5 µl of dilutions of each library onto Lb plates with spectinomycin and kanamycin) and total actual transformed cells were also estimated at ~$2x10^5$ CFU/mL. The 1 mL library was added to 4 mL Lb with spectinomycin and kanamycin and incubated overnight. This host, *E. coli* 10G with pHT7Helper1 and pHRec1-Lib, was the host used for Optimized Recombination during ORACLE.

ORACLE - Engineering T7 Acceptor Phages

Acceptor phages were assembled using YAC rebooting (Ando et al., 2015b; Jaschke et al., 2012), which requires yeast transformation of relevant DNA segments, created as follows. A prs415 yeast centromere plasmid was split into three segments by PCR, separating the centromere and leucine selection marker, which partially limits recircularization and improved assembly efficiency (Kuijpers et al., 2013). Wildtype T7 segments were made by PCR using wildtype T7 as template. At the site of recombination the acceptor phage contains, in order, lox71 sites with an m2 spacer (Langer et al., 2002) to facilitate one way recombinase mediated cassette exchange (RMCE), a fixed sequence that was derived from sfGFP with a nonsense mutation, a short region mimicking *gp17* to allow detection of acceptor phages by deep sequencing (5' NGS in Figure 1 – Figure Supplement 1), a 3' 'pad' to facilitate deep sequencing, and lox66 sites with a wt spacer (see Figure 1 – Figure Supplement 1). This entire region was turned into one DNA segment by serial SOE reactions.

Method Note: reversable or two way RMCE using wildtype loxP sites could feasibly increase recombination efficiency, as one way RMCE is not necessarily required for ORACLE.

Method Note: PCR, including PCR for deep sequencing, behaved inconsistently at lox sites. We theorize this may be because these sites are inverted repeats. Our inclusion of the 3' 'pad' region corrected this problem and facilitated acceptor phage detection by deep sequencing.

DNA parts were combined together (0.1 pmol/segment) and transformed into *S. cerevisiae* BY4741 using the a high efficiency yeast transformation protocol (Gietz and Woods, 2002) using SD-Leu selection. After 2-3 days colonies were picked and directly assayed by multiplex colony PCR to assay assembly. Multiplex PCR interrogated junctions in the YAC construct and was an effective way of distinguishing correctly assembled YACs. Correctly assembled YACs were purified and transformed into *E. coli* 10G cells containing pHT7Helper1, and after recovery 400 µl was used to inoculate 4.6 mL Lb. This culture was incubated until lysis, after which phages were purified to create the acceptor phage stock.

ORACLE - Optimized Recombination

Recombination was performed by adding T7 Acceptor phages (MOI ~5) to 15 mL exponential phase 10G with pHT7Helper1 and pHRec1-Lib (shown as the donor plasmid in Figure 1), split across three 5 mL cultures. A high MOI is used to allow for one effective infection cycle. Cultures were incubated until lysis (~30 minutes). Lysed cultures were combined and purified. This phage population constitutes the initial recombined phage population and contained an estimated $2x10^7$ variants/mL in a total phage population of ~$2x10^{10}$ PFU/mL. The remainder of the phages are acceptor phages. A schematic of the recombination is shown in Figure 1 – Figure Supplement 1. Note pHT7Helper1 ensures progeny should remain viable by providing *gp17 in trans*.

ORACLE – Accumulation

Accumulation was performed by adding ~MOI of 0.2 of recombined phages (50 µl or ~1x10^9 total phages) to 5 mL of stationary phase *E. coli* 10G with pHT7Helper1 and pHCas9-3 resuspended in fresh Lb with kanamycin and spectinomycin. Cultures are incubated until lysis (~3.5 hours), then phages are purified. This MOI was chosen to target 1% of acceptor phages remaining in the final population as an internal control – the remainder of the phage population is accumulated variant phages. Stationary phase was used because it was more inhibitory based on EOP (Figure 1 - Figure Supplement 2A). Note pHT7Helper1 still ensures progeny should remain viable by providing *gp17 in trans* and progeny from accumulation do not fully express variant genes.

Methods Note: During ORACLE, the library gene is not actively repressed and some fraction of progeny phages are likely assembled with the variant gene, or contain chimeric tail fibers with both proteins. This may account for some degree of the skew we see during accumulation, although skewed residues were not consistent with enrichment patterns on 10G. Here, this skew was not significant, but this could be an influencing factor in future studies. The bias can feasibly be further reduced by a number of methods of varying complexity including repressing the genomic variant, increasing the amount of wild type provided *in trans*, or simply reducing the number of replication events during ORACLE.

Method Note: See Figure 1 - Figure Supplement 2A for inhibition results for versions of pHCas9. When sequenced, individual plaques after selection had

mutations in the region targeted by each gRNA, as expected for how resistance to Cas9 occurs. Note acceptor phages are not actively destroyed but are rather inhibited and maintained at the same concentration. Selection may be improved by using multiple guide RNAs or using sgRNA platforms.

ORACLE - Library Expression

Library expression was performed by adding the accumulated DMS library to 5 mL *E. coli* 10G (with no plasmid) at an MOI of ~1. Cultures are incubated for 30 minutes, then 200 µl chloroform is added to the culture to lyse any remaining cells and phages are purified. This constitutes the final phage variant library with full expression of the variant *gp17* tail fibers. This phage population is directly sequenced to establish the pre-selection library population.

Method Note: This MOI and culture conditions are chosen to prevent phages from undergoing more than 1 replication cycle. Additional replication cycles would result in skew towards variants that grow better on the host used for expression. At an MOI of 1, progeny from the first replication cycle already comprise the majority of the possible concentration of phages for T7. Chloroform is added at 30 minutes to halt any rogue second infections in process.

Method Note: Two points bear additional mention regarding ORACLE as a whole to create variant libraries. First, the importance of retaining variants that do not grow well on the host used to create the library cannot be overstated. These

variants are critical for mapping functional regions. For example, we used 10G to grow our library, which happened to have the most significant selection of the susceptible hosts and had depletion of many functional regions. The resolution of this assay would have been deeply impacting if these variants had been lost. Second, due to possible depletion and skew, it is critical to assay library distribution after insertion and expression of the variants in the phage, instead of, for example, assaying distribution in plasmid before phage insertion. While ORACLE is designed to avoid this problem, any selection that occurs during library construction needs to be identified prior to selection experiments.

DMS Selection

All DMS selection results besides single infection cycle experiments on resistant hosts noted later were performed in the same way. The T7 variant library was added to 5 mL of exponential host at an MOI of ~10^{-2} and the culture was allowed to fully lyse (typically 40 to 80 minutes depending on the host). Phage lysate was purified and then the titer established for the host the phage was being selected on. This process was then repeated using the selected phage lysate. An MOI of 10^{-2} was chosen to allow phages which grow slower a chance to replicate. For reference under these conditions we expect wildtype to complete four infection cycles on a susceptible host. Phage lysate from the second selection was retained and used as template for deep sequencing to establish the post-selection phage population. The entire process was repeated in biological triplicate for each host. Single infection cycle results cited in Figure 4 – Figure Supplement 4 and Figure 4 – Source Data 1 and 2 were performed on BW25113Δ*rfaG* and BW25113Δ*rfaD*

using an MOI of ~1 then waiting for the culture to fully lyse, after which phage lysate was purified and sequenced directly.

Method Note: The number of infection cycles should be carefully considered due to the effect of multiple exponential replication events on the phage population. As noted in the text, we expected less consistency as we increased infection cycles, especially on severely bottlenecked populations as seen on BW25113Δ*rfaD*, but additional infection cycles allows for increased competition between productive members.

Deep Sequencing Preparation and Analysis

We used deep sequencing to evaluate phage populations. We first amplified the tip domain by two step PCR, or tailed amplicon sequencing, using KAPA HiFi. Primers for deep sequencing attach to constant regions adjacent to the tip domain (the target region is 304 bp total, between the 5' NGS region and 3' pad on Figure 1 – Figure Supplement 1). Constant regions are also installed in the fixed region of the acceptor phages for the same size amplicon so acceptor phages can also be detected. The first PCR reaction adds an internal barcode (used for technical replicates to assay PCR skew), a variable N region (to assist with nucleotide diversity during deep sequencing, this is essential for DMS libraries due to low nucleotide diversity at each position), and the universal Illumina adapter. Undiluted phages are used as template. Four forward and four reverse primers were used in each reaction, each with a variable N count (0, 2, 4, or 8). Primers were mixed at equimolar ratios and total primers used was per recommended

primer concentration. PCR was performed using 12 total cycles in the first PCR reaction, then the product of this reaction was then purified. The second PCR reaction adds an index and the Illumina ‘stem’. 1 µl of purified product from the first reaction was used as template using 8 total PCR cycles. The product of this reaction was purified and was used directly for deep sequencing. Each phage population was sampled at least twice using separate internal barcodes, and no PCR reactions were pooled. Total PCR cycles overall for each sample was kept at 20x to avoid PCR skew. All phage samples were deep sequenced using an Illumina Miseq System, 2x250 read length using MiSeq Reagent Kit v2 or v2 Nano according to manufacturer documentation.

Paired-end Illumina sequencing reads were merged with FLASH (Fast Length Adjustment of SHort reads) using the default software parameters (Edgar and Flyvbjerg, 2015). Phred quality scores (Q scores) were used to compute the total number of expected errors (E) for each merged read and reads exceeding an Emax of 1 were removed (Magoč and Salzberg, 2011)**.** Wildtype, acceptor phages, and each variant were then counted in the deep sequencing output. We correlated read counts for each technical replicate to determine if there was any notable skew from PCR or deep sequencing. Replicates correlated extremely well ($R \geq 0.98$ for all samples) indicating no relevant PCR skew. Besides wildtype and acceptor counts, we included only sequences with single substitutions in our analysis. While this limited the scope of the analysis, it greatly reduced the possibility of deep sequencing error resulting in an incorrect read count for a variant, as virtually every relevant error would result in at least a double substitution in our library. With this in mind, and to avoid missing low abundance members after selection, we simply used a low read cutoff of 2 and did not utilize a pseudo-count of 1 for each position.

Of the 1660 variants, three (S487P, L524M and R542N) fell below our limit of detection in the variant pool before selection. These positions were excluded from analysis as a pre-selection population could not be accurately determined, although both S487P and L542M subsequently emerged in several post-selection populations indicating they were present below the limit of detection. Technical replicates of each biological replicate were aggregated and each biological replicate was correlated to determine reproducibility, reported for each relevant figure as a Figure Supplement. Possible outliers were identified by data sorting, and lower correlation in 10G was noted to primarily be a result of differences in enrichment of several polar, uncharged substitutions and proline substitutions in the first biological replicate. These substitutions are proximally located near exterior loop HI and may indicate an unknown variable or growth condition in the first replicate that produces a slightly different enrichment pattern. Correlation was otherwise robust and excluding only G479Q produces R=0.94, R=0.95, and R=0.99 as displayed in Figure 2 – Figure Supplement 1. Positively charged, downward facing substitutions were universally enriched in BW25113Δ*rfaD*, but correlation was reduced due to inconsistencies in which particular substitutions were the most enriched in a given replicate. While the same substitutions were enriched in all three replicates, suggesting reproducibility of results, the scale of enrichment varied considerably. Correlation was otherwise robust, excluding the single most enriched substitution of each replicate (G521R, A500R and G521H) produces R=0.90, R=0.93, and R=0.86 as displayed.

Method Note: BW25113Δ*rfaD* is the most resistant host by EOP, and we hypothesize these inconsistencies in F_N may arise due severe loss of diversity, ‘bottlenecking’ the population and causing stochastic differences in enrichment to

become magnified with multiple rounds of selection across independent experiments. Future work with very resistant hosts may benefit from fewer rounds of selection to prevent significant stochastic divergence. Single infection cycles are more consistent but provide fewer cycles for variants to directly compete with one another, as noted in methods for DMS selection.

To score enrichment for each variant we used a basic functional score (F), averaging results of the three biological replicates where $F = \bar{x}\ \frac{Variant\ \%_{Post-Passage}}{Variant\ \%_{Pre-Passage}}$. To compare variant performance across hosts we normalized functional score (F_N) to wildtype, where $F_N = \bar{x}\ \frac{Variant\ \%_{Post-Passage}}{Variant\ \%_{Pre-Passage}} \Big/ \frac{WT\ \%_{Post-Passage}}{WT\ \%_{Pre-Passage}}$.

Classifying Variants and Isolating Variants

To define variant behavior on *E. coli* 10G, BL21, and BW25113 we considered variants depleted if F_N was below 0.1 (i.e. performed ten times worse than wildtype), tolerated if between 0.1 and 2, and enriched if above 2 (i.e. performed twice as good as wildtype) (Figure 2). As wildtype T7 effectively grows on all three hosts, we reasoned that it would be more challenging for an enriched mutant to surpass wildtype than it would be for a mutant to become depleted. These cutoffs were supported based on preliminary plaque assay results and the extent of standard deviation across biological replicates. For BW25113Δ*rfaD* and BW25113Δ*rfaG* we further defined significantly enriched variants as

performing at least ten times better ($F_N \geq 10$) than wildtype because wildtype does not grow effectively on these strains (Figure 4).

We compared variant F_N across 10G, BL21 and BW25113 to further characterize each variant and find functional variants (Figure 3). We sought to identify variants that had meaningfully different performance on different hosts, which would be strong evidence that either the wildtype residue or the variant substitution was important in a host-specific context. In addition to providing direct insight intro structure-function relationships, such substitutions or positions are ideal engineering targets for altering host range or increasing activity in engineered phages. We considered substitutions that were depleted ($F_N < 0.1$) on all three hosts to be intolerant, while substitutions that were tolerated or enriched ($F_N \geq 0.1$) in all three hosts to be generally tolerated. Substitutions that were depleted on one host but tolerated or enriched on another were considered functional. To broadly characterize each position, we counted the number of substitutions at that position that fell into each category, and colored positions (Figure 3C) as functional if over 33% of substitutions at that position were functional, intolerant if over 50% of substitutions were intolerant, an tolerant otherwise. We found these cutoffs to effectively group residues of interest, although we note there remain substitutions that could be tolerant and relevant in different contexts or intolerant for these three hosts but not others.

For defining ideal host constriction mutants (Figure 6) we first constricted F_N values that were greater than 1 to reduce the impact of higher scores on this comparison. Specifically we generated Functional Difference (F_D) , where if $F_N < 1$, $F_D = F_N$, and where $F_N \geq 1$, $F_D = \frac{F_N - 1}{\max(Strain\ F_N) - 1} + 1$). F_D thus ranged from 0 to 2 for each substitution where scores above 1 are normalized to the maximum value for that host and fall between 1 and

2, minimizing but not eliminating weight for enrichment. We reasoned that for the purposes of finding host constriction mutants, the extent of enrichment for a substitution is less relevant than if that substitution did poorly on another host. Put another way, it does not matter if a substitution is tolerated or enriched so long as it is depleted on a different host. For example, V544R has an F_N of 9.09 in *E. coli* 10G but 0.07 in *E. coli* BW25113, while G479E has an F_N of 1.73 in *E. coli* 10G and falls below the limit of detection for *E. coli* BW25113. For host constriction both positions should be scored highly, as the mutations can be tolerated or enriched in one host but are depleted in another. In contrast A500H has an F_N of 7.46 in *E. coli* 10G and 1.2 in *E. coli* BW25113. While F_N differs significantly and the substitution is enriched on one host, it is still tolerated in the other host and thus makes a poor host constriction target. After generating F_D we simply subtracted the substitution's F_D on one host from the other in a pairwise comparison (Figure 6). Substitutions for host range constriction were considered ideal candidates if $|F_D| \geq 1$.

Variants with individual substitutions tested for EOP (Figures 4, 5 and 6) were either picked from plaques or created using SDM on pHRec1, which was used to create the variant using ORACLE.

Rosetta ΔΔG and Physicochemical Property Comparison and Calculations

The crystal structure of the *gp17* tip domain was obtained (PDB ID: 4A0T) and water molecules removed before calculations run. All modeling calculations were performed using the Rosetta molecular modeling suite v3.9. Substitutions were generated

using the standard ddg_monomer application (Kellogg et al., 2011) to enable local conformational to minimize energy. For comparison to F_D (Figure S5), a ΔΔG of 10 or greater was considered destabilizing and ΔΔG values between 10 and 30 were transformed to values between 0 and 1, with any ΔΔG greater than 30 set to 1. ΔΔG values below 10 were transformed to values between 0 and -1 based on a range to -9.29, the most stabilizing ΔΔG value calculated. We calculated F_D for this plot using the maximum F_D value from *E. coli* 10G, BL21 or BW25113. The maximum was used because any substitution that has a high F_D on any host was theorized to require a stable protein. Positions considered destabilizing (right side) are expected to have a very low maximum F_D, whereas stabilizing positions (left side) may have a low or high F_D based on tolerance of the substitution. Positions that were tolerated or functional that Rosetta predicted to be unstable on the right of the plot may be due to errors in stability calculations or actual structural distortions that are either smaller perturbations that don't affect fitness or are accommodated by quaternary arrangement. Alternatively, structural instability could be beneficial in some cases, allowing for enhanced receptor binding by, for example, exposing critical residues.

To compare physicochemical properties for *E. coli* 10G, BL21 and BW25113 we binned depleted ($F_N \leq 0.1$) tolerated ($F_N > 0.1$ and < 2) or enriched ($F_N \geq 2$) substitutions and derived the change in mass, hydrophilicity, and hydrophobicity for each substitution (Li et al., 2016b). For BW25113Δ*rfaG* and BW25113Δ*rfaD* we binned using and F_N of 10 for the cutoff for enrichment instead. Packages ggplot2 and ggpubr in R were then used to compare the means of the three groups using a Kruskal-Wallis test (Hollander and Wolfe, 1973), while subsequent pairwise comparisons were made using a Wilcoxon test

(Bauer, 1972). Effect size (r) was calculated r and can be summarized as r = abs(zscore)/sqrt(n) where n is the total number of observations among the two groups.

Statistical Analysis and Source Code

Alluvial plots (Figure 2 – Figure Supplement 2, Figure 4) and Figure 2F parallel plot were generated with RawGraphs. Violin plots for physicochemical properties are output from R. Significance for F_N values in Figure 2 – Source Data 2 and Figure 4 – Source Data 2 was defined as an average F_N two or more standard deviations from wild type or below the limit of detection in all biological replicates. P-values for EOP graphs compare plaque capability on the tested host to the reference host for the EOP graph. Hierarchical clustering in Figure 2- Figure Supplement 3 is performed in R using heatmap.2 without scaling. P-values to compare functional data and EOP measurements was performed using two tailed t-tests where values below the limit of detection were considered as the limit of detection. All other calculations and plots were made in Excel. Relevant statistical details of experiments can be found in the corresponding figure legends or relevant methods section. Source code is available at https://github.com/raman-lab/oracle.

REFERENCES

Abedon, S.T. (2011). Lysis from without. Bacteriophage *1*, 46–49. https://doi.org/10.4161/bact.1.1.13980.

Álvarez, B., Mencía, M., de Lorenzo, V., and Fernández, L.Á. (2020). In vivo diversification of target genomic sites using processive base deaminase fusions blocked by dCas9. Nat Commun *11*, 6436. https://doi.org/10.1038/s41467-020-20230-z.

Ando, H., Lemire, S., Pires, D.P., and Lu, T.K. (2015a). Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. Cell Systems *1*, 187–196. https://doi.org/10.1016/j.cels.2015.08.013.

Ando, H., Lemire, S., Pires, D.P., and Lu, T.K. (2015b). Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. Cell Syst *1*, 187–196. https://doi.org/10.1016/j.cels.2015.08.013.

Andrews, B., and Fields, S. (2021). Balance between promiscuity and specificity in phage λ host range. ISME J *15*, 2195–2205. https://doi.org/10.1038/s41396-021-00912-2.

Arimoto-Kobayashi, S., Anma, N., Yoshinaga, Y., Douki, T., Cadet, J., and Hayatsu, H. (2000). Oxidative damage and induced mutations in M13mp2 phage DNA exposed to N-nitrosopyrrolidine with UVA radiation. Mutagenesis *15*, 473–477. https://doi.org/10.1093/mutage/15.6.473.

Arthur, M., Arbeit, R.D., Kim, C., Beltran, P., Crowe, H., Steinbach, S., Campanelli, C., Wilson, R.A., Selander, R.K., and Goldstein, R. (1990). Restriction fragment length polymorphisms among uropathogenic Escherichia coli isolates: pap-related sequences compared with rrn operons. Infect Immun *58*, 471–479. .

Assad-Garcia, N., D'Souza, R., Buzzeo, R., Tripathi, A., Oldfield, L.M., Vashee, S., and Fouts, D.E. Cross-Genus "Boot-Up" of Synthetic Bacteriophage in Staphylococcus aureus by Using a New and Efficient DNA Transformation Method. Applied and Environmental Microbiology *88*, e01486-21. https://doi.org/10.1128/AEM.01486-21.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol *2*, 2006.0008. https://doi.org/10.1038/msb4100050.

Badran, A.H., and Liu, D.R. (2015). Development of potent in vivo mutagenesis plasmids with broad mutational spectra. Nat Commun *6*, 8425. https://doi.org/10.1038/ncomms9425.

Bauer, D.F. (1972). Constructing Confidence Sets Using Rank Statistics. Journal of the American Statistical Association *67*, 687–690. https://doi.org/10.1080/01621459.1972.10481279.

Bertozzi Silva, J., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. FEMS Microbiol Lett *363*. https://doi.org/10.1093/femsle/fnw002.

Canfield, G.S., and Duerkop, B.A. (2020). Molecular mechanisms of enterococcal-bacteriophage interactions and implications for human health. Curr. Opin. Microbiol. *56*, 38–44. https://doi.org/10.1016/j.mib.2020.06.003.

Chen, M., Zhang, L., Xin, S., Yao, H., Lu, C., and Zhang, W. (2017). Inducible Prophage Mutant of Escherichia coli Can Lyse New Host and the Key Sites of Receptor Recognition Identification. Front. Microbiol. *8*. https://doi.org/10.3389/fmicb.2017.00147.

Chen, Z., Guo, L., Zhang, Y., Walzem, R.L., Pendergast, J.S., Printz, R.L., Morris, L.C., Matafonova, E., Stien, X., Kang, L., et al. (2014). Incorporation of therapeutically modified bacteria into gut microbiota inhibits obesity. J. Clin. Invest. *124*, 3391–3406. https://doi.org/10.1172/JCI72517.

Clokie, M.R.J., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011). Phages in nature. Bacteriophage *1*, 31–45. https://doi.org/10.4161/bact.1.1.14942.

Cravens, A., Jamil, O.K., Kong, D., Sockolosky, J.T., and Smolke, C.D. (2021). Polymerase-guided base editing enables in vivo mutagenesis and rapid protein engineering. Nat Commun *12*, 1579. https://doi.org/10.1038/s41467-021-21876-z.

Csörgő, B., León, L.M., Chau-Ly, I.J., Vasquez-Rifo, A., Berry, J.D., Mahendra, C., Crawford, E.D., Lewis, J.D., and Bondy-Denomy, J. (2020). A compact Cascade–Cas3 system for targeted genome engineering. Nat Methods *17*, 1183–1190. https://doi.org/10.1038/s41592-020-00980-w.

Dedrick, R.M., Guerrero-Bustamante, C.A., Garlena, R.A., Russell, D.A., Ford, K., Harris, K., Gilmour, K.C., Soothill, J., Jacobs-Sera, D., Schooley, R.T., et al. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nature Medicine *25*, 730. https://doi.org/10.1038/s41591-019-0437-z.

Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M.J., et al. (2019). Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Reports *29*, 1336-1350.e4. https://doi.org/10.1016/j.celrep.2019.09.062.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., et al. (2008). The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. J Bacteriol *190*, 2597–2606. https://doi.org/10.1128/JB.01695-07.

Edgar, R.C., and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics *31*, 3476–3482. https://doi.org/10.1093/bioinformatics/btv401.

Elena, S.F., Agudelo-Romero, P., and Lalić, J. (2009). The Evolution of Viruses in Multi-Host Fitness Landscapes. Open Virol J *3*, 1–6. https://doi.org/10.2174/1874357900903010001.

Emond, S., Petek, M., Kay, E.J., Heames, B., Devenish, S.R.A., Tokuriki, N., and Hollfelder, F. (2020). Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. Nat Commun *11*, 3469. https://doi.org/10.1038/s41467-020-17061-3.

Eskenazi, A., Lood, C., Wubbolts, J., Hites, M., Balarjishvili, N., Leshkasheli, L., Askilashvili, L., Kvachadze, L., van Noort, V., Wagemans, J., et al. (2022). Combination of pre-adapted bacteriophage therapy and antibiotics for treatment of fracture-related infection due to pandrug-resistant Klebsiella pneumoniae. Nat Commun *13*, 302. https://doi.org/10.1038/s41467-021-27656-z.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011). A System for the Continuous Directed Evolution of Biomolecules. Nature *472*, 499–503. https://doi.org/10.1038/nature09929.

Faber, M.S., Van Leuven, J.T., Ederer, M.M., Sapozhnikov, Y., Wilson, Z.L., Wichman, H.A., Whitehead, T.A., and Miller, C.R. (2020). Saturation Mutagenesis Genome Engineering of Infective ΦX174 Bacteriophage via Unamplified Oligo Pools and Golden Gate Assembly. ACS Synth Biol *9*, 125–131. https://doi.org/10.1021/acssynbio.9b00411.

Favor, A.H., Llanos, C.D., Youngblut, M.D., and Bardales, J.A. (2020). Optimizing bacteriophage engineering through an accelerated evolution platform. Sci Rep *10*, 13981. https://doi.org/10.1038/s41598-020-70841-1.

Fister, S., Robben, C., Witte, A.K., Schoder, D., Wagner, M., and Rossmanith, P. (2016). Influence of Environmental Factors on Phage–Bacteria Interaction and on the Efficacy and Infectivity of Phage P100. Front Microbiol *7*. https://doi.org/10.3389/fmicb.2016.01152.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods *11*, 801–807. https://doi.org/10.1038/nmeth.3027.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507. https://doi.org/10.1016/j.jmb.2006.03.043.

Fraser, J.S., Maxwell, K.L., and Davidson, A.R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr. Opin. Microbiol. *10*, 382–387. https://doi.org/10.1016/j.mib.2007.05.018.

Garcia, E., Elliott, J.M., Ramanculov, E., Chain, P.S.G., Chu, M.C., and Molineux, I.J. (2003). The genome sequence of Yersinia pestis bacteriophage phiA1122 reveals an intimate history with the coliphage T3 and T7 genomes. J. Bacteriol. *185*, 5248–5262. .

Garcia-Doval, C., and Raaij, M.J. van (2012). Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. PNAS *109*, 9390–9395. https://doi.org/10.1073/pnas.1119719109.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464–471. https://doi.org/10.1038/nature24644.

Gebhart, D., Williams, S.R., and Scholl, D. (2017). Bacteriophage SP6 encodes a second tailspike protein that recognizes Salmonella enterica serogroups C2 and C3. Virology *507*, 263–266. https://doi.org/10.1016/j.virol.2017.02.025.

Gelman, S., Fahlberg, S.A., Heinzelman, P., Romero, P.A., and Gitter, A. (2021). Neural networks to learn protein sequence–function relationships from deep mutational scanning data. Proceedings of the National Academy of Sciences *118*, e2104878118. https://doi.org/10.1073/pnas.2104878118.

Gietz, R.D., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Meth. Enzymol. *350*, 87–96. .

González-García, V.A., Bocanegra, R., Pulido-Cid, M., Martín-Benito, J., Cuervo, A., and Carrascosa, J.L. (2015). Characterization of the initial steps in the T7 DNA ejection process. Bacteriophage *5*. https://doi.org/10.1080/21597081.2015.1056904.

Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R., and Gordon, J.I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe *6*, 279–289. https://doi.org/10.1016/j.chom.2009.08.003.

Grigonyte, A.M., Harrison, C., MacDonald, P.R., Montero-Blay, A., Tridgett, M., Duncan, J., Sagona, A.P., Constantinidou, C., Jaramillo, A., and Millard, A. (2020). Comparison of CRISPR and Marker-Based Methods for the Engineering of Phage T7. Viruses *12*, 193. https://doi.org/10.3390/v12020193.

Guo, H., Tse, L.V., Nieh, A.W., Czornyj, E., Williams, S., Oukil, S., Liu, V.B., and Miller, J.F. (2011). Target Site Recognition by a Diversity-Generating Retroelement. PLOS Genetics *7*, e1002414. https://doi.org/10.1371/journal.pgen.1002414.

Halperin, S.O., Tou, C.J., Wong, E.B., Modavi, C., Schaffer, D.V., and Dueber, J.E. (2018). CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. Nature *560*, 248–252. https://doi.org/10.1038/s41586-018-0384-8.

Hatfull, G.F., Dedrick, R.M., and Schooley, R.T. (2022). Phage Therapy for Antibiotic-Resistant Bacterial Infections. Annu Rev Med *73*, 197–211. https://doi.org/10.1146/annurev-med-080219-122208.

Heineman, R.H., Springman, R., and Bull, J.J. (2008). Optimal foraging by bacteriophages through host avoidance. Am. Nat. *171*, E149-157. https://doi.org/10.1086/528962.

Hess, G.T., Frésard, L., Han, K., Lee, C.H., Li, A., Cimprich, K.A., Montgomery, S.B., and Bassik, M.C. (2016). Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat Methods *13*, 1036–1042. https://doi.org/10.1038/nmeth.4038.

Hollander, M., and Wolfe, D.A. (1973). Nonparametric Statistical Methods (Wiley).

Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (2020). A continuous evolution system for contracting the host range of bacteriophage T7. Scientific Reports *10*, 1–8. https://doi.org/10.1038/s41598-019-57221-0.

Hoshiga, F., Yoshizaki, K., Takao, N., Miyanaga, K., and Tanji, Y. (2019). Modification of T2 phage infectivity toward Escherichia coli O157:H7 via using CRISPR/Cas9. FEMS Microbiology Letters *366*, fnz041. https://doi.org/10.1093/femsle/fnz041.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. Human Immunology *82*, 801–811. https://doi.org/10.1016/j.humimm.2021.02.012.

Huss, P., and Raman, S. (2020). Engineered bacteriophages as programmable biocontrol agents. Curr Opin Biotechnol *61*, 116–121. https://doi.org/10.1016/j.copbio.2019.11.013.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. (2021). Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. ELife *10*, e63775. https://doi.org/10.7554/eLife.63775.

Hyman, P., and Abedon, S.T. (2010). Bacteriophage Host Range and Bacterial Resistance. Advances in Applied Microbiology *70*, 217–248. https://doi.org/10.1016/S0065-2164(10)70007-1.

Jaschke, P.R., Lieberman, E.K., Rodriguez, J., Sierra, A., and Endy, D. (2012). A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. Virology *434*, 278–284. https://doi.org/10.1016/j.virol.2012.09.020.

Jensen, J.D., Parks, A.R., Adhya, S., Rattray, A.J., and Court, D.L. (2020). λ Recombineering Used to Engineer the Genome of Phage T7. Antibiotics (Basel) *9*, 805. https://doi.org/10.3390/antibiotics9110805.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotechnology *31*, 233–239. https://doi.org/10.1038/nbt.2508.

Jones, G.M., Stalker, J., Humphray, S., West, A., Cox, T., Rogers, J., Dunham, I., and Prelich, G. (2008). A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae. Nat Methods *5*, 239–241. https://doi.org/10.1038/nmeth.1181.

de Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., and Dutilh, B.E. (2019). Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends in Microbiology *27*, 51–63. https://doi.org/10.1016/j.tim.2018.08.006.

Kantor, A., McClements, M.E., and MacLaren, R.E. (2020). CRISPR-Cas9 DNA Base-Editing and Prime-Editing. International Journal of Molecular Sciences *21*, 6240. https://doi.org/10.3390/ijms21176240.

Kellogg, E.H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins *79*, 830–838. https://doi.org/10.1002/prot.22921.

Khan Mirzaei, M., and Nilsson, A.S. (2015). Isolation of Phages for Phage Therapy: A Comparison of Spot Tests and Efficiency of Plating Analyses for Determination of Host Range and Efficacy. PLoS One *10*. https://doi.org/10.1371/journal.pone.0118557.

Kilcher, S., and Loessner, M.J. (2019). Engineering Bacteriophages as Versatile Biologics. Trends in Microbiology *27*, 355–367. https://doi.org/10.1016/j.tim.2018.09.006.

Kilcher, S., Studer, P., Muessner, C., Klumpp, J., and Loessner, M.J. (2018a). Cross-genus rebooting of custom-made, synthetic bacteriophage genomes in L-form bacteria. PNAS *115*, 567–572. https://doi.org/10.1073/pnas.1714658115.

Kilcher, S., Studer, P., Muessner, C., Klumpp, J., and Loessner, M.J. (2018b). Cross-genus rebooting of custom-made, synthetic bacteriophage genomes in L-form bacteria. PNAS *115*, 567–572. https://doi.org/10.1073/pnas.1714658115.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424. https://doi.org/10.1038/nature17946.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. *110*, 14024–14029. https://doi.org/10.1073/pnas.1301301110.

Kuijpers, N.G., Solis-Escalante, D., Bosman, L., van den Broek, M., Pronk, J.T., Daran, J.-M., and Daran-Lapujade, P. (2013). A versatile, efficient strategy for assembly of

multi-fragment expression vectors in Saccharomyces cerevisiae using 60 bp synthetic recombination sequences. Microb Cell Fact *12*, 47. https://doi.org/10.1186/1475-2859-12-47.

Kutter, E.M., Kuhl, S.J., and Abedon, S.T. (2015). Re-establishing a place for phage therapy in western medicine. Future Microbiol *10*, 685–688. https://doi.org/10.2217/fmb.15.28.

Landsberger, M., Gandon, S., Meaden, S., Rollie, C., Chevallereau, A., Chabas, H., Buckling, A., Westra, E.R., and van Houte, S. (2018). Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. Cell *174*, 908-916.e12. https://doi.org/10.1016/j.cell.2018.05.058.

Langer, S.J., Ghafoori, A.P., Byrd, M., and Leinwand, L. (2002). A genetic screen identifies novel non-compatible loxP sites. Nucleic Acids Res *30*, 3067–3077. .

Langridge, G.C., Phan, M.-D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., et al. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res *19*, 2308–2316. https://doi.org/10.1101/gr.097097.109.

Lee, J.M., Huddleston, J., Doud, M.B., Hooper, K.A., Wu, N.C., Bedford, T., and Bloom, J.D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. PNAS *115*, E8276–E8285. https://doi.org/10.1073/pnas.1806133115.

Li, C.-L.F., Santhanam, B., Webb, A.N., Zupan, B., and Shaulsky, G. (2016a). Gene discovery by chemical mutagenesis and whole-genome sequencing in Dictyostelium. Genome Res. *26*, 1268–1276. https://doi.org/10.1101/gr.205682.116.

Li, Z., Tang, J., and Guo, F. (2016b). Identification of 14-3-3 Proteins Phosphopeptide-Binding Specificity Using an Affinity-Based Computational Approach. PLOS ONE *11*, e0147467. https://doi.org/10.1371/journal.pone.0147467.

Lin, T.-Y., Lo, Y.-H., Tseng, P.-W., Chang, S.-F., Lin, Y.-T., and Chen, T.-S. (2012). A T3 and T7 Recombinant Phage Acquires Efficient Adsorption and a Broader Host Range. PLOS ONE *7*, e30954. https://doi.org/10.1371/journal.pone.0030954.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507.

Marinelli, L.J., Piuri, M., Swigoňová, Z., Balachandran, A., Oldfield, L.M., Kessel, J.C. van, and Hatfull, G.F. (2008). BRED: A Simple and Powerful Tool for Constructing Mutant and Recombinant Bacteriophage Genomes. PLOS ONE *3*, e3957. https://doi.org/10.1371/journal.pone.0003957.

Marinelli, L.J., Piuri, M., and Hatfull, G.F. (2019). Genetic Manipulation of Lytic Bacteriophages with BRED: Bacteriophage Recombineering of Electroporated DNA. Methods Mol. Biol. *1898*, 69–80. https://doi.org/10.1007/978-1-4939-8940-9_6.

Medhekar, B., and Miller, J.F. (2007). Diversity-generating retroelements. Current Opinion in Microbiology *10*, 388–395. https://doi.org/10.1016/j.mib.2007.06.004.

Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., and Lenski, R.E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. Science *335*, 428–432. https://doi.org/10.1126/science.1214449.

Miller, S.M., Wang, T., and Liu, D.R. (2020). Phage-assisted continuous and non-continuous evolution. Nat Protoc *15*, 4101–4127. https://doi.org/10.1038/s41596-020-00410-3.

Mizuno, C.M., Luong, T., Cederstrom, R., Krupovic, M., Debarbieux, L., and Roach, D.R. (2020). Isolation and Characterization of Bacteriophages That Infect Citrobacter rodentium, a Model Pathogen for Intestinal Diseases. Viruses *12*. https://doi.org/10.3390/v12070737.

Molineux, I.J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol. Microbiol. *40*, 1–8. https://doi.org/10.1046/j.1365-2958.2001.02357.x.

Moore, C.L., Papa, L.J., and Shoulders, M.D. (2018). A Processive Protein Chimera Introduces Mutations across Defined DNA Regions In Vivo. J. Am. Chem. Soc. *140*, 11560–11564. https://doi.org/10.1021/jacs.8b04001.

Mutalik, V.K., and Arkin, A.P. (2022). A Phage Foundry Framework to Systematically Develop Viral Countermeasures to Combat Antibiotic-Resistant Bacterial Pathogens. IScience *25*, 104121. https://doi.org/10.1016/j.isci.2022.104121.

Mutalik, V.K., Adler, B.A., Rishi, H.S., Piya, D., Zhong, C., Koskella, B., Kutter, E.M., Calendar, R., Novichkov, P.S., Price, M.N., et al. (2020). High-throughput mapping of the phage resistance landscape in E. coli. PLOS Biology *18*, e3000877. https://doi.org/10.1371/journal.pbio.3000877.

Nayeemul Bari, S.M., Walker, F.C., Cater, K., Aslan, B., and Hatoum-Aslan, A. (2017). Strategies for Editing Virulent Staphylococcal Phages Using CRISPR-Cas10. ACS Synth Biol *6*, 2316–2325. https://doi.org/10.1021/acssynbio.7b00240.

Nobrega, F.L., Vlot, M., de Jonge, P.A., Dreesens, L.L., Beaumont, H.J.E., Lavigne, R., Dutilh, B.E., and Brouns, S.J.J. (2018). Targeting mechanisms of tailed bacteriophages. Nature Reviews Microbiology *16*, 760–773. https://doi.org/10.1038/s41579-018-0070-8.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq; high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods *6*, 767–772. https://doi.org/10.1038/nmeth.1377.

Oppenheim, A.B., Rattray, A.J., Bubunenko, M., Thomason, L.C., and Court, D.L. (2004). In vivo recombineering of bacteriophage lambda by PCR fragments and single-strand oligonucleotides. Virology *319*, 185–189. https://doi.org/10.1016/j.virol.2003.11.007.

Pagnout, C., Sohm, B., Razafitianamaharavo, A., Caillet, C., Offroy, M., Leduc, M., Gendre, H., Jomini, S., Beaussart, A., Bauda, P., et al. (2019). Pleiotropic effects of rfa-gene mutations on Escherichia coli envelope properties. Sci Rep *9*. https://doi.org/10.1038/s41598-019-46100-3.

Pires, D.P., Cleto, S., Sillankorva, S., Azeredo, J., and Lu, T.K. (2016). Genetically Engineered Phages: a Review of Advances over the Last Decade. Microbiol. Mol. Biol. Rev. *80*, 523–543. https://doi.org/10.1128/MMBR.00069-15.

Pires, D.P., Monteiro, R., Mil-Homens, D., Fialho, A., Lu, T.K., and Azeredo, J. (2021). Designing P. aeruginosa synthetic phages with reduced genomes. Sci Rep *11*, 2164. https://doi.org/10.1038/s41598-021-81580-2.

Principi, N., Silvestri, E., and Esposito, S. (2019). Advantages and Limitations of Bacteriophages for the Treatment of Bacterial Infections. Front. Pharmacol. *10*. https://doi.org/10.3389/fphar.2019.00513.

Pryor, J.M., Potapov, V., Bilotti, K., Pokhrel, N., and Lohman, G.J.S. (2022). Rapid 40 kb Genome Construction from 52 Parts through Data-optimized Assembly Design. ACS Synth. Biol. https://doi.org/10.1021/acssynbio.1c00525.

Pulkkinen, E.M., Hinkley, T.C., and Nugen, S.R. (2019). Utilizing in vitro DNA assembly to engineer a synthetic T7 Nanoluc reporter phage for Escherichia coli detection. Integrative Biology *11*, 63–68. https://doi.org/10.1093/intbio/zyz005.

Qimron, U., Marintcheva, B., Tabor, S., and Richardson, C.C. (2006). Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. Proc. Natl. Acad. Sci. U.S.A. *103*, 19039–19044. https://doi.org/10.1073/pnas.0609428103.

Raman, S., Taylor, N., Genuth, N., Fields, S., and Church, G.M. (2014). Engineering Allostery. Trends Genet *30*, 521–528. https://doi.org/10.1016/j.tig.2014.09.004.

Ramirez-Chamorro, L., Boulanger, P., and Rossier, O. (2021). Strategies for Bacteriophage T5 Mutagenesis: Expanding the Toolbox for Phage Genome Engineering. Frontiers in Microbiology *12*. .

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012). Going viral: next generation sequencing applied to human gut phage populations. Nat Rev Microbiol *10*, 607–617. https://doi.org/10.1038/nrmicro2853.

Romero, P.A., Tran, T.M., and Abate, A.R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. PNAS *112*, 7159–7164. https://doi.org/10.1073/pnas.1422285112.

Rousset, F., Cui, L., Siouve, E., Becavin, C., Depardieu, F., and Bikard, D. (2018). Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. PLOS Genetics *14*, e1007749. https://doi.org/10.1371/journal.pgen.1007749.

Rustad, M., Eastlund, A., Marshall, R., Jardine, P., and Noireaux, V. (2017). Synthesis of Infectious Bacteriophages in an E. coli-based Cell-free Expression System. JoVE (Journal of Visualized Experiments) e56144. https://doi.org/10.3791/56144.

Salmond, G.P.C., and Fineran, P.C. (2015). A century of the phage: past, present and future. Nat Rev Microbiol *13*, 777–786. https://doi.org/10.1038/nrmicro3564.

Santos, A.L., Moreirinha, C., Lopes, D., Esteves, A.C., Henriques, I., Almeida, A., Domingues, M.R.M., Delgadillo, I., Correia, A., and Cunha, Â. (2013). Effects of UV Radiation on the Lipids and Proteins of Bacteria Studied by Mid-Infrared Spectroscopy. Environ. Sci. Technol. *47*, 6306–6315. https://doi.org/10.1021/es400660g.

Sausset, R., Petit, M.A., Gaboriau-Routhiau, V., and De Paepe, M. (2020). New insights into intestinal phages. Mucosal Immunology *13*, 205–215. https://doi.org/10.1038/s41385-019-0250-5.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., et al. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrob. Agents Chemother. *61*, e00954-17. https://doi.org/10.1128/AAC.00954-17.

Selle, K., Fletcher, J.R., Tuson, H., Schmitt, D.S., McMillan, L., Vridhambal, G.S., Rivera, A.J., Montgomery, S.A., Fortier, L.-C., Barrangou, R., et al. In Vivo Targeting of Clostridioides difficile Using Phage-Delivered CRISPR-Cas3 Antimicrobials. MBio *11*, e00019-20. https://doi.org/10.1128/mBio.00019-20.

Shen, T.-C.D., Albenberg, L., Bittinger, K., Chehoud, C., Chen, Y.-Y., Judge, C.A., Chau, L., Ni, J., Sheng, M., Lin, A., et al. (2015). Engineering the gut microbiota to treat hyperammonemia. J Clin Invest *125*, 2841–2850. https://doi.org/10.1172/JCI79214.

Shibai, A., Takahashi, Y., Ishizawa, Y., Motooka, D., Nakamura, S., Ying, B.-W., and Tsuru, S. (2017). Mutation accumulation under UV radiation in Escherichia coli. Sci Rep *7*, 14531. https://doi.org/10.1038/s41598-017-15008-1.

Shin, J., Jardine, P., and Noireaux, V. (2012). Genome replication, synthesis, and assembly of the bacteriophage T7 in a single cell-free reaction. ACS Synth Biol *1*, 408–413. https://doi.org/10.1021/sb300049p.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

Silva, Y.J., Costa, L., Pereira, C., Mateus, C., Cunha, A., Calado, R., Gomes, N.C.M., Pardo, M.A., Hernandez, I., and Almeida, A. (2014). Phage therapy as an approach to prevent Vibrio anguillarum infections in fish larvae production. PLoS ONE *9*, e114197. https://doi.org/10.1371/journal.pone.0114197.

Simon, A.J., Morrow, B.R., and Ellington, A.D. (2018). Retroelement-Based Genome Editing and Evolution. ACS Synth. Biol. *7*, 2600–2611. https://doi.org/10.1021/acssynbio.8b00273.

Sordi, L.D., Lourenço, M., and Debarbieux, L. (2019). The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract. Cell Host & Microbe *25*, 210–218. https://doi.org/10.1016/j.chom.2019.01.018.

Stemmer, W.P.C. (1994). Rapid evolution of a protein in vitro by DNA shuffling. Nature *370*, 389–391. https://doi.org/10.1038/370389a0.

Tran, N.Q., Rezende, L.F., Qimron, U., Richardson, C.C., and Tabor, S. (2008). Gene 1.7 of bacteriophage T7 confers sensitivity of phage growth to dideoxythymidine. Proc Natl Acad Sci U S A *105*, 9373–9378. https://doi.org/10.1073/pnas.0804164105.

Valvano, M.A., Messner, P., and Kosma, P. (2002). Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. Microbiology, *148*, 1979–1989. https://doi.org/10.1099/00221287-148-7-1979.

Wetmore, K.M., Price, M.N., Waters, R.J., Lamson, J.S., He, J., Hoover, C.A., Blow, M.J., Bristow, J., Butland, G., Arkin, A.P., et al. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. MBio *6*, e00306-15. https://doi.org/10.1128/mBio.00306-15.

Wrenbeck, E.E., Klesmith, J.R., Stapleton, J.A., Adeniran, A., Tyo, K.E.J., and Whitehead, T.A. (2016). Plasmid-based one-pot saturation mutagenesis. Nat Methods *13*, 928–930. https://doi.org/10.1038/nmeth.4029.

Wu, Z., Kan, S.B.J., Lewis, R.D., Wittmann, B.J., and Arnold, F.H. (2019). Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences *116*, 8852–8858. https://doi.org/10.1073/pnas.1901979116.

Yehl, K., Lemire, S., Yang, A.C., Ando, H., Mimee, M., Torres, M.D.T., de la Fuente-Nunez, C., and Lu, T.K. (2019). Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. Cell *179*, 459-469.e9. https://doi.org/10.1016/j.cell.2019.09.015.

Yosef, I., Goren, M.G., Globus, R., Molshanski-Mor, S., and Qimron, U. (2017). Extending the Host Range of Bacteriophage Particles for DNA Transduction. Molecular Cell *66*, 721-728.e3. https://doi.org/10.1016/j.molcel.2017.04.025.

Zheng, K., Wang, Y., Li, N., Jiang, F.-F., Wu, C.-X., Liu, F., Chen, H.-C., and Liu, Z.-F. (2018). Highly efficient base editing in bacteria using a Cas9-cytidine deaminase fusion. Commun Biol *1*, 1–6. https://doi.org/10.1038/s42003-018-0035-5.

Chapter 5

Publication

Engineering Bacteriophage T7 with Metagenomic Motif Curation

Authors: Phil Huss[1,2,3], Kristopher Kieft[2,3], Anthony Meger[1], Kyle Nishikawa[1], Karthik Anantharaman[2] and Srivatsan Raman[1,2,4*]

Contributions:

P.H.: Conceptualization, Data curation, Software, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing

K.K.: Conceptualization, Data curation, Software, Formal analysis, Investigation, Methodology, Writing - review and editing

A.M.: Software

K.N.: Software

K.A.: Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing - review and editing

S.R.: Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing - review and editing

Affiliations:

[1]Department of Biochemistry, University of Wisconsin-Madison

[2]Department of Bacteriology, University of Wisconsin-Madison

[3]Microbiology Doctoral Training Program, University of Wisconsin-Madison

[4]Department of Chemical and Biological Engineering, University of Wisconsin-Madison

[*] Correspondence: sraman4@wisc.edu (Srivatsan Raman)

Abstract

Bacteriophages are bacterial viruses that are valuable tools for targeted killing of drug-resistant bacteria and precision editing of the microbiome. Bacteriophage specificity and activity can be improved and tailored by changing the phage genome to engineer the phage. Here, we establish a new process called Motif Curation to find sequence motifs in phage metagenomes datasets that drive phage activity. This approach is not reliant on overall protein homology, allowing us to deeply mine rich metagenomic datasets with diverse phage genomes. Using Motif Curation we identify over 10^4 activity-driving motifs for the T7 phage receptor binding protein. Screening these phages revealed hundreds of phages with novel host specificity with most active motifs sourced from distantly related phages. We examined position and substitution preference to trace key combinations of substitutions that drive activity and host range and shape the sequence-function landscape. We reveal epistatic combinations of substitutions that are individually deleterious but when combined show dramatic increases in activity and identify variants that eliminate pathogenic *Escherichia coli* causing urinary tract infections. Motif Curation allows identification of motifs that are highly divergent from the wildtype sequence and lays a foundation for identifying motifs in diverse metagenomic datasets that influence phage activity. We envision Motif Curation as a useful tool for tapping into the ever-increasing wealth of phage metagenomic data to explain sequence-function relationships and engineer phages.

Highlights

- We establish Motif Curation as a high throughput approach for mining >10^4 motifs from diverse metagenomic databases to understand sequence-function relationships and engineer phages.
- Motif Curation is not reliant on protein homology but instead uses seeded scores from DMS to search for small motifs, allowing for deep exploration of metagenomic datasets.
- Metagenomic motifs from distantly related phages drove more phage activity in the T7 RBP, indicating mining deep into metagenomic space was essential for creating active phage variants.
- Using Motif Curation, we created hundreds of T7 variants with novel host specificity and trace activity to key positions and combinations of substitutions.
- We reveal unique epistatic interactions in the RBP and generate T7 variants that eliminate insensitive pathogenic *E. coli* causing UTIs.

Introduction

Bacteriophages, or ‘phages’, are bacterial viruses that play an essential role in microbial ecosystems. Phages typically have a narrow host range and only infect and kill specific bacteria without affecting other bacteria or mammalian cells. This specificity has made them valuable tools for applications including targeted killing of drug-resistant bacteria, precision modulation of the microbiome, as gene delivery devices, and as diagnostic sensors for pathogens (Canfield and Duerkop, 2020; Chen et al., 2017; Clokie et al., 2011; Dedrick et al., 2019; Dunne et al., 2019; Eskenazi et al., 2022; Garcia et al., 2003; Hatfull et al., 2022; Heineman et al., 2008; Holtzman et al., 2020; Huss and Raman, 2020; Kilcher and Loessner, 2019; Kutter et al., 2015; Meyer et al., 2012; Schooley et al., 2017; Yehl et al., 2019). Phage specificity and activity can be improved and tailored by engineering the phage. However, determining exactly what sequence changes to make to tailor phage activity, and the extent to which activity can be changed by doing so, has remained a challenge. Here, we establish a new process called Motif Curation to find sequence motifs in metagenomic datasets. We demonstrate how these short motifs can improve and tailor phage activity. This approach is not reliant on overall protein homology, allowing us to deeply mine metagenomic datasets to find motifs that alter phage function from diverse groups of phages.

Recent years have seen a staggering increase in metagenomic data for bacteriophages. Phage metagenomes has been harvested from diverse sources, including the gut microbiome (Abeles and Pride, 2014; Benler et al., 2021; Shkoporov and Hill, 2019; Shkoporov et al., 2019), from oceans (Brum and Sullivan, 2015; Coutinho et al., 2017; Luo et al., 2022; Winter et al., 2014), and from lakes and soil environments

(Bruder et al., 2016; Gu et al., 2021; Pratama and van Elsas, 2018; Wu et al., 2021). With over 10^{31} phages estimated to exist on earth, even this wealth of information represents only a small fraction of total phage diversity (Cobián Güemes et al., 2016). These metagenomic datasets represent a treasure trove of sequences that can be mined to engineer new function and alter activity in phages.

We sought to tap into these vast metagenomic datasets to identify sequences that influence phage activity by inserting metagenomic motifs into the receptor binding protein (RBP) of T7 bacteriophage. RBPs are responsible for interacting with bacterial host receptors to position the phage on the host surface and initiate infection (Bertozzi Silva et al., 2016; de Jonge et al., 2019; Nobrega et al., 2018). Because they are crucial determinants of phage activity, RBPs have proven to be excellent targets for engineering and altering phage activity (Ando et al., 2015; Andrews and Fields, 2021; Dedrick et al., 2019; Dunne et al., 2019; Holtzman et al., 2020; Huss et al., 2021, 2021; Yehl et al., 2019). Metagenomic sequences of similar phages have been used to exchange key domains in RBPs (Ando et al., 2015; Chen et al., 2017; Gebhart et al., 2017; Yosef et al., 2017). These exchanges can cause dramatic shifts in phage activity but have been greatly limited by the small number of available homologous sequences.

To mine the enormous depth of metagenomic databases we thus looked beyond related or homologous phages. Phages are highly diverse and have no universally conserved genes by which to readily identify similar phages (Hatfull, 2008). Furthermore, many sequences in metagenomic databases have little or no homology to previously characterized viruses, with most sequences being completely novel (Coutinho et al., 2017). Finding phages that may harbor a relevant sequence in complex metagenomic

datasets is made even more challenging for RBPs because these proteins are themselves highly diverse (Fraser et al., 2006). However, recombination and exchange of short motifs is a viable approach for adapting phages to different hosts (Fraser et al., 2006, 2007; Lin et al., 2012). These small motifs are responsible for critical receptor interactions that have enormous influence on phage activity. Motifs that may influence phage activity may therefore be hidden in non-homologous proteins that make up the bulk of metagenomic datasets. We reasoned that searching for motifs known to be relevant for phage activity was an effective approach to mine metagenomic datasets for such motifs without relying on overall protein homology.

Here, we created a process called Motif Curation to identify these relevant motifs in diverse metagenomic datasets. We used Motif Curation to create T7 phage variants with metagenomic motifs inserted into the tip domain of the RBP. We leveraged data from Deep Mutational Scanning (DMS) of this domain previously performed on multiple bacterial hosts (Huss et al., 2021). Using this dataset, we seed scores for every possible substitution to determine which theoretical motifs are most relevant for phage activity. We then determine the frequency of the best motifs in NCBI and IMG/VR metagenomic databases to curate a final list of candidate motifs. Motif Curation identified over 10^4 relevant motifs from diverse metagenomic phages located throughout the tip domain. We show how these motifs are found in greater abundance in structural proteins and finding motifs requires correctly seeded scores from the DMS set, indicating motifs play a critical functional role. Our process allowed us to identify motifs that are highly divergent from the wildtype sequence which can be sourced from diverse, non-homologous phage sequences.

Screening these phages revealed hundreds of phages with novel host specificity. Furthermore, we saw that metagenomic motifs from distantly related phages primarily drove phage activity, indicating deep mining of metagenomic datasets was essential for tuning host range. We examined position and substitution preference across different hosts to reveal regions in the T7 RBP more tolerant to multiple substitutions and traced key combinations of substitutions that drive activity and host range. We reveal epistatic combinations of substitutions that are individually deleterious but when combined show dramatic increases in activity. Finally, we identify several variants in our library that eliminate pathogenic *Escherichia coli* causing urinary tract infections (UTIs) that are insensitive to wildtype T7, indicating our approach can be used to tune host specificity and increase phage activity on many different hosts.

Overall, our study lays a foundation for identifying motifs in diverse metagenomic datasets that influence phage activity. This process can be used to mine diverse motifs from non-homologous phage proteins that can tune host range and improve phage activity. We paint a broader picture of the malleability of the tip domain of the T7 RBP and how changes in the sequence-function landscape facilitate receptor interaction. We envision Motif Curation as a useful tool for tapping into the ever-increasing wealth of phage metagenomic data to explain sequence-function relationships and engineer phages.

Results

Motif curation identifies motifs from diverse phages

We identified three challenges that needed to be overcome to identify metagenomic motifs that influence phage activity. First, metagenomic proteins that contain relevant motifs but have little sequence homology to the original phage protein cannot be found by simply comparing the overall protein sequence. Structural proteins are very diverse and relying on overall protein similarity is highly restrictive (Hatfull, 2008). For example, a BLAST search of the T7 RBP tip domain showed only 83 non-identical sequences, barely scratching the surface of available metagenomic sequences. Second, motifs that influence phage activity are likely to be substantially different than the wildtype sequence. Identifying these motifs requires more information than just the wildtype sequence and knowing which substitutions are most relevant is critical for finding motifs that influence activity. Third, we needed an approach that could determine where motifs needed to be inserted in a protein to influence phage function. We surmised any approach that was unable to surmount these challenges would not be able to effectively mine diverse metagenomic sequences for relevant motifs.

To overcome these challenges we developed Motif Curation, a high-throughput approach for identifying metagenomic motifs that leverages data from deep mutational scanning to find motifs that influence phage activity (Figure 1A). Motif Curation is not reliant on overall sequence homology because it queries metagenomic databases at the motif level. This allows us to reach deep into metagenomic space and mine motifs from phages that are not closely related to the wildtype phage. Further, by using seed data from deep mutational scanning, Motif Curation enables the identification of motifs that

significantly differ from the wildtype sequence but are still relevant to phage activity. These motifs can be found because every possible substitution at every position in candidate motifs is evaluated, allowing significant deviation from the wildtype sequence. Finally, Motif Curation precisely identifies where motifs need to be inserted in a sequence to influence phage activity because the position of the DMS seed scores is retained and used to place the motif in the most relevant location. Motif Curation is thus well suited to identify metagenomic motifs that influence phage activity even for highly variable proteins like RBPs.

