UNIVERSITY OF CALIFORNIA,
IRVINE

Examining the Processes of Microbial Genotypic and Phenotypic Adaptation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Tiffany Nada Batarseh

Dissertation Committee:
Professor Brandon S. Gaut, Chair
Associate Professor Adam Martiny
Assistant Professor Alejandra Rodríguez-Verdugo

2022

DEDICATION

To

my parents

Adela Ester Batarseh and Issa Eid Batarseh

Thank you for your impressive dedication, work ethic, and love.

This dissertation is dedicated to all first-generation students.

You belong in academia, and you are valued.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Brandon S. Gaut for being an amazing mentor, role model, and incredible scientist. Your unconditional support was absolutely incredible, and I was very fortunate for the opportunity to join your lab. I have learned so much from you and I strive to embody your values and have become a better person from working in your lab. I also would like to thank Rebecca Gaut for her constant support and laughs in and out of the lab.

Thank you to my committee members: Dr. Adam Martiny, Dr. Alejandra Rodriguez-Verdugo, and Dr. J.J. Emerson. Your support, guidance, and input were invaluable to my development as a scientist. Thank you for your time and sharing your knowledge with me to help with improving my dissertation research.

I would also like to thank my parents, family, friends, and colleagues for all of the support over the years. I could not have taken on this degree without having a community support structure behind me, and I am thankful to have such a caring, thoughtful group of people in my life.

Chapter 1: We thank R. Gaut, P. McDonald and the UC Irvine Genomics High-Throughput Facility for contributing to data generation. We also thank A. Rodríguez-Verdugo for comments on the manuscript. T. Batarseh was supported by the National Science Foundation Graduate Research Fellowship Program. S. Hug was supported by the National Institute of Biomedical Imaging and Bioengineering, National Research Service Award EB009418 from the University of California, Irvine, Center for Complex Biological Systems. The work was supported by National Science Foundation grant DEB-0748903.

Chapter 2: We thank R. Gaut and the UC Irvine Genomics High-Throughput Facility for contributing to data generation. We also thank A. Martiny for helpful input on methodology. T. Batarseh was supported by the National Science Foundation Graduate Research Fellowship Program and the University of California Irvine President's Dissertation Year Fellowship.

Chapter 3: We thank three anonymous reviewers for comments. We also thank R. Gaut, INRA-CFBP, and the UC Irvine Genomics High-Throughput Facility for contributing to data generation. We also thank J.J. Emerson, A. Martiny, E. Solares, and C.I. Rodriguez for helpful input on methodology. T.N.B. was supported by the National Science Foundation Graduate Research Fellowship Program and the University of California Irvine President's Dissertation Year Fellowship. This work was supported by National Science Foundation grant no. 1741627 to B.S.G. and California Department of Food and Agriculture agreement numbers 15-0218-SA and 18-0328-000-SA to M.C.R.

VITA

Tiffany Nada Batarseh

EDUCATION

Doctor of Philosophy in Biological Sciences. University of California Irvine 2022
Master of Science in Biological Sciences. University of California Irvine 2019
Bachelor of Science in Biological Sciences. University of California Irvine 2016

PUBLICATIONS

Tiffany N Batarseh, Shaun M Hug, Sarah N Batarseh, Brandon S Gaut, Genetic Mutations That Drive Evolutionary Rescue to Lethal Temperature in *Escherichia coli*, *Genome Biology and Evolution*, Volume 12, Issue 11, November 2020, Pages 2029–2044, https://doi.org/10.1093/gbe/evaa174

GRANTS, FELLOWSHIPS, AND FUNDING

NSF Postdoctoral Research Fellowship in Biology 2022-2023
Rose Hills Foundation STEM Scholarship Fall 2021
UC Irvine President's Dissertation Year Fellowship Award 2021-2022 Academic year
UC Irvine Graduate Completion Fellowship Award Spring 2021
UC Irvine Howard A. Schneiderman Fellowship Award Spring 2021
NSF Graduate Research Fellowship Program Award Oct 2018 - Oct 2021
UC Irvine Diversity Recruitment Fellowship Award Fall 2016
NIH Initiative for Maximizing Student Development (IMSD) Funding Sep 2016 - Dec 2017
NIH Minority Health & Health Disparities International Research Funding June 2016 - Sep 2016
NIH Minority Access to Research Careers (MARC) Scholarship Aug 2015 - June 2016
NIH Minority Biomedical Research Support Program (MBRS) Research Funding Apr 2015 - Aug 2015

HONORS AND AWARDS

UC Irvine LEAD Award for Excellence in Research for the School of Biological Sciences 2022
Society for Molecular Biology and Evolution Young Investigator Award 2020
Sigma Xi Conference Graduate Student Research Presentation Winner November 12, 2016
AAAS Student Poster Competition Winner February 13-14, 2016
Annual Biomedical Research Conference for Minority Scientists Poster Award November 14, 2015
IANAS/UNESCO/IHP Funding Sep 3-5, 2015
UC Irvine Dean's Honor List Spring 2015, Fall 2015, Winter 2016, Spring 2016
El Camino College Award for Academic Achievement in the Field of Biology May 2014

TEACHING

Discussion Section Leader (x3) Bio 94: Organisms to Ecosystems, UCI Summer Session 2021

Discussion Section Leader (x3) Bio 94: Organisms to Ecosystems, UCI Winter 2019

Certifications

Mentoring Excellence Certificate given by UC Irvine
Winter 2022

The Inclusive STEM Teaching Project Certificate, Boston University (INCLU1x BUx)
August 2021

TA Professional Development Program at UC Irvine
Included FERPA training. Winter 2018

RESEARCH SUPERVISION AND STUDENT TRAINING

Student, Degree Awarding School	*Year Trained*
Sarah Batarseh, Bachelor of Science student, UCI	2019-2021
Paul Hernandez, Bachelor of Science student, UCI	2017-2018
Veronica Castrejon-Villegas, Bachelor of Science student, UCI	2017
Melissa Emami, Bachelor of Science student, UCI	2017
Alyza Roman, Bachelor of Science student, UCI	2017
Kyelo Torres, Bachelor of Science student, UCB	2017

SERVICE

UCI EEB Anti-Racism, Diversity, Equity, and Inclusion Working Group 2020-Present
UCI EEB Departmental Seminar Series Committee 2020-2021
UCI School of Biological Sciences NSF GRFP Workshop Panelist 2018-2021
UCI EEB Faculty Search Committee 2019-2020
UCI School of Biological Sciences NSF GRFP Writing Tutor 2018-2019
UCI EEB Graduate Student Invited Speaker Committee 2018-2019
UCI Bridges to Baccalaureate Workshop Instructor Summer 2017
UCI IMSD Workshop Instructor Spring 2017
UCI Minority Sciences Program Symposium Poster Judge Fall 2017
UCI Minority Sciences Program Orientation Panelist 2017
UCI Minority Sciences Program Peer Mentor 2016

OUTREACH

UCI SACNAS Chapter Volunteer 2018-2020
Irvine Unified School District Ask a Scientist Night Volunteer 2018-2019
Irvine Unified School District Science Fair Judge 2018-2019
Tech Trek Volunteer: Lab tours and science demonstrations to visiting middle school girls 2018
King's College London Volunteer: Science demonstrations for visiting secondary school students 2016
El Camino Chemistry Club Volunteer: Science activities for visiting K-12 students 2014

PRESENTATIONS

SACNAS National Diversity in STEM Digital Conference: Poster Presenter

"The Influence of Adaptive History on Microbial Evolutionary Trajectories." Virtual. Oct 25-29, 2021

American Society for Microbiology World Microbe Forum: iPoster Presenter

"Evolutionary History Influences Microbial Adaptive Evolution in a New Environment." Virtual. June 20-24, 2021

Microbial Ecology and Evolution Virtual Conference: Poster Presenter

"Genetic Mutations that Drive Evolutionary Rescue to Lethal Temperature in *Escherichia Coli."* Virtual. August 12-14, 2020

Southern California Evolutionary Genetics and Genomics Meeting: Poster Presenter

"Genome Content Evolution in *Xylella fastidiosa*." UC Irvine, California. October 19, 2019

Keystone Symposia Antimicrobials and Resistance Opportunities and Challenges: Poster Presenter

"Antibiotic Resistance Genes and Heavy Metal Resistance Genes on a Plasmid Present in *Klebsiella pneumoniae."* Santa Fe, New Mexico. October 30, 2018

Sigma Xi Annual Conference: Poster Presenter

"Characterization of peripheral blood B cells in allergic asthma and the effect of rhinovirus infection." Atlanta, Georgia. November 12, 2016

Minority Sciences Program Research Symposium: Oral Presenter

"Characterization of peripheral blood B cells in allergic asthma and the effect of rhinovirus infection." UC Irvine, California. September 26-27, 2016

MHIRT Symposium: Oral Presenter

"Characterization of peripheral blood B cells in allergic asthma and the effect of rhinovirus infection." King's College London, England. August 25, 2016

AAAS Annual Meeting: Poster Presenter

"Nine plasmids discovered in a multidrug resistant *Escherichia coli* isolate." Washington, D.C. February 13-14, 2016

Annual Biomedical Research Conference for Minority Students (ABRCMS): Poster Presenter

"Nine plasmids discovered in a multidrug resistant *Escherichia coli* isolate." Seattle, Washington. November 11-14, 2015

Minority Sciences Program Research Symposium: Oral Presenter

"Multiple plasmids discovered in a multidrug resistant *Escherichia coli* isolate." UC Irvine, California. September 21-22, 2015

ABSTRACT OF THE DISSERTATION

Examining the Processes of Microbial Genotypic and Phenotypic Adaptation

by

Tiffany Nada Batarseh

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2022

Professor Brandon S. Gaut, Chair

Adaptation by natural selection is a fundamental process in evolution, yet there is a deficit in our understanding of the mechanisms of adaptation at the genomic level and how genetic changes translate to phenotypic change. For my dissertation, I addressed questions about evolution using genomic and experimental data to better understand the phenotypic and genotypic changes underlying adaptation and to investigate the consequences of adaptation utilizing bacteria as my study system.

In my first chapter, I investigated the mechanisms of adaptation that underlie evolutionary rescue in *Escherichia coli*. In my experiment, rescue occurred for 9% of populations, and I found that one mutation in either the *rpoBC* (RNA polymerase) or *hslVU* (heat shock protease) operon was sufficient for rescue. Overall, this chapter demonstrated that a single mutation in the *rpoBC* or *hslVU* operon allowed for rescue through similar changes in gene expression, and that adaptation by rescue may be qualitatively different from adaptation to non-lethal stress.

In my second chapter, I studied evolutionary contingency and its effects on adaptive potential. To study contingency, I expanded on a large evolution experiment previously

conducted in the Gaut lab. In this experiment, over 100 originally identical populations of *E. coli* adapted to thermal stress (42.2°C) through two distinct pathways. By conducting a second evolution experiment in a novel thermal environment (19.0°C), I contrasted the evolution of a subset of the *E. coli* populations descended from either adaptive pathway. I found evidence to suggest that the adaptive history of a population may significantly influence future genotypic evolution and even phenotypic outcomes to an extent.

Finally, in my third chapter, I investigated the effects of evolution and adaptation on the genome of the plant pathogen *Xylella fastidiosa*. This bacterium causes devastating disease in many economically important crops around the world. Using maximum likelihood methods, I estimated the ratio of nonsynonymous to synonymous substitutions (dN/dS) in over 5,000 core and accessory genes found in the *Xylella* genus. By screening for positive selection using dN/dS, I identified both core and accessory genes that may affect pathogenicity, including genes involved in biofilm formation.

INTRODUCTION

Examining the processes of microbial genotypic and phenotypic adaptation

Adaptation by natural selection is a central tenet of evolution. Understanding the mechanisms of adaptation and their effects on phenotype is essential to the study of biology and all of its diversity. Adaptation is the shift of a population towards the phenotypes that are a better fit for their environment through heritable, genetic change. Historically, adaptation has been studied through the observations of phenotype, however recent advances in genomic sequencing have allowed the pairing of phenotype with genotypic data, thereby advancing our knowledge of the processes and consequences of adaptive evolution (Orr 2005). Despite these innovations, we still do not have a complete understanding of the mechanisms of adaptation at the genomic level and how those genetic changes translate to phenotypic change. In my dissertation research, I focused on identifying and investigating the genotypic changes that arose in response to environmental challenges, and also on how those changes directly affected phenotype or evolutionary consequences.

Through the observation of natural populations, we have identified clear examples of adaptation often by evidence of convergent evolution. Field studies have been instrumental in demonstrating adaptive evolution through the comparison of distinct populations that experience similar environments or selective pressures. Biologists have observed that similar traits can arise in different populations experiencing analogous environmental pressures, therefore providing evidence for adaptation. Such examples include beak size and shape in finches (Grant et al. 2004), limb length in lizards (Hagey et al. 2017; Kohlsdorf et al. 2001), and spine length in stickleback fish (McKinnon & Rundle 2002). These studies have elucidated the repeatability of evolution in similar populations, however, we still have some deficits in our knowledge of the mechanisms behind these adaptations.

With the advances of genomic sequencing, efforts have been made to identify the causative unit of selection behind these examples of convergent evolution. In the case of stickleback fishes, in which convergent evolution has occurred in freshwater lakes colonized by marine sticklebacks, researchers have identified adaptive alleles at the *Eda* locus that underlie the convergent phenotype. Sequencing confirmed that the adaptive alleles are present at low frequency in ancestral populations, suggesting that convergent evolution was driven from standing genetic variation (Colosimo et al. 2005). In regard to Darwin's finches, sequencing and phylogenetic analyses have identified the *ALX1* gene as strongly associated with beak shape diversity, however, it is likely not the only causative locus (Lamichhaney et al. 2015). Additionally, extensive gene flow and hybridization between finches has been identified as a mechanism underlying their adaptive radiations, therefore similarly implicating standing genetic variation as an important component for adaptive evolution. While these are striking examples of adaptation with both genotypic and phenotypic evidence, they may not be reflective of evolution and adaptation across all levels of life, especially when standing genetic variation may not be available for selection or *de novo* mutation is required.

Experimental evolution provides a different approach to study adaptation and test hypotheses by tracking genotypic and phenotypic changes in response to a controlled selection pressure (Long et al. 2015). Evolution experiments using macroorganisms, like the fruit fly *Drosophila melanogaster* (Phillips et al. 2016; Rose 1984), and microorganisms, like bacteria or yeast (Zeyl 2006; Johnson et al. 2021; Good et al. 2017), have allowed for the exploration of long-standing questions about the dynamics and modes of adaptive evolution. Microorganisms, however, provide at least three advantages to study evolution and adaptation. First, bacteria have large population sizes and short generation times, which allow for experimental replication and the ability to observe natural selection in real time because microbes can be evolved for thousands of generations (Lenski et al. 1991). Second, bacteria can be stored in a non-evolving state (typically by storage in glycerol at -50 to -80°C), which allows for the direct comparison between derived and ancestral populations (Howard 1956; Wiser & Lenski 2015). Finally, bacterial genomes are small relative to other organisms, which allow for robust chromosome assembly using next-generation sequencing technology and the ability to identify the genetic targets of selection. The bacteria *Escherichia coli* is frequently used in evolution experiments and has a genome that is only 4.6Mb on average while the genome of *Saccharomyces cerevisiae* is more than double the size (12Mb) and the *D. melanogaster* genome is more than 35x larger (180Mb). Altogether, bacteria are effective tools in the study of evolution and adaptation.

Experimental evolution with bacteria has greatly expanded our knowledge about evolutionary trajectories and the genetic changes associated with such change. The Long Term Evolution Experiment (LTEE) started by Richard Lenski at UC Irvine in 1988 provides a famous example of experimental evolution of *E. coli* (Lenski et al. 1991). The LTEE consisted of 12 initially identical *E. coli* populations that were propagated through daily serial transfer at 37°C. The populations have now experienced over 75,000 generations of evolution and have been instrumental in expanding our knowledge about the tempo and mode of genotypic and phenotypic evolution (Good et al. 2017; Tenaillon et al. 2016). Specifically, the LTEE has elucidated general trends describing the trajectories of fitness change over evolutionary time and have successfully connected a phenotypic innovation to the causative genetic changes (de Visser & Lenski 2002; Quandt et al. 2015).

To explore the breadth of molecular changes associated with adaptive evolution, Tenaillon et al. (2012) conducted an experiment consisting of ~115 *E. coli* lines evolved at the stressful temperature of 42.2°C. This experiment allowed researchers to investigate whether independent populations that adapt to a fixed environment converge to a similar adaptive pathway by identical mutations or by alternative genetic pathways. In this system, adaptation occurred through two distinct paths in response to the selection pressure: roughly half of the evolved lines carried mutations in *rpoB*, the gene encoding the beta-subunit of RNA polymerase, and about one third of the evolved lines carried mutations in *rho*, the transcriptional terminator (Tenaillon et al. 2012). Individually, mutations in *rpoB* and *rho* altered gene expression levels of >1000 genes (Rodriguez-Verdugo *et al.* 2016; Gonzalez-Gonzalez *et al.* 2017), with substantial overlaps in the altered genes. Interestingly, mutations in *rpoB* and *rho* occurred together in the evolved lines less often than expected by chance, possibly owing to negative epistatic interactions. These results, along with studies of the LTEE lines and other experimental systems, suggest that interactions between mutations are pervasive and have a profound effect on phenotypic outcomes (Khan et al. 2011; Kryazhimskiy et al. 2014). These observations suggest that any understanding of adaptation requires consideration of potential interactions between adaptive mutations. They also suggest that adaptation may be contingent on history because the effects of a new mutation may rely on its interactions with existing mutations.

The foundational knowledge gained from experimental evolution has only opened more research avenues regarding adaptation, its mechanisms, and its consequences on future evolution. Advances in sequencing technology have allowed for fine-scale tracking of allele frequency trajectories over evolutionary time, which has revealed the existence of coexisting lineages and the molecular dynamics underscoring phenomena like clonal interference (Lang et al. 2013; Maddamsetti et al. 2015). Additionally, by leveraging sequential evolution experiments, researchers have begun to address the effects of an organism's evolutionary history on its future adaptive trajectories. For example, Plucain et al. (2016) used a two-phase experimental evolution strategy to investigate whether history influenced adaptation in a new environment using *E. coli*. The first phase of evolution consisted of initially identical populations of *E. coli* evolving under four different environmental conditions followed by a second phase of evolution in which the populations all evolved under the same environmental conditions. The researchers found that historical contingency significantly affects phenotypic adaptation to the second environment, but they did not identify contingency with respect to genetic changes (Plucain et al. 2016). The LTEE provides another example of history affecting evolution. After over 30,000 generations of evolution, one of the twelve *E. coli* populations evolved the ability to utilize citrate which was always present in the media. To understand if this novel trait was influenced by history, the researchers repeated the evolution experiment with frozen stock of earlier generations. The authors found that the evolution of this trait was contingent on particular genetic changes, suggesting that evolutionary history can significantly affect trait evolution (Blount et al. 2020, 2008). Similarly, in the case of antibiotic resistance evolution in bacteria, the starting genetic background significantly influenced the set of future adaptive mutations and resistance profile suggesting a significant effect of historical contingency on both genotypic and phenotypic evolution (Card et al. 2021). In contrast, two-step evolution experiments with yeast have suggested that phenotypic changes are not influenced by evolutionary history and that fitness will follow predictable trajectories and phenotypic convergence is expected (Kryazhimskiy et al. 2014). The contrasting results between experiments highlights areas in which our knowledge of evolution and adaptation can be expanded.

With advances in sequencing technology, we are better equipped to study the genetic signatures of adaptation and their consequences. Adaptive evolution not only influences genome content through point mutations but also by shifts in gene content and structural variation. Within many bacterial species, there is immense variation in genome content that can influence how bacteria function by conferring alternative modes of metabolism, resistance mechanisms, or virulence factors (Ochman et al. 2000). Investigation of bacterial genomes using comparative genomics and phylogenetic methods have elucidated extensive mosaicism between bacteria of the same species. These genomes typically consist of core genes responsible for housekeeping and general cellular functions and a set of variable accessory genes that often confer virulence functions (Welch et al. 2002; Rasko et al. 2008; Chen et al. 2018). By investigating bacterial genomes, we may identify genetic determinants underlying virulence or resistance mechanisms which could be targeted for pathogen management. More generally, understanding how adaptation shapes bacterial genome content at all levels is important for understanding bacterial biodiversity across all habitats.

The goals of my dissertation are to investigate the genotypic changes that underlie evolution and adaptation and to understand how these changes may affect phenotype or future evolutionary change. In my first chapter, I first investigated evolutionary rescue which is the phenomenon by which adaptation by natural selection saves a population from extinction under lethal or rapidly deteriorating environmental conditions. To do so, I used the results from a short-term evolution experiment previously conducted in my lab to study rescue that resulted in a set of rescue populations for analysis. I isolated single mutants harboring putative adaptive mutations and performed fitness assays and mRNA sequencing to study the effects of rescue mutations and to identify a mechanism of adaptation. In my second chapter, I investigated whether adaptation may be contingent on history. I used experimental evolution to study how and to what extent previous evolutionary history affects future evolutionary and adaptive potential by expanding on and conducting a second phase of experimental evolution that followed the evolution experiment previously conducted and described in Tenaillon et al. (2012). Finally in my third chapter, I studied genome content evolution and how it contributes to host-parasite interactions by analyzing whole genomes of the plant pathogen *Xylella fastidiosa*. I used comparative genomics and phylogenetic methods to investigate whether the genomes of *X. fastidiosa* exhibit evidence of host specificity through gene content variation or by evidence of positive selection in either core or accessory genes. Altogether, my dissertation expands on our knowledge about bacterial evolution, adaptation, and its consequences.

References

Blount ZD et al. 2020. Genomic and phenotypic evolution of Escherichia coli in a novel citrate-only resource environment Rainey, PB & Wittkopp, PJ, editors. eLife. 9:e55414. doi: 10.7554/eLife.55414.

Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc. Natl. Acad. Sci. 105:7899–7906. doi: 10.1073/pnas.0803151105.

Card KJ, Thomas MD, Graves JL, Barrick JE, Lenski RE. 2021. Genomic evolution of antibiotic resistance is contingent on genetic background following a long-term experiment with Escherichia coli. Proc. Natl. Acad. Sci. 118:e2016886118. doi: 10.1073/pnas.2016886118.

Chen NWG et al. 2018. Horizontal gene transfer plays a major role in the pathological convergence of Xanthomonas lineages on common bean. BMC Genomics. 19:606. doi: 10.1186/s12864-018-4975-4.

Colosimo PF et al. 2005. Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. Science. 307:1928–1933. doi: 10.1126/science.1107239.

Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The Dynamics of Molecular Evolution Over 60,000 Generations. Nature. 551:45–50. doi: 10.1038/nature24287.

Grant PR, Grant BR, Markert JA, Keller LF, Petren K. 2004. CONVERGENT EVOLUTION OF DARWIN’S FINCHES CAUSED BY INTROGRESSIVE HYBRIDIZATION AND SELECTION. Evolution. 58:1588–1599. doi: 10.1554/04-016.

Hagey TJ, Harte S, Vickers M, Harmon LJ, Schwarzkopf L. 2017. There’s more than one way to climb a tree: Limb length and microhabitat use in lizards with toe pads. PLoS ONE. 12:e0184641. doi: 10.1371/journal.pone.0184641.

Howard DH. 1956. THE PRESERVATION OF BACTERIA BY FREEZING IN GLYCEROL BROTH, 12. J. Bacteriol. 71:625.

Johnson MS et al. 2021. Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations Verstrepen, KJ, Wittkopp, PJ, Verstrepen, KJ, & Hodgins-Davis, A, editors. eLife. 10:e63910. doi: 10.7554/eLife.63910.

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. Science. 332:1193–1196. doi: 10.1126/science.1203801.

Kohlsdorf T, Garland T, Navas CA. 2001. Limb and tail lengths in relation to substrate usage in Tropidurus lizards. J. Morphol. 248:151–164. doi: 10.1002/jmor.1026.

Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 344:1519–1522. doi: 10.1126/science.1250939.

Lamichhaney S et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature. 518:371–375. doi: 10.1038/nature14181.

Lang GI et al. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature. 500:571–574. doi: 10.1038/nature12344.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. Am. Nat. 138:1315–1341.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat. Rev. Genet. 16:567–582. doi: 10.1038/nrg3937.

Maddamsetti R, Lenski RE, Barrick JE. 2015. Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia coli. Genetics. 200:619–631. doi: 10.1534/genetics.115.176677.

McKinnon JS, Rundle HD. 2002. Speciation in nature: the threespine stickleback model systems. Trends Ecol. Evol. 17:480–488. doi: 10.1016/S0169-5347(02)02579-X.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405:299–304. doi: 10.1038/35012500.

Orr HA. 2005. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6:119–127. doi: 10.1038/nrg1523.

Phillips MA et al. 2016. Genome-wide analysis of long-term evolutionary domestication in Drosophila melanogaster. Sci. Rep. 6:39281. doi: 10.1038/srep39281.

Plucain J et al. 2016. Contrasting effects of historical contingency on phenotypic and genomic trajectories during a two-step evolution experiment with bacteria. BMC Evol. Biol. 16:86. doi: 10.1186/s12862-016-0662-8.

Quandt EM et al. 2015. Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment Kliebenstein, DJ, editor. eLife. 4:e09696. doi: 10.7554/eLife.09696.

Rasko DA et al. 2008. The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. J. Bacteriol. 190:6881–6893. doi: 10.1128/JB.00619-08.

Rose MR. 1984. Laboratory Evolution of Postponed Senescence in Drosophila melanogaster. Evolution. 38:1004–1010. doi: 10.2307/2408434.

Tenaillon O et al. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature. 536:165–170. doi: 10.1038/nature18959.

Tenaillon O et al. 2012. The Molecular Diversity of Adaptive Convergence. Science. 335:457–461. doi: 10.1126/science.1212986.

de Visser JAG, Lenski RE. 2002. Long-term experimental evolution in Escherichia coli. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. BMC Evol. Biol. 2:19. doi: 10.1186/1471-2148-2-19.

Welch RA et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. 99:17020–17024. doi: 10.1073/pnas.252529799.

Wiser MJ, Lenski RE. 2015. A Comparison of Methods to Measure Fitness in Escherichia coli. PLOS ONE. 10:e0126210. doi: 10.1371/journal.pone.0126210.

Zeyl C. 2006. Experimental evolution with yeast. FEMS Yeast Res. 6:685–691. doi: 10.1111/j.1567-1364.2006.00061.x.

CHAPTER 1

Genetic Mutations That Drive Evolutionary Rescue to Lethal Temperature in *Escherichia coli*

1.1 Abstract

Evolutionary rescue occurs when adaptation restores population growth against a lethal stressor. Here we studied evolutionary rescue by conducting experiments with *Escherichia coli* at the lethal temperature of 43.0°C, to determine the adaptive mutations that drive rescue and to investigate their effects on fitness and gene expression. From hundreds of populations, we observed that ~9% were rescued by genetic adaptations. We sequenced 26 populations and identified 29 distinct mutations. Of these populations, 21 had a mutation in the *hslVU* or *rpoBC* operon, suggesting that mutations in either operon could drive rescue. We isolated seven strains of *E. coli* carrying a putative rescue mutation in either the *hslVU* or *rpoBC* operon to investigate the mutations' effects. The single rescue mutations increased *E. coli's* relative fitness by an average of 24% at 42.2°C, but they decreased fitness by 3% at 37.0°C, illustrating that antagonistic pleiotropy likely affected the establishment of rescue in our system. Gene expression analysis revealed only 40 genes were upregulated across all seven mutations, and these were enriched for functions in translational and flagellar production. As with previous experiments with high temperature adaptation, the rescue mutations tended to restore gene expression towards the unstressed state, but they also caused a higher proportion of novel gene expression patterns. Overall, we find that rescue is infrequent, that it is facilitated by a limited number of mutational targets, and that rescue mutations may have qualitatively different effects than mutations that arise from evolution to non-lethal stressors.