The first step of Motif Curation is weighing every possible substitution at every possible position in the domain to determine a ‘seed score’ for functionality (Figure 1B and Supplementary File 1). We leveraged data from a previous study where we performed DMS on this region to generate these scores (Huss et al., 2021). In that study, we used ORACLE (Optimized Recombination, Accumulation, and Library Expression) to create an unbiased DMS library of phage variants that could be tested on multiple hosts. This allowed us to compare the activity of each substitution between hosts and establish functional roles for substitutions across the tip domain. The seed score is then generated based on two observations. First, how that substitution performed compared to wildtype and other substitutions at that position, and second how differently the substitution scored for various hosts, which indicates functional importance (see methods). Scores are well distributed throughout the tip domain and favor positions that are known to influence phage activity (Figure 1B). Weighing these results allowed us to refine sets of substitutions that are contextually relevant to the tip domain while retaining the position of the seed substitution to determine where motifs need to be inserted to influence activity.

The next step of Motif Curation is mining for seeded motifs in metagenomic datasets to determine which motifs are most abundant. We reasoned those motifs that influence phage activity are more likely to be found in greater abundance. Seed scores were used to screen possible 6-AA and 10-AA motifs mined from ~60,000 unique metagenomic structural proteins sourced from NCBI and IMG/VR databases. Motifs were filtered based on a hit threshold to ensure motifs were sourced multiple times. We found the number of metagenomic hits varied dramatically for different motifs (see Figure S1A). After applying our hit threshold, we successfully identified 15,565 candidate 6mer motifs and 3,549 candidate 10mer motifs from these metagenomic structural proteins (Figure 1C).

We then used two approaches to evaluate if Motif Curation was identifying relevant motifs. First, we first repeated motif mining using the same DMS seed scores in a comparably sized set of metagenomic non-structural proteins. Given that we sought to find motifs relevant for receptor recognition, we reasoned that we should mine many fewer motifs in non-structural proteins. Indeed, we mined fewer than ten times as many motifs searching this set of proteins (Figure 1C). We next evaluated the importance of the seed scores using two randomization schemes. We expected that both the position and identity of each substitution seed score was critical for accurately mining motifs. Altering the seed scores should result in fewer hits in the metagenomic database as scores lose relevance and mining becomes more random. We curated motifs using randomized seed scores across the entire tip domain and using randomized scores within each position (see Supplementary File 1). Both randomization approaches mined less than 10 times fewer motifs compared to the correctly seeded scores (Figure 1C). Overall, these results

showed that motif curation was able to identify more metagenomic motifs over the hit threshold when mining from relevant structural proteins and with correct DMS seeding, and many more motifs than similar proteins using BLAST, indicating curated motifs were promising candidates for influencing phage activity.

We next sought to verify that our approach was not restricted to curating motifs from only Podoviridea phages closely related to T7 phages. We examined protein taxonomy from the NCBI portion of our dataset, which included 28,989 source proteins that collectively informed 84,650 source motifs. Examining the taxonomy of proteins from which motifs were mined, we discovered that motifs were derived from a diverse set of phages, most significantly from the families Siphoviridae and Myoviridae (Figure 1D). The taxonomic origin of mined motifs closely paralleled the overall distribution of the metagenomic protein library, indicating our motif mining approach was able to reach deeply into metagenomic datasets to find motifs from many diverse phages and was not restricted only to closely related phages.

Finally, we generated our curated motif library by swapping the metagenomic motif for the native sequence using ORACLE. Our final phage library consisted of the top ~10^4 variants, with variants having between 2 and 10 substitutions with an average of 6 substitutions (see Figure S1B and Supplementary File 2). Our approach was thus able to successfully curate a large phage library with motifs both similar to and highly divergent from the wildtype sequence.

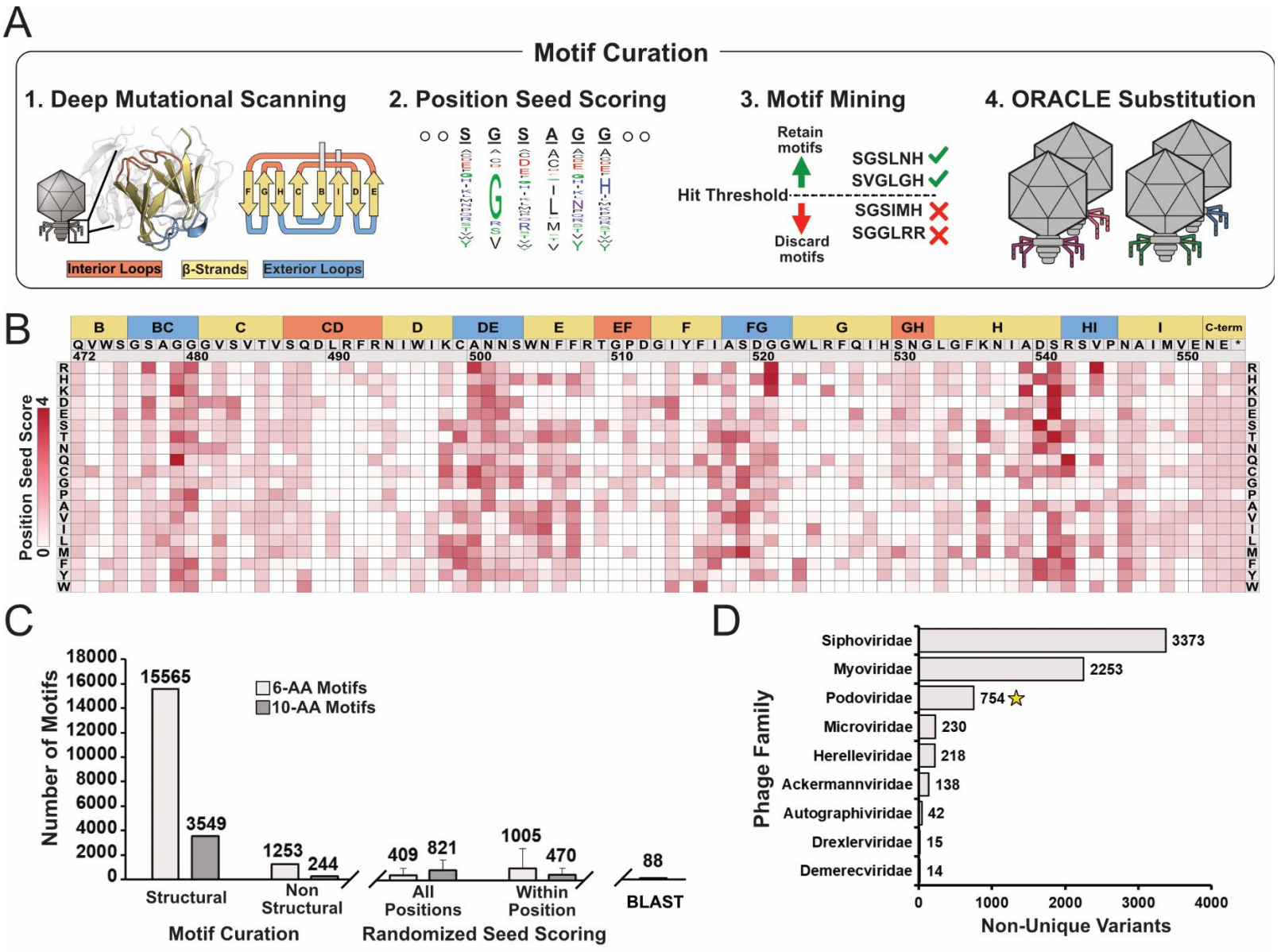

A
Motif Curation
1. Deep Mutational Scanning
2. Position Seed Scoring
3. Motif Mining
4. ORACLE Substitution
Interior Loops
β-Strands
Exterior Loops
Retain motifs
Hit Threshold
Discard motifs
SGSLNH
SVGLGH
SGSIMH
SGGLRR
B
Position Seed Score
C
Number of Motifs
6-AA Motifs
10-AA Motifs
15565
3549
1253
244
409
821
1005
470
88
Structural
Non Structural
All Positions
Within Position
BLAST
Motif Curation
Randomized Seed Scoring
D
Phage Family
Siphoviridae
Myoviridae
Podoviridae
Microviridae
Herelleviridae
Ackermannviridae
Autographiviridae
Drexlerviridae
Demerecviridae
3373
2253
754
230
218
138
42
15
14
Non-Unique Variants

Figure 1. Metagenomic motif curation generates diverse phage variants

(A) Illustration of motif curation. Functional mapping from deep mutational scanning previously performed on the tip domain of the T7 bacteriophage receptor binding protein (crystal structure and secondary structure topology shown color coded with interior loops as red, β-sheets as yellow and exterior loops as blue) (1) is used to seed scores for probable motifs (2), which are then mined in metagenomic databases (3). Motifs passing a hit threshold are substituted into phages in a phage library using ORACLE (4). **(B)** Heat map showing seed score (red gradient) associated with every substitution in the tip domain. Substitutions are shown top to bottom while position (based on PDB 4A0T), wild type amino acid and secondary structure topology is shown left to right. **(C)** Number of 6-AA (light grey) or 10-AA (dark grey) motifs found passing hit threshold using motif curation and metagenomic phage structural proteins or non-structural proteins (left), using motif curation with metagenomic structural proteins after fully randomizing seed scores or randomizing seed scores at each position (middle, mean ± SD of triplicate randomizations), or number of hits using BLAST (right). **(D)** Number of non-unique variants sourced from different phage families derived from proteins annotated in NCBI. The family for T7 phage (podoviridae) is marked with a yellow star.

Library selection reveals family and location preferences for active phage variants

We evaluated the activity of phage variants in our library by selecting the library on a panel of five *E. coli* hosts. We used the same host panel used to assess the seed DMS library to allow for direct comparison to phage variants with single substitutions. Three hosts, B strain derivative BL21, K-12 derivative BW25113, and DH10B derivative 10G, are susceptible to wildtype T7 infection. The remaining two hosts, *E. coli* deletion strains BW25113Δ*rfaG* and BW25113Δ*rfaD,* are resistant to wildtype T7. The deleted host genes are required for the biosynthesis of LPS, the known receptor for T7 in E. coli, and deletion of these genes creates truncated LPS that reduces T7 phage's ability to infect these strains by several orders of magnitude (González-García et al., 2015; Huss et al., 2021; Molineux, 2001; Qimron et al., 2006). The pooled library was selected on each host for an estimated two replication cycles, and then each variant was given a functional score, F, based on the ratio of their relative abundance before and after selection (Figure 2A). The functional score of each variant was then normalized to wildtype to yield F_N, where wildtype $F_N = 1$ (see Methods). Selection on each host gave good correlation across biological triplicates (Figure S2).

We anticipated strong selection pressure in our library because each motif was curated based on the ability of individual substitutions to significantly impact phage activity. Indeed, we found that only ~6% (607 of 10073) variants were detectable after selection on any one host, indicating substantial selective pressure. To establish if active variants derive motifs from phages more closely related to T7 phage we examined the annotated NCBI portion of our dataset. Surprisingly variants that sourced motifs from Podoviridea, the family of T7 phage, had the least activity with only 6.2% (47 of 754) of non-unique

variants demonstrating activity on any host, closely followed by Ackermannviridae with 11.6% (16 of 138) of non-unique variants demonstrating activity (Figure 2B). Variants with motifs sourced from Autographiviridae, Demerecviridae and Drexlerviridae made up a small portion of the library but were highly active, with 73.8% (31 of 42), 64.3% (9 of 14), and 73.3% (11 of 15) of non-unique variants active on any host. These results indicate those motifs from closely related phages were less tolerated than motifs from more distantly related phages. Mining deeper in metagenomic space from distantly related phages had thus been essential for creating active phage variants.

We then compared where in the tip domain motifs had been selected for to determine if there were trends in positional preference for active variants. In the pre-selection library motifs were located throughout the tip domain (figure 2C). Most motifs were located near exterior loops or proximal β-sheets, areas we previously identified as functionally relevant (Huss et al., 2021). A clear majority of motifs were substituted in and around the DE exterior loop and E β-sheet. These positions were known to be relevant for function and tolerant of individual substitutions based on our previous DMS results. After selection the distribution of motifs shifted, and the majority of motifs in active phage variants were instead located in and around the BC exterior loop and C β-sheet (Figure 2D). We hypothesize this region was generally more tolerant of motifs because the T7 RBD is a trimer and this region faces outward away from possible interactions with other monomers (Garcia-Doval and Raaij, 2012). As such this region may allow for multiple substitutions without impacting overall protein stability or secondary structure. These results indicate broad positional preferences for motif substitution across the tip domain.

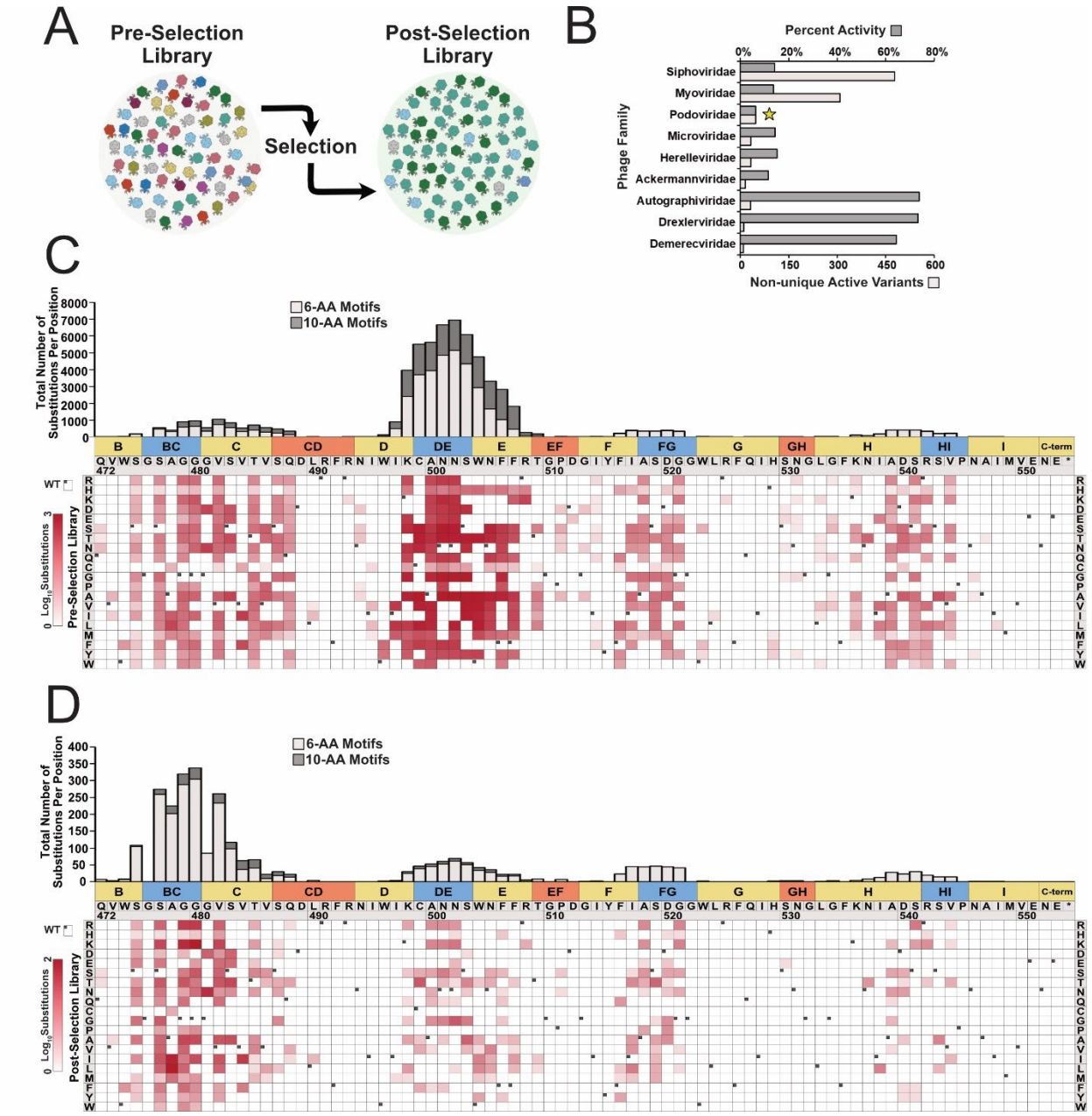

A
Pre-Selection Library
Selection
Post-Selection Library
B
Percent Activity
0%
20%
40%
60%
80%
Phage Family
Siphoviridae
Myoviridae
Podoviridae
Microviridae
Herelleviridae
Ackermannviridae
Autographiviridae
Drexlerviridae
Demerecviridae
0
150
300
450
600
Non-unique Active Variants
C
Total Number of Substitutions Per Position
6-AA Motifs
10-AA Motifs
B
BC
C
CD
D
DE
E
EF
F
FG
G
GH
H
HI
I
C-term
WT
Pre-Selection Library
Log10Substitutions
D
Post-Selection Library

Figure 2. Library Selection Reveals Positional Preference for Substitutions

(A) Library variants are scored by comparing variant abundance pre- and post-selection on different hosts. **(B)** Number of non-unique variants active on any host (bottom, light grey) and percent of active non-unique variants compared to the pre-selection library (top, dark grey) sourced from different phage families derived from proteins annotated in NCBI. The family for T7 bacteriophage (Podoviridea) is marked with a yellow star. **(C-D)** Heat map showing total count **(C)** pre-selection and **(D)** post-selection of specific substitutions listed top to bottom (log_{10}, red gradient) with substituted residue number (based on PDB 4A0T) and secondary structure shown left to right. Wildtype amino acids shown blank with black dot upper left. Total number of substitutions are shown top at every position as stacked bar graph with substitutions in 10-AA motifs shown dark grey and 6-AA motifs shown in light grey.

Hierarchical Clustering Reveals Patterns in Motif Selection

We next sough to identify common features amongst the phage variants that were responsible for activity on different hosts. We grouped phage variants based on activity on each host using agglomerative hierarchical clustering, plotting the $\log_2 F_N$ score of each variant for the five bacterial hosts (Figure 3A). Spearman's correlation and ward's minimum variance method for linkage was used for clustering because these approaches are less sensitive to outliers and allow for correlation-based distances as a dissimilarity measure. The cophentic correlation of this approach was 0.91, indicating clustering was valid, and variants were partitioned into 21 clusters based on gap-statistic analysis. Hierarchical clustering showed remarkably distinct clusters among the 607 variants active on any host with significant alterations to host range for many variants.

Examining clustering showed the degree to which host specificity could be tailored using metagenomic motifs. Some clusters of variants show some degree on all strains while other clusters were highly specific to one or more strains but not others. The largest cluster (cluster 8) was defined by variants that had high F_N on insensitive strains, moderate activity on *E. coli* 10G and BL21, and poor performance on *E. coli* BW25113. These results largely align with our DMS assay, which saw a similar activity profile for many single substitutions with BW25113 as an outlier. However, other clusters, like cluster 4, retained activity on all strains but saw less activity on BL21 and more activity on BW25113, contrasting our results with single substitutions. Some clusters showed remarkably narrow host specificity, including cluster 2, 9, 18, and 20, which all had variants with activity on only a single strain. Host range could be seen to expand to two strains in clusters 11 or 19 or three strains in clusters 10 or 16. While some clusters have

a small number of variants overall these results indicated that tailored host specificity was achievable using metagenomic motifs, far exceeded our results with single substitutions.

We then examined where substitutions were incorporated for each cluster to explore how the position of substitutions across the tip domain drove activity (Figure 3B). Given that the majority of tolerated substitutions were near the BC loop and C β-strand we anticipated that all clusters would contain variants with substitutions in this region. Surprisingly, while many clusters had a substantial portion of variants with substitutions in this region, several clusters aggregated variants with substitutions at other positions. For example, we investigated what differences caused the shift in preference from BL21 to BW25113 when examining clusters 8 and cluster 4. We noted that variants in cluster 8 almost completely eschew substitution between positions 498 and 503 near the DE exterior loop with only 1.5% of variants (4 of 266) having a substitution at these positions. In contrast, cluster 4 has 45% (15 of 33) variants with a substitution in this region and avoids variants with substitutions in the early BC exterior loop commonly seen in cluster 8. Other clusters show similar trends, such as cluster 10 and 11. Cluster 11, which contains variants active only on BL21 and BW25113, has 31.3% (5 of 16) of variants with substitutions near the DE exterior loop. In contrast, cluster 10, which has variants active on BL21, BW2513 and adds activity to 10G, has no variants with substitutions in this area. These results indicate the position of the substituted motif can substantially contribute to functional differentiation and influence phage activity.

We further explored drivers of functional differentiation by examining the substitution profile at various positions for different clusters. For example, the majority of substitutions for cluster 8 (223 of 266, 84%) and cluster 18 (30 of 36, 83%) had

substitutions between positions 478 and 483 and substitution frequency at many positions was similar (Figure 3C). However, these clusters show dramatic functional differentiation with variants in cluster 18 only active on BL21. Closer examination of substitution preferences shows that while both clusters favored substitutions with large, positively charged arginine and lysine substitutions, particularly at position 479 and 480, variants in cluster 18 typically also had a negatively charged aspartic or glutamic acid between positions 479 and 483 or a serine substitution at position 480. These particular substitutions were quite rare in cluster 8, indicating they may drive functional differentiation when combined. Many of these substitutions are not well tolerated individually but the unique combination enabled specificity for only BL21 for variants in cluster 18.

Other regions provided similarly intriguing comparisons. We examined substitutions between positions 498 and 503 for variants in cluster 9, which showed activity only on *E. coli* BW25113, and cluster 4 to investigate what causes specificity for BW25113 (Figure 3D). Glycine substitutions are tolerated at positions 500 and 501 only in cluster 9, while cluster 4 prefers larger positively charged arginine and lysine throughout this region and tolerates glycine substitutions only at positions 502 and 503. We hypothesize the preference for glycine at different positions changes the relative conformation of the tip domain, preventing adequate receptor engagement for all strains besides BW25113.

Comparison between cluster 8 and cluster 9 between positions 516 and 521 reveals further details for substitution preference (Figure 3E). A total of 8.6% of variants in cluster 8 (23 of 266) had substitutions in this region, compared to cluster 9 with 10.6%

(9 of 85) variants. Our DMS screen showed that this region was largely intolerant to substitution for all strains besides BW25113, but key substitutions at G521 or A518 were tolerated and enabled function on insensitive strains and improved function on 10G and BL21. Motifs in this region in cluster 8 are dominated by lysine and arginine substitutions at either G521 and A518, recapitulating our DMS results, and then contain many other substitutions that were not tolerated individually, notably proline substitutions which we would expect to change the local structure of the loop. This contrasts cluster 9, which contains no lysine or arginine substitutions at G521 and lacks abundant proline substitutions in the region. This indicates the presence of the enabling substitution can allow substantial changes to the local architecture of the loop, expanding allowable substitutions and influencing host range.

Overall, several key lessons emerged from this analysis. First, Motif Curation was able to create variants with unique combinations of substitutions that drove distinct patterns of host specificity. Many variants have substitutions that were individually poorly tolerated but when combined provide a novel path for altering phage activity. Second, the location of substitutions was important for determining host specificity, with the area near the BC exterior loop and C β-strand overall more tolerant of substitution. Finally, comparing the identity of specific amino acids substitutions across clusters reveals novel combinations of substitutions that drive host specificity and allow for substantial changes to the local architecture of the tip domain.

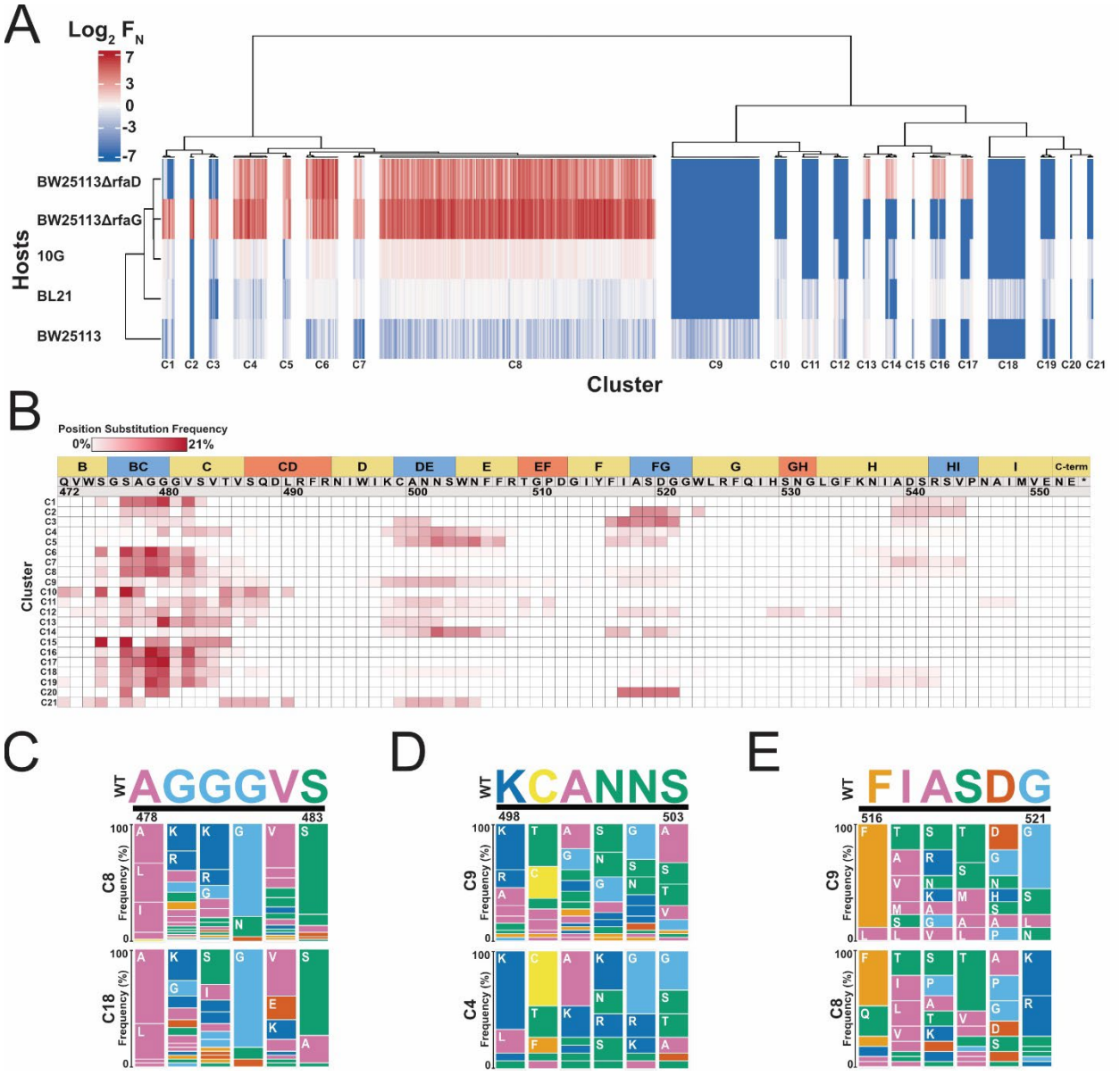


Figure 3. Hierarchical clustering reveals patterns in motif selection

(A) Hierarchical clustering of normalized functional scores ($\log_2 F_N$, blue to red gradient, wildtype $F_N = 0$) for active phage variants on five *E. coli* hosts (clustered left). F_N clustered is the average of three biological replicates. **(B)** Heat map showing substitution frequency (red gradient) at every position in the tip domain for each cluster. Clusters are organized top to bottom, while position, wildtype amino acid and secondary structure topology is

shown left to right. **(C-E)** Frequency plots comparing substitution frequency for (C) clusters 8 and 18 at positions 478 through 483, (D) clusters 9 and 4 at positions 498 through 503, and (E) clusters 9 and 8 at positions 516 through 521. The wildtype sequence and position is shown top and amino acids are colored based on similarity of physicochemical properties.

Motif curation reveals unique epistatic interactions in the tip domain

We next investigated if our motif library could reveal specific epistatic interactions in the tip domain. Epistasis is a phenomenon that occurs when multiple mutations combined together alter a protein's biochemical function differently than the sum of the individual mutations. Specific epistasis is caused by physical interactions between mutations that drive differences in protein function. Epistatic interactions are well studied and understanding how local networks of residues interact using epistatic data provides insights into the sequence-function landscape of the protein (Dickinson et al., 2013; Domingo et al., 2019; Miton et al., 2021; Nishikawa et al., 2021). We examined variant activity on insensitive strains BW25113Δ*rfaD* and BW25113Δ*rfaG* for epistasis. These insensitive strains were chosen because selection on these strains was more significant, and we specifically looked for variants where every constituent substitution was individually deleterious but when combined caused an increase in activity. We found 40 variants for BW25113Δ*rfaD* and 27 variants for BW25113Δ*rfaG* where the variant F_N was much larger than the sum of the F_N of individual substitutions, indicating epistasis was occurring (Figure 4A and 4B, see Supplementary File 2). In many cases epistasis was dramatic, with most individual substitutions having no activity on the host while the combination of substitutions considerably outperformed wild type.

We examined these epistatic variants to evaluate what interactions were dictating the change in activity when substitutions were combined. For both BW25113Δ*rfaD* and BW25113Δ*rfaG* most epistatic variants had substitutions between positions A478 and S483 in the BC exterior loop and C β-sheet (26 variants BW25113Δ*rfaD* and 16 for BW25113Δ*rfaG*) or between positions N502 and F507 in the DE exterior loop and E β-

sheet (11 variants BW25113Δ*rfaD* and 6 for BW25113Δ*rfaG*). Examining substitution frequency between positions A478 and S483 revealed different residue preferences that drove activity (Figure 4C). This region faces outward on the tip domain and has fewer probable interactions with other monomers (Supplementary Figure 3A and 3B). Notable substitutions in epistatic variants include large, hydrophilic lysine, histidine or arginine substitutions at either G479 for BW25113Δ*rfaD* or V482 for BW25113Δ*rfaG*. Our DMS results had shown that these substitutions increased activity in nearby positions, especially at G480, but were not tolerated themselves (Huss et al., 2021). We predict these large, hydrophilic substitutions remained responsible for the gain in activity on these hosts and other substitutions changed the local structure of the tip domain to enable a key binding event with these substitutions.

Epistatic motifs with substitutions between positions N502 and F507 showed a similar substitution pattern for both hosts which substantially altered the local region (Figure 4D). This region lies on the opposite side of the tip domain, faces inward, and likely has many interactions with other monomers (Supplementary Figure 3A and 3C). Epistatic motifs primarily had lysine and arginine substitutions at position N502 and replaced polar, uncharged residues or larger hydrophobic residues with small hydrophobic residues. Our DMS screen had shown that large hydrophobic residues recovered function at A500 and N501 but were not tolerated at N502. We predict that the smaller hydrophobic substitutions stabilize interactions with the other monomers, allowing the large hydrophilic substitutions to successfully interact with the host receptor.

Overall, epistatic variants demonstrated patterns of substitution that emphasized the importance of large hydrophilic substitutions to activity on either insensitive strain.

Other substitutions likely change the local architecture of the monomer or affect how each monomer interacts in the trimer to facilitate a key interaction with the large hydrophilic substitution. These results paint a broader picture of the malleability of the tip domain and the flexibility of the sequence-function landscape to facilitate receptor interaction.

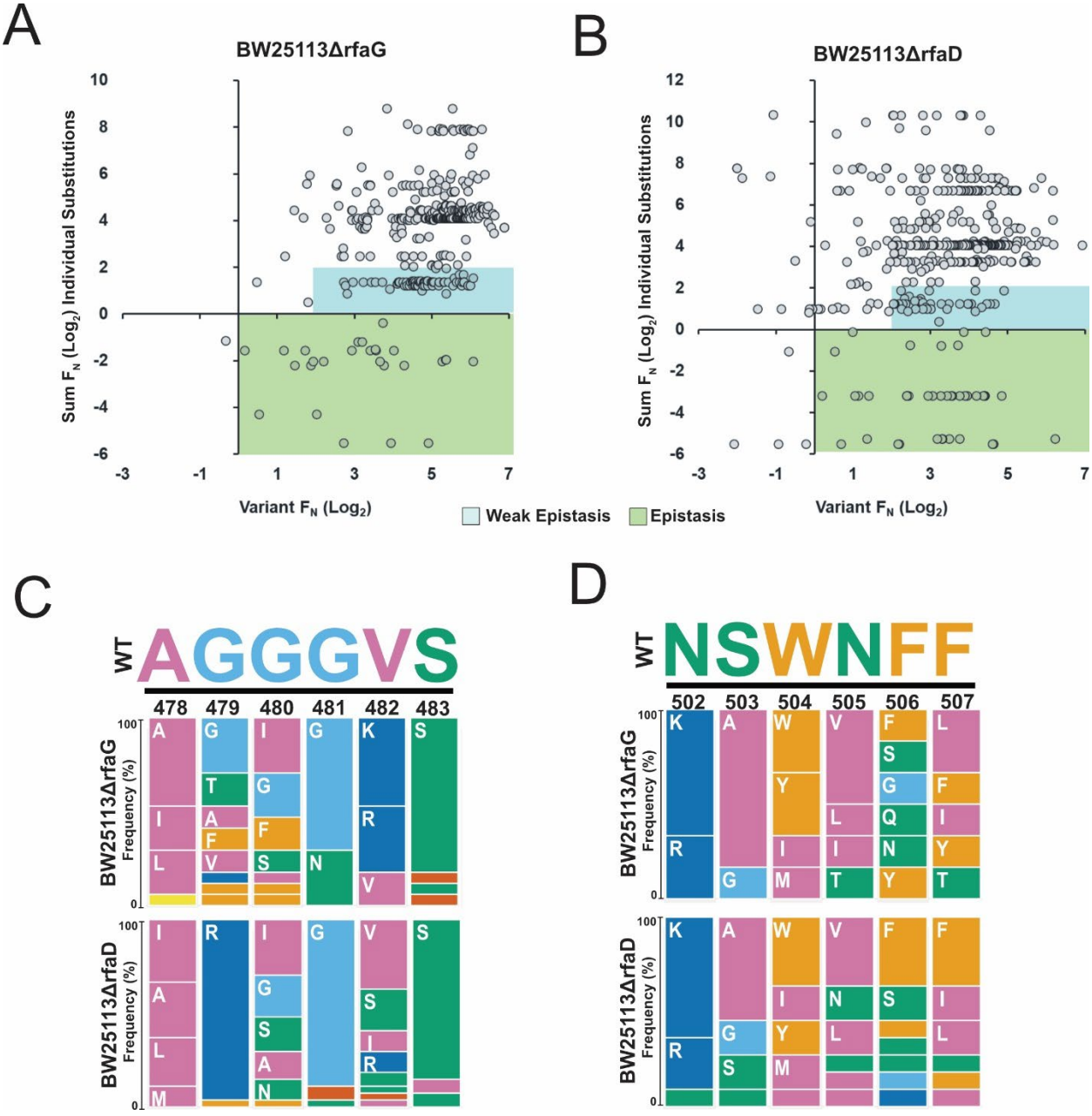


Figure 4. Activity on insensitive variants reveals epistatic interactions

(A-B) Comparison of variant activity (log_2 F_N) to the sum of activity of individual substitutions that comprised the variant motif on (A) BW25113Δ*rfaG* and (B) BW25113Δ*rfaD*. Shaded areas represent variants showing signs of weak epistasis (blue) or strong epistasis (green), indicating substitutions in that motif show greater activity when combined than individually. Points are shown only if variant activity was greater than the limit of detection. **(C-D)** Frequency plots comparing substitution frequency for BW25113Δ*rfaG* and BW25113Δ*rfaD* at (C) positions 478 through 483 and (D) at positions 502 through 507 for variants showing strong epistasis. Wildtype sequence and position is shown top and amino acids are colored based on similarity of physicochemical properties.

Curated motifs enhance activity on E. coli causing UTIs

Because our library had been curated to find motifs that impacted overall phage activity, we expected that the library could improve activity on strains that were not included in the original host panel used for DMS. We evaluated this hypothesis by passaging the library on two insensitive strains of *E. coli* that cause urinary tract infections, here labeled UTI33 and UTI46 (Arthur et al., 1990). Wild type T7 phage has a limited ability to plaque on both strains and plaques are much smaller than on susceptible strains. As before the pooled library was selected on each host in biological triplicates for an estimated two replication cycles and each variant was given a F_N score. Selection on each host was greater than the original host panel but retained good correlation across replicates (Figure S4A-B). Variant F_N scores trended the same for both strains with notable variants that significantly outperformed wild type during passage (Figure 5A). Substitution and location preference was also similar for both strains, heavily favoring the BC loop and C β-strand region (Figure S4C). These results showed that our approach was able to find motifs that influence activity outside of the DMS panel.

To validate our analysis, we isolated two variants that performed exceedingly well in the pooled selection to determine if they could overcome the plaquing deficiency seen with wildtype T7. The first variant contained the motif FGALKG, replacing the wildtype region SGSAGG from positions S475 to G480 (UTI33 $\log_2 F_N = 5.1 \pm 0.1$, UTI46 $\log_2 F_N = 5.5 \pm 0.1$). The second variant contained the motif WCKENK, replacing the wildtype region SAGGGV from positions S477 through V482 (UTI33 $\log_2 F_N = 3.6 \pm 0.3$, UTI46 $\log_2 F_N = 3.7 \pm 0.1$). Wildtype T7 has a 2-log total reduction in plaque efficiency on UTI33 and a 4-log total plaque deficiency on UTI46 (Figure 5B). Wildtype plaques are otherwise hazy

and small, with a diameter of ~1-2 mm compared to a 5-6 mm diameter on susceptible strains. The T7 variant with the motif FGALKG restored plaque morphology to wild type but retained a ~1 log deficiency on both strains, while the variant with the motif WCKENK completely restored plaque efficiency and plaque morphology on both strains (Figure 5B). This experiment demonstrated that Motif Curation can be used to create variants that restore activity in clinically relevant strains, indicating our approach was successfully identifying functional hotspots in the tip domain and finding motifs that are relevant to phage activity in diverse bacterial hosts.

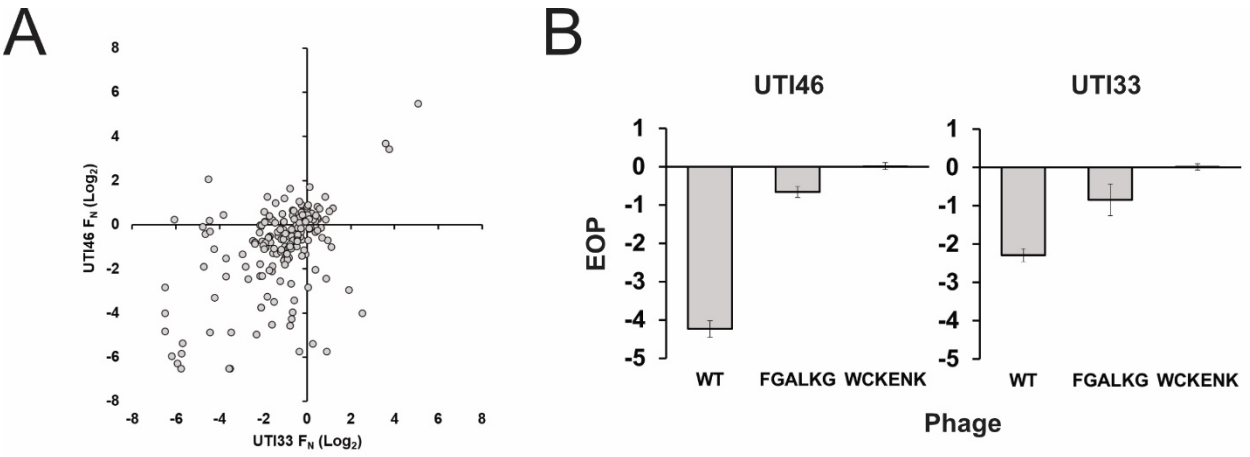


Figure 5. Curated motifs enhance activity on E. coli causing UTIs

(A) Comparison of variant activity (log_2 F_N) on *E. coli* UTI46 and *E. coli* UTI33. **(B)** Efficiency of Plating (mean ± SD, biological triplicates) on *E. coli* UTI46 and *E. coli* UTI33 for wildtype T7 (WT), T7 variant with FGALKG replacing the wildtype region SGSAGG, and T7 variant contained the motif WCKENK replacing the wildtype region SAGGGV. *E. coli* 10G with a helper plasmid constitutively producing wildtype receptor binding protein is used as a reference host. Wildtype plaques are atypically small and hazy.

Discussion

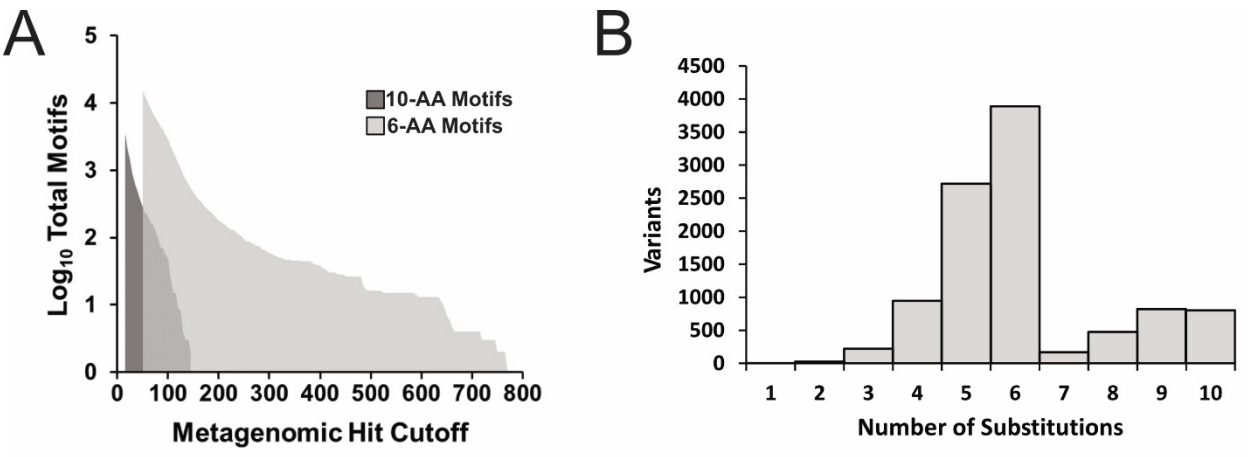
In this study we devised Motif Curation as a new approach to find metagenomic motifs that influence phage activity. Using Motif Curation, we found more than 10^4 complex motifs for the T7 RBP. These motifs were significantly different from the wildtype sequence and were found in diverse and distantly related metagenomic sequences. By screening phage variants with these motifs, we revealed hundreds of phages with novel host specificity and broke down how motifs drove changes in the sequence-function landscape across tip domain.

Several key insights emerged from our results. First, cross-comparison of results from our previous DMS assay was successful in seeding scores to find motifs that broadly influenced phage activity and altered host specificity (Huss et al., 2021). This approach was able to generate phage variants that improved activity on UTI strains that had not been included in the host panel used for the DMS assay and created variants that had novel host specificity when compared to single substitutions. Second, changes in phage activity and host range were dependent on curating motifs from distantly related phages. Motifs from Podoviridea, the family of T7 phage, performed notably poorly on the host panel while more motifs sourced from more distantly related phages drove activity and host specificity. This emphasized the importance of being able to reach deeply into metagenomic space into non-homologous phages. Third, both the position and composition of motifs were important drivers of phage activity in the tip domain of the T7 RBP. The BC exterior loop and C β-strand were overall more tolerant of substitution while motifs inserted at other positions could drive host specificity. Novel combinations of substitutions appear to allow for substantial changes to the local architecture of the tip domain and revealed epistatic interactions that enable key residues to contact the host

receptor. Altogether these results improve our understanding of the sequence-function landscape of the T7 RBP and identify motifs that can drive highly specific changes in host activity.

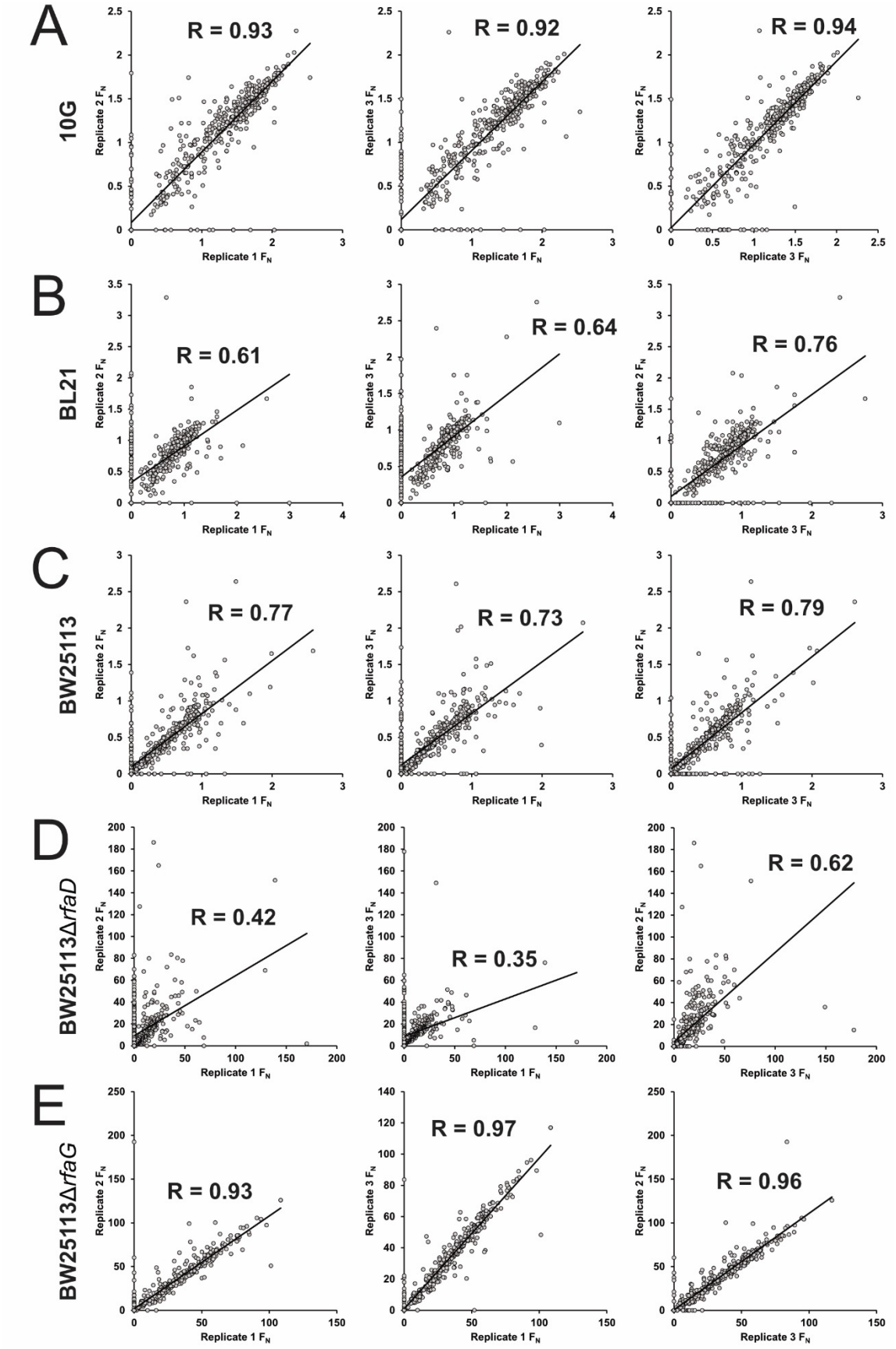
Vast troves of phage metagenomic data are becoming increasingly available from every conceivable biome. However, the enormous breadth of phage diversity makes querying these databases for relevant sequences a challenging prospect. Motif Curation is designed as a process that can tap into these complex datasets to identify the most relevant motifs that drive phage activity and be used to engineer phages. This process is not limited to the RBP but can be used on any region with suitable DMS results to seed and score potential motifs, including other structural proteins, lysins or holins. By leveraging these metagenomic motifs we envision Motif Curation as a platform to better understand sequence-function relationships and engineer phages.

Supplementary Figures



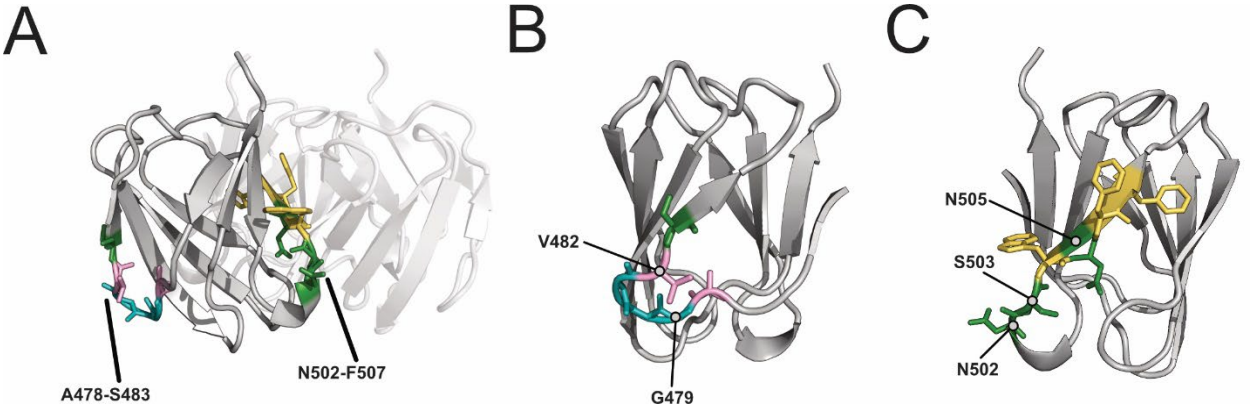
Supplementary Figure 1. Total curated motifs based on metagenomic hit cutoff and number of substitutions in each variant in the final phage library

(A) Total number of curated motifs (log_{10}) found using different metagenomic hit cutoffs for 10-AA motifs (dark grey) and 6-AA motifs (light grey). The lowest cutoff indicated is 15 hits for 10-AA motifs and 50 hits for 6-AA motifs. **(B)** Count of substitutions for each phage variant in the final motif phage library.


A
10G
R = 0.93
R = 0.92
R = 0.94
B
BL21
R = 0.61
R = 0.64
R = 0.76
C
BW25113
R = 0.77
R = 0.73
R = 0.79
D
BW25113ΔrfaD
R = 0.42
R = 0.35
R = 0.62
E
BW25113ΔrfaG
R = 0.93
R = 0.97
R = 0.96
Replicate 1 F_N
Replicate 2 F_N
Replicate 3 F_N

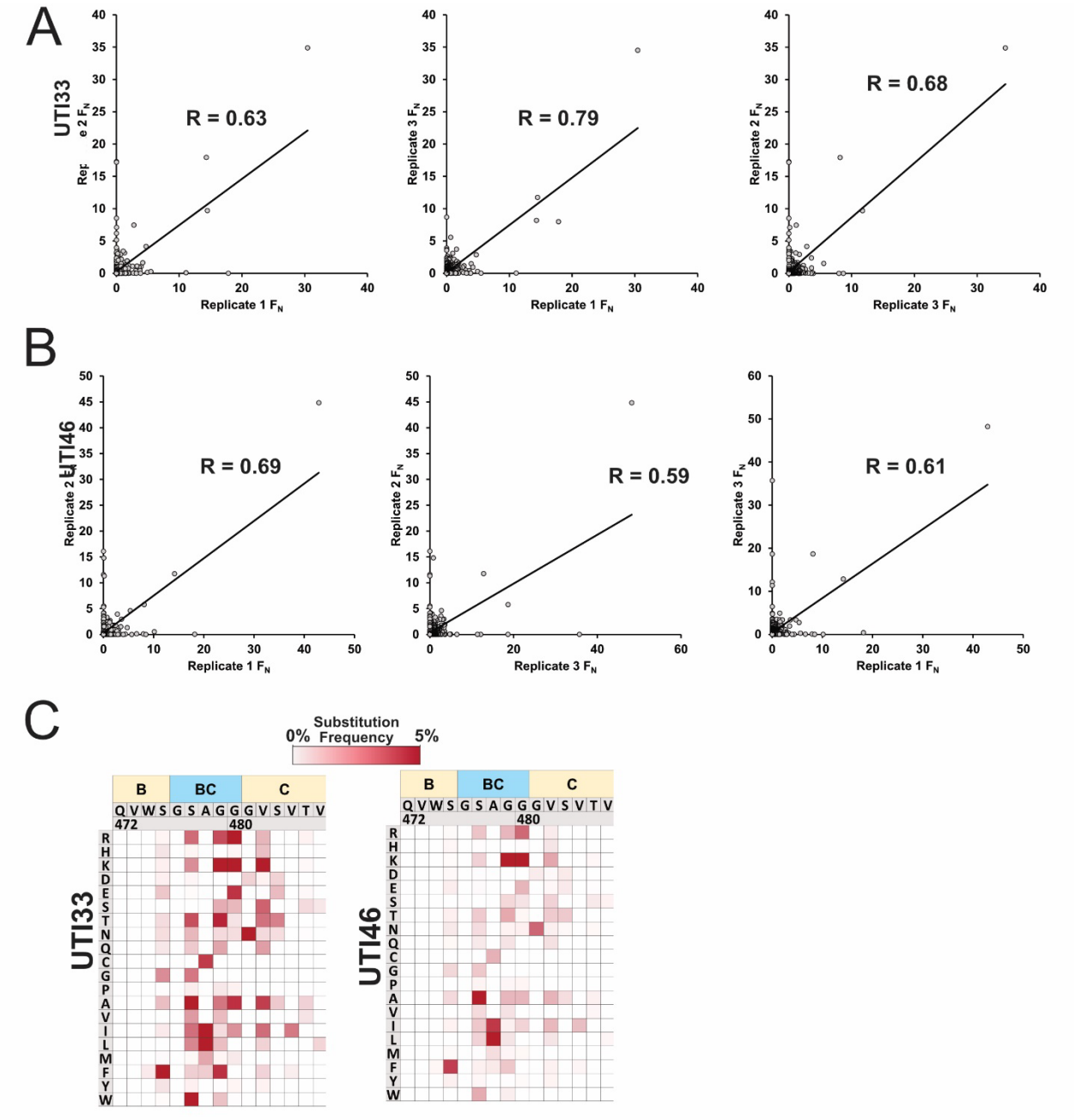
Supplementary Figure 2. Correlation between biological replicates after selection of phage variant library on the E. coli host panel

(A-E) Correlation of F_N scores between biological replicates of the phage variant library on (A) *E. coli* 10G, (B) BL21, (C) BW25113, (D) BW25113Δ*rfaD* and (E) BW25113Δ*rfaG*. R values and trendlines are displayed for all replicates as a black line.



Supplementary Figure 3. Position of epistatic motifs in the tip domain

(A) Crystal structure of the tip domain with regions A478-S483 and N502-F507 in color on one monomer with the other monomers of the receptor binding protein shown transparent. Amino acids are colored based on similarity of physicochemical properties. **(B-C)** Rotated views of regions (B) A478-S483 and (C) N502-F507.


A
UTI33
R = 0.63
Replicate 1 F_N
R = 0.79
Replicate 3 F_N
Replicate 1 F_N
R = 0.68
Replicate 2 F_N
Replicate 3 F_N
B
UTI46
R = 0.69
Replicate 2 F_N
Replicate 1 F_N
R = 0.59
Replicate 2 F_N
Replicate 3 F_N
R = 0.61
Replicate 3 F_N
Replicate 1 F_N
C
Substitution Frequency
0%
5%
B
BC
C
Q V W S G S A G G G V S V T V
472
480
UTI33
UTI46
R H K D E S T N Q C G P A V I L M F Y W

Supplementary Figure 4. Correlation between biological replicates after selection of phage variant library on E. coli UTI strains and variant substitution preference

(A-B) Correlation of F_N scores between biological replicates of the phage variant library on (A) *E. coli* UTI33 and (B) UTI46. R values and trendlines are displayed for all replicates as a black line. **(C)** Substitution frequency (red gradient, total percentage of substitutions) for active variants on *E. coli* UTI33 (left) and UTI46 (right). Specific substitutions are listed top to bottom with substituted residue number (based on PDB 4A0T) and secondary structure shown left to right. Wildtype amino acids shown blank with black dot upper left.

List of Supplementary Files

Supplementary File 1. Seed scores for each substitution in the tip domain and randomized seed scores used to validate Motif Curation.

Supplementary File 2. List of phage variants and the wild type region the motif replaced, the motif sequence and associated substitutions.

Supplementary File 3. Seed scores for each substitution in the tip domain and randomized seed scores used to validate Motif Curation.

Supplementary File 4. List of Primers Used for Experiment

Methods

Essential Resources Table				
Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
strain, strain background (*Escherichia coli*)	*E. coli* 10G	Lucigen	Lucigen:60107-1	
strain, strain background (*Escherichia coli*)	*E. coli* BL21	ATCC	ATCC:BAA-1025	
strain, strain background (*Escherichia coli*)	*E. coli* BW25113	(Baba et al., 2006)	BW25113	
strain, strain background (*Escherichia coli*)	*E. coli* BW25113 Δ*rfaG*	(Baba et al., 2006)	BW25113Δ*rfaG*	
strain, strain background (*Escherichia coli*)	*E. coli* BW25113 Δ*rfaD*	(Baba et al., 2006)	BW25113 Δ*rfaD*	
strain, strain background (*Escherichia coli*)	*E. coli* UTI33	(Arthur et al., 1990)	UTI33	
strain, strain background (*Escherichia coli*)	*E. coli* UTI46	(Arthur et al., 1990)	UTI46	

strain, strain background (T7 Bacteriophage)	T7 Bacteriophage	ATCC	ATCC:BAA-1025-B2	
strain, strain background (T7 Bacteriophage)	T7 receptor binding protein (gp17) Accepter phage	(Huss et al., 2021)	T7 Acceptor Phage	
strain, strain background (T7 Bacteriophage)	T7 Bacteriophage motif variants	This paper	Available on request	Variant with metagenomic motifs, available from the Raman lab
commercial assay or kit	KAPA HiFi PCR Kit	Roche	Roche:KK2101	
commercial assay or kit	Golden Gate Assembly Kit (BsaI-HFv2)	NEB	NEB:E1601L	
recombinant DNA reagent	pHT7Helper1 (plasmid)	This paper		Helper with T7 gp17. See methods for full details.
recombinant DNA reagent	pHRec2 and derivatives (plasmids)	This paper		Recombination plasmid. See methods for details
recombinant DNA reagent	pHCas9-3	(Huss et al., 2021)		Plasmid with Cas9 targeting acceptor phage. See methods for full details.

software, algorithm	R scripts for hierarchical clustering	This paper	N/A	Available here https://github.com/PhilHuss/MotifMining
software, algorithm	R scripts for Motif Curation	This paper	N/A	Available here https://github.com/PhilHuss/MotifMining

Microbes and Culture Conditions

T7 bacteriophage was obtained from ATCC (ATCC® BAA-1025-B2). T7 acceptor phages used in ORACLE were previously generated in our lab (Huss et al., 2021). *Escherichia coli* BL21 is a lab stock, *E. coli* 10G is a highly competent DH10B derivative (Durfee et al., 2008) originally obtained from Lucigen (60107-1). *E. coli* BW25113, BW25113Δ*rfaD* and BW25113Δ*rfaG* were obtained from Doug Weibel (University of Wisconsin, Madison) and are derived from the Keio collection (Baba et al., 2006). UTI33 and UTI46 were obtained from Rod Welch (University of Wisconsin, Madison) and originates from a UTI collection (Arthur et al., 1990).

All bacterial hosts are grown in and plated on Lb media (1% Tryptone, 0.5% Yeast Extract, 1% NaCl in dH2O, plates additionally contain 1.5% agar, while top agar contains only 0.5% agar) and Lb media was used for all experimentation and was used to recover host and phages after transformation. Kanamycin (50 µg/ml final concentration, marker for pHT7Helper1) and spectinomycin (115 µg/ml final concentration, marker for pHRec2, pHRec1-MLib and pHCas9 and derivatives) was added as needed. All incubations of bacterial cultures were performed at 37°C, with liquid cultures shaking at 200-250 rpm unless otherwise specified. Bacterial hosts were streaked on appropriate Lb plates and stored at 4°C.

T7 bacteriophage was propagated using *E. coli* BL21 after initial receipt from ATCC and then as described on various hosts in methods. All phage experiments were performing using Lb and culture conditions as described for bacterial hosts. Phages were stored in Lb at 4°C.

For long term storage all microbes were stored as liquid samples at -80°C in 10% glycerol, 90% relevant media.

SOC (2% tryptone, 0.5% yeast extract, 0.2% 5M NaCl, 0.25% 1M KCL, 1% 1M $MgCl_2$, 1% 1M $MgSO_4$, 2% 1M glucose in dH2O) was used to make competent cells.

General Cloning Methods

PCR was performed using KAPA HiFi (Roche KK2101) for all experiments. Golden Gate assembly was performed using New England Biosciences (NEB) Golden Gate Assembly Kit (BsaI-HFv2, E1601L). Restriction enzymes were purchased from NEB. DNA purification was performed using EZNA Cycle Pure Kits (Omega Bio-tek D6492-01) using the centrifugation protocol. Gibson assembly was performed according to the Gibson Assembly Protocol (NEB E5510) but Gibson Assembly Master Mix was made in lab (final concentration 100 mM Tris-HCL pH 7.5, 20 mM $MgCl_2$, 0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 10 mM dTT, 5% PEG-8000, 1 mM NAD^+, 4 U/ml T5 exonuclease, 4 U/µl Taq DNA Ligase, 25 U/mL Phusion polymerase). All cloning was performed according to manufacturer documentation except where noted in methods.

PCR reactions use 1 µl of ~1 ng/µl plasmid or ~0.1 ng/µl DNA fragment as template for relevant reactions. PCR reactions using phage as template use 1 µl of undiluted phage

stock (genomic extraction was unnecessary) and have an extended 5 minute 95°C denaturation step.

DpnI digest was performed on all PCR that used plasmid as template. Digestion was performed directly on PCR product immediately before purification by combining 1.5 µl DpnI (30 units), 15 µl 10x CutSmart Buffer, 98 ul PCR product, and 35.5 ul dH2O, incubating at 37°C for 2 hours and 30 minutes then heat inactivating at 80C for 20 minutes.

Electroporation of plasmids was performed using a Bio-rad MicroPulser (165-2100), Ec2 setting (2 mm cuvette, 2.5 kV, 1 pulse) using 50 µl competent cells and 100 ng of plasmid or 20 ul of golden gate reaction for transformation. Electroporated cells were immediately recovered with 950 µl Lb, then incubated at 37°C for 1 to 1.5 hours and plated or grown in relevant media.

E. coli 10G competent cells were made by adding 8 mL overnight 10G cells to 192 mL SOC (with antibiotics as necessary) and incubating at 21°C and 200 rpm until ~OD_{600} of 0.4 as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation (actual incubation time varies based on antibiotic, typically overnight). Cells are centrifuged at 4°C, 1000g for 20 minutes, the supernatant is discarded, and cells are resuspended in 50 mL 10% glycerol. Centrifugation and washing are repeated three times, then cells are centrifuged at 2000g for 20 minutes then resuspended in a final volume of ~1 mL 10% glycerol and are aliquoted and stored at -80°C.

Detailed protocols for cloning are available on request. All primers used in experiments in this publication are listed in supplementary file 4.

Plasmid Cloning and Descriptions

pHT7Helper1 is used during optimized recombination and during accumulation in ORACLE to prevent library bias and depletion of variants that grow poorly on *E. coli* 10G. There are no changes to this plasmid from its initial construction as documented in our previous DMS assay, plasmid details are provided here for completeness (Huss et al., 2021). This plasmid contains a pBR backbone, kanamycin resistance cassette, mCherry, and the T7 receptor binding protein *gp17*. Both mCherry and *gp17* are under constitutive expression. *Gp17* was combined with promoter apFAB47 (Kosuri et al., 2013) using SOE and the plasmid assembled by Gibson assembly. There is a single nucleotide deletion in the promoter that has no effect on plaque recovery for phages that require *gp17* to plaque.

pHRec2 is used as template to generate the variant library containing metagenomic motifs, referred to as pHRec2-Motif and used during optimized recombination in ORACLE. This plasmid has been updated since its initial construction in our previous DMS assay, where it was called pHRec1. As before this plasmid contains an SC101 backbone, Cre recombinase, a spectinomycin resistance cassette, and the T7 tail fiber *gp17* flanked by Cre lox66 sites with an m2 spacer, a 3' pad region and lox71 sites with a wt spacer (Langer et al., 2002). Cre recombinase is under constitutive expression. This plasmid was altered using sequential PCR and Gibson assembly to include a larger constant region and random, 20 bp barcode between the end of *gp17* and the Cre-lox sites. The total region inserted was 105 bp. This barcode was added to facilitate smaller, rather less expensive NGS reads but was ultimately not used in this experiment. Note insertion of the barcode requires a 3 or 4 part golden gate reaction instead of a 2 part reaction as with pHRec1.

pHCas9-3 is used during Accumulation in ORACLE. There are no changes to this plasmid from its initial construction as documented in our previous DMS assay, plasmid details are provided here for completeness (Huss et al., 2021). This plasmid contains an SC101 backbone, a spectinomycin resistance cassette and cas9 cassette with a previously cloned gRNA targeting the T7 acceptor phage (Jiang et al., 2013). pHCas9 was created with Gibson assembly, while pHCas9-3 was assembled by phosphorylation and annealing gRNA oligos (100 uM forward and reverse oligo, 5 µl T4 Ligase buffer, 1 µl T4 PNK, to 50 µl dH2O, incubate at 37°C for 1 hour, 96C for 6 minutes then 0.1C/s temperature reduction to 23C), then Golden Gate cloning (1 µl annealed oligo, 75 ng pHCas9, 2 µl T4 DNA Ligase Buffer, 1 µl Golden Gate Enzyme Mix, dH20 to 20 ul. Incubation at 37°C for 1 hour then 60C for 5 minutes, followed by direct transformation of 1 ul, plated on Lb with spectinomycin). All plasmid backbones and gene fragments are lab stocks.

General Bacteria and Phage Methods

Bacterial concentrations were determined by serial dilution of bacterial culture (1:10 or 1:100 dilutions made to 1 mL in 1.5 microcentrifuge tubes in Lb) and subsequent plating and bead spreading of 100 µl of a countable dilution (targeting 50 colony forming units) on Lb plates. Plates were incubated overnight and counted the next morning. Typically, two to three dilution series were performed for each host to initially establish concentration at different OD_{600} and subsequent concentrations were confirmed with a single dilution series for later experiments.

Stationary phase cultures are created by growing bacteria overnight (totaling ~20-30 hours of incubation) at 37°C. Cultures are briefly vortexed then used directly. Exponential phase culture consists of stationary culture diluted 1:20 in Lb then incubated at 37°C until an OD_{600} of ~0.4-0.8 is reached (as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation), typically taking 40 minutes to 1 hour and 20 minutes depending on the strain and antibiotic, after which cultures are briefly vortexed and used directly.

Phage lysate was purified by centrifuging phage lysate at 16g, then filtering supernatant through a 0.22 uM filter. Chloroform was not used.

To establish titer, phage samples were typically serially diluted (1:10 or 1:100 dilutions made to 1 mL in 1.5 microcentrifuge tubes) in Lb to a 10^{-8} dilution for preliminary titering by spot assay. Spot assays were performed by mixing 250 µl of relevant bacterial host in stationary phase with 3.5 mL of 0.5% top agar, briefly vortexing, then plating on Lb plates warmed to 37°C. After plates solidified (typically ~5 minutes), 1.5 µl of each dilution of phage sample was spotted in series on the plate. Plates were incubated and checked after overnight incubation (~20-30 hours) to establish a preliminary titer. After a preliminary titer was established, phage samples were serially diluted in triplicate for efficiency of plating (EOP) assays.

EOP assays were performed using whole plates instead of spot plates to avoid inaccurate interpretation of results due to spotting error (Khan Mirzaei and Nilsson, 2015). To perform the whole plate EOP assay, 400 µl of bacterial host in exponential phase was mixed with between 5 to 50 µl of phages from a relevant dilution targeting 50 plaque forming units (PFUs) after overnight incubation. The phage and host mixture was briefly

vortexed, briefly centrifuged, then added to 3.5 mL of 0.5% top agar, which was again briefly vortexed and immediately plated on Lb plates warmed to 37°C. After plates solidified (typically ~5 minutes), plates were inverted and incubated overnight. PFUs were typically counted after overnight incubation (~20-30 hours) and the total overnight PFU count used to establish titer of the phage sample. PFU totals between 10 and 300 PFU were typically considered acceptable, otherwise plating was repeated for the same dilution series. This was repeated at least in triplicate for each phage sample on each relevant host to establish phage titer.

EOP was determined using *E. coli* 10G with pHT7Helper1 as a reference host. EOP values were generated for each of the three dilutions by taking the phage titer on the test host divided by the phage titer on the reference host, and this value was subsequently $\log_{10}$ transformed. Values are reported as mean ± SD.

MOI was calculated by dividing phage titer by bacterial concentration. MOI for the T7 variant library after the variant gene is expressed was estimated by titering on 10G with pHT7Helper1.

Position Seed Scoring

To seed substitution scores the maximum F_N of each biological replicate for each substitution from susceptible strains 10G, BL21, and BW25113 in our DMS assay was compressed where if $F_N \leq 1$ compressed F_N (CF_N) = F_N and if $F_N > 1$ then $CF_N = (F_N - 1 / \max(\text{strain } F_N) - 1) + 1$. Scores were compressed to reduce the impact of variants that performed exceedingly well, as we wanted to weight scores higher for variants with

functionally different scores across strains (i.e. it matters less if a variant does very well on one strain and comparable to wildtype on other strains, it matters more if it does poorly on one strain and comparable to wildtype or better than wildtype on another, as that is more functionally distinct.)

If the difference between CF_N on all susceptible strains was less than 0.5, indicating the functional difference between strains was small, motif seed score MFF = CF_N, else MFF = CF_N + 0.5. This served to seed scores higher if there was a functional difference between susceptible strains. Finally max(⌊log(F_N⌋) of each biological replicate for each substitution from resistant strains BW25113Δ*rfaG* and BW25113Δ*rfaD* was calculated and added to MFF. This further weighed substitutions higher if they had recovered function on resistant strains. These final values were used to seed scores for every possible substitution to curate motifs.

Database construction

NCBI databases (release July 2019) were queried for the term “prokaryotic virus” and limited to sequences of at least 3 kb. The resulting sequences were manually curated by removing non-genomic sequences (e.g., containing the terms ‘open reading frame’ or ‘ORF’) and non-viral sequences. Additionally, the IMG/VR v2.1 (October 2018) database was downloaded and limited to sequences originating from human, wastewater, and animal environments according to IMG/VR metadata.

For each dataset, open reading frames were predicted using Prodigal v2.6.3 (-p meta) and proteins were dereplicated using CD-HIT v4.7 (-c 1). To obtain proteins of interest, HMM profiles for 373 viral hallmark proteins were selected from VOGDB v94

(vogdb.org) according to viral hallmark proteins designated by VIBRANT v1.2.1. The selected HMM profiles represented annotations containing the terms *tail, capsid, spike, baseplate, sheath, lysin,* and *holin*. Hmmsearch v3.1b (-e 1e-5) was used to query the database proteins to the selected HMM profiles. In total, 34682 and 26335 non-redundant proteins from NCBI and IMG/VR databases, respectively, were selected for final analysis. Source code is available at https://github.com/PhilHuss/MotifMining.

Motif search (methods for f-score table construction before this)

A custom motif search tool (*motif_finder_tool.py*) was used to query each of the protein databases separately for 6-mer motifs (-n 6 -c 1 -e 1e-50 -t 0.8) and 10-mer motifs (-n 10 -c 1 -e 1e-5 -t 0.045). The motif search tool scans n-mer windows of input database amino acids against n-mer windows of input matrix scores. The scores per amino acid window are compared to a set cutoff and predetermined constants. The maximum score per amino acid window across all matrix windows is summed to generate an overall score per protein. The overall score and e-value, calculated based on the number of windows searched, are compared to set thresholds. Any protein passing these thresholds is considered as a candidate protein that is similar to the motifs found within the original input matrix. All amino acid windows used in the generation of the overall score for that protein are extracted. To limit the results to a practical search space, 6-mer and 10-mer motifs were filtered to those that were extracted at least 50 and 15 times, respectively, by the motif search tool across all proteins. Source code is available at https://github.com/PhilHuss/MotifMining.

Motif Variant Plasmid Library Preparation

To create the metagenomic motif variant plasmid library, oligos were first designed and ordered from Agilent as a SurePrint Oligonucleotide Library (OLS 230mers, Barcode 33049511001 OLS, internal freezer stock #2). Every oligo contained a single motif that replaced the wild type region in the tip domain. We used the most frequently found codon for each amino acid in the *gp17* tail fiber to define the codon for each substitution. Oligos contained BsaI sites at each end to facilitate Golden Gate cloning. Two pools of oligos were used for this library. Oligo pools were amplified by PCR using 0.005 pmol total oligo pool as template (~0.5 ng total DNA) and 15 total cycles to prevent PCR bias, then pools were purified. pHRec2 was used as template in a PCR reaction to create two backbones for each of the three pools, which was then DpnI digested and purified. Backbones are not otherwise pretreated before Golden Gate Assembly. Golden gate assembly was performed using ~300 ng of relevant pool backbone and a 1x molar ratio for oligos (~20 ng) and a small ~100 bp fragment containing constant regions and a 20 bp randomized barcode (~15 ng), combined with 2 µl 10x T4 DNA ligase buffer, 1 µl NEB Golden Gate Enzyme mix and dH2O to 20 ul. For the 3' pool this is a 4 part assembly, for the 5' pool this is a 3 part assembly. These reactions were cycled from 37°C to 16°C over 5 minutes, 60x, then held at 60°C for 5 minutes to complete Golden Gate assembly and held at 4°C overnight. The following day the reaction was held at 60°C for 5 minutes before dialysis in accordance with NEB guidance. Membrane drop dialysis was then performed on each library pool for 90 minutes to enhance transformation efficiency. The entire 20 µl reaction

was then transformed into 50 µl competent *E. coli* 10G cells. Drop plates were made at this point (spotting 2.5 µl of dilutions of each library on Lb plates with spectinomycin) and total actual transformed cells were estimated at ~5x10^5 CFU/mL for the 5' pool at ~1x10^6 for the 3' pool. Each 1 mL pool was added to 4 mL Lb with spectinomycin and incubated overnight, then plasmids were purified. Plasmids concentration was determined by nanodrop and pools were then combined at an equimolar ratio to create the final phage variant pool, denoted as pHRec2-Motif. pHRec2-Motif was transformed into *E. coli* 10G with pHT7Helper1 by transforming 50 ng of plasmid into 60 ul of *E. coli* 10G competent cells. Drop plates were made (spotting 2.5 µl of dilutions of each library onto Lb plates with spectinomycin and kanamycin) and total actual transformed cells were estimated at ~6.7x10^5 CFU/mL. The 1 mL library was added to 4 mL Lb with spectinomycin and kanamycin and incubated overnight. This was repeated once to ensure >100 coverage of the entire library and all preparations were combined. This host, *E. coli* 10G with pHT7Helper1 and pHRec1-Motif, was the host used for Optimized Recombination during ORACLE.

Phage Library Preparation Using ORACLE

T7 Acceptor phages from our previous DMS study for the RBP tip domain were used for ORACLE (Huss et al., 2021). Optimized Recombination was performed by adding T7 Acceptor phages (MOI ~1) to 50 mL exponential phase 10G with pHT7Helper1 and pHRec2-Motif. Cultures were incubated until lysis and phages were purified, titer was estimated by spot plate at ~5x10^7 variants/mL in a total phage population of ~1x10^{10} PFU/mL. Accumulation was performed by adding ~MOI of 0.2 of recombined phages to

50 mL of stationary phase *E. coli* 10G with pHT7Helper1 and pHCas9-3 resuspended in fresh Lb with kanamycin and spectinomycin. Cultures were incubated until lysis then phages were purified. Phage titer was estimated to be ~$3x10^9$ PFU/mL on stationary *E. coli* 10G with pHT7Helper1 and ~$1.2x10^9$ on stationary *E. coli* 10G. Library expression was performed by adding the accumulated library to 5 mL *E. coli* 10G (with no plasmid) at an MOI of ~2. Cultures are incubated until lysis and phages were purified. This constitutes the final phage variant library with full expression of the variant *gp17* tail fibers. This phage population is directly sequenced to establish the pre-selection library population.

Motif Library Selection

All motif library selection experiments were performed in the same way. The T7 variant library was added to 5 mL of exponential host at an MOI of ~$2x10^{-2}$ and the culture was allowed to fully lyse. This MOI was chosen to give an estimated 2 replication cycles for the phage library. The entire process was repeated in biological triplicate for each host. Phages were then purified and sequenced.

Deep Sequencing Preparation and Analysis

We used deep sequencing to evaluate phage populations. We first amplified the tip domain by two step PCR, or tailed amplicon sequencing, using KAPA HiFi. Primers for deep sequencing attach to constant regions adjacent to the tip domain, 3' of the new barcode in pHT7Rec2. Constant regions are also installed in the fixed region of the acceptor phages for the similarly size amplicon so acceptor phages can also be detected.

The first PCR reaction adds an internal barcode (used for technical replicates to assay PCR skew), a variable N region (to assist with nucleotide diversity during deep sequencing, this is essential for DMS libraries due to low nucleotide diversity at each position), and the universal Illumina adapter. Four forward and four reverse primers were used in each reaction, each with a variable N count (0, 2, 4, or 8). Primers were mixed at equimolar ratios and total primers used was per recommended primer concentration. PCR was performed using 13 total cycles in the first PCR reaction. The product of this reaction was used directly in the second PCR reaction, which adds an index and the Illumina 'stem'. This PCR was run for 9 total PCR cycles. The product of this reaction was purified and was used directly for deep sequencing. Each phage population was sampled at least twice using separate internal barcodes, and no PCR reactions were pooled. Total PCR cycles overall for each sample was kept at 22x to avoid PCR skew. All phage samples were deep sequenced using an Illumina Miseq System, 2x250 read length using MiSeq Reagent Kit v2 or v2 Nano according to manufacturer documentation.

Paired-end Illumina sequencing reads were merged with PEAR using the default software parameters (Zhang et al., 2014). Phred quality scores (Q scores) were used to compute the total number of expected errors (E) for each merged read and reads exceeding an Emax of 1 were removed (Magoč and Salzberg, 2011). Wildtype, acceptor phages, and each variant were then counted in the deep sequencing output. We correlated read counts for each technical replicate to determine if there was any notable skew from PCR or deep sequencing. Technical PCR replicates correlated extremely well ($R \geq 0.98$ for all samples) indicating no relevant PCR skew and were aggregated for each biological replicate. Besides wildtype and acceptor counts, we included only sequences

with known motifs in our analysis, greatly reducing the possibility of deep sequencing error resulting in an incorrect read count for a variant. With this in mind, and to avoid missing low abundance members after selection, we used a low read cutoff of 4. A pseudo-count was not utilized. Possible outliers were identified by data sorting for UTI strains.

To score enrichment for each variant we used a basic functional score (F), averaging results of the three biological replicates where $F = \bar{x}\ \frac{Variant\ \%_{Post-Passage}}{Variant\ \%_{Pre-Passage}}$. To compare variant performance across hosts we normalized functional score (F_N) to wildtype, where $F_N = \bar{x}\ \frac{Variant\ \%_{Post-Passage}}{Variant\ \%_{Pre-Passage}} \Big/ \frac{WT\ \%_{Post-Passage}}{WT\ \%_{Pre-Passage}}$.

Other Statistical Analysis and Source Code

Hierarchical Clustering in Figure 3 was performed using the Complex Heatmaps package in R (Gu et al., 2016). Different distance values were computed using get_dist in R, clustering approaches were calculated using hclust in R, and cophenetic scores were calculated using cophenetic in R. See documentation on github for exact code used. Frequency plots in Figure 3 and Figure 4 were made using SequenceLogoVis (https://github.com/ISA-tools/SequenceLogoVis) (Maguire et al.). P-values for EOP graphs compare plaque capability on the tested host to the reference host for the EOP graph. Counting motif variants was done in Pandas. All other calculations and plots were made in Excel. Relevant statistical details of experiments can be found in the corresponding figure legends or relevant methods section. Source code is available at https://github.com/PhilHuss/MotifMining.

Bibliography

Abeles, S.R., and Pride, D.T. (2014). Molecular Bases and Role of Viruses in the Human Microbiome. J. Mol. Biol. *426*, 3892–3906. https://doi.org/10.1016/j.jmb.2014.07.002.

Ando, H., Lemire, S., Pires, D.P., and Lu, T.K. (2015). Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. Cell Syst. *1*, 187–196. https://doi.org/10.1016/j.cels.2015.08.013.

Andrews, B., and Fields, S. (2021). Balance between promiscuity and specificity in phage λ host range. ISME J. *15*, 2195–2205. https://doi.org/10.1038/s41396-021-00912-2.

Arthur, M., Arbeit, R.D., Kim, C., Beltran, P., Crowe, H., Steinbach, S., Campanelli, C., Wilson, R.A., Selander, R.K., and Goldstein, R. (1990). Restriction fragment length polymorphisms among uropathogenic Escherichia coli isolates: pap-related sequences compared with rrn operons. Infect. Immun. *58*, 471–479. .

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008. https://doi.org/10.1038/msb4100050.

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A.B., Pevzner, P., and Koonin, E.V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome *9*, 78. https://doi.org/10.1186/s40168-021-01017-w.

Bertozzi Silva, J., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. FEMS Microbiol. Lett. *363*. https://doi.org/10.1093/femsle/fnw002.

Bruder, K., Malki, K., Cooper, A., Sible, E., Shapiro, J.W., Watkins, S.C., and Putonti, C. (2016). Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges. Evol. Bioinforma. Online *12*, 25–33. https://doi.org/10.4137/EBO.S38549.

Brum, J.R., and Sullivan, M.B. (2015). Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat. Rev. Microbiol. *13*, 147–159. https://doi.org/10.1038/nrmicro3404.

Canfield, G.S., and Duerkop, B.A. (2020). Molecular mechanisms of enterococcal-bacteriophage interactions and implications for human health. Curr. Opin. Microbiol. *56*, 38–44. https://doi.org/10.1016/j.mib.2020.06.003.

Chen, M., Zhang, L., Xin, S., Yao, H., Lu, C., and Zhang, W. (2017). Inducible Prophage Mutant of Escherichia coli Can Lyse New Host and the Key Sites of Receptor Recognition Identification. Front. Microbiol. *8*. https://doi.org/10.3389/fmicb.2017.00147.

Clokie, M.R.J., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011). Phages in nature. Bacteriophage *1*, 31–45. https://doi.org/10.4161/bact.1.1.14942.

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as Winners in the Game of Life. Annu. Rev. Virol. *3*, 197–214. https://doi.org/10.1146/annurev-virology-100114-054952.

Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D., Dutilh, B.E., and Thompson, F.L. (2017). Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat. Commun. *8*, 15955. https://doi.org/10.1038/ncomms15955.

Dedrick, R.M., Guerrero-Bustamante, C.A., Garlena, R.A., Russell, D.A., Ford, K., Harris, K., Gilmour, K.C., Soothill, J., Jacobs-Sera, D., Schooley, R.T., et al. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nat. Med. *25*, 730. https://doi.org/10.1038/s41591-019-0437-z.

Dickinson, B.C., Leconte, A.M., Allen, B., Esvelt, K.M., and Liu, D.R. (2013). Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. Proc. Natl. Acad. Sci. *110*, 9007–9012. https://doi.org/10.1073/pnas.1220670110.

Domingo, J., Baeza-Centurion, P., and Lehner, B. (2019). The Causes and Consequences of Genetic Interactions (Epistasis). Annu. Rev. Genomics Hum. Genet. *20*, 433–460. https://doi.org/10.1146/annurev-genom-083118-014857.

Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M.J., et al. (2019). Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Rep. *29*, 1336-1350.e4. https://doi.org/10.1016/j.celrep.2019.09.062.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., et al. (2008). The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. J. Bacteriol. *190*, 2597–2606. https://doi.org/10.1128/JB.01695-07.

Eskenazi, A., Lood, C., Wubbolts, J., Hites, M., Balarjishvili, N., Leshkasheli, L., Askilashvili, L., Kvachadze, L., van Noort, V., Wagemans, J., et al. (2022). Combination of pre-adapted bacteriophage therapy and antibiotics for treatment of fracture-related infection due to pandrug-resistant Klebsiella pneumoniae. Nat. Commun. *13*, 302. https://doi.org/10.1038/s41467-021-27656-z.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507. https://doi.org/10.1016/j.jmb.2006.03.043.

Fraser, J.S., Maxwell, K.L., and Davidson, A.R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr. Opin. Microbiol. *10*, 382–387. https://doi.org/10.1016/j.mib.2007.05.018.

Garcia, E., Elliott, J.M., Ramanculov, E., Chain, P.S.G., Chu, M.C., and Molineux, I.J. (2003). The genome sequence of Yersinia pestis bacteriophage phiA1122 reveals an intimate history with the coliphage T3 and T7 genomes. J. Bacteriol. *185*, 5248–5262. .

Garcia-Doval, C., and Raaij, M.J. van (2012). Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. Proc. Natl. Acad. Sci. *109*, 9390–9395. https://doi.org/10.1073/pnas.1119719109.

Gebhart, D., Williams, S.R., and Scholl, D. (2017). Bacteriophage SP6 encodes a second tailspike protein that recognizes Salmonella enterica serogroups C2 and C3. Virology *507*, 263–266. https://doi.org/10.1016/j.virol.2017.02.025.

González-García, V.A., Bocanegra, R., Pulido-Cid, M., Martín-Benito, J., Cuervo, A., and Carrascosa, J.L. (2015). Characterization of the initial steps in the T7 DNA ejection process. Bacteriophage *5*. https://doi.org/10.1080/21597081.2015.1056904.

Gu, C., Liang, Y., Li, J., Shao, H., Jiang, Y., Zhou, X., Gao, C., Li, X., Zhang, W., Guo, C., et al. (2021). Saline lakes on the Qinghai-Tibet Plateau harbor unique viral assemblages mediating microbial environmental adaption. IScience *24*, 103439. https://doi.org/10.1016/j.isci.2021.103439.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

Hatfull, G.F. (2008). Bacteriophage Genomics. Curr. Opin. Microbiol. *11*, 447–453. https://doi.org/10.1016/j.mib.2008.09.004.

Hatfull, G.F., Dedrick, R.M., and Schooley, R.T. (2022). Phage Therapy for Antibiotic-Resistant Bacterial Infections. Annu. Rev. Med. *73*, 197–211. https://doi.org/10.1146/annurev-med-080219-122208.

Heineman, R.H., Springman, R., and Bull, J.J. (2008). Optimal foraging by bacteriophages through host avoidance. Am. Nat. *171*, E149-157. https://doi.org/10.1086/528962.

Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (2020). A continuous evolution system for contracting the host range of bacteriophage T7. Sci. Rep. *10*, 1–8. https://doi.org/10.1038/s41598-019-57221-0.

Huss, P., and Raman, S. (2020). Engineered bacteriophages as programmable biocontrol agents. Curr. Opin. Biotechnol. *61*, 116–121. https://doi.org/10.1016/j.copbio.2019.11.013.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. (2021). Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. ELife *10*, e63775. https://doi.org/10.7554/eLife.63775.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. *31*, 233–239. https://doi.org/10.1038/nbt.2508.

de Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., and Dutilh, B.E. (2019). Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends Microbiol. *27*, 51–63. https://doi.org/10.1016/j.tim.2018.08.006.

Khan Mirzaei, M., and Nilsson, A.S. (2015). Isolation of Phages for Phage Therapy: A Comparison of Spot Tests and Efficiency of Plating Analyses for Determination of Host Range and Efficacy. PLoS ONE *10*. https://doi.org/10.1371/journal.pone.0118557.

Kilcher, S., and Loessner, M.J. (2019). Engineering Bacteriophages as Versatile Biologics. Trends Microbiol. *27*, 355–367. https://doi.org/10.1016/j.tim.2018.09.006.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *110*, 14024–14029. https://doi.org/10.1073/pnas.1301301110.

Kutter, E.M., Kuhl, S.J., and Abedon, S.T. (2015). Re-establishing a place for phage therapy in western medicine. Future Microbiol. *10*, 685–688. https://doi.org/10.2217/fmb.15.28.

Langer, S.J., Ghafoori, A.P., Byrd, M., and Leinwand, L. (2002). A genetic screen identifies novel non-compatible loxP sites. Nucleic Acids Res. *30*, 3067–3077. .

Lin, T.-Y., Lo, Y.-H., Tseng, P.-W., Chang, S.-F., Lin, Y.-T., and Chen, T.-S. (2012). A T3 and T7 Recombinant Phage Acquires Efficient Adsorption and a Broader Host Range. PLOS ONE *7*, e30954. https://doi.org/10.1371/journal.pone.0030954.

Luo, E., Leu, A.O., Eppley, J.M., Karl, D.M., and DeLong, E.F. (2022). Diversity and origins of bacterial and archaeal viruses on sinking particles reaching the abyssal ocean. ISME J. *16*, 1627–1635. https://doi.org/10.1038/s41396-022-01202-1.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507.

Maguire, E., Rocca-Serra, P., Sansone, S.-A., and Chen, M. Redesigning the Sequence Logo with Glyph-based Approaches to Aid Interpretation. 5. .

Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., and Lenski, R.E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. Science *335*, 428–432. https://doi.org/10.1126/science.1214449.

Miton, C.M., Buda, K., and Tokuriki, N. (2021). Epistasis and intramolecular networks in protein evolution. Curr. Opin. Struct. Biol. *69*, 160–168. https://doi.org/10.1016/j.sbi.2021.04.007.

Molineux, I.J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol. Microbiol. *40*, 1–8. https://doi.org/10.1046/j.1365-2958.2001.02357.x.

Nishikawa, K.K., Hoppe, N., Smith, R., Bingman, C., and Raman, S. (2021). Epistasis shapes the fitness landscape of an allosteric specificity switch. Nat. Commun. *12*, 5562. https://doi.org/10.1038/s41467-021-25826-7.

Nobrega, F.L., Vlot, M., de Jonge, P.A., Dreesens, L.L., Beaumont, H.J.E., Lavigne, R., Dutilh, B.E., and Brouns, S.J.J. (2018). Targeting mechanisms of tailed bacteriophages. Nat. Rev. Microbiol. *16*, 760–773. https://doi.org/10.1038/s41579-018-0070-8.

Pratama, A.A., and van Elsas, J.D. (2018). The “Neglected” Soil Virome - Potential Role and Impact. Trends Microbiol. *26*, 649–662. https://doi.org/10.1016/j.tim.2017.12.004.

Qimron, U., Marintcheva, B., Tabor, S., and Richardson, C.C. (2006). Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. Proc. Natl. Acad. Sci. U. S. A. *103*, 19039–19044. https://doi.org/10.1073/pnas.0609428103.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., et al. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrob. Agents Chemother. *61*, e00954-17. https://doi.org/10.1128/AAC.00954-17.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host Microbe *26*, 527-541.e5. https://doi.org/10.1016/j.chom.2019.09.009.

Winter, C., Garcia, J.A.L., Weinbauer, M.G., DuBow, M.S., and Herndl, G.J. (2014). Comparison of Deep-Water Viromes from the Atlantic Ocean and the Mediterranean Sea. PLOS ONE *9*, e100600. https://doi.org/10.1371/journal.pone.0100600.

Wu, R., Davison, M.R., Nelson, W.C., Graham, E.B., Fansler, S.J., Farris, Y., Bell, S.L., Godinez, I., Mcdermott, J.E., Hofmockel, K.S., et al. (2021). DNA Viral Diversity, Abundance, and Functional Potential Vary across Grassland Soils with a Range of Historical Moisture Regimes. MBio *12*, e0259521. https://doi.org/10.1128/mBio.02595-21.

Yehl, K., Lemire, S., Yang, A.C., Ando, H., Mimee, M., Torres, M.D.T., de la Fuente-Nunez, C., and Lu, T.K. (2019). Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. Cell *179*, 459-469.e9. https://doi.org/10.1016/j.cell.2019.09.015.

Yosef, I., Goren, M.G., Globus, R., Molshanski-Mor, S., and Qimron, U. (2017). Extending the Host Range of Bacteriophage Particles for DNA Transduction. Mol. Cell *66*, 721-728.e3. https://doi.org/10.1016/j.molcel.2017.04.025.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620. https://doi.org/10.1093/bioinformatics/btt593.

Chapter 6

Publication

Investigating Bacteriophage T7 Interactions in Microgravity

Authors: Phil Huss[1,2,3], Chutikarn Chitboonthavisuk[1,2,3], Anthony Meger[1], Kyle Nishikawa[1], Heath Mills[4], R.P. Oates[4], Olivia Holzhaus[4], and Srivatsan Raman[1,2,5*]

Contributions:

P.H.: Conceptualization, Data curation, Software, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing
C.C.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review and editing
A.M.: Software
K.N.: Software
H.M., R.O., O.H.: Resources, Project administration, Methodology, Writing - review and editing
S.R.: Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing - review and editing

Affiliations:

[1]Department of Biochemistry, University of Wisconsin-Madison
[2]Department of Bacteriology, University of Wisconsin-Madison
[3]Microbiology Doctoral Training Program, University of Wisconsin-Madison
[4]Rhodium Scientific
[4]Department of Chemical and Biological Engineering, University of Wisconsin-Madison
[*] Correspondence: sraman4@wisc.edu (Srivatsan Raman)

Abstract

Bacteriophages, or ‘phages’, are bacterial viruses that are the most abundant organism on Earth. While phage research has progressed considerably terrestrially, how phages and bacteria interact in microgravity is largely unknown. Microgravity presents enormous challenge for microorganisms, which are greatly affected by gravity to mediate cell-cell and cell-phage interactions. To elucidate these interactions, we explore how T7 bacteriophage interacts with *E. coli* BL21 in microgravity onboard the International Space Station (ISS). We incubated samples in gravity and microgravity for short- and long-term experiments, finding that lysis is delayed in microgravity conditions and suggesting this phage is inhibited in microgravity. We identified novel mutations in phage proteins that influence activity in microgravity using whole genome sequencing and examine how microgravity influences selection in phage structural proteins. Our study provides a preliminary examination of the influence of microgravity on phage-host interactions. The success of the approach used in this study provides a foundation for future research in microgravity to explore interactions between phages and bacteria that define microbial communities.

Introduction

Bacteriophages, or ‘phages’, are bacterial viruses that are the most abundant organism on Earth (Cobián Güemes et al., 2016). Phages play an essential role in microbial ecosystems and have become increasingly valuable tools for applications ranging from targeted killing of antibiotic resistant bacteria to precision modification of the microbiome (Canfield and Duerkop, 2020; Chen et al., 2017; Clokie et al., 2011; Dedrick et al., 2019; Dunne et al., 2019; Eskenazi et al., 2022; Garcia et al., 2003; Hatfull et al., 2022; Heineman et al., 2008; Holtzman et al., 2020; Huss and Raman, 2020; Kilcher and Loessner, 2019; Kutter et al., 2015; Meyer et al., 2012; Schooley et al., 2017; Yehl et al., 2019). While phage research has progressed considerably terrestrially, how phages and bacteria interact in microgravity is largely unknown. Microgravity presents enormous challenge for microorganisms, which are greatly affected by gravity to mediate cell-cell and cell-phage interactions. To elucidate these interactions, we explore how T7 bacteriophage interacts with *E. coli* BL21 in microgravity onboard the International Space Station (ISS).

Microgravity poses a challenge for phages and bacteria and can significantly alter how phages are able to interact with their bacterial hosts. Phages do not actively seek out their bacterial hosts but are reliant on random motion to encounter host cells (Joiner et al., 2019; Kasman and Porter, 2022). Several fundamental forces are altered in microgravity. In gravity, sedimentation causes materials to move downward in response to gravitational pull. Sedimentation is lacking in microgravity and affects diffusion of nutrients to bacterial cells and cell motility (Benoit and Klaus, 2005; Freed and Vunjak-Novakovic, 2002; Klaus et al., 1997; Todd, 2007). Convection, the movement of fluids

and gasses, is also limited in microgravity. This movement is contingent on temperature changes that cause changes in density, and is therefore reliant on the presence of gravity (Klaus et al., 1997; Nickerson et al., 2004). This effect could cause waste products to accumulate near cells and influence growth rates. These changed conditions can influence a phages ability to grow on a host and force phages to rely entirely on only Brownian motion to encounter host cells. With limited predation by phages, the microbial ecosystem could be dramatically impacted.

Here, we explored how microgravity influences T7 phage and susceptible host *E. coli* BL21. We incubated samples in gravity and microgravity for short- and long-term experiments to evaluate if lysis of the bacterial host was delayed and found that lysis is delayed in microgravity conditions, suggesting this phage is inhibited in microgravity. We analyzed results for long term incubation using Whole Genome Sequencing (WGS) to evaluate what mutations accumulate in phage and bacteria over long term incubation. We identified many novel mutations in phage proteins, with most mutations occurring in structural proteins that facilitate interaction with bacterial host receptors. We finally examine how microgravity influences selection in a Deep Mutational Scanning (DMS) library of T7 phage receptor binding protein (RBP) tip domain variants (Huss et al., 2021). We identify a distinct pattern for substitutions that differs from gravity condition.

Our study provides a preliminary examination of the influence of microgravity on phage-host interactions. By identifying mutations that accumulate in microgravity we highlight genes and regions of the phage that can be targeted to engineer phages to alter phage activity. The success of the approach used in this study also provides a foundation

for future research into interactions between phages and bacteria that define microbial communities.

Results

Preparing bacteriophage and bacterial samples for experimentation on the ISS

Experiments involving phages and bacteria onboard the ISS have several practical challenges that must be overcome. Classic experiments to evaluate phage activity, like plaque assays or growth curves, are not possible onboard the ISS. Instead, all experiments must be carried out in closed vessels to limit risk of exposure. Furthermore, both space and time are highly limited in this environment and all samples require rigorous testing to ensure astronaut safety. Our experimental outline was designed to accommodate these restrictions.

We devised two sets of samples, one set for incubation in microgravity and one control set used for incubation terrestrially in gravity. In brief, all samples were prepared terrestrially by mixing phage and/or bacteria in cryovials then immediately freezing samples at -80C. Samples were then transported to NASA, where samples were delivered to the ISS for incubation in microgravity. After incubation samples were refrozen and transported back to Earth, where they were thawed and immediately tested for plaque and bacterial counts and prepared for sequencing (Figure 1). The duration of freezing and incubation was recorded, and a second set of samples was tested terrestrially using the same incubation and freezing times (see Supplementary File 1 for incubation and freezing records).

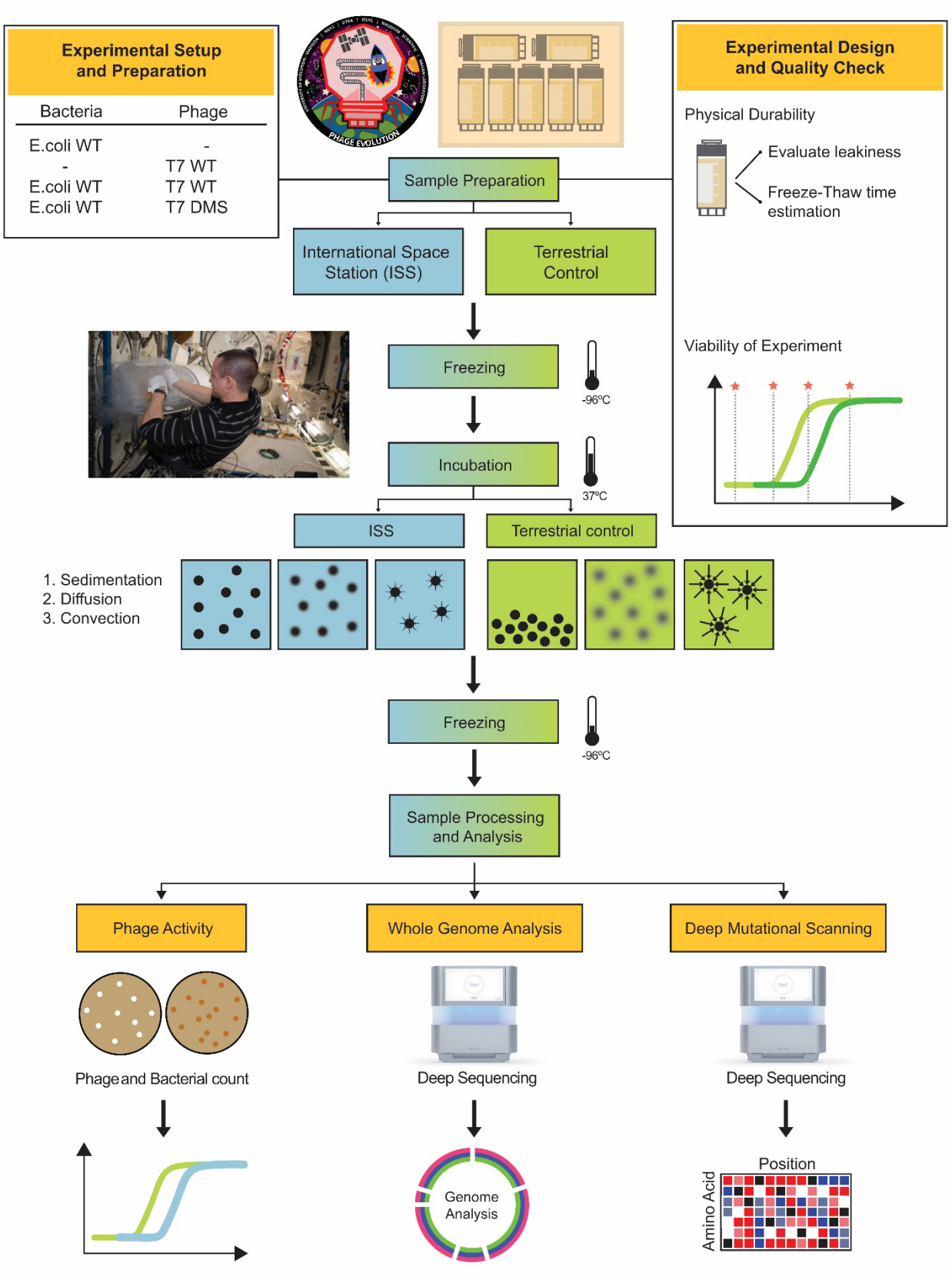

Experimental Setup and Preparation
Bacteria
Phage
E.coli WT
-
T7 WT
E.coli WT
T7 WT
E.coli WT
T7 DMS
Experimental Design and Quality Check
Physical Durability
Evaluate leakiness
Freeze-Thaw time estimation
Viability of Experiment
Sample Preparation
International Space Station (ISS)
Terrestrial Control
Freezing
-96°C
Incubation
37°C
ISS
Terrestrial control
1. Sedimentation
2. Diffusion
3. Convection
Freezing
-96°C
Sample Processing and Analysis
Phage Activity
Whole Genome Analysis
Deep Mutational Scanning
Phage and Bacterial count
Deep Sequencing
Deep Sequencing
Genome Analysis
Position
Amino Acid

Figure 1. Outline of terrestrial and ISS experimental process

Illustration of experimental process to prepare and test samples terrestrially and in microgravity onboard the International Space Station (ISS). Samples are prepared terrestrially, and quality checked to ensure samples cannot leak and cryovial integrity is unaffected by the freeze-thaw process. Two sets of samples are eventually tested, one in microgravity aboard the ISS (left) and one terrestrially under the influence of microgravity (right). Samples are frozen, then thawed and incubated in either condition for specified intervals. Samples are then re-frozen and evaluated terrestrially for phage and bacterial counts, for whole genome sequencing, and for phages deep mutational scanning of the T7 receptor binding protein tip domain.

To ensure sample integrity and to meet NASA standards for astronaut safety, the cryovial container underwent biocompatibility and leakiness testing (see Supplementary File 2). Briefly, biocompatibility was validated by the change in sample weight after a period of freezing and incubation. Samples were prepared, immediately frozen at -80C for all timepoints used in the experiment, and then were weighed to ensure no liquid had escaped the container. All samples passed biocompatibility and safety testing.

Bacteriophage T7 growth is inhibited in microgravity

We first assessed if *E. coli* BL21 was susceptible to T7 infection in microgravity. Under terrestrial conditions this host is susceptible to T7 infection and infection of the host is typically completed in ~30 minutes with a burst size of roughly 100 (Chitboonthavisuk et al., 2021). We hypothesized infection would be significantly slower under conditions of microgravity because phages would have less chances to randomly encounter and infect bacteria. To evaluate this hypothesis, we mixed phages and bacteria at MOIs of 10^{-6} and 10^{-4} in triplicate and targeted incubation times for 1 hour, 2 hours, 3 hours and 20 days, with actual incubation times on the ISS being 1 hour 12 minutes, 2 hours 11 minutes, 4 hours 17 minutes, and 23 days 1hour and 30 minutes (Figure 2A). Controls consisting of phages or bacteria alone at a titer of 10^8 at each time point. Time points were selected to cover a wide range of possible outcomes, as we had no initial indication of how delayed, if at all, host lysis would be under microgravity. Additionally, we anticipated loss of cell and phage viability after multiple freeze-thaws, as no cryoprotectant could be included with samples.

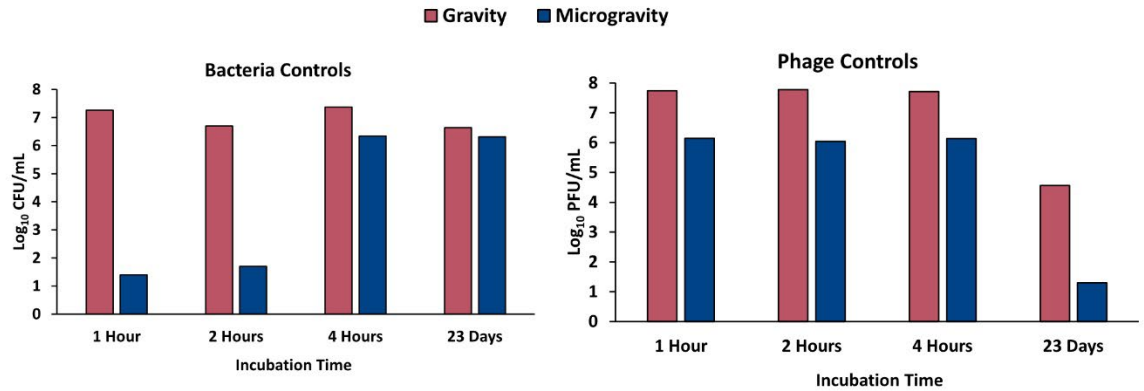
Analysis of bacterial controls indicated bacterial viability was greatly reduced when counted after 1 and 2 hours of incubation in microgravity (Figure 2B). However, incubation for 4 hours or 23 days in microgravity was saw less impact on the countable population of bacteria with only a 0.5-1 log loss of bacteria. Phages saw a two-log loss in titer in earlier timepoints in microgravity and a significant decrease in viability after 23 days in both conditions. Titers in mixed cultures for both MOIs saw phage population soared at 4 hours and after 23 days in gravity but only after 23 days in microgravity, with a significant drop in viable bacteria at all timepoints (Figure 2C-D).

While both bacteria and phages were significantly impacted by the freeze-thaw cycles used in this experiment, the extended time points allow us to approximate phage activity in microgravity. T7 phage was only able to productively lyse *E. coli* BL21 at some point after 4 hours in microgravity but was able to productively lyse this host after only 4 hours in gravity, indicating lysis was delayed in microgravity. Future experiments using this model can narrow time to lyse host in microgravity further by examining later timepoints.

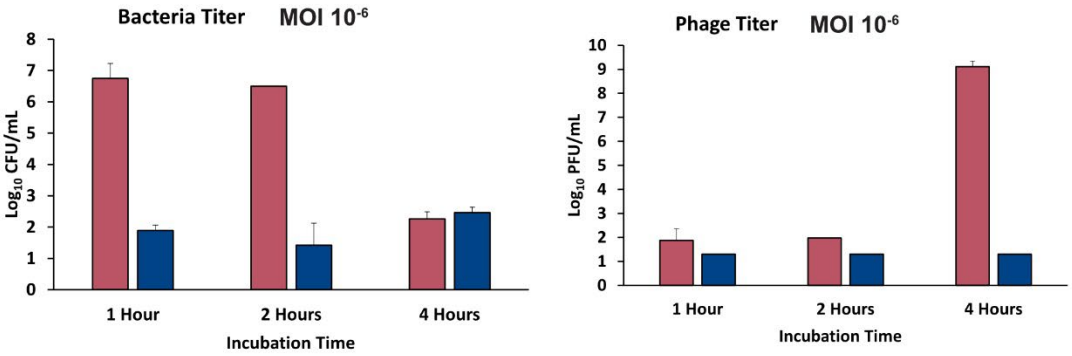
A

Sample Composition	1 Hour (1:12 Final)	2 Hours (2:11 Final)	3 Hours (4:17 Final)	25 Days (23:01:30 Final)
Phage / Bacteria – MOI 10^{-6}	3x	3x	3x	
Phage / Bacteria – MOI 10^{-4}	3x	3x	3x	3x
Bacteria Control (No Phage)	1x	1x	1x	1x
Phage Control (No Bacteria)	1x	1x	1x	1x
T7 tail fiber library (tip domain DMS, MOI 10^{-2})				3x

B



C



D

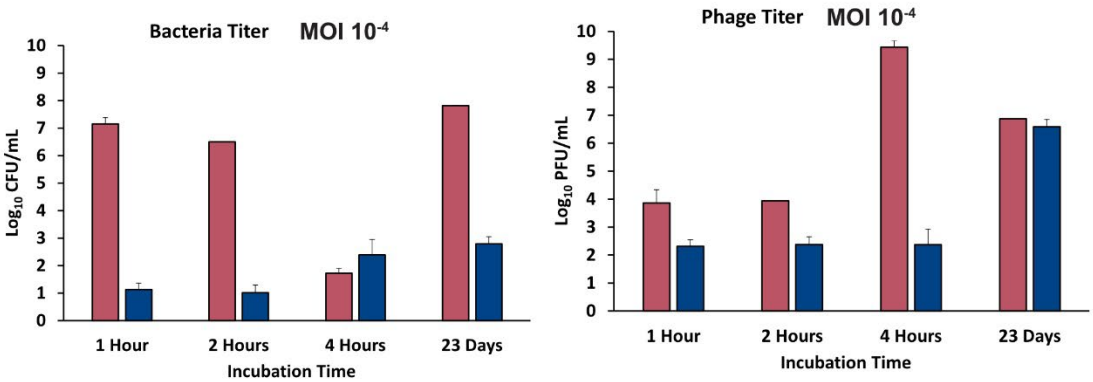


Figure 2. Bacteriophage T7 growth is inhibited by microgravity

(A) Summary of samples and timepoints used in experiments. Time in parenthesis indicates actual incubation time on the ISS, which was subsequently used for terrestrial samples. **(B-D)** Titer of bacterial and phage samples at different incubation times (log_{10} PFU or CFU/mL) for gravity (red) and microgravity (blue) conditions for (B) control samples, (C) MOI^{-6} sample series, and (D) MOI^{-4} sample series. Triplicate samples represented as mean +/- SD.

Incubation in microgravity reveals novel mutations

We next sought to assess what mutations arose in microgravity that could influence phage-host host relationships. To facilitate this analysis, we extracted phage genomes from our 23-day samples and performed whole genome sequencing (WGS). Samples were sequenced at a significant sequencing depth to find mutations that were lowly abundant, with average coverage of 40,000x. This was critical, as it was unlikely mutations would be so beneficial to completely overtake either population and be found as consensus reads in the population. Mutation abundance was determined by comparing mutation frequency to unpassaged phage controls. Analysis of WGS data revealed several hotspots of activity for across both gravity and microgravity conditions.

We found many mutations in structural proteins of T7 as well as several uncharacterized proteins. Structural proteins are proteins located on the phage virion and include tail and tail fiber structures responsible for host recognition (Huss et al., 2021; Molineux, 2001) (Figure 2A). Besides structural genes, the genome of bacteriophage T7 can roughly be grouped into three parts based on if genes are transcribed early in infection (class I), shortly after infection (class II), or late in infection (class III) (Yang et al., 2014) (Figure 2B). We saw a predominance of mutations accumulate in structural proteins in both conditions, particularly in *gp11, gp12 and gp17* which form parts of the tail and tail fiber (Cuervo et al., 2019; González-García et al., 2015). In contrast, mutations were noted to be much more predominant in *gp7.3* in microgravity. This protein is ejected into the cell upon adsorption but is not considered essential for a successful infection (Kemp et al., 2005), making the difference in mutations between conditions notable. Given that this protein is expected to make contact with the cell surface during infection,

we theorize these mutations may have been helpful for retaining the phage to the cell surface in the difficult conditions of microgravity, but otherwise provided no fitness advantage in gravity. Intriguingly the majority of mutations in *gp7.3* were 3 and 6 amino acid deletions located between positions 39 and 47 of the gene, while no deletions were seen in microgravity. Altogether these results indicate *gp7.3* may have a greater role to play in phage activity than previously known.

Other mutations that arose that were distinct across conditions included mutations in *gp4.7* in gravity and highly abundant mutations in *gp0.5* in microgravity. *Gp0.5* is an unknown protein with a putative assigned function as a transmembrane protein. Mutations in this gene are notable because both microgravity replicates with mutations in this gene had the same mutation, V26A, that overtook the entire phage population, indicating it caused strong selection pressure. *gp4.7* has no putative assigned function, but both proteins can be found in many different phages, indicating they may play a larger role in phage activity.

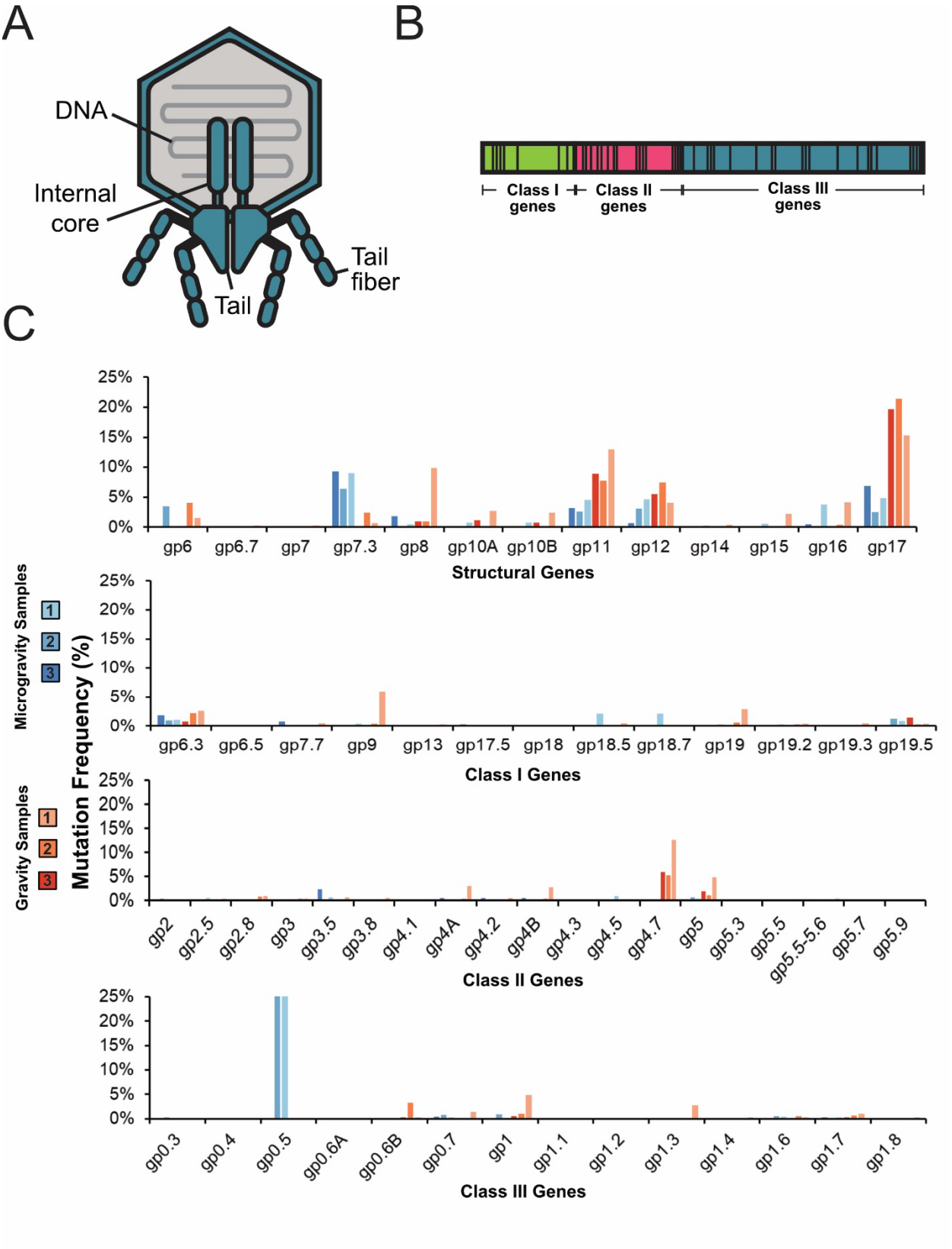

A
DNA
Internal core
Tail
Tail fiber
B
Class I genes
Class II genes
Class III genes
C
Mutation Frequency (%)
Microgravity Samples
Gravity Samples
1
2
3
25%
20%
15%
10%
5%
0%
gp6 gp6.7 gp7 gp7.3 gp8 gp10A gp10B gp11 gp12 gp14 gp15 gp16 gp17
Structural Genes
gp6.3 gp6.5 gp7.7 gp9 gp13 gp17.5 gp18 gp18.5 gp18.7 gp19 gp19.2 gp19.3 gp19.5
Class I Genes
gp2 gp2.5 gp2.8 gp3 gp3.5 gp3.8 gp4.1 gp4A gp4.2 gp4B gp4.3 gp4.5 gp4.7 gp5 gp5.3 gp5.5 gp5.5-5.6 gp5.7 gp5.9
Class II Genes
gp0.3 gp0.4 gp0.5 gp0.6A gp0.6B gp0.7 gp1 gp1.1 gp1.2 gp1.3 gp1.4 gp1.6 gp1.7 gp1.8
Class III Genes

Figure 3. Incubation in microgravity reveals novel mutations

(A) Illustration of bacteriophage T7, highlighting key structural proteins in the virion. **(B)** Illustration of the genome of bacteriophage T7, divided into early class I genes (green), middle class II genes (pink) and late class II genes (blue). **(C)** Total percentage mutation frequency for structural genes and other class I, II and II genes for each replicate incubated in microgravity (blue, shaded) and gravity (red, shaded).