1.2 Introduction

Under severe environmental stress, a population will decline rapidly and may face extinction. However, populations can adapt genetically; if an individual appears with an adaptation to the severe stress, the population may recover. This process of decline and recovery results in a U-shaped pattern of population dynamics that defines the phenomenon of evolutionary rescue (Bell 2017). It is important to understand the frequency and dynamics of rescue events, both because they affect our understanding of species' survival and also because they have practical implications for medicine, agriculture, and conservation biology. For example, evolutionary rescue drives some of the dynamics of bacterial antibiotic resistance. When exposed to potentially lethal concentrations of antibiotics, a bacterial population declines, but a resistance mutation can restore population growth (Orr and Unckless 2008; Baquero and Cantón 2017). This evolutionary response often occurs because antibiotics target a specific enzyme or structure (e.g., the ribosome), so that a single beneficial mutation inhibits the antibiotic's mechanism of action (Blair et al. 2015). Similar dynamics contribute to fungicide and pesticide resistance in agriculture, where there is often a simple genetic basis to resistance (Délye et al. 2013; Lucas et al. 2015).

Sometimes environmental challenges affect more complex physiological traits, and this issue has been addressed to some extent in experimental studies of evolutionary rescue. In one study, yeast populations were grown in a lethal concentration of salt (Bell and Gonzalez 2009). Salt tolerance in yeast is a complex, polygenic trait (Dhar et al. 2011), suggesting that evolutionary rescue could result from mutations in one of several genes or perhaps even require multiple genic changes. Studies have shown that adaptive mutations can rescue yeast populations from lethal salt conditions, but such rescue occurs infrequently, and the probability of rescue varies by population size (Bell and Gonzalez 2009). Similarly, green alga (*Chlamydomonas*) has also been used to study evolutionary rescue to another complex trait, low-light conditions (Bell 2013), showing again that rescue is infrequent and depends on population characteristics. Together, these studies highlight that rescue can alter evolutionary outcomes for complex traits. However, neither set of studies identified the genetic basis of rescue (Bell and Gonzalez 2009; Bell 2013; Gonzalez and Bell 2013), which is an important precursor for understanding the dynamics of rescue and its underlying mechanisms.

Here we study evolutionary rescue in *Escherichia coli* that has been challenged with a lethal temperature. Temperature is a complex environmental variable because it governs the rates of biological reactions that underlie respiration, growth, and reproduction (Somero 1978; Cooper et al. 2001). Furthermore, characterizing the evolutionary response to severe thermal stress is important for understanding adaptation to global climate change (Holt 1990). Previous work has investigated adaptation to both non-lethal and lethal heat stress using *E. coli*. As an example of non-lethal stress, Tenaillon *et al.* (2012) subjected *E. coli* to a high but sustainable temperature (42.2°C) and found >1,000 putatively adaptive mutations, illustrating the genetic diversity of adaptive responses. These mutations occurred within dozens of genes, but there were also clear patterns. Mutations were especially frequent in genes that modify transcription, such as the RNA polymerase subunit β (*rpoB*) gene. Some of these *rpoB* mutations conveyed a fitness benefit in high heat but fitness trade-offs at lower temperatures (<20.0°C), indicating antagonistic pleiotropy (Rodríguez-Verdugo et al. 2014).

Previous work has also subjected *E. coli* to lethal temperatures that resulted in rare rescue dynamics (Bennett and Lenski 1993). For example, Mongold *et al.* (1999) characterized patterns of evolutionary recovery at 44.0°C, using *E. coli* strains that had been adapted to 32.0°C, 37.0°C, and 41.0-42.0°C (Mongold et al. 1999). They found that rescue events at 44.0°C occurred in 8% of populations but only in populations derived from ancestors that had been previously adapted to high temperature (41.0-42.0°C), suggesting that pre-adaptation contributes to evolutionary rescue. Moreover, they found that the rescued populations exhibited a fitness cost at elevated, but non-lethal temperatures, suggesting that at least some rescue mutations were antagonistically pleiotropic. Here again, however, the underlying adaptive mutations were not identified.

In this study we perform *E. coli* growth experiments to better understand the dynamics, mechanism, and fitness consequences of evolutionary rescue. Beginning with an ancestor derived from a single colony, we carry out replicated evolution experiments at 43.0°C, which typically results in population extinction under our growth conditions. After observing and noting the frequency of rescue events, we identify mutations within the rescue populations. With these mutations in hand, we ask the following three sets of questions. First, can these mutations drive evolutionary rescue to lethal temperature in *E. coli*? Discriminating between driving and hitchhiking mutations is a major challenge in evolutionary biology (Rosenzweig and Sherlock 2014), and hence unambiguous identification of drivers is an important goal. Second, what is the fitness effect of driver mutations, and do they have trade-offs that affect their population dynamics? We are specifically interested in antagonistic pleiotropy, a phenomenon that has been shown to be common (Williams 1957; Cooper and Lenski 2000; MacLean et al. 2004) but not universal in evolution experiments. Finally, can we glean any insights into the molecular effects and mechanisms of rescue? To do so, we study gene expression changes introduced by driver mutations, to try to better understand their downstream effects. We also assess whether gene expression shifts back toward an unstressed physiological state (Carroll and Marx, 2013) or toward novel expression patterns.

1.3 Materials and Methods

Evolutionary Rescue Experiments: A frozen glycerol stock was prepared from a single colony of *Escherichia coli* B strain REL1206 possessing a neutral Ara- marker. This strain had been propagated previously at 37.0°C for 2,000 generations in Davis minimal medium supplemented with glucose at 25 mg/L (DM25) and was thus adapted to the growth medium (Lenski et al. 1991). To isolate the single colony, REL1206 was streaked from frozen onto a tetrazolium-arabinose (TA) plate and incubated overnight at 37.0°C. The single colony was inoculated into Luria-Bertani medium (LB) and grown overnight. To prepare a frozen reference stock, 900 µL of culture was mixed with 900 µL of 80% glycerol and frozen at -80°C. We term this REL1206 frozen stock the "rescue ancestor" (Figure 1.1A). A backup rescue ancestor stock was prepared from the same LB culture.

As is common practice (Bennett and Lenski 1993; Lenski and Travisano 1994; Rodríguez-Verdugo et al. 2014; Hug and Gaut 2015), we first acclimated the rescue ancestor (REL1206) to mild laboratory conditions to allow it to recover from being frozen. The rescue ancestral stock was inoculated into 100 mL LB and grown for eight hours in an Infors HT Minitron incubator at 37.0°C and 120 RPM (Figure 1.1A). 10 µL of this culture was then inoculated into 100 mL DM25 and grown for 24 hours in an Infors HT Minitron incubator at 37.0°C and 120 RPM. We inoculated 100 μL of the 37.0°C, DM25 culture into each of 44 culture tubes containing 9.9 mL DM25. An additional four culture tubes containing 9.9 mL DM25 were used as contamination controls and cell density blanks. The total set of 44 tubes were placed into an Innova 3100 water bath shaker (New Brunswick Scientific) and grown for 24 hours at 120 RPM and at the experimental temperature of 43.0°C. Tube cultures were serially propagated over the course of five days by inoculating 100 μL of culture into 9.9 mL DM25 after each day of growth (Figure 1.1A). Note that the 44 inoculated tubes did not constitute independent experiments, because they all derived from the same overnight, 100 mL DM25 culture. However, the procedure was repeated independently across seven weeks, for a total of (7 weeks x 44 =) 308 populations. Our interpretations make use of both the independent and non-independent features of the design.

To measure cell densities from individual populations, 50 μL of each culture (including blanks) were inoculated into cuvettes containing 9.9 mL Isoton II Diluent (Beckman Coulter) on each of the five days. These samples were analyzed using a Multisizer 3 Coulter counter (Beckman Coulter) to determine cell densities (particles/mL). The average particle density of the four blanks was subtracted from each sample's cell density. Of 308 experimental populations, 26 were excluded due to technical failure with the Coulter counter, leaving 296 measured populations. We defined a rescue event as a population whose cell density increased by at least an order of magnitude over the previous day's measurements. Following cell density measurement, rescue populations were saved as frozen glycerol stocks on their initial day of recovery, as well as any subsequent days. To prepare frozen stocks of rescue recovery events, 900 μL of culture was mixed with 900 µL of 80% glycerol and frozen at -80°C.

DNA Extraction and Population Sequencing: Most samples for DNA extraction and sequencing were derived from day five of the experiment, but two samples (populations #1 and #19) were derived from day four. Each of the frozen rescue populations was inoculated into ten separate culture tubes containing 9.9 mL DM25 and incubated in an Innova 3100 water bath shaker (New Brunswick Scientific) overnight at 37.0°C and 120 RPM. Cells from all ten tubes were pooled, and genomic DNA was extracted from these samples using Wizard Genomic DNA Purification Kits (Promega). Pooling was employed to filter new mutations that might have risen during the process of recovery. DNA from the rescue ancestor was extracted in the same manner but pooled from four tubes rather than 10. Genomic DNA libraries were prepared using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina). The 26 rescue populations were multiplexed and sequenced in two lanes of an Illumina HiSeq 2500 in rapid mode at UC Irvine's Genomics High-Throughput Facility (https://ghtf.biochem.uci.edu). Two ancestral samples—one working stock (Figure 1.1A) and one backup stock—were also sequenced on an Illumina HiSeq 3000 at the Bioinformatics Core Facility at the UC Davis Genome Center.

Mutations and mutation frequencies were called using breseq (Deatherage and Barrick 2014) in polymorphism mode, using the *E. coli* B REL606 genome as a reference, which differs from REL1206 in six positions that were excluded from our analyses (Barrick et al. 2009; Tenaillon et al. 2012). In theory, breseq provides information about duplications and deletions by reporting novel junctions. No evidence of novel junctions or sequencing coverage was found for large deletions in our data set, but breseq did provide some novel junction evidence for the presence of large duplications. To assess duplications more formally, we compared unique reads (mapping quality >5 in samtools 1.3) across 10 kb regions of the genome, defining duplications as regions with more than twice the average genome coverage.

Isolating Single Mutants from Rescue Populations: To purify isolates carrying a single mutation in either the *hslVU* or *rpoBC* operon, we selected populations that had a single fixed mutation in *hslVU* or *rpoBC*; that is, occurring at a frequency of 85% or greater as called through the breseq analysis. These populations were streaked from frozen stock onto TA plates and incubated at 37.0°C. Multiple single colonies were picked per line and subsequently purified on new TA plates. Purified isolates were then grown in LB and incubated at 37.0°C and 120 RPM. This culture was used to prepare frozen glycerol stocks of the purified isolates.

For each purified isolate, we designed PCR primers (https://www.idtdna.com) for all mutations found in the population at >10% frequency. We submitted the PCR products for Sanger sequencing to determine mutation presence or absence for each screened gene. From the PCR and Sanger sequencing results, we identified putative single mutant genotypes. Isolates that were positive for any non-fixed (background) mutations were eliminated from further study. To determine the validity of these putative single mutant genotypes, we performed whole genome sequencing. Total genomic DNA from the putative single mutants was extracted using the Promega Wizard Genomic DNA Purification kit, and DNA concentrations were measured with Qubit dsDNA HS Assay kits. Genomic DNA libraries were prepared for sequencing using the Illumina Nextera DNA Flex Library

Preparation kit. Libraries were multiplexed and sequencing was carried out on a single lane of Illumina HiSeq 4000 at the UCI Genomics High-Throughput Facility (https://ghtf.biochem.uci.edu). To call mutations in these isolates, the Illumina reads were mapped against the reference sequence, REL606, using breseq as described above (Deatherage and Barrick 2014).

Relative Fitness Measures: To measure the relative fitness of the single mutants, we competed populations or single mutants against the ancestral line by growing them together in the same culture tube. We used REL1207 as the ancestral strain, because it is identical to the REL1206 strain except for a neutral marker, Ara+. High temperature fitness assays were performed at 42.2°C, because REL1207 does not survive at higher temperatures in our system.

To perform assays, we revived REL1207 and either the rescue populations or single mutants from frozen into 10 mL LB and incubated at 37.0°C with 120 RPM. The next day, we diluted the overnight cultures 100-fold in saline and transferred 100 μL of this dilution into 9.9 mL DM25 media. This was then incubated at 37.0°C with 120 RPM to acclimate from frozen conditions (Bennett and Lenski 1993). After 24 hours of incubation, the cultures were transferred to fresh DM25 media and incubated at 42.2°C to acclimate to high temperature stress. The following day, we mixed the ancestral strain and a rescue mutant or population 9:1 into sterile DM25 media. This mixture was plated onto TA solid media to count the initial cell densities before competition. This mixture was incubated at 42.2°C with 120 RPM. The cells were left to compete for 24 hours, and we quantified the final cell densities of the ancestor and rescue line by plating on TA plates and counting colonies. To perform competitions at 37.0°C, we began the competition on the day following the first acclimation step in DM25 and mixed the ancestral and rescue lines at a 1:1 ratio.

To quantify relative fitness, w_r, we used the methods as in Lenski et al. (1991) and Tenaillon et al. (2012). The fitness of a mutant or population relative to the ancestor is estimated as: $w_r = [\log_2(N^M_f/N^M_i)]/[\log_2(N^A_f/N^A_i)]$, where N^M_i and N^A_i represent the initial cell densities of the mutant (or population) and the ancestor before competition, and N^M_f and N^A_f represent the final cell densities after one day of competition.

RNA Harvest, Isolation and Sequencing: In order to harvest cells for RNA extraction, we grew the single mutants to the mid-exponential phase of their growth curve. To do so, we acclimated the single mutants from frozen stock in 10 mL LB media at 37.0°C with 120 RPM in an Innova 3100 water bath. We then diluted these cultures 10,000-fold into DM25 media and incubated the cultures at 37.0°C with 120 RPM. Following 24-hours of incubation, the cultures were diluted 1,000-fold into DM25 media and acclimated to either 42.2°C or 43.0°C with 120 RPM for 24 hours as is customary to acclimate cells to stressful temperatures (Bennett and Lenski 1993; Rodríguez-Verdugo et al. 2014). The following day the growth curve was started by transferring 100 μL of the culture to 24 tubes with 9.9 mL of DM25 and incubated at either 42.2°C or 43.0°C. Cell density was measured using the Multisizer 3 Coulter counter (Beckman Coulter) in volumetric mode by diluting 50 μL of cell culture into 9.9 mL of Isoton II diluent. We measured the cell density every 30 minutes following the first five hours of growth until the cells reached the mid-exponential growth phase based on the electronic counts. Cells were concentrated through vacuum filtration of

150-200 mL of culture onto cellulose nitrate membrane filters with 0.2 μm pore size. The cells were washed off from the filters and pelleted for storage at -20.0°C in a mixture of 2 mL of Qiagen RNA Protect Bacterial reagent and DM25 media. Three replicates per single mutant line were harvested for both temperatures (7 mutants × 2 temperatures × 3 replications = 42 samples), and three replicates of the ancestral line, REL1206, were harvested at 42.2°C, for a total of 45 RNAseq samples.

The cell pellets were thawed and treated with lysozyme for 5 minutes before extracting total RNA using Qiagen RNeasy kits. RNA concentrations were measured with Qubit RNA HS assay kits and RNA quality was assessed by running an Agilent RNA-Nano chip on a bioanalyzer. We enriched for mRNA by the removal of rRNA using NEBNext rRNA depletion kits for bacteria. We prepared the RNA for Illumina sequencing using the NEB Ultra II Directional RNA Library Prep kit. All samples were uniquely barcoded and multiplexed for sequencing with Illumina NovaSeq at the UCI Genomics High Throughput Facility (https://ghtf.biochem.uci.edu).

Gene Expression Analyses: RNA sequencing reads from our study and previously sequenced reads from REL1206 grown at 42.2°C and 37°C from Rodríguez-Verdugo et al. (2016) were filtered with a custom Perl script to a quality cut-off of 20. The filtered reads were then mapped to the REL606 reference sequence using BWA version 0.7.8 with default parameters (Li and Durbin 2009). Uniquely mapping reads were used as input into HTSeq, which counts the number of uniquely mapped reads to annotation features (Anders et al. 2015). Analysis of the RNAseq counts was carried out in R (R Core Team 2019). We normalized the RNAseq counts and identified differentially expressed genes using the

DESeq2 package (Love et al. 2014). We followed previous studies (Rodríguez-Verdugo et al. 2016; González-González et al. 2017) by identifying differentially expressed genes (DEGs) as significant at *padj*<0.001 and also exhibiting log2-fold change > 2 between samples. Gene ontology enrichment analyses were performed at the online website (http://geneontology.org) using *E. coli* as the reference list (Ashburner et al. 2000; The Gene Ontology Consortium 2019).

Because we used new and previously published RNAseq data for expression analyses, we were concerned about the potential for batch effects. To assess batch effects, we compared analyses comparing single mutants at 42.2°C to previous REL1206 data at 42.2°C (*n*=2; Rodríguez-Verdugo et al. 2016) and to our new RNAseq data (*n*=3) of REL1206 at 42.2°C. The new data resulted in 20% more detected DEGs than the old data, perhaps reflecting differences in power with different sample sizes (*n* = 3 vs. 2). Importantly, however, 93% of DEGs were shared between the two analyses, and the two datasets led to qualitatively identical GO-enrichment analyses. Based on this comparison, we concluded that batch effects did not dramatically alter overall conclusions about the types and direction of genic shifts in expression. We therefore combined old and new REL1206 42.2°C samples, so that all reported comparisons to the 42.2°C ancestor were based on *n*=5 replicates.

Once detected, changes in DEG gene expression were categorized into one of four directions (restored, reinforced, novel, or unrestored) as previously described (Carroll and Marx 2013; Rodríguez-Verdugo et al. 2016; González-González et al. 2017). These directions represent the change in gene expression of a rescue mutant relative to the ancestor's gene expression at 42.2°C and 37.0°C, where 42.2°C represents a stressed state for the ancestor and 37.0°C is an unstressed state (Supplementary Table S1). Briefly, a gene was restored if the ancestral expression level was significantly different from itself at 37.0°C and 42.2°C, and the mutant expression level was significantly different and in the opposite direction to that of the ancestral gene expression at 42.2°C. A gene was reinforced if the ancestral expression level while stressed at 42.2°C was significantly different from its unstressed expression level at 37.0°C, and the mutant's expression level of the gene was significantly different and exaggerated in the same direction to that of the ancestral gene expression at 42.2°C. A gene had novel expression if the ancestral expression level was not significantly different from itself at 37.0°C and 42.2°C, but the mutant had significantly differential expression to the ancestral expression level at both temperatures. Finally, a gene was unrestored if the ancestral expression level was significantly different from itself at 37.0°C and 42.2°C, and the mutant did not have a significant difference in expression to the ancestral gene expression level at 42.2°C.

1.4 Results

Mutations associated with rescue events. We ran experiments that started with an overnight culture of the REL1206 ancestor at 37.0°C in low nutrient DM25 media (Figure 1.1A). We then transferred 100 μL of the overnight culture into 44 tubes of fresh DM25 media. These 44 cultures represented distinct populations which we maintained for five days by 1:100 daily serial dilution at 43.0°C in a precisely controlled shaking water bath (Figure 1.1A). This experiment was repeated over seven separate weeks, for a total of 308 (= 7 × 44) experimental populations. Of these, 296 populations were monitored for cell density over a period of five days to determine whether the population went extinct or rebounded to rescue. Altogether, we identified rescue events in six of the seven weeks and 26 populations. Thus, the frequency of rescue was 8.8% of populations (i.e., 26 of 296). Among the rescue events, three were detectable on day three, 12 more on day four, and the rest on day five (Figure 1.1B). Neither the number nor the timing of rescue events were correlated with initial cell densities (r = -0.24, *p* = 0.61; r = 0.50, *p* = 0.32), suggesting that results were not driven by variation in initial conditions across weeks.

To characterize the genomic changes associated with rescue events, we sequenced each of the 26 rescued populations and identified the frequencies of mutations. We focused on mutations that reached near-fixation, which we defined as >85% frequency. The 26 populations contained 1.8 fixed mutations on average, but 11 populations had just one fixed mutation, making these genetic changes the likely drivers of population recovery. One rescue population (#2) evolved a mutator phenotype due to a small deletion in the *mutT* gene and contained six fixed mutations, the most of any population in the experiment (Figure 1.1C). Four populations contained two distinct large duplications based on their sequencing coverage profiles; one duplication included the *groEL* and *groES* genic region, and the second contained a duplication that included the *hslVU* operon.

In total, the 26 populations yielded 29 distinct point or small indel mutations within 20 different genic and intergenic regions (Table 1.1). However, three regions were particularly notable. The first was the *clpA*/*serW* intergenic region, in which the same point mutation appeared across four separate weeks (Table 1.1). Note, however, that this mutation always appeared with other fixed mutations, making it unclear whether it was sufficient to drive rescue. The second was the *rpoBC* operon, where four distinct mutations were identified across five of seven weeks and six of 26 populations. All of these point mutations caused nonsynonymous changes, including one *rpoC* mutation (W1020G) that was fixed across three separate weeks. Two of the *rpoBC* mutations (*rpoB* H447L, *rpoC* W1020G) were the only fixed mutations in at least one population, suggesting that they were sufficient to rescue a population.

Finally, the most mutations were observed in the *hslVU* operon, which encodes a heat shock protease system (Missiakas et al. 1996; Bochtler et al. 2000). In addition to the duplication of this region mentioned previously, for which both copies apparently had a one bp indel frameshift (Table 1.1, population 5-21), the operon had eight distinct point or indel mutations across five of the seven weeks. Another fixed mutation altered the 3' intergenic region of this operon (Table 1.1). Altogether, *hslVU* mutations were found in 62% (16/26) of rescued populations and were the only fixed mutation in at least seven populations. Five of the mutations within *hslVU* caused frameshifts, suggesting that interruption of function was adaptive. Interestingly, the populations with fixed *rpoBC* mutations were distinct from those with *hslVU* mutations; no populations contained fixed mutations in both operons, even though mutations in both operons were identified during weeks 3 and 5 (Table 1.1; Figure 1.1C).

We evaluated four additional features of fixed mutations, focusing on populations that contained potential driver mutations in *rpoBC* and *hslVU*. First, because rescue events occurred at different times during the course of the five-day experiments, we assessed whether the identities of fixed mutations were related to the day of rescue. We found no obvious relationship (Mann-Whitney U test comparing *hslVU* and *rpoBC* populations: p = 0.34). Second, we tested for a relationship between fixed mutations and cell densities after population recovery; *rpoBC* populations had significantly higher cell densities than all other recovered populations on both day four and day five (Mann-Whitney U test: day four p = 0.0044, day five $p = 3.71 \times 10^{-4}$) (Figure 1.1B). Third, to assess whether the fixed mutations arose as a consequence of thermal stress, we also sequenced two control cultures that were maintained at 37.0°C in DM25 (see Methods). Both controls were sequenced to >2,000x, but we found no fixed mutations relative to the REL1206 genome, only three mutations present at >10% frequency, and an average variant frequency of 1.3%. Of the three mutations at >10%, two were not shared with any of the rescue populations. The remaining mutation was a four-nucleotide indel within the *ECB_01992* gene which is likely to be hypermutable because it is part of a motif of seven four-nucleotide repeats (Tenaillon et al. 2016). The four-nucleotide indel was found at frequencies of 19% and 21% in the two ancestral samples and also between 32% and 34% frequency across eight of the 26 rescue populations (Supplementary Table S1.2). The fact that it was not fixed in any population suggests it was likely not adaptive.

Fitness properties of single mutations in *hslVU* and *rpoBC*. Repeated mutations in *rpoBC* and *hslVU* across weeks and populations suggested that specific mutations in these operons drive evolutionary rescue. To verify this conjecture, we isolated seven clones containing single mutations in either the *rpoBC* operon or in the *hslVU* operon (Tables 1.1 and 1.2).

For each of the seven single mutants, we measured their relative fitness (w_r) against the ancestor using competition assays. To measure w_r we competed the ancestor against each of the single mutants at 42.2°C, because the ancestor does not survive at higher temperatures. Six of the seven mutants conferred a significant advantage ($w_r > 1.0$; Figure

1.2A and Table 1.2), and collectively they had a 24% average w_r increase, (one-tailed t-test: $p = 3.44 \times 10^{-12}$). The two *rpoBC* mutants had values 26% and 30% fitness advantages, while the *hslVU* mutants ranged from an estimated 5% to 41% advantage. We note that w_r increases for single mutants were nearly identical to those based on w_r estimates based on competing population samples against the ancestor (Supplementary Table S1.2). On average, the 26 rescued populations had a w_r increase of 24% relative to the ancestor, with an average of 28% (n=7; range 14% to 35%) and 23% (n=16; range 4% to 47%) for populations that contained fixed *rpoBC* and *hslVU* populations, respectively (Figure 1.2B).

In contrast, experiments at the ancestral optimum temperature, 37.0°C, showed that the single mutants had an average w_r disadvantage of 3.5% (one-tailed t-test: $p = 0.0002$; Figure 1.2A). The *rpoB* H447L mutant had a particularly low w_r value of 0.905, and the *hslVU* mutations had w_r values ranging from 0.933 to 0.993 (Table 1.2). These single mutant results were again nearly identical to results based on population samples, because the 26 rescue populations had a fitness decrease of 3% relative to the ancestor, with 14 of 26 populations having relative fitness values significantly < 1.0 (one-tailed t-test: $p < 0.05$, Figure 1.2B). Populations with fixed *rpoBC* mutations had an average w_r decrease of 8% (one-tailed t-test: $p = 0.0082$; $n = 7$; range 0.82 to 0.96), which was significantly lower than all other populations (one-tailed t-test, unequal variance: $p = 0.019$). Although some *hslVU* populations had fitness values significantly < 1.0 at 37.0°C (Supplementary Table S2), *hslVU* populations had an average w_r decrease of 1% relative to the ancestor (n=16; range 0.92 to 1.03), which was not significantly different from 1.0 (one-tailed t-test: $p = 0.054$).