Deep Mutational Scanning profiles beneficial substitutions in the RBP tip domain

We next examined how activity of individual substitutions was impacted in microgravity using a deep mutational scanning (DMS) library of the tip domain of the receptor binding protein (RBP). The RPB of T7 phage consists of six short non-contractile tails that form a homotrimer composed of a rigid shaft ending with a β-sandwich tip domain (Garcia-Doval and Raaij, 2012). This domain is a major determinant of phage activity and interacts with host receptor lipopolysaccharide (LPS) and position the phage for successful, irreversible binding with the host (Fraser et al., 2006, 2007; González-García et al., 2015; Lin et al., 2012; Molineux, 2001; Qimron et al., 2006). We have previously used this library to functionally characterize the tip domain in short passage experiments (Huss et al., 2021). One replicate of our 23-day samples for each condition retained sufficient phage titer for analysis, which we deep sequenced and scored.

This phage DMS library consists of a pool of variants with a single substitution in the tip domain, encompassing all possible substitutions at each codon spanning residue

positions 472-554 in the RBP (residue numbering based on PDB 4A0T). Deep sequencing in each condition revealed a fairly selective environment (Figure 4A-B). Mutations in gravity largely mirror our previous results, with the exception of positive and negatively charged mutations in the FG exterior loop. Positively charged substitutions at this position have increased activity in hosts with truncated LPS, but negatively charged substitutions at this position are novel.

Substitutions that performed the best in microgravity accumulated in three distinct locations. Large, positively charged substitutions were preferred in the BC exterior loop, mirroring results from our initial study for strains that have truncated LPS. Similar substitutions are preferred in the GH interior loop and proximal H β-sheet. These positions face inwards and away from putative receptor binding interfaces in the tip domain, suggesting their presence positively influences the structure of the tip domain in some way. Finally, a host of mutations are present at N-terminus of the protein replacing the stop codon. In our construct replacement of the stop codon results in a three amino acid -DAR extension. Its apparent this extension increases activity in microgravity, making small extensions of the tip domain a feasible avenue for engineering the phage to increase activity in the future.

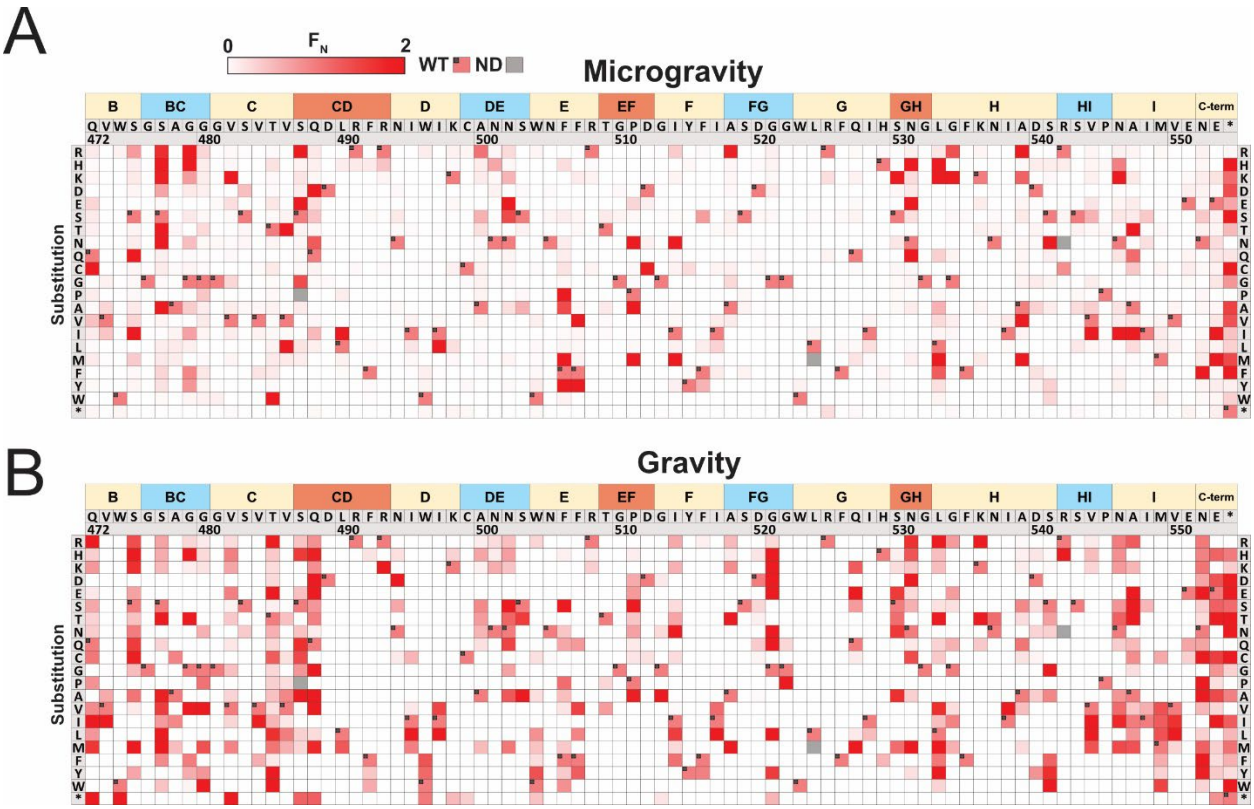


Figure 4. Deep Mutational Scanning profiles beneficial substitutions in the RBP tip domain

(A-B) Heatmaps showing normalized functional scores (F_N) of all substitutions (red gradient) and wildtype amino acid (F_N=1 and black dot upper left) at every position after incubation with *E. coli* BL21 in (**A**) microgravity and (B) gravity. Residue numbering (based on PDB 4A0T), wildtype amino acid and secondary structure topology are shown above left to right, substitutions listed top to bottom.

Discussion

The interplay between phages and bacteria largely defines microbial ecosystems. If microgravity influences this interaction there may be enormous repercussions for how microbial communities develop in microgravity. Here, we describe a preliminary study to evaluate how microgravity influences interactions between T7 bacteriophage and *E. coli* BL21. Our results indicate this phage is inhibited in microgravity and has a delayed lysis phenotype. To explore changes in phage sequence after incubation in microgravity we examine WGS and DMS data from phages incubated over 23 days with their bacterial host. We identified many novel mutations in phage proteins, with most mutations occurring in structural proteins that facilitate interaction with bacterial host receptors and identified mutations in as yet uncharacterized proteins that may influence phage function. Microgravity also influenced selection of DMS variants in the T7 phage RBP, with novel mutations accumulating in the c-terminal region and new areas near the GH interior loop.

Overall, our study provides a preliminary examination of the influence of microgravity on phage-host interactions. Investigating phages in new environments sheds light on different genes that can influence phage activity. Investigation of these leads may provide avenues for engineering phages terrestrially in future experiments. The success of this approach lays the groundwork for future phage studies in microgravity aboard the ISS.

Supplementary Files

Supplementary File 1 – Incubation and Freezing records for samples

Supplementary File 2 – Rhodium biocompatibility protocol

Supplementary File 2 – List of Primers using in experiment

Short Methods

T7 bacteriophage was obtained from ATCC (ATCC® BAA-1025-B2). *Escherichia coli* BL21 is a lab stock. All bacterial hosts are grown in and plated on Lb media (1% Tryptone, 0.5% Yeast Extract, 1% NaCl in dH2O, plates additionally contain 1.5% agar, while top agar contains only 0.5% agar) and Lb media was used for all experimentation and was used to recover hosts. All incubations of bacterial cultures were performed at 37°C without shaking. T7 bacteriophage was propagated using *E. coli* BL21 after initial receipt from ATCC and then as described on various hosts in methods. All phage experiments were performing using Lb and culture conditions as described for bacterial hosts. Phages were stored in Lb at 4°C. For freezing microbes were stored at -80°C in 100% relevant media.

PCR was performed using KAPA HiFi (Roche KK2101) for all experiments. All cloning was performed according to manufacturer documentation except where noted in methods. Whole Genome Sequencing was prepared using Illumina DNA Prep. PCR reactions use 1 µl of ~1 ng/µl plasmid or ~0.1 ng/µl DNA fragment as template for relevant reactions. PCR reactions using phage as template use 1 µl of undiluted phage stock (genomic extraction was unnecessary) and have an extended 5 minute 95°C denaturation step. Detailed protocols for cloning are available on request. All primers used in experiments in this publication are listed in supplementary file 3.

Samples were prepared by mixing ~$1x10^8$ CFU of exponential phase *E. coli* BL21 with the noted amount of T7 phages in Rhodium Cryotubes.

Bacterial concentrations were determined by serial dilution of bacterial culture (1:10 or 1:100 dilutions made to 1 mL in 1.5 microcentrifuge tubes in Lb) and subsequent

plating and bead spreading of 100 µl of a countable dilution (targeting 50 colony forming units) on Lb plates. Plates were incubated overnight and counted the next morning. Typically, two to three dilution series were performed for each host to initially establish concentration at different OD_{600} and subsequent concentrations were confirmed with a single dilution series for later experiments.

Exponential phase culture consists of stationary culture diluted 1:20 in Lb then incubated at 37°C until an OD_{600} of ~0.4-0.8 is reached (as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation). Phage lysate was purified by centrifuging phage lysate at 16g, then filtering supernatant through a 0.22 uM filter. Chloroform was not used.

To establish titer, phage samples were typically serially diluted (1:10 or 1:100 dilutions made to 1 mL in 1.5 microcentrifuge tubes) in Lb to a 10^{-8} dilution for preliminary titering by spot assay. Spot assays were performed by mixing 250 µl of relevant bacterial host in stationary phase with 3.5 mL of 0.5% top agar, briefly vortexing, then plating on Lb plates warmed to 37°C. After plates solidified (typically ~5 minutes), 1.5 µl of each dilution of phage sample was spotted in series on the plate. Plates were incubated and checked after overnight incubation (~20-30 hours) to establish a titer.

MOI was calculated by dividing phage titer by bacterial concentration. MOI for the T7 variant library after the variant gene is expressed was estimated by titering on 10G with pHT7Helper1.

We used deep sequencing to evaluate phage populations as previously described (Huss et al., 2021). Breseq was used to identify mutations for WGS (Deatherage and Barrick, 2014).

References

Benoit, M., and Klaus, D. (2005). Can genetically modified Escherichia coli with neutral buoyancy induced by gas vesicles be used as an alternative method to clinorotation for microgravity studies? Microbiology (Reading) *151*, 69–74. https://doi.org/10.1099/mic.0.27062-0.

Canfield, G.S., and Duerkop, B.A. (2020). Molecular mechanisms of enterococcal-bacteriophage interactions and implications for human health. Curr. Opin. Microbiol. *56*, 38–44. https://doi.org/10.1016/j.mib.2020.06.003.

Chen, M., Zhang, L., Xin, S., Yao, H., Lu, C., and Zhang, W. (2017). Inducible Prophage Mutant of Escherichia coli Can Lyse New Host and the Key Sites of Receptor Recognition Identification. Front. Microbiol. *8*. https://doi.org/10.3389/fmicb.2017.00147.

Chitboonthavisuk, C., Luo, C.H., Huss, P., Fernholz, M., and Raman, S. (2021). Engineering a dynamic, controllable infectivity switch in bacteriophage T7.

Clokie, M.R.J., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011). Phages in nature. Bacteriophage *1*, 31–45. https://doi.org/10.4161/bact.1.1.14942.

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as Winners in the Game of Life. Annu Rev Virol *3*, 197–214. https://doi.org/10.1146/annurev-virology-100114-054952.

Cuervo, A., Fàbrega-Ferrer, M., Machón, C., Conesa, J.J., Fernández, F.J., Pérez-Luque, R., Pérez-Ruiz, M., Pous, J., Vega, M.C., Carrascosa, J.L., et al. (2019). Structures of T7 bacteriophage portal and tail suggest a viral DNA retention and ejection mechanism. Nat Commun *10*, 3746. https://doi.org/10.1038/s41467-019-11705-9.

Deatherage, D.E., and Barrick, J.E. (2014). Identification of mutations in laboratory evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol *1151*, 165–188. https://doi.org/10.1007/978-1-4939-0554-6_12.

Dedrick, R.M., Guerrero-Bustamante, C.A., Garlena, R.A., Russell, D.A., Ford, K., Harris, K., Gilmour, K.C., Soothill, J., Jacobs-Sera, D., Schooley, R.T., et al. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nature Medicine *25*, 730. https://doi.org/10.1038/s41591-019-0437-z.

Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M.J., et al. (2019). Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Reports *29*, 1336-1350.e4. https://doi.org/10.1016/j.celrep.2019.09.062.

Eskenazi, A., Lood, C., Wubbolts, J., Hites, M., Balarjishvili, N., Leshkasheli, L., Askilashvili, L., Kvachadze, L., van Noort, V., Wagemans, J., et al. (2022). Combination

of pre-adapted bacteriophage therapy and antibiotics for treatment of fracture-related infection due to pandrug-resistant Klebsiella pneumoniae. Nat Commun *13*, 302. https://doi.org/10.1038/s41467-021-27656-z.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507. https://doi.org/10.1016/j.jmb.2006.03.043.

Fraser, J.S., Maxwell, K.L., and Davidson, A.R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr. Opin. Microbiol. *10*, 382–387. https://doi.org/10.1016/j.mib.2007.05.018.

Freed, L.E., and Vunjak-Novakovic, G. (2002). Spaceflight bioreactor studies of cells and tissues. Adv Space Biol Med *8*, 177–195. https://doi.org/10.1016/s1569-2574(02)08019-x.

Garcia, E., Elliott, J.M., Ramanculov, E., Chain, P.S.G., Chu, M.C., and Molineux, I.J. (2003). The genome sequence of Yersinia pestis bacteriophage phiA1122 reveals an intimate history with the coliphage T3 and T7 genomes. J. Bacteriol. *185*, 5248–5262. .

Garcia-Doval, C., and Raaij, M.J. van (2012). Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. PNAS *109*, 9390–9395. https://doi.org/10.1073/pnas.1119719109.

González-García, V.A., Bocanegra, R., Pulido-Cid, M., Martín-Benito, J., Cuervo, A., and Carrascosa, J.L. (2015). Characterization of the initial steps in the T7 DNA ejection process. Bacteriophage *5*. https://doi.org/10.1080/21597081.2015.1056904.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

Hatfull, G.F., Dedrick, R.M., and Schooley, R.T. (2022). Phage Therapy for Antibiotic-Resistant Bacterial Infections. Annu Rev Med *73*, 197–211. https://doi.org/10.1146/annurev-med-080219-122208.

Heineman, R.H., Springman, R., and Bull, J.J. (2008). Optimal foraging by bacteriophages through host avoidance. Am. Nat. *171*, E149-157. https://doi.org/10.1086/528962.

Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (2020). A continuous evolution system for contracting the host range of bacteriophage T7. Scientific Reports *10*, 1–8. https://doi.org/10.1038/s41598-019-57221-0.

Huss, P., and Raman, S. (2020). Engineered bacteriophages as programmable biocontrol agents. Curr Opin Biotechnol *61*, 116–121. https://doi.org/10.1016/j.copbio.2019.11.013.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. (2021). Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. ELife *10*, e63775. https://doi.org/10.7554/eLife.63775.

Joiner, K.L., Baljon, A., Barr, J., Rohwer, F., and Luque, A. (2019). Impact of bacteria motility in the encounter rates with bacteriophage in mucus. Sci Rep *9*, 16427. https://doi.org/10.1038/s41598-019-52794-2.

Kasman, L.M., and Porter, L.D. (2022). Bacteriophages. In StatPearls, (Treasure Island (FL): StatPearls Publishing), p.

Kemp, P., Garcia, L.R., and Molineux, I.J. (2005). Changes in bacteriophage T7 virion structure at the initiation of infection. Virology *340*, 307–317. https://doi.org/10.1016/j.virol.2005.06.039.

Kilcher, S., and Loessner, M.J. (2019). Engineering Bacteriophages as Versatile Biologics. Trends in Microbiology *27*, 355–367. https://doi.org/10.1016/j.tim.2018.09.006.

Klaus, D., Simske, S., Todd, P., and Stodieck, L. (1997). Investigation of space flight effects on Escherichia coli and a proposed model of underlying physical mechanisms. Microbiology (Reading) *143 (Pt 2)*, 449–455. https://doi.org/10.1099/00221287-143-2-449.

Kutter, E.M., Kuhl, S.J., and Abedon, S.T. (2015). Re-establishing a place for phage therapy in western medicine. Future Microbiol *10*, 685–688. https://doi.org/10.2217/fmb.15.28.

Lin, T.-Y., Lo, Y.-H., Tseng, P.-W., Chang, S.-F., Lin, Y.-T., and Chen, T.-S. (2012). A T3 and T7 Recombinant Phage Acquires Efficient Adsorption and a Broader Host Range. PLOS ONE *7*, e30954. https://doi.org/10.1371/journal.pone.0030954.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507.

Maguire, E., Rocca-Serra, P., Sansone, S.-A., and Chen, M. Redesigning the Sequence Logo with Glyph-based Approaches to Aid Interpretation. 5. .

Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., and Lenski, R.E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. Science *335*, 428–432. https://doi.org/10.1126/science.1214449.

Molineux, I.J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol. Microbiol. *40*, 1–8. https://doi.org/10.1046/j.1365-2958.2001.02357.x.

Nickerson, C.A., Ott, C.M., Wilson, J.W., Ramamurthy, R., and Pierson, D.L. (2004). Microbial Responses to Microgravity and Other Low-Shear Environments. Microbiol Mol Biol Rev *68*, 345–361. https://doi.org/10.1128/MMBR.68.2.345-361.2004.

Qimron, U., Marintcheva, B., Tabor, S., and Richardson, C.C. (2006). Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. Proc. Natl. Acad. Sci. U.S.A. *103*, 19039–19044. https://doi.org/10.1073/pnas.0609428103.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., et al. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrob. Agents Chemother. *61*, e00954-17. https://doi.org/10.1128/AAC.00954-17.

Todd, P. (2007). GRAVITY-DEPENDENT PHENOMENA AT THE SCALE OF THE SINGLE CELL. Gravitational and Space Research *2*. .

Yang, H., Ma, Y., Wang, Y., Yang, H., Shen, W., and Chen, X. (2014). Transcription regulation mechanisms of bacteriophages. Bioengineered *5*, 300–304. https://doi.org/10.4161/bioe.32110.

Yehl, K., Lemire, S., Yang, A.C., Ando, H., Mimee, M., Torres, M.D.T., de la Fuente-Nunez, C., and Lu, T.K. (2019). Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. Cell *179*, 459-469.e9. https://doi.org/10.1016/j.cell.2019.09.015.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620. https://doi.org/10.1093/bioinformatics/btt593.

Chapter 7

Summary and Suggestions for Future Research

Summary

Understanding the sequence-function landscape of phages reveals how phages interact with their bacterial hosts and allows us to engineer phages to increase activity and tailor host range. Phages are poised to become valuable tools for applications ranging from treating antibiotic resistant bacterial infections to precision modification of the microbiome (Dedrick et al., 2019; Hatfull et al., 2022; Huss and Raman, 2020; Mutalik and Arkin, 2022; Schooley et al., 2017). A key limitation for understanding and engineering phages has been a dearth of high-throughput methods to systematically and comprehensively profile sequence-function relationships in phages.

In **Chapter 4**, we presented a new method called ORACLE (**O**ptimized **R**ecombination, **Ac**cumulation and **L**ibrary **E**xpression) to create abundant, unbiased libraries of phages. Creating these phage libraries allows us to test more than 10^5 phages, all at once, to establish a generalized framework to describe sequence-function relationships. ORACLE overcame two key problems of library abundance and library bias that would have undermined this analysis. Our ability to create these libraries was essential for work presented in **Chapter 4** and in **Chapter 5** and **Chapter 6.**

We used ORACLE to study the tip domain of the T7 phage receptor binding protein (RBP), a key region that mediates host receptor interaction and determines phage activity (Garcia-Doval and Raaij, 2012; González-García et al., 2015; Molineux, 2001). T7 phage is an obligate lytic podovirus that infects *E. coli*. This phage tolerates engineering well, including insertion of genetic cassettes and editing gene expression (as demonstrated in **Appendix 1** and **Appendix 2)**, and is well positioned to be engineered for different roles targeting *E. coli* and related bacteria. As document in **Chapter 4**, we used ORACLE to

create a library capable of deep mutational scanning (DMS) of the tip domain to functionally profile this region in the RBP (Huss et al., 2021). We selected this library on three susceptible and two resistant *E. coli* hosts to determine the relative fitness of each variant and mapped activity-enhancing substitutions and host-specific substitutional patterns, many of which were previously unknown. Comparison of DMS results created a functional profile of the tip domain, establishing a high-resolution map of genetic changes that drive phage activity. This functional profile precisely revealed regions that are ideal engineering targets for tailoring host range and improving activity in T7 phage.

In **Chapter 5**, we leveraged this functional profile to find sequence motifs that were responsible for receptor recognition in metagenomic datasets. Increasingly available phage metagenomic datasets represent a treasure trove of sequences that can be mined to engineer new function and alter activity in phages (Abeles and Pride, 2014; Benler et al., 2021; Gu et al., 2021; Luo et al., 2022; Shkoporov and Hill, 2019; Shkoporov et al., 2019). Our functional profile of the tip domain gave us the information needed to identify short motifs relevant for phage activity, allowing us to search these datasets for such motifs without relying on overall protein homology. We created a new process called Motif Curation that uses our functional mapping to seed scores for every possible motif and found over 10^4 relevant motifs for the T7 RBP tip domain in NCBI and IMG/VR metagenomic databases using this approach. Motif Curation allowed us to identify motifs that are highly divergent from the wildtype sequence in diverse, non-homologous phage sequences. We used ORACLE to create a phage library with these motifs and screened the library on different *E. coli* hosts, revealing hundreds of phages with novel host specificity. We traced combinations of substitutions that drove activity and host range and

isolated variants that eliminate pathogenic *E. coli* causing urinary tract infections that are insensitive to wildtype T7. Motif Curation was thus able to tap into the wealth of phage metagenomic data to explain sequence-function relationships and engineer phages.

In **Chapter 6** we explored how T7 phage interacted with *E. coli* BL21 in microgravity aboard the International Space Station. Microgravity provides a unique environment for phages to interact with their hosts which can highlight new regions in the phage genome that alter phage activity and new targets for engineering the phage. We used whole genome sequencing with enormous depth to identify mutations that increased activity in this unique environment and examined how our RBP tip domain DMS library fared compared to under gravity. These experiments highlighted the value of assessing phages in different growth conditions and identified targets for engineering T7 phage.

Several avenues of research are viable future directions for the work presented in this dissertation. Using DMS to functionally map the remainder of the RBP in T7 and other key genes is a natural extension of our work here. Expanding ORACLE to other phages and other genes will increase our body of knowledge for engineering phages. Several approaches could continue to leverage our functional mapping of the tip domain, including using these results to train machine learning models and creating new combinatorial libraries of mutants to parallel our work in Chapter 6. There are also several technical aspects of ORACLE that can readily be improved. Most excitingly, our approach for functionally characterizing and engineering T7 phages has shown promising results on two clinically relevant strains, *E. coli* serogroup O121 and multidrug resistant Enterobacter cloacae. Suggestions for these lines of future research are expanded upon below.

Suggestions for Future Research

Future directions for DMS Screening

While the tip domain of the RBP is enormously important for phage activity, we should expand our DMS analysis to the remainder of the RBP and to other genes in T7. Single amino acid substitutions in other regions earlier in the RBP and other structural genes, particularly in the tail nozzle and other tail proteins, have been shown to change the activity of T7 phage (Cuervo et al., 2019; Holtzman et al., 2020). These regions and other genes in T7 were also mutated in our microgravity study in **Chapter 6**, emphasizing their potential value.

One limitation of our original approach for deep sequencing and scoring the DMS library is the length of the sequencing read. While the ~600 bp afforded by short read sequencing is sufficient for the tip domain and other small domains, this approach is not workable for larger proteins. A viable alternative is incorporation of a short barcode alongside the variant sequence, then mapping the variant to the barcode using long read sequencing platforms such as PacBio or Oxford Nanopore (Hu et al., 2021). The barcode region can then be sequenced using (more affordable) short read sequencing platforms. To facilitate barcode incorporation and mapping I have already redesigned the recombination plasmid originally used in ORACLE to incorporate these barcodes.

Using this new approach, I created, but have not yet screened DMS libraries for the entirety of the RBP, for the tail nozzle, and for relatively uncharacterized structural genes *gp6.7, gp7* and *gp7.3*. These uncharacterized genes are related to the tail structure and *gp6.7* and *gp7.3* and may be ejected into the bacterial host upon infection (Kemp et al., 2005; Studier, 1975). These small proteins were noted to affect adsorption and plaque size, and several mutations in *gp7.3* were apparent in our microgravity study. DMS of

these proteins could improve our understanding of their role in T7 and reveal if engineering these proteins is a viable approach for tailoring activity.

Another valuable avenue of research is expanding ORACLE and DMS screens to other phages. While T7 is a useful chassis phage for targeting *E. coli* and other closely related strains, establishing this process in other phages will improve the flexibility of our engineering approach. Phages such as T3 or K1 phages are closely related to T7 and would require minimal effort to assay with ORACLE. K1 and K1-related phages contain depolymerases which are ideally suited for targeted clinically relevant strains that produce biofilms, making them ideal targets to begin expanding the panel of functionally characterized phage (Leiman et al., 2007; Lin et al., 2017; Scholl and Merril, 2005).

More approaches for leveraging DMS data for complex combinatorial libraries

Our approach for Motif Curation outlined in **Chapter 5** was an effective way to create complex combinatorial mutants using short motifs found in metagenomic databases. However, this is not the only approach that can leverage our DMS data to create combinatorial mutants. A more direct approach is combining functional mutants directly from the DMS results. Several variations of this scheme are feasible. First, combining substitutions that were most directly responsible for changes in phage activity across the tip domain. These substitutions can be selected so every variant contains substitutions that impacted function at different positions in the tip domain, potentially greatly enhancing the variants activity or host specificity. I have created one such library of phages that include variants with two or three of the most important substitutions across

the tip domain (Figure 1). This library has been screened on the five-member host panel used to in the original DMS assay and results are being analyzed.

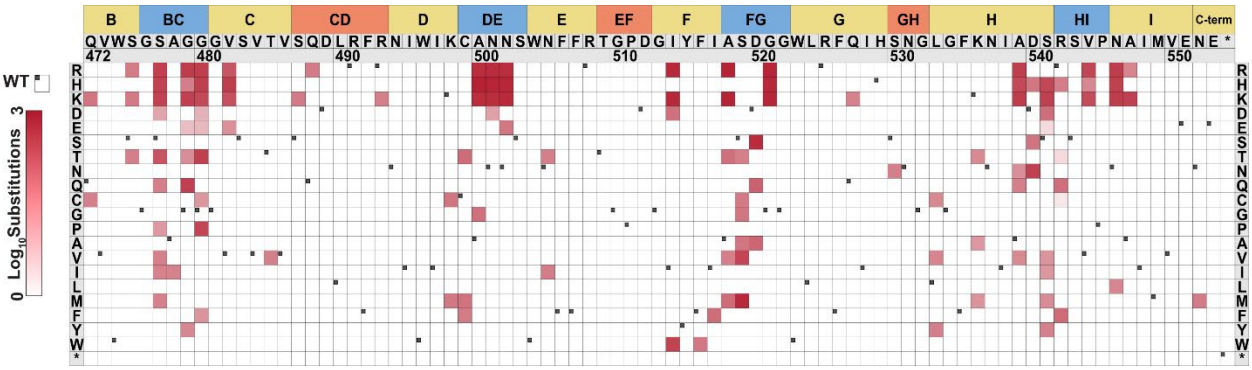


Figure 1. Heat Map of Combinatorial Substitutions

(B) Heat map showing the number of substitutions (log_{10}, red gradient) made in the combinatorial phage library. Substitutions are shown top to bottom while position (based on PDB 4A0T), wild type amino acid and secondary structure topology is shown left to right.

Another approach is targeting specific hosts and excluding others by selecting substitutions that do well on one host but poorly on another. By installing several such substitutions across the tip domain creating engineered variants with user-specified host range becomes straightforward. Several libraries of phage variants with up to 6 substitutions have been made to specifically target *E. coli* 10G, BL21, or BW25513 as a proof of concept, and this testing is in process. These and other combinatorial processes can also feed into machine learning models to identify non-obvious specific traits that and combinatorial variants that influence host range.

Targeting Pathogenic *E. coli* using engineered T7 phage

An exciting extension of this work is creating engineered T7 phages that can specifically target and eliminate clinically relevant *E. coli* strains. Preliminary experiments in several strains have shown promising results for T7 DMS and combinatorial libraries. One such strain is *E. coli* serogroup O121. This is a Shiga toxin-producing *E. coli* (STEC) that is one of the major non-O157:H7 serotypes frequently responsible for foodborne outbreaks (Caprioli et al., 2014; Tarr et al., 2005). Treatment options for STEC are limited as Shiga toxin production is triggered by the bacterial SOS response, hampering treatment with most antibiotics (Kakoullis et al., 2019; Mühlen and Dersch, 2020). Using phages to target and eliminate STEC could provide an opportunity to inhibit Shiga toxin production at the source by either directly interfering with protein production or by delivering Shiga-toxin neutralizing agents such as peptides or shiga-receptor mimics (Bernedo-Navarro et al., 2014; Hostetter et al., 2014). Despite wildtype T7 having limited activity, both T7 DMS and combinatorial libraries have shown promising results on this strain.

First, we applied the T7 RBP tip domain DMS library to this strain to evaluate if any substitutions in the tip domain could improve phage activity. We created three versions of the library and selected each three times using an MOI of 0.01, then deep sequenced the libraries as in **Chapter 4** (Figure 2). Analysis of these results showed dramatic selection in the tip domain that was far more specific than we had previously seen on susceptible or insensitive hosts. Variants with negatively charge substitutions were

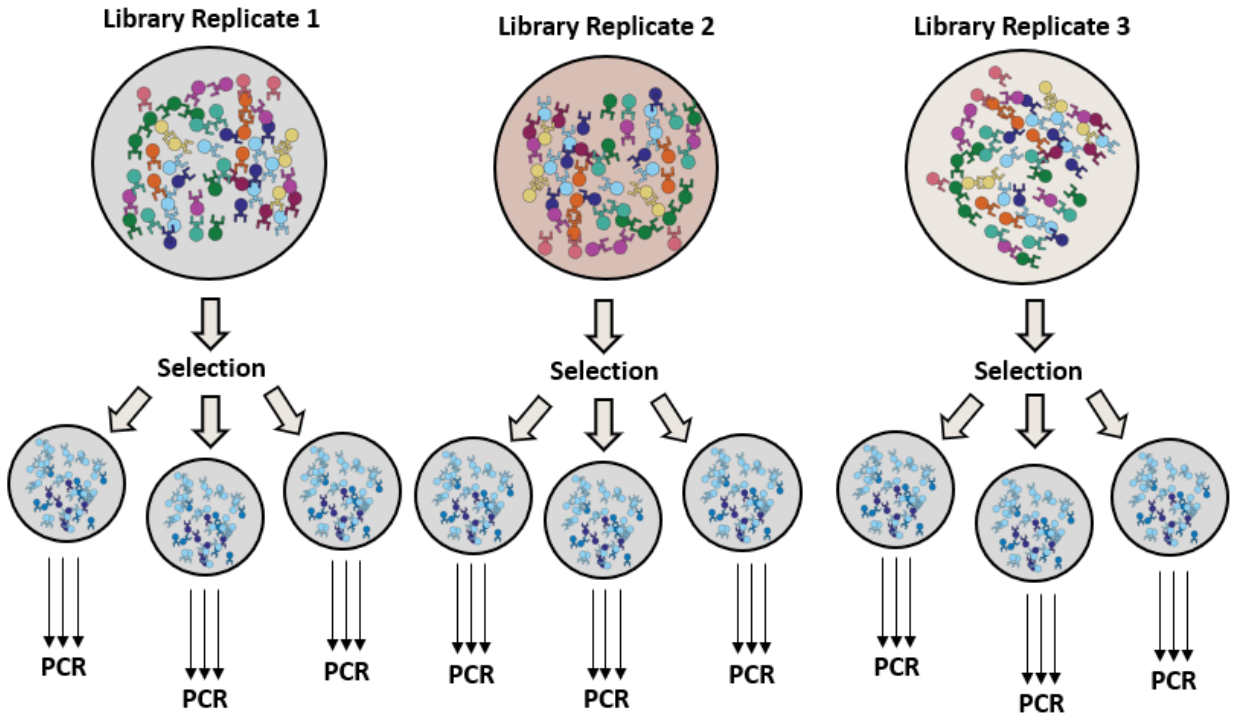


Figure 2. *E. coli*, Serogroup O121 Selection Scheme

Selection scheme for E. coli, Serogroup O121. The DMS library is made in triplicate and each library is passaged three times before being deep sequenced and scored.

overwhelmingly selected for, primarily in interior loops or positions facing away from expected host receptor interface (Figure 3 and Figure 4). These results painted a very different functional profile for the tip domain for this strain. We postulated these substitutions affecting the secondary structure of the RBP in some way to facilitate receptor binding in the host and are perhaps necessary to bind to hosts with O-antigens, which T7 does not normally infect. We are currently developing combinatorial libraries that use these mutations as ‘gate-keepers’ to see if these mutations in combination with other substitutions in the tip domain further increase activity on this strain.

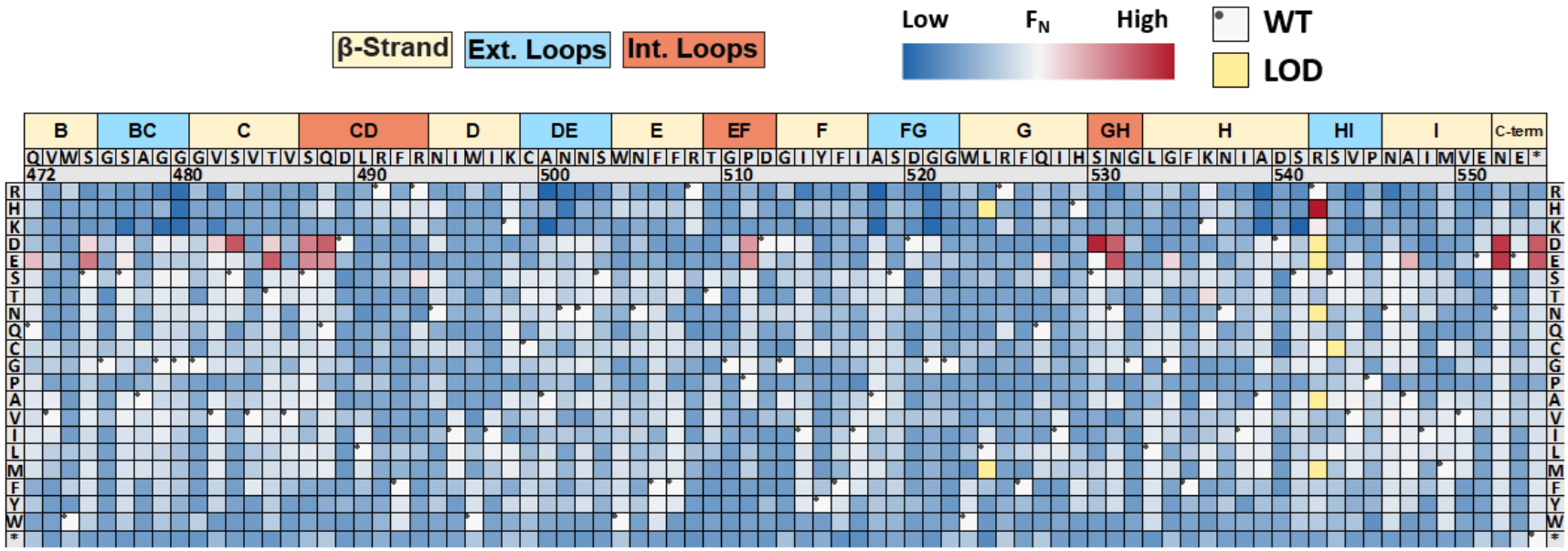

β-Strand
Ext. Loops
Int. Loops
Low
F_N
High
WT
LOD
B
BC
C
CD
D
DE
E
EF
F
FG
G
GH
H
HI
I
C-term

Figure 3. Deep mutational scanning shows specific preference for negatively charged substitutions

Heatmap showing normalized functional scores (F_N) of all substitutions (blue to red gradient) and wildtype amino acid (F_N=1 and black dot upper left) at every position after selection on *E. coli* serotype O121. Residue numbering (based on PDB 4A0T), wildtype amino acid and secondary structure topology are shown above left to right, substitutions listed top to bottom.

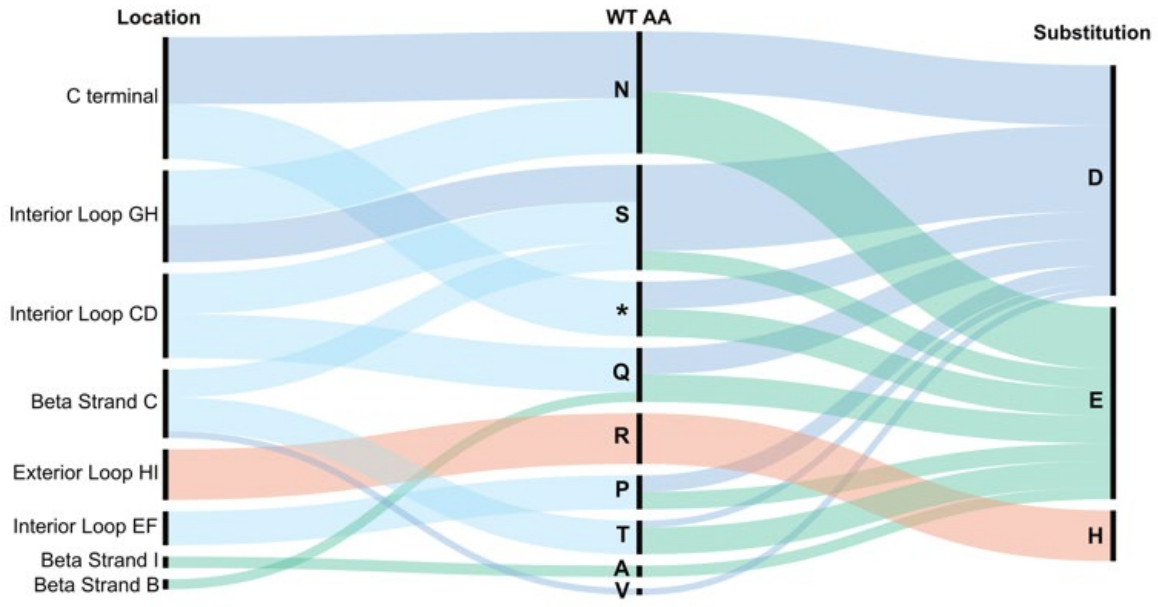


Figure 4. Alluvial plot for enriched variants after selection on *E. coli* O121

Alluvial plot showing enriched variants ($F_N \geq 1$) and topological location (left), wildtype amino acid (middle), and substitution (right) on *E. coli* O121. Each alluvial colored based on the substituted amino acid and scaled by F_N.

We have also been intrigued by results passaging our 2-3 combinatorial libraries on multidrug resistant *Enterobacter cloacae*. *E. cloacae* is global threat that has demonstrated resistance to an enormous number of antibiotics, making phage therapy a suitable alternative strategy (Annavajhala et al., 2019; Davin-Regli and Pagès, 2015). The particular strain of *E. cloacae* that we evaluated our T7 libraries against is resistant to penicillins, carbapenems, quinolones, aminoglycosides, and several last-line antibiotics including ceftazidime-avibactam, meropenem-vaborbactam, ceftolozane-tazobactam,

and imipenem-relebactam. This strain is partially susceptible to only two known antibiotics, colistin and cefiderocol, both last-line antibiotics. Finding alternative treatment options for this strain is therefore essential. Initial screens have shown that wildtype T7 has little to no activity on this strain, but our combinatorial library was able to lyse the strain initially before the strain became insensitive after 8 hours (Figure 5). We selected phages from this initial passage and combined them into three phage pools, all of which were capable of lysing the strain without insensitivity arises. These results are promising leads, and we are investigating which combinations of substitutions are most effective. One can envision how powerful this approach could be after functionally characterizing more genes in T7 to create highly active phages against this infectious strain.

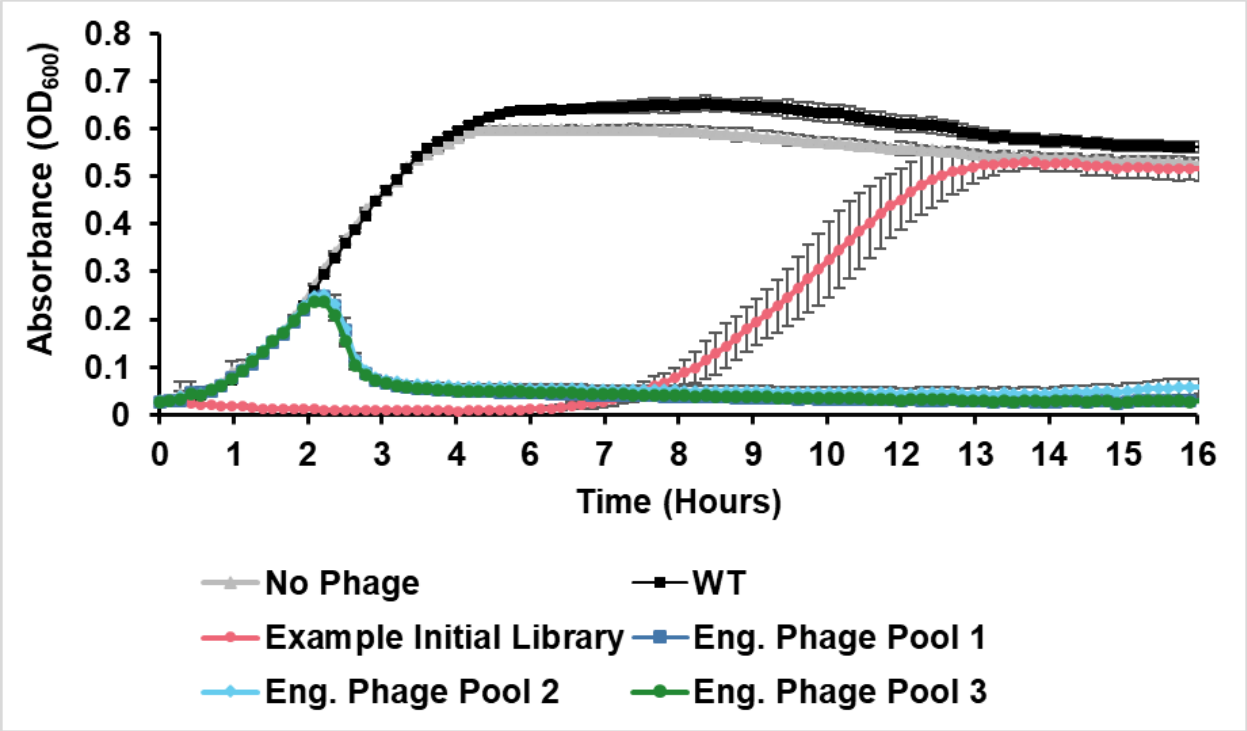


Figure 5. Growth time course for *Enterobacter cloacae* and T7 phages

Growth time course of *E. cloacae* subject to no phage (grey), wildtype T7 (black), combinatorial library (pink) and selected variants (blue, light blue, and green).All phages are applied at time zero. Wildtype and initial combinatorial library are applied at an MOI of 10, selected variants are applied an MOI of ~10^{-2}.

Improving Recombination Efficiency in ORACLE

ORACLE is the technical foundation upon which the entirety of our results are made possible. Many of the variants critical for function would have otherwise been lost to library bias, and any improvements to ORACLE should be kept in mind as a valuable future direction. One approach for improving ORACLE is increasing recombination efficiency. Recombination in ORACLE currently occurs at a conservative rate of ~10^{-3} recombined phages/host passage. Improving this rate will increase the library size that can be easily tested and decrease rounds of accumulation required.

Recombination efficiency may be improving in ORACLE by manipulating the Cre-*loxP* sites used to mediate recombination. Cre is a site-specific recombinase which is reliant on two 34 bp *loxP* elements flanking the sequence that will be recombined. These *loxP* sites consist of two 13 bp inverted repeat 'arms' which themselves flank an 8 bp spacer region. The T7 acceptor phages, pHT7Rec1 and pHT7Rec2 plasmid libraries are all designed with lox66/m2/lox71 mutant sequences (see Figure 6). One spacer has a mutation designated m2 and different arms have mutations designated lox66 and lox71 on the acceptor phage and the donor plasmid.

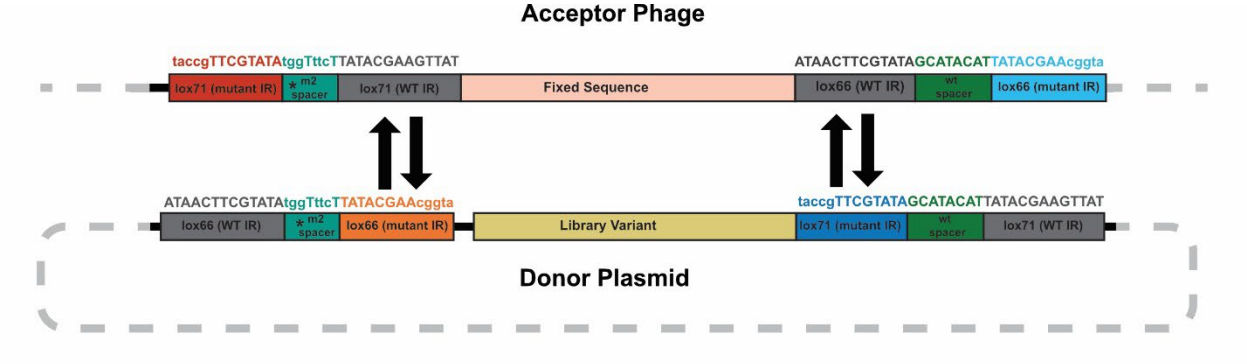


Figure 6. LoxP sites responsible for recombination

Schematic illustration of sequence rearrangements in acceptor phage and donor plasmid during recombinase-mediated cassette exchange (RMCE). Mutant sequences are shown lower case. Recombination shifts inner lox66 and lox71 inverted repeats (IR) and library variant into the acceptor phage.

The purpose of the mutant m2 spacer is to prevents intramolecular excision of DNA and confers directionality to the recombination event. Recombination only occurs between homologous spacers and having the mutated m2 spacer on one flank of the DNA to be recombined and a wildtype spacer on the other flank forces intermolecular recombination (Albert et al., 1995; Hoess et al., 1986; Langer et al., 2002). The lox66 and lox71 mutations in the arms of the *loxP* site additively weaken affinity for Cre recombinase and largely prevent recombination when combined. Because both mutations are present after recombination in the phage sequence, Cre can no longer viably interact with recombined phages. The result is one way recombinase mediated cassette exchange (RMCE), preventing a reverse reaction from occurring to swap the variant out of the phage genome.

These mutations were not originally designed for recombination in phages and were isolated without modern sequencing technology that can effectively query larger sequence spaces, leaving room for improvement for ORACLE.

One-way RMCE is not ideal for creating phage libraries, and two-way RMCE is likely a superior approach which can improve recombination efficiency. One-way RMCE was originally designed to ensure unidirectionality in recombination to accumulate recombined product, but in doing so the lox66 and lox71 mutations reduce the rate of recombination (Albert et al., 1995). This approach is ideal when recombination plasmids are retained in bacterial or mammalian cells over long periods of time. In contrast, recombination in ORACLE occurs inside a ~20-minute period while the phage is actively killing the host cell. Any increase in the speed of recombination would greatly facilitate overall recombination efficiency. Furthermore, the majority of two-way RMCE interactions are likely to be net neutral instead of deleterious if plasmid copy number is kept high as recombined phages are more likely to swap in the same library member. Investigation of wild type *loxP* arms to facilitate recombination is therefore a potentially valuable future direction.

Evaluating a large set of mutant *loxP* arms to determine effects on recombination rate is also feasible if the sequence space queried is restricted to regions known to be essential for recombination. The best regions for this experiment are likely the 5’ and 3’ ends of the *loxP* site. Using this approach, the entire sequence space could be restricted to 10^4 or 10^5 variants. The efficiency of recombination for different variants could be assessed by inserting wildtype *gp17* into T7 acceptor phages in a similar experiment to

the recombination rate assay presented in **Chapter 2**. Variants of the m2 spacer could be assessed in a similar manner to determine impact on recombination efficiency. In addition to improving ORACLE, these results would provide footing for greater understanding of how sequence changes influence Cre recombination.

Improving Library Skew in ORACLE

Another improvement to ORACLE would be increasing the effectiveness of the helper plasmid used to prevent library bias. While ORACLE largely prevents losing library members to bias, some skew does occur as the recombined variant is still expressed in the host cell. Attenuating this skew would help create more balanced libraries that are easier to screen. The helper plasmid used in ORACLE, pHT7Helper1, can easily be improved by increasing the amount of wildtype protein provided *in trans* and creating an inducible system for expression. Increasing the amount of wildtype protein decreases the impact of any variant protein expressed from the phage that could contribute to library bias.

The current system is constructed on a mid-copy pBR plasmid. Simply increasing the copy number of this plasmid has a measurable impact on library skew in ORACLE (Figure 7). Using a high copy number pUC backbone (pHT7Helper11) during Accumulation had a significant impact on leveling the library distribution after expression. For comparison the most abundant library members were 5.3%, 4%, and 3.5% abundant using pHT7Helper1, whereas the most abundant members using pHT7Helper11 were 0.71%, 0.70%, and 0.66% abundant after expression. Another approach would be

inserting a repressor into the phage genome to reduce variant expression, although this would require additional engineering of the phage. Finally, the helper plasmid could be converted to an inducible system. An inducible system would facilitate using ORACLE with phage proteins that are deleterious to the host and cannot be expressed constitutively. These improvements would increase the flexibility of ORACLE and prove useful for ensuring adequate coverage as library sizes increase.

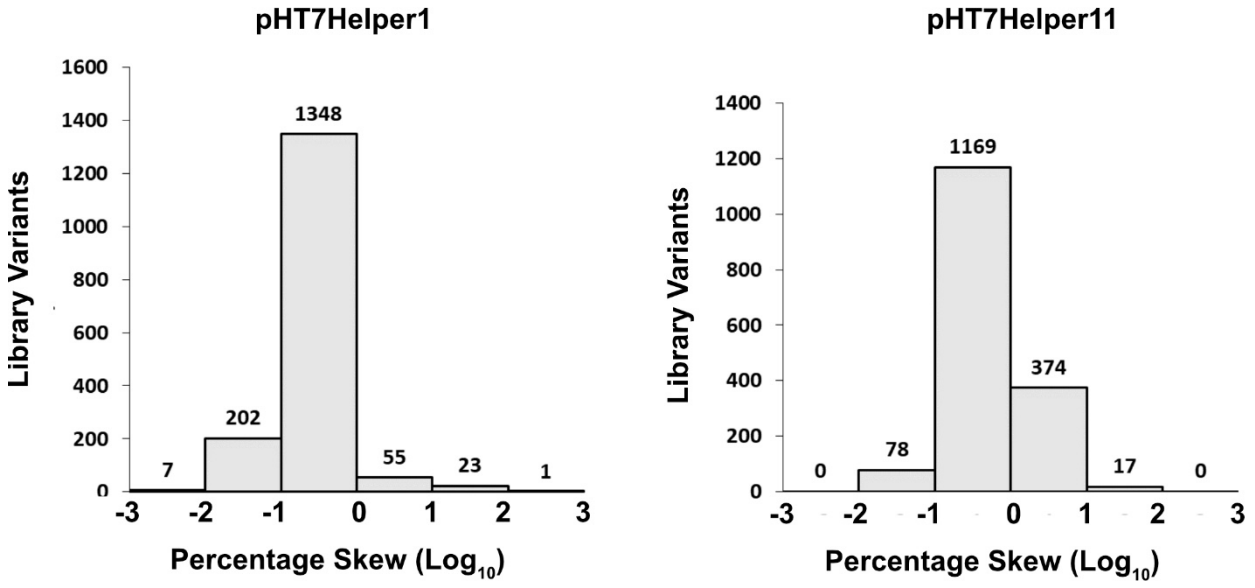


Figure 7. Library Skew using different copy number helper plasmids

Histogram of abundance of variants in the expression library using pHT7Helper1 during accumulation (left) or pHT7Helper11 during accumulation (right) binned using log proportion centered on equal representation. The DMS library in **Chapter 2** was used for this experiment.

References

Abeles, S.R., and Pride, D.T. (2014). Molecular Bases and Role of Viruses in the Human Microbiome. J Mol Biol *426*, 3892–3906. https://doi.org/10.1016/j.jmb.2014.07.002.

Albert, H., Dale, E.C., Lee, E., and Ow, D.W. (1995). Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. Plant J. *7*, 649–659. .

Annavajhala, M.K., Gomez-Simmonds, A., and Uhlemann, A.-C. (2019). Multidrug-Resistant Enterobacter cloacae Complex Emerging as a Global, Diversifying Threat. Frontiers in Microbiology *10*. .

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A.B., Pevzner, P., and Koonin, E.V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome *9*, 78. https://doi.org/10.1186/s40168-021-01017-w.

Bernedo-Navarro, R.A., Miyachiro, M.M., da Silva, M.J., Reis, C.F., Conceição, R.A., Gatti, M.S.V., and Yano, T. (2014). Peptides derived from phage display libraries as potential neutralizers of Shiga toxin-induced cytotoxicity in vitro and in vivo. J Appl Microbiol *116*, 1322–1333. https://doi.org/10.1111/jam.12451.

Caprioli, A., Scavia, G., and Morabito, S. (2014). Public Health Microbiology of Shiga Toxin-Producing Escherichia coli. Microbiol Spectr *2*. https://doi.org/10.1128/microbiolspec.EHEC-0014-2013.

Cuervo, A., Fàbrega-Ferrer, M., Machón, C., Conesa, J.J., Fernández, F.J., Pérez-Luque, R., Pérez-Ruiz, M., Pous, J., Vega, M.C., Carrascosa, J.L., et al. (2019). Structures of T7 bacteriophage portal and tail suggest a viral DNA retention and ejection mechanism. Nat Commun *10*, 3746. https://doi.org/10.1038/s41467-019-11705-9.

Davin-Regli, A., and Pagès, J.-M. (2015). Enterobacter aerogenes and Enterobacter cloacae; versatile bacterial pathogens confronting antibiotic treatment. Frontiers in Microbiology *6*. .

Dedrick, R.M., Guerrero-Bustamante, C.A., Garlena, R.A., Russell, D.A., Ford, K., Harris, K., Gilmour, K.C., Soothill, J., Jacobs-Sera, D., Schooley, R.T., et al. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nature Medicine *25*, 730. https://doi.org/10.1038/s41591-019-0437-z.

Garcia-Doval, C., and Raaij, M.J. van (2012). Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. PNAS *109*, 9390–9395. https://doi.org/10.1073/pnas.1119719109.

González-García, V.A., Bocanegra, R., Pulido-Cid, M., Martín-Benito, J., Cuervo, A., and Carrascosa, J.L. (2015). Characterization of the initial steps in the T7 DNA ejection process. Bacteriophage *5*. https://doi.org/10.1080/21597081.2015.1056904.

Gu, C., Liang, Y., Li, J., Shao, H., Jiang, Y., Zhou, X., Gao, C., Li, X., Zhang, W., Guo, C., et al. (2021). Saline lakes on the Qinghai-Tibet Plateau harbor unique viral assemblages mediating microbial environmental adaption. IScience *24*, 103439. https://doi.org/10.1016/j.isci.2021.103439.

Hatfull, G.F., Dedrick, R.M., and Schooley, R.T. (2022). Phage Therapy for Antibiotic-Resistant Bacterial Infections. Annu Rev Med *73*, 197–211. https://doi.org/10.1146/annurev-med-080219-122208.

Hoess, R.H., Wierzbicki, A., and Abremski, K. (1986). The role of the loxP spacer region in P1 site-specific recombination. Nucleic Acids Res *14*, 2287–2300. .

Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (2020). A continuous evolution system for contracting the host range of bacteriophage T7. Scientific Reports *10*, 1–8. https://doi.org/10.1038/s41598-019-57221-0.

Hostetter, S.J., Helgerson, A.F., Paton, J.C., Paton, A.W., and Cornick, N.A. (2014). Therapeutic use of a receptor mimic probiotic reduces intestinal Shiga toxin levels in a piglet model of hemolytic uremic syndrome. BMC Res Notes *7*, 331. https://doi.org/10.1186/1756-0500-7-331.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. Human Immunology *82*, 801–811. https://doi.org/10.1016/j.humimm.2021.02.012.

Huss, P., and Raman, S. (2020). Engineered bacteriophages as programmable biocontrol agents. Curr Opin Biotechnol *61*, 116–121. https://doi.org/10.1016/j.copbio.2019.11.013.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. (2021). Mapping the functional landscape of the receptor binding domain of T7 bacteriophage by deep mutational scanning. ELife *10*, e63775. https://doi.org/10.7554/eLife.63775.

Kakoullis, L., Papachristodoulou, E., Chra, P., and Panos, G. (2019). Shiga toxin-induced haemolytic uraemic syndrome and the role of antibiotics: a global overview. J Infect *79*, 75–94. https://doi.org/10.1016/j.jinf.2019.05.018.

Kemp, P., Garcia, L.R., and Molineux, I.J. (2005). Changes in bacteriophage T7 virion structure at the initiation of infection. Virology *340*, 307–317. https://doi.org/10.1016/j.virol.2005.06.039.

Langer, S.J., Ghafoori, A.P., Byrd, M., and Leinwand, L. (2002). A genetic screen identifies novel non-compatible loxP sites. Nucleic Acids Res *30*, 3067–3077. .

Leiman, P.G., Battisti, A.J., Bowman, V.D., Stummeyer, K., Mühlenhoff, M., Gerardy-Schahn, R., Scholl, D., and Molineux, I.J. (2007). The Structures of Bacteriophages K1E and K1-5 Explain Processive Degradation of Polysaccharide Capsules and Evolution of New Host Specificities. Journal of Molecular Biology *371*, 836–849. https://doi.org/10.1016/j.jmb.2007.05.083.

Lin, H., Paff, M.L., Molineux, I.J., and Bull, J.J. (2017). Therapeutic Application of Phage Capsule Depolymerases against K1, K5, and K30 Capsulated E. coli in Mice. Front Microbiol *8*. https://doi.org/10.3389/fmicb.2017.02257.

Luo, E., Leu, A.O., Eppley, J.M., Karl, D.M., and DeLong, E.F. (2022). Diversity and origins of bacterial and archaeal viruses on sinking particles reaching the abyssal ocean. ISME J *16*, 1627–1635. https://doi.org/10.1038/s41396-022-01202-1.

Molineux, I.J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol. Microbiol. *40*, 1–8. https://doi.org/10.1046/j.1365-2958.2001.02357.x.

Mühlen, S., and Dersch, P. (2020). Treatment Strategies for Infections With Shiga Toxin-Producing Escherichia coli. Front Cell Infect Microbiol *10*, 169. https://doi.org/10.3389/fcimb.2020.00169.

Mutalik, V.K., and Arkin, A.P. (2022). A Phage Foundry Framework to Systematically Develop Viral Countermeasures to Combat Antibiotic-Resistant Bacterial Pathogens. IScience *25*, 104121. https://doi.org/10.1016/j.isci.2022.104121.

Scholl, D., and Merril, C. (2005). The Genome of Bacteriophage K1F, a T7-Like Phage That Has Acquired the Ability To Replicate on K1 Strains of Escherichia coli. J Bacteriol *187*, 8499–8503. https://doi.org/10.1128/JB.187.24.8499-8503.2005.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., et al. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrob. Agents Chemother. *61*, e00954-17. https://doi.org/10.1128/AAC.00954-17.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host & Microbe *26*, 527-541.e5. https://doi.org/10.1016/j.chom.2019.09.009.

Studier, F.W. (1975). Gene 0.3 of bacteriophage T7 acts to overcome the DNA restriction system of the host. Journal of Molecular Biology *94*, 283–295. https://doi.org/10.1016/0022-2836(75)90083-2.

Tarr, P.I., Gordon, C.A., and Chandler, W.L. (2005). Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. Lancet *365*, 1073–1086. https://doi.org/10.1016/S0140-6736(05)71144-2.

Appendix 1

Publication

Virus-associated organosulfur metabolism in human and environmental systems

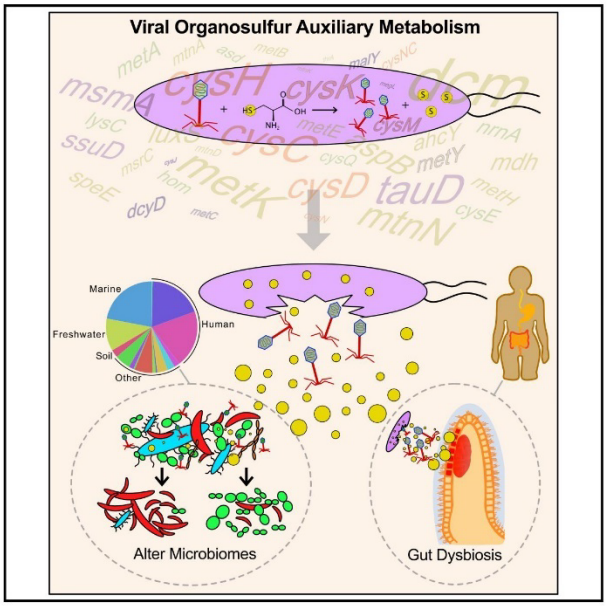
Reprinted from Cell Rep. 2021 Aug 3;36(5):109471. Article link:
https://doi.org/10.1016/j.celrep.2021.109471

Cell Reports

Article

Virus-associated organosulfur metabolism in human and environmental systems

Graphical abstract



Highlights

- Thousands of diverse viruses encode genes that manipulate organic sulfur metabolism
- Infection in the presence of sulfide increases bacteriophage production
- Engineered phage T7 retains cysteine synthase (*cysK*) over multiple generations
- These viruses can influence human gut dysbiosis, microbiomes, and biogeochemistry

Authors

Kristopher Kieft, Adam M. Breister, Phil Huss, ..., Srivatsan Raman, Simon Roux, Karthik Anantharaman

Correspondence

karthik@bact.wisc.edu

In brief

Kieft et al. identify thousands of viruses that encode genes for organosulfur metabolism. They provide evidence for potential benefits of these genes to viruses, namely the role of organosulfur-derived hydrogen sulfide in increasing viral fitness. Their results have implications in exacerbation of human gut dysbiosis and global biogeochemical sulfur cycling.

Kieft et al., 2021, Cell Reports *36*, 109471
August 3, 2021
https://doi.org/10.1016/j.celrep.2021.109471

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Virus-associated organosulfur metabolism in human and environmental systems

Kristopher Kieft,[1] Adam M. Breister,[1] Phil Huss,[1,2] Alexandra M. Linz,[3] Elizabeth Zanetakos,[1] Zhichao Zhou,[1] Janina Rahlff,[4,6] Sarah P. Esser,[4] Alexander J. Probst,[4] Srivatsan Raman,[1,2] Simon Roux,[5] and Karthik Anantharaman[1,7,*]

[1]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA
[3]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI, USA
[4]Department of Chemistry, Environmental Microbiology and Biotechnology, University of Duisburg-Essen, Essen, Germany
[5]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[6]Present address: Centre for Ecology and Evolution in Microbial Model Systems, Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden
[7]Lead contact
*Correspondence: karthik@bact.wisc.edu
https://doi.org/10.1016/j.celrep.2021.109471

SUMMARY

Viruses influence the fate of nutrients and human health by killing microorganisms and altering metabolic processes. Organosulfur metabolism and biologically derived hydrogen sulfide play dynamic roles in manifestation of diseases, infrastructure degradation, and essential biological processes. Although microbial organosulfur metabolism is well studied, the role of viruses in organosulfur metabolism is unknown. Here, we report the discovery of 39 gene families involved in organosulfur metabolism encoded by 3,749 viruses from diverse ecosystems, including human microbiomes. The viruses infect organisms from all three domains of life. Six gene families encode for enzymes that degrade organosulfur compounds into sulfide, whereas others manipulate organosulfur compounds and may influence sulfide production. We show that viral metabolic genes encode key enzymatic domains, are translated into protein, and are maintained after recombination, and sulfide provides a fitness advantage to viruses. Our results reveal viruses as drivers of organosulfur metabolism with important implications for human and environmental health.

INTRODUCTION

Biological sulfur cycling is one of the oldest and most influential biochemical processes on Earth and is primarily driven by microbial reduction of sulfate to produce hydrogen sulfide (Andreae, 1990; Fike et al., 2015; Wacey et al., 2011). Sulfide plays dynamic roles in the degradation of infrastructure and souring of oil reserves (Ma et al., 2000; Voordouw et al., 1996), microbial respiration and essential biosynthesis processes, and manifestation of human gastrointestinal disorders such as colitis, inflammatory bowel diseases (IBD), and colorectal cancer (CRC) (Guo et al., 2016). Much of our knowledge of sulfur cycling focuses on a small subset of microbes that are capable of respiring and transforming inorganic sulfur compounds, a process known as dissimilatory metabolism (Anantharaman et al., 2018). Consequently, the cycling of sulfur-containing organic (organosulfur) compounds and resulting sulfide production from more widespread biological mechanisms and sources has largely been ignored.

Two mechanisms of sulfide production include the degradation of organosulfur compounds and assimilatory sulfur metabolism. Sulfide production from microbial-driven degradation of organosulfur compounds, such as the amino acid cysteine, has been noted as a significant contributor to sulfide concentrations in environmental and human systems (Carbonero et al., 2012; Morra and Dick, 1991; Xia et al., 2017). However, there exists no comprehensive analysis of the specific microbes involved. Assimilatory sulfur metabolism, a common strategy used by many microbes and some eukaryotes to incorporate sulfide into biological compounds, has similarly been routinely discounted as a mechanism of significant sulfide release into either environmental or human systems. Notably, the role of viruses in these processes has not been explored.

Microbial viruses, mainly comprising bacteriophages (phages) are extraordinarily abundant on Earth. Microbial viruses are known to redirect and recycle nutrients on the scale of ecosystems by infecting and lysing host cells (Gobler et al., 1997; Jiao et al., 2010; Jover et al., 2014; Wilhelm and Suttle, 1999). In the oceans alone, the number of viral infections per second exceeds the number of stars in the known universe, which likely leads to the lysis of over 20% of all microbes per day (Manojlović, 2015; Suttle, 2007). In addition to lysis, viruses can actively redirect host metabolism during infection that manipulates major biogeochemical cycles, including carbon, nitrogen, and sulfur. One such mechanism involves viruses "stealing" metabolic genes from their host in order to gain fitness advantages during infection (Sullivan et al., 2006). Such host-derived viral genes are termed auxiliary metabolic genes (AMGs) and are expressed

CellPress
OPEN ACCESS

Cell Reports
Article

during infection to modulate microbial respiration, biosynthesis processes, and/or direct intracellular nutrients toward virus replication and virion production (Anantharaman et al., 2014; Breitbart et al., 2007; Hurwitz et al., 2013, 2015; Kieft et al., 2021; Mann et al., 2003; Roux et al., 2014; Suttle, 2005; Thompson et al., 2011; Trubl et al., 2018). For example, some viruses of Cyanobacteria encode core photosystem proteins that augment host metabolism in order to increase the biosynthesis of dNTPs that are utilized for viral genome replication (Thompson et al., 2011). The viral auxiliary metabolism of iron-sulfur clusters, central carbon metabolism, nitrification, methane oxidation, and other metabolic processes could also provide viruses with a multi-faceted method of manipulating nutrients within their host cell to enable efficient, rapid, or otherwise a more improved viral replication cycle (Ahlgren et al., 2019; Chen et al., 2020; Hurwitz and U'Ren, 2016; Hurwitz et al., 2015).

In spite of the importance and global prevalence of viruses, nothing is known about their contribution and impact on AMG-driven organosulfur metabolism in the environment. Moreover, the role of AMGs in human microbiomes has been largely unexplored. Here, we investigated environmental and human microbiomes for the presence of viruses involved in production of hydrogen sulfide and manipulation of organosulfur metabolism. By screening publicly available partial and complete viral genomes from cultivated and uncultivated viruses, we identified genes involved in direct and indirect sulfide production from organosulfur degradation and assimilatory sulfur metabolism. We followed this up with experiments to validate the impacts of genes for organosulfur metabolism as well as hydrogen sulfide on viral fitness.

RESULTS

Metabolic pathways for organosulfur metabolism driven by viral AMGs

We queried a comprehensive dataset of ~135,000 partial and complete viral genomes (contigs) publicly available on the Integrated Microbial Genomes/Viruses (IMG/VR) database (Paez-Espino et al., 2016, 2017) and the National Center for Biotechnology Information (NCBI) database, and two metagenomic studies from Lake Mendota, WI (Linz et al., 2018), for the presence of virally encoded proteins for organosulfur metabolism. In total, we identified 4,103 viral AMGs representative of 39 unique gene families. All genes identified are categorized as class I AMGs, or those for central metabolic functions but auxiliary to productive viral infection (Hurwitz and U'Ren, 2016). These AMGs were detected on 3,749 non-redundant viral genomes from all major bacterial double-stranded DNA (dsDNA) viral families (*Myoviridae*, *Podoviridae*, and *Siphoviridae*) including viruses infecting an archaea (Rahlff et al., 2020) and eukaryote (amoeba) (Schulz et al., 2020). Therefore, AMGs for organosulfur metabolism were identified on viruses infecting all three domains of life, representing a shared metabolic constraint regardless of host domain. The viruses represent cultivated and uncultivated viruses, linear and circular genomes, and lytic and lysogenic cycles of viral replication across a vast range of environmental and human microbiomes. Of these viruses, 164 have been isolated and cultivated on hosts spanning nine major bacterial lineages (Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Cyanobacteria, Actinobacteria, Firmicutes, Bacteroidetes, Verrucomicrobia, and Deinococcus-Thermus) as well as an amoeba (*Vermamoeba vermiformis*) (Table S1). The isolation of viruses encoding organosulfur metabolism AMGs indicates that the identification of such viral driven metabolism is not an artifact of metagenomic analysis.

Viral AMGs are putatively associated with five distinct processes: sulfide production from organic sulfur, the assimilatory sulfate reduction pathway, sulfite production from organic sulfur, metabolism of organic sulfur, and sulfur-related amino acid metabolism (Figure 1; Table 1). Six different AMG families (*cysK*, *cysM*, *malY*, *dcyD*, *metC*, and *metY*) encode for enzymes able to directly produce sulfide from the degradation of cysteine and homocysteine, which are important organosulfur compounds and central sources of sulfur in the environment and human body (Chiku et al., 2009; Fitzgerald, 1976). Six other AMG families (*cysD*, *cysN*, *cysC*, bifunctional-*cysNC*, *cysH*, and *cysJ*) are components of the assimilatory sulfate reduction pathway, which is widely utilized across all three domains of life for incorporation of sulfide into cysteine. Sulfite can be directly produced from the breakdown of several organosulfur compounds (e.g., taurine) by three families of AMGs (*tauD*, *ssuD*, and *msmA*) and successively fed into dissimilatory and assimilatory sulfate reduction. Eleven of the AMG families (*aspB*, *metB*, *metH*, *metE*, *msrC*, *metK*, *megL*, *dcm*, *mtnN*, *ahcY*, and *luxS*) are inferred to indirectly produce sulfide by manipulating abundant organosulfur compounds (e.g., methionine and cystathionine) that funnel into the synthesis of cysteine or homocysteine. Finally, indirect organosulfur metabolism by the remaining thirteen AMG families (*lysC*, *thrA*, *asd*, *hom*, *metA*, *cysE*, *cysQ*, *nrnA*, *speE*, *mdh*, *mtnD*, *mtnA*, and *mtnK*) would influence the synthesis of organosulfur compounds (e.g., synthesis of cysteine using serine) that feed into sulfide producing reactions.

Viruses encoding AMGs for organosulfur metabolism are globally distributed

Uncultivated viruses encoding AMGs for organosulfur metabolism were recovered from diverse environmental (marine, freshwater, engineered, soil, hydrothermal vent, non-marine saline and alkaline, deep subsurface, wetland, and thermal spring), non-human host-associated (mammalian gut, other animal-associated and plant-associated) and human host-associated (gastrointestinal, oral, and vaginal) microbiomes (Figure 2A). Cultivated and well-characterized viruses exhibited likewise microbiome dispersal because they were recovered from more than one ecosystem (e.g., food production, marine, freshwater, soil, engineered, hot springs, animal-associated, plant-associated, as well as human-associated gastrointestinal, oral, and skin) (Table S1). These results encompassed every ecosystem category, with the exception of air, in which viruses are routinely identified. This displays evidence that viruses encoding AMGs for sulfide production are ubiquitous on Earth.

Next, we estimated the proportion of viral richness in each ecosystem category found to encode organosulfur metabolism AMGs. Viruses encoding at least one AMG were found to be highly abundant in human vaginal, gastrointestinal, and oral

Cell Reports
Article

CellPress
OPEN ACCESS

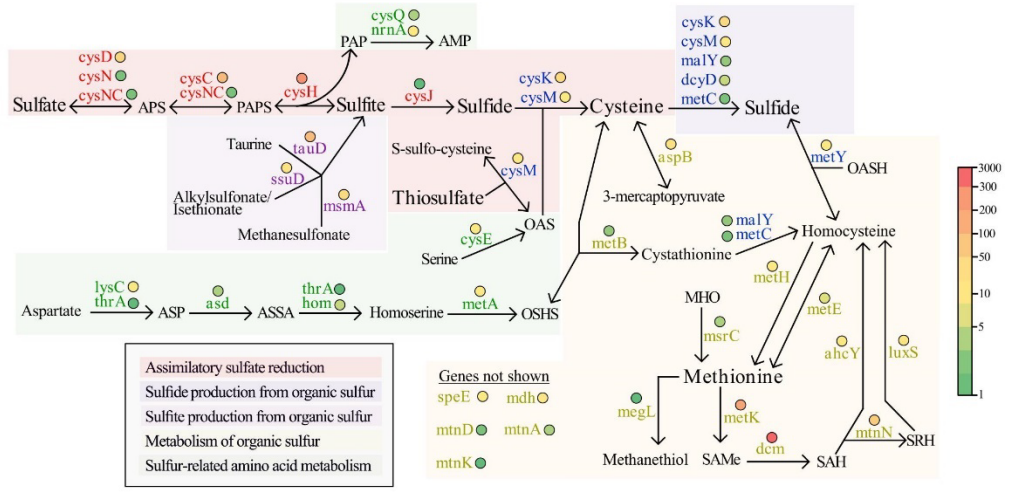


Figure 1. Reaction diagram of organosulfur transformations mediated by viruses

All genes shown have been identified on viruses and are colored coordinated respective to the process with which they are putatively associated. Colored circles represent the abundance of each AMG across all viral genomes according to the color scale (heatmap) on the right. Complete reactions and full names of acronyms are listed in Table 1.

microbiomes comprising 8%, 6%, and 3% of all identified viruses, respectively. Mammalian-associated, other animal-associated, and plant-associated microbiomes likewise had significant AMG-encoding virus abundances of 8%, 6%, and 6%, respectively. Notably, previous reports have determined that expanded viral richness in the gastrointestinal tract is correlated with the manifestation of IBD (Norman et al., 2015), and our results support the possibility of this being in part due to the metabolic potential of viruses, such as for sulfide production. This points to an important distinction that the collective metabolic potential of viruses in these host-associated environments, in conjunction with measuring total viral richness, could have significant implications for host health. Viruses encoding organosulfur AMGs beyond host-associated microbiomes may also impact ecosystem health. Major environmental systems, such as the deep subsurface (6%), engineered (3%), soil (3%), freshwater (2%), wetlands (2%), marine (2%), and hydrothermal vents (2%), likewise display significant richness of organosulfur AMG encoding viruses (Table S2C). The net impact of viral metabolism on organic and inorganic sulfur compound concentrations in these environments is unknown, but it is nonetheless striking that up to 8% of all resident viruses may be involved.

Viruses recovered from non-human microbiomes also displayed extensive geographical and niche distributions, which demonstrate their relevance in global sulfur biochemistry (Figure 2B). Individual distributions of abundant AMGs (e.g., *dcm*, *cysC*, *cysK*, *cysH*, *metK*, and *tauD*) likewise had no geographical or environmental restrictions (Figures S1A–S1F). For example, *cysH* that encodes a critical enzyme for assimilatory sulfur metabolism was found in every ecosystem except the deep subsurface. CysK, a predominant enzyme involved in sulfide generation from cysteine degradation was also broadly dispersed in marine, freshwater, engineered, hydrothermal vent, and host-associated environments. Even *msmA* that was only identified in marine environments showed strong geographical dispersal (Figure S1G).

AMG distributions between environments may depend on different factors, such as how universal the AMG function is (e.g., CysH and CysK are common among bacteria) or the nutrient landscape in a specific environment (e.g., MsmA is capable of degrading methanesulfonate, a common compound in marine environments) (Henriques and De Marco, 2015). However, human-associated samples contained the greatest fraction of identified *cysH* and *cysD* AMGs overall, whereas marine and freshwater environments contained nearly all of the identified *cysC*. In human-associated samples, nearly 97% of AMGs were *cysD*, *cysH*, *metK*, *mtnN*, *luxS*, and *dcm* that encompass essential steps of cysteine and methionine degradation (Figure 2C). The uneven distribution of these assimilatory sulfate reduction AMGs suggests that further constraints on nutrient availability or variance in rate limiting steps based on thermodynamics in different environments play a role in determining the distribution of organosulfur metabolism AMGs.

Viral organosulfur AMGs result in likely functional proteins and provide a fitness advantage to the virus

To overcome the challenge of assigning conclusive function to protein sequences in the absence of biochemical evidence, we analyzed functional and conserved domains of AMG-encoded proteins with biochemically characterized bacterial homologs. Overall, we examined 24 AMG families and found broad

CellPress
OPEN ACCESS

Cell Reports
Article

Table 1. Complete reaction(s) performed by each AMG-encoded protein

Pathway	Protein	Reaction(s)
Assimilatory sulfate reduction	CysC	APS + ATP ↔ PAPS + ADP + H^+
	CysN	SO_4^{2-} + ATP + H^+ ↔ APS+ $P_2O_7^{4-}$
	CysD	SO_4^{2-} + ATP + H^+ ↔ APS' $P_2O_7^{4-}$
	CysH	PAP + SO_3^{2-} + an oxidized Trdx + 2 H^+ ↔ PAPS + a reduced Trdx
	CysNC	PAP + ATP ↔ PAPS + ADP + H^+
		SO_4^{2-} + ATP + H^+ ↔ APS + $P_2O_7^{4-}$
	CysJ	SO_3^{2-} + 3 NADPH + 5 H^+ → H_2S + 3 $NADP^+$ + 3 H_2O
Direct sulfide production	CysK	OAS + H_2S → L-cysteine + acetate + H^+
		L-cysteine + H_2O → pyruvate + H_2S + NH_4^+
	CysM	OAS + $S_2O_3^{2-}$ ↔ S-sulfo-L-cysteine + acetate + H^+
		OAS + H_2S → L-cysteine + acetate + H^+
		L-cysteine + H_2O → pyruvate + H_2S + NH_4^+
	MalY	L-cystathionine + H_2O → L-homocysteine + pyruvate + NH_4^+
		L-cysteine + H_2O → pyruvate + H_2S + NH_4^+
	DcyD	D-cysteine + H_2O → NH_4^+ + pyruvate + H_2S
		3-chloro-D-alanine + thioglycolate → S-carboxymethyl-D-cysteine + Cl^- + H^+
	MetC	L-cystathionine + H_2O → L-homocysteine + pyruvate + NH_4^+
		L-cysteine + H_2O → pyruvate + H_2S + NH_4^+
	MetY	OASH + H_2S ↔ L-homocysteine + acetate + H^+
Direct sulfite production	TauD	taurine + 2-OG + O_2 → SO_3^{2-} + 2-aminoacetaldehyde + succinate + CO_2 + H^+
	SsuD	an alkylsulfonate + $FMNH_2$ + O_2 → an aldehyde + SO_3^{2-} + FMN + H_2O + 2H^+
		isethionate + $FMNH_2$ + O_2 → glycolaldehyde + SO_3^{2-} + FMN + H_2O + 2H^+
	MsmA	methanesulfonate + NADH + O_2 → formaldehyde + SO_3^{2-} + NAD^+ + H_2O
Indirect sulfide production	MetB	OSHS + L-cysteine ↔ L-cystathionine + succinate + H^+
		OSHS + H_2O → 2-oxobutanoate + succinate + NH_4^+ + H^+
	MetH	L-homocysteine + a 5-methyl-THF → L-methionine + a THF
	MetE	L-homocysteine + 5-methyl-THP-3G ↔ L-methionine + THP-3G
	MetK	ATP + L-methionine + H_2O → SAMe + PO_4^{3-} + $P_2O_7^{4-}$
	MtnN	SAH + H_2O → SRH + adenine
		MTA + H_2O → MTR + adenine
	Dcm	SAMe + a cytosine in DNA → a 5-methylcytosine in DNA + SAH + H^+
	AhcY	SAH + H_2O → L-homocysteine + adenosine
	LuxS	SRH → L-homocysteine + autoinducer 2
	MsrC	MHO + a reduced Trdx → L-methionine + an oxidized Trdx + H_2O
	MegL	L-methionine + H_2O → methanethiol + 2-oxobutanoate + NH_4^+
	AspB	L-aspartate + 2-OG ↔ oxaloacetate + L-glutamate
		L-cysteine + 2-OG ↔ 3-mercaptopyruvate + L-glutamate
Indirect sulfur metabolism	CysE	L-serine + acetyl-CoA → OAS + CoA
	NrnA	PAP + H_2O → AMP + PO_4^{3-}
	SpeE	putrescine + dAdoMT ↔ spermidine + MTA + H^+
		cadaverine + dAdoMT → aminopropylcadaverine + MTA + H^+
	MetA	L-homoserine + succinyl-CoA → OSHS + CoA
	MtnK	ATP + MTR → ADP + 5-MTR-1-phosphate + H^+
	MtnA	5-MTR-1-phosphate → 5-(MT)-ribulose 1-phosphate
	MtnD	DHK-MTPene + O_2 → 4-(MT)-2-oxobutanoate + formate + H^+
		DHK-MTPene + O_2 → 3-(MT)propanoate + formate + CO + H^+
	LysC	L-aspartate + ATP → ASP + ADP
	ThrA	L-aspartate + ATP → ASP + ADP

(*Continued on next page*)

Cell Reports
Article

CellPress
OPEN ACCESS

Table 1. *Continued*

Pathway	Protein	Reaction(s)
		ASSA + NAD(P)H + H^+ → L-homoserine + NAD(P)$^+$
	Asd	ASSA + $NADP^+$ + PO_4^{3-} ↔ ASP + NADPH + H^+
	Hom	ASSA + NAD(P)H + H^+ → L-homoserine + NAD(P)$^+$
	Mdh	(S)-malate + NAD^+ ↔ oxaloacetate + NADH + H^+
	CysQ	PAP + H_2O → AMP + PO_4^{3-}

Each protein is grouped respective to the main organosulfur metabolism pathway in which it is involved. Abbreviations are as follows: PAP, adenosine 3′,5′-bisphosphate; APS, adenosine 5′-phosphosulfate; PAPS, 3′-phosphoadenosine-5′-phosphosulfate; CoA, coenzyme A; OG, oxoglutarate; OAS, O-acetyl-L-serine; OASH, O-acetyl-L-homoserine; OSHS, O-succinyl-L-homoserine; SAMe, S-adenosyl-L-methionine; dAdoMT, S-adenosyl 3-(methylsulfanyl)propylamine; MTA, S-methyl-5′-thioadenosine; MTR, 5-(methylsulfanyl)-α-D-ribose; MT, methylsulfanyl; SAH, S-adenosyl-L-homocysteine; SRH, S-ribosyl-L-homocysteine; DHK-MTPene, 1,2-dihydroxy-5-(methylsulfanyl)pent-1-en-3-one; ASSA, L-aspartate 4-semialdehyde; ASP, L-aspartyl-4-phosphate; MHO, L-methionine-(R)-S-oxide; Trdx, thioredoxin; THF, tetrahydrofolate; THP-3G, tetrahydropteroyl tri-L-glutamate.

conservation of whole protein sequence and functional amino acid residues (Figure S2). For example, viral sequences encode specific domains for: CysC, ATP binding (gsGKss) and required motifs (dgD) (Poyraz et al., 2015); CysK, cofactor pyridoxal phosphate binding (KDR, NtG, GT/SgGT and SS/AG), substrate binding (T/SSGN and QF), and phosphate recognition (GI/V) (Ishikawa et al., 2010); MetK, substrate binding (egHPDk, acE, gEit, GDqG, DaK, TgRKi, sGKd, and kvDrs) (Komoto et al., 2004); CysH, iron-sulfur cluster motif (CC…CxxC) (Chartron et al., 2006); and TauD, nitrogen and oxygen binding (e.g., WH and H) (Knauer et al., 2012). Conserved amino acid residues that are not functional are likely preserved for structural features. The retention of AMGs on viral genomes despite strong selective pressures for reduced genome size suggests that most of these AMGs are functional (Bragg and Chisholm, 2008). In addition to functional and conserved domain analysis, we calculated the ratio of non-synonymous to synonymous nucleotide differences (*dN*/*dS*) for a subset of the abundant viral AMG families. A *dN*/*dS* value less than one would suggest that the virus is under selective pressures to retain a functional AMG. *dN*/*dS* calculations for *cysK*, *cysC*, *cysD*, *cysH*, *tauD*, *msmA*, *metK*, *mtmN*, and *luxS* AMG pairs revealed that viral AMGs appear to be under purifying selective pressures to retain function of the encoded AMGs (Figure S3).

To assess if viral AMGs are active in the environment, we queried a comprehensive metagenomic and metatranscriptomic dataset from Lake Mendota, WI. We identified 23 AMGs representative of six gene families (*aspB*, *cysC*, *cysH*, *metK*, *speE*, and *tauD*) that were actively expressed by 22 different viruses over a 48-h time period (Table S3). One *cysC* in particular was expressed by a virus with a 210 kb genome that was bioinformatically determined to be complete and circular. Analysis of the genome's GC-skew, a metric to evaluate genome replication patterns using nucleotide coverage (Sernova and Gelfand, 2008), was used to determine that the virus performs rolling circle replication (i.e., unidirectional) that is a common method utilized by viruses (Olm et al., 2017a) (Figure S4A). To assess if the virus was actively replicating when *cysC* was expressed, we used a metagenomic read mapping approach to estimate the genome's *in situ* index of replication (iRep) (Brown et al., 2016). The genome's iRep value of 1.54 falls within the range of typical values of growing populations and indicates that the virus was actively replicating its genome in the environment when *cysC* was expressed (Figure S4B). Analyses of other host-virus systems with transcriptomic data enabled the identification of *cysH* expression by Enterobacteria phage Lambda during infection of *Escherichia coli* MG1655 (Liu et al., 2013). The activity and expression of viral AMGs in various systems provides further evidence that they are likely utilized for a specific function during infection.

To validate that AMGs are in fact transcribed during infection, we developed a model host-virus system with *Lactococcus lactis* C10 and its *cysK*-encoding virus Lactococcus phage P087. The transcript abundance of *cysK* was measured in a culture of either *L. lactis* C10 grown alone (control) or with P087 at time points 15-, 60-, and 120-min post infection (Figure S5; Table S4A). At 120 min, the host cells in the infection condition had mostly lysed from viral infection. Transcript abundance of *L. lactis* C10 *cysK* was found to be comparable at 15 min and 60 min in either the uninfected control or infected with P087. At 120 min, transcripts of *L. lactis* C10 *cysK* were 4× greater than at 60 min in the control but were undetectable in the infected condition. This suggests that *L. lactis* C10 *cysK* transcripts are greatly reduced during mid to late infection by P087. The transcript abundance of P087 *cysK* follows a similar trend as *L. lactis* C10 *cysK*. At 15 min, P087 *cysK* transcripts were near zero and by 60 min were in ~2× greater abundance compared to transcripts of the host. By 120 min, P087 *cysK* transcripts were likewise reduced to nearly initial levels. There was no detection of P087 *cysK* transcripts within the uninfected control. Although P087 *cysK* transcript abundance never exceeded that of *L. lactis* C10 *cysK*, we provide further evidence that the viral AMG *cysK* is actively transcribed during infection and potentially replaced host *cysK* to an extent with the greatest abundance during mid infection rather than early or late infection.

To validate that transcribed AMGs in fact produce protein, we further leveraged the *L. lactis* and P087 system. Using untargeted mass spectrometry at the endpoint of virus infection (i.e., lysis) we identified that P087's AMG *cysK* produces protein and at ~1.5× greater abundance than *L. lactis* C10 *cysK* (Table S4B). The higher ratio of virus CysK to host CysK suggests the virus gains a fitness advantage from compensation of CysK

CellPress
OPEN ACCESS

Cell Reports
Article

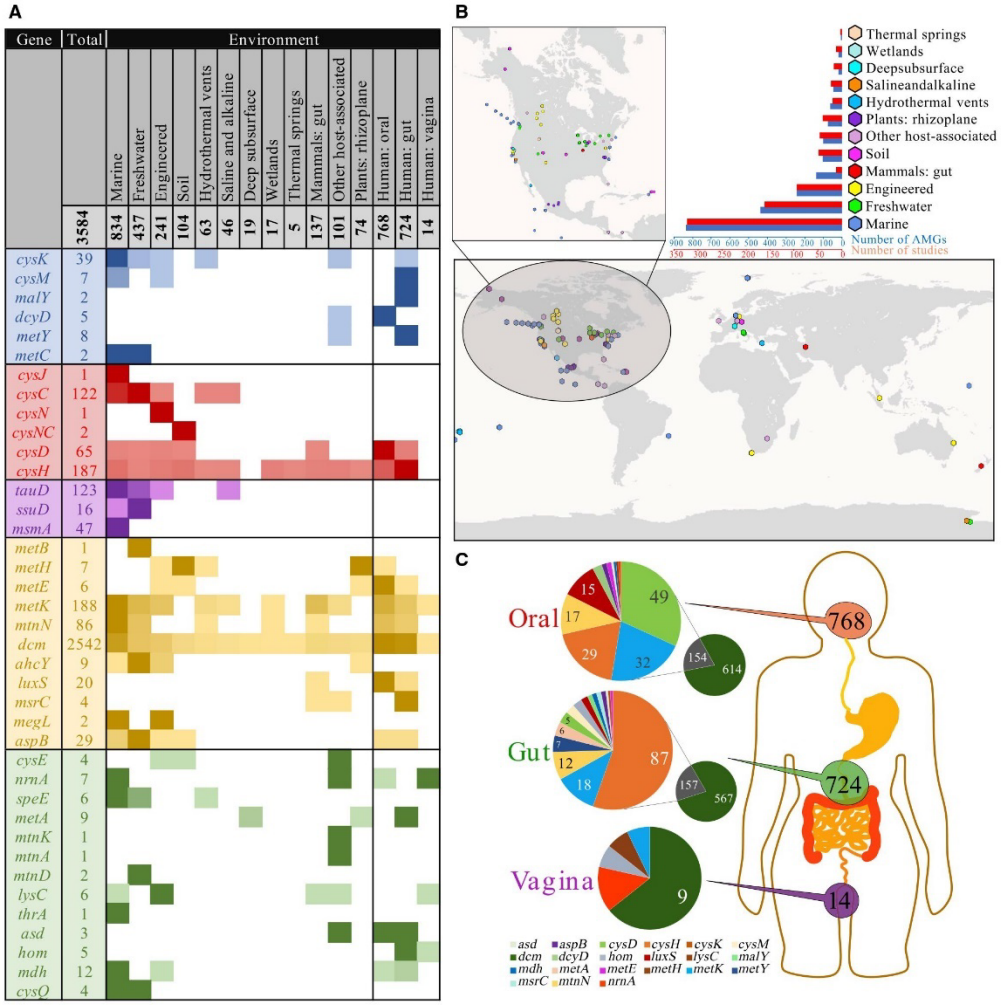


Figure 2. Distribution of viral AMGs in environmental and human microbiomes

(A) Heatmap of each AMG's relative abundance in environmental and human systems with colors coordinated by the AMG's pathway respective to Figure 1. Per AMG, darker colors represent greater abundance. A total of 3,584 AMGs derived from IMG/VR are shown.

(B) Global distribution of viruses encoding AMGs, color coordinated by environment classification. The bar graphs represent the number of AMGs and IMG studies from which viruses were recovered. See Tables S2A and S2B for exact abundances for (A) and (B), respectively. Only studies with published coordinates and environment categories are shown.

(C) Abundance of AMGs derived from incomplete or uncultivated viruses from human oral, gastrointestinal, and vaginal microbiomes. Only values greater than five are shown.

levels in the cell. These findings build upon the results from our qPCR-based analysis of transcript abundance in which host transcripts were more abundant than viral but may be explained by higher stability of either viral CysK or *cysK* transcripts. Moreover, because viruses demand a substantial fraction of cellular resources during infection (Mahmoudabadi et al., 2017), the high viral CysK levels measured here supports our hypothesis that CysK is actively utilized during productive infection in

contrast to being metabolically inactive. The presence of the gene on the genome in conjunction with transcription and translation measurements is consistent with the AMG providing a fitness advantage, which has been modeled to be as much as a 4% gain for some AMGs (Bragg and Chisholm, 2008). The mechanism(s) by which this functions is likely different than what has been observed previously for AMGs. For example, AMGs for photosynthesis were found to have differential effects during light-dark cycles as well as transcript compensatory effects over an ~8 h time period (Thompson et al., 2011). Conversely, P087 is not influenced by light-dark cycles and complete lysis can occur within ~2.5 h. Beyond providing evidence that AMGs can be remarkably active during infection, this further underlines the diverse nature by which AMGs are utilized by viruses. In addition, the identification of similar gene families on genomes of diverse, geographically spread viruses strongly supports the hypothesis that organosulfur metabolism AMGs play a functional role during infection (Roux et al., 2014).

Viruses encoding organosulfur AMGs are phylogenetically diverse

To investigate the diversity of AMGs, we conducted phylogenetic analysis of encoded amino acid sequences for five gene families. Phylogeny of CysH from complete viral genomes show close relationships between viruses and their known hosts, supporting previous observations that AMGs are most often acquired from the host (Sullivan et al., 2006) (Figure S6A). One clade in particular encoded an addition domain of unknown function (DUF3440) that suggests a shared evolutionary history. Analysis of CysH phylogeny of viral contigs with no known host revealed a similar clustering of viruses with their putative bacterial hosts (phyla Bacteroidetes and Firmicutes) (Figure S6B). In contrast to CysH, phylogenetic analysis for several abundant AMG protein sequences (CysC, CysK, TauD, and MetK) on complete and incomplete viral genomes displayed clustering of viral sequences in separate clades from bacterial homologs with few exceptions of the virus clustering with a putative host (Figures S6C–S6F).

Separate clustering would suggest that viruses may have acquired AMGs beyond their current or known host range, which is supported by the observation that viruses can encode an AMG that their host does not (e.g., *cysC* for *Xylella* phage Sano), and AMGs can cluster separately from their host (e.g., CysH for *Vibrio* phages). However, based on the CysH phylogeny of complete viral genomes, another likely explanation for distinct viral clustering is that the full range of host sequences has yet to be identified. Within the human microbiome alone, thousands of novel bacterial genomes have been identified recently and may provide further insight into host ranges or origins of AMG transfer (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019). Even so, in comparison to human microbiomes, little is known about the breadth and diversity of environmental or human viromes. Analysis of all AMGs suggests they have collectively been derived from bacteria (with the exceptions of the archaeal and eukaryotic viruses) affiliated with the phyla Firmicutes, Bacteroidetes, Alphaproteobacteria, and Gammaproteobacteria, which is supported by the host range of cultivated AMG-encoding viruses (Table S1).

Directed recombination and AMG sequence conservation validates proposed mechanism of AMG transfer and retention

The proposed mechanism of AMG acquisition by viruses in nature is the transfer of a host metabolic gene to the virus by recombination. Over multiple replication cycles of the viral genome, the AMG is retained as a functional gene. To verify this proposed mechanism, we engineered *Escherichia coli* phage T7 by inserting the host gene *cysK* (T7::*cysK*) to simulate a recombination event. Following successful insertion, T7::*cysK* was passaged, in three biological replicates, for nine complete infection cycles to simulate infection in nature over time. After passaging, the T7::*cysK* construct was sequenced to check for retention of the AMG in the viral population. Sequencing confirmed retention of the gene, indicating that recombination of a host metabolic gene onto a viral genome (i.e., AMG acquisition) can lead to stable retention of an AMG over time. Furthermore, between three biological replicates no mutations from the wild-type *cysK* sequence were observed.

Importantly, these observations show that a recombination event can occur without environmental triggers (e.g., nutrient limitation during infection) or fitness constraints (e.g., metabolic bottlenecks in the host), which provides further credibility for the proposed mechanism that AMG transfer occurs frequently and randomly in nature. If the AMG provides sufficient fitness benefits, or a lack of detrimental effects on viral replication, it will be retained over multiple infection cycles. In the system developed here, conditions resulting in a fitness benefit (e.g., greater burst size or faster replication) for the T7::*cysK* virus compared to wild-type T7 were not identified.

Sulfide can provide a fitness advantage to viruses

Because active expression and function of AMGs likely can result in the production of sulfide in the environment and human microbiome, we sought to determine if sulfide does indeed confer a fitness advantage to viruses. A highly plausible method for viruses to achieve this would be through the degradation of cysteine that is present in nearly all environments. As a result, we hypothesized the *cysK*-encoding virus P087 would have the capacity to gain a fitness advantage in the presence of sulfide. Theoretically, P087 would be involved in the direct degradation of intracellular cysteine via the action of virally encoded CysK under some conditions. To elucidate if sulfide alone confers a fitness advantage, we exogenously added sulfide during P087 infection of *L. lactis* and quantified the impact on virus and host growth. We found that viable virus production increased linearly with the addition of physiologically relevant concentrations of sulfide (Figure 3A) with no significant observed differences in host growth (Figure 3B). This indicates that, under the conditions tested, P087 benefits from increased production of sulfide in the system through either AMG or host-driven mechanisms, and the resulting fitness gain is not due to a simple increase in host abundance. We performed the same experiment with exogenously added cysteine but did not observe any effect on viral fitness (data not shown). This has significant biological implications because microorganisms contain high intracellular concentrations of cysteine, with *L. lactis* species reported to contain ~3.5 mM intracellular cysteine (Li et al., 2005). Likewise,

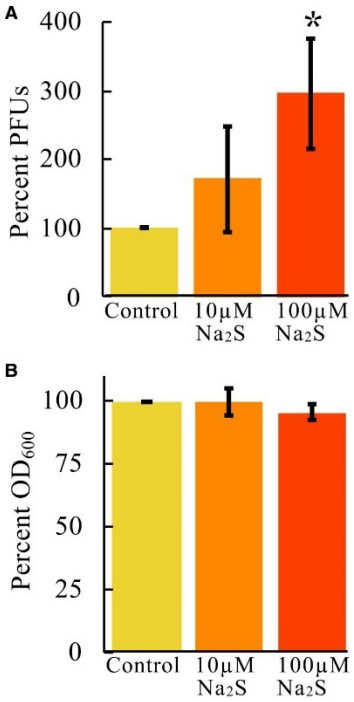


Figure 3. Increased viral fitness is associated with sulfide concentrations

(A) Impact of varying sulfide concentrations on Lactococcus phage P087 virus production as measured by plaque forming units (PFUs). The average of three independent experiments each with three biological replicates with SD error bars are shown.

(B) Corresponding uninfected host growth as an average of two biological replicates. Experimental conditions are normalized to percent of control. Asterisk represents statistical significance ($p < 0.02$) compared to the control.

Escherichia coli has a free cysteine pool of ~150 μM (Park and Imlay, 2003). We believe other viruses encoding organosulfur metabolism AMGs would likewise derive a fitness advantage under similar conditions, and this phenotype is not restricted to the ability to directly produce sulfide from cysteine degradation.

Viral organosulfur auxiliary metabolism associated with human gut bacteria

Among viruses with known hosts, 107 viruses were found to be associated with 35 different bacterial species known to be commensal or pathogenic residents of the human gastrointestinal tract (Table S1). These viruses encode five AMGs (*cysE*, *cysH*, *cysK*, *dcm*, and *metK*) for both the assimilation of sulfur and the capacity to degrade organosulfur compounds into sulfide. Most of these viruses were isolated from a variety of dairy, soil, sewage, and wastewater environments indicating a potential for environmental reservoirs of sulfide producing viruses, or in the case of wastewater environments, the viruses may have been resident in human gastrointestinal tracts. Five AMG-encoding viruses of the pathogens *Salmonella enterica*, *Staphylococcus aureus*, *Vibrio cholerae*, and *Clostridium difficile* were isolated from human fecal samples indicating transmission and replication in human gastrointestinal tracts likely does occur and may contribute to dysbiosis via the production of sulfide or altering the organosulfur metabolic potential of the pathogenic host.

Uncultivated viruses from the human gastrointestinal tract encoding AMGs putatively involved in direct sulfide production (*cysM*, *malY*, and *metY*) had high protein identity (>97%) to *Alistipes putredinis*, *Alistipes obesi*, *Alistipes finegoldii*, *Bacteroides uniformis*, and *Bacteroides vulgatus* suggesting they are viruses closely associated with these human gut bacteria from the order *Bacteroidales* (phylum Bacteroidetes) (Fenner et al., 2007; Hugon et al., 2013; Patrascu et al., 2017; Schirmer et al., 2018). Viruses encoding *metK*, *mtnN*, and *metE* (i.e., capacity for methionine degradation to sulfide) in human gastrointestinal samples were likewise inferred to be closely associated with the human gut bacteria *Alistipes ihumii*, *Faecalibacterium prausnitzii*, *Flavonifractor sp.*, *Bacteroides intestinalis*, *Bacteroides xylanisolvens*, *Bacteroides uniformis*, *Bacteroides thetaiotaomicron*, *Haemophilus parainfluenzae*, *Aggregatibacter sp.*, and *Eubacterium sp.* based on high protein identity (Bakir et al., 2006; Costea et al., 2017; Curtis et al., 2014; Jiang et al., 2015; Kuang et al., 2017; Martín et al., 2017; Pfleiderer et al., 2014; Qin et al., 2010; Veiga et al., 2014). At lower protein identity (96%–80%), viruses encoding *metK*, *luxS*, and *mtnN* were inferred to be in some part associated with the gut bacteria *Prevotella spp.* (*Bacteroidales*), *Butyricicoccus spp.*, and *Clostridiales sp.* (Eeckhaut et al., 2013; Larsen, 2017; Patrascu et al., 2017) (Table S5).

Many of these *Bacteroidales* (i.e., *Alistipes spp.*, *Bacteroides spp.*, and *Prevotella spp.*) and some members of the phylum Firmicutes (e.g., *Haemophilus parainfluenzae* and *Butyricicoccus spp.*) have been strongly associated with IBD (Eeckhaut et al., 2013; Lucke et al., 2006; Schirmer et al., 2018; Veiga et al., 2014) and their role in inflammation may be in part attributed to virus-mediated or influenced production of sulfide. Importantly, viruses of these *Bacteroidales*, including *Prevotella* megaphages with high coding capacity, have been shown to be dominant and abundant in human gastrointestinal tracts that could promote the continuous viral-driven production of sulfide to exacerbate inflammation (Devoto et al., 2019; Dutilh et al., 2014).

Comparative genomics displays diversity of viral genome organization

We used comparative genomics to examine the diversity of viruses found to be associated with human microbiomes. We identified four distinct uncultivated virus contigs encoding *dcm* from human oral samples to be closely related to known *Streptococcus pneumoniae* viruses based on genome sequence identity (Figure S7A). However, there are large stretches of dissimilarity between some of the genomes that may indicate evidence for large genetic exchange between viruses that frequently share the same niche and not the same host, which has been demonstrated before between *Lactococcus* and

Cell Reports
Article

CellPress
OPEN ACCESS

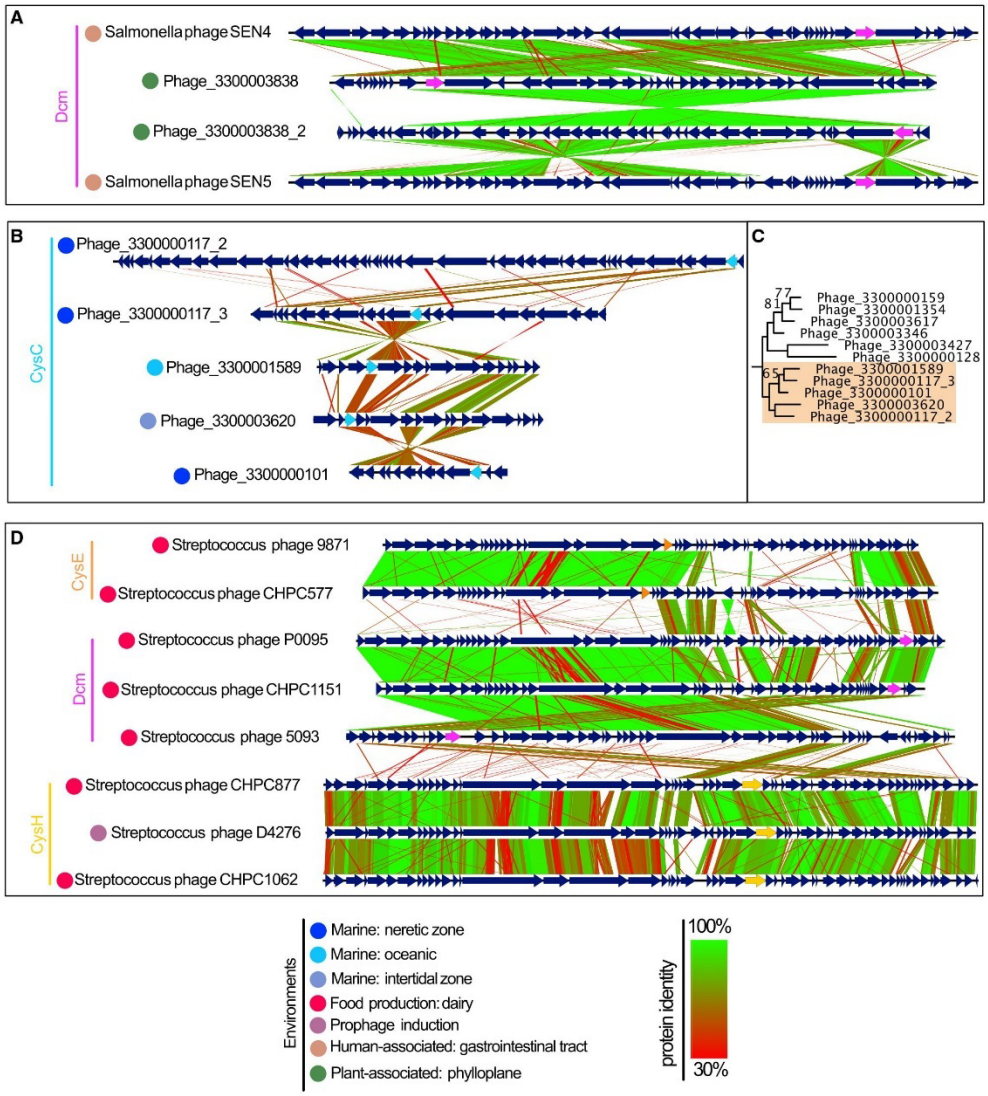


Figure 4. Genome comparisons of viruses encoding AMGs

Comparisons of (A) uncultivated viruses and complete *Salmonella enterica* viruses encoding *dcm* (pink), (B) uncultivated viruses encoding *cysC* (cyan) with (C) respective protein phylogeny (orange highlighting, refer to Figure S6 for full phylogenetic tree), and (D) complete *Streptococcus thermophilus* viruses encoding *cysE* (orange), *dcm* (pink) or *cysH* (yellow). For all comparisons, predicted open readings frames are annotated by dark blue arrows and genomes are connected with lines according to protein identity by tblastx alignment. Colored circles refer to the environment in which the virus was isolated or identified.

Enterococcus viruses (Villion et al., 2009). This observation supports the likelihood of AMG transfer between viruses in human and environmental microbiomes. Furthermore, two plant-associated viruses were identified to be closely related to known *Salmonella enterica* viruses originally derived from human fecal samples (Figure 4A). These plant-associated viruses may represent examples of environmental reservoirs for AMG-encoding viruses in the human gastrointestinal tract.

However, for either case above, the exact nature of viral transfer of AMGs is challenging to determine because AMG sequences that closely share evolutionary history can be encoded on dissimilar and geographically diverse viruses. For example, five *cysC*-encoding viruses that group closely by CysC phylogeny conversely depict dissimilarity by genome comparison and are geographically dispersed in marine environments (Figures 4B and 4C). The same is true for six different *metK*-encoding viruses in which MetK shows phylogenetic similarity but the genomes are diverse and geographically spread (Figure S7B).

To further investigate the relationships of AMGs on viral genomes, we examined the prevalence of multiple AMG copies on individual genomes. In total, we identified 285 viral genomes that contained multiple copies. Although most such genes encoded for identical functions (i.e., two copies of protein from the same gene family), some with connected (e.g., *metK* and *dcm*, *luxS* and *mtnN*) or disparate functions (e.g., *dcm* and *cysM*, *cysH* and *mtnN*) were also found. These findings suggest viruses may utilize these genes for diverse regulation of host organosulfur metabolism to fit their individual requirements (Table S6). For example, a single virus may augment both assimilatory sulfate reduction (e.g., using CysH) as well as methionine degradation (e.g., using MetK) during infection by encoding and expressing both AMGs.

We next compared viral genome organization to identify relationships in the physical location of AMGs between different viral genomes and interpret affiliations with other encoded genes. We found no universal organization of AMGs that were broadly encoded in various locations, such as between structural genes, adjacent to lysis factors, near genes for genome replication or nucleotide metabolism, and within regions comprising genes of unknown function (Figure 5). Additionally, no pattern associated with encoding specific AMGs was detected according to virus classification, genome length, or isolation source. There were a small number of outliers, such as a comparison of 10 complete viral genomes encoding *cysH* that indicated a trend toward co-location of the AMG with genome replication and/or nucleotide metabolism genes to suggest similar transcriptional regulation or function of this AMG across different viruses (Figure S7C).

The model that viruses acquire AMGs from diverse sources and for disparate functions is further supported by looking at AMG-encoding viruses that share the same host but not the same AMG. There are several different variations in which this occurs. One example involves *Bacillus cereus* phages PBC5, Basilisk, BCU4, and PBC6 where the viruses have low sequence similarity between genomes and AMG sequences (i.e., *cysH*) (Figure S7D). Another example involves *Streptococcus suis* phages phiJH1301-2, phiSC070807, phiNJ3, and phiD12 where the viruses have very similar genome sequences but encode multiple AMGs with similarity shared only among a subset of them (i.e., *metK* and *dcm*) (Figure S7E). A final example involves *Streptococcus thermophilus* phages 9871, CHPC577, P0095, CHPC1151, 5093, CHPC877, D4276, and CHPC1062 where the viruses group separately according to the single AMG each encodes (*cysE*, *cysH*, or *dcm*) (Figure 4D). Taken together, these three examples indicate that viruses are able to employ separate strategies to accomplish a similar function of manipulating host organosulfur metabolism. This may be in the form of acquiring the same AMG from different sources to perform a shared task or acquiring disparate AMGs to perform separate tasks toward the same objective, such as sulfide production.

DISCUSSION

The metabolic potential of viruses, the most abundant biological entities on Earth, is all too often overlooked because viruses do not independently conduct metabolic transformations. Here, we show that viral manipulation of host metabolism in contrast to solely measurements of viral richness and host range is likely important to the environmental sulfur cycle and human health. Furthermore, we propose that assimilatory sulfur metabolism, a ubiquitous method of fixing sulfur and manipulating organosulfur compounds, is frequently modulated by viruses during infection of organisms from all three domains and in almost all microbiomes on Earth. This poses an important question. What have we been overlooking in viromes by frequently assessing sequence reads instead of metagenomically assembled genomes that encode AMGs? Are we giving enough emphasis on viruses as core drivers in the metabolism of microbiomes?

AMG-driven organosulfur metabolism mediated by viruses may lead to sulfide production in the gastrointestinal tract during infection or following microbial lysis. The result would be a sulfide-induced inflammatory response in conjunction with the activity of resident microbiota or invading pathogens, although the extent to which this occurs in human or environmental systems has yet to be quantified. Indeed, it has been observed that infected bacterial cells have manipulated and "rewired" sulfur assimilation that will impact cysteine metabolism and likely sulfide production (Howard-Varona et al., 2020). Furthermore, viruses encoding sulfur assimilation AMGs may be short-circuiting the assimilatory sulfur pathway by reducing the steps necessary for assimilation of sulfur into organosulfur compounds. This concept is supported by the observation that *cysH* is the most abundant organosulfur metabolism AMG, which plays a role in both the canonical sulfate assimilation pathway as well as direct sulfonation of organic molecules (Moran and Durham, 2019). The latter mechanism may explain the high abundance of *cysH* on viral genomes.