Gene expression differences between rescue mutants and REL1206. To attempt to elucidate the molecular mechanisms that lead to evolutionary rescue, we contrasted gene expression between REL1206 and the seven rescue mutants. We gathered replicated RNAseq data for each mutant at two temperatures: 42.2°C and 43.0°C. We used 42.2°C because it allowed a direct comparison to REL1206 under the same conditions, and 43.0°C because it is the experimental temperature at which rescue occurred. To assess whether the difference between high stress (42.2°C) and rescue (43.0°C) conditions mattered, we performed two analyses. First, we contrasted gene expression between the two temperatures for each mutant. Six of seven mutants had <100 significant differentially expressed genes (DEGs; *padj* < 0.001), and the *hslU* frameshift had 231 DEGs between temperatures (Supplementary Table S1.3). Second, we compared the two temperatures for mutants to REL1206 at 42.2°C. We detected 491 DEGs with the 43.0°C mutant data and 450 DEGs with the 43.0°C data, with 75% of DEGs shared between the two analyses (Supplementary Table S1.4). Overall, these results suggest some temperature-specific differences between high stress and rescue temperatures. However, comparisons using the mutant data at the two temperatures led to identical trends and qualitative conclusions about functional enrichment and directional changes in gene expression. Hence, or simplicity, we focused on comparisons between REL1206 at 42.2°C and the single mutants at which rescue occurred - i.e., 43.0°C.

At 43.0°C, single mutations in *rpoBC* or *hslVU* exhibited from 58 to 250 upregulated DEGs and between 156 to 369 downregulated DEGs relative to REL1206 at 42.2°C. We assessed common sets of DEGs among single mutants within specific operons. For example, the five *hslVU* mutants shared 127 highly downregulated genes (of 294 total) in common, and these exhibited no GO-based enrichment for specific biological processes (Figure 1.3A). Similarly, the two *rpoBC* mutations shared 250 down-regulated genes (Figure 1.3A) that were enriched for transmembrane transporters and catabolic processes (Supplementary Table S1.5). Finally, all seven mutants shared 113 downregulated DEGs (Figure 1.3A), a set that could contain genes critical to rescue. GO analyses of this gene set did not reveal a significant enrichment for any biological processes, leaving it difficult to infer which (if any) of these genes contributed to rescue events.

For upregulated genes in the mutants compared to the REL1206 ancestor, we found 185 DEGs shared by *rpoBC* mutants (of 288 total upregulated genes) and 44 for all five *hslVU* mutants (of 137 total upregulated genes). Interestingly, 40 of the 44 were also upregulated in the *rpoBC* mutants (Figure 1.3B). GO analyses on this set revealed enrichment for flagellum assembly and motility (Supplementary Table S1.6). More specifically, 16 of these 40 genes were annotated to be directly involved in flagellum regulation, assembly, or motility (Liu and Ochman 2007; Kaundal et al. 2020). Through manual investigation of the remaining twenty-four genes, we found eleven genes that were involved in membrane transport and thirteen genes involved in translational processes, amino acid synthesis, and nucleotide synthesis (Supplementary Table S1.7).

Rescue predominantly restores gene expression. An ongoing question about molecular adaptation is whether it restores physiological and molecular processes from a stressed state back toward the unstressed, wild-type state or whether it instead tends to drive the evolution of novelty (Carroll and Marx 2013). Previous studies have suggested the former, because studies have shown that *E. coli* adapts to high temperature stress (42.2°C) by restoring both gene expression (Rodríguez-Verdugo et al. 2016; González-González et al. 2017) and phenotypic characteristics (Hug and Gaut 2015) toward that of the unstressed ancestor.

Following previous studies, we investigated gene expression among mutants, the ancestor at 42.2°C and the ancestor at 37.0°C. Similar to those studies, we found a strong negative correlation between ratios that measure the degree of gene expression change in the mutant relative to the two states of the ancestor. For example, the *rpoB* H447L mutation exhibited a correlation of -0.834 (Figure 1.4A), illustrating that the mutant tended to move gene expression back from the stressed (42.2°C) state toward the wild-type (37.0°C) state. Similar negative correlations were obtained with the other seven mutants (Figure 1.4B-C and Supplementary Figure S1.1), but the negative correlations were generally stronger for the *rpoBC* mutations than the *hslVU* mutations.

We also counted the number of genes that fell into one of four expression categories: restored, unrestored, reinforced, or novel (see Methods). The predominant category was restored, which suggests that the mutant shifted gene expression back toward the unstressed state (Table 1.3; Figure 1.4). Five of seven mutants had >50% of their genes in this category, while the *hslU* frameshift mutant and the *hslU* G60D mutants had 38% and 40% of their genes restored, respectively. The next highest category was unrestored, which represented 23-53% of the genes in the rescue mutants. Perhaps the most striking aspect of our analyses was that each rescue mutant had >100 genes that exhibited novel expression patterns. This category had far fewer genes in previous studies; for example, high temperature adaptive mutations in *rpoB* and *rho* caused <60 genes to have novel expression patterns (Rodríguez-Verdugo et al. 2016; González-González et al.

2017). Given this apparent difference, we compared the average number of novel genes across all seven mutations to previous studies that used the same methods (Rodríguez-Verdugo et al. 2016; González-González et al. 2017). We found that rescue mutations had a significantly different proportion of genes in the four categories to previously studied *rpoB* and *rho* mutations ($p < 2.2 \times 10^{-16}$, contingency test; Figure 1.4D). A total of 43 genes displayed novel expression patterns across all seven rescue mutants; according to GO analyses, these were enriched for transmembrane transporters for carbohydrates, such as glucose and mannose (Supplementary Table S1.8).

1.5 Discussion

We have performed experiments to characterize the genetic mutations that contribute to the rescue of *E. coli* populations from an otherwise lethal temperature of 43.0°C. Overall, we have found that rescue is infrequent, because it occurred for only 8.8% (26 of 296) of our experimental populations. Although not all of our populations were independent (Figure 1A and Methods), the observed rescue frequency is similar to that of Mongold et al. (1999), who found that 10% of *E. coli* populations recovered from a 44.0°C treatment (Mongold et al. 1999). One difference is that they observed rescue only in populations that were pre-adapted to thermal stress, whereas our populations were not pre-adapted. Nonetheless, the two studies are consistent in showing that evolutionary rescue is infrequent but, somewhat paradoxically, frequent enough to be a potent source of evolutionary innovation (Bell 2017).

We have studied thermal stress because it has complex effects on a wide variety of physiological functions, implying that rescue adaptations could be genetically diverse.

Among the 26 populations that exhibited U-shaped rescue dynamics (Figure 1.1B), we have identified 29 distinct fixed mutations (Table 1.1). There were clear patterns among these mutations, because some mutations were found in parallel across presumably independent experiments. For example, the same point mutation in ECB_00530 was found in three separate weeks, as was a nonsynonymous mutation in *rpoC* (Table 1.1). Other common locations of fixed mutations included the *hslVU* heat shock protease operon, the *rpoBC* RNA polymerase operon, and an intergenic region between *clpA* and *serW,* which encode a component of a protease system similar to that encoded by *hslVU* (Kwon et al. 2004) and a serine-bearing tRNA.

Mutations in some of our genes have been identified in previous experiments of *E. coli* temperature adaptation under non-lethal conditions (Tenaillon et al. 2012; Deatherage et al. 2017). For example, mutations in *mrdA* and *rpoBC* have been identified in numerous evolution experiments, both in low-nutrient conditions and under temperature adaptation (Conrad et al. 2010; Tenaillon et al. 2012; Long et al. 2015; Deatherage et al. 2017). Similarly, multiple mutations in *hslVU* were identified in an experiment that evolved REL1206 for 2,000 generations at several temperature regimes, including 37.0°C and 42.0°C (Deatherage et al. 2017). However, the *hslVU* mutations were primarily observed at 37.0°C, not at the stressful temperature, and the 37.0°C mutations did not obviously inhibit function via frameshifts or premature stop codons. Interestingly, mutations in *hslVU* were not found commonly during evolution at 42.2°C by Tenaillon et al. (2012). Of their 115 lines and >1000 mutations, only one line carried a nonsynonymous mutation in *hslU*. These observations suggest that slight differences in conditions (i.e., from stressful to lethal temperature) may have large effects on the set of potentially adaptive mutations.

Single mutations drive rescue. Of our 26 rescued populations, 10 had only a single fixed variant, suggesting they were drivers of evolutionary rescue. To explore the fitness effects and the potential mechanistic basis of these potential drivers, we isolated single mutant genotypes for seven mutations within the *rpoBC* and *hslVU* operons (Table 1.2). We isolated two nonsynonymous mutations in the former, which encodes the β and β' subunits of RNA polymerase (RNAP). Mutations in RNAP must maintain enzyme function due to its central role in transcription, but it is also known that nonsynonymous mutations in RNAP can have numerous effects on cellular properties like fitness, growth rate, and patterns of gene expression (Herring et al. 2006; Rodríguez-Verdugo et al. 2014; Carroll et al. 2015; Rodríguez-Verdugo et al. 2016).

We also isolated three nonsynonymous mutations and two frameshifts within the *hslVU* operon (Table 1.1), which encodes two heat shock proteins that form an ATP-dependent protease complex. In *hslU*, one nonsynonymous mutation (G60D) is in the N-terminal domain that has ATPase activity, the other (L163R) is in the I-intermediate domain that recognizes substrates, and the frameshift is in the C-terminal domain that interacts with the *hslV* protein product (Bochtler et al. 2000; Lien et al. 2009). Previous studies have detailed the effects of nonsynonymous mutations in the N- and C-terminal domains of *hslU* and concluded that most mutations cause the loss of ATP hydrolyzing ability and protease activity (Shin et al. 1996; Bochtler et al. 2000). Similarly, research has shown that nonsynonymous mutations throughout the sequence of *hslV* causes reduced protease activity (Yoo et al. 1997; Yoo et al. 1998). Taken together, these studies suggest that the *hslVU* mutations in our study likely reduce or completely knock out their heat shock protease activity.

We assayed the relative fitness of each of seven mutations to confirm that they can drive rescue dynamics. Six of the seven have w_r values significantly greater than 1.0 at high temperature, with the last borderline significant ($p = 0.073$). The w_r estimates range from a ~6% fitness increase for *hslU* G60D to 30% or higher fitness increase for three of the seven mutations (the *hslU* frameshift, *hslU* L163R, and *rpoB* H447L; Table 1.2 and Figure 1.3). This range of w_r values is not particularly unexpected, even for mutations within the same gene (Barrick et al. 2010; Conrad et al. 2010; LaCroix et al. 2015). For example, Rodríguez-Verdugo et al. (2014) assessed the fitness of four adaptive mutations in two different codons of *rpoB*, and their fitness benefits varied from 17% to 37%. Similarly, González-González et al. (2017) found that adaptive mutations within the *rho* gene varied in fitness increases from 8% to 26%. Together, the w_r estimates of our seven mutations, coupled with the fact that each was the lone fixed mutation in at least one rescue population, clearly establish that each mutation is sufficient for rescue.

Population dynamics of rescue mutations. There were, however, at least two interesting differences in the observed patterns of *hslVU* and *rpoBC* mutations. First, fixed mutations in these two operons were not found together in the same population, which is statistically improbable given their respective frequencies across populations ($p < 0.02$). The lack of co-occurring *rpoBC* and *hslVU* mutations suggests that clonal competition canalizes the initial adaptive response. To the extent that w_r values at 42.2°C reflect fitnesses at 43.0°C, the relative fitness assays suggest that the two *rpoBC* mutations would outcompete at least three of the *hslVU* mutations when both are present (Figure 1.2A). Second, patterns of parallelism differed between *hslVU* and *rpoBC* mutations. Specific *hslVU* mutations were found across different weeks and among non-independent populations within weeks (Figure 1.1C). In contrast, *rpoBC* mutations were identified across weeks but typically in only a single population.

What might drive these apparently different patterns? To address this question, we first recognize that evolutionary rescue can act on standing genetic variation prior to the introduction of the lethal stressor (Bell 2013; Bell 2017). Indeed, our experiments were unlikely to be severely mutation-limited (e.g., Lang et al. 2013) because the experiment for each week began in an overnight DM25 culture at 37.0°C (Figure 1.1A). We nonetheless believe two factors may have contributed to different patterns for *hslVU* and *rpoBC* rescue mutations. First, there is strong constraint on function for *rpoBC* mutations, whereas *hslVU* knockouts are adaptive at lethal temperatures. Hence, we suspect that there are more potential rescue mutations in *hslVU* due to fewer functional constraints. This difference may alone explain why fixed *hslVU* mutations were more numerous than *rpoBC* mutations (Table 1.1), but it does not fully explain why *hslVU* mutations tended to be more common across populations in one week.

Second, we suspect that trade-offs contribute to the pattern across populations in one week. We posit that fitness costs at 37.0°C affect the frequency of individual mutations in the overnight culture, which in turn affects the probability of a mutation being sampled into multiple populations in any given week (Figure 1.1A). We thus assessed w_r for the single mutants at 37.0°C, the temperature of the initial batch cultures, for trade-offs. We found that fitness costs are not a universal feature of the rescue mutations, at least within the power of our experiments to detect such differences. Only two of the seven mutations exhibited significant fitness deficits: *rpoB* H447L and the *hslU* frameshift mutation (Table 1.2). These observations contribute to a growing consensus that trade-offs, and specifically antagonistic pleiotropy, are common but not universal (Cooper and Lenski 2000; MacLean et al. 2004; Rodríguez-Verdugo et al. 2014; Deatherage et al. 2017). The pattern of trade-offs does not fully support our model, because the *hslU* frameshift was common across populations in week 3 despite its low fitness at 37.0°C (Table 1.1). Nonetheless, w_r values based on populations consistently show that the populations with fixed *rpoBC* mutations do exhibit trade-offs (Figure 1.2B). We thus continue to suspect that trade-offs play a large role in the population dynamics of our experiment, because mutations with trade-offs are at low(er) frequency in the 37.0°C overnight culture and thus less likely to be sampled into multiple populations.

We add two additional points about the dynamics of rescue in our experiment. First, the fact that we uncover clear and repeatable patterns of mutations in only a small subset of genes and operons suggests that the universe of potential rescue mutations is small, especially given that the experiment should not have been severely mutation-limited. One potential explanation for parallel mutations across weeks - such as the *clpA/serW* mutation and several others (Table 1.1) - is that the experiment selected for low-frequency mutations that were present in the common, frozen ancestral stock (Figure 1.1A). However, this possibility does not contradict (but rather reinforces) the conjecture that there are only a few major mutational targets for adaptive rescue. In this context, it is interesting to muse whether there is in fact a very small universe of mutations that are capable of rescue or whether there is a large universe of such mutations but most do not establish in our populations because they have severe fitness costs at 37.0°C. We cannot yet distinguish between these two alternatives, but knowing the prevalence of trade-offs is important to furthering our understanding about the dynamics of evolutionary rescue. Second, we need to mention an important distinction between our *E. coli* experiment and evolutionary rescue in natural populations, particularly in size-limited, non-bacterial populations subjected to stressors like climate change. The yeast experiment under lethal salt conditions is illustrative, because it showed that rescue occurs less frequently in populations of small size (Bell and Gonzalez 2009). It thus seems likely that the frequency of rescue in most plant and animal populations, which tend to be relatively small, will be far less than the 8% to 10% estimated for *E. coli* under lethal temperature stress.

Insights into rescue mechanisms and evolutionary direction. We have established that the seven mutations are capable of rescue. However, an overarching question is about their function—i.e., what do they do and how do they drive rescue events? To begin to address this question, we generated gene expression data for the seven single rescue mutants. For each clone we measured gene expression at the exponential phase of growth at two temperatures (42.2°C and 43.0°C) to compare to the REL1206 ancestor at 42.2°C and 37.0°C. Our goals were: *i*) to find sets of DEGs in common across the entire set of mutants, in the hope that they yield clues to mechanisms and *ii*) to characterize the overall direction of gene expression changes with respect to the stressed and unstressed state of the ancestor.

Our first goal was formulated under the hypothesis that rescue mutants may affect common pathways that lead to rescue. We proffer this hypothesis knowing that the differences in w_r among mutants reflects the fact that many DEGs vary among them and also that different mutants may have utilized different pathways to achieve rescue. Nonetheless, we first focused on shared DEGs between the two *rpoBC* mutants. They share a set of 435 DEGs relative to the ancestor (Figure 1.3), based on a conservative measure of differential expression that includes both adjusted *p*-values < 0.001 and two-fold differences in $\log_2$ expression (see Methods). The high number of DEGs is not surprising, because previous work has shown that single mutations in RNAP can alter the gene expression of ~1,000 or more genes (Conrad et al. 2010; Carroll et al. 2015; Rodríguez-Verdugo et al. 2016). The number of common DEGs was much lower for *hslVU* mutations, at 171 total; the union between the two sets yielded 40 up- and 113 down-regulated genes (Figure 3).

The set of 113 down-regulated genes provided no clear patterns with regard to function, based on GO analyses and manual investigation of gene annotations. However, the 40 upregulated genes yielded two notable observations. First, 24 of the 40 were involved in transport, ribosomal assembly, and amino acid and nucleotide pathways. This suggests that changes in gene expression were supporting translation processes, perhaps to enhance efficiency at high temperature. Surprisingly, GO analyses revealed that this 40 gene set is also enriched for genes involved in flagellar assembly and motility, a process that is known to be energetically costly (Soutourina and Bertin 2003). We manually verified that at least 16 of the 40 genes have functions associated with flagella. Given this observation, one must ask about the potential adaptive benefit of flagellar production and activity. It is hard to answer this question directly, but it has been shown that *E. coli* produce flagella in low nutrient environments (Shi et al. 1992; Sim et al. 2017). Hence, one hypothesis is that increased expression of flagella leads to enhanced motility and nutrient acquisition. It seems doubtful, however, that enhanced motility is an advantage in our well-mixed system. Interestingly, previous work has shown that first-step adaptive mutations also increased flagellar gene expression, only to be attenuated by subsequent compensatory mutations (Rodríguez-Verdugo et al. 2016). These observations suggest that flagellar production may be a disadvantageous by-product of other major and adaptive shifts in physiological processes. Altogether, then, it is not clear which—if any—of the common DEGs contribute to the rescue phenotype. We suspect, however, that shifts in the expression of translation-related genes are more critical for adaptation than the upregulation of flagella-related genes. In the future, proteomic analyses may provide further insights into changes introduced by rescue mutations and evolutionary mechanisms.

It is worth briefly considering the potential effects of *hslVU* mutations separately, because the frameshift mutations lead to the somewhat paradoxical conclusion that knockouts of heat shock-related proteins are beneficial at lethal temperatures. In this context, it is helpful to know that the *hslVU* protein degrades the σ^{32} factor. In *E. coli*, σ^{32} typically exists as an RNA with secondary structure that unfolds under high temperature to allow translation, and then the σ^{32} protein regulates the transcription of genes needed to carry out the heat shock response (Roncarati and Scarlato 2017). By degrading σ^{32}, the *hslVU* protein indirectly inhibits the production of other proteins in the heat shock cascade (Kanemori et al. 1997). We propose that the *hslVU* rescue mutations decrease or knock out the protease function, thereby facilitating an uninhibited heat shock response through σ^{32}. One interesting fact is that σ^{32} regulates a series of genes—like *dnaK*, *dnaJ*, and *grpE* (Nonaka et al. 2006)—that are also required for the control of flagellar synthesis through another sigma factor (σ^{28}) (Shi et al. 1992), suggesting a mechanistic link between *hslVU* and flagellar genes. These observations support our hypothesis that loss of *hslVU* activity enhances some aspects of the heat shock response and strengthens the possibility that the upregulation of flagellar genes is not a direct feature of adaptation.

In a second goal, we used expression data to characterize the directional change for all genes (Carroll and Marx 2013). The question is whether our rescue mutations produce patterns similar to previously studied mutations that contributed to temperature adaptation (Rodríguez-Verdugo et al. 2016; González-González et al. 2017). The previous mutations predominantly moved gene expression from a stressed physiological condition toward the unstressed, wild-type condition. Our rescue mutations are similar, because they also predominantly shift expression toward restoring the unstressed state (Figure 1.4A-C). However, there is an important difference: the rescue mutations yielded significantly more genes with novel gene expression patterns (Figure 1.4D), particularly for transmembrane transporters. This result is not dependent on the 43.0°C conditions, because it is also evident at the slightly modified temperature of 42.2°C, under conditions identical to the previous experiments (Supplementary Figures S1.2-3 and Supplementary Table S1.9 for 42.2°C results).

These data suggest that there could be a qualitative difference between rescue mutations and mutations that contribute to adaptation under non-lethal conditions. This conclusion is of course subject to caveats. For example, although we followed procedures identical to previous studies, the differences may nonetheless reflect experiment-specific effects, such as batch effects in RNAseq data (although we explicitly examined such effects; see Materials and Methods), rather than differences in the dynamics of adaptation. We also assayed only a single point in the growth phase, which may not represent the crucial point in the cell cycle for rescue adaptations. As a consequence, the novel expression patterns we have observed may represent noise rather than changes that contribute to (or are necessary for) adaptation. However, it is possible that phenotypic novelty is an important feature of evolutionary rescue; that is, when challenged with a lethal stressor, it is not sufficient to move toward restoring expression toward the unstressed state. Ours is only a first observation, but it opens an interesting question for future research: do mutations that drive evolutionary rescue differ qualitatively from adaptations in non-lethal environments?

Figures

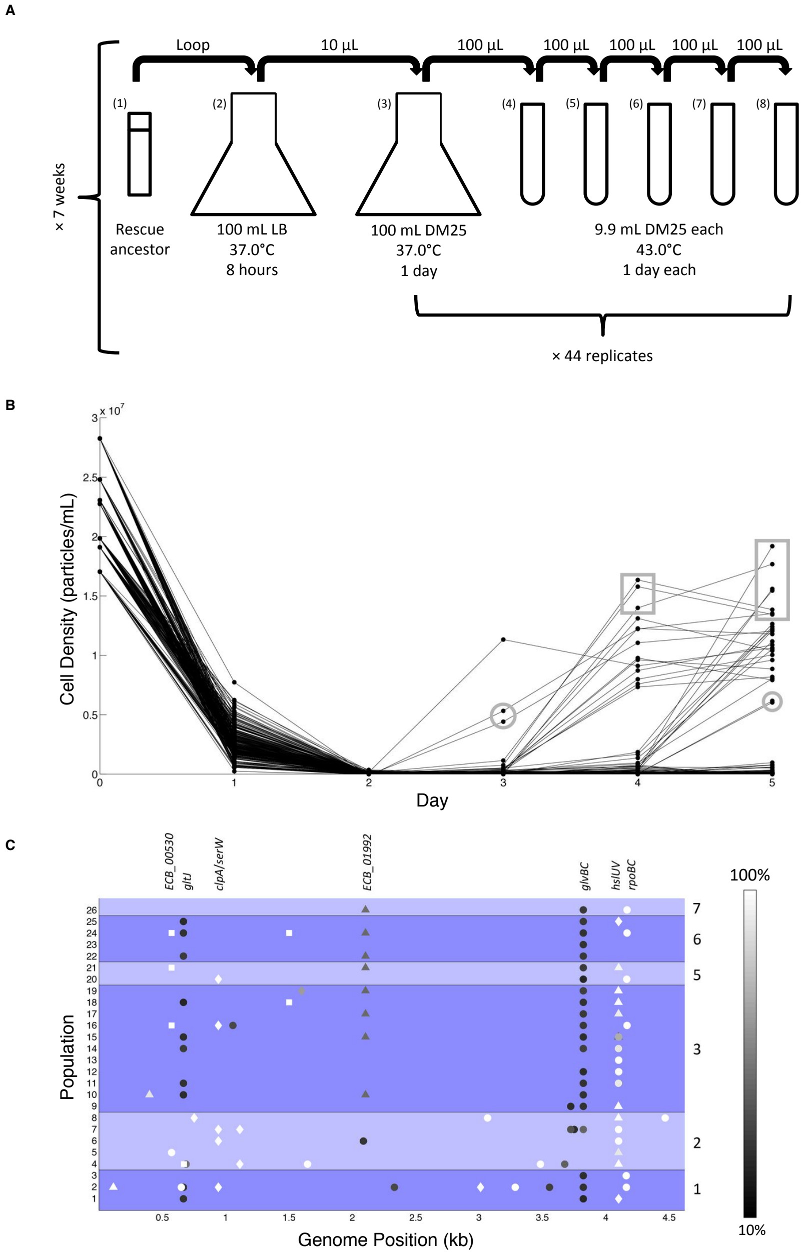

Figure 1.1: Evolutionary rescue experimental design and dynamics. A) Experimental design for producing and observing rescue events. Bacteria were propagated from frozen through two flasks to acclimate them and to produce enough cells for experimental replication. Samples of flask culture were transferred to 44 replicates that were propagated through 1:100 serial dilution for five days. This procedure was repeated across seven different weeks. B) Population cell densities over time. Most populations went extinct over the course of five days. A total of 26 rescue events were observed across the third, fourth, and fifth days of growth. The timing of rescue events was determined by the day at which cell density increased by an order of magnitude over the previous day. Populations

possessing *rpoBC* mutations are indicated by rectangles. The three populations possessing duplications are circled. C) Genome-wide distribution of mutations in rescue populations. Populations 1 to 26 are labelled on the left and different weeks are separated into groups and labeled at the right. Mutations are colored by their frequency in the population according to the scale at the right. Synonymous, nonsynonymous, indel, and intergenic mutations are represented by squares, circles, triangles, and diamonds, respectively. Only mutations at frequencies >10% are shown. Mutations occurring in more than two populations are labeled at the top.

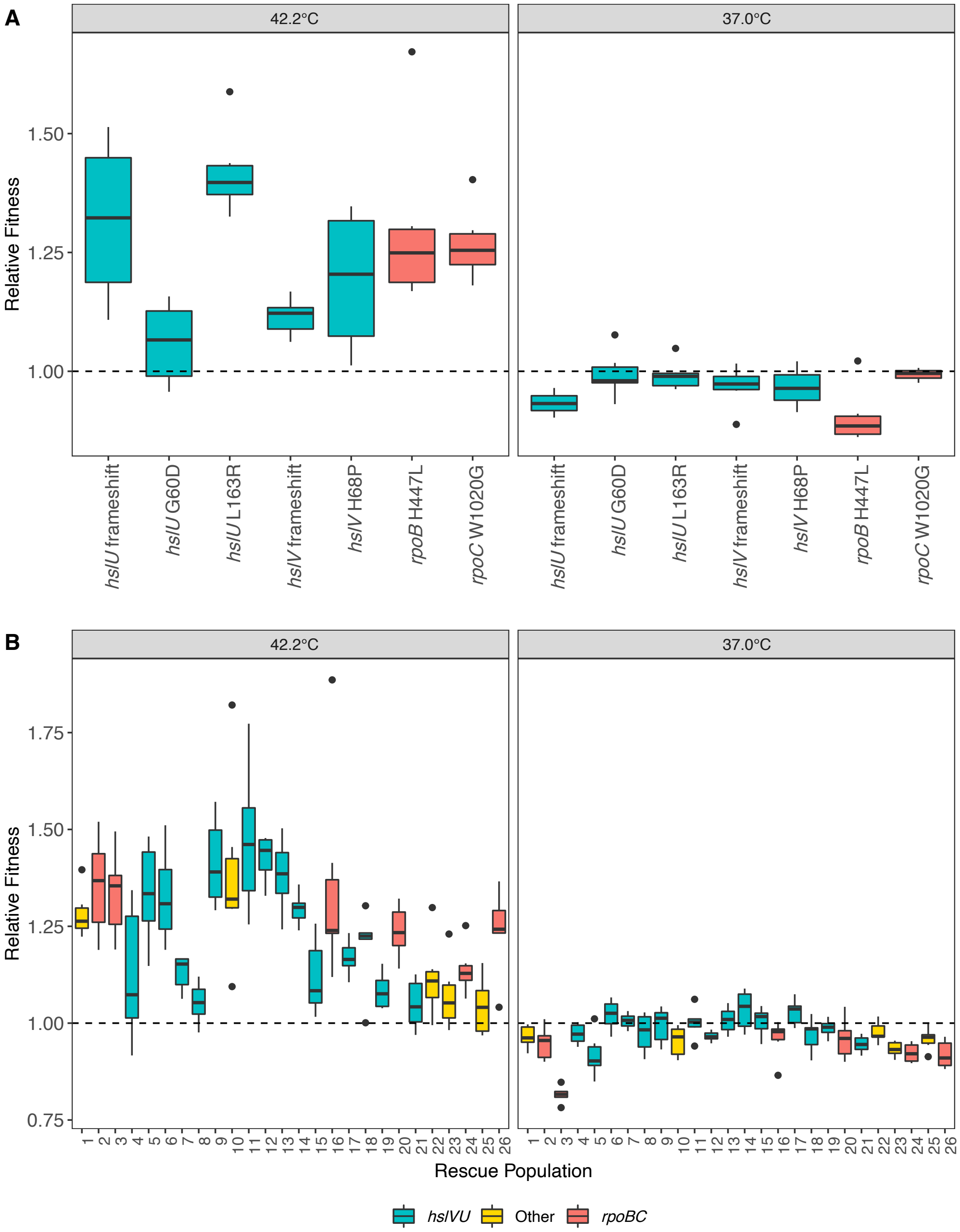

Figure 1.2: Relative fitness of the single mutants and populations at 42.2°C and 37.0°C. A) Relative fitness of the single mutants in competition with the ancestor at 42.2°C and 37.0°C. B) Relative fitness of the 26 rescue populations in competition with the ancestor at 42.2°C and 37.0°C. Boxplots represent the relative fitness (w_r) values of all replicates for each single mutant or population in competition with the ancestor. A w_r value near or at 1.0 indicates similar fitness to that of the ancestor; values > 1.0 indicate higher fitness than the ancestor, and values < 1.0 indicate lower fitness than the ancestor.

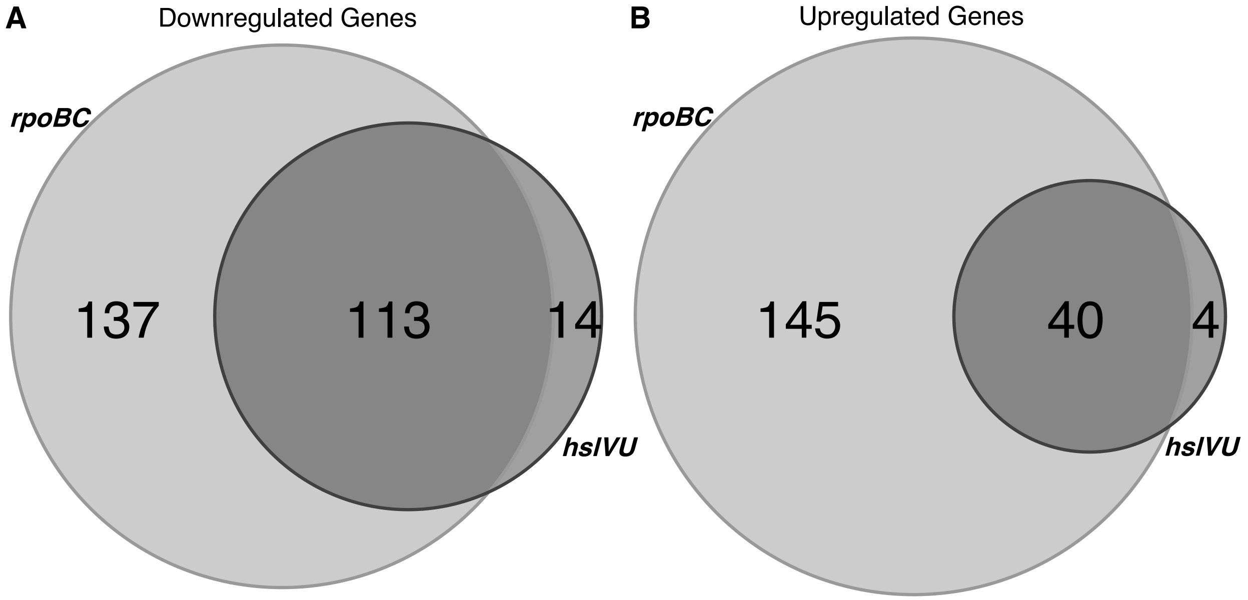

Figure 1.3: The number of upregulated and downregulate genes in the single mutants grown at 43.0°C relative to the ancestor grown at 42.2°C. For the Venn diagrams in both A and B, the *rpoBC* circle represents the number of differentially expressed genes shared between the two *rpoBC* mutants, and the *hslVU* circle represents genes shared among the five *hslVU* mutants. A) The number of downregulated genes relative to the ancestor. B) The number of upregulated genes relative to the ancestor.

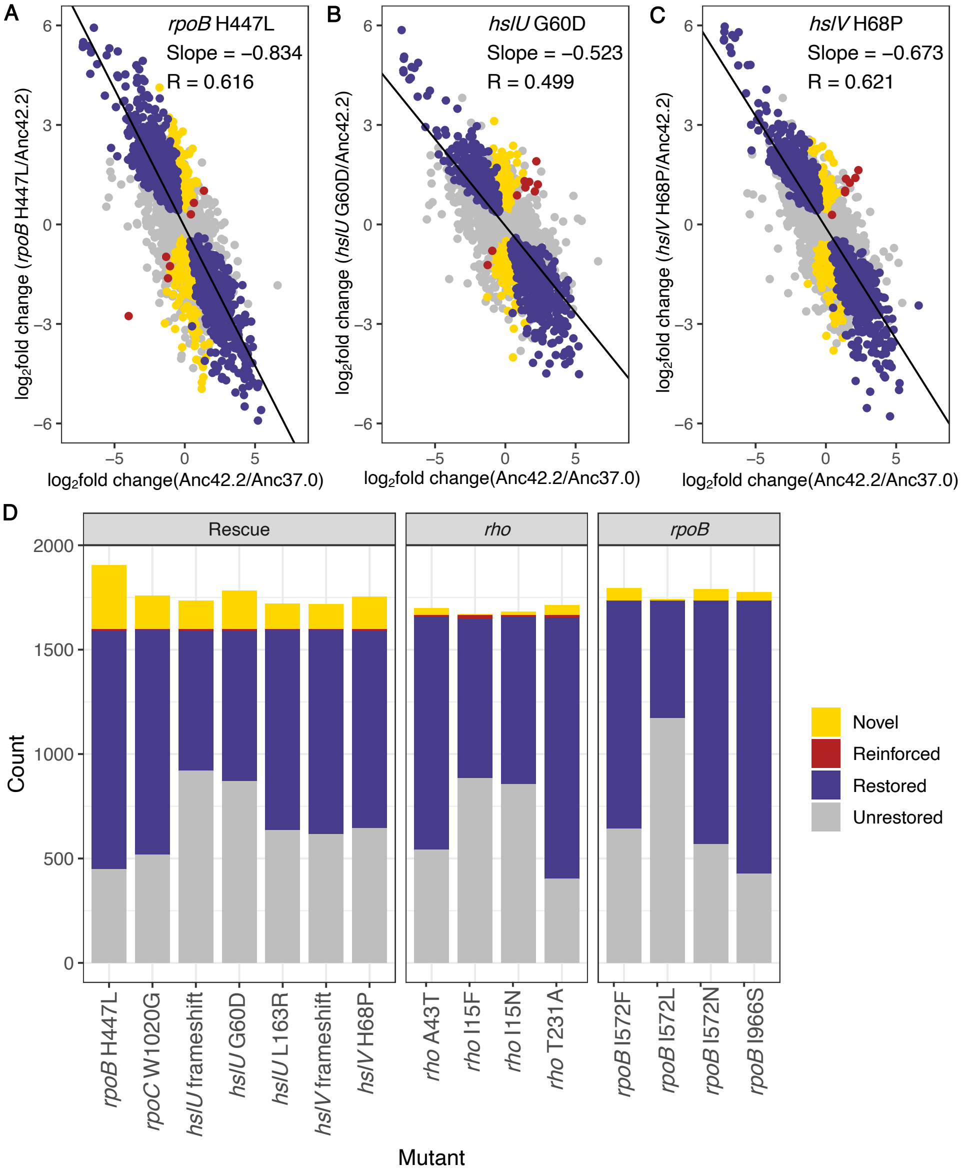

Figure 1.4: Direction of gene expression change in single rescue mutants. A-C) The *y*-axis represents the direction of gene expression change of the three single rescue mutants, (A) *rpoB* H447L, (B) *hslU* G60D, and (C) *hslV* H68P, when grown at 43.0°C compared to the ancestor grown at 42.2°C. These three mutations were chosen as illustrative, with the remaining four mutations shown in Supplementary Figure 2. The x-axis represents the ancestral changes in gene expression when it is grown at 42.2°C compared to when it is grown at 37.0°C. The black line represents the linear regression fitted to the data in each graph. D) Comparison of the number of genes in each category of gene expression change for the rescue mutants and for the *rpoB* and *rho* mutants studied in Rodríguez-Verdugo et al. 2016 and in González-González et al. 2017. The *rpoB* and *rho* mutations were adaptive to high but non-lethal temperatures.

Tables

Table 1.1: Fixed mutations in rescued populations

Gene(s)[1]	**Week-Population[2]**	**Position[3]**	**Mutation[4]**	**Mutation type[5]**
arcB	1-2	3285997	T-->G	E755A (GAA-->GCA)
clpA/serW	1-2, 2-6, 2-7, 3-26, 3-16, 5-20	942604	A-->G	intergenic (+359/+339)
ECB_00530	3-16, 5-21, 6-24	573246	G-->C	L81L (CTC-->CTG)
ECB_00530	2-5	573229	T-->G	N87T (AAC-->ACC)
ECB_02812/ ECB_02813	1-2	3012076	T-->G	intergenic (-516/-58)
hslU	3-9, 3-18, 3-19	4100115	+C	coding (1116/1332 nt)
hslU	2-5, 2-8	4100512	Δ1 bp	coding (719/1332 nt)
hslU	3-11	4101052	C-->T	G60D (GGT-->GAT)
hslU	2-6, 2-7	4101159	C-->A	K24N (AAG-->AAT)
hslU	3-14	4100743	A-->C	L163R (CTG-->CGG)
hslV	2-4, 5-21	4101760	+T	coding (11/531 nt)
hslV	3-17	4101363	Δ2 bp	coding (408/531 nt and 409/531nt))
hslV	3-12, 3-13	4101568	T-->G	H68P (CAT-->CCT)
hslV/ftsN	1-1, 6-25	4101844	Δ1 bp	intergenic (-74/+19)
insE-1/serX	2-4, 2-7	1111967	A-->G	intergenic (-236/+166)
lnt	2-4	671609	A-->C	G456G (GGT-->GGG)
mrdA	1-2	649901	T-->G	I301L (ATC-->CTC)

mutT	1-2	114029	Δ1 bp	coding (182/390 nt)
pepA	2-8	4468692	T-->G	T163P (ACC-->CCC)
rhsE	3-18, 6-24	1500351	T-->G	G25G (GGT-->GGG)
rpoB	1-3	4162195	A-->T	H447L (CAC-->CTC)
rpoB	1-2	4163133	A-->C	N760H (AAC-->CAC)
rpoC	5-20	4165883	A-->G	D308G (GAT-->GGT)
rpoC	3-16, 6-24, 7-26	4168018	T-->G	W1020G (TGG-->GGG)
rtcA	2-4	3484800	G-->A	S217F (TCC-->TTC)
secF	3-10	398683	+GGT	coding (756/972 nt)
ybgG/cydA	2-8	751779	T-->G	intergenic (+286/-561)
yghS	2-8	3067389	T-->G	H219P (CAC-->CCC)
ynfL	2-4	1646703	T-->A	I29F (ATT-->TTT)

[1] Gene names as defined in the REL606 annotation. Bolded names represent single mutations isolated for further study (see Table 2).
[2] Provides information about the week and population in which the mutation was fixed. This column shows, for example, that the same *clpA/serW* intergenic point mutation was present across four independent weeks and six total populations.
[3] Location of mutations in the REL606 reference.
[4] Provides mutation type from REL1206 for point mutations. + and Δ represents an insertion and deletion relative to REL1206.
[5] Provides information about codon change for nonsynonymous mutations. Coding variants with + and Δ represent frameshifts.

Table 1.2: Seven single mutants and their relative fitness values

	Relative Fitness 42.2°C		Relative Fitness 37.0°C	
Mutant	**w_r**	**w_r *p*-value**	**w_r**	**w_r *p*-value**
hslU frameshift	1.317	0.003	0.933	0.033
hslU G60D	1.060	0.073	0.993	0.378
hslU L163R	1.419	4.88E-05	0.992	0.272
hslV frameshift	1.115	0.0004	0.967	0.062
hslV H68P	1.192	0.011	0.966	0.195
rpoB H447L	1.303	0.0054	0.905	0.006
rpoC W1020G	1.268	0.0002	0.993	0.257

Table 1.3: Number of genes in each category of gene expression change

	Restored	**Reinforced**	**Unrestored**	**Novel**	**Total**
hslU frameshift	669	8	923	135	1735
hslU G60D	719	9	872	183	1783
hslU L163R	960	4	636	121	1721
hslV H68P	947	7	646	154	1754
hslV frameshift	977	4	619	118	1718
rpoB H447L	1144	7	449	305	1905
rpoC W1020G	1077	3	520	159	1759

Supplemental Information

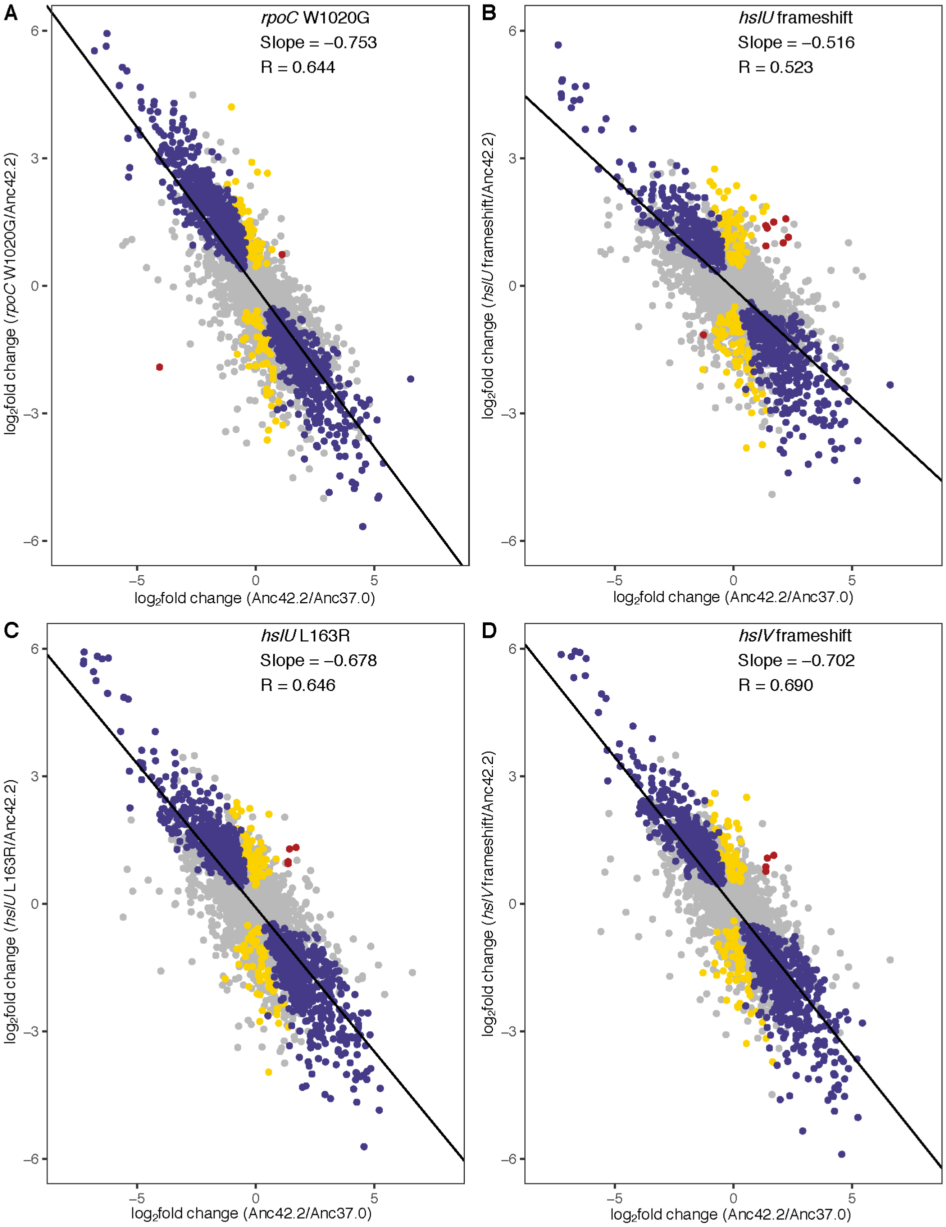

Figure S1.1: Direction of gene expression change at 43.0°C. The x-axis represents the ancestral changes in gene expression when it is grown at 37.0°C compared to when it is grown at 42.2°C. The y-axis represents the direction of gene expression change of 4 single rescue mutants, (A) *rpoC* W1020G, (B) *hslU* frameshift, (C) *hslU* L163R, and (D) *hslV* frameshift, when grown at 43.0°C compared to the ancestor grown at 42.2°C. The black line represents the linear regression fitted to the data in each graph.

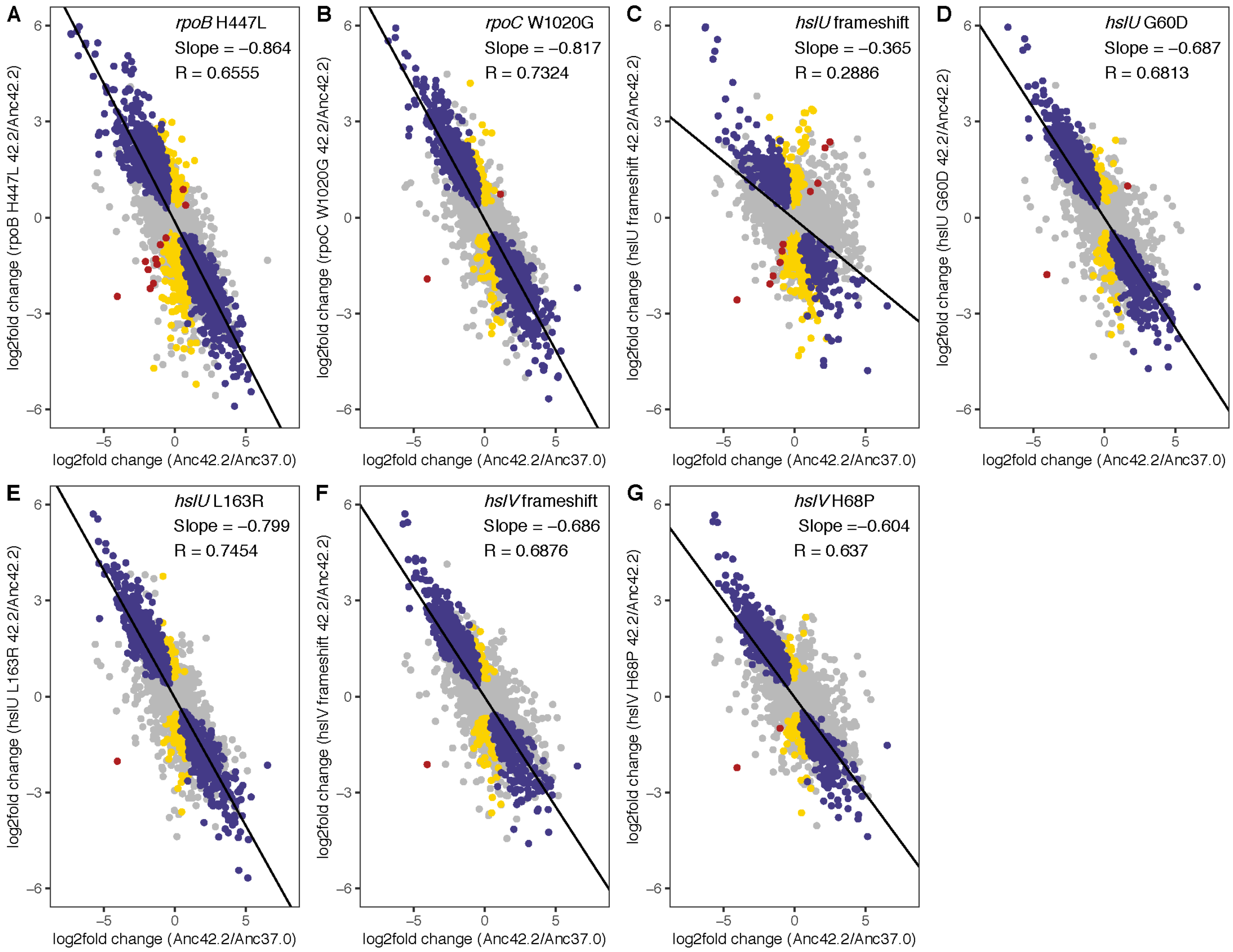

Figure S1.2: Direction of gene expression change at 42.2°C. The x-axis represents the ancestral changes in gene expression when it is grown at 37.0°C compared to when it is grown at 42.2°C. The y-axis represents the direction of gene expression change of 7 single rescue mutants, (A) *rpoB* H447L, (B) *rpoC* W1020G, (C) *hslU* frameshift, (D) *hslU* G60D, (E) *hslU* L163R, (F) *hslV* frameshift, and (G) *hslV* H68P, when grown at 42.2°C compared to the ancestor grown at 42.2°C. The black line represents the linear regression fitted to the data in each graph.

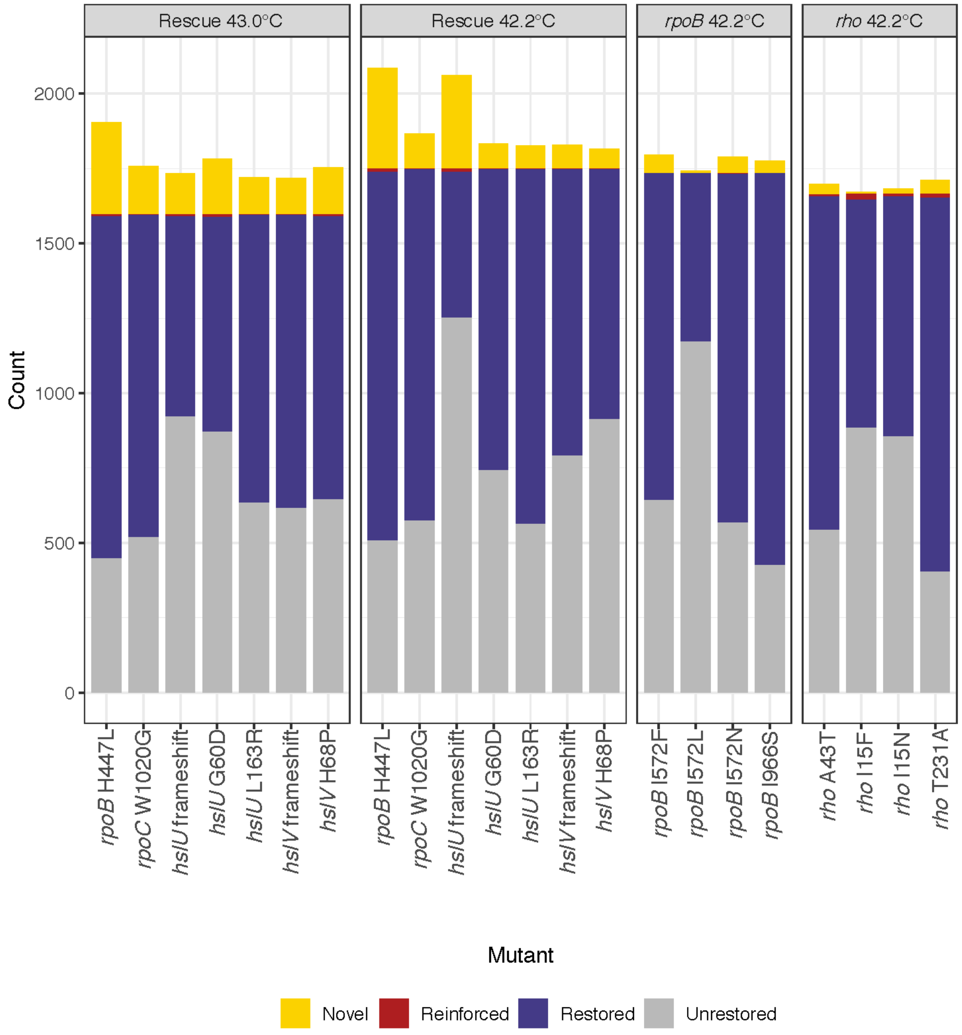

Figure S1.3: Direction of gene expression change. Comparison of the number of genes in each category of gene expression change for the rescue mutants grown at 43.0°C, 42.2°C, and for the *rpoB* and *rho* mutants studied in Rodríguez-Verdugo et al. 2016 and in González-González et al. 2017. The *rpoB* and *rho* mutations were adaptive to high but non-lethal temperatures.