The evidence presented here strongly points toward sulfide production as a component of viral organosulfur auxiliary metabolism, either directly or indirectly by AMG activity, which could provide many fitness advantages for viruses (Figure 6A). As obligate intracellular pathogens, viruses could benefit from the survival and enhanced growth of their host, which could be achieved by responding to sulfur starvation signals, assimilating sulfide for biosynthesis (e.g., for sulfolipids), upregulating sulfide utilization (e.g., sulfide oxidation), antibiotic stress response (Figure 6A, 1), or redox balance and free radical scavenging

Cell Reports
Article

CellPress
OPEN ACCESS

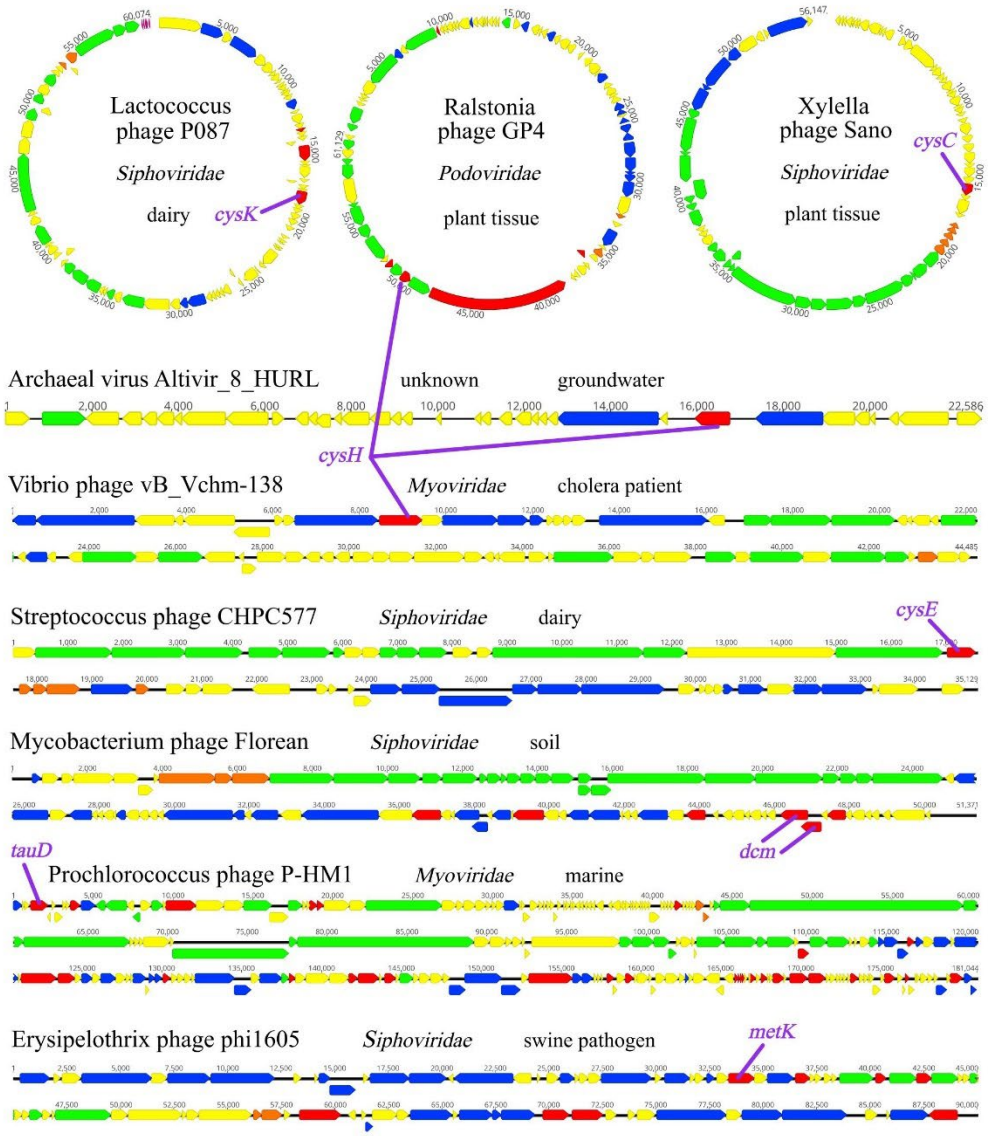


Figure 5. Genome organization of 9 complete viral genomes encoding organosulfur AMGs

Genome representation of circular and linear viruses. Arrows indicate open reading frames and are annotated by general function: virion structural assembly (green), auxiliary metabolism and general functions (red), nucleotide metabolism and genome replication (blue), lysis (orange), and unknown function (yellow). AMGs are annotated in purple.

CellPress
OPEN ACCESS

Cell Reports
Article

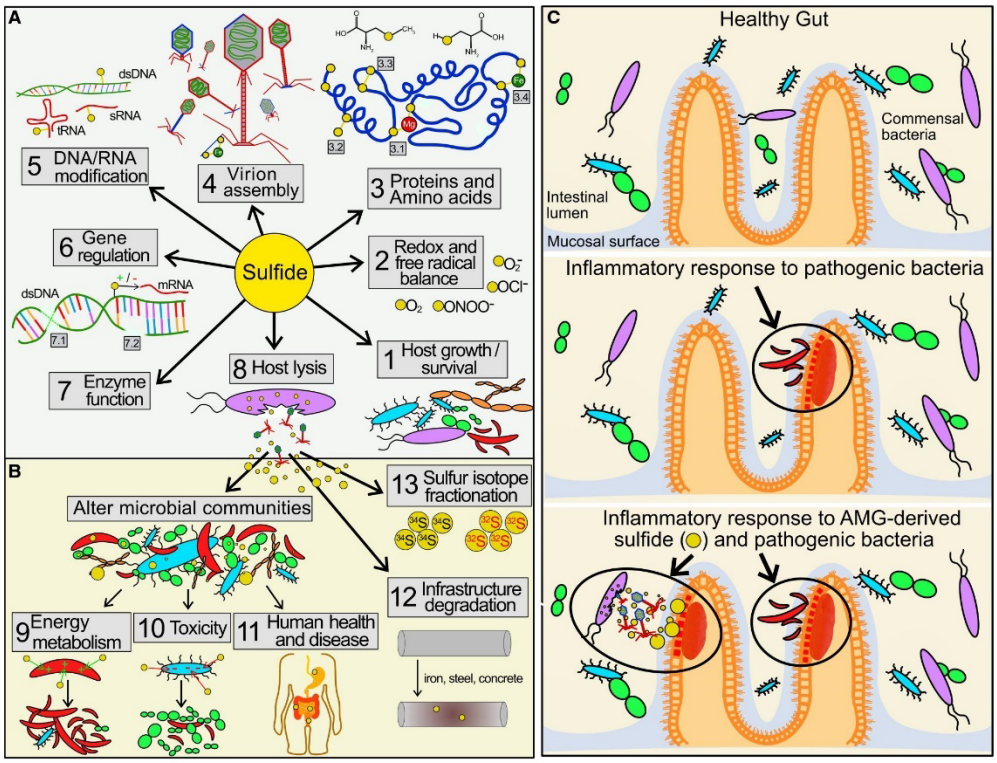


Figure 6. Virus-driven production of sulfide and its effects on human health, viral fitness, and microbial communities

(A and B) Mechanisms by which sulfide could (A) benefit viral fitness and (B) effect microbial communities, human health, and environmental conditions.

(C) Proposed impact of viral-driven production of sulfide in conjunction with activity of pathogenic bacteria on inflammation in the gastrointestinal tract and its implications in IBD and CRC.

(Figure 6A, 2) (Anantharaman et al., 2014; Gyaneshwar et al., 2005; Mahmoudabadi et al., 2017; Nambi et al., 2015; Pal et al., 2018; Roux et al., 2016; Xia et al., 2017). To benefit the virus directly, sulfide could be utilized for amino acid synthesis or protein function, such as for co-factor binding (e.g., metal ions) (Figure 6A, 3.1), persulfidation of cysteine residues for signaling (Figure 6A, 3.2), structural sulfide bridge formation (Figure 6A, 3.3), iron-sulfur cluster formation (Figure 6A, 3.4), or for viral structural proteins in virion assembly (Figure 6A, 4) (Peng et al., 2017; Tam et al., 2013). Furthermore, thiol modification of nucleic acids (i.e., dsDNA, tRNA, and sRNA) could provide an avenue for responding to stresses (Figure 6A, 5) or regulating gene expression for the virus or host (Figure 6A, 6) (Damon et al., 2015; Hsu et al., 1967; de Lira et al., 2018; Peng et al., 2017; Shimizu et al., 2017; Yang et al., 2017). Another method of nucleic acid modification that viruses may rely on is dsDNA recombination or integration (Figure 6A, 7.1), or dsDNA repair (Figure 6A, 7.2), which can be enabled by essential thiol components of enzymes (Jessop et al., 2000; Kessler, 2006; Yeeles et al., 2009). Sulfide may even be a key component in the ability of viruses to effectively lyse their host (Figure 6A, 8) (Propst-Ricciuti, 1976).

However, due to the diversity of functions encoded by AMGs (e.g., degradation of organosulfur compounds directly into sulfide or sulfite, manipulation of organosulfur compound forms, or fixing sulfur) it is likely that host physiology and local environmental conditions drive their acquisition and function. Regardless of the utility of AMGs employed by individual viruses, the eventual lysis and release of virus-derived sulfide or virus-influenced sulfur chemistry could have significant impacts on the surrounding environment and local microbial communities (Figure 6B). Increased sulfide concentrations could either enhance the growth of sulfide oxidizing organisms (Figure 6B, 9) or act as a toxin to inhibit the growth of others (Figure 6B, 10) (Pal et al., 2018). Likewise, in both environmental and human systems, intracellular content released through viral lysis could alter nutrient availability and sulfide concentrations in the microbial

community (Figure 6B, 11) or lead to the degradation of iron, steel, and concrete in infrastructure (Figure 6B, 12).

In humans, balancing organic and inorganic sulfur concentrations is pivotal to both the health of the gastrointestinal tract and the resident microbiota (Yin et al., 2016), and our evidence suggests that viruses may interfere with this equilibrium. Moreover, dozens of microbial species have been linked to accumulation of sulfide within the human gut via the degradation of organosulfur compounds (e.g., cysteine and taurine) and implicated in CRC and IBD (Carbonero et al., 2012; Guo et al., 2016), but the role of viruses in facilitating or upregulating these processes is unknown. Specifically, virus-mediated sulfide production could accelerate the development of sulfide-associated gastrointestinal disorders such as colitis, IBD, and CRC (Figure 6C).

Our discovery of AMGs for organosulfur metabolism and sulfide production also has widespread ramifications for interpreting Earth history (Figure 6B, 13). Sulfur isotope fractionation ($^{34}S/^{32}S$) analysis is widely used to interpret geological records and estimate rates of microbial processes such as sulfate reduction (Habicht and Canfield, 1997; Sim et al., 2019; Thode et al., 1953). Microbial assimilatory sulfate reduction and viral auxiliary metabolism have been ignored as contributors to fractionation in the environment, mainly because sulfide is incorporated into organosulfur compounds instead of being exported into the environment as it is in dissimilatory reactions. As a result, assimilatory fractionation appears to be negligible ($\sim 3‰$), whereas dissimilatory fractionation is frequently measured closer to $47‰$ (Chambers and Trudinger, 1979; Kaplan and Rittenberg, 1964). Without the incorporation of sulfide into organosulfur compounds, assimilatory sulfite to sulfide reduction fractionates up to $36‰$–$42‰$ in *Salmonella*, *Clostridium*, and *Bacillus* species (Chambers and Trudinger, 1979). We propose that virus-mediated sulfide production can directly impact the observed fraction of ^{32}S-enriched sulfide at scales relevant to dissimilatory sulfate reduction.

Overall, the global distribution and diversity of viruses encoding organosulfur transforming AMGs represents a so-far unexplored cog in the global organic and inorganic sulfur cycles. By modulating organic and inorganic sulfur compound concentrations, viruses likely play important roles in infrastructure degradation, human disease and ecosystem health. Beyond viral organosulfur metabolism, this study serves as a model for elucidating the impacts of virus-driven degradation of amino acids, whose fate is an important driver in human health and biotechnology and associated with ecosystem services in agriculture.

Limitations of study

This study provides preliminary evidence for the function of organosulfur AMGs and viral influences on organosulfur compounds, namely hydrogen sulfide, in environmental and human microbiomes. One limitation is that the direct roles and interactions of AMGs within organosulfur metabolic frameworks and the elucidation of incurred benefits of hydrogen sulfide for some viruses was not shown. We show that the AMG *cysK* can be conserved evolutionarily over time that points toward, rather than measures, a fitness benefit of retaining the AMG. Furthermore, our experimental evidence for the benefit of sulfide to Lactococcus Phage P087, despite identifying viral CysK protein translation, did not distinguish if the measured fitness effect was the result of viral CysK or due to other unknown viral or host factors. Finally, our attempts to purify viral CysK protein and measure its activity in degrading cysteine was unsuccessful.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Lactococcus system growth conditions
  - T7 system growth conditions
- METHOD DETAILS
  - Identification of viral genomes
  - AMG identification and annotation
  - Sequence alignment and dN/dS analysis
  - Sequence phylogeny
  - Protein functional analysis
  - Protein Reactions
  - Viral transcriptomics and growth rates
  - Virus growth and fitness assay
  - Virus and host cysK qPCR assay
  - Mass spectrometry and protein identification
  - Genome organization and comparisons
  - Geographical distributions
  - Host classification
  - T7 recombination: cloning
  - T7 recombination: engineering T7 with cysK
  - T7 recombination: passaging and AMG retention
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Virus growth and fitness
  - Virus and host cysK qPCR

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109471.

ACKNOWLEDGMENTS

We thank Anna-Louise Reysenbach and Katherine D. McMahon for helpful discussions and suggestions. We thank the University of Wisconsin-Office of the Vice Chancellor for Research and Graduate Education, University of Wisconsin-Department of Bacteriology, and University of Wisconsin-College of Agriculture and Life Sciences for their support. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231. K.K. is supported by a Wisconsin Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison. This work was partly supported by the National Science Foundation (OCE-2049478 to K.A.). A.J.P. was supported by the Ministry of Culture and Science of North Rhine-Westphalia (Nachwuchsgruppe "Dr. Alexander Probst") and the NOVAC project of the German Science Foundation (DFG PR1603/2-1).

AUTHOR CONTRIBUTIONS

K.K. and K.A. designed the study. K.K., E.Z., P.H., and A.M.B. performed host-virus experiments. K.K. and K.A. conducted bioinformatic and metabolic analyses. K.K. and A.M.L. performed metatranscriptomic analyses. K.K. and K.A. drafted the manuscript. All authors reviewed the results and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as living with a disability. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Received: April 24, 2019
Revised: January 7, 2021
Accepted: July 9, 2021
Published: August 3, 2021

REFERENCES

Ahlgren, N.A., Fuchsman, C.A., Rocap, G., and Fuhrman, J.A. (2019). Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. ISME J. *13*, 618–631.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anantharaman, K., Duhaime, M.B., Breier, J.A., Wendt, K.A., Toner, B.M., and Dick, G.J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. Science *344*, 757–760.

Anantharaman, K., Hausmann, B., Jungbluth, S.P., Kantor, R.S., Lavy, A., Warren, L.A., Rappé, M.S., Pester, M., Loy, A., Thomas, B.C., and Banfield, J.F. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. ISME J. *12*, 1715–1728.

Ando, H., Lemire, S., Pires, D.P., and Lu, T.K. (2015). Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. Cell Syst. *1*, 187–196.

Andreae, M.O. (1990). Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. Mar. Chem. *30*, 1–29.

Bakir, M.A., Kitahara, M., Sakamoto, M., Matsumoto, M., and Benno, Y. (2006). Bacteroides intestinalis sp. nov., isolated from human faeces. Int. J. Syst. Evol. Microbiol. *56*, 151–154.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bragg, J.G., and Chisholm, S.W. (2008). Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. PLoS ONE *3*, e3550.

Breitbart, M., Thompson, L., Suttle, C., and Sullivan, M. (2007). Exploring the Vast Diversity of Marine Viruses. Oceanography (Wash. D.C.) *20*, 135–139.

Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res. *43*, D571–D577.

Brown, C.T., Olm, M.R., Thomas, B.C., and Banfield, J.F. (2016). Measurement of bacterial replication rates in microbial communities. Nat. Biotechnol. *34*, 1256–1263.

Carbonero, F., Benefiel, A.C., Alizadeh-Ghamsari, A.H., and Gaskins, H.R. (2012). Microbial pathways in colonic sulfur metabolism and links with health and disease. Front. Physiol. *3*, 448.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., et al. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. *40*, D742–D753.

Chambers, L.A., and Trudinger, P.A. (1979). Microbiological fractionation of stable sulfur isotopes: A review and critique. Geomicrobiol. J. *1*, 249–293.

Chartron, J., Carroll, K.S., Shiau, C., Gao, H., Leary, J.A., Bertozzi, C.R., and Stout, C.D. (2006). Substrate recognition, protein dynamics, and iron-sulfur cluster in Pseudomonas aeruginosa adenosine 5′-phosphosulfate reductase. J. Mol. Biol. *364*, 152–169.

Chen, L.-X., Méheust, R., Crits-Christoph, A., McMahon, K.D., Nelson, T.C., Warren, L.A., and Banfield, J.F. (2020). Large freshwater phages with the potential to augment aerobic methane oxidation. Nat. Microbiol. *5*, 1504–1515.

Chiku, T., Padovani, D., Zhu, W., Singh, S., Vitvitsky, V., and Banerjee, R. (2009). H2S biogenesis by human cystathionine γ-lyase leads to the novel sulfur metabolites lanthionine and homolanthionine and is responsive to the grade of hyperhomocysteinemia. J. Biol. Chem. *284*, 11601–11612.

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2016). GenBank. Nucleic Acids Res. *44* (*D1*), D67–D72.

Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960.

Curtis, M.M., Hu, Z., Klimko, C., Narayanan, S., Deberardinis, R., and Sperandio, V. (2014). The gut commensal Bacteroides thetaiotaomicron exacerbates enteric infection through modification of the metabolic landscape. Cell Host Microbe *16*, 759–769.

Damon, J.R., Pincus, D., and Ploegh, H.L. (2015). tRNA thiolation links translation to stress responses in Saccharomyces cerevisiae. Mol. Biol. Cell *26*, 270–282.

de Lira, N.P.V., Pauletti, B.A., Marques, A.C., Perez, C.A., Caserta, R., de Souza, A.A., Vercesi, A.E., Paes Leme, A.F., and Benedetti, C.E. (2018). BigR is a sulfide sensor that regulates a sulfur transferase/dioxygenase required for aerobic respiration of plant bacteria under sulfide stress. Sci. Rep. *8*, 3508.

Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J., Archie, E.A., Turnbaugh, P.J., Seed, K.D., Blekhman, R., et al. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat. Microbiol. *4*, 693–700.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., 3rd, Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., et al. (2008). The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. J. Bacteriol. *190*, 2597–2606.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 4498.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Eeckhaut, V., Machiels, K., Perrier, C., Romero, C., Maes, S., Flahou, B., Steppe, M., Haesebrouck, F., Sas, B., Ducatelle, R., et al. (2013). Butyricicoccus pullicaecorum in inflammatory bowel disease. Gut *62*, 1745–1752.

Fenner, L., Roux, V., Ananian, P., and Raoult, D. (2007). Alistipes finegoldii in blood cultures from colon cancer patients. Emerg. Infect. Dis. *13*, 1260–1262.

Fike, D.A., Bradley, A.S., and Rose, C.V. (2015). Rethinking the Ancient Sulfur Cycle. Annu. Rev. Earth Planet. Sci. *43*, 593–622.

Fitzgerald, J.W. (1976). Sulfate ester formation and hydrolysis: a potentially important yet often ignored aspect of the sulfur cycle of aerobic soils. Bacteriol. Rev. *40*, 698–721.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gietz, R.D., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol. *350*, 87–96.

Gobler, C.J., Hutchins, D.A., Fisher, N.S., Cosper, E.M., and Sañudo-Wilhelmy, S.A. (1997). Release and bioavailability of C, N, P Se, and Fe following viral lysis of a marine chrysophyte. Limnol. Oceanogr. *42*, 1492–1504.

Guo, F.-F., Yu, T.-C., Hong, J., and Fang, J.-Y. (2016). Emerging Roles of Hydrogen Sulfide in Inflammatory and Neoplastic Colonic Diseases. Front. Physiol. *7*, 156.

Gyaneshwar, P., Paliy, O., McAuliffe, J., Popham, D.L., Jordan, M.I., and Kustu, S. (2005). Sulfur and nitrogen limitation in Escherichia coli K-12: specific homeostatic responses. J. Bacteriol. *187*, 1074–1090.

Habicht, K.S., and Canfield, D.E. (1997). Sulfur isotope fractionation during bacterial sulfate reduction in organic-rich sediments. Geochim. Cosmochim. Acta *61*, 5351–5361.

Henriques, A.C., and De Marco, P. (2015). Methanesulfonate (MSA) Catabolic Genes from Marine and Estuarine Bacteria. PLoS ONE *10*, e0125735.

Howard-Varona, C., Lindback, M.M., Bastien, G.E., Solonenko, N., Zayed, A.A., Jang, H., Andreopoulos, B., Brewer, H.M., Glavina Del Rio, T., Adkins, J.N., et al. (2020). Phage-specific metabolic reprogramming of virocells. ISME J. *14*, 881–895.

Hsu, W.T., Foft, J.W., and Weiss, S.B. (1967). Effect of bacteriophage infection on the sulfur-labeling of sRNA. Proc. Natl. Acad. Sci. USA *58*, 2028–2035.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680–682.

Hugon, P., Ramasamy, D., Lagier, J.-C., Rivet, R., Couderc, C., Raoult, D., and Fournier, P.-E. (2013). Non contiguous-finished genome sequence and description of Alistipes obesi sp. nov. Stand. Genomic Sci. *7*, 427–439.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Hurwitz, B.L., and U'Ren, J.M. (2016). Viral metabolic reprogramming in marine ecosystems. Curr. Opin. Microbiol. *31*, 161–168.

Hurwitz, B.L., Hallam, S.J., and Sullivan, M.B. (2013). Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol. *14*, R123.

Hurwitz, B.L., Brum, J.R., and Sullivan, M.B. (2015). Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. ISME J. *9*, 472–484.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Ishikawa, K., Mino, K., and Nakamura, T. (2010). New function and application of the cysteine synthase from archaea. Biochem. Eng. J. *48*, 315–322.

Jaschke, P.R., Lieberman, E.K., Rodriguez, J., Sierra, A., and Endy, D. (2012). A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. Virology *434*, 278–284.

Jessop, L., Bankhead, T., Wong, D., and Segall, A.M. (2000). The amino terminus of bacteriophage lambda integrase is involved in protein-protein interactions during recombination. J. Bacteriol. *182*, 1024–1034.

Jiang, W., Wu, N., Wang, X., Chi, Y., Zhang, Y., Qiu, X., Hu, Y., Li, J., and Liu, Y. (2015). Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. Sci. Rep. *5*, 8096.

Jiao, N., Herndl, G.J., Hansell, D.A., Benner, R., Kattner, G., Wilhelm, S.W., Kirchman, D.L., Weinbauer, M.G., Luo, T., Chen, F., and Azam, F. (2010). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. Nat. Rev. Microbiol. *8*, 593–599.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

Jover, L.F., Effler, T.C., Buchan, A., Wilhelm, S.W., and Weitz, J.S. (2014). The elemental composition of virus particles: implications for marine biogeochemical cycles. Nat. Rev. Microbiol. *12*, 519–528.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J. Mol. Biol. *428*, 726–731.

Kaplan, I.R., and Rittenberg, S.C. (1964). Microbiological Fractionation of Sulphur Isotopes. J. Gen. Microbiol. *34*, 195–212.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kessler, D. (2006). Enzymatic activation of sulfur for incorporation into biomolecules in prokaryotes. FEMS Microbiol. Rev. *30*, 825–840.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90.

Kieft, K., Zhou, Z., Anderson, R.E., Buchan, A., Campbell, B.J., Hallam, S.J., Hess, M., Sullivan, M.B., Walsh, D.A., Roux, S., and Anantharaman, K. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat. Commun. *12*, 3503.

Knauer, S.H., Hartl-Spiegelhauer, O., Schwarzinger, S., Hänzelmann, P., and Dobbek, H. (2012). The Fe(II)/α-ketoglutarate-dependent taurine dioxygenases from Pseudomonas putida and Escherichia coli are tetramers. FEBS J. *279*, 816–831.

Komoto, J., Yamada, T., Takata, Y., Markham, G.D., and Takusagawa, F. (2004). Crystal structure of the S-adenosylmethionine synthetase ternary complex: a novel catalytic mechanism of S-adenosylmethionine synthesis from ATP and Met. Biochemistry *43*, 1821–1831.

Kuang, Y.-S., Lu, J.-H., Li, S.-H., Li, J.-H., Yuan, M.-Y., He, J.-R., Chen, N.-N., Xiao, W.-Q., Shen, S.-Y., Qiu, L., et al. (2017). Connections between the human gut microbiome and gestational diabetes mellitus. Gigascience *6*, 1–12.

Kuijpers, N.G., Solis-Escalante, D., Bosman, L., van den Broek, M., Pronk, J.T., Daran, J.-M., and Daran-Lapujade, P. (2013). A versatile, efficient strategy for assembly of multi-fragment expression vectors in Saccharomyces cerevisiae using 60 bp synthetic recombination sequences. Microb. Cell Fact. *12*, 47.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Larsen, J.M. (2017). The immune response to Prevotella bacteria in chronic inflammatory disease. Immunology *151*, 363–374.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Li, Y., Hugenholtz, J., Sybesma, W., Abee, T., and Molenaar, D. (2005). Using Lactococcus lactis for glutathione overproduction. Appl. Microbiol. Biotechnol. *67*, 83–90.

Lillehaug, D. (1997). An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. J. Appl. Microbiol. *83*, 85–90.

Linz, A.M., He, S., Stevens, S.L.R., Anantharaman, K., Rohwer, R.R., Malmstrom, R.R., Bertilsson, S., and McMahon, K.D. (2018). Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. PeerJ *6*, e6075.

Linz, A.M., Aylward, F.O., Bertilsson, S., and McMahon, K.D. (2020). Time-series metatranscriptomes reveal conserved patterns between phototrophic and heterotrophic microbes in diverse freshwater systems. Limnol. Oceanogr. *65*, S101–S112.

Liu, X., Jiang, H., Gu, Z., and Roberts, J.W. (2013). High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proc. Natl. Acad. Sci. USA *110*, 11928–11933.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402–408.

Lucke, K., Miehlke, S., Jacobs, E., and Schuppler, M. (2006). Prevalence of Bacteroides and Prevotella spp. in ulcerative colitis. J. Med. Microbiol. *55*, 617–624.

Ma, H., Cheng, X., Li, G., Chen, S., Quan, Z., Zhao, S., and Niu, L. (2000). The influence of hydrogen sulfide on corrosion of iron under different conditions. Corros. Sci. *42*, 1669–1683.

Mahmoudabadi, G., Milo, R., and Phillips, R. (2017). Energetic cost of building a virus. Proc. Natl. Acad. Sci. USA *114*, E4324–E4333.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. Nature *424*, 741.

Manojlović, L.M. (2015). Photometry-based estimation of the total number of stars in the Universe. Appl. Opt. *54*, 6589–6591.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. *45* (*D1*), D200–D203.

Markowitz, V.M., Chen, I.-M.A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I., Tringe, S., et al. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res. *42*, D568–D573.

Martín, R., Miquel, S., Benevides, L., Bridonneau, C., Robert, V., Hudault, S., Chain, F., Berteau, O., Azevedo, V., Chatel, J.M., et al. (2017). Functional Characterization of Novel *Faecalibacterium prausnitzii* Strains Isolated from Healthy Volunteers: A Step Forward in the Use of *F. prausnitzii* as a Next-Generation Probiotic. Front. Microbiol. *8*, 1226.

Moran, M.A., and Durham, B.P. (2019). Sulfur metabolites in the pelagic ocean. Nat. Rev. Microbiol. *17*, 665–678.

Morra, M.J., and Dick, W.A. (1991). Mechanisms of h(2)s production from cysteine and cystine by microorganisms isolated from soil by selective enrichment. Appl. Environ. Microbiol. *57*, 1413–1417.

Nambi, S., Long, J.E., Mishra, B.B., Baker, R., Murphy, K.C., Olive, A.J., Nguyen, H.P., Shaffer, S.A., and Sassetti, C.M. (2015). The oxidative stress network of Mycobacterium tuberculosis reveals coordination between radical detoxification systems. Cell Host Microbe *17*, 829–837.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell *160*, 447–460.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44* (*D1*), D733–D745.

Olm, M.R., Brown, C.T., Brooks, B., Firek, B., Baker, R., Burstein, D., Soenjoyo, K., Thomas, B.C., Morowitz, M., and Banfield, J.F. (2017a). Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Res. *27*, 601–612.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017b). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425–430.

Paez-Espino, D., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., et al. (2017). IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. *45* (*D1*), D457–D465.

Pal, V.K., Bandyopadhyay, P., and Singh, A. (2018). Hydrogen sulfide in physiology and pathogenesis of bacteria and viruses. IUBMB Life *70*, 393–410.

Park, S., and Imlay, J.A. (2003). High levels of intracellular cysteine promote oxidative DNA damage by driving the fenton reaction. J. Bacteriol. *185*, 1942–1950.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

Patrascu, O., Béguet-Crespel, F., Marinelli, L., Le Chatelier, E., Abraham, A.-L., Leclerc, M., Klopp, C., Terrapon, N., Henrissat, B., Blottière, H.M., et al. (2017). A fibrolytic potential in the human ileum mucosal microbiota revealed by functional metagenomic. Sci. Rep. *7*, 40248.

Peng, H., Shen, J., Edmonds, K.A., Luebke, J.L., Hickey, A.K., Palmer, L.D., Chang, F.J., Bruce, K.A., Kehl-Fie, T.E., Skaar, E.P., and Giedroc, D.P. (2017). Sulfide Homeostasis and Nitroxyl Intersect via Formation of Reactive Sulfur Species in *Staphylococcus aureus*. MSphere *2*, 2.

Pfleiderer, A., Mishra, A.K., Lagier, J.-C., Robert, C., Caputo, A., Raoult, D., and Fournier, P.-E. (2014). Non-contiguous finished genome sequence and description of Alistipes ihumii sp. nov. Stand. Genomic Sci. *9*, 1221–1235.

Poyraz, Ö., Brunner, K., Lohkamp, B., Axelsson, H., Hammarström, L.G.J., Schnell, R., and Schneider, G. (2015). Crystal structures of the kinase domain of the sulfate-activating complex in Mycobacterium tuberculosis. PLoS ONE *10*, e0121494.

Propst-Ricciuti, B. (1976). The effect of host-cell starvation on virus-induced lysis by MS2 bacteriophage. J. Gen. Virol. *31*, 323–330.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Rahlff, J., Turzynski, V., Esser, S.P., Monsees, I., Bornemann, T.L.V., Figueroa-Gonzalez, P.A., Schulz, F., Woyke, T., Klingl, A., Moraru, C., et al. (2021). Lytic archaeal viruses infect abundant primary producers in Earth's crusts. Nat. Commun, in press. https://doi.org/10.1038/s41467-021-24803-4.

Rambaut, A. (2009). FigTree version 1.4.3. http://tree.bio.ed.ac.uk/software/figtree/.

Roux, S., Hawley, A.K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S.J., and Sullivan, M.B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. eLife *3*, e03125.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., et al.; Tara Oceans Coordinators (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature *537*, 689–693.

Schirmer, M., Franzosa, E.A., Lloyd-Price, J., McIver, L.J., Schwager, R., Poon, T.W., Ananthakrishnan, A.N., Andrews, E., Barron, G., Lake, K., et al. (2018). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat. Microbiol. *3*, 337–346.

Schulz, F., Andreani, J., Francis, R., Boudjemaa, H., Bou Khalil, J.Y., Lee, J., La Scola, B., and Woyke, T. (2020). Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. mSystems *5*, e00048-20.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Sernova, N.V., and Gelfand, M.S. (2008). Identification of replication origins in prokaryotic genomes. Brief. Bioinform. *9*, 376–391.

Shimizu, T., Shen, J., Fang, M., Zhang, Y., Hori, K., Trinidad, J.C., Bauer, C.E., Giedroc, D.P., and Masuda, S. (2017). Sulfide-responsive transcriptional repressor SqrR functions as a master regulator of sulfide-dependent photosynthesis. Proc. Natl. Acad. Sci. USA *114*, 2355–2360.

Sim, M.S., Ogata, H., Lubitz, W., Adkins, J.F., Sessions, A.L., Orphan, V.J., and McGlynn, S.E. (2019). Role of APS reductase in biogeochemical sulfur isotope fractionation. Nat. Commun. *10*, 44.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. *4*, e234.

Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011). Easyfig: a genome comparison visualizer. Bioinformatics *27*, 1009–1010.

Suttle, C.A. (2005). Viruses in the sea. Nature *437*, 356–361.

Suttle, C.A. (2007). Marine viruses–major players in the global ecosystem. Nat. Rev. Microbiol. *5*, 801–812.

Tam, W., Pell, L.G., Bona, D., Tsai, A., Dai, X.X., Edwards, A.M., Hendrix, R.W., Maxwell, K.L., and Davidson, A.R. (2013). Tail tip proteins related to bacteriophage λ gpL coordinate an iron-sulfur cluster. J. Mol. Biol. *425*, 2450–2462.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. *44*, 6614–6624.

The UniProt Consortium (2018). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *46*, 2699.

Thode, H.G., Macnamara, J., and Fleming, W.H. (1953). Sulphur isotope fractionation in nature and geological and biological time scales. Geochim. Cosmochim. Acta *3*, 235–243.

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A.U., Stubbe, J., and Chisholm, S.W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc. Natl. Acad. Sci. USA *108*, E757–E764.

Trubl, G., Jang, H.B., Roux, S., Emerson, J.B., Solonenko, N., Vik, D.R., Solden, L., Ellenbogen, J., Runyon, A.T., Bolduc, B., et al. (2018). Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. mSystems *3*, e00076-18.

Veiga, P., Pons, N., Agrawal, A., Oozeer, R., Guyonnet, D., Brazeilles, R., Faurie, J.-M., van Hylckama Vlieg, J.E.T., Houghton, L.A., Whorwell, P.J., et al. (2014). Changes of the human gut microbiome induced by a fermented milk product. Sci. Rep. *4*, 6328.

Villion, M., Chopin, M.-C., Deveau, H., Ehrlich, S.D., Moineau, S., and Chopin, A. (2009). P087, a lactococcal phage with a morphogenesis module similar to an Enterococcus faecalis prophage. Virology *388*, 49–56.

Voordouw, G., Armstrong, S.M., Reimer, M.F., Fouts, B., Telang, A.J., Shen, Y., and Gevertz, D. (1996). Characterization of 16S rRNA genes from oil field microbial communities indicates the presence of a variety of sulfate-reducing, fermentative, and sulfide-oxidizing bacteria. Appl. Environ. Microbiol. *62*, 1623–1629.

Wacey, D., Kilburn, M.R., Saunders, M., Cliff, J., and Brasier, M.D. (2011). Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. Nat. Geosci. *4*, 698–702.

Wilhelm, S.W., and Suttle, C.A. (1999). Viruses and Nutrient Cycles in the Sea. Bioscience *49*, 8.

Xia, Y., Lü, C., Hou, N., Xin, Y., Liu, J., Liu, H., and Xun, L. (2017). Sulfide production and oxidation by heterotrophic bacteria under aerobic conditions. ISME J. *11*, 2754–2766.

Yang, Y., Xu, G., Liang, J., He, Y., Xiong, L., Li, H., Bartlett, D., Deng, Z., Wang, Z., and Xiao, X. (2017). DNA Backbone Sulfur-Modification Expands Microbial Growth Range under Multiple Stresses by its anti-oxidation function. Sci. Rep. *7*, 3516.

Yeeles, J.T.P., Cammack, R., and Dillingham, M.S. (2009). An iron-sulfur cluster is essential for the binding of broken DNA by AddAB-type helicase-nucleases. J. Biol. Chem. *284*, 7746–7755.

Yin, J., Ren, W., Yang, G., Duan, J., Huang, X., Fang, R., Li, C., Li, T., Yin, Y., Hou, Y., et al. (2016). L-Cysteine metabolism and its nutritional implications. Mol. Nutr. Food Res. *60*, 134–146.

CellPress
OPEN ACCESS

Cell Reports
Article

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Lactococcus phage P087	Université Laval's Félix d'Hérelle Reference Center for Bacterial Viruses	HER#361
Lactococcus lactis subs. lactic C10	Université Laval's Félix d'Hérelle Reference Center for Bacterial Viruses	HER#1361
Escherichia phage T7	ATCC	Cat#BAA-1025-B2
Saccharomyces cerevisiae BY4741	This paper	Available on request
Escherichia coli BL21	This paper	Available on request
Escherichia coli 10G	Lucigen	Cat#60107-1
Escherichia coli BW25113 and BW25113Δ*cysK*	Doug Weibel, UW-Madison	N/A
Critical commercial assays		
PureLink RNA Mini Kit	Ambion	Cat#12183020
DNase Max Kit	QIAGEN	Cat#15200-50
High Capacity cDNA Reverse Transcription Kit	Applied Biosystems	Cat#4368814
Qubit dsDNA BR Assay Kit	Invitrogen	Cat#Q32850
Long Orbitrap LC/MS/MS	University of Wisconsin-Madison Biotechnology Center	N/A
KAPA HiFi	Roche	Cat#KK2101
KAPA2G Robust PCR kits	Roche	Cat#KK5005
EZNA Cycle Pure Kits	Omega Bio-tek	Cat#D6492-01
YeaStar Genomic DNA Extraction kit	Zymo Research	Cat#D2002
Bio-rad MicroPulser	Bio-rad	Cat#165-2100
Deposited data		
Viral genome and AMG sequences; sequence alignments	This paper	https://doi.org/10.5281/zenodo.4947151
Lake Mendota sequence and raw data	Linz et al., 2020	See Table S1 for details
IMG/VR sequences	Paez-Espino et al., 2017	See Table S1 for details
NCBI sequences	NCBI	See Table S1 for details
Protein Data Bank sequences	https://www.rcsb.org/; Berman et al., 2000	4BZQ, 4BZP, 2GOY, 3ZEI, 3SWT, 1RG9
Oligonucleotides		
Primers for qPCR	This paper	See Table S7 for details
Primers for T7 recombination	This paper	See Table S7 for details
Recombinant DNA		
T7::cysK	This paper	N/A
Plasmid: pRS415 for recombination	This paper	N/A
Software and algorithms		
Prodigal v2.6.3	Hyatt et al., 2010	https://github.com/hyattpd/Prodigal
Prokka v1.13.3	Seemann, 2014	https://github.com/tseemann/prokka
Integrated Microbial Genomes and Microbiomes pipeline	Markowitz et al., 2014	https://img.jgi.doe.gov/index.html
InterProScan v65.0, v71.0	Jones et al., 2014	https://www.ebi.ac.uk/interpro/search/sequence/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
VIBRANT v1.2.1	Kieft et al., 2020	https://github.com/AnantharamanLab/VIBRANT
VirSorter v1.0.3	Roux et al., 2015	https://github.com/simroux/VirSorter
Blast	Altschul et al., 1990; Marchler-Bauer et al., 2017	https://blast.ncbi.nlm.nih.gov/Blast.cgi
MAFFT v7.388	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/software/
HMMER v3.1	Eddy, 1998	http://hmmer.org/
BlastKOALA v2.1	Kanehisa et al., 2016	https://www.kegg.jp/blastkoala/
CD-HIT	Fu et al., 2012; Huang et al., 2010; Li and Godzik, 2006	http://weizhongli-lab.org/cd-hit/
Geneious Prime 2019.0.3	N/A	https://www.geneious.com/
dRep v2.6.2	Olm et al., 2017b	https://github.com/MrOlm/drep
RAxML v8.2.4	Stamatakis, 2014	https://github.com/stamatak/standard-RAxML
FigTree v1.4.3	Rambaut, 2009	http://tree.bio.ed.ac.uk/software/figtree/
iRep	Brown et al., 2016	https://github.com/christophertbrown/iRep
Bowtie2 v2.3.4.1	Langmead and Salzberg, 2012	https://github.com/BenLangmead/bowtie2
EasyFig v2.2.2	Sullivan et al., 2011	https://mjsull.github.io/Easyfig/
GhostKOALA v2.0	Kanehisa et al., 2016	https://www.kegg.jp/ghostkoala/
Basemap v1.2.0	Hunter, 2007	https://matplotlib.org/basemap/
Seaborn v0.8.1	N/A	https://seaborn.pydata.org/
Matplotlib v3.0.0	Hunter, 2007	https://matplotlib.org/

RESOURCE AVAILABILITY

Lead contact

Further information and any resources requests should be directed to and will be fulfilled by the lead contact Karthik Anantharaman (karthik@bact.wisc.edu).

Materials availability

The recombinant phage line generated in this study is available upon request.

Data and code availability

All sequences used in this study are publicly available and can be found at their original sources. The genomic and protein sequences of viruses highlighted in this study and respective AMG protein sequences identified can be found on GitHub (https://github.com/AnantharamanLab/Kieft_et_al_2021_organosulfur_AMGs) and Zenodo (https://doi.org/10.5281/zenodo.4947151). This paper does not report original code. Any additional information required to analyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Lactococcus system growth conditions

Lactococcus lactis subs. lactis C10 and Lactococcus phage P087 were obtained from Université Laval's Félix d'Hérelle Reference Center for Bacterial Viruses (Canada, https://www.phage.ulaval.ca). *L. lactis* C10 was grown without agitation at 30°C in M17 broth (Oxoid) supplemented with 0.5% glucose (GM17). Infections were supplemented with 10mM $CaCl_2$ and incubated without agitation at room temperature.

T7 system growth conditions

T7 phage was obtained from ATCC (ATCC® BAA-1025-B2). *Saccharomyces cerevisiae* BY4741 and *E. coli* BL21 are lab stocks, and *E. coli* 10G is a highly competent DH10B derivative (Durfee et al., 2008) originally obtained from Lucigen (60107-1). *E. coli* BW25113 and BW25113Δ*cysK* were obtained from Doug Weibel (University of Wisconsin, Madison).

All bacterial hosts were grown in and plated on LB media (1% Tryptone, 0.5% Yeast Extract, 1% NaCl in dH_2O, plates additionally contain 1.5% agar, while top agar contained 0.5% agar) and LB media was used for all experimentation. All incubations of bacterial

CellPress
OPEN ACCESS

Cell Reports
Article

cultures were performed at 37°C, with liquid cultures shaking at 200-250 rpm unless otherwise specified. Bacterial hosts were streaked on appropriate LB plates and stored at 4°C. *S. cerevisiae* BY4741 was grown on YPD (2% peptone, 1% yeast extract, 2% glucose in dH_2O, plates additionally contain 2.4% agar), after Yeast Artificial Chromosomes (YAC) transformation *S. cerevisiae* BY4741 was grown on SD-Leu (0.17% yeast nitrogen base, 0.5% ammonium sulfate, 0.162% amino acids – Leucine [Sigma Y1376], 2% glucose in dH_20, plates additionally contain 2% agar). All incubations of *S. cerevisiae* were performed at 30°C, with liquid cultures shaking at 200-250 rpm. *S. cerevisiae* BY4741 was streaked on YPD or SD-Leu plates as appropriate and stored at 4°C.

T7 phage was propagated using *E. coli* BL21 after initial receipt from ATCC and then as described on various hosts in methods. All phage experiments were performing using LB and culture conditions as described for bacterial hosts. Phages were stored in LB at 4°C. For long term storage all microbes were stored as liquid samples at −80°C in 10% glycerol, 90% relevant media. SOC (2% tryptone, 0.5% yeast extract, 0.2% 5M NaCl, 0.25% 1M KCL, 1% 1M $MgCl_2$, 1% 1M $MgSO_4$, 2% 1M glucose in dH_2O) was used to recover host and phages after transformation.

For infection experiments, stationary phase cultures were created by growing bacteria overnight (totaling ~20-30 hours of incubation) at 37°C. Exponential phase culture consisted of stationary culture diluted 1:20 in LB then incubated at 37°C until an OD_{600} of ~0.4-0.8 was reached, typically after 40 minutes. Phage lysate was purified by centrifuging phage lysate at 16 g, then filtering supernatant through a 0.22 μM filter. To establish titer, phage samples were serially diluted (1:10 or 1:100 dilutions made to 1 mL in 1.5 mL microcentrifuge tubes) in LB to a 10^{-8} dilution for titering by spot assay. Spot assays were performed by mixing 250 μL of relevant bacterial host in the stationary phase with 3.5 mL of 0.5% top agar, briefly vortexing, then plating on LB plates pre-warmed to 37°C. After plates solidified (typically ~5 minutes), 1.5 μL of each dilution of phage sample was spotted in series on the plate. Plates were incubated and checked every 2-4 hours or overnight (~20-30 hours) to establish a preliminary titer. MOI was estimated by calculated by dividing phage titer by estimated bacterial concentration.

METHOD DETAILS

Identification of viral genomes

A total of 125,842 viral genomes from the Integrated Microbial Genomes/Virus (IMG/VR) (Paez-Espino et al., 2017) v1 database were used for analysis (accessed October 2017). Only publicly available genomes > 5kb analyzed by Paez-Espino et al. (2016) were used in this study. Open reading frames were predicted using Prodigal with default parameters (v2.6.3) (Hyatt et al., 2010). All viral genomes were annotated using a combination of Prokka (v1.13.3) (Seemann, 2014), Integrated Microbial Genomes and Microbiomes pipeline (Markowitz et al., 2014), and InterProScan (v65.0) (Jones et al., 2014). Contigs with a high ratio of bacterial to viral protein annotations were manually identified and discarded. Contigs were further validated and annotated using a combination of VIBRANT (v1.2.1) and VirSorter (v1.0.3, virome database, categories 1, 2, 4, 5) (Kieft et al., 2020; Roux et al., 2015). All viral genomes encoding AMGs were manually inspected. Additional viral genomes were identified on the National Center for Biotechnology Information (NCBI) RefSeq (Brister et al., 2015; O'Leary et al., 2016; Tatusova et al., 2016) or GenBank database (Clark et al., 2016) (accessed Jan 2019) by querying viral genomes for AMGs of interest by blastp domain analysis (Altschul et al., 1990; Marchler-Bauer et al., 2017). Approximately 9,500 genomes corresponding to the viral classification *Caudovirales* were searched. VIBRANT and VirSorter were used to identify viruses > 5kb from Lake Mendota, WI.

AMG identification and annotation

In-house hidden Markov model (HMM) profiles were built corresponding to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of organosulfur Metabolism as well as Cysteine and Methionine Metabolism (accessed December 2018) (Kanehisa and Goto, 2000). The two pathways' KEGG Orthology (KO) numbers (189 total) were used to access corresponding proteins from the UniProt database (release 2018_11) (The UniProt Consortium, 2018). The resulting proteins were aligned with MAFFT (v7.388, default parameters) (Katoh and Standley, 2013) and HMM profiles were built using hmmbuild (HMMER v3.1, default parameters) (Eddy, 1998). HMM profiles for CysC and CysH were built in the same manner, except manually verified viral CysC and CysH sequences, respectively, were added to the alignment for robustness. Hmmsearch (HMMER v3.1, evalue < 1e-5) was used to scan proteins on viral genomes. Proteins identified by the in-house HMM profiles were uploaded to the KEGG BlastKOALA server (v2.1) (Kanehisa et al., 2016) and queried under "prokaryotes" taxonomy and "genus prokaryotes" database for best hit annotations. Proteins annotated according to the original 189 KO numbers were selected for further verification. Manual verification of several representatives of each KO number (i.e., protein family) was done to curate the results using blastp (NCBI non-redundant database, accessed Jan 2019) and InterProScan (v71.0) to check for the presence of all expected conserved domains. Individual proteins and protein families of irrelevance or incorrect annotation were removed.

Sequence alignment and dN/dS analysis

Alignment of CysH, CysK, CysC, TauD and MetK sequences was performed using MAFFT (v7.388, default parameters). For *cysH*-encoding genomes identified from NCBI, all viral sequences were used. Host genomes were scanned, by annotation and blastp domain analysis, for multiple copies of *cysH* and all those identified were used, along with non-host bacterial sequences that were found to be highly similar to viral sequences according to pairwise identity. For the remaining alignments, all viral AMG protein

Cell Reports
Article

CellPress
OPEN ACCESS

sequences that shared at least 95% pairwise identity were restricted to one representative using CD-HIT (accessed Jan 2019) (Fu et al., 2012; Huang et al., 2010; Li and Godzik, 2006) and aligned. Viral CysK and CysH sequences were limited to lengths 200-330 and 117-600 amino acids, respectively. To obtain bacterial representatives, the majority consensus sequence of aligned viral proteins was queried against the NCBI RefSeq database by blastp (evalue < 1e-5). In order to ensure broad phylogenetic distribution of blastp results, the output was restricted to the top 500 hits from each of five phylogenetic groups based on NCBI categorization: [1] Proteobacteria, [2] Terrabacteria, [3] FCB superphlylum, [4] PVC superphylum and [5] a group containing all other phyla. The resulting sequences were manually limited to specific lengths to match viral sequences (CysC: 210-360, CysH: 150-600, CysK: 269-400, TauD: 314-400 amino acids, MetK: all) and reduced to one representative per 50% pairwise identity using CD-HIT. Viral and bacterial representatives were aligned together using MAFFT (default parameters) and gaps were stripped by 98%. The resulting alignments were used for phylogenetic analysis. Visualization of alignments was done using Geneious Prime 2019.0.3. For reference to full virus protein name and genome, see Table S1.

The AMGs for *cysK, cysC, cysD, cysH, tauD, msmA, metK, mtmN and luxS* were used to calculate *dN/dS* ratios. dRep (v2.6.2) was used to compare AMG sequences separately (dRep compare–SkipMash–S_algorithm goANI) and dnds_from_drep.py was used to calculate *dN/dS* ratios from the AMG pairs (Olm et al., 2017b). The *dN/dS* ratios were visualized with Seaborn (v0.8.1) and Matplotlib (v3.0.0).

Sequence phylogeny

Phylogenetic analysis was performed using protein alignments of CysH, CysK, CysC, TauD and MetK as described above. To infer phylogenetic relationships RAxML (v8.2.4) (Stamatakis, 2014) was used with the following parameters: raxmlHPC-PTHREADS -N 100 -f a -m PROTCATLG. Resulting best trees were used and rooted by manual identification of most distant (outgroup) taxa. Trees were visualized using FigTree (v1.4.3) (Rambaut, 2009).

Protein functional analysis

For domain and residue analysis, phylogenetic trees were used as a reference to select representative viral and bacterial sequences, which were then aligned using MAFFT (default parameters). Annotations of functional amino acid residues were labeled according to the Protein Data Bank (PDB, accessed January 2019) (Berman et al., 2000) with the following identification numbers: 4BZQ and 4BZP (CysC), 2GOY (CysH), 3ZEI (CysK), 3SWT (TauD), and 1RG9 (MetK). For alignments with no phylogenetic tree, up to five viral sequences and five PDB homologs (when available) were randomly selected for all AMGs with abundance of five or greater. The PDB sequences used for annotation were added to the alignment. N- and C-terminal ends of protein alignments were manually removed for clarity and gaps were stripped by 90% (for alignments with phylogenetic trees) or 80% (for all others). Residues were highlighted according to 85% pairwise identity between sequences, excluding sequence gaps. An identity graph, generated by Geneious, was fitted to the alignment to visualize pairwise identity of 100% (green), 99%–30% (yellow) and 29%–0% (red).

Protein Reactions

Enzymatic reactions, diagrams and pathways were created by referencing KEGG and MetaCyc (v22.6) (Caspi et al., 2012) annotations.

Viral transcriptomics and growth rates

Publicly available metatranscriptomic data from Lake Mendota, WI was assessed for AMGs by querying annotation names (Linz et al., 2020). This gene expression data comprises a two-day time series and is accompanied by metagenomic assemblies (IMG Taxon Object IDs 3300013004 and 3300013005). Metatranscriptomic reads were mapped to a custom, non-redundant database of freshwater reference data, including the metagenome assemblies; annotations in this study are derived from the annotations of the reference database. We used read counts normalized to transcripts per liter as the input for our study, and we searched for AMGs in the metagenomic assemblies as described above.

The growth rate of the *cysC*-encoding Lake Mendota virus was identified using index of replication (iRep) with default parameters (Brown et al., 2016). Metagenomic assembly reads used for iRep are available on IMG under the Taxon Object ID 3300013005. Reads were mapped to the viral genome using Bowtie2 (v2.3.4.1) (Langmead and Salzberg, 2012). GC-skew to indicate rolling circle replication of the viral genome was likewise completed using the iRep toolkit.

Virus growth and fitness assay

Approximately 10^8 plaque forming units (PFUs) of Lactococcus phage P087 (approximate multiplicity of infection (MOI) of 1 were used to infect 1mL of *L. lactis* C10 which had been brought to an optical density (OD_{600}) of approximately 0.15 in GM17 broth. For fitness experiments, either vehicle control (water), 10μM Na_2S or 100μM Na_2S was supplemented to the media at time of infection. Infections were incubated without agitation at room temperature for approximately three hours. Additional cultures of uninfected *L. lactis* C10 with all other variables identical were measured for growth at the endpoint of infections using OD_{600}. To end infections, *L. lactis* C10 were spun out of solution at 10,000 r*cf.* and the supernatant (i.e viral fraction) was removed and cooled to 4°C. Plaque assays were done using the standard double agar method (Lillehaug, 1997) with diluted viral fraction and *L. lactis* C10 brought to high concentration. A 1% bottom agar and 0.4% top agar of GM17 were used, both supplemented with 0.5% glycine and 10mM $CaCl_2$.

Virus and host cysK qPCR assay

An overnight culture of *L. lactis* C10 was diluted in GM17 broth to OD 0.08 and grown at 30°C for ~2 hours until OD reached 0.15. In a batch culture 10mM $CaCl_2$ was added. Two different conditions were assayed, each in duplicate (biological replicates): (1) *L. lactis* C10 control and (2) *L. lactis* C10 plus Lactococcus phage P087. For infection conditions, Lactococcus phage P087 was added at a MOI of 1 (time 0 minutes). RNA was extracted using the PureLink RNA Mini Kit (Ambion) from 500μL of the cellular fraction at 15, 60 and 120 minutes post-infection. RNA was then treated with DNase with the DNase Max Kit (QIAGEN) and converted to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). qPCR of viral and host *cysK* was performed using *Power* SYBR Green PCR Master Mix (Applied Biosystems) with 7ng of cDNA template and the following primer sets (IDT): *L. lactis* C10 forward (CCTTCGTTGGCTCTGCTTTG), *L. lactis* C10 reverse (TGGCATCATCTCCTTTGACCC), Lactococcus phage P087 forward (CAGAAACTATCGGAAACACACCAC), and Lactococcus phage P087 reverse (TTGAGTGAATGACCTGCTCCA) (Table S10). The concentration of template cDNA was measured with the Qubit dsDNA BR Assay Kit (Invitrogen). The viral and host *cysK* sequences were sufficiently dissimilar in sequence identity (< 60% at the protein level) to allow for accurate distinction by qPCR and the primers selected.

Mass spectrometry and protein identification

L. lactis C10 was grown without agitation at 30°C in modified M17 broth supplemented with 0.5% glucose (mGM17). mGM17 was made by adding 1.25 g glucose, 0.625 g tryptone, 1.25 g peptone, 0.125 g yeast extract, 0.125 g ascorbic acid, 0.0626 g anhydrous magnesium sulfate and 4.75 g disodium glycerophosphate to 250mL deionized water. Approximately 10^8 PFUs of Lactococcus phage P087 were used to infect 3mL of *L. lactis* C10 which had been brought to OD_{600} of approximately 0.15 and supplemented with 10mM $CaCl_2$. Infections proceeded to complete lysis without agitation at room temperature for approximately three hours. To end the infection, *L. lactis* C10 were spun out of solution at 10,000 *rcf.* and the supernatant was removed and stored at 4°C. The supernatant was size fractionated by filtration for the 100kDa to 10kDa size fraction before trypsin solution digestion and analysis by Long Orbitrap LC/MS/MS (University of Wisconsin-Madison Biotechnology Center).

Genome organization and comparisons

Genome organization was visualized using Geneious Prime. Genes were manually colored by referencing functions according to NCBI RefSeq or GenBank annotation, or blastp search. Viral genomes in GenBank format were compared and visualized with EasyFig (v2.2.2) (Sullivan et al., 2011) using the tblastx function. Only tblastx (v2.8.1+) hits with percent identities greater than 30% and e-values less than 0.001 are shown. Remaining analysis parameters were set to default. Circular sequences were visualized linearly for ease of comparison.

Geographical distributions

IMG Taxon Object ID numbers were used to identify global coordinates of studies in which AMGs were identified. Coordinates were mapped using Matplotlib's Basemap (v1.2.0) (Hunter, 2007). Human studies were excluded from coordinate maps.

Host classification

GhostKOALA (v2.0) (Kanehisa et al., 2016) with the "genus prokaryotes" database was used to query all 3,794 AMG-encoded proteins identified from IMG/VR derived viruses (3,421 annotated and used for taxonomy). To benchmark accuracy of the analysis, all 282 AMG-encoded proteins identified from NCBI-derived viruses with known hosts were queried in the same manner (278 were annotated and used for taxonomy) and compared to the taxonomy of hosts.

T7 recombination: cloning

All primers can be found in Table S10. PCR was performed using KAPA HiFi (Roche) for all experiments with the exception of multiplex PCR for screening Yeast Artificial Chromosomes (YACs), which was performed using KAPA2G Robust PCR kits (Roche). DNA purification was performed using EZNA Cycle Pure Kits (Omega Bio-tek) using the centrifugation protocol. YAC extraction was performed using YeaStar Genomic DNA Extraction kits (Zymo Research). All cloning was performed according to manufacturer documentation except where noted in methods. PCR reactions using phage as template use 1 μL of undiluted phage stock, with extension of the 95°C denaturation step to 5 minutes.

Electroporation of YACs was performed using a Bio-rad MicroPulser (165-2100), Ec2 setting (2 mm cuvette, 2.5 kV, 1 pulse) using 50 μL competent cells and 2 μL YAC DNA for transformation. Electroporated cells were immediately recovered with 950 μL SOC, then incubated at 37°C for 1 to 1.5 hours and plated or grown in Lb.

E. coli 10G competent cells were made by adding 8 mL overnight 10G cells to 192 mL SOC (with antibiotics as necessary) and incubating at 21°C and 200 rpm until ~OD_{600} of 0.4 as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation (actual incubation time varies based on antibiotic, typically overnight). Cells are centrifuged at 4°C, 800-1000 g for 20 minutes, the supernatant is discarded, and cells are resuspended in 50 mL 10% glycerol. Centrifugation and washing are repeated three times, then cells are resuspended in a final volume of ~1 mL 10% glycerol and are aliquoted and stored at −80°C. Cells are competent for plasmid and YACs. All primers used in experiments in this publication are listed in supplemental.

Cell Reports
Article

CellPress
OPEN ACCESS

T7 recombination: engineering T7 with cysK

Phages were assembled using YAC rebooting (Ando et al., 2015; Jaschke et al., 2012), which requires yeast transformation of relevant DNA segments, created as follows. A prs415 yeast centromere plasmid was split into three segments by PCR, separating the centromere and leucine selection marker, which partially limits recircularization and improved assembly efficiency (Kuijpers et al., 2013). Wild-type T7 segments were made by PCR using wild-type T7 as template. *CysK* segments were made by colony PCR of BW25113. *CysK* was inserted into two locations to create two phage constructs. The first location was replacement of gp1.7 to establish *CysK* in early Class II genes. This insertion causes a two amino acid extension (YE) of the immediate 5′ gene gp1.6 that was not anticipated to have an effect on phage viability. The second location was inserted adjacent to gp6.3 to establish *CysK* in early class III genes and leverages a copy of phage promoter phi6.5 for expression.

DNA parts were combined together (0.1 pmol/segment) and transformed into *S. cerevisiae* BY4741 using a high efficiency yeast transformation protocol (Gietz and Woods, 2002) using SD-Leu selection. After 2-3 days colonies were picked and directly assayed by multiplex colony PCR to assay assembly. Multiplex PCR interrogated junctions in the YAC construct and was an effective way of distinguishing correctly assembled YACs. Correctly assembled YACs were purified and transformed into *E. coli* 10G cells and these cultures incubated until lysis, after which phages were purified to create the initial phage stock.