Supplemental Tables

Table S1.1: Change in gene expression direction classification

Category	**Anc42.2 vs Anc37.0**	**Mut43.0/42.2 vs Anc42.2**	**Mut43.0/42.2 vs Anc37.0**
Restored	Significant upregulation Significant downregulation	Significant downregulation Significant upregulation	- -
Reinforced	Significant upregulation Significant downregulation	Significant upregulation Significant downregulation	Significant upregulation Significant downregulation
Unrestored	Significant upregulation Significant downregulation	Non-significant Non-significant	- -
Novel	Non-significant Non-significant	Significant upregulation Significant downregulation	Significant upregulation Significant downregulation

Significant: significantly differentially expressed gene (*padj* < 0.001)
Non-significant: not significantly differentially expressed gene (*padj* > 0.001)
Adapted from Rodríguez-Vergudo *et al.* 2016

Table S1.2: Mutations present in populations at frequencies >10% and mean fitness values ($\overline{w}$) of populations relative to their ancestor at 42.2°C and 37.0°C

Week	Pop.	Affected Region (Frequency in Population)	$\overline{w}_{42.2}$ (Standard Deviation)	*p*-value[a]	$\overline{w}_{37.0}$ (Standard Deviation)	*p*-value[a]
1	1	*hslV/ftsN* (1), *gltJ* (0.136), *glvBC* (0.108)	1.28 (0.06)	5.24E-05	0.96 (0.03)	1.58E-02
1	2	*mutT* (1), *mrdA* (1), *clpA/serW* (1), *ECB_02812/ECB_02813* (1), *arcB* (1), *rpoB* (1), *yfbM* (0.205), *yhhI* (0.178), *gltJ* (0.146), *glvBC* (0.136)	1.35 (0.13)	5.20E-04	0.95 (0.04)	1.56E-02
1	3*	*rpoB* (1), *glvBC* (0.126)	1.33 (0.11)	3.71E-04	0.82 (0.02)	2.29E-06
2	4	*lnt* (1), *insE1/serX* (1), *ynfL* (1), *rtcA* (1), *hslV* (1), *nagE* (0.352), *yiaN* (0.292)	1.12 (0.18)	7.35E-02	0.97 (0.03)	2.46E-02
2	5	*ECB_00530* (1), *hslU* (0.877)	1.34 (0.13)	6.97E-04	0.92 (0.06)	7.62E-03
2	6	*clpA/serW* (1), *hslU* (1), *yegM* (0.154)	1.33 (0.12)	6.06E-04	1.02 (0.04)	1.13E-01
2	7	*clpA/serW* (1), *insE1/serX* (1), *hslU* (1), *glvBC* (0.309), *gpsA* (0.237), *dfp* (0.121)	1.13 (0.05)	4.52E-04	1.01 (0.02)	2.52E-01
2	8	*ybgG/cydA* (1), *yghS* (1), *hslU* (1), *pepA*	1.05 (0.05)	2.96E-02	0.98 (0.05)	1.41E-01

		(1)				
3	9*	*hslU* (0.986), *glvBC* (0.138), *gpsA* (0.117)	1.41 (0.12)	1.62E-04	1 (0.05)	4.09E-01
3	10	*secF* (0.872), *ECB_01992* (0.332), *glvBC* (0.137), *gltJ* (0.116)	1.38 (0.24)	5.91E-03	0.95 (0.04)	1.78E-02
3	11*	*hslU* (0.893), *glvBC* (0.144), *gltJ* (0.136), *hslV* (0.106)	1.47 (0.19)	8.52E-04	1 (0.04)	4.69E-01
3	12	*hslV* (1), *glvBC* (0.122)	1.43 (0.06)	5.56E-06	0.97 (0.01)	6.53E-04
3	13*	*hslV* (1)	1.38 (0.09)	8.31E-05	1.01 (0.03)	2.93E-01
3	14*	*hslU* (0.874), *glvBC* (0.178), *gltJ* (0.165), *hslU* (0.111)	1.30 (0.04)	5.46E-06	1.03 (0.05)	7.82E-02
3	15	*hslV* (0.611), *ECB_01992* (0.326), *hslU* (0.286), *gltJ* (0.133), *glvBC* (0.133)	1.12 (0.10)	1.57E-02	1 (0.04)	3.84E-01
3	16	*ECB_00530* (1), *clpA/serW* (1), *rpoC* (1), *appB* (0.212), *glvBC* (0.157)	1.35 (0.28)	1.29E-02	0.96 (0.05)	5.28E-02
3	17*	*hslV* (0.909), *ECB_01992* (0.329), *glvBC* (0.164)	1.17 (0.04)	1.20E-04	1.03 (0.03)	4.62E-02
3	18	*rhsE* (1), *hslU* (0.987), *glvBC* (0.135), *gltJ* (0.109)	1.20 (0.10)	2.49E-03	0.97 (0.04)	7.41E-02
3	19	*hslU* (0.989), *ydfJ/ydfK* (0.529),	1.08 (0.05)	3.85E-03	0.99 (0.02)	1.05E-01

		ECB_01992 (0.318), *glvBC* (0.182)				
5	20	*clpA/serW* (1), *rpoC* (1), *glvBC* (0.113)	1.24 (0.07)	1.79E-04	0.96 (0.05)	5.90E-02
5	21[b]	*ECB_00530* (1), *hslV* (0.956), *ECB_01992* (0.33), *glvBC* (0.171)	1.05 (0.06)	6.14E-02	0.95 (0.02)	8.25E-04
6	22[c]	*ECB_01992* (0.316), *gltJ* (0.173), *glvBC* (0.159)	1.12 (0.10)	1.90E-02	0.98 (0.03)	4.40E-02
6	23[c]	*glvBC* (0.137)	1.07 (0.09)	5.33E-02	0.93 (0.02)	1.89E-04
6	24	*ECB_00530* (1), *rhsE* (1), *rpoC* (1), *ECB_01992* (0.33), *glvBC* (0.144), *gltJ* (0.143)	1.14 (0.06)	1.51E-03	0.92 (0.02)	3.05E-04
6	25[c]	*hslV/ftsN* (1), *glvBC* (0.156), *gltJ* (0.121)	1.04 (0.07)	1.06E-01	0.96 (0.03)	1.04E-02
7	26*	*rpoC* (1), *ECB_01992* (0.341), *glvBC* (0.155)	1.24 (0.11)	1.55E-03	0.92 (0.04)	1.36E-03

[a] Significance of relative fitness increases compared to a value of 1.0 was determined using one-tailed t-tests and six replicate fitness estimates per population.
[b] Population also contained a duplication of ~80 kb.
[c] Population also contained a duplication of ~20 kb.
*Population used for isolating single mutants.

Table S1.3: Number of differentially expressed genes (DEGs) between the single mutants grown at 43.0°C relative to 42.2°C

Comparison of Single Mutants at 43.0˚C relative to 42.2˚C			
Rescue Mutant	**DEGs, *padj*<0.001**	**Highly upregulated genes (*padj*<0.001, log$_2$-fold change > 2)**	**Highly downregulated genes (*padj*<0.001, log$_2$-fold change < -2)**
hslU G60D	77	0	6
hslU indel	231	3	83
hslU L163R	0	0	0
hslV H68P	45	10	1
hslV indel	1	0	0
rpoB H447L	3	0	0
rpoC W1020G	20	0	0

Table S1.4: Comparison of the number and overlap of DEGs between the rescue mutants grown at 42.2°C relative to the ancestor at 42.2°C and the rescue mutants grown at 43°C relative to the ancestor at 42.2°C

	Rescue Mutant at 43˚C relative to Ancestor at 42.2˚C		**Rescue Mutant at 42.2˚C relative to Ancestor at 42.2˚C**		**Comparing gene sets**	
Rescue Mutant	**Highly upregulated genes: *padj*<0.001, $\log_2$-fold change > 2**	**Highly downregulated genes: *padj*<0.001, $\log_2$-fold change < -2**	**Highly upregulated genes: *padj*<0.001, $\log_2$-fold change > 2**	**Highly downregulated genes: *padj*<0.001, $\log_2$-fold change < -2**	**Shared upregula-ted genes**	**Shared downregul-ated genes**
hslU G60D	65	167	142	132	50	78
hslU indel	58	156	62	71	26	38
hslU L163R	104	231	194	189	91	140
hslV H68P	101	241	115	71	76	63
hslV indel	123	222	143	111	110	78
rpoB H447L	250	369	246	370	199	286
rpoC W1020G	223	268	220	230	169	171

Table S1.5: GO analysis on the set of highly downregulated genes shared by *rpoB* H447L and *rpoC* W1020G single mutants

GO biological process complete	**Reflist (4391)**	**Observed**	**Expected**	**Over/under**	**Fold Enrichment**	**p-value**
monosaccharide transmembrane transport (GO:0015749)	43	11	2.14	+	5.13	3.78E-02
alpha-amino acid catabolic process (GO:1901606)	60	13	2.99	+	4.34	3.34E-02
carbohydrate transport (GO:0008643)	129	27	6.43	+	4.2	2.71E-06
cellular amino acid catabolic process (GO:0009063)	72	15	3.59	+	4.18	1.22E-02
carbohydrate transmembrane transport (GO:0034219)	87	17	4.34	+	3.92	6.59E-03
small molecule catabolic process (GO:0044282)	290	47	14.46	+	3.25	4.89E-09
carbohydrate catabolic process (GO:0016052)	137	22	6.83	+	3.22	4.86E-03
carboxylic acid catabolic process (GO:0046395)	185	25	9.23	+	2.71	1.64E-02
organic acid catabolic process (GO:0016054)	193	26	9.63	+	2.7	1.15E-02
organic substance catabolic process (GO:1901575)	482	53	24.04	+	2.2	1.13E-04
catabolic process (GO:0009056)	498	54	24.84	+	2.17	9.93E-05
cellular catabolic process (GO:0044248)	380	40	18.95	+	2.11	1.89E-02
Unclassified (UNCLASSIFIED)	939	36	46.83	-	0.77	0.00E+00

cellular nitrogen compound metabolic process (GO:0034641)	1032	26	51.47	-	0.51	3.69E-02
macromolecule metabolic process (GO:0043170)	1079	25	53.81	-	0.46	2.38E-03
cellular macromolecule metabolic process (GO:0044260)	900	19	44.89	-	0.42	5.85E-03
cellular biosynthetic process (GO:0044249)	955	19	47.63	-	0.4	7.21E-04

Table S1.6: GO analysis on the set of highly upregulated genes shared by all rescue mutants

GO biological process complete	Reflist	Observed	Expected	Over/under	Fold Enrichment	p-value
bacterial-type flagellum-dependent swarming motility (GO:0071978)	14	6	0.13	+	47.05	1.44E-05
bacterial-type flagellum organization (GO:0044781)	26	8	0.24	+	33.78	3.59E-07
bacterial-type flagellum assembly (GO:0044780)	17	5	0.15	+	32.29	1.21E-03
bacterial-type flagellum-dependent cell motility (GO:0071973)	40	10	0.36	+	27.44	9.51E-09
archaeal or bacterial-type flagellum-dependent cell motility (GO:0097588)	44	10	0.4	+	24.95	2.16E-08
cilium or flagellum-dependent cell motility (GO:0001539)	44	10	0.4	+	24.95	2.16E-08
localization of cell (GO:0051674)	50	11	0.46	+	24.15	2.49E-09
cell motility (GO:0048870)	50	11	0.46	+	24.15	2.49E-09

Table S1.7: List of highly upregulated genes and their function

Gene	Function/process	Reference
asnA	Aspartate--ammonia ligase, asparagine synthetase	(Nakamura et al., 1981)
bipA	GTPase, has chaperone like activity and assists ribosome assembly	(Choi & Hwang, 2018)
deaD	ATP-dependent RNA helicase, involved in ribosome assembly	(Toone et al., 1991)
fecC	Fe(3+) dicitrate transport system permease protein	(Pressler et al., 1988)
fecD	Membrane bound protein for iron transport	(Pressler et al., 1988)
flgA	Flagella	(Liu & Ochman, 2007)
flgB	Flagella	(Liu & Ochman, 2007)
flgC	Flagella	(Liu & Ochman, 2007)
flgD	Flagella	(Liu & Ochman, 2007)
flgE	Flagella	(Liu & Ochman, 2007)
flgF	Flagella	(Liu & Ochman, 2007)
flgG	Flagella	(Liu & Ochman, 2007)
flgH	Flagella	(Liu & Ochman, 2007)
flgI	Flagella	(Liu & Ochman, 2007)
flgJ	Flagella	(Liu & Ochman, 2007)
flgM	Flagella	(Liu & Ochman, 2007)
flgN	Flagella	(Liu & Ochman, 2007)
flhA	Flagella	(Liu & Ochman, 2007)
flhB	Flagella	(Liu & Ochman, 2007)

flhE	Flagella	(Liu & Ochman, 2007)
glnA	Glutamine synthetase	(Reitzer & Magasanik, 1986)
guaB	Inosine-5'-monophosphate dehydrogenase, required for synthesis of GMP from the common purine precursor	(Husnain et al., 2009)
purE	N5-carboxyaminoimidazole ribonucleotide mutase, de novo purine nucleotide synthesis	(Watanabe et al., 1989)
rhlE	RNA helicase regulates the function of related RNA helicases during ribosome assembly, works with *deaD* gene	(Jain, 2008)
rplC	L3 ribosomal protein	(Riley, 1993)
rplI	50s ribosomal protein	(Schnier et al., 1986)
rpsI	30s ribosomal protein	(Aseev et al., 2016)
secG	Protein-export membrane protein, promotes protein export across the inner membrane, protein translation and correct transport to periplasmic space	(Belin et al., 2015)
serA	D-3-phosphoglycerate dehydrogenase, catalyzes the first committed step in the "phosphorylated" pathway of L-serine	(Tobey & Grant, 1986)
sotB	Sugar efflux transporter	(Condemine, 2000)
xseA	Exodeoxyribonuclease 7 large subunit,	(Riley, 1993)
ycaD	Uncharacterized MFS-type transporter	https://www.uniprot.org/uniprot/P21503
ydhP	Inner membrane transport protein	https://www.uniprot.org/uniprot/P77389

ydjN	CysB regulon, which plays a central role in sulfur assimilation and cysteine metabolism, L-cystine transporter.	(Yamazaki et al., 2016)
yecR	Uncharacterized protein, upregulated by *flhDC* which is involved in flagella regulation	https://www.uniprot.org/uniprot/P76308
yeeF	Hypothetical transport protein	(Kaundal et al., 2020)
yeiB	Uncharacterized protein, likely involved in transport	https://www.uniprot.org/uniprot/P25747
yhjE	Inner membrane metabolite transport protein	https://www.uniprot.org/uniprot/P37643
yicE	Putative transport protein	(Karatza & Frillingos, 2005)
yjcD	Guanine/hypoxanthine permease	(Papakostas et al., 2013)

Table S1.8: GO analysis on the set of genes that have novel expression levels in all rescue mutants

GO biological process complete	Reflist (4391)	Observed	Expected	Over/under	Fold Enrichment	p-value
mannose transmembrane transport (GO:0015761)	3	3	0.03	+	> 100	1.18E-02
N-acetylglucosamine transport (GO:0015764)	4	3	0.03	+	89.01	2.05E-02
hexose import across plasma membrane (GO:0140271)	4	3	0.03	+	89.01	2.05E-02
glucose import across plasma membrane (GO:0098708)	4	3	0.03	+	89.01	2.05E-02
carbohydrate import across plasma membrane (GO:0098704)	4	3	0.03	+	89.01	2.05E-02
fructose import (GO:0032445)	5	3	0.04	+	71.21	3.26E-02
fructose transmembrane transport (GO:0015755)	11	4	0.09	+	43.15	5.88E-03
Unclassified (UNCLASSIFIED)	939	3	7.91	-	0.38	0.00E+00

Table S1.9: Direction of gene expression change for single mutants at 42.2°C

	Restored	**Reinforced**	**Unrestored**	**Novel**	**Total**
hslU indel	488	10	1253	310	2061
hslU G60D	1004	2	745	82	1833
hslU L163R	1186	1	564	76	1827
hslV H68P	835	2	914	64	1815
hslV indel	957	1	793	79	1830
rpoB H447L	1231	11	509	335	2086
rpoC W1020G	1174	2	575	115	1866

References

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25:25–29.

Baquero F, Cantón R. 2017. Evolutionary Biology of Drug Resistance. In: Mayers DL, Sobel JD, Ouellette M, Kaye KS, Marchaim D, editors. Antimicrobial Drug Resistance: Mechanisms of Drug Resistance, Volume 1. Cham: Springer International Publishing. p. 9–36. Available from: https://doi.org/10.1007/978-3-319-46718-4_2

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461:1243–1247.

Barrick JE, Kauth MR, Strelioff CC, Lenski RE. 2010. Escherichia coli rpoB Mutants Have Increased Evolvability in Proportion to Their Fitness Defects. *Mol Biol Evol* 27:1338–1347.

Bell, G. 2013. Evolutionary rescue of a green alga kept in the dark. *Biol. Lett.* 9(1):20120823

Bell G. 2017. Evolutionary Rescue. *Annu. Rev. Ecol. Evol. Syst.* 48:605–627.

Bell G, Gonzalez A. 2009. Evolutionary rescue can prevent extinction following environmental change. *Ecol. Lett.* 12:942–948.

Bennett AF, Lenski RE. 1993. Evolutionary Adaptation to Temperature II. Thermal Niches of Experimental Lines of Escherichia coli. *Evolution* 47:1–12.

Blair JMA, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJV. 2015. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* 13:42–51.

Bochtler M, Hartmann C, Song HK, Bourenkov GP, Bartunik HD, Huber R. 2000. The structures of HsIU and the ATP-dependent protease HsIU-HsIV. *Nature* 403:800–805.

Carroll SM, Chubiz LM, Agashe D, Marx CJ. 2015. Parallel and Divergent Evolutionary Solutions for the Optimization of an Engineered Central Metabolism in Methylobacterium extorquens AM1. *Microorganisms* 3:152–174.

Carroll SM, Marx CJ. 2013. Evolution after introduction of a novel metabolic pathway consistently leads to restoration of wild-type physiology. *PLoS Genet.* 9:e1003427.

Conrad TM, Frazier M, Joyce AR, Cho B-K, Knight EM, Lewis NE, Landick R, Palsson BØ. 2010. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. *Proc. Natl. Acad. Sci. U. S. A.* 107:20500–20505.

Cooper VS, Bennett AF, Lenski RE. 2001. EVOLUTION OF THERMAL DEPENDENCE OF GROWTH RATE OF ESCHERICHIA COLI POPULATIONS DURING 20,000 GENERATIONS IN A CONSTANT ENVIRONMENT. *Evolution* 55:889–896.

Cooper VS, Lenski RE. 2000. The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature* 407:736–739.

Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol. Clifton NJ* 1151:165–188.

Deatherage DE, Kepner JL, Bennett AF, Lenski RE, Barrick JE. 2017. Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. *Proc. Natl. Acad. Sci.* 114:E1904–E1912.

Délye C, Jasieniuk M, Le Corre V. 2013. Deciphering the evolution of herbicide resistance in weeds. *Trends Genet.* 29:649–658.

Dhar R, Sägesser R, Weikert C, Yuan J, Wagner A. 2011. Adaptation of Saccharomyces cerevisiae to saline stress through laboratory evolution. *J. Evol. Biol.* 24:1135–1153.

Gonzalez A, Bell G. 2013. Evolutionary rescue and adaptation to abrupt environmental change depends upon the history of stress. *Philos. Trans. R. Soc. B Biol. Sci.* [Internet] 368. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3538446/

González-González A, Hug SM, Rodríguez-Verdugo A, Patel JS, Gaut BS. 2017. Adaptive Mutations in RNA Polymerase and the Transcriptional Terminator Rho Have Similar Effects on Escherichia coli Gene Expression. *Mol. Biol. Evol.* 34:2839–2855.

Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, et al. 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* 38:1406–1412.

Holt RD. 1990. The microevolutionary consequences of climate change. *Trends Ecol. Evol.* 5:311–315.

Hug SM, Gaut BS. 2015. The phenotypic signature of adaptation to thermal stress in Escherichia coli. *BMC Evol. Biol.* 15:177.

Kanemori M, Nishihara K, Yanagi H, Yura T. 1997. Synergistic roles of HslVU and other ATP-dependent proteases in controlling in vivo turnover of sigma32 and abnormal proteins in Escherichia coli. *J. Bacteriol.* 179:7219–7225.

Kaundal S, Deep A, Kaur G, Thakur KG. 2020. Molecular and Biochemical Characterization of YeeF/YezG, a Polymorphic Toxin-Immunity Protein Pair From Bacillus subtilis. *Front. Microbiol.* [Internet] 11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00095/full

Kwon A-R, Trame CB, McKay DB. 2004. Kinetics of protein substrate degradation by HslUV. *J. Struct. Biol.* 146:141–147.

LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM. 2015. Use of Adaptive Laboratory Evolution To Discover Key Mutations Enabling Rapid Growth of Escherichia coli K-12 MG1655 on Glucose Minimal Medium. *Appl. Environ. Microbiol.* 81:17–30.

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500:571–574.

Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 109:E2774-2783.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *Am. Nat.* 138:1315–1341.

Lenski RE, Travisano M. 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci.* 91:6808–6814.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760.

Lien H-Y, Shy R-S, Peng S-S, Wu Y-L, Weng Y-T, Chen H-H, Su P-C, Ng W-F, Chen Y-C, Chang P-Y, et al. 2009. Characterization of the Escherichia coli ClpY (HslU) Substrate Recognition Site in the ClpYQ (HslUV) Protease Using the Yeast Two-Hybrid System. *J. Bacteriol.* 191:4218–4231.

Liu R, Ochman H. 2007. Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci.* 104:7116–7121.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat. Rev. Genet.* 16:567–582.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* [Internet] 15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/

Lucas JA, Hawkins NJ, Fraaije BA. 2015. The evolution of fungicide resistance. *Adv. Appl. Microbiol.* 90:29–92.

MacLean RC, Bell G, Rainey PB. 2004. The evolution of a pleiotropic fitness tradeoff in Pseudomonas fluorescens. *Proc. Natl. Acad. Sci. U. S. A.* 101:8072.

Missiakas D, Schwager F, Betton JM, Georgopoulos C, Raina S. 1996. Identification and characterization of HsIV HsIU (ClpQ ClpY) proteins involved in overall proteolysis of misfolded proteins in Escherichia coli. *EMBO J.* 15:6899–6909.

Mongold JA, Bennett AF, Lenski RE. 1999. EVOLUTIONARY ADAPTATION TO TEMPERATURE. VII. EXTENSION OF THE UPPER THERMAL LIMIT OF ESCHERICHIA COLI. *Evol. Int. J. Org. Evol.* 53:386–394.

Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA. 2006. Regulon and promoter analysis of the E. coli heat-shock factor, σ32, reveals a multifaceted cellular response to heat stress. *Genes Dev.* 20:1776–1789.

Orr HA, Unckless RL. 2008. Population Extinction and the Genetics of Adaptation. *Am. Nat.* 172:160–169.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rodríguez-Verdugo A, Carrillo-Cisneros D, González-González A, Gaut BS, Bennett AF. 2014. Different tradeoffs result from alternate genetic adaptations to a common environment. *Proc. Natl. Acad. Sci.* 111:12121–12126.

Rodríguez-Verdugo A, Tenaillon O, Gaut BS. 2016. First-Step Mutations during Adaptation Restore the Expression of Hundreds of Genes. *Mol. Biol. Evol.* 33:25.

Roncarati D, Scarlato V. 2017. Regulation of heat-shock genes in bacteria: from signal sensing to gene expression output. *FEMS Microbiol. Rev.* 41:549–574.

Rosenzweig F, Sherlock G. 2014. Editorial: Experimental Evolution: Prospects and Challenges. *Genomics* 104:v–vi.

Shi W, Zhou Y, Wild J, Adler J, Gross CA. 1992. DnaK, DnaJ, and GrpE are required for flagellum synthesis in Escherichia coli. *J. Bacteriol.* 174:6256–6263.

Shin DH, Yoo SJ, Shim YK, Seol JH, Kang MS, Chung CH. 1996. Mutational analysis of the ATP-binding site in HslU, the ATPase component of HslVU protease in Escherichia coli. *FEBS Lett.* 398:151–154.

Sim M, Koirala S, Picton D, Strahl H, Hoskisson PA, Rao CV, Gillespie CS, Aldridge PD. 2017. Growth rate control of flagellar assembly in Escherichia coli strain RP437. *Sci. Rep.* [Internet] 7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5259725/

Somero GN. 1978. Temperature Adaptation of Enzymes: Biological Optimization Through Structure-Function Compromises. *Annu. Rev. Ecol. Syst.* 9:1–29.

Soutourina OA, Bertin PN. 2003. Regulation cascade of flagellar expression in Gram-negative bacteria. *FEMS Microbiol. Rev.* 27:505–523.

Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The Molecular Diversity of Adaptive Convergence. *Science* 335:457–461.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D., Lenski, R. E. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536: 165–170.

The Gene Ontology Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47:D330–D338.

Williams GC. 1957. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* 11:398–411.

Yoo SJ, Kim HH, Shin DH, Lee CS, Seong IS, Seol JH, Shimbara N, Tanaka K, Chung CH. 1998. Effects of the cys mutations on structure and function of the ATP-dependent HslVU protease in Escherichia coli. The Cys287 to Val mutation in HslU uncouples the ATP-dependent proteolysis by HslvU from ATP hydrolysis. *J. Biol. Chem.* 273:22929–22935.

Yoo SJ, Shim YK, Seong IS, Seol JH, Kang MS, Chung CH. 1997. Mutagenesis of two N-terminal Thr and five Ser residues in HslV, the proteolytic component of the ATP-dependent HslVU protease. *FEBS Lett.* 412:57–60.

CHAPTER 2

Genotypic evolution and fitness outcomes are contingent on the adaptive strategy to heat stress

2.1 Abstract

There exists ambiguity as to whether an organism's previous selection environment affects its future adaptive potential due to the historical nature of evolution. Here, we studied how history affects future evolution by performing a two-phase evolution experiment. The first experiment consisted of 115 lines of *Escherichia coli* adapted to the stressful temperature of 42.2°C. Evolution occurred through two adaptive pathways: mutation of *rpoB*, encoding RNA polymerase, or through *rho*, a transcriptional terminator. Thus, the pathways represent two distinct evolutionary histories resulting from natural selection. We contrasted these two histories in a second phase of evolution. In Phase 2, we evolved a subset of the Phase 1 *rpoB* and *rho* genotypes, along with the Phase 1 Founder, representing a total of 72 evolved populations, for 1,000 generations at 19.0°C. All evolved populations increased their relative fitness by 8% on average in response to evolving in Phase 2. The magnitude of fitness change was not significantly influenced by the adaptive pathway, as evidenced by an overall convergence towards similar fitness values across all evolved populations. However, we did observe that the initial genotype significantly affected relative fitness, suggesting that phenotypic changes can be influenced by different codon mutations in the same gene, therefore suggesting an influence of evolutionary history. We sequenced whole populations and identified 1,387 mutations that arose during Phase 2 evolution, of which 119 were fixed. While the total number and type of mutations did not differ between isolates from different adaptive pathways, we identified genes

frequently mutated in evolved populations descended from the same starting adaptive pathway, suggesting an influence of history on genotypic changes. Our results suggest that evolutionary history can significantly influence adaptive potential and may be an important factor towards evolutionary forecasting.

2.2 Introduction

Phenotypic and genotypic evolution are influenced by the processes of mutation, genetic drift, recombination, and natural selection (Bobay & Ochman 2017). These evolutionary forces do not act in isolation, rather, the combination of these forces push an organism along a particular evolutionary trajectory. In this regard, evolution is an inherently historical process, and the evolutionary history of an organism can often be traced by analyzing genomic content. However, the extent to which evolutionary history itself can have an impact on future adaptation and evolutionary events is not well established. In other words, is evolution contingent on past events?

Famously, Stephen Jay Gould argued that history is an essential feature of evolution. If evolution was not contingent on historical processes, then natural selection would be able to overcome history such that the highest-fitness solution to any environmental challenges could always be selected for, rendering evolution deterministic and predictable. If history does have an impact on future evolution and adaptation, then evolution would be unpredictable and unrepeatable as evolutionary trajectories would be contingent on the previous events. This is an important distinction because many biological questions rely on predicting evolutionary outcomes. For example, predictive questions are crucial for identifying species that may or may not survive dramatic environmental change, to produce chemotherapeutic agents against cancerous tumors and prevent subsequent resistance against the treatment, and to forecast pathogen variation and mutation which is incredibly relevant given the Covid-19 pandemic (Vlachostergios & Faltas 2018; Leray et al. 2021; Bay et al. 2017). Understanding if and how evolution is contingent on past events will allow for more reliable predictions of evolutionary outcomes.

The extent to which historical contingency impacts evolution has been studied in natural populations and in the laboratory. In natural populations, experiments have focused on established populations of vertebrates and tried to infer contingency from convergent evolution events (Losos 2011). For example, one such experiment tested brown anole lizard populations that were subjected to living on narrow perches. All lizard populations evolved shorter limbs in response to this selection pressure (Kolbe et al. 2012). Similarly, in an experiment with guppies, male guppies across different populations evolved shorter life histories in the absence of predators (Reznick & Bryga 1987). Some use these convergent phenotypes as evidence that natural selection is predictable and that contingencies of history therefore do not play a major role in evolution. However, this interpretation is debated, particularly because these observational experiments may rely on standing genetic variation in the ancestral population that can increase the probability of parallel responses (Blount et al. 2018).

In the laboratory, controlled experiments can be conducted to investigate the extent of contingency on evolution, especially when utilizing isogenic populations or identical environmental conditions (Blount et al. 2018). With careful consideration of experimental design, questions regarding evolutionary contingency can be addressed. Bacterial populations offer a good way to study the effects of contingency because of their amenability to experimentation. Several studies have investigated contingency in microbial populations utilizing the methods of experimental evolution. However, no clear pattern has emerged. For example, in the Long Term Evolution Experiment (LTEE), phenotypic convergence among 12 replicate *Escherichia coli* populations was evidenced by increased fitness, faster growth, and larger cells (Bennett & Lenski 1993; Wiser et al. 2013). However, some adaptations were unique among populations, most notably the ability to utilize citrate present in the media (Blount et al. 2008). To discern whether this adaptation occurred due to the appearance of a rare mutation or was contingent on previous adaptations, Blount et al. (2008) investigated contingency by "analytically replaying" evolution with populations re-founded from frozen samples of previous generations. They found that populations restarted from generation 20,000 onward were more likely to develop citrate utilization, indicating that adaptations present in the later generations were essential for this trait to evolve, showing that contingency had an impact on the evolutionary trajectory - both genotypically and phenotypically. Similarly, antibiotic resistance evolution was studied using the LTEE lines with results suggesting that both the level of resistance and the types of resistance mutations that occurred were contingent on the starting genetic background (Card et al. 2021).

In other experimental systems, the influence of contingency on future evolution is not as clear. For example, Plucain et al. (2016) evolved 16 populations of *E. coli* in four different chemical environments for 1,000 generations before being propagated for another 1,000 generations in a single, new environment (Plucain et al. 2016). Phenotypic evolution, as measured by fitness in this two-phase experiment, was found to be contingent on the evolutionary history of each population. However, in contrast to results observed in the LTEE, the authors found that the mutations that arose during the second phase of evolution were not contingent upon the mutations that arose during the first phase. In addition, a two-step evolution experiment with yeast has suggested that phenotypic and genotypic changes are not influenced by evolutionary history and that fitness will follow predictable trajectories (Kryazhimskiy et al. 2014).

Here, we perform an evolution experiment in two phases to study evolutionary contingency. We build on an evolution experiment previously described in Tenaillon et al. 2012. In this first phase of evolution, 115 initially identical lines of *E. coli* were evolved at the stressful temperature of 42.2°C (Tenaillon et al. 2012) from a single *E. coli* founder strain. After 2000 generations of evolution, a clone from each population was sequenced, and these sequences revealed that adaptation occurred through two adaptive pathways: mutation the RNA polymerase subunit beta gene (*rpoB*), or in the transcriptional terminator (*rho*) gene. Intriguingly, these two pathways were each significantly associated additional but distinct sets of mutations, thus representing two distinct evolutionary histories that resulted from natural selection. For example, clones that had mutations in *rpoB* also tended to have mutations in *rod, ILV* and *RSS* genes, whereas mutations in these genes were rare in clones that had mutations in *rho*. This first phase of evolution represents a unique opportunity to study contingency, because natural selection resulted in two distinct adaptive pathways (therefore, represent different evolutionary histories) that can be leveraged to study contingency in a second phase of evolution.

By performing evolution in two phases, we have investigated the impact of evolutionary history on future adaptive trajectories at both the genotypic and phenotypic level. Our study is unique as evolution during Phase 1 occurred in a single environment and resulted in two distinct adaptive histories due to natural selection that can be directly contrasted in a second evolution regime referred to as Phase 2. By evolving a subset of lines from the two adaptive histories (*rpoB* or *rho* genotypes), we address questions about the effect of historical contingency on phenotypic and genotypic evolution. We ask: Do the *rpoB* and *rho* lines differ in their response to selection in Phase 2? Do the *rpoB* and *rho* adapted lines differ in their phenotypic evolution as measured through fitness? When placed in a new environment, do the *rpoB* and *rho* adapted lines evolved by different pathways? Using both phenotypic and genotypic data, we address the impact of historical contingency on future evolution.

2.3 Methods

Two-Phase Evolution Experiment Isolate Criteria and Selection. To study evolutionary contingency, we relied on two evolution experiments, referred to as Phase 1 and Phase 2. The first evolution experiment, Phase 1, was conducted previously (Tenaillon *et al.* 2012). Briefly, 114 initially identical lines of *E. coli* strain B REL1206 were serially propagated at the stressful temperature of 42.2°C. Natural selection in this thermal environment resulted in evolution through two distinct adaptive pathways: the *rpoB* pathway or the *rho* pathway. Adaptive mutations in *rpoB* and *rho* were not specific to particular codons, so we selected representative genotypes carrying different codon mutations in both *rpoB* and *rho* for evaluation. Altogether, these two adaptive pathways represent two distinct evolutionary histories that arose by natural selection that can be contrasted in a second phase of evolution in order to investigate if and how historical contingency influences evolution.

The second phase of evolution was conducted at 19°C, which is towards the lower limit of the temperature niche for the REL1206 ancestor (Rodríguez-Verdugo et al. 2014). The Phase 2 experiment was founded from a subset of high temperature evolved ancestors from Phase 1, referred to as Phase 2 founders. We selected 5 representative individuals from the *rpoB* and *rho* adaptive pathways to compare their evolution in this new thermal environment. To determine the founders (specific *rpoB* and *rho* genotypes) for Phase 2, we developed four criteria. First, the isolate had to carry a single mutation in either *rpoB* or *rho* but no mutations in both genes. Second, the isolates should represent the breadth of mutations, so we were careful to select isolates with different codon mutations in *rpoB* and *rho*. Third, we identified a collection of isolates that reflect the range of fitness values at both 19.0°C and 42.2°C for the Phase 1 evolved lines as previously measured in Rodriguez-Verdugo *et al.* 2014 (Table 2.1). Fourth, the isolates had to survive a nine day extinction test, in which we grew and maintained the potential isolates in liquid culture through daily transfer at 19.0°C. Briefly, we grew the isolates from frozen stock in Luria-Bertani medium (LB) and incubated them at 37.0°C for one day to acclimate from frozen conditions (Rodríguez-Verdugo et al. 2014; Bennett & Lenski 1993; Lenski & Travisano 1994). The overnight culture was diluted 1,000-fold in saline and this dilution was transferred into fresh Davis Minimal (DM) Media supplemented with 25mg/L of glucose and grown for one day at 37.0°C. Following incubation, 100ul of the culture was transferred into 9.9ml of fresh DM media and incubated at 19.0°C and serially propagated for at least nine days. Each day we measured the cell density to determine if extinctions had occurred. To measure the cell density, 50ul of overnight culture was diluted in 9.9ml of Isoton II Diluent (Beckman Coulter) and measured in volumetric mode on a Multisizer 3 Coulter Counter (Beckman

Coulter). An isolate was determined to survive if its cell density measurements were maintained over the course of the test while allowing for fluctuations of +/- 1x10^6 cells.

Altogether, we selected 10 different *rpoB* and *rho* genotypes for evolution at 19.0°C with 6 replicates each (Figure 2.1, Table 2.1). These 10 starting genotypes are referred to as the Phase 2 Founders. We also included 12 replicates of the original ancestral *E. coli* REL1206 from Phase 1 (Phase 1 Founder), which represents a third evolutionary history with which we can use for comparison in Phase 2.

Evolution Experiment at 19.0°C. To prepare the isolates for the Phase 2 experiment, the selected evolved lines from Phase 1 and the ancestor were grown from frozen stock in 10 ml of LB at 37.0°C with 120 RPM. After 24 hours of incubation, the overnight cultures were diluted 10,000-fold and plated onto TA plates and incubated at 37.0°C. On the next day, single colonies were picked from the plates and inoculated into 10 mL of fresh LB and incubated at 37.0°C with 120 RPM. We started six replicate lines of each *rho* and *rpoB* genotype, as well as 12 replicate lines of the Phase 1 Founder resulting in 72 lines for evolution in Phase 2. The next day, we transferred 100 µl of the bacterial culture into 9.9 ml of fresh DM25 media, which was incubated at 37.0°C at 120 RPM for 24 hours to acclimate to experimental conditions, following common practice (Lenski & Travisano 1994; Bennett & Lenski 1993; Rodríguez-Verdugo et al. 2014). After incubation, we began the Phase 2 evolution experiment by transferring 100 µl of culture into 9.9 ml of fresh DM25 and incubated the tubes at 19.0°C with 120 RPM and incubated for 24 hours.

Each day, the cultures were transferred daily into fresh media via a 100-fold dilution. At regular intervals (at generation 100 and roughly every 200 generations after that), we mixed 800ul of each line with 800 ul of 80% glycerol to prepare whole population frozen stocks. We began the experiment in January 2020, but due to the Covid-19 pandemic, we had to pause the experiment after 297 bacterial generations. To restart the evolution experiment, we revived the bacterial populations by transferring 100ul of thawed glycerol stock into 9.9 mL of fresh DM25 media. We incubated the lines in Phase 2 experimental conditions, and we transferred the lines daily as previously described until the bacteria had grown for 1000 generations or 152 days.

Measuring Relative Fitness. We performed competition experiments to measure the relative fitness of the Phase 2 evolved lines after 1000 generations of evolution. We competed the Phase 2 evolved lines against the Phase 1 Founder and their respective Phase 2 Founder at both 19.0°C and 42.2°C. To perform the competitions, we mixed the cells in a single glass culture tube and plated the mixture to count the colonies before and after 24 hours of competition. We used the neutral Ara+ marker to differentiate between the two lines when plating on tetrazolium-arabinose (TA) plates. To generate Ara+ mutants from the Phase 2 Founders for competitions, we followed previously published methods as described in (Lenski et al. 1991). To validate neutrality, we competed the Ara+ mutants against the original Ara- stock using the methods described below.

To perform competition assays, bacteria from frozen glycerol stocks were revived with a loop into 10 mL of LB and incubated at 37°C with 120 RPM for 24 hours. After incubation, the overnight cultures were vortexed and 100 µl of each were diluted in 9.9 ml of 0.0875% saline solution. From each dilution tube, 100 µl was transferred to 9.9ml DM25 to incubate at 37.0°C with 120 RPM for 24 hours. Following incubation and in order for the bacteria to acclimate to the experimental temperature, we transferred 100 µl of the overnight cultures into 9.9 ml of DM25 and incubated the tubes at the experimental temperature (19.0°C or 42.2°C) with 120 RPM for 24 hours (Bennett & Lenski 1993). The next day, we mixed the Ara- and Ara+ competitor strains into sterile DM25 media. For competitions at 19.0°C, we mixed the bacteria 1:1. For competitions at 42.2°C, we mixed the bacteria 1:1 or we adjusted the ratio to 1:3 if the original ratio resulted in too few colonies (<20) on the plate for either competitor. The mixture was incubated at the experimental temperature with 120 rpm for 24 hours. After allowing the cells to compete, we quantified the cell density of each competitor by plating the overnight culture onto tetrazolium-arabinose (TA) plates and counting the number of colonies. All competitions were performed in at least triplicate, resulting in roughly 600 competitions.

Using the methods described in Lenski *et al.* (1991) and Tenaillon *et al.* (2012), we quantified the relative fitness, w_r. The fitness of a Phase 2 evolved line relative to its competitor was estimated by:

$$w_r = [\log_2(N^M{}_f/N^M{}_i)]/[\log_2(N^A{}_f/N^A{}_i)]$$

Where $N^M{}_i$ and $N^A{}_i$ represent the initial cell densities of the two competing clones, and $N^M{}_f$ and $N^A{}_f$ represent the final cell densities of the two clones after one day of competition.

DNA Library Preparation and DNA Sequencing. In order to sequence the 1000 generation evolved populations, we revived ~10 ul of frozen glycerol stock in 10ml of DM media supplemented with 1000mg/L of glucose. The culture tubes were incubated at 19.0°C with 120 RPM. Because some populations went extinct (see below), we extracted from 65 bacterial populations using the Promega Wizard Genomic DNA Purification kit.

DNA concentrations were measured with Qubit dsDNA HS Assay kits. We prepared our DNA sequencing libraries with the Illumina Nextera DNA Flex Library Preparation kit. The libraries were multiplexed and sequenced using the Illumina NovaSeq on an S4 flow cell to generate 100bp paired-end reads at UC Irvine's Genomics High-Throughput Facility (https://ghtf.biochem.uci.edu). Sequencing read quality was assessed with FastQC v. 0.11.9 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), trimmed with fastp v. 0.23.2 (Chen et al. 2018), and visualized with MultiQC v. 1.9 (Ewels et al. 2016).

Variant Detection. We detected mutations and their respective frequencies in each evolved Phase 2 population using breseq v. 0.35.5 (Deatherage & Barrick 2014). We performed the breseq analysis in polymorphism mode with two different reference genomes. First, we performed breseq analysis using *E. coli* strain B REL606 as the reference genome. This *E. coli* strain differs from the Phase 1 Founder, REL1206, in seven positions (*topA*, *spoT K662I*, *glmU*/*atpC*, *pykF*, *yeiB*, *fimA* and the *rbs* operon) that were excluded from our analysis (Barrick et al. 2009; Tenaillon et al. 2012). We performed a first-round of variant detection using breseq in polymorphism mode on all of the 1000-generation evolved populations. We also performed breseq analysis in consensus mode on the Phase 2 Founders that were previously isolated and sequenced following Phase 1 evolution in Tenaillon et al. 2012.

Following this first-step of the analysis, we generated a mutated reference for each Phase 2 Founder using the gdtools APPLY command in breseq. We then ran the breseq analysis again using the respective mutated reference to verify mutation predictions, as described in Deatherage and Barrick 2014. Using gdtools available through breseq, we compiled the mutation information into readable tables and as an alignment file in PHYLIP format. A phylogeny was constructed using IQ-tree and the PHYLIP alignment as input (Nguyen et al. 2015).

We generated VCF files representing the fixed SNPs among the Phase 1 Founder, Phase 2 Founders, and the Phase 2 1000-generation evolved populations using Snippy v. 4.4.0 with the trimmed sequencing reads as input (Seemann 2022). The VCF files generated from Snippy were used as input for Plink2 for principal components analysis (Chang et al. 2015).

Genotypic Statistical Analysis.. To statistically test for associations between the mutation patterns observed in Phase 2 and their initial adaptive pathway, we first built a distance matrix from the presence and absence matrix of Phase 2 mutations in R v 4.0.2 (R Core Team 2019). Using the vegan package v 2.5-7 in R, we directly tested for associations between the distance matrix of Phase 2 mutations and the adaptive pathway or mutated codon with ANOSIM (Oksanen et al. 2020). We also built a Neighbor-Joining (NJ) tree based on the presence-absence matrix of accessory genes. To do so, we first calculated the Euclidean distances from the presence-absence matrix of the accessory genes using the *dist* function in R. We then built the NJ tree from the Euclidean distances using the ape package in R (Paradis & Schliep 2019). To identify genes or intergenic regions that were more frequently mutated in populations descended from one adaptive history but not the other, we built a 2 × 2 Fisher's Exact Test for each mutated gene or intergenic region in R.

2.4 Results

We performed a two-phase evolution experiment to study the effects of contingency on future evolution and adaptive potential (Figure 2.1). In the first phase of evolution, 115 initially identical lines of *E. coli* were evolved at the stressful temperature of 42.2°C (Tenaillon et al. 2012). After 2,000 generations of evolution, the lines experienced fitness gains of about 42% on average relative to the REL1206 ancestor. Full genome sequencing revealed that the populations traversed one of two evolutionary trajectories. Of the 115 lines, 76 carried a mutation in *rpoB* and 43 lines carried a mutation in *rho*. The *rpoB* or *rho* trajectories were each significantly associated with mutations in different sets of genes. To study contingency, we relied upon this initial phase of evolution and performed a second phase of evolution founded from a subset of the high temperature adapted lines. We selected 5 representative genotypes of the *rpoB* and *rho* pathways to serve as the Phase 2 Founders (6 replicates each per Phase 2 Founder) and 12 replicates of the ancestral *E. coli* REL1206 (referred to as the Phase 1 Founder for the remainder of the text) for experimental evolution at 19.0°C, a new thermal environment. In total, we evolved 72 *E. coli* populations at 19.0°C in minimal media for 1,000 generations. Among these 72, we observed that 7 populations that went extinct during the course of evolution: 1 population descended from the Phase 1 Founder, 4 populations descended from Phase 2 Founder Line 3 (*rpoB* I966S), and 2 populations descended from Phase 2 Founder Line 142 (*rpoB* I572L). Therefore, the following analyses were performed on the set of 65 populations that survived throughout the 1000 generation experiment.

Changes in relative fitness at 19.0°C. Historical contingency may affect the rate of fitness change thus impacting phenotypic outcomes in new environments. To investigate whether evolutionary contingency affected phenotypic evolution in our system, we measured the changes in fitness of the Phase 2 evolved populations. First, we measured the changes in relative fitness (w_r) of the Phase 2 populations by competing them against their respective Phase 2 Founder at 19.0°C. By competing the evolved populations with their Phase 2 Founder, we assessed phenotypic changes as a result of adaptation to only the Phase 2 environment, thus allowing for independent investigation and comparison of the populations descended from differing starting adaptive pathways. On average, the Phase 2 populations had w_r = 1.08, thus reflecting on average an 8% fitness advantage at the end of the experiment. Lines descended from *rpoB* genotypes experienced a 10% fitness advantage on average, while those descended from *rho* founders had a 7% fitness advantage. However the difference was not significant (P = 0.0798, unpaired t-test; Figure 2.2A). We also investigated the changes in w_r at the level of *rho* or *rpoB* genotype and found that 8 of the 10 genotypes had significant fitness advantages at the end of Phase 2 (Table 2.2; Figure 2.3). Using ANOVA, we found a significant effect of the genetic background at the level of adaptive codon on relative fitness ($P = 1.54 \times 10^{-6}$, ANOVA).

Second, we measured the fitness of all Phase 2 Founders and Phase 2 Evolved Populations against the Phase 1 Founder at 19.0°C (Figure 2.4). With this method, we could directly compare each population's fitness changes relative to a single competitor. Of the Phase 2 Founders, only one (*rpoB* I966S background) had a significant fitness decline at 19.0°C relative to the Phase 1 Founder (Figure 2.4), and the effect of adaptive codon on relative fitness was statistically significant as detected by ANOVA (P = 0.019, ANOVA; Table 2.1).

In total, the evolved populations increased in fitness by 3.8%, on average, at the end of Phase 2 relative to the Phase 1 Founder (Figure 2.4). Investigation at the level of adaptive pathway revealed that, on average, lines descended from *rpoB* backgrounds increased by 1%, *rho* lines decreased by 6%, and those descended from the Phase 1 Founder increased by 5%. The difference in relative fitness among the groups was significant at the level of adaptive pathway (P = 0.0007337, Kruskal-Wallis test) and at the level of codon mutation (P = 0.0001793, Kruskal-Wallis test). The Wilcoxon Rank Sum Test revealed that Phase 2 evolved populations descended from *rpoB* founders had significantly different average w_r to evolved populations descended from *rho* (P = 0.0007). Moreover, at the level of the adaptive codon we found significant variation of average w_r among the Phase 2 evolved populations (P = 0.000529, ANOVA).

High temperature fitness tradeoffs after Phase 2 evolution relative to Phase 2 Founder. Fitness tradeoffs are pervasive in evolution experiments, and previous research has demonstrated significant differences in tradeoff dynamics between adaptive genotypes (Rodríguez-Verdugo et al. 2014). Thus, evolutionary contingency may influence the tradeoff dynamics that occur after organisms are exposed to new environments. To measure tradeoff dynamics, we competed the Phase 2 evolved populations against their respective Phase 2 Founder at 42.2°C. Altogether, the Phase 2 evolved populations had an average w_r = 0.8846 reflecting significant fitness declines at 42.2°C (P = 3.338×10^{-10}, t-test). The difference in average w_r between *rho* and *rpoB* populations was not statistically significant despite *rho* experiencing a fitness decline of 8.4% and *rpoB* by 15.4% on average (P = 0.05112, t-test; Figure 2.2B).

We were interested in investigating whether the background at the level of adaptive pathway or adaptive codon influenced the tradeoffs dynamics at high temperature (Figure 2.5). However, the starting genotype significantly influenced average w_r at 42.2°C at the level of adaptive codon ($P = 6.83 \times 10^{-10}$, Kruskal-Wallis test). Of the ten different adaptive backgrounds differentiated by the specific codon mutation in *rho* or *rpoB*, seven backgrounds had significantly lower fitness at 42.2°C compared to their Phase 2 Founder counterpart (Table 2.2). Lines descended from the *rpoB* G4446S and *rho* V206A adaptive backgrounds had the lowest fitness at 42.2°C with $w_r = 0.501$ ($P = 1.34 \times 10^{-6}$, t-test) and $w_r = 0.795$ (P = 0.01242, t-test). Additionally, Phase 2 evolved populations descended from the *rpoB* I572L, *rpoB* I966N, *rho* I15N_1, *rho* A43T, and *rho* T231A backgrounds all had significant decreases in relative fitness (Table 2.2).

Genome sequencing of Phase 2 evolved populations. Next, we investigated the effect of historical contingency at the level of genomic evolution. At the end of Phase 2 evolution to 19.0°C, we sequenced the DNA from whole populations. We identified genomic variants in the evolved populations with breseq, which identifies variants in haploid genomes using a reference sequence. First, we investigated the presence of Phase 1 mutations in the Phase 2 evolved lines and found that all mutations that arose during Phase 1 were maintained in the Phase 2 evolved lines. Next, we identified the mutations that arose during Phase 2 evolution at 19.0°C. In total, we identified 1,387 novel mutations occurring at a 5% frequency or higher in our Phase 2 evolved populations (Figure 2.6). The distribution of mutation frequencies among all Phase 2 populations illustrated that almost half of all mutations (45%) were present at a frequency of < 10% in the populations (Supplemental

Figure S2.1). Among the 1,387 mutations, we identified 119 fixed mutations, which were found at a frequency of 85% or higher in populations, in 53 of the 65 sequenced populations. Of the fixed mutations, 98 were at 100% frequency.

The largest portion of mutations occurred in intergenic regions of the genome: 742 mutations were in intergenic regions and 618 mutations were in genic regions (Figure 2.7). In the intergenic regions, an overwhelming majority of mutations were point mutations (95.8%, 711/742) with the rest composed of indels (4.2%, 31/742), which are short insertions or deletions less than 50 bp long. Of the point mutations in intergenic regions, the majority were transversions (93.7%, 666/711); transitions made up only 6.3% of point mutations in intergenic regions.

Next, we investigated the 618 mutations that occurred within genes, as these mutations are most likely to drive adaptation. We first classified the mutations by type across all populations and found 413 mutations were point mutations (66.8%) and 205 were indels (33.2%). We further categorized the point mutations into either nonsynonymous or synonymous mutations. A majority, 89.8%, of point mutations that arose in genes were nonsynonymous mutations (371/413) while 10.2% were synonymous mutations (42/413). Of the indels, insertions were more common (84.4%, 173/205) over deletions (15.6%, 32/205). Nonsynonymous mutations and indels are most likely to drive adaptation as they directly interfere with the protein sequence by either changing the amino acid or causing a frameshift, respectively.

We investigated the phylogenetic relationships between the evolved populations using the polymorphism information. The resultant phylogeny illustrated that all Phase 2 Founders and their descendants were appropriately clustered with each other, suggesting that contamination did not occur during the Phase 2 (Supplemental Figure S2.2). We further visualized the isolates through principal components analysis (Supplemental Figure S2.3). In plots of the first and second principal components, most populations clustered together, but five *rpoB*-descended populations were differentiated based on the identity of the Phase 2 Founder. Visualizing the second and third principal components further differentiated the lines based on the Phase 2 Founder identity.

Effects of historical contingency on genomic evolution. To investigate the effects of historical contingency on genomic evolution, we performed statistical analysis to contrast the mutational patterns found in the lines descended from the *rpoB* genotypes, *rho* genotypes, or the Phase 1 Founder. We were interested in testing whether the lines significantly differed in the patterns and types of mutations that arose in their genomes during Phase 2. First, we explored any potential differences in the proportion of mutational variant types, namely: intergenic, frameshift, nonsynonymous, and synonymous mutations and large deletions spanning over 50 base pairs. We found no significant effect of contingency due to the adaptive pathway on the total number of each mutation type occurring in the evolved populations at a frequency of 5% or higher (P = 0.7809, contingency test; Figure 2.8). This previous result considered all mutations that could be detected at a 5% frequency or higher which could contribute excess noise, so we next investigated whether the mutations that reached fixation in these populations differed due to contingency at the level of the adaptive pathway. We considered the set of fixed mutations separately as they were likely drivers of adaptation. Similarly, we found no significant difference in the proportion of each mutation type for those mutations that reached fixation (P = 0.0803, contingency test; Figure 2.8).

We also tested whether the specific genes and intergenic regions that were mutated during Phase 2 evolution differed due to historical contingency defined by the adaptive pathways of Phase 1 founders. To do so, we built a distance matrix and NJ tree from the Phase 2 populations' presence-absence pattern of genic and intergenic mutations and tested for an association with Phase 1 founder pathways using ANOSIM (Figure 2.9). We found a significant association between the mutations that arose during Phase 2 and the adaptive history at the level of pathway (descended from *rpoB*, *rho*, or Phase 1 Founder), suggesting that the mutations that arose during Phase 2 evolution were influenced by historical contingency (ANOSIM R = 0.139, P = 4 × 10^{-4}). Additionally, we tested the association between the mutation presence-absence patterns and the adaptive history at the level of codon mutation (Table 2.1). We hypothesized that the specific codon mutation in *rpoB* or *rho* would significantly influence genomic evolution, as previous research has shown significant differences in phenotypic outcome between different SNPs in *rpoB* and *rho* (González-González et al. 2017; Rodríguez-Verdugo et al. 2014). The ANOSIM test was significant for the effect of adaptive history at the level of mutated codon on the mutational patterns observed at the end of Phase 2 evolution (ANOSIM R = 0.2398, P = 1 × 10^{-4}).

Our results suggest that the mutations that arose in the Phase 2 experiment were influenced by historical contingency. In order to determine if the different starting adaptive history influenced the mode of adaptation during Phase 2, we identified specific mutations that exhibit contingency among the adaptive histories. Using Fisher's Exact Test, we identified six genes or intergenic regions that were more frequently mutated in Phase 2 evolved populations descended from one adaptive history but not the other (P < 0.05, Table 2.3). We note that significance levels were lost (P > 0.05) after P-value correction with FDR, perhaps reflecting low statistical power due to sample size.