T7 recombination: passaging and AMG retention

Either T7Δ1.7::cysK or T7::cysK phages were added to 5 mL exponential phase BW25113 or BW25113Δ*cysK* at an estimated MOI of 10^{-4} to allow for an estimated three phage passages. After the culture had fully lysed, typically ~1 hour and 30 minutes, lysate was purified and then the titer established by spot assay. This process was then repeated twice for a total of an estimated 9 phage passages assuming at least 100 phage progeny per host. Phage lysate from the final passage was used as template for sequencing to determine if the *cysK* insert remained as the consensus sequence in the phage population. The entire process was repeated in biological triplicate for both host and phage combinations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Virus growth and fitness

The number of resulting plaques from the growth and fitness assays were normalized to 100% of controls for each experiment. Three independent experiments with three biological infection replicates and two biological growth replicates each was performed. Further information of experiments can be found in Method Details below.

Virus and host cysK qPCR

For each replicate of the two conditions assayed both primer sets were used for qPCR. To analyze the qPCR results, the Cq readings were averaged between the three replicates for each treatment at each time point to obtain a single datapoint per treatment:primer pair per time point, termed *average Cq*. Using time point zero for the uninfected *L. lactis* C10 condition with *L. lactis* C10 *cysK* primers as the baseline *control*, delta-delta-Cq values were calculated by subtracting the *control* value from the *average Cq* values. This result calculates the expression of *L. lactis* C10 *cysK* at time point zero to be normalized to zero (delta-delta-Cq of zero). Finally, all delta-delta-Cq values were transformed using the formula $2^{-(\text{delta-delta-Cq})}$ (Livak and Schmittgen, 2001). All raw Cq values and normalized results, including equations, can be found in Table S6. Further information of experiments can be found in Method Details below.

Cell Reports, Volume *36*

Supplemental information

Virus-associated organosulfur metabolism in human and environmental systems

Kristopher Kieft, Adam M. Breister, Phil Huss, Alexandra M. Linz, Elizabeth Zanetakos, Zhichao Zhou, Janina Rahlff, Sarah P. Esser, Alexander J. Probst, Srivatsan Raman, Simon Roux, and Karthik Anantharaman

Virus-associated organosulfur metabolism in human and environmental systems

Kristopher Kieft[1], Adam M. Breister[1], Phil Huss[1,2], Alexandra M. Linz[3], Elizabeth Zanetakos[1], Zhichao Zhou[1], Janina Rahlff[4†], Sarah P. Esser[4], Alexander J. Probst[4], Srivatsan Raman[1,2], Simon Roux[5], Karthik Anantharaman[1,6,*]

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA
[2]Department of Biochemistry, University of Wisconsin–Madison, Madison, WI, USA
[3]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI, USA
[4]Department of Chemistry, Environmental Microbiology and Biotechnology, University of Duisburg-Essen, Essen, Germany
[5]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[6]Lead Contact
[*]Correspondence: karthik@bact.wisc.edu, Address: 4550 MSB, 1550 Linden Dr., Madison, WI, 53706
[†]Present address: Janina Rahlff, Centre for Ecology and Evolution in Microbial Model Systems, Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden

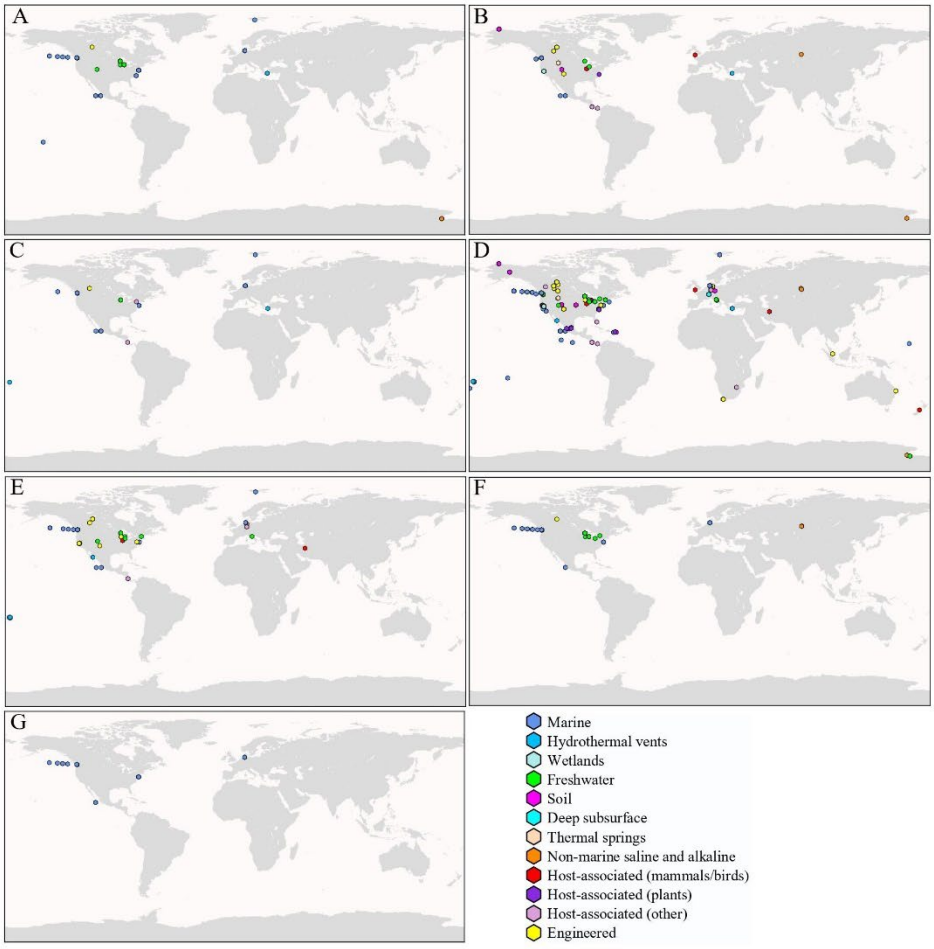


Figure S1. Distribution of individual AMGs in the environment, related to Figure 2. Global distribution of viral populations encoding (A) *cysC*, (B) *cysH*, (C) *cysK*, (D) *dcm*, (E) *metK*, (F) *tauD* or (G) *msmA* identified on the IMG/VR database, color coordinated by environment classification.

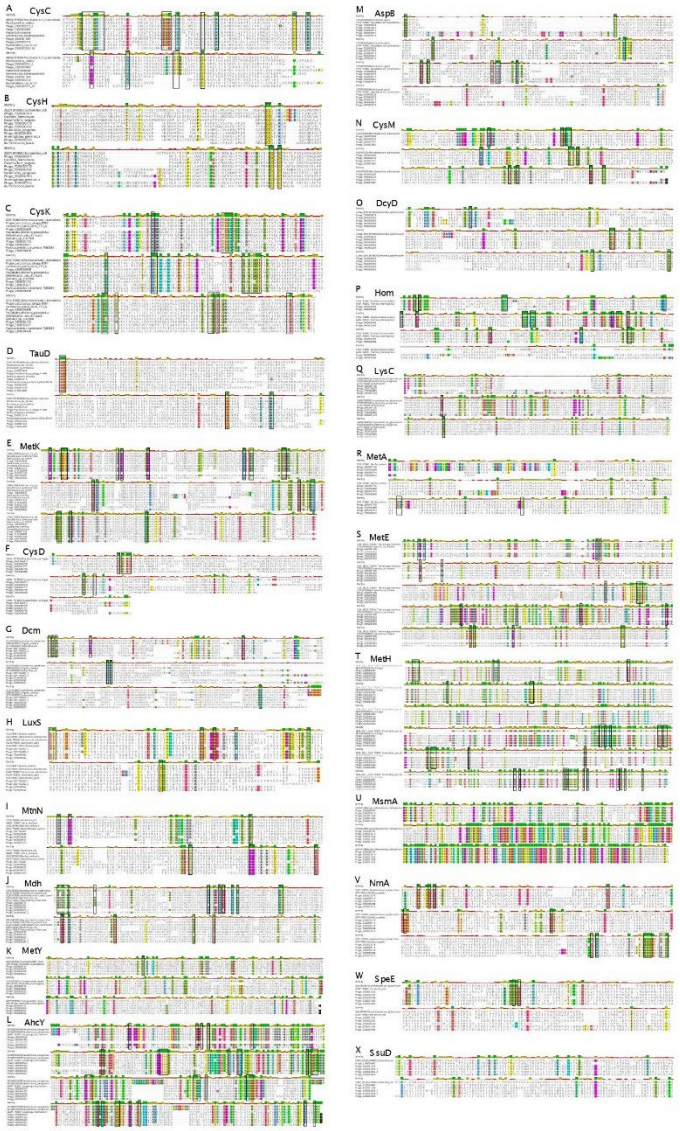


Figure S2. Amino acid alignments of proteins encoded by AMGs, related to Figure 1 and Table 1. Alignment of representative viral and bacterial sequences for all AMGs with abundances greater than five. Viruses are indicated by the preface "Phage" followed by their respective IMG Taxon Object ID number. See Table S1 for full genome names. Refer to Figurer S4 for phylogeny of the represented sequences for A-E. Highlighted amino acids indicate conservation in >85% of sequences shown. Black boxes indicate amino acid residues that are biochemically verified as functional according to the given PDB reference sequence.

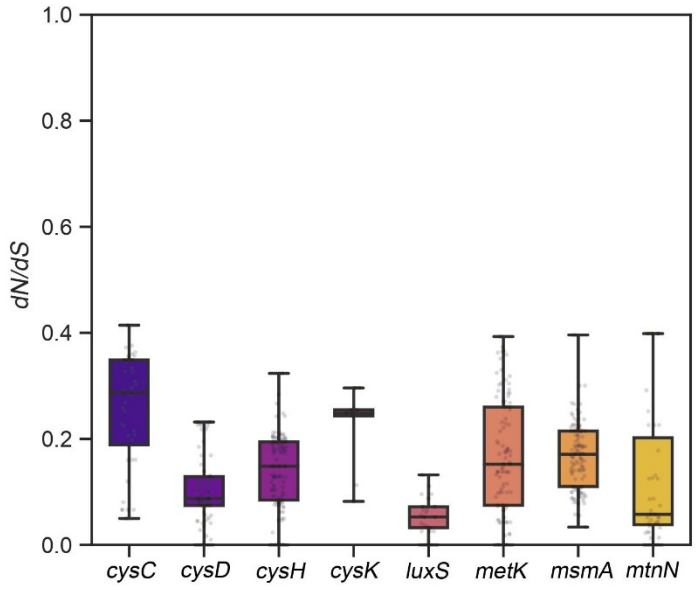


Figure S3. *dN/dS* ratio calculations for cysK, cysC, cysD, cysH, tauD, msmA, metK, mtmN and luxS AMG pairs, related to Figures 1, 4 and 5. Each data point overlaid on the standard box plot represents a single AMG pair. The figure was generated using seaborn and matplotlib.

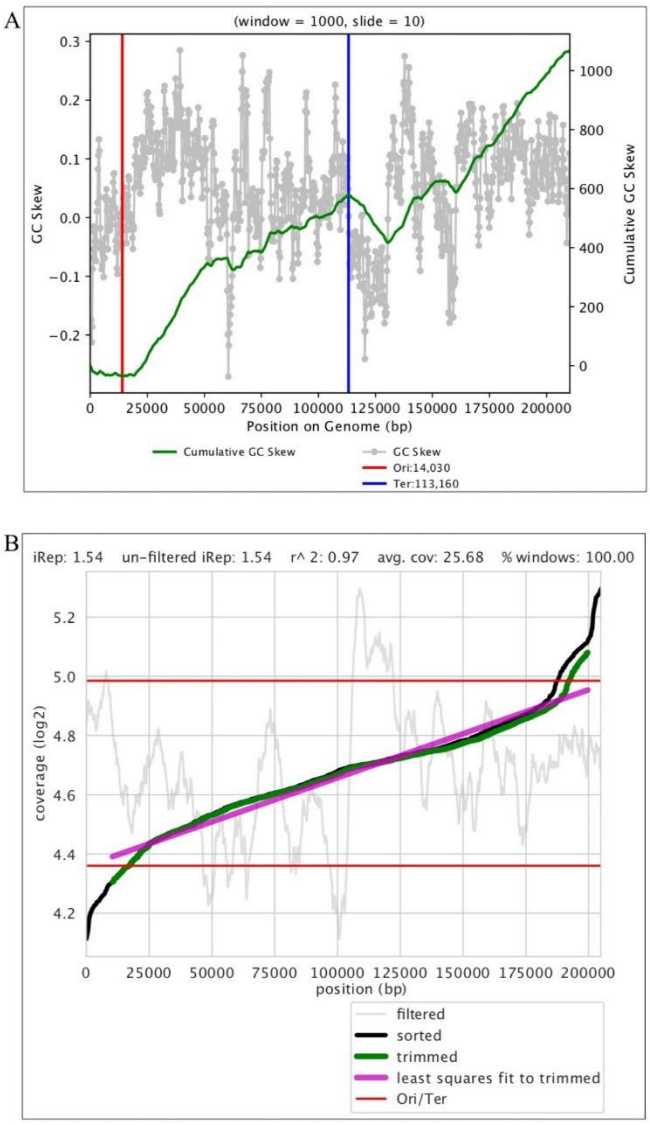


Figure S4. Genome and growth statistics of a complete virus identified to express *cysC* in Lake Mendota, WI, related to Figure 6. The (A) GC-skew and (B) index of replication for a complete, circular virus identified in Lake Mendota, WI. GC-skew and replication statistics indicate that the virus was actively replicating at time of sampling and likely undergoes rolling circle replication.

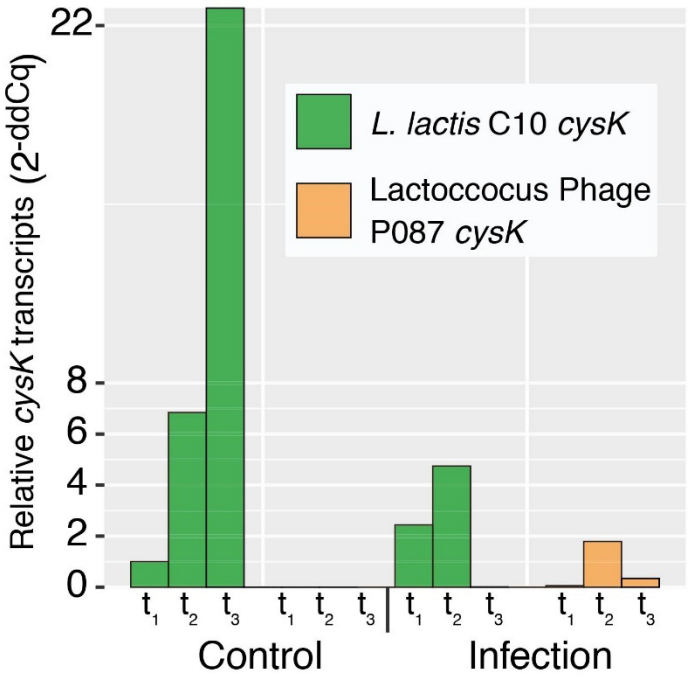


Figure S5. qPCR results of *cysK* transcript relative abundance for *L. lactis* C10 and phage P087, related to Figure 3. Relative transcript abundance is provided in the normalized 2^{-ddCq} metric. Control is *L. lactis* C10 (host) alone and infection includes host plus phage P087. The times shown are t_1 (15 minutes), t_2 (60 minutes) and t_3 (120 minutes), each in duplicate.

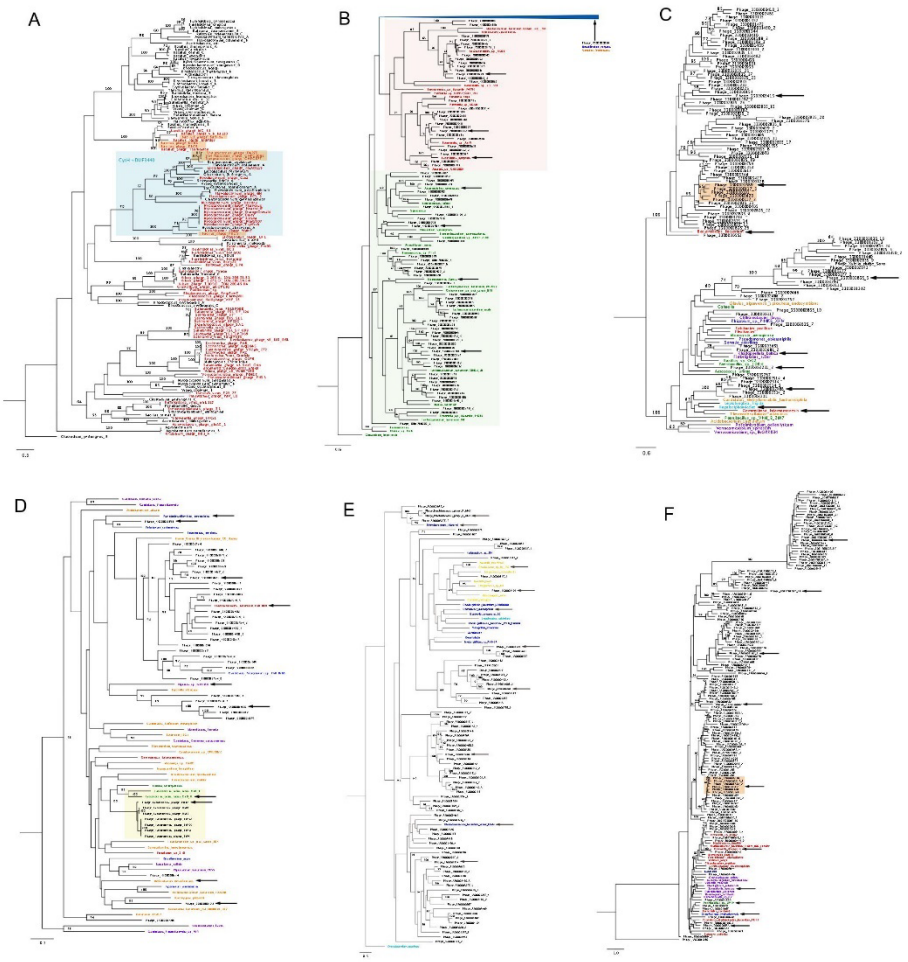


Figure S6. Phylogeny of viral AMG encoded protein sequences, related to Figures 1, 4 and 5. (A) Phylogenetic tree for CysH of complete viruses with known bacterial hosts. Viruses are in red and bacteria are in black, and proteins with an additional DUF4440 domain are highlighted in blue. Bacteria with multiple copies of *cysH* are appended with a letter ("A" through "D"). Refer to Table S1 for virus-host associations. Also shown are phylogenetic trees of uncultivated viruses with bacterial homologs and select cultivated viruses for (B) CysH (red and green highlighting refers to putative virus-host associations; compressed blue clade contains 36 viruses and 89 bacteria from several phyla), (C) CysC, (D) CysK (yellow highlighting refers to known virus-host associations), (E) TauD and (F) MetK. For trees (B-F) colored names refer to viruses (black), Bacteroidetes and other members of the FCB superphylum (red), Cyanobacteria (cyan), Verrucomicrobia and Planctomycetes and other members of the PVC superphylum (purple), Actinobacteria (yellow), and all other phyla in orange. For all trees, bootstrap values greater than 60 are shown, orange highlighting indicates respective genomes used for comparative genomics (see Figures 4 and S6), and arrows indicate sequences used for protein alignments (see Figure S2).

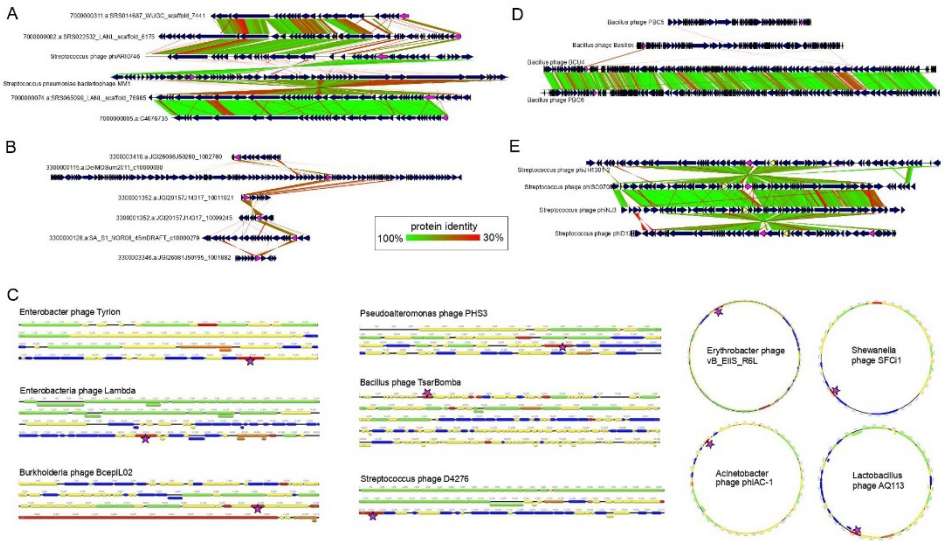


Figure S7. Genome comparisons of viruses encoding AMGs, related to Figures 4 and 5. (A) genome alignments of incomplete viruses and complete *Streptococcus pneumoniae* viruses encoding *dcm* (pink), (B) genome alignments of incomplete viruses encoding *metK* (pink), (C) genome organization of linear and circular complete viral genomes that encode *cysH*, (D) genome alignments of complete *Bacillus cereus* viruses encoding *cysH* (pink), and (E) genome alignments of complete *Streptococcus suis* viruses encoding *metK* (yellow) and *dcm* (pink). For genome alignments, predicted open readings frames are annotated by dark blue arrows and genomes are connected with lines according to tblastx identity. Refer to Figure S6 for phylogeny of AMGs for (B) and (D) (orange highlighting). For the genome organizations, arrows indicate open reading frames and are annotated by general function: virion structural assembly (green), auxiliary metabolism and general functions (red), nucleotide metabolism and genome replication (blue), lysis (orange) and unknown function (yellow). Pink stars indicate the location of *cysH*. Refer to Table S1 for virus details.

Appendix 2

Publication

Engineering a Dynamic Controllable Infectivity Switch in Bacteriophage T7

ACS SyntheticBiology

pubs.acs.org/synthbio Research Article

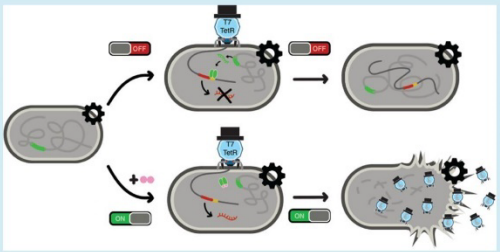
Engineering a Dynamic Controllable Infectivity Switch in Bacteriophage T7

Chutikarn Chitboonthavisuk,[#] Chun Huai Luo,[#] Phil Huss, Mikayla Fernholz, and Srivatsan Raman*

Cite This: *ACS Synth. Biol.* 2022, 11, 286−296 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Transcriptional repressors play an important role in regulating phage life cycle. Here, we examine how synthetic transcription repressors can be used in bacteriophage T7 to create a dynamic, controllable infectivity switch. We engineered T7 phage by replacing a large region of the early phage genome with different combinations of ligand-responsive promoters and ribosome binding sites (RBS) designed to control the phage RNA polymerase, *gp1*. Phages with engineered infectivity switch are fully viable at levels comparable to wildtype T7, when not repressed, indicating the phage can be engineered without loss of fitness. The most effective switch used a TetR-responsive promoter and an attenuated RBS, resulting in a 2-fold increase in latent period and a 10-fold decrease in phage titer when repressed. Phage activity can be further tuned using different inducer concentrations. Our study provides a proof of concept for how a simple synthetic circuit introduced into the phage genome enables user control over phage infectivity.



KEYWORDS: *engineered bacteriophages, ligand-responsive repression system, pseudolysogenic, enterobacteria phage T7*

INTRODUCTION

Bacteriophages (or "phages") are obligate parasites that require a bacterial host to complete their life cycle.[1] Once a phage infects its host, a choreographed cascade of phage genes is expressed to regulate subsequent steps in the phage's life cycle. Gene regulation in phages is frequently controlled by transcriptional repressors.[2] Transcriptional repressors act as switches that determine lifestyle decisions (e.g., lytic vs lysogenic) by silencing or activating different sets of genes. This is best exemplified in the transcriptional repressor-based genetic switch in phage lambda and other temperate phages with similar mechanisms (P22,[3] 434,[4] ϕC31[5]). Phage lambda persists in lysogenic state when transcriptional repressor CI represses early stage lambda promoters, halting the transcriptional cascade of lytic genes. To activate lytic genes and release the prophage, a different repressor Cro counteracts CI through differential binding at the same promoters. Transcriptional repressors also control activation and inactivation of non-integrating, episomal phage genomes, called pseudolysogens, in a nutrient dependent manner.[6] Transcriptional repressors are valuable tools for phages because their mechanism of gene regulation simply relies on steric obstruction of the RNA polymerase making it largely host-independent.[7] Further, a transcriptional repressor with strong affinity for its promoter can exert tight regulation over multiple co-cistronic open reading frames commonly found in long phage operons. While natural transcription regulation in phages is well studied, the engineering rules of introducing synthetic regulation into phages has not been developed.

Bacteria have long served as a popular chassis for exploring and prototyping synthetic regulation. This choice was largely driven by the availability of tools for bacterial genome engineering and the application goal of bacterial biomanufacturing.[8] In recent years, a strong impetus has emerged for engineering phages, driven similarly by the availability of new tools for phage genome engineering and their potential applications in medicine and biotechnology.[9−11] Phage genomes can be now edited with high precision from single base to kilobase resolution using yeast cloning, homology-directed repair, and enzymatic recombination.[12,13] Phages engineered using these approaches could be powerful tools for killing antibiotic resistant bacteria and precisely manipulating microbiomes with applications in agriculture, livestock, medicine, and environment. Engineered phages have many advantages over their natural counterparts including superior efficacy, greater programmability, higher compositional stability, and easier scalability of production.[14] Despite this promise,

Received: August 27, 2021
Published: January 5, 2022

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

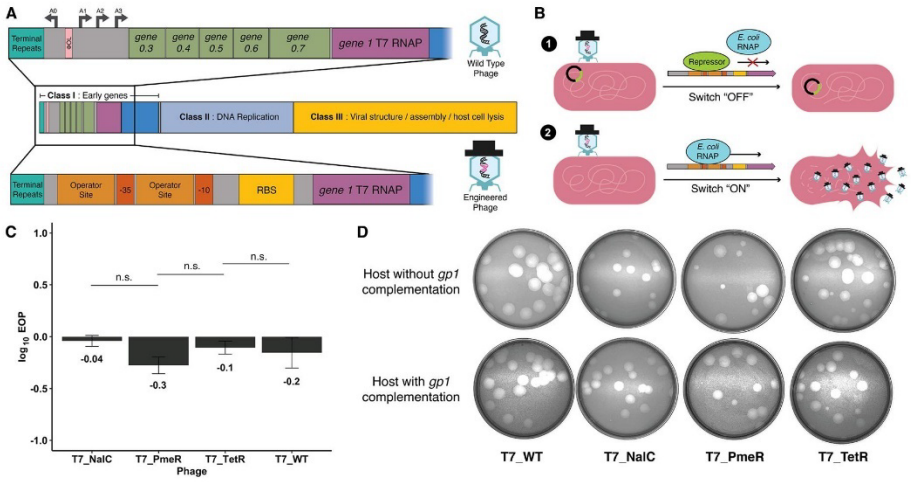


Figure 1. Engineered phages retain wildtype infectivity. (A) Schematic illustration of the T7 bacteriophage genome (middle) with early genes shown in close-up in wildtype T7 (above) and engineered T7 (below). The portion of the phage genome shown boxed in black is engineered with a replacement synthetic promoter and ribosome binding site (RBS). (B) illustration of the phage "OFF" state, where repressor (green) prevents RNA polymerase (RNAP, blue) binding and blocks *gp1* transcription, preventing host lysis. In the "ON" state, infection without the repressor allows RNAP to transcribe *gp1*, resulting in phage replication and host lysis. (C) Ability of engineered T7 phages to infect *E. coli* by Efficiency of Plating (EOP) using *E. coli* complementing *gp1* as a reference host. Compared to wildtype T7 (T7_WT), T7 engineered with NalC, PmeR, and TetR (T7_NalC, T7_PmeR, and T7_TetR, respectively) show no significant difference (n.s.) in ability to plaque. Data are represented as mean ± SD of biological triplicates. (D) Plaque morphology for wildtype and engineered phages shown by plaque assay on wildtype *E. coli* and *E. coli* with *gp1* complemented host after a 19-h incubation. Plaques for engineered phages retain wildtype plaque morphology.

even basic rules of engineering new regulation into phages have not been developed. This is in stark contrast to the wealth of research on engineering promoters, switches, circuits, and pathways in bacteria. Rules for engineering bacteria are not directly transferrable to phages due to substantial differences in genome compactness, compartmentalization of regulation, and kinetics of growth and replication. Thus, rules for engineering new regulation for phages must be developed by systematic design-build-test-learn analysis.

With this goal in mind, we examined how bacterial transcriptional repressors could be used to engineer synthetic genetic regulation in phages. We sought to engineer a repressor-based infectivity switch that can dynamically control the activity of an obligate lytic phage in a ligand-dependent manner. Such a system would provide direct user control of phage activity which would otherwise undergo unregulated exponential amplification upon infecting a host.[15] Natural phages such as lambda are not ligand-inducible but instead rely on stochastic differences in the concentration of CI vs Cro to determine lysogenic vs lytic choice.[16] A dynamic controllable infectivity switch would be a valuable tool for activating a phage at user-defined times to carry out microbiome editing in natural or synthetic communities. These engineered phages could be tailored to remove pathogens or to target specific strains in the gut microbiome. Engineered promoters could be designed to induce activation of phages only in the presence of specific metabolites found in certain regions of the gut or that are produced by targeted strains. Furthermore, inducible phage would be an effective form of biocontainment, as the phages only remain active while the ligand is provided but otherwise remain inert.

Here, we engineered an obligate lytic T7 phage with a synthetic regulatory switch to control its infection cycle in a ligand-dependent manner. To introduce synthetic regulation, we removed a large tract of native regulatory sequence from the phage genome and replaced this region with a short ligand-responsive bacterial promoter that regulates the T7 RNA polymerase gene. We saw no appreciable loss of fitness in the synthetic phage relative to wildtype T7 (T7_WT). We tested different ligand-responsive bacterial promoters and ribosome-binding sites (RBS) representing a range of expression levels to characterize how these variables affect phage activity. To attenuate phage replication, we measured the phage latent period and change in phage titer, and found that the strongest attenuation of phage replication, or the "OFF" state, occurred with a Tet-regulated promoter and a very weak RBS, which resulted in a 2-fold increase in latent period and an approximately 10-fold decrease in a change in phage titer. Ligand induced activation to the "ON" state restored activity of synthetic phages to levels comparable to unregulated T7_WT phage. Our study provides a basic proof-of-concept for recoding a phage genome with synthetic regulation, paving the way of engineering more sophisticated circuitry to enable phages to carry out complex, user-defined tasks.

■ RESULTS

Refactored Synthetic Phages Retain Wildtype Infectivity. We sought to create a simple, dynamically controllable infectivity switch for phages using ligand-responsive transcriptional repressors naturally utilized for gene regulation in bacteria. To test our synthetic switch, we chose bacteriophage T7, a well characterized, prototypical obligate lytic phage that infects *Escherichia coli*. Bacterial

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

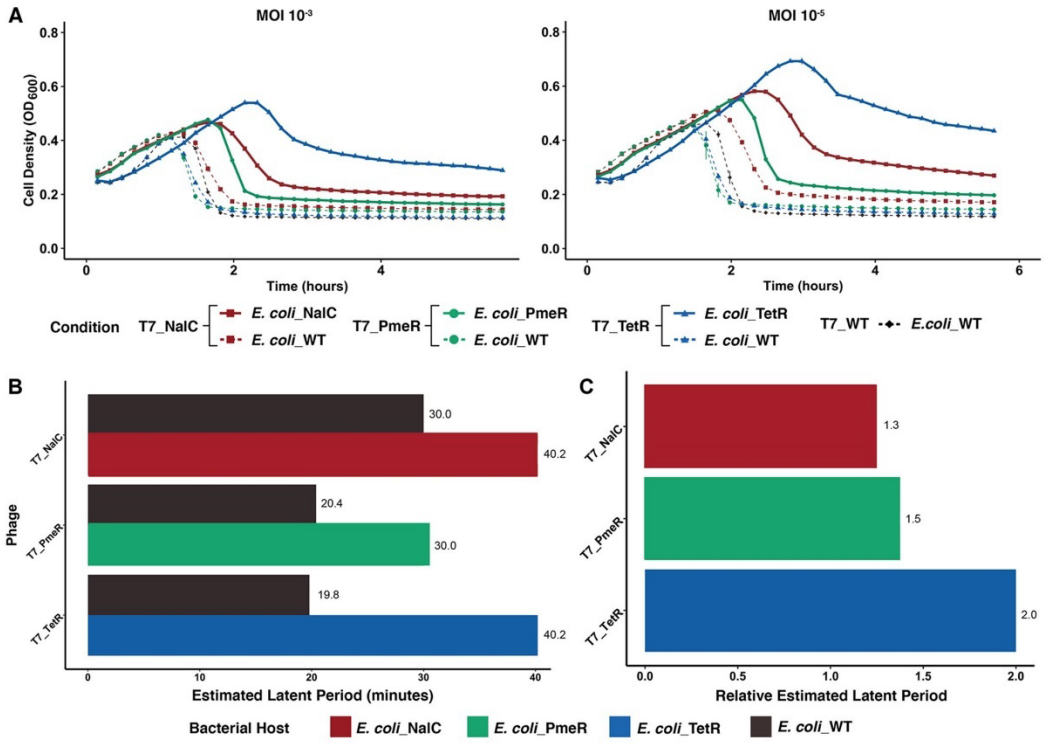


Figure 2. Repression of *gp1* reduces activity in engineered phages. (A) Bacterial cell density over time measured by absorbance (OD_{600}) of wildtype host (*E. coli*_WT; dotted lines) and host expressing NalC, PmeR, or TetR repressor (*E. coli*_NalC, *E. coli*_PmeR, *E. coli*_TetR, respectively; solid lines) after addition of T7_NalC (red), T7_PmeR (green), or T7_TetR (blue) repressors or T7_WT (black). Phages were applied at time 0 at a MOI of 10^{-3} and 10^{-5}. Phages with synthetic promoters showed delayed lysis on hosts expressing the cognate repressor but have a comparable estimated latent period in hosts without repressor. All data represented as mean OD_{600} ± SD in technical triplicates. (B) The estimated latent period of engineered phages on wildtype host compared to host expressing the cognate repressor, as determined by comparing the inflection in the growth curves between two MOIs where the difference is equal to the phage burst size. (C) The relative estimated latent period of engineered phage on hosts expressing the cognate repressor compared to wildtype host. See also Figure S1.

systems are routinely engineered with inducible gene expression by placing ligand-responsive promoters upstream of gene of interest.[17] Bacterial genomes are generally tolerant to promoter substitutions as most substitutions have marginal effects on bacterial fitness.[18] In contrast, phage genomes have evolved to maintain compact genomes containing mostly essential genes with overlapping regulation.[19−21] This makes it challenging to identify a suitable T7 gene whose promoter can be replaced by ligand-responsive bacterial promoter.

The T7 phage genome is approximately 40 kb long and is roughly partitioned into early, middle and late-stage genes denoting their roles at different stages of T7 lifecycle.[22] Transcription of phage genes occurs in two stages after infecting its host. First, the host *E. coli* RNA polymerase transcribes T7 RNA polymerase and other early phage genes.[23−26] After early genes have been expressed, the host RNA polymerase is inhibited and T7 RNA polymerase assumes responsibility for transcribing middle and late-stage genes.[27,28] T7 RNA polymerase is thus a "lynchpin" gene and we reasoned that a ligand-responsive promoter regulating this gene (*gp1*) would be an optimal site to exert systemic control. The region upstream of *gp1* gene is comprised of several parts (Figure 1A). The beginning of the genome contains elements used for replication, including terminal repeats, promoter A0 recognized by the host RNA polymerase, and origin of replication øOL.[29] Downstream of these components lies early promoters A1, A2, and A3, all of which are recognized by the host RNA polymerase, followed by early phage genes *gp0.3*, *gp0.4*, *gp0.5*, *gp0.6*, and *gp0.7*. These early promoters are used for transcription-dependent internalization of the phage genome by the *E. coli* host RNA polymerase and these genes are involved in host suppression and inhibition of host defenses.[30−36]

We assembled engineered T7 phage genomes in yeast without the genomic segments covering early promoters A0, A1, A2, A3, origin of replication øOL, and gene products *gp0.3*, *gp0.4*, *gp0.5*, *gp0.6*, and *gp0.7*, thereby retaining the essential replication elements but simplifying the early genome to create an effective regulatory switch. This approximately 3 kb long genomic tract was replaced with 95−100 bp long ligand-responsive promoter NalC, PmeR, or TetR (T7_NalC, T7_PmeR, and T7_TetR, respectively) placed immediately

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

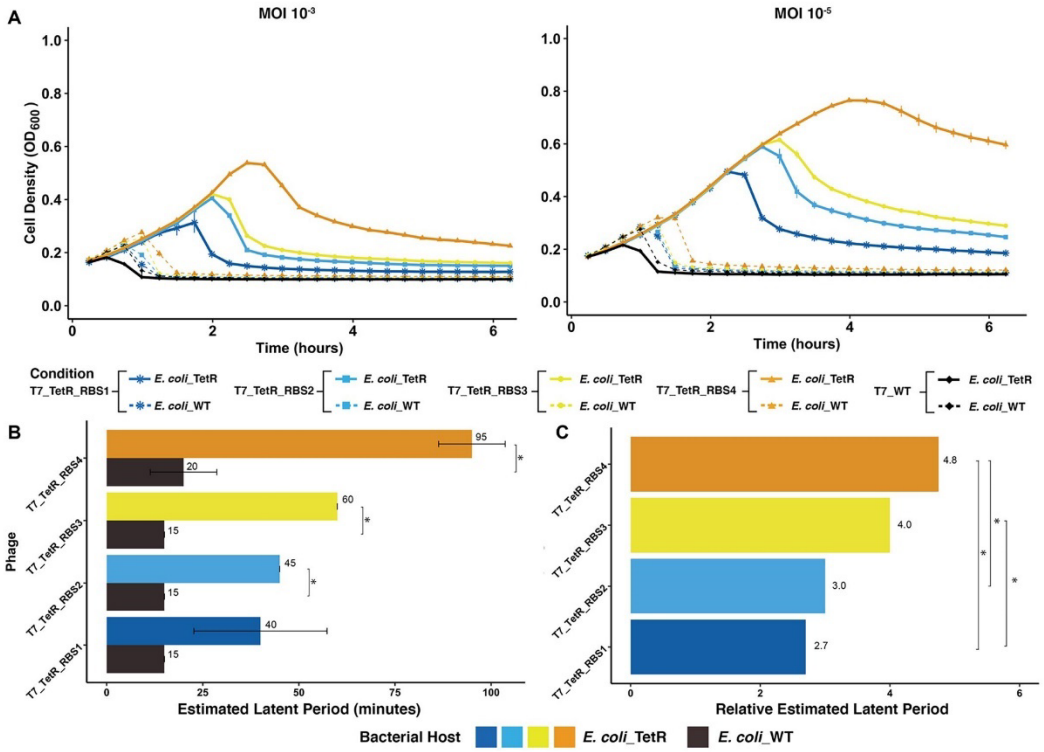


Figure 3. Engineered ribosome binding sites further enhance control over phage activity. (A) Bacterial cell density over time measured by absorbance (OD_{600}) for wildtype (*E. coli*_WT; dotted lines) and host expressing TetR repressor (*E. coli*_TetR; solid lines) after application of engineered T7 with an RBS at 100% strength (T7_TetR_RBS1, dark blue), 39% strength (T7_TetR_RBS2, light blue), 3% strength (T7_TetR_RBS3, yellow), and 1% strength (T7_TetR_RBS4, orange). Phages were added at time 0 at a MOI of 10^{-3} and 10^{-5}. The delay in estimated latent period increases as the RBS strength decreases. All data represented as mean OD_{600} ± SD in biological triplicates. (B) The estimated latent period of engineered phages on *E.* coli_WT compared to hosts expressing the TetR repressor. (C) The relative estimated latent period of engineered phage on host expressing the TetR receptor compared to wildtype host. Single asterisk (*) represents significant difference (*p*-value <0.05). Nonsignificant difference not shown. See also Figure S2.

upstream of *gp1* (T7 RNA polymerase) (Figure 1A). Because the promoters inserted are *E. coli* promoters, we reasoned that the phage could leverage these promoters for necessary transcription-dependent internalization. Engineered phages were subsequently "rebooted" by transforming their genome into an *E. coli* host.[37] Once an engineered phage infects a host expressing the repressor, the repressor will bind to its cognate promoter and block the expression of *gp1*, attenuating phage activity (Figure 1B). In the presence of the inducer or in a host where cognate repressor is absent, the repressor is unable to bind to the promoter, initiating transcription of *gp1* and reactivating the phage (Figure 1B).

Since phage amplification occurs rapidly within the host, we wanted to evaluate different repressor-promoter systems for their OFF-states. We chose to test repressor-promoter systems NalC, PmeR, and TetR as these systems have been previously engineered with tight and inducible gene regulation.[38−40] However, before comparing the ability of different repressor-promoters to regulate the phage, we first sought to determine if our engineered phages (T7_NalC, T7_PmeR, and T7_TetR) were viable after removing approximately 3 kb (~7%) of native genes and regulatory regions. The engineered phages were viable despite the substantial genomic disruption. Efficiency of plating (EOP) assays showed no significant difference (*p*-value >0.05) in ability to plaque between our engineered phages and T7_WT (Figure 1C). Furthermore, the plaque morphology of the engineered phages was indistinguishable from T7_WT in *E. coli* with and without *gp1* complementation (Figure 1D). Taken together, our engineered phages were able to maintain viability and infectivity on susceptible bacterial host compared to T7_WT.

Repressing *gp1* Diminishes Activity of Synthetic Phages. We applied T7_NalC, T7_PmeR, and T7_TetR on *E. coli* host expressing the cognate repressor from a plasmid (Figure 2A, *E. coli*_NalC, *E. coli*_PmeR, and *E. coli*_TetR, respectively) or control wildtype *E. coli* host without repressor (Figure 2A, *E. coli*_WT). For each engineered phage, we evaluated the estimated latent period (the time required to complete one infection cycle) by comparing cell densities at two different multiplicities of infection (MOI or the phage-to-bacteria ratio). The latent period can be estimated from the inflection in the growth curves between two MOIs where the

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

difference is equal to the phage burst size (see Materials and Methods) (Figure 2A). All three synthetic phages showed a delayed latent period in host expressing the cognate repressor compared to *E. coli*_WT lacking the repressor (Figure 2A and 2B). The TetR repressor had the strongest attenuation of phage activity among the three repressors based on estimated lysis time (Figure 2A). In *E. coli*_WT, T7_PmeR and T7_TetR required approximately 20 min to lyse the host, which was comparable to that of T7_WT (Figure S1). This is also consistent with similar EOPs for wildtype and engineered phages (Figure 1B). T7_NalC required a longer time of 30 min to lyse the host, indicating that the NalC promoter may be affecting phage fitness even in the absence of the repressor. When engineered phages were applied on hosts expressing their corresponding repressors, the estimated lysis time of hosts *E. coli*_PmeR, *E. coli*_NalC, and *E. coli*_TetR was approximately 40, 30, and 40 min, respectively. As expected, the repressor was able to delay the activity of engineered phages (Figure 2B). We found that the TetR system gave the longest delay (2-fold) in estimated lysis time after normalizing for differences in basal activity of each engineered phage on *E. coli*_WT (Figure 2C). We also examined residual bacterial density (OD_{600}) as it is indicative of the fraction of host cells that remain unlysed. After 6 h of incubation with the engineered phages, bacterial density was highest for *E. coli*_TetR, followed by *E. coli*_NalC and then *E. coli*_PmeR (Figure 2A). Therefore, synthetic switches engineered into the phages able to reduce phage efficacy and delay host cell lysis. Since the TetR system gave the longest delay, we chose *E. coli*_TetR and T7_TetR for further optimization.

Engineering Ribosome Binding Sites Further Enhances Phage Control. Although T7_TetR performed better than T7_PmeR and T7_NalC, T7_TetR still showed relatively high activity in the OFF-state (Figure 2C). We hypothesized this could be due to high basal expression of *gp1* from the TetR-regulated promoter. To address this, we investigated if modifications to the ribosome binding site (RBS) could further attenuate phage activity in the OFF-state by reducing basal expression of *gp1*. We engineered T7_TetR with three RBS variants of the TetR-regulated promoter. The chosen RBS variants spanned a range of translational activities based on a large-scale experimental study of *E. coli* RBS variants.[38,41] Relative to the strength of the original RBS at 100% strength (T7_TetR_RBS1), we engineered T7_TetR with RBS variants at 39% strength (T7_TetR_RBS2), 3% strength (T7_TetR_RBS3), and 1% strength (T7_TetR_RBS4) (see Materials and Methods). On *E. coli*_WT without repressor, all T7_TetR RBS variants required an estimated 15−20 min to lyse the host, comparable to T7_WT, indicating that strength of RBS had no measurable impact on phage activity in the absence of repression (Figure 3A and 3B). T7_TetR_RBS4 also retains viability when grown in different temperatures and in starvation conditions without repression (Figure S4). In contrast, when the engineered phages were applied on *E. coli*_TetR, there was substantial delay in time required to lyse the host compared to *E. coli*_WT (Figure 3B). The time required to lyse *E. coli*_TetR was inversely related to the strength of the RBS. T7_TetR_RBS4 (the weakest RBS) required 95 min to lyse *E. coli*_TetR, the longest time and a 4.75-fold greater delay relative to *E. coli*_WT (Figure 3C). T7_TetR_RBS4 also retained a high residual bacterial density after 6 h, suggesting lowering basal expression of *gp1* successfully attenuated T7 activity (Figure 3A). Weaker RBSs are thus able to provide enhance control of phage replication by further restricting phage activity in the OFF-state.

Repression Significantly Delays Replication in Engineered Phages. To further characterize the effectiveness of our synthetic switch, we examined the number of progeny phages over one infection cycle using one-step growth curves for T7_WT and T7_TetR_RBS4 on *E. coli*_WT and *E. coli*_TetR (Figure 4A). By counting the number of phages

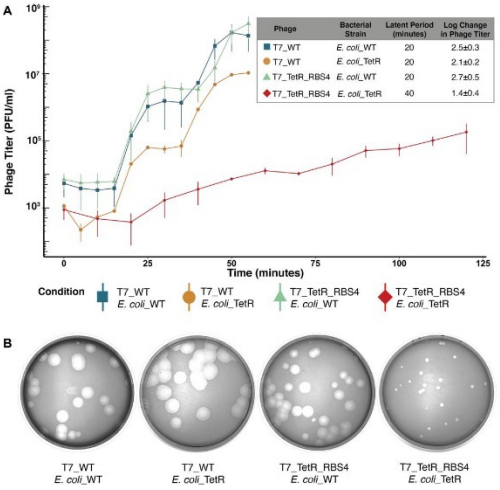


Figure 4. Repression significantly delays replication in engineered phages. (A) One step growth assay showing phage titer (PFU/mL) of wildtype (T7_WT) and engineered (T7_TetR_RBS4) phage on wildtype (*E. coli*_WT) and engineered (*E. coli*_TetR) host. Inset table shows the latent period (minutes) and estimated log change in phage titer produced from the first replication cycle. All data shown as mean ± SD in biological triplicates. (B) Plaque assay of wildtype (T7_WT) and T7 with TetR and a 1% strength ribosome binding site (T7_TetR_RBS4) phages on wildtype host (*E. coli*_WT) and host expressing TetR repressor (*E. coli*_TetR) host after 19 h of incubation.

produced at different times, we determined the actual latent period (the time required to complete one infection cycle), and compared the log change in phage titer over one infection cycle. In prior literature, the latent period for T7_WT is approximately 20 min with an expected 2−2.5 log increase in the total phage population after one infection cycle.[42,43] Since the TetR promoter and RBS does not have an impact on T7 fitness, we expected similar values for T7_TetR_RBS4 on *E. coli*_WT.

Our one step assay confirmed that both T7_WT and T7_TetR_RBS4 have a 20 min latent period on *E. coli*_WT and T7_WT has a 20 min latent period on *E. coli*_TetR (Figure 4A). After one infection cycle, the average increase in total phages for T7_WT on *E. coli*_WT and *E. coli*_TetR was approximately 2.5 ± 0.3 and 2.1 ± 0.2 log, respectively, while T7_TetR_RBS4 saw a comparable increase in total phages of 2.7 ± 0.5 log on *E. coli*_WT. These results confirm that the TetR repression system had no apparent effect on phage activity for T7_WT or for T7_TetR_RBS4 in the absence of repression. In contrast, T7_TetR_RBS4 infecting *E. coli*_TetR

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

had greatly extended latent period of 40 min with only a 1.4 ± 0.4 log increase in total phages (Figure 4A), indicated a dramatic delay of phage activity with very gradual production of phage progeny during each replication cycle. We next determined how the TetR_RBS4 repression system affects the ability of the phage to plaque on *E. coli*_TetR compared to *E. coli*_WT using an EOP assay. T7_TetR_RBS4 phages had an EOP of −0.2 ± 0.1 while T7_WT phage had an EOP of 0.2 ± 0.1. The plaque activity of T7_TetR_RBS4 and T7_WT was slightly significantly different (p-value = 0.05) and the T7_TetR_RBS4 plaques were significantly smaller than T7_WT plaques after 19 hours of incubation, indicative of much slower phage activity and consistent with our one-step results (Figure 4B). In summary, engineered T7_TetR_RBS4 phages have a significant delay in replication compared to wildtype when the synthetic repressor is present, and phage activity can be fully recovered if the synthetic repressor is not present or under induction.

Phage Infectivity Can Be Further Controlled by Tuning Inducer Concentration. To assess if phage activity could be dynamically controlled using a small molecule inducer, we compared the activities of T7_TetR_RBS4 with and without the inducer anhydrotetracycline (aTC).[44] Comparison of spot plates showed that repressed T7_TetR_RBS4 (Figure 5A, left) regained activity comparable to T7_WT when maximally induced at 1 μM aTC (Figure 4A, right, Figure S2). A control experiment confirmed that aTC had no impact on activities of T7_TetR_RBS4 on *E. coli*_WT host (Figure S3). However, host cells grown over a gradient of aTC concentrations experienced a minor fitness deficit at higher aTC concentrations suggesting inducer toxicity may have partially contributed to cell death.[45] To evaluate if the activity of T7_TetR_RBS4 could be tuned by the inducer in a dose-dependent manner without cell toxicity, we performed a time course experiment over a range of aTC concentrations from 0 to 1 μM (Figure 5B). T7_TetR_RBS4 showed switch-like behavior from no activity to high activity with modest dose-dependent activity over a narrow range of aTC concentration from 15 to 23 nM. Above 23 nM, T7_TetR_RBS4 was fully switched "ON" and activity was comparable to activity at maximal induction at 1 μM (Figure 5B). Below 15 nM, the activity of T7_TetR_RBS4 was nearly the same as the activity with no inducer. The tunable concentration range (15−23 nM) showed large variations which arise due to high stochasticity in the expression of *gp1* across the host population at these concentrations. Since T7 RNA polymerase can be recycled for phage gene expression,[46] even stochastic bursts of expression would eventually lead to bacterial lysis. Altogether these results show that a moderate aTC concentration can provide a tunable level of control over phage activity.

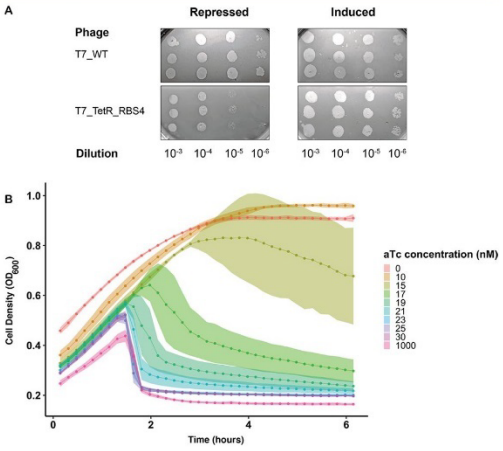


Figure 5. Phage infectivity can be tuned by changing inducer concentration. (A) Spot plate assay measuring the ability of wildtype T7 (T7_WT) and T7 with TetR and a 1% strength ribosome binding site (T7_TetR_RBS4) to infect *E. coli* expressing the TetR repressor without inducer (Repressed) or with 1 μM anhydrotetracycline (aTC) inducer (Induced). Phages are spotted from 10^{-3} to 10^{-6} dilution and spots are imaged after 4 h of incubation. Induction restores the phage ability to spot comparable to wildtype. (B) Bacterial cell density over time measured by absorbance (OD_{600}) for *E. coli* expressing TetR repressor. T7_TetR_RBS4 phage is added at time point 0 at a MOI of 10^{-5}, and each culture contains aTC inducer ranging from a concentration of 0 to 1000 nM (color gradient). See also Figure S3.

DISCUSSION

In this study, we engineered synthetic gene regulation into T7 phage to create a phage-bacterial system with ligand-regulated infectivity. We deleted a large tract of the genome from T7 phage and inserted a suite of ligand-regulated promoters to control the expression of *gp1*, the T7 RNA polymerase. The deletion of phage early promoters and genes and replacement with a ligand-regulated promoter resulted in a viable phage with infectivity comparable to T7_WT. Transcription-independent translocation drives initial DNA internalization that would include our engineered tetR promoter driving *gp1*.[31] We hypothesize transcription-dependent translocation then proceeds using this promoter. Our most optimized engineered phage, T7_TetR_RBS4, had a 2-fold delay in estimated latent period and an approximately 10-fold decrease in total phages after one replication cycle compared to T7_WT. The activity of the engineered phage could be fully recovered in dose-dependent manner by adding the inducer ligand.

Our engineered phage sets the stage for several improvements to further refine and control phage infectivity. Because the infectivity of our engineered phage was determined solely by control over T7 RNAP, even basal levels of expression would allow the phage to escape repression proceed with infection. Adding multiple levels of repression throughout the phage genome could substantially improve our system and decrease the activity the phage in the OFF-state.[47,48] Critical genes such as late structural genes would make natural targets for improving control over the phage replication. Our study and other previous reports suggest that the phage genome may be tolerant to such changes without affecting viability.[42,49]

An engineered phage with a dynamic controllable infectivity switch can be used as a potent tool for bacterial community control.[50,51] Utilizing different repression systems across different bacterial hosts, phages can be redirected to different bacteria in the same community.[52] Phages could also be continuously propagated in a community using a "feeder" host whose susceptibility could be turned ON or OFF as needed. By tuning inducer concentrations,[53,54] phages could be engineered to repress but not eliminate specific bacterial

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

hosts in a complex community. Alternatively, our engineered phages can act as pseudolysogenic phages, maintaining a low burden on the targeted microbe until being triggered to eliminate the host. This could be useful for applications like starter cultures, where timed bacterial lysis is a critical consideration for cheese maturation.[55] Our study presents a simple genetic regulatory technique that can be further engineered to create a more controllable phage-bacterial system to precisely manipulate microbial communities.[56−58]

MATERIALS AND METHODS

Microbes and Culture Conditions. *Escherichia coli* (*E. coli*) 10G, a highly competent derivative of DH10β was obtained from Lucigen (60107−1),[59] T7 bacteriophage was obtained from ATCC (ATCC BAA-1025-B2), and *Saccharomyces cerevisiae* BY4741[60] is a laboratory stock. *Escherichia coli* (*E. coli*) MG1655 is a laboratory stock.

Bacterial cultures were grown in LB (Luria−Bertani) media (1% Tryptone, 0.5% Yeast Extract, 1% NaCl). For plating, LB agar contains 1.5% agar, while top agar contains 0.5% agar (Teknova). Spectinomycin (115 μg/mL final concentration, GoldBio) was added to media for selections of pSC101_NalC, pSC101_PmeR and pSC101_TetR. All incubations of bacterial and bacteriophage cultures were performed at 37 °C, with the liquid culture shaking at 200−250 rpm consistently, unless otherwise specified. For the host starvation experiment, bacterial cultures were grown in minimal media M9 supplemented with glucose and thiamine (33.7 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.55 mM NaCl, 9.35 mM NH_4Cl, 1 mM $MgSO_4$, 0.3 mM $CaCl_2$, 0.4% glucose, 1 μg thiamine).

T7 bacteriophage was grown and propagated using *E. coli* 10G in LB media. Phage stocks were tittered using plaque assay and stored in LB at 4 °C.

S. cerevisiae BY4741 was grown in YPD (2% Peptone, 1% Yeast Extract, 2% Glucose) media prior to transformation. Yeast transformants were selectively grown on SD-Leu (0.17% Yeast Nitrogen Base, 0.5% Ammonium Sulfate, 0.162% Amino Acid-Leucine, [Sigma Y1376], 2% Glucose). YPD and SD-Leu plates contain 2.4% and 2% agar additionally, respectively. Yeast incubation was performed at 30 °C, with liquid culture shaking at 200−250 rpm.

Short-term storage of liquid culture and plates were performed at 4 °C and long-term storages of bacterial and yeast stock culture were performed at −80 °C in screw-capped cryotubes, with 25% glycerol added as a cryoprotectant.

Bacterial and phage transformants were recovered in SOC (2% Tryptone, 0.5% Yeast Extract, 0.2% 5 M NaCl, 0.25% 1 M KCl, 1% 1 M $MgCl_2$, 1% 1 M $MgSO_4$, 2% Glucose) liquid media.

General Cloning Procedure. PCR and cloning were adapted and performed using standard laboratory procedures.[12] Briefly, PCR amplification was performed using KAPA HiFi (Roche KK2101) for all amplifications with plasmid or phage templates. KAPA2G Robust PCR kits (Roche KK5005) were used to perform colony PCR and multiplex PCR for a screening of Yeast Artificial Chromosomes (YACs). All primer oligos were obtained from IDT. Golden Gate assembly was performed using New England Biosciences (NEB) Golden Gate Assembly Kit (BsaI-HFv2, E1601L). DNA purification was performed using EZNA Cycle Pure Kits (Omega Biotek D6492−01) using centrifugation protocol. YAC extraction was performed using YeaStar Genomic DNA Extraction Kits (Zymo Research D2002). Gibson Assembly mixture was made in the laboratory (final concentration 100 mM Tris-HCl pH 7.5, 20 mM $MgCl_2$, 0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 0.2 mM dTTP, 10 mM DTT, 5% PEG-8000, 1 mM NAD^+, 4 U/mL T5 exonuclease, 4 U/μL Taq DNA Ligase, 25 U/mL Phusion polymerase). PCR product visualization was performed using agarose gel electrophoresis with appropriate agarose concentration and SYBR Safe DNA Gel Stain (Invitrogen).

PCR amplification using plasmid templates was performed with 0.1 ng DNA template. Phage fragment amplification was performed using 1 μL phage crude lysis treated at 65 °C for 10 min as a template. Deletions and insertions of the T7 genome were performed using PCR primers skipping or adding desired sequences, respectively. All plasmid-template PCR products were treated with DpnI (NEB) following standard protocol. Briefly, purified PCR product was combined with 5 μL 10× CutSmart Buffer, 1 μL DpnI, and dH_2O to 50 μL reaction. Digestion was performed at 37 °C for at least 2 h, followed by heat inactivation at 80 °C, 20 min. PCR purification was performed afterward. All PCR products were quantified using NanoDrop 2000 Spectrophotometer (Thermo Scientific).

E. coli 10G competent cells were made by mixing 192 mL SOC with 8-mL overnight culture and incubating at 21 °C and shaking at 200 rpm until $OD_{600} \sim 0.4$ was reached, determined by the Agilent Cary 60 UV−vis Spectrometer using manufacturer documentation. Cells were centrifuged at 4 °C, 800*g* to 1000*g* for 20 min, the supernatant was discarded, and cells were resuspended in 50 mL of precooled 10% glycerol. Centrifugation and washing were repeated three times. Cells were resuspended in a final volume of ∼1 mL 10% glycerol and were aliquoted and stored at −80 °C. Cells made by this protocol are competent for both plasmid and YAC transformation using electroporation.

Electroporation was performed using 40 μL 10G competent cell for both plasmids and YACs using a Biorad MicroPulser (165−2100), EC2 setting with 2 mm cuvette, 2.5 kV, single pulse. All cuvettes and Eppendorf tubes were chilled prior to the electroporation. After electroporation, recovery was performed by adding 950 μL prewarmed SOC and incubated at 37 °C for 1 h and plated on relevant selective media.

Detailed protocols for cloning are available on request. All primers and plasmids used in this study are listed in supporting document (Table S1).

Phage Engineering. A large segment (2961 bp) in the left end region of the wild type (WT) T7 phage genome, including promoters A0, A1, A2, A3 and *gp0.3−0.7* genes, was deleted. Two identical operator (repressor binding) sites were inserted, one upstream of −35 and the other in between the −35 and −10 consensus sequences replacing the spacer sequence of promoter A1 driving the expression of *gp1* gene. The three operator sites are NalC1, PmeR2, and TetO correspond to repressors, NalC, PmeR and TetR, respectively, in bacterial system. The Bujard RBS (Ec-TTL-R111) in TetO engineered phage was replaced by Bba_J61106 (Ec-TTL-R065), Bba_J61133 (Ec-TTL-R003), or DeadRBS (Ec-TTL-R001) with 39%, 3%, and 1% translational strength relative to the Bujard RBS as 100% strength,[41] respectively. The sequences of all consensus sequences, operator sites, RBS, and repressors used are included in supporting document (Table S2).

Engineered phage genomes were assembled using yeast assembly,[4,5] which requires yeast transformation of relevant DNA segments. A prs315 yeast centromere plasmid was split

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

into three segments by PCR, separating the centromere and leucine selection marker, which has been shown to improve assembly efficiency by limiting recircularization events.[6] T7 genomic segments were made by PCR using WT T7 as a template.

Relevant DNA fragments were mixed together (0.1 pmol/fragment) and transformed into *S. cerevisiae* BY4741 using the high efficiency yeast transformation protocol.[7] Successfully assembled YACs were selected using SD-Leu media. After 2−3 days of incubation at 30 °C, colonies were picked and directly assayed by multiplex colony PCR to screen for colonies with correctly assembled YACs. Multiplex PCR was an effective way of distinguishing correctly assembled YACs by interrogating junctions in the YACs. Correctly assembled YACs were purified and transformed into *E. coli* 10G cells, and after 1-h recovery, 400 μL was inoculated in 4.6 mL LB. This culture was incubated until a complete lysis, engineered phages were purified, sequence-confirmed, and stored at 4 °C in LB.

Plasmid Descriptions. pSC101_NalC, pSC101_PmeR, and pSC101_TetR contain a pSC101 backbone, spectinomycin resistance cassette, and repressors NalC, PmeR, and TetR under constitutive promoter apFAB61, respectively.

pHT7_gp1 contains a pBR backbone, kanamycin resistance cassette, mCherry fluorescence marker, and the T7 RNAP gene, *gp1*. Both mCherry marker and *gp1* are under constitutive expression.

Bacterial Methods and Phage Titer Quantification. Bacterial concentration was determined using a drop plate with 10-fold serial dilution of bacterial culture. Ten-microliters of bacterial dilution were dropped on LB plate in triplicate and counted after overnight incubation as colony-forming unit per milliliter (CFU/mL). Optical density (OD) measurement for all microplate-reader experiments were performed using microplate reader (Synergy HTX) at 600 nm wavelength, except the preparation of electrocompetent cell and the determination of log phase of bacterial culture that were performed using an Agilent Cary 60 UV−vis Spectrometer.

Bacterial culture used for bacteria-phage interaction experiment and phage quantification was obtained at log phase when the initial phage was added, unless otherwise specified. Briefly, overnight bacterial culture was diluted 1:20 in liquid LB media. Bacterial culture was collected when the OD_{600} reaches ∼0.4−0.6 as determined using an Agilent Cary 60 UV−vis Spectrometer.

Bacterial culture used for investigating phage viability under various conditions, including bacterial growth phase (log and stationary phase), incubation temperature (25, 30, 37, 45 °C), and host starvation (minimal M9 with 0.4% glucose and thiamine), was obtained at log phase and stationary phase from relevant media and temperature conditions throughout the entire experiment. *E. coli* MG1655 was used for this experiment as *E. coli* 10G cannot productively grow in minimal media. Briefly, overnight bacterial culture in each condition and media was used as a stationary phase host and was diluted 1:20 in liquid media to obtain log phase host when the OD_{600} reaches ∼0.4−0.6 as determined using an Agilent Cary 60 UV−vis Spectrometer. Plaque assays were performed under relevant media and incubation temperature.

Phage stock was produced by a complete infection of phage in *E. coli* 10G. Bacterial lysate was centrifuged at 4400*g* for 5 min and supernatant was filtered through a 0.22 μm filter (Celltreat 151205−051). Phage titer was determined by plaque assay with a 10-fold serial dilution of phage lysate. Typically, 10−20 μL of phage dilution is mixed thoroughly with 250 μL log-phase bacterial culture and 3.5-mL 0.5% top agar and plated on 37 °C prewarmed LB plate. Plaque number was counted after overnight incubation as plaque-forming unit per milliliter (PFU/mL). Spot assay was considered as a preliminary quantification of phage titer, which was performed occasionally prior to whole-plate plaque assay. Log-phase bacterial culture was mixed with 3.5-mL top agar and plated on prewarmed LB plate. One-and-a-half microliters of phage dilution from 10^0 to 10^{-8} were spotted in triplicate on set top agar. The number of PFU was approximated determined at 4−6 h of incubation and the appropriate dilution was chosen to perform whole-plate plaque assay.

Multiplicity of Infection (MOI) was determined by dividing phage titer by bacterial concentration. MOI used in this study was approximately 10^{-3} and 10^{-5}, except for one-step growth curve which was performed with MOI ∼ 0.1.

Efficiency of Plating (EOP) was calculated using wildtype *E. coli* 10G as a reference host. After performing a plaque assay using desired bacterial hosts and phages, EOP was determined by dividing the experimental phage titer to the phage titer on the reference bacterial host. Typically, a 10-fold serial dilution was performed with phage stock in triplicate and 10 μL appropriate phage dilutions were applied for plaque assay as described in desired bacterial host.

Infection Time Course Curves. Preliminary change in phage activity of engineered phages was determined by evaluating bacterial growth after addition of different phages. Seven microliters of overnight bacterial culture were inoculated in 133 μL of relevant media as a 1:20 dilution with corresponding antibiotics if necessary or 1 μM aTC to the total volume of 140 μL in a 96-well plate. Ten microliters of phages were added in triplicate at an MOI of $\sim 10^{-3}$ and $\sim 10^{-5}$. The culture was continuously incubated in microplate reader (Synergy HTX) at 37 °C with OD_{600} measured every 10 min for 6 h. The bacterial growth curve was constructed by averaging OD_{600}.

Phage Latent Period Estimation with Infection Time Courses. With the assumption that T7 phage produces an average of 100 progeny phage at 37 °C when infecting 10G,[43] time to lysis was estimated by comparing the inflection point of bacterial growth (or the point at which the culture begins to lyse) for two cultures after addition of two phage titers spanning 2 orders of magnitude. An MOI of 10^{-3} and 10^{-5} were chosen for this estimation. The estimated latent period was calculated as the difference between the last time point immediately before the inflection point in OD_{600} between the two cultures.

Phage Growth Quantification: Latent Period and Phage Titer Change Estimation. One-step phage growth assays were performed to construct one-step growth curves using an adaptation of a standard protocol.[61,62] Briefly, Seven-hundred-50 microliters of an overnight bacterial were added into 15-mL LB media with antibiotics if necessary. The culture was incubated shaking at 250 rpm, 37 °C in a 50-mL flask. Bacterial culture was collected when the OD_{600} ∼ 0.25 was reached in 15-mL tube and concentrated into 1.5 mL by centrifugation at 4400*g* for 5 min. Bacterial culture was transferred into a 2-mL Eppendorf tube, phage was added into the culture at MOI ∼ 0.1. Phage-bacteria culture was incubated for absorption without shaking at 37 °C for 5 min. Culture was washed four times by centrifugation at 10 000*g* for 30 s, discarding the supernatant, and resuspending with 1-mL LB

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

media, with antibiotic if necessary. In between washes, culture was resuspended by vortexing. After the final wash, 1-mL media was added and transferred into 14-mL prewarmed media in a 50-mL flask. Eight-hundred microliters of culture was collected immediately and centrifuged at 10 000*g* for 30 s. Seven-hundred-microliter supernatant was filtered through 0.22 μm (Celltreat 151205−051) into a new Eppendorf tube. Fifteen-milliliter culture was incubated shaking at 250 rpm at 37 °C. The culture sample was collected every 5 or 10 min, depending on the type of phage culture for 60 or 120 min. Phage titer was quantified by plaque assay using wildtype *E. coli* 10G. Plated culture was incubated at 37 °C for 12 h and the numbers of plaques were counted and calculated. The dilution of the phage lysate was made if necessary. The growth curve of each bacteria-phage combination was constructed in biological triplicates.

The latent period of each phage-bacteria combination was quantified by the first phage replication cycle[62] using one-step phage growth curve. The mid-time-point at log-phase of replication cycle in between plateau phage titer was considered as a latent period.

The change in total phage titer was estimated using the phage growth curve by calculating the log difference of number of phage progenies at the plateau after the first replication cycle and at the initial plateau where phage titer added into the experimental culture.

Inducer Dependency Assay. The activity of phage in response to a range of aTC inducer concentrations was determined by quantifying bacterial growth using OD_{600}. Briefly, 7 μL overnight bacterial culture was inoculated in 133 μL of relevant medium as 1:20 dilution to the total volume of 140 μL in a 96-well plate. Media used was prepared with a range of aTC concentration from 0 to 1 μM. Culture was incubated in a microplate reader (Synergy HTX), and OD was measured at 600 nm wavelength. When bacterial $OD_{600} \sim 0.25$ was reached, 10 μL of desired phage or LB as control was added into the culture for MOI $\sim 10^{-5}$. Culture was continuously incubated for 7 h using vertical shaking mode, and OD_{600} was measured every 10 min. The experiment was performed in biological triplicates.

Statistical Analysis. All data analyses were performed in Microsoft Excel 2020 and R v4.0.4.[63] Bonferroni pairwise *t*-test[64,65] was used to detect the difference among conditions in all experiments. Statistical analyses were considered significant at $p < 0.05$.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.1c00414.

Control experiments; primer sequences; plasmid sequences; promoter and ribosome binding site sequences (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Srivatsan Raman − *Department of Biochemistry, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States; Department of Bacteriology and Department of Chemical & Biological Engineering, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0003-2461-1589; Email: sraman4@wisc.edu

Authors

Chutikarn Chitboonthavisuk − *Department of Biochemistry, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States; Department of Bacteriology and Microbiology Doctoral Training Program, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States*

Chun Huai Luo − *Department of Biochemistry, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States; Department of Bacteriology, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States*

Phil Huss − *Department of Biochemistry, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States; Department of Bacteriology and Microbiology Doctoral Training Program, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States*

Mikayla Fernholz − *Department of Biochemistry, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.1c00414

Author Contributions

#C.C. and C.H.L. contributed equally to this work.

Author Contributions

S.R., C.C., C.H.L., P.H., and M.F. conceptualized and designed the study. C.H.L. and M.F. constructed plasmids. C.C. and C.H.L. performed the experiment, curated the data, performed the analysis, wrote the original manuscript, and designed the figures. S.R. and P.H. supervised the overall experiment. S.R., C.C., C.L., P.H., and M.F. reviewed and edited the manuscript. All authors discussed the results and provided critical feedback on the manuscript.

Notes

The authors declare the following competing financial interest(s): S.R. is on the scientific advisory board of MAP/PATH LLC. All other authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by NIAID grant R21AI156785 (to S.R.) and by the Anandamahidol Foundation Scholarship, Thailand (to C.C).

■ REFERENCES

(1) Azam, A. H.; Tanji, Y. Bacteriophage-host arm race: an update on the mechanism of phage resistance in bacteria and revenge of the phage with the perspective for phage therapy. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 2121−2131.

(2) Yang, H. H.; Ma, Y.; Wang, Y.; Yang, H. H.; Shen, W.; Chen, X. Transcription regulation mechanisms of bacteriophages: Recent advances and future prospects. *Bioeng. Bugs.* **2014**, *5*, 300−304.

(3) Prell, H. H. Regulation of gene expression in Salmonella phage P22 - II. Regulation of expression of late functions. *Mol. Gen. Genet.* **1975**, *136*, 351−360.

(4) Bushman, F. D.; Ptashne, M. Activation of transcription by the bacteriophage 434 repressor (in vitro transcription/RNA polymerase/

ethylation interference/protein-protein interaction). *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83*, 9353−9357.
(5) Kuhstoss, S.; Rao, R. N. Analysis of the integration function of the streptomycete bacteriophage φC31. *J. Mol. Biol.* **1991**, *222*, 897−908.
(6) Court, D. L.; Oppenheim, A. B.; Adhya, S. L. A new look at bacteriophage λ genetic networks. *J. Bacteriol.* **2007**, *189*, 298−304.
(7) Rojo, F. Repression of Transcription Initiation in Bacteria. *J. Bacteriol.* **1999**, *181*, 2987.
(8) Guzmán-Trampe, S.; Ceapa, C. D.; Manzo-Ruiz, M.; Sánchez, S. Synthetic biology era: Improving antibiotic's world. *Biochem. Pharmacol.* **2017**, *134*, 99−113.
(9) Lemire, S.; Yehl, K. M.; Lu, T. K. Phage-Based Applications in Synthetic Biology. *Annu. Rev. Virol.* **2018**, *5*, 453−476.
(10) Brown, R.; Lengeling, A.; Wang, B. Phage engineering: how advances in molecular biology and synthetic biology are being utilized to enhance the therapeutic potential of bacteriophages. *Quant. Biol.* **2017**, *5*, 42−54.
(11) Huss, P.; Raman, S. Engineered bacteriophages as programmable biocontrol agents. *Curr. Opin. Biotechnol.* **2020**, *61*, 116−121.
(12) Huss, P.; Meger, A.; Leander, M.; Nishikawa, K.; Raman, S. Mapping the functional landscape of the receptor binding domain of t7 bacteriophage by deep mutational scanning. *eLife* **2021**, DOI: 10.7554/eLife.63775.
(13) Kilcher, S.; Loessner, M. J. Engineering Bacteriophages as Versatile Biologics. *Trends Microbiol.* **2019**, *27*, 355−367.
(14) Pires, D. P.; Cleto, S.; Sillankorva, S.; Azeredo, J.; Lu, T. K. Genetically Engineered Phages: a Review of Advances over the Last Decade. *Microbiol. Mol. Biol. Rev.* **2016**, *80*, 523−543.
(15) Kulczyk, A. W.; Richardson, C. C. The Replication System of Bacteriophage T7. *Enzymes* **2016**, *39*, 89−136.
(16) Casjens, S. R.; Hendrix, R. W. Bacteriophage lambda: Early pioneer and still relevant. *Virology* **2015**, *479*−*480*, 310−330.
(17) Chen, Y.; Ho, J. M. L.; Shis, D. L.; Gupta, C.; Long, J.; Wagner, D. S.; Ott, W.; Josić, K.; Bennett, M. R. Tuning the dynamic range of bacterial promoters regulated by ligand-inducible transcription factors. *Nat. Commun.* **2018**, *9*, 1−8.
(18) Gruber, T. M.; Gross, C. A. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **2003**, *57*, 441−66.
(19) Pachl, C. A.; Young, E. T. Detection of polycistronic and overlapping bacteriophage T7 late transcripts by in vitro translation. *Proc. Natl. Acad. Sci. U. S. A.* **1976**, *73*, 312.
(20) Teif, V. B. Predicting Gene-Regulation Functions: Lessons from Temperate Bacteriophages. *Biophys. J.* **2010**, *98*, 1247.
(21) Jack, B. R.; Boutz, D. R.; Paff, M. L.; Smith, B. L.; Wilke, C. O. Transcript degradation and codon usage regulate gene expression in a lytic phage. *Virus Evol.* **2019**, DOI: 10.1093/ve/vez055.
(22) Dunn, J. J.; Studier, F. W.; Gottesman, M. Complete Nucleotide Sequence of Bacteriophage T7 DNA and the Locations of T7 Genetic Elements. *J. Mol. Biol.* **1983**, *166*, 477−535.
(23) Kemp, P.; Gupta, M.; Molineux, I. J. Bacteriophage T7 DNA ejection into cells is initiated by an enzyme-like mechanism. *Mol. Microbiol.* **2004**, *53*, 1251−1265.
(24) Kemp, P.; Garcia, L. R.; Molineux, I. J. Changes in bacteriophage T7 virion structure at the initiation of infection. *Virology* **2005**, *340*, 307−317.
(25) García, L. R.; Molineux, I. J. Rate of translocation of bacteriophage T7 DNA across the membranes of Escherichia coli. *J. Bacteriol.* **1995**, *177*, 4066.
(26) Savalia, D.; Robins, W.; Nechaev, S.; Molineux, I.; Severinov, K. The Role of the T7 Gp2 Inhibitor of Host RNA Polymerase in Phage Development. *J. Mol. Biol.* **2010**, *402*, 118−126.
(27) McAllister, W. T.; Wu, H. L. Regulation of transcription of the late genes of bacteriophage T7. *Proc. Natl. Acad. Sci. U. S. A.* **1978**, *75*, 804.
(28) Studier, F. W.; Dunn, J. J. Organization and expression of bacteriophage T7 DNA. *Cold Spring Harb. Symp. Quant. Biol.* **1983**, *47* (2), 999−1007.
(29) Kelly, T. J.; Thomas, C. A. An Intermediate in the Replication of Bacteriophage T7 DNA Molecules. *J. Mol. Biol.* **1969**, *44*, 459−475.
(30) McAllister, W. T.; Barrett, C. L. Roles of the Early Genes of Bacteriophage T7 in Shutoff of Host Macromolecular Synthesis. *J. Virol.* **1977**, *23*, 543.
(31) Struthers-Schlinke, J. S.; Robins, W. P.; Kemp, P.; Molineux, I. J. The internal head protein Gp16 controls DNA ejection from the bacteriophage T7 virion. *J. Mol. Biol.* **2000**, *301*, 35−45.
(32) Hirsch-Kauffmann, M.; Herrlich, P.; Ponta, H.; Schweiger, M. Helper function of T7 protein kinase in virus propagation. *Nature* **1975**, *255*, 508−510.
(33) King, T.; Kocharunchitt, C.; Gobius, K.; Bowman, J. P.; Ross, T. Global Genome Response of Escherichia coli O157:H7 Sakai during Dynamic Changes in Growth Kinetics Induced by an Abrupt Temperature Downshift. *PLoS One* **2014**, *9*, e99627.
(34) Hasan, C. M. M.; Shimizu, K. Effect of temperature up-shift on fermentation and metabolic characteristics in view of gene expressions in Escherichia coli. *Microb. Cell Fact.* **2008**, *7*, 35.
(35) White-Ziegler, C. A.; Um, S.; Pérez, N. M.; Berns, A. L.; Malhowski, A. J.; Young, S. Low temperature (23 degrees C) increases expression of biofilm-, cold-shock- and RpoS-dependent genes in Escherichia coli K-12. *Microbiology* **2008**, *154*, 148−166.
(36) Jones, T.; Gill, C. O.; McMullen, L. M. The behaviour of log phase Escherichia coli at temperatures that fluctuate about the minimum for growth. *Lett. Appl. Microbiol.* **2004**, *39*, 296−300.
(37) Ando, H.; Lemire, S.; Pires, D. P.; Lu, T. K. Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. *Cell Syst.* **2015**, *1*, 187.
(38) Liu, X.; Gupta, S. T P; Bhimsaria, D.; Reed, J. L; Rodriguez-Martinez, J. A; Ansari, A. Z; Raman, S. De novo design of programmable inducible promoters. *Nucleic Acids Res.* **2019**, *47*, 10452−10463.
(39) Ramos, J. L.; Martínez-Bueno, M.; Molina-Henares, A. J.; Terán, W.; Watanabe, K.; Zhang, X.; Gallegos, M. T.; Brennan, R.; Tobes, R. The TetR Family of Transcriptional Repressors. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 326−356.
(40) Lutz, R.; Bujard, H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I 1-I 2 regulatory elements. *Nucleic Acids Res.* **1997**, *25*, 1203−1210.
(41) Kosuri, S.; Goodman, D. B.; Cambray, G.; Mutalik, V. K.; Gao, Y.; Arkin, A. P.; Endy, D.; Church, G. M. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 14024−14029.
(42) Nguyen, H. M.; Kang, C. Lysis Delay and Burst Shrinkage of Coliphage T7 by Deletion of Terminator Tφ Reversed by Deletion of Early Genes. *J. Virol.* **2014**, *88*, 2107−2115.
(43) You, L.; Suthers, P. F.; Yin, J. Effects of Escherichia coli Physiology on Growth of Phage T7 In Vivo and In Silico. *J. Bacteriol.* **2002**, *184*, 1888−1894.
(44) Gossen, M.; Bujard, H. Anhydrotetracycline, a novel effector for tetracycline controlled gene expression systems in eukaryotic cells. *Nucleic Acids Res.* **1993**, *21*, 4411−4412.
(45) Halling-Sørensen, B.; Sengeløv, G.; Tjørnelund, J. Toxicity of Tetracyclines and Tetracycline Degradation Products to Environmentally Relevant Bacteria. *Arch. Environ. Contam. Toxicol.* **2002**, *42*, 263−271.
(46) Ferrari, R.; Rivetti, C.; Dieci, G. Transcription reinitiation properties of bacteriophage T7 RNA polymerase. *Biochem. Biophys. Res. Commun.* **2004**, *315*, 376−380.
(47) Ponta, H.; Rahmsdorf, H. J.; Pai, S. H.; Hirsch-Kauffmann, M.; Herrlich, P.; Schweiger, M. Control of gene expression in bacteriophage T7: Transcriptional controls. *Mol. Gen. Genet.* **1974**, *134*, 281−287.
(48) Endy, D.; Kong, D.; Yin, J. Intracellular Kinetics of a Growing Virus: A Genetically Structured Simulation for Bacteriophage T7. *Biotechnol. Bioeng.* **1997**, *55*, 375−389.
(49) Chan, L. Y.; Kosuri, S.; Endy, D. Refactoring bacteriophage T7. *Mol. Syst. Biol.* **2005**, DOI: 10.1038/msb4100025.

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

(50) Hu, Y. O. O.; Hugerth, L. W.; Bengtsson, C.; Alisjahbana, A.; Seifert, M.; Kamal, A.; Sjöling, Å.; Midtvedt, T.; Norin, E.; Du, J.; Engstrand, L. Bacteriophages Synergize with the Gut Microbial Community To Combat Salmonella. *mSystems* **2018**, *3*, e00119-18.
(51) Ganeshan, S. D.; Hosseinidoust, Z. Phage Therapy with a Focus on the Human Microbiota. *Antibiotics* **2019**, *8*, 131.
(52) Hennig, S.; Rödel, G.; Ostermann, K. Artificial cell-cell communication as an emerging tool in synthetic biology applications. *J. Biol. Eng.* **2015**, *9*, 1−12.
(53) Bertram, R.; Hillen, W. The application of Tet repressor in prokaryotic gene regulation and expression. *Microb. Biotechnol.* **2008**, *1*, 2−16.
(54) Schmitter, S.; Fieseler, L.; Klumpp, J.; Bertram, R.; Loessner, M. J. TetR-dependent gene regulation in intracellular Listeria monocytogenes demonstrates the spatiotemporal surface distribution of ActA. *Mol. Microbiol.* **2017**, *105*, 413−425.
(55) Lortal, S.; Chapot-Chartier, M. P. Role, mechanisms and control of lactic acid bacteria lysis in cheese. *Int. Dairy J.* **2005**, *15*, 857−871.
(56) Vrancken, G.; Gregory, A. C.; Huys, G. R. B.; Faust, K.; Raes, J. Synthetic ecology of the human gut microbiota. *Nat. Rev. Microbiol.* **2019**, *17*, 754−763.
(57) Mabwi, H. A.; Kim, E.; Song, D. G.; Yoon, H. S.; Pan, C. H.; Komba, E. V. G.; Ko, G. P.; Cha, K. H. Synthetic gut microbiome: Advances and challenges. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 363−371.
(58) de Souza, R. S. C.; Armanhi, J. S. L.; Arruda, P. From Microbiome to Traits: Designing Synthetic Microbial Communities for Improved Crop Resiliency. *Front. Plant Sci.* **2020**, *11*, 1179.
(59) Durfee, T.; Nelson, R.; Baldwin, S.; Plunkett, G.; Burland, V.; Mau, B.; Petrosino, J. F.; Qin, X.; Muzny, D. M.; Ayele, M.; Gibbs, R. A.; Csörgo, B.; Pósfai, G.; Weinstock, G. M.; Blattner, F. R. The complete genome sequence of Escherichia coli DH10B: Insights into the biology of a laboratory workhorse. *J. Bacteriol.* **2008**, *190*, 2597−2606.
(60) Baker Brachmann, C.; Davies, A.; Cost, G. J.; Caputo, E.; Li, J.; Hieter, P.; Boeke, J. D. Designer Deletion Strains derived from Saccharomyces cerevisiae S288C: a Useful set of Strains and Plasmids for PCR-mediated Gene Disruption and Other Applications. *Yeast* **1998**, *14*, 115−132.
(61) Yu, Y.-P.; Gong, T.; Jost, G.; Liu, W.-H.; Ye, D.-Z.; Luo, Z.-H. Isolation and characterization of five lytic bacteriophages infecting a Vibrio strain closely related to Vibrio owensii. *FEMS Microbiol. Lett.* **2013**, *348*, 112−119.
(62) Chaudhry, W. N.; Haq, I. U.; Andleeb, S.; Qadri, I. Characterization of a virulent bacteriophage LK1 specific for Citrobacter freundii isolated from sewage water. *J. Basic Microbiol.* **2014**, *54*, 531−541.
(63) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018. https://www.R-project.org/.
(64) Goeman, J. J.; Solari, A. Multiple hypothesis testing in genomics. *Stat. Med.* **2014**, *33*, 1946−1978.
(65) Bland, j. M.; Altman, D. G. Multiple significance tests: The Bonferroni method. *BMJ.* **1995**, *310*, 170.

https://doi.org/10.1021/acssynbio.1c00414
ACS Synth. Biol. 2022, 11, 286−296

Supporting Information

Engineering A Dynamic Controllable Infectivity Switch In Bacteriophage T7

Chutikarn Chitboonthavisuk[1,2,3,‡], Chun Huai Luo[1,2,‡], Phil Huss[1,2,3], Mikayla Fernholz[1]

and Srivatsan Raman[1,2,4,*]

1 Dept. of Biochemistry, Univ. of Wisconsin-Madison, Madison, WI, 53706, USA
2 Dept. of Bacteriology, Univ. of Wisconsin-Madison, Madison, WI, 53706, USA
3 Microbiology Doctoral Training Program, University of Wisconsin-Madison
4 Dept. of Chemical & Biological Eng., Univ. of Wisconsin-Madison, Madison, WI, 53706, USA
‡ These authors contributed equally to this work
* Corresponding author: sraman4@wisc.edu

Supporting Information

Figure S1. Expression of repressors has no effect on wildtype T7

Bacterial cell density over time measured by absorbance (OD_{600}) of wildtype T7 (T7_WT) on wildtype host (*E. coli*_WT; black) and host expressing NalC (*E. coli*_NalC, red), PmeR (*E. coli*_PmeR, green) or TetR (*E. coli*_TetR, blue) repressor. Phages were applied at time 0 at a MOI of 10^{-3} and 10^{-5}. There is no delay of estimated latent period when any repressor is expressed compared to wildtype host. All data represented as mean OD_{600} ± SD in technical triplicates.

Figure S2. Activity of engineered phages is rescued by aTC inducer

Bacterial cell density over time measured by absorbance (OD_{600}) for *E. coli* expressing TetR repressor after application of no phages (T7_None, grey), wildtype phage (T7_WT, black), engineered T7 with an RBS at 100% strength (T7_TetR_RBS1, dark blue), 39% strength (T7_TetR_RBS2, light blue), 3% strength (T7_TetR_RBS3, yellow) and 1% strength (T7_TetR_RBS4, orange). Phages and 1 μM aTC was added at timepoint. Addition of aTC rescues phage function and results in comparable time to lysis for engineered phages and T7_WT. All data represented as mean OD_{600} ± SD in biological triplicates.

Figure S3. Inducer has no effect on bacterial growth or phage activity

Bacterial cell density over time measured by absorbance (OD_{600}) for **(A)** *E. coli* expressing TetR (*E. coli*_TetR) without phage (T7_None), **(B)** wildtype *E. coli* (*E. coli*_WT) and engineered phage with a 1% strength RBS (T7_TetR_RBS4), and **(C)** *E. coli* expressing

TetR (*E. coli*_TetR) and wildtype phage (T7_WT). Anhydrotetracycline (aTC) is added at concentrations from 0-1000 nM (color gradient) at time zero. *E. coli* is able to grow productively under all aTC concentrations and application of phages results in a reduction of cell density at the same time for all aTC concentrations. All data represented as mean OD_{600} ± SD in biological triplicates.

Figure S4. Engineered phages are viable under different growth conditions

Phage titer quantified by plaque assay for engineered phage T7_TetR_RBS4 on *E. coli* MG1655 under different temperature (25ºC, 30ºC, 37ºC, 45ºC) and media conditions (LB rich media, Minimal M9 media with glucose and thiamine) in **(A)** exponential phase culture and **(B)** stationary phase culture. Compared to standard incubation in LB at 37ºC, T7_TetR_RBS4 demonstrates activity in all growth conditions with only a slightly reduced titer in exponential culture grown at 45ºC. All data represented as mean PFU/ml ± SD.

Table S1. Primers and Plasmids used in this study

Table S2. Promoter and Ribosome Binding Site sequences

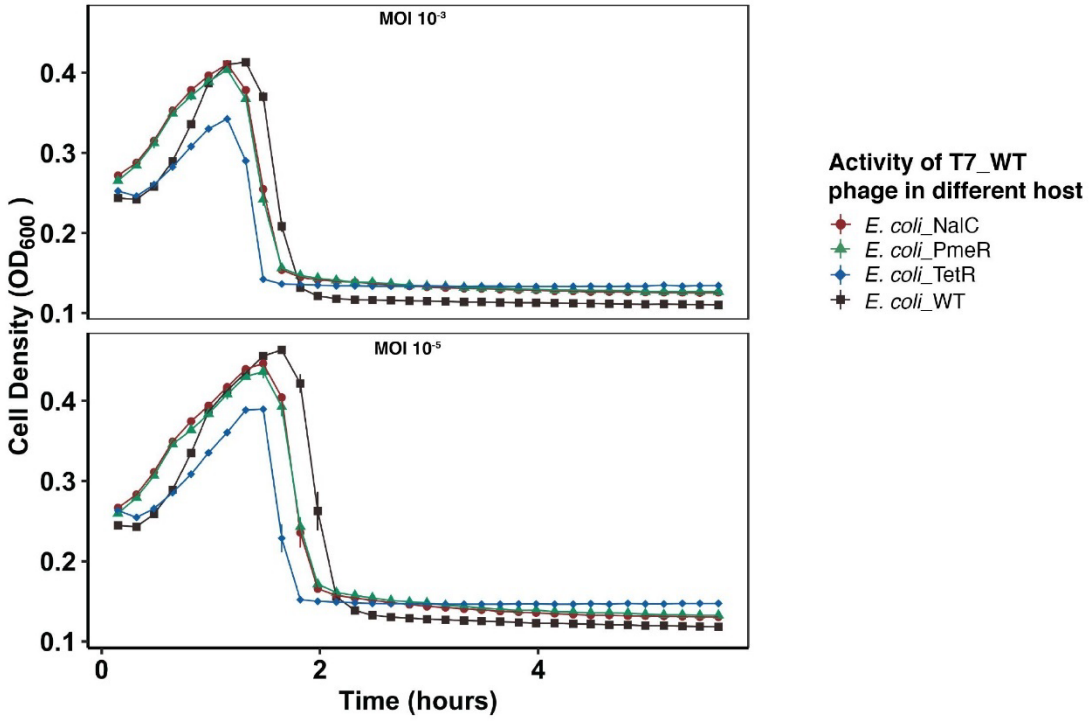


Figure S1. Expression of repressors has no effect on wildtype T7

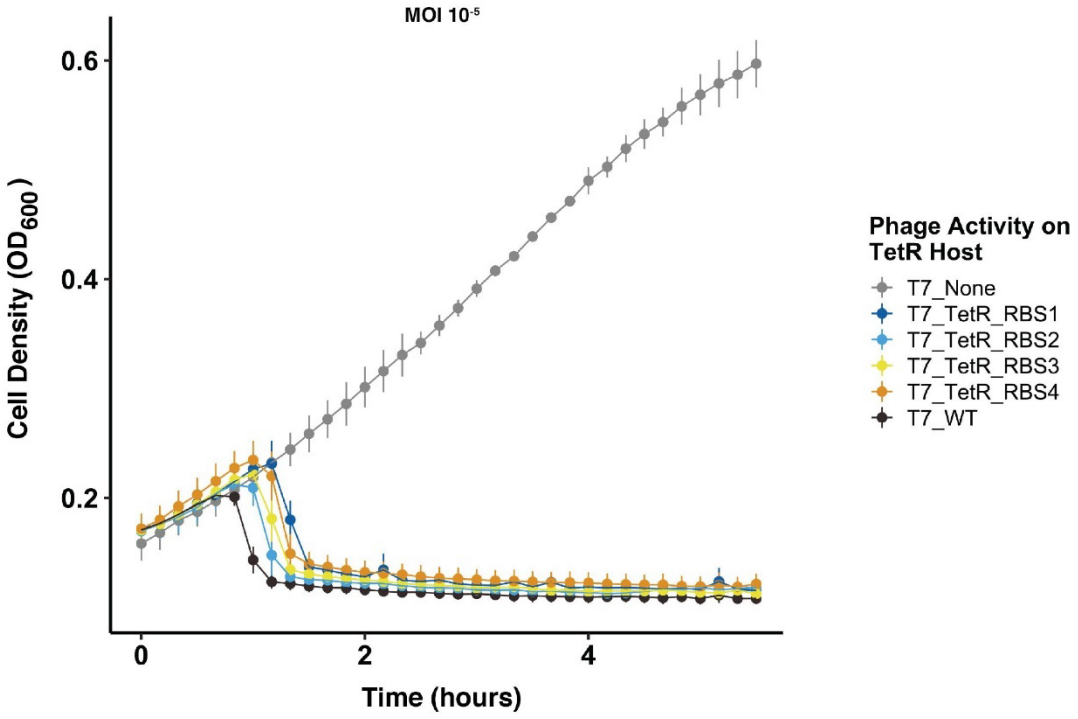


Figure S2. Activity of engineered phages is rescued by aTC inducer

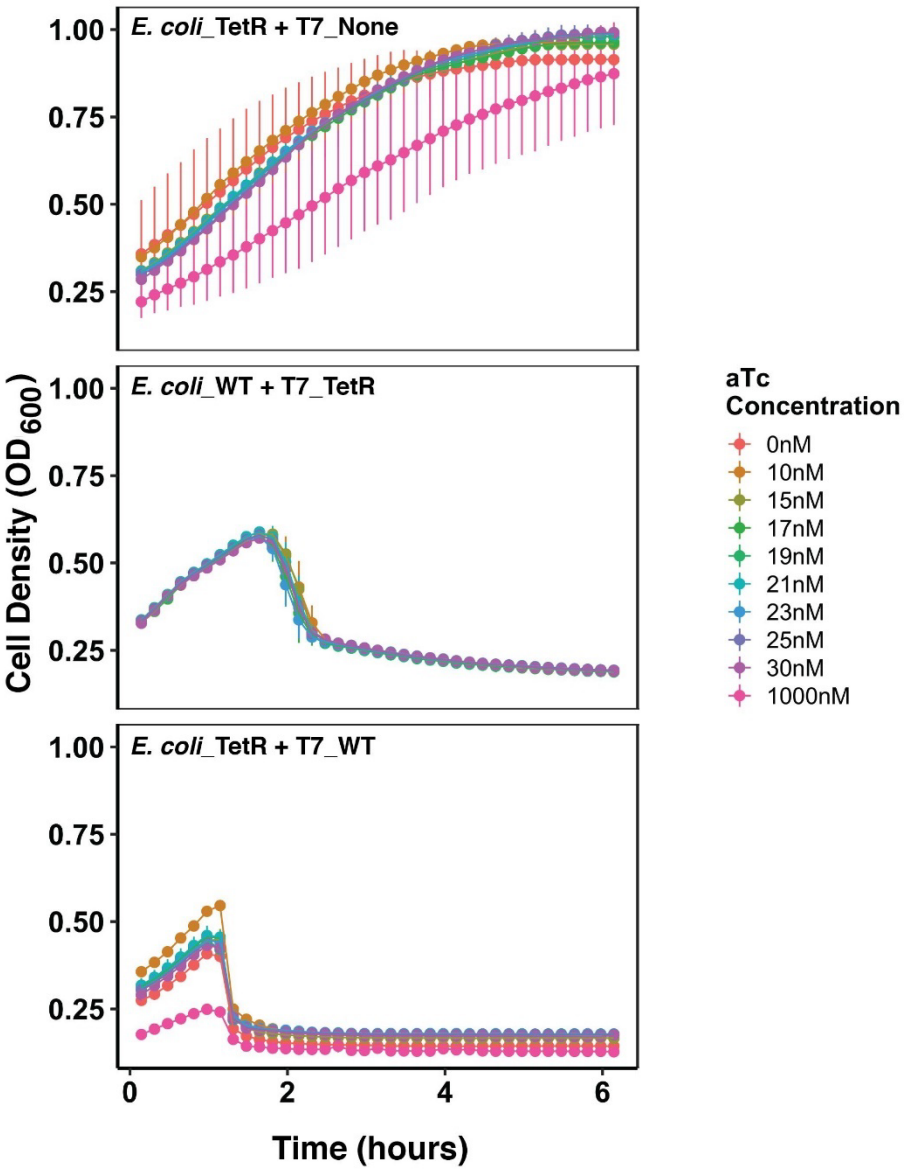


Figure S3. Inducer has no effect on bacterial growth or phage activity

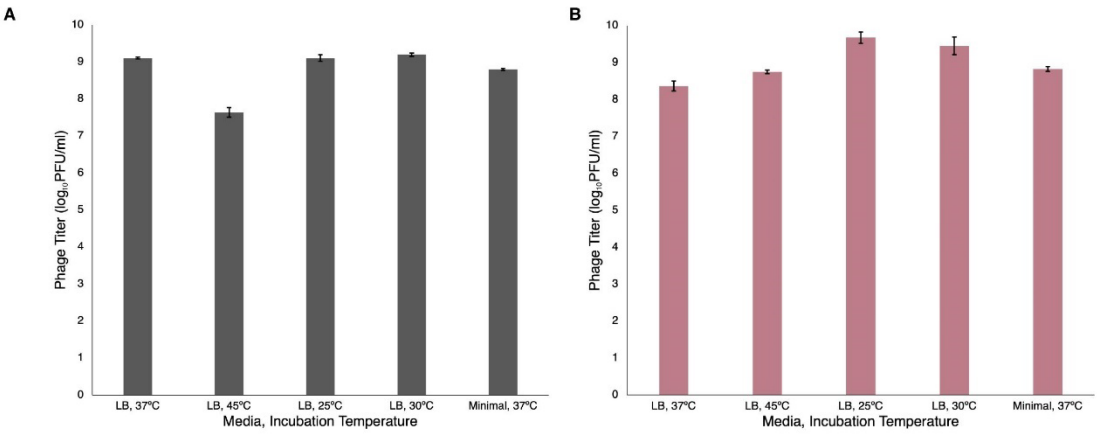


Figure S4. Engineered phages are viable under different growth conditions

Name	Sequence	Purpose
#540	TGTCTCATGAGCGGATACATATTTGAATGT	prs315 backbone segment for acceptor YAC cloning. Contains the CEN
#1361	GCATGTGTCAGAGGTTTTCACCG	prs315 backbone segment for acceptor YAC cloning. Contains the CEN
#1362	CGCGTTTCGGTGATGACGG	prs315 backbone segment for acceptor YAC cloning. Seperates the CEN and Leu marker to inhibit recircularization
#1363	GGGTAGACGAAACTATATACGC	prs315 backbone segment for acceptor YAC cloning. Seperates the CEN and Leu marker to inhibit recircularization
#1364	GCGTATATAGTTTCGTCTACCC	prs315 backbone segment for acceptor YAC cloning. Contains the Leu marker
#541	CCTTGTTCATGTGTGTTCAAAAACGTTATA	prs315 backbone segment for acceptor YAC cloning. Contains the Leu marker
#542	TATAACGTTTTTGAACACACATGAACAAGGT CTCACAGTGTACGGACCTAAAGTTCCCCC	T7 genome segment
#543	ATTACGCGATGACAGTAGACAACCTTTCCG	T7 genome segment
#544	TGCAGCAATACCGGAAAGGTTGTCTACTGT	T7 genome segment
#545	ATATGTCTCCTCATAGATGTGCCTATGTGG	T7 genome segment
#546	ACTTGTGACTCCACATAGGCACATCTATGA	T7 genome segment
#547	GAATAACCTGAGGGTCAATACCCTGCTTGT	T7 genome segment
#548	GACATGATGGACAAGCAGGGTATTGACCCT	T7 genome segment
#549	ACATTCAAATATGTATCCGCTCATGAGACAA GGGACACAGAGAGACACTCAAGGTAACAC	T7 genome segment
#3548	TAATTCACCATACACATATGCATATGAACACGATTAACATCGCTAAGAA	Used to insert 1% Dead RBS
#3549	ATATGCATATGTGTATGGTGAATTACTGCTGATGTGCTCAGTATC	Used to insert 1% Dead RBS
#3550	AAGACCCGAGACACCATATGCATATGAACACGATTAACATCGCTAAG	Used to insert 3% J61133 RBS
#3551	TATGGTGTCTCGGGTCTTTCTCTAGAAATTACTGCTGATGTGCTCAGTATC	Used to insert 3% J61133 RBS
#3554	AGATAGGAGACACCATATGCATATGAACACGATTAACATCGCTAAGA	Used to insert 39% J61106 RBS
#3555	ATGCATATGGTGTCTCCTATCTTTCTCTAGAAATTACTGCTGATGTGCTCAGTATC	Used to insert 39% J61106 RBS
#3955	CCGTACAGTTGACATGTGTACCGTACCGTACAGATACTGAGCACATCAGCAGTAATTG	Used to insert NalC1 opetator site
#3956	GGTACACATGTCAACTGTACGGTACGGTACACAGACGTTAGGAGGTGACTTTAGG	Used to insert NalC1 opetator site
#3957	GGATGTATGTTGACATACATTCGTGGATGTATGATACTGAGCACATCAGCAGTAATTG	Used to insert PmeR2 opetator site
#3958	CCACGAATGTATGTCAACATACATCCACGAATGTAGACGTTAGGAGGTGACTTTAGGAG	Used to insert PmeR2 opetator site
#2401	TCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGATAGAGATACTGAGC	Used to insert TetO opetator site
#3959	CTCCTAAAGTCACCTCCTAACGTCTCCCTATCAGTGATAGAGAT	Used to insert TetO opetator site
#3960	CACTGATAGGGAGACGTTAGGAGGTGACTTTAGGAGG	Used to insert TetO opetator site
#2171	GAGGTTTCCACTGAGGACGAAG	Multiplex primers used to assess YAC assembly
#2172	CCTTTGGTAACCATGTCCTCGG	Multiplex primers used to assess YAC assembly
#2173	GCGAATCAAGGTGACTTCCTC	Multiplex primers used to assess YAC assembly
#2174	GGAGTTCCTATGGATAGGTGTTCAGG	Multiplex primers used to assess YAC assembly
#996	ATGGGCGCTATGGGTTCAAA	Multiplex primers used to assess YAC assembly
#286	CATGCTGCTTCATGTGGTCC	Multiplex primers used to assess YAC assembly
#1061	CTCCGTGAGTCCATGCAGTT	Multiplex primers used to assess YAC assembly
#1062	GGCGATCCCCCTAGAGTCTT	Multiplex primers used to assess YAC assembly
#1002	GCCCCTGTGTGTTCTCGTTA	Multiplex primers used to assess YAC assembly
#1003	GTGTGGTGACTCACTGTCGT	Multiplex primers used to assess YAC assembly
#3961	CTAGCTGCGGAGGGTTTCATATGAACACGATTAACATCGCTAAG	Used to amplify gp1 to create pHT7_gp1
#3962	ATCCGAGACGACCTGCATTATTACGCGAACGCGAAGTCCG	Used to amplify gp1 to create pHT7_gp1
#3963	TAATGCAGGTCGTCTCGGATCGAG	Used to amplify pBR backbone to create pHT7_gp1
#3964	AATCTTAATCTAGCTGCGGAGGGTTTCAT	Used to amplify pBR backbone to create pHT7_gp1

Table S1. Primers and Plasmids used in this study

Sequence type	Name	Name_on_paper	Sequence
promoter	minus 35 concensus sequence used in all engineered phages		TTGACA
promoter	minus 10 concensus sequence used in all engineered phages		GATACT
repressor	NalC	*E.coli* _NalC	ATGAACGATGCGAGCCCGCGTCTGACCGAGCGTGGCCGTCAGCGTCGTCGTGCGATGCTGGATGCGGCGACCCAA GCGTTCCTGGAGCACGGTTTTGAAGGCACCACCCTGGATATGGTGATCGAACGTGCGGGTGGTAGCCGTGGCACC CTGTACAGCAGCTTCGGTGGCAAAGAGGGTCTGTTTGCGGCGGTTATTGCGCACATGATCGGTGAAATTTTTGAC GATAGCGCGGACCAGCCGCGTCCGGCGGCGACCCTGAGCGCGACCCTGGAGCACTTCGGTCGTCGTTTTCTGACCA GCCTGCTGGACCCGCGTTGCCAGAGCCTGTATCGTCTGGTGGTTGCGGAAAGCCCGCGTTTCCCGGCGATCGGCAA GAGCTTTTACGAGCAAGGCCCGCAGCAAAGCTATCTGCTGCTGAGCGAACGTCTGGCGGCGGTGGCGCCGCACAT GGACGAGGAAACCCTGTACGCGGTTGCGTGCCAGTTCCTGGAGATGCTGAAGGCGGACCTGTTTCTGAAAGCGCT GAGCGTGGCGGATTTCCAACCGACGATGGCGCTGCTGGAGACCCGTCTGAAACTGAGCGTTGATATCATTGCGTG CTATCTGGAACACCTGAGCCAGAGCCCGGCGCAAGGC
repressor	PmeR	*E.coli* _PmeR	ATGGTGCGTCGTACCAAGGAAGAGGCGCAGATCACCCGTAGCCAAATTCTGGAGGCGGCGGAACAGGCGTTCTAT GAGCGTGGTGTGGCGCGTACCACCCTGGCGGATATTGCGACCCTGGCGGGTGTTACCCGTGGCGCGATTTATTGG CACTTTAACAACAAAGCGGACCTGGTTCAGGCGATGCTGGATAGCCTGCAAGAGCCGCTGGACGAAATGGCGCAG GCGAGCCAAAGCGAAGACGAGGAAGATCCGCTGGGCTGCATGCGTAACCTGCTGATCCACCTGTTCCATGAGCTG GCGCTGGACCCGAAGACCCGTCGTATCAACGAAATTCTGTTCCACAAATGCGAGTTTACCGACGAAATGTGCGATTT TCGTCGTCAGCGTCAAGATAACGCGATCCAGTGCCACGACCGTATTACCCTGGGTCTGAGCAACGCGGTGCGTCAG GGTCAACTGCCGCAAGAACTGGATACCGGTCGTGCGGCGGTGGCGCTGTTCAGCTACGTTAACGGCATCATTTATC AGTGGCTGCTGGTTCCGGACAGCTTTAGCCTGCCGGCGGAGGCGGAACAACTGGTGGACGTTTGCCTGGATATGC TGCGTTTCAGCCCGACCCTGCGTGTGAAGAACAGCCCGGAACTGAACGTTGAGCGTGAACGTCTG
repressor	TetR	*E.coli* _TetR	ATGTCTAGATTAGATAAAAGTAAAGTGATTAACAGCGCATTAGAGCTGCTTAATGAGGTCGGAATCGAAGGTTTA ACAACCCGTAAACTCGCCCAGAAGCTAGGTGTAGAGCAGCCTACATTGTATTGGCATGTAAAAAATAAGCGGGCTT TGCTCGACGCCTTAGCCATTGAGATGTTAGATAGGCACCATACTCACTTTTGCCCTTTAGAAGGGGAAAGCTGGCAA GATTTTTTACGTAATAACGCTAAAAGTTTTAGATGTGCTTTACTAAGTCATCGCGATGGAGCAAAAGTACATTTAGG TACACGGCCTACAGAAAAACAGTATGAAACTCTCGAAAATCAATTAGCCTTTTTATGCCAACAAGGTTTTTCACTAGA GAATGCATTATATGCACTCAGCGCTGTGGGGCATTTTACTTTAGGTTGCGTATTGGAAGATCAAGAGCATCAAGTC GCTAAAGAAGAAAGGGAAACACCTACTACTGATAGTATGCCGCCATTATTACGACAAGCTATCGAATTATTTGATCA CCAAGGTGCAGAGCCAGCCTTCTTATTCGGCCTTGAATTGATCATATGCGGATTAGAAAAACAACTTAAATGTGAAA GTGGGTCT

Table S2. Promoter and Ribosome Binding Site Sequences (1)

Sequence type	Name	Name_on_paper	Sequence
operator site	NalC1	T7_NalC	TGTGTACCGTACCGTACAG
operator site	PmeR2	T7_PmeR	TACATTCGTGGATGTATG
operator site	TetO	T7_TetR	TCCCTATCAGTGATAGAGA
ribosome binding site	Dead 1%	T7_TetR_RBS4	CACCATACACATATG
ribosome binding site	J61133 3%	T7_TetR_RBS3	TCTAGAGAAAGACCCGAGACACCATATG
ribosome binding site	J61106 39%	T7_TetR_RBS2	TCTAGAGAAAGATAGGAGACACCATATG
ribosome binding site	Bujard 100%	T7_TetR_RBS1	GAATTCATTAAAGAGGAGAAAGGT
gp1 gene	gp1	gp1	ATGAACACGATTAACATCGCTAAGAACGACTTCTCTGACATCGAACTGGCTGCTATCCCGTTCAACACTCTGGCTGACCATTACGGTG AGCGTTTAGCTCGCGAACAGTTGGCCCTTGAGCATGAGTCTTACGAGATGGGTGAAGCACGCTTCCGCAAGATGTTTGAGCGTCA ACTTAAAGCTGGTGAGGTTGCGGATAACGCTGCCGCCAAGCCTCTCATCACTACCCTACTCCCTAAGATGATTGCACGCATCAACGAC TGGTTTGAGGAAGTGAAAGCTAAGCGCGGCAAGCGCCCGACAGCCTTCCAGTTCCTGCAAGAAATCAAGCCGGAAGCCGTAGCGT ACATCACCATTAAGACCACTCTGGCTTGCCTAACCAGTGCTGACAATACAACCGTTCAGGCTGTAGCAAGCGCAATCGGTCGGGCCA TTGAGGACGAGGCTCGCTTCGGTCGTATCCGTGACCTTGAAGCTAAGCACTTCAAGAAAAACGTTGAGGAACAACTCAACAAGCGC GTAGGGCACGTCTACAAGAAAGCATTTATGCAAGTTGTCGAGGCTGACATGCTCTCTAAGGGTCTACTCGGTGGCGAGGCGTGGT CTTCGTGGCATAAGGAAGACTCTATTCATGTAGGAGTACGCTGCATCGAGATGCTCATTGAGTCAACCGGAATGGTTAGCTTACAC CGCCAAAATGCTGGCGTAGTAGGTCAAGACTCTGAGACTATCGAACTCGCACCTGAATACGCTGAGGCTATCGCAACCCGTGCAGG TGCGCTGGCTGGCATCTCTCCGATGTTCCAACCTTGCGTAGTTCCTCCTAAGCCGTGGACTGGCATTACTGGTGGTGGCTATTGGGC TAACGGTCGTCGTCCTCTGGCGCTGGTGCGTACTCACAGTAAGAAAGCACTGATGCGCTACGAAGACGTTTACATGCCTGAGGTGT ACAAAGCGATTAACATTGCGCAAAACACCGCATGGAAAATCAACAAGAAAGTCCTAGCGGTCGCCAACGTAATCACCAAGTGGAAG CATTGTCCGGTCGAGGACATCCCTGCGATTGAGCGTGAAGAACTCCCGATGAAACCGGAAGACATCGACATGAATCCTGAGGCTCT CACCGCGTGGAAACGTGCTGCCGCTGCTGTGTACCGCAAGGACAAGGCTCGCAAGTCTCGCCGTATCAGCCTTGAGTTCATGCTTG AGCAAGCCAATAAGTTTGCTAACCATAAGGCCATCTGGTTCCCTTACAACATGGACTGGCGCGGTCGTGTTTACGCTGTGTCAATGT TCAACCCGCAAGGTAACGATATGACCAAAGGACTGCTTACGCTGGCGAAAGGTAAACCAATCGGTAAGGAAGGTTACTACTGGCT GAAAATCCACGGTGCAAACTGTGCGGGTGTCGATAAGGTTCCGTTCCCTGAGCGCATCAAGTTCATTGAGGAAAACCACGAGAAC ATCATGGCTTGCGCTAAGTCTCCACTGGAGAACACTTGGTGGGCTGAGCAAGATTCTCCGTTCTGCTTCCTTGCGTTCTGCTTTGAG TACGCTGGGGTACAGCACCACGGCCTGAGCTATAACTGCTCCCTTCCGCTGGCGTTTGACGGGTCTTGCTCTGGCATCCAGCACTTCT CCGCGATGCTCCGAGATGAGGTAGGTGGTCGCGCGGTTAACTTGCTTCCTAGTGAAACCGTTCAGGACATCTACGGGATTGTTGCT AAGAAAGTCAACGAGATTCTACAAGCAGACGCAATCAATGGGACCGATAACGAAGTAGTTACCGTGACCGATGAGAACACTGGTG AAATCTCTGAGAAAGTCAAGCTGGGCACTAAGGCACTGGCTGGTCAATGGCTGGCTTACGGTGTTACTCGCAGTGTGACTAAGCG TTCAGTCATGACGCTGGCTTACGGGTCCAAAGAGTTCGGCTTCCGTCAACAAGTGCTGGAAGATACCATTCAGCCAGCTATTGATTC CGGCAAGGGTCTGATGTTCACTCAGCCGAATCAGGCTGCTGGATACATGGCTAAGCTGATTTGGGAATCTGTGAGCGTGACGGTG GTAGCTGCGGTTGAAGCAATGAACTGGCTTAAGTCTGCTGCTAAGCTGCTGGCTGCTGAGGTCAAAGATAAGAAGACTGGAGAG ATTCTTCGCAAGCGTTGCGCTGTGCATTGGGTAACTCCTGATGGTTTCCCTGTGTGGCAGGAATACAAGAAGCCTATTCAGACGCG CTTGAACCTGATGTTCCTCGGTCAGTTCCGCTTACAGCCTACCATTAACACCAACAAAGATAGCGAGATTGATGCACACAAACAGGA GTCTGGTATCGCTCCTAACTTTGTACACAGCCAAGACGGTAGCCACCTTCGTAAGACTGTAGTGTGGGCACACGAGAAGTACGGAA TCGAATCTTTTGCACTGATTCACGACTCCTTCGGTACCATTCCGGCTGACGCTGCGAACCTGTTCAAAGCAGTGCGCGAAACTATGGT TGACACATATGAGTCTTGTGATGTACTGGCTGATTTCTACGACCAGTTCGCTGACCAGTTGCACGAGTCTCAATTGGACAAAATGCC AGCACTTCCGGCTAAAGGTAACTTGAACCTCCGTGACATCTTAGAGTCGGACTTCGCGTTCGCGTAA
Terminal repeat and SRL	T7 left-end (LE) region	T7 LE region	TCTCACAGTGTACGGACCTAAAGTTCCCCCATAGGGGGTACCTAAAGCCCAGCCAATCACCTAAAGTCAACCTTCGGTTGACCTTGA GGGTTCCCTAAGGGTTGGGGATGACCCTTGGGTTTGTCTTTGGGTGTTACCTTGAGTGTCTCTCTGTGTCCCTATCTGTTACAGTCT CCTAAAGTATCCTCCTAAAGTCACCTCCTAACGTC

Table S2. Promoter and Ribosome Binding Site Sequences (2)