2.5 Discussion

Evolution is an inherently historical process, but the influence of historical contingency on future evolution and adaptation is not well characterized. We performed a two-phase evolution experiment to investigate the influence of historical contingency on bacterial genotypic and phenotypic evolution. Phase 1 of the evolution experiment, previously described in Tenaillon et al. (2012), consisted of over 100 initially identical populations of *E. coli* evolved under the stressful temperature of 42.2°C. Evolution in this phase resulted in the *E. coli* populations adapting by either one of two pathways and represented two adaptive histories: *rpoB* genotypes or *rho* genotypes (Rodríguez-Verdugo et al. 2014). Here, a subset of the populations from Phase 1 were selected to serve as founders of Phase 2 evolution at 19.0°C (Table 2.1). With this two-phase experiment, we have compared phenotypic and genotypic evolution between the two adaptive histories to assess the effects of contingency on adaptation.

Phenotypic evolution. To assess phenotypic evolution in our system, we measured relative fitness by competing evolved populations against a reference genotype, either the Phase 1 Founder or the respective Phase 2 Founder (Figure 2.1). First, we characterized the changes in fitness of the evolved populations at 19.0°C, the Phase 2 thermal environment, against the Phase 2 Founders. In general, the Phase 2 evolved populations experienced an

8% fitness increase on average, indicating adaptive evolution in our system, however, the difference between average relative fitness between *rpoB* and *rho* descended populations was not significant (Figure 2.2). Similarly to experiments using yeast, we found that phenotypic evolution may converge despite initial differences in adaptive history or starting genotype (Kryazhimskiy et al. 2014). In fact, one might expect that the 3% fitness difference observed between *rho* and *rpoB* descendants after 1,000 generations of evolution would dampen with time if let to evolve in the same environment. Likewise using *E. coli*, Plucain et al. (2016) found that the evolution trajectory a population may take when adapting to a new environment may be contingent, but certain phenotypes, like growth rate, may improve over time to a point that they do not appear to be influenced by contingency (Plucain et al. 2016).

While the general trends between *rho* and *rpoB* descended populations do not indicate an influence of historical contingency, we examined the differences in fitness change at the level of initial genotype designated by the specific codon mutation in either *rho* or *rpoB* (Table 2.1). The starting genotype significantly influenced phenotypic evolution based on relative fitness at 19.0°C and 42.2°C, suggesting that bacteria with different mutations in the same gene could experience significantly different evolutionary trajectories. Previous work in *E. coli* supports the notion that different mutations in the same gene can influence evolutionary trajectories, because single mutations in both *rpoB* and *rho* significantly influenced gene expression and fitness in unique ways (Rodríguez-Verdugo et al. 2014; González-González et al. 2017; Batarseh et al. 2020). Altogether, our results suggest that the trajectory of phenotypic evolution may be influenced by historical contingency at the level of genotype such that mutational differences in the same gene may significantly alter trajectories, however, the phenotype may converge with time.

Evolutionary theory, specifically Fisher's geometric model (Fisher 1930), may explain some of the dynamics in our system. This model predicts that genotypes further from the local optimum experience larger leaps in phenotypic change in a shorter amount of time due to large effect mutations, compared to those closer to the optimum that will accumulate smaller effect mutations and thus smaller phenotypic advancements (Fisher 1930). We were interested in exploring whether the initial differences in fitness at the start of Phase 2 at 19.0°C drove phenotypic evolution (Figure 2.4). To do so, we measured the fitness of all Phase 2 Founders and Phase 2 Evolved Populations at 19.0°C relative to a single genotype: the Phase 1 Founder REL1206 (Figure 2.1). Only one founder, the *rpoB* I966S genotype, had a significant fitness disadvantage relative to the Phase 1 Founder. Intriguingly, the evolved populations descended from the *rpoB* I966S had the greatest fitness increase at 19.0°C relative to its Phase 2 Founder (w_r = 1.14506; Table 2.2).

We investigated the tradeoff dynamics in our system by measuring the changes in fitness of the Phase 2 Evolved Populations at 42.2°C, the Phase 1 environment, relative to the Phase 2 Founders. In agreement with our observations at 19.0°C, the difference in fitness at 42.2°C between the *rho* and *rpoB* genotypes was not statistically significant, suggesting that the adaptive history at the level of pathway did not influence the overall phenotypic changes (Figure 2.2). However, there was a significant effect of the initial genotype on relative fitness at high temperature supporting the notion that tradeoff dynamics can vary based on genotypic composition even when the same gene is mutated but in different codon locations (Rodríguez-Verdugo et al. 2014).

Our results suggest that phenotype is influenced by contingency at the level of adaptive codon. In particular, Phase 2 evolved lines descended from the *rpoB* I966S genotype experienced the highest change in fitness relative to its ancestral genotype at 19.0°C, and the ancestral *rpoB* I966S genotype was also observed to have the lowest initial starting fitness (Table 2.1) and has been previously implicated as being associated with significant tradeoffs at 18.0°C (Rodríguez-Verdugo et al. 2014). Together this suggests that the initial phenotypic values at the level of individual genotype may be a predictor of evolutionary responses or rate of phenotypic change and is likely to be a more reliable predictor than just considering the identity of the mutated genes, therefore, following Fisher's geometric model (Fisher 1930).

Genotypic evolution. To characterize the influence of adaptive history on future genotypic change, we performed whole genome sequencing of the evolved populations and identified variants that arose during Phase 2 evolution to 19.0°C. We identified over 1,000 mutations occurring at a 5% frequency and 119 fixed mutations. There were no significant differences in the types of mutations that occurred between the adaptive pathways, suggesting that the adaptive history did not influence the rates of any particular mutation type that occurred in our system. The majority of mutations were in intergenic regions, which has been previously observed in other evolution experiments with *E. coli*, but we also observed that 44% of the mutations occurred in genes (Lenski 2017). We considered mutations in both genes and intergenic regions for our contingency analyses, because intergenic regions have been previously implicated as drivers for adaptation in bacteria (Khademi et al. 2019).

To examine evolutionary contingency with regard to genotype, we tested the association between the presence absence patterns of the Phase 2 mutations and the adaptive pathway or initial genotype. We found a significant association between the mutational pattern from Phase 2 and both the adaptive pathway and the initial genotype, suggesting that the mutations that arose during Phase 2 was contingent on the adaptive history. We noted that the ANOSIM R-statistic was higher for the effect of initial genotype (ANOSIM R = 0.240) compared to the R-statistic estimated for the effect of adaptive pathway (ANOSIM R = 0.139). This suggests that the initial genotype may have a greater influence on the mutations that arise during evolution than the general adaptive pathway.

The association tests revealed that the mutations that arose were influenced by historical contingency, which is in contrast to Plucain et al. (2016) but in agreement with Card et al. (2021). We identified six regions of the genome that were more frequently mutated in Phase 2 evolved populations descended from one adaptive history but not the other. Specifically, four regions had a significantly higher proportion of mutations in *rpoB* descended populations and two regions were associated with *rho* descended populations (Table 2.3). Intriguingly, mutations in *rpoC* and *rho* occurred more often in *rpoB* backgrounds, along with mutations in the gene *hepA*. The gene *rpoC*, similarly to *rpoB*, codes for beta subunit of RNA polymerase and was statistically more likely to be mutated in *rpoB* derived populations (Conrad et al. 2010; Trinh et al. 2006). The RNA polymerase associated protein *hepA* (also known as *rapA*), was also found to be mutated in evolved populations descended from the *rpoB* background, and this gene is a general transcription factor with ATPase activity (Sukhodolets et al. 2001). The genes *valY* and *lysV* are tRNA synthetases (Andersen et al. 1997; Ruan et al. 2011; Agrawal et al. 2014). Phase 2 evolved populations descended from *rho* backgrounds had significantly more mutations in the gene *ECB_01992* and within the intergenic region *ybcW/ECB_01526,* however, all have unknown function (Table 2.3).

It is interesting to note that lines descended from *rpoB* backgrounds were more likely to mutate genes that are key players in transcription, such as gaining a second mutation in *rpoB* or mutating other important genes like *rpoC* or *rho*, while *rho* descended lines were not enriched for these mutations. These observations together with our ANOSIM association test results suggests that the evolutionary history at the level of adaptive pathway (either *rho* or *rpoB*) influenced the types of genes mutated in the second phase of evolution. We note that Phase 2 evolved populations descended from the Phase 1 founder did not display any evidence of having mutations in *rpoB* or *rpoC*, further supporting our notion that mutating such key players in transcription during Phase 2 evolution was an adaptive strategy beneficial to *rpoB* descended lines and contingent on evolutionary history.

Evolutionary history influences evolutionary trajectories and genotypic change. We performed a sequential, two-phase evolution experiment using *E. coli* to investigate the influence of evolutionary history on evolutionary outcomes. We assessed both phenotypic and genotypic evolution by measuring relative fitness and performing whole genome sequencing of the evolved populations after evolution to two different environments. Our results suggest that contingency does significantly influence future evolution at the genotypic level, but that contingency may not be as important to influence phenotypic evolution in the case of fitness.

We saw no differences in phenotype (measured by relative fitness) based on the identity as either part of the *rho* or *rpoB* adaptive pathway, but we did see an influence of the initial genotype on phenotypic change. This suggests that evolutionary trajectories may be influenced by the initial genotypic composition, such that two genotypes with mutations in the same gene but different codons may experience different evolutionary trajectories or different rates of fitness change. With enough evolutionary time, those phenotypic differences are likely to converge, which explains the lack of a statistical difference between the average relative fitness of the two pathways. Measuring and tracking the relative fitness from the earlier generations warrants further research as we may disentangle the differences among evolutionary trajectories influenced by initial genotype.

The genotypic results displayed significant associations between the genetic changes that occurred in Phase 2 to the evolutionary history at the level of adaptive pathway and initial genotype. Similarly to the phenotypic results, we saw a stronger influence of contingency at the level of initial genotype. However, this observation may be less surprising for genotypic changes, as we may expect contingency to have a greater effect on mutational change due to genetic complexity such as pleiotropic effects and widespread epistatic interactions (Chou et al. 2014). We did observe particular regions of the genome that were enriched for mutations in either the *rho* or *rpoB* adaptive pathway, and we found that the mutations that were enriched could be found in evolved populations derived from two or more Phase 2 founders. For example, the gene *hepA* was more likely to be mutated in evolved populations derived from three different *rpoB* Phase 2 Founders: Line 3, Line 94, and Line 137 (Table 2.1). Here, the adaptive pathway (i.e. mutation in *rpoB*) did significantly influence the genetic changes in a new environment.

Evolutionary contingency has the potential to significantly influence evolutionary trajectories and outcomes which can affect our ability to forecast evolutionary change. Using experimental methods, we assessed the impact of contingency on genotypic and phenotypic evolution. Our results suggest that both genotypic and phenotypic changes are influenced by evolutionary history. Further experimentation over a gradient of selection pressures or in more diverse environments should be considered in order to disentangle the effects of contingency in conjunction with other evolutionary forces.

Figures

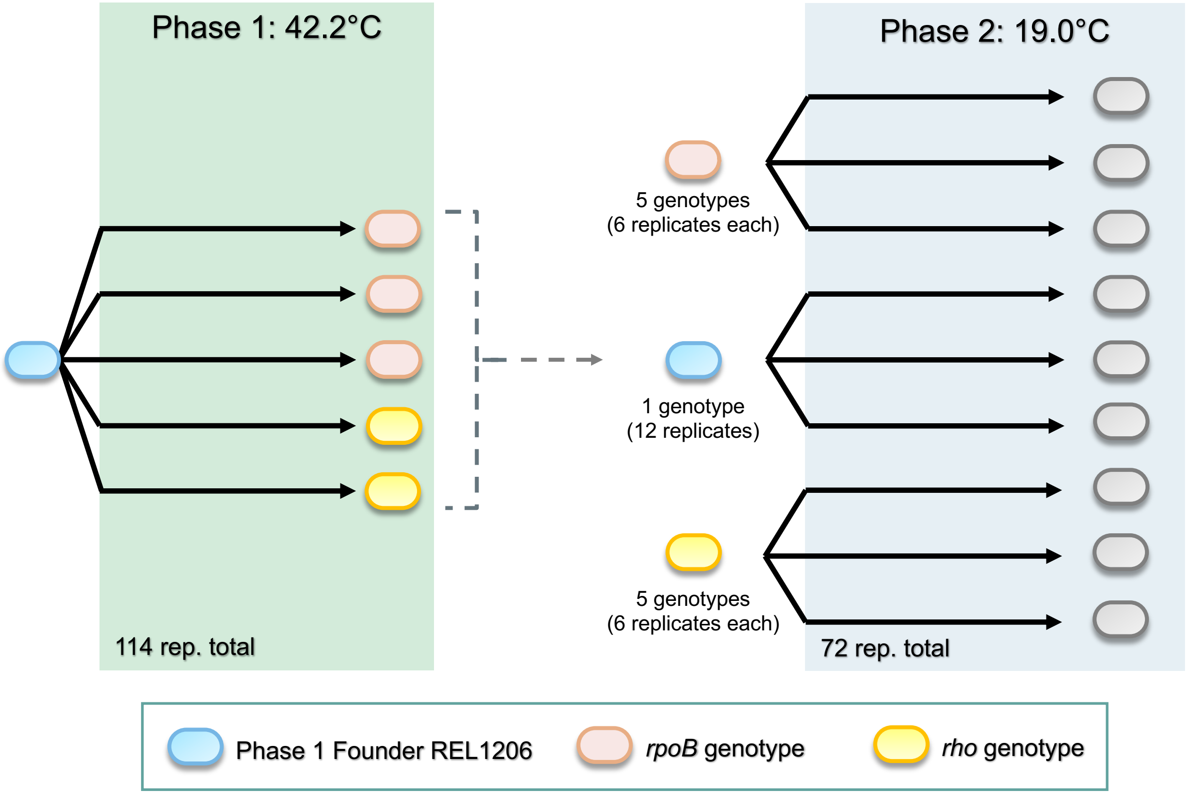

Figure 2.1: Sequential evolution experiment design to study contingency. The first phase of evolution was previously conducted and described in Tenaillon et al. (2012). The second phase of evolution was founded from a subset of evolved lines from Phase 1

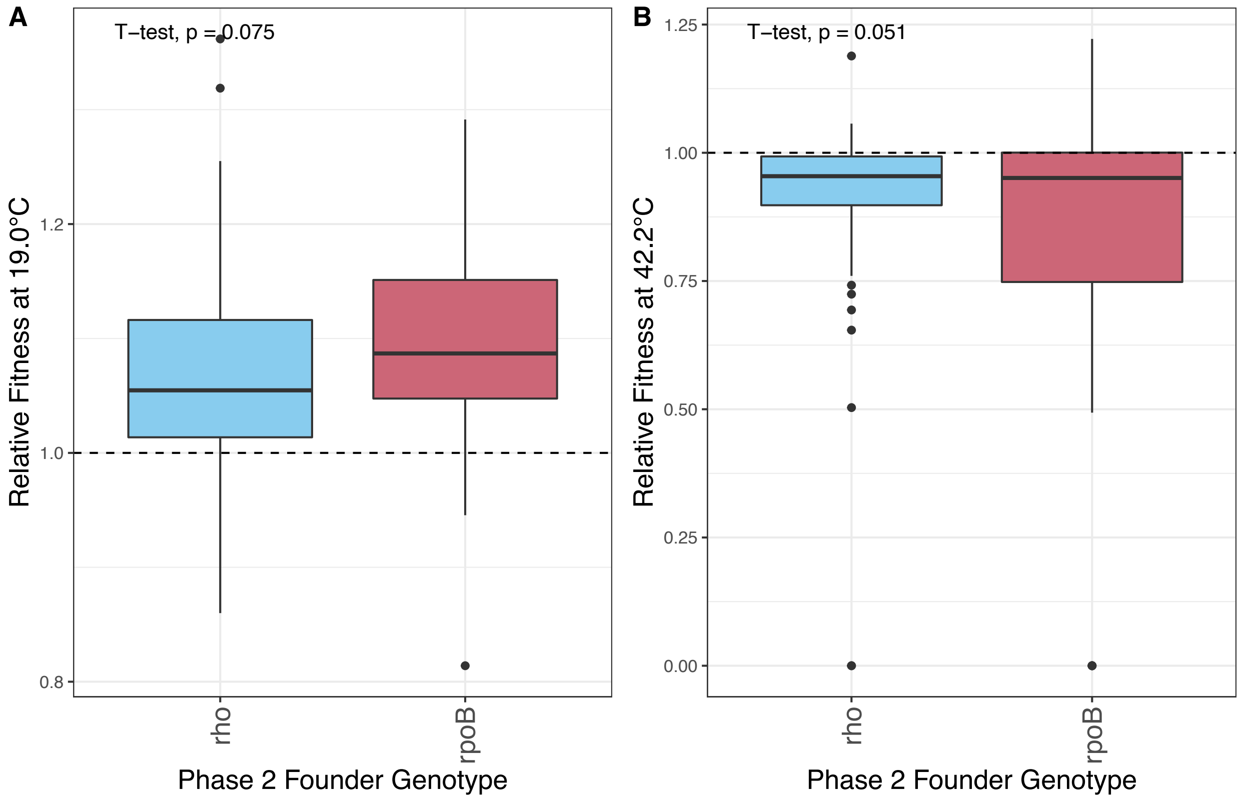

Figure 2.2: Relative fitness of evolved populations at two temperatures: A) 19.0°C and B) 42.2°C relative to their Phase 2 Founder.

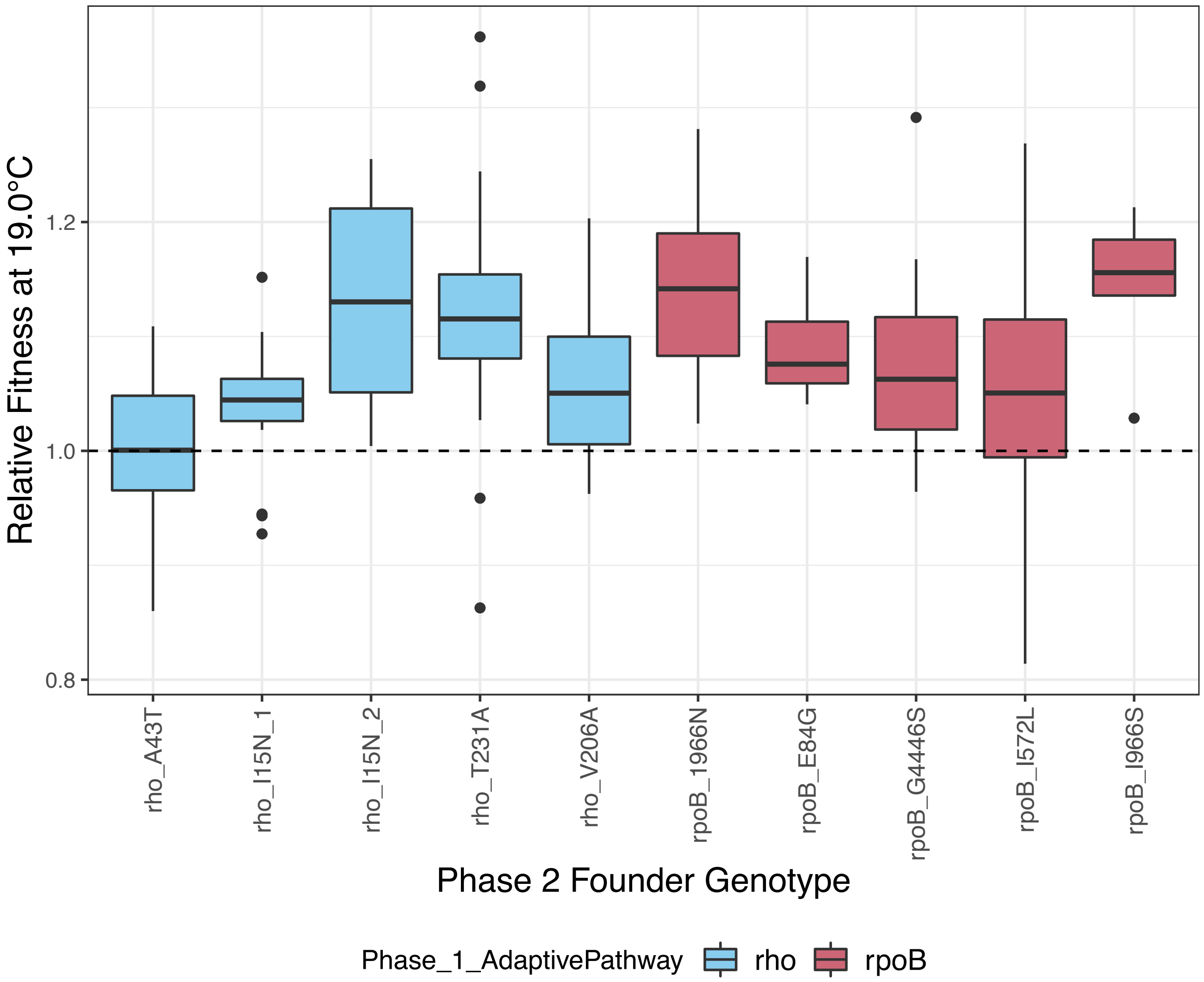

Figure 2.3: Phase 2 Evolved Populations Relative Fitness at 19.0°C. Assays were performed by directly competing a Phase 2 evolved population against their respective Phase 2 Founder.

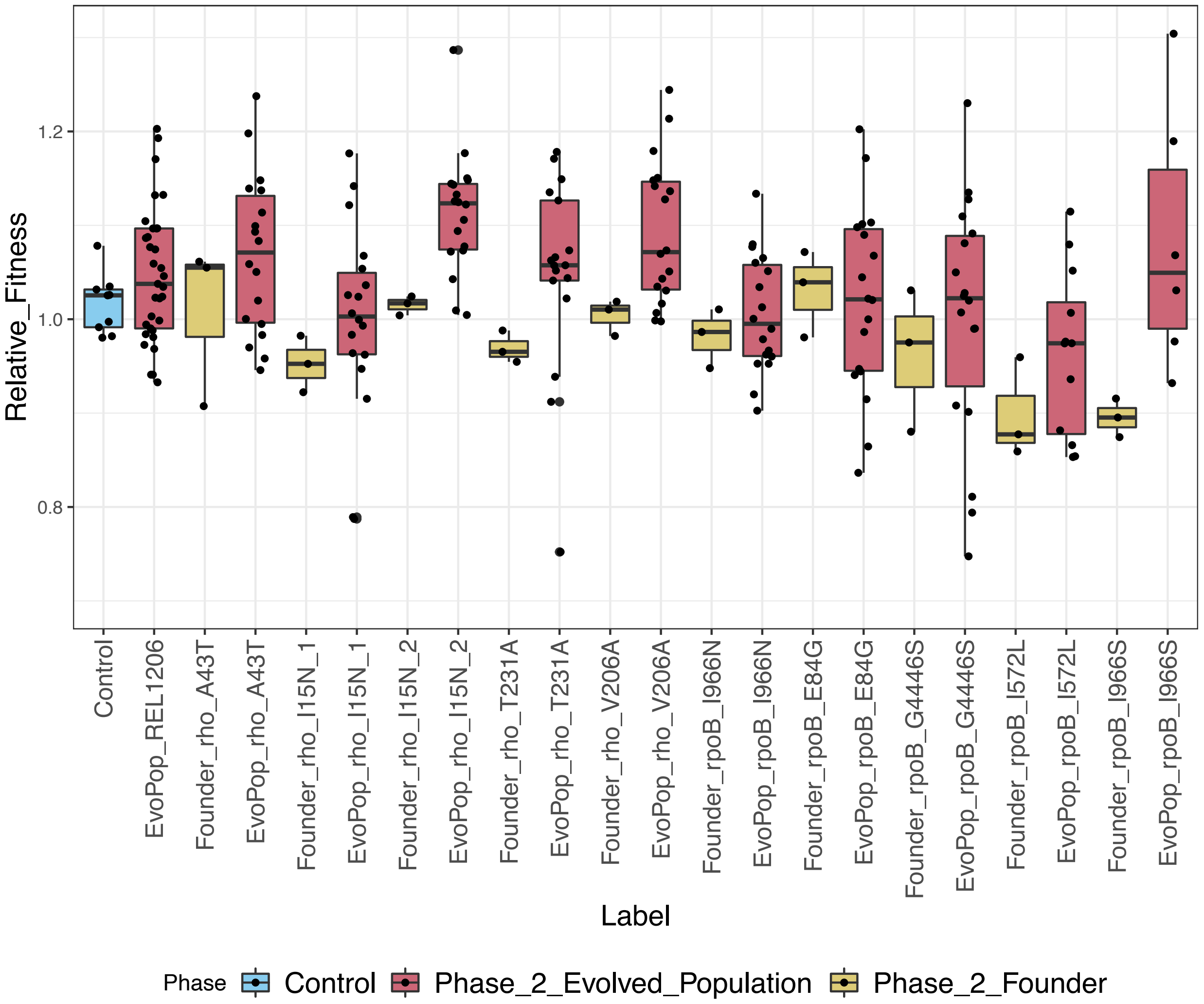

Figure 2.4: Relative fitness of Phase 2 Evolved Populations, Phase 2 Founders, and ancestral controls relative to the Phase 1 Founder genotype at 19.0°C.

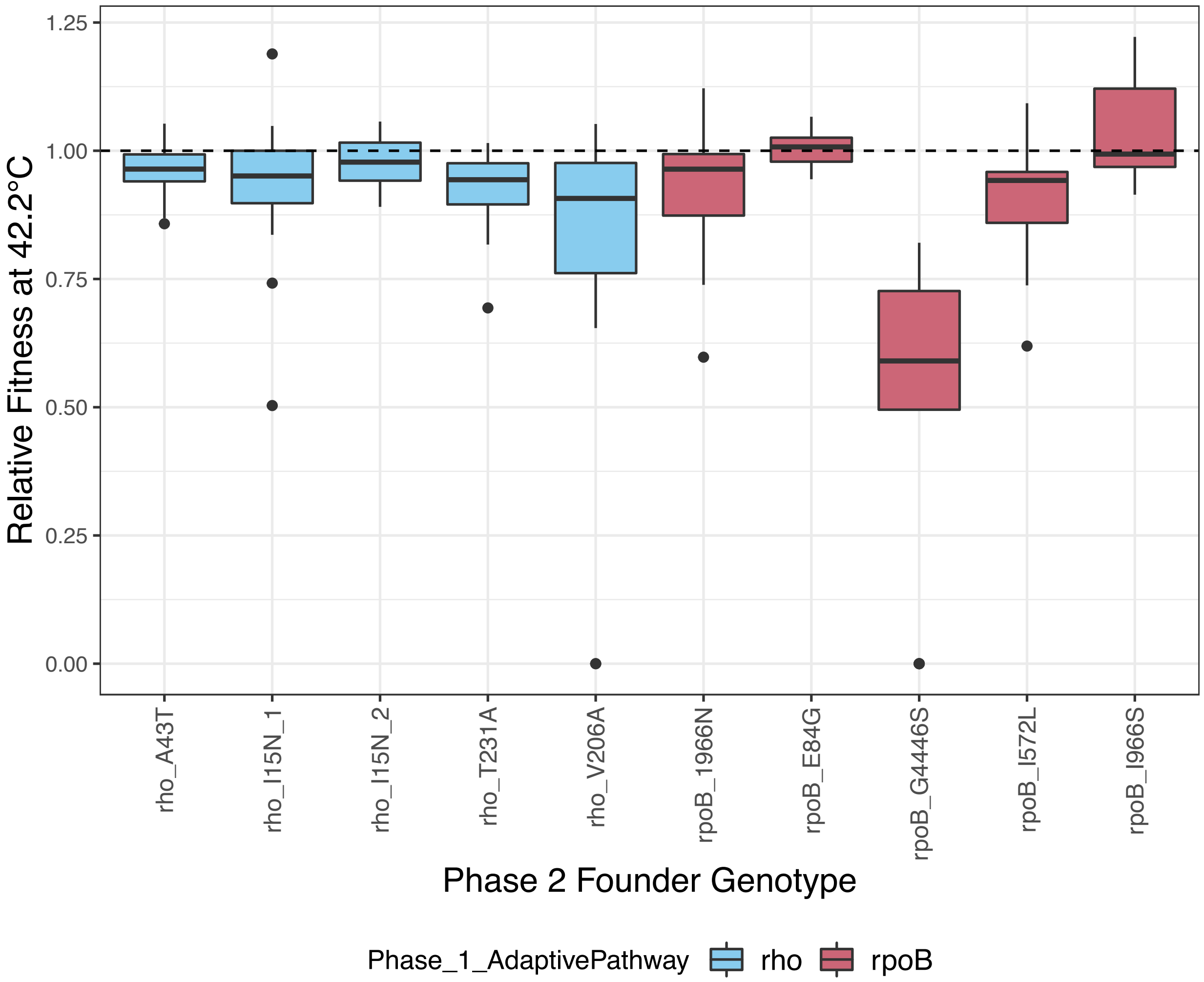

Figure 2.5: Phase 2 Evolved Populations Relative Fitness at 42.2°C. Assays were performed by directly competing a Phase 2 evolved population against their respective Phase 2 Founder.

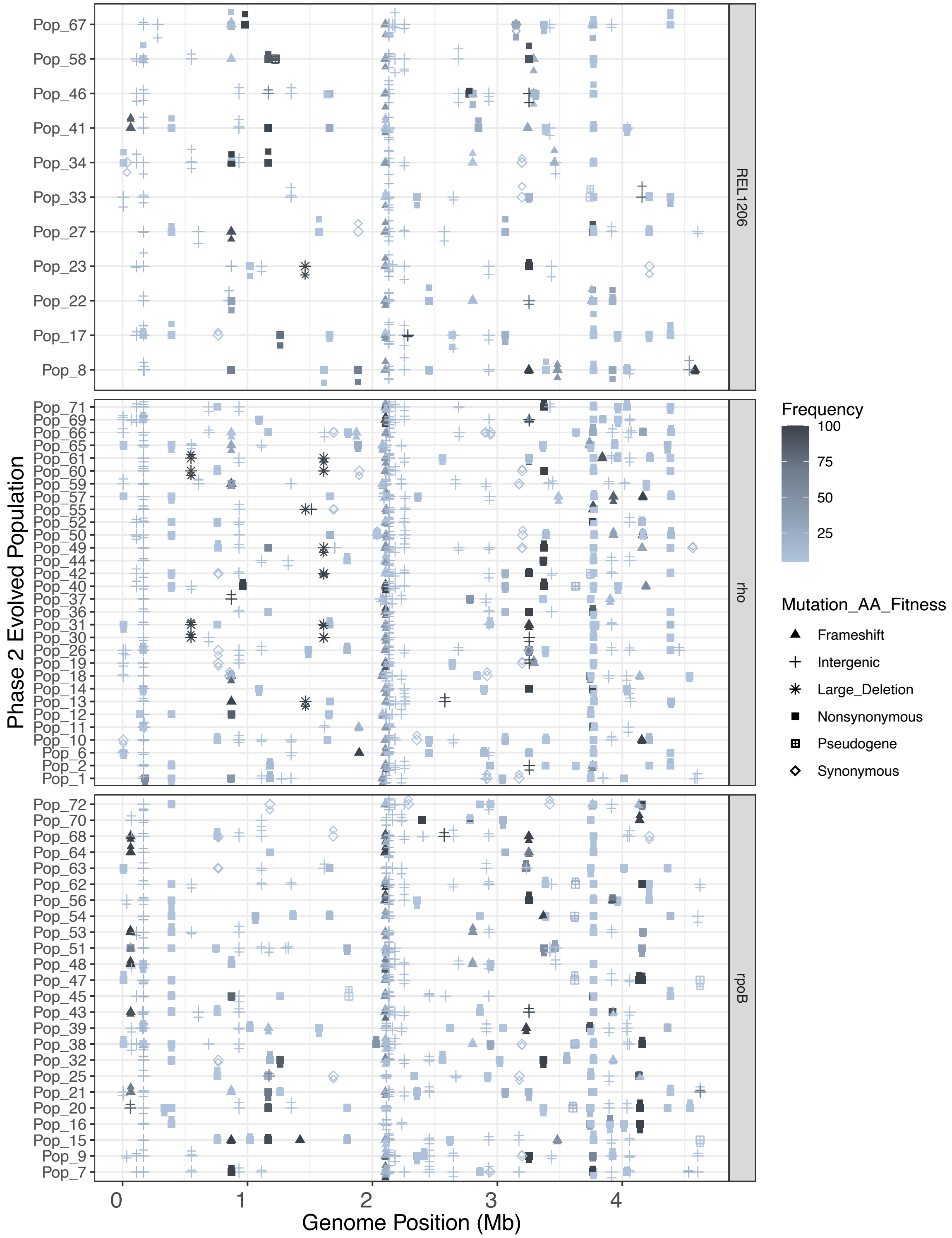

Figure 2.6: Mutations in Phase 2 Evolved Populations plotted along the genome. The graph is faceted by adaptive pathway and displays over 1,000 mutations identified using breseq in Phase 2 evolved populations. Only mutations occurring at a 5% frequency or higher are shown.

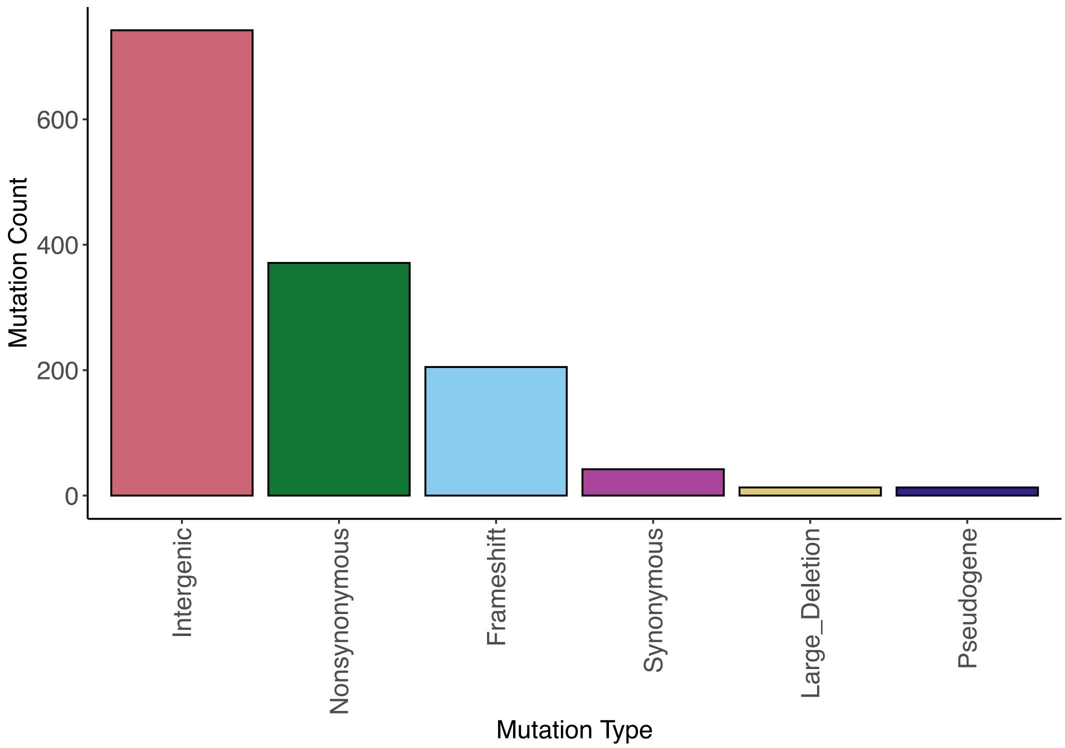

Figure 2.7: Number of mutations by type across all evolved populations. Mutations considered were intergenic, nonsynonymous, frameshift, synonymous, large deletions (deletions greater than 50 bp), and pseudogene mutations.

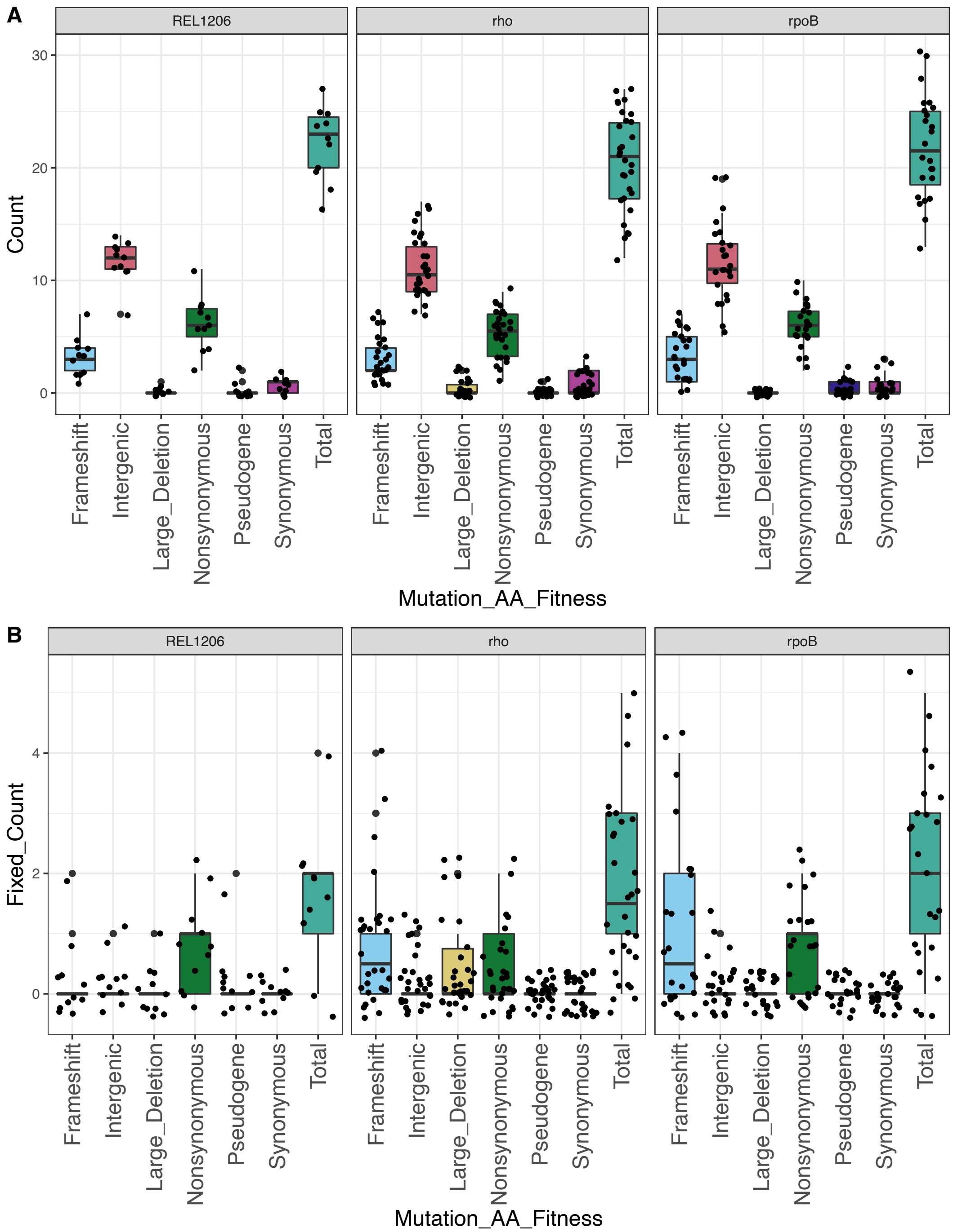

Figure 2.8: Types of mutations in Phase 2 Evolved Populations faceted by adaptive pathway. A) All mutations occurring at a 5% frequency or higher and B) fixed mutations (85% frequency or higher).

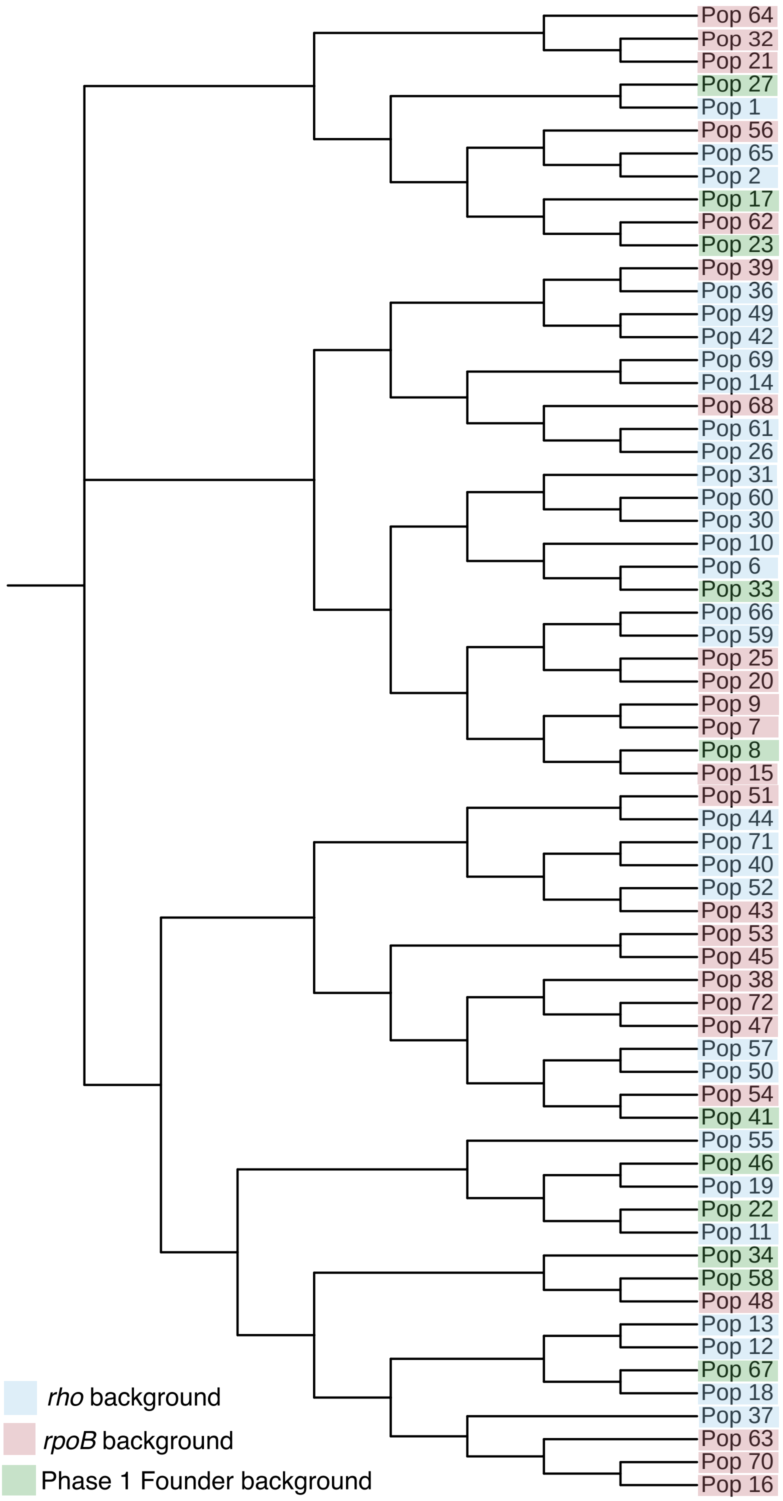

Figure 2.9: Neighbor-joining tree built from presence absence patterns of mutations that arose in Phase 2 evolved populations.

Tables

Table 2.1: Phase 2 Founder Adaptive History and Fitness

Phase 1 Evolved Line*	**Phase 1 Adaptive Pathway**	**Phase 1 Adaptive Pathway Genotype**	**Mean Absolute Fitness at 19.0°C[1]**	**Mean Relative Fitness at 19.0°C[2]**	**Mean Relative Fitness at 42.2°C[1]**	**Number of replicates**
2	*rho*	T231A	0.018	0.970	1.484	6
3	*rpoB*	I966S	0.022	0.895[3]	1.257	6
34	*rpoB*	G446S	-0.082	0.962	1.510	6
66	*rho*	V206A	0.053	1.004	1.430	6
82	*rho*	I15N_1	-0.003	0.952	1.498	6
87	*rho*	I15N_2	0.044	1.015	1.703	6
94	*rpoB*	E84G	0.070	1.031	1.767	6
134	*rho*	A43T	0.070	1.008	1.380	6
137	*rpoB*	I966N	0.046	0.982	1.349	6
142	*rpoB*	I527L	0.106	0.899	1.609	6
REL1206	Phase 1 Founder	NA	-0.004	1	1	12

* Line numbers designated in Tenaillon et al. 2012

[1] Relative fitness value data for 42.2°C and absolute fitness value data for 19.0°C originally measured in Rodríguez-Verdugo et al. 2014

[2] Relative fitness value data at 19.0°C obtained from this study

[3] Statistically significant for a fitness decline relative to competitor, t-test

Table 2.2: Relative fitness measurements for Phase 2 Evolved Populations at 19.0°C and 42.2°C

		Phase 2 Founder Competitor		**Phase 1 Founder Competitor**		**Phase 2 Founder Competitor**	
Phase 1 Evolved Line*	**Phase 1 Adaptive Codon Background**	**Average w_r (19.0°C)**	**P-value[1] (19.0°C)**	**Average w_r (19.0°C)**	**P-value[1] (19.0°C)**	**Average w_r (42.2°C)**	**P-value[1] (42.2°C)**
1	*rho* T231A	1.123	**< 0.01**	1.016	0.692	0.919	**< 0.01**
3	*rpoB* I966S	1.146	**< 0.01**	1.083	0.203	1.041	0.444
34	*rpoB* G4446S	1.078	**< 0.01**	1.003	0.934	0.500	**< 0.01**
66	*rho* V206A	1.056	**< 0.01**	1.092	**< 0.01**	0.795	**0.0124**
82	*rho* I15N_1	1.041	**< 0.01**	1.000	0.990	0.928	**0.0433**
87	*rho* I15N_2	1.129	**< 0.01**	1.113	**< 0.01**	0.977	0.0597
94	*rpoB* E84G	1.085	**< 0.01**	1.020	0.417	1.005	0.594
134	*rho* A43T	1.008	0.604	1.069	**< 0.01**	0.962	**< 0.01**
137	*rpoB* 1966N	1.140	**< 0.01**	1.006	0.711	0.923	**0.0204**
142	*rpoB* I572L	1.050	0.170	0.964	0.190	0.908	**0.0310**
NA	*REL1206*	NA	NA	1.047	**< 0.01**	NA	NA

* Line numbers designated in Tenaillon et al. 2012

[1] T-test, bolded p-values indicate statistically significant increase or decrease in relative fitness

Table 2.3: Genic or intergenic regions with evidence of contingency due to adaptive history

Gene or Intergenic Region	**Adaptive pathway**	**Fisher's Exact Test P-value**	**Adjusted P-value (FDR)**
nmpC/dsbG	*rpoB*	8.20E-05	0.0134
hepA	*rpoB*	0.00195	0.160
ECB_01992	*rho*	0.00521	0.2123
valY/lysV	*rpoB*	0.00521	0.212
rpoC	*rpoB*	0.0223	0.726
ybcW/ECB_01526	*rho*	0.0282	0.766
rho	*rpoB*	0.0336	0.782

Supplemental Information

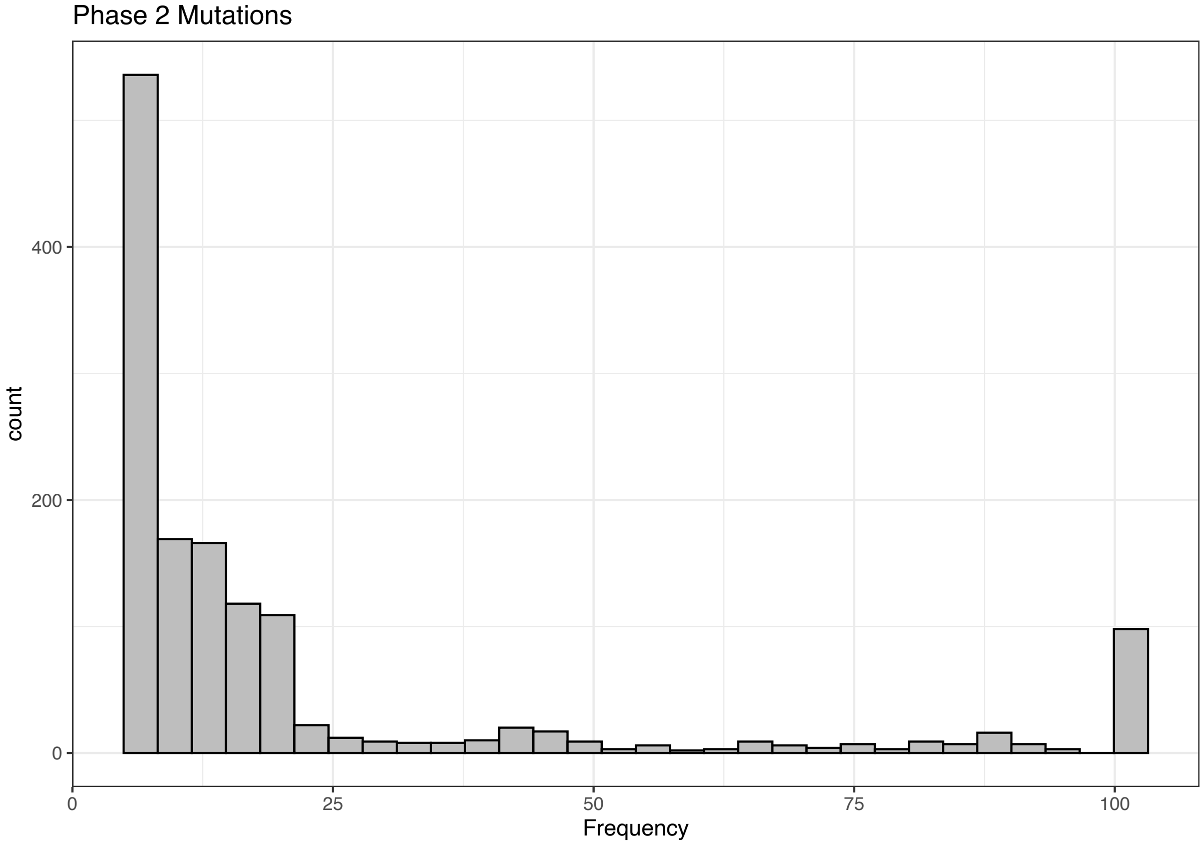

Figure S2.1: Distribution of mutation frequencies in Phase 2 evolved populations. Mutations that arose during Phase 2 and their frequencies based on percentage of individuals in the population carrying the mutation.

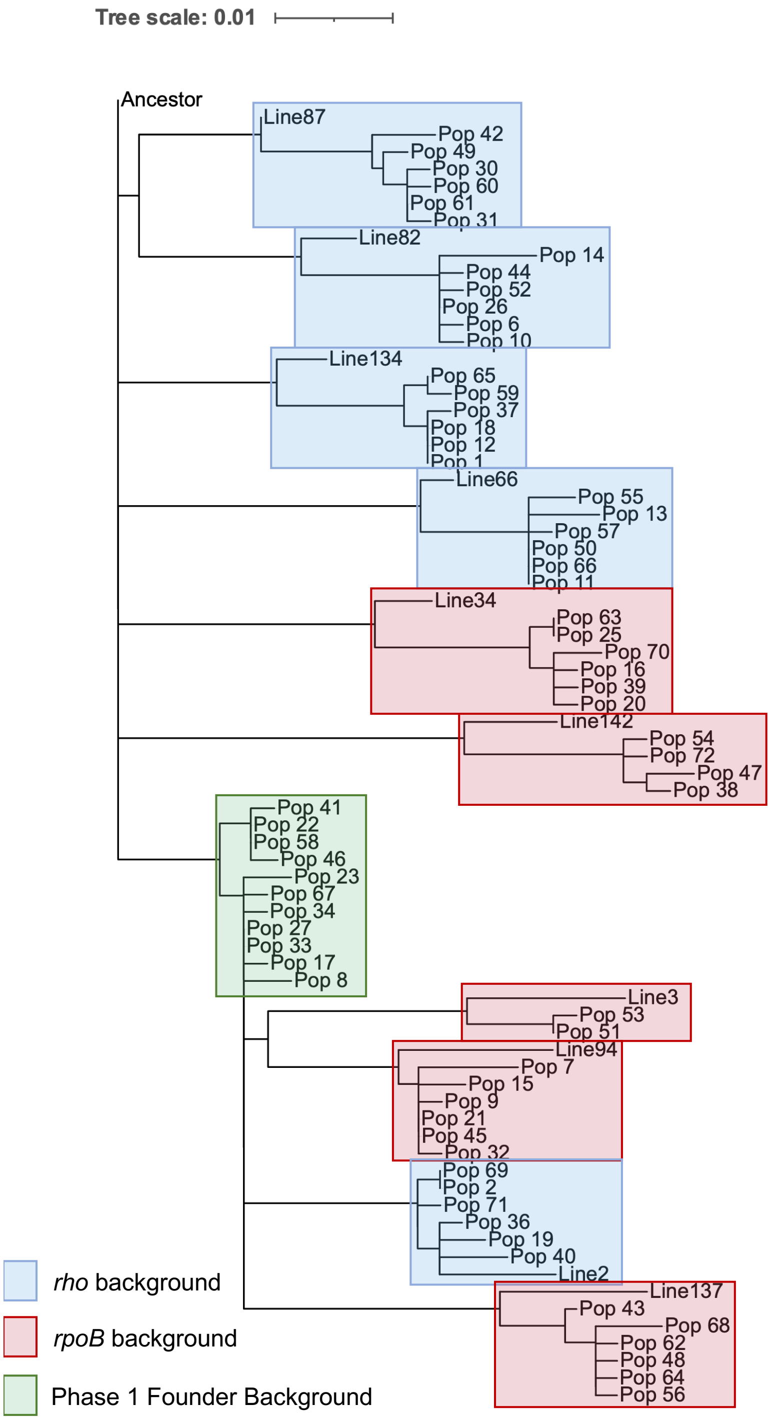

Figure S2.2: Phylogeny of Phase 1 Founder, Phase 2 Founders, and Phase 2 Evolved Populations. Phylogeny is built from genotypic information.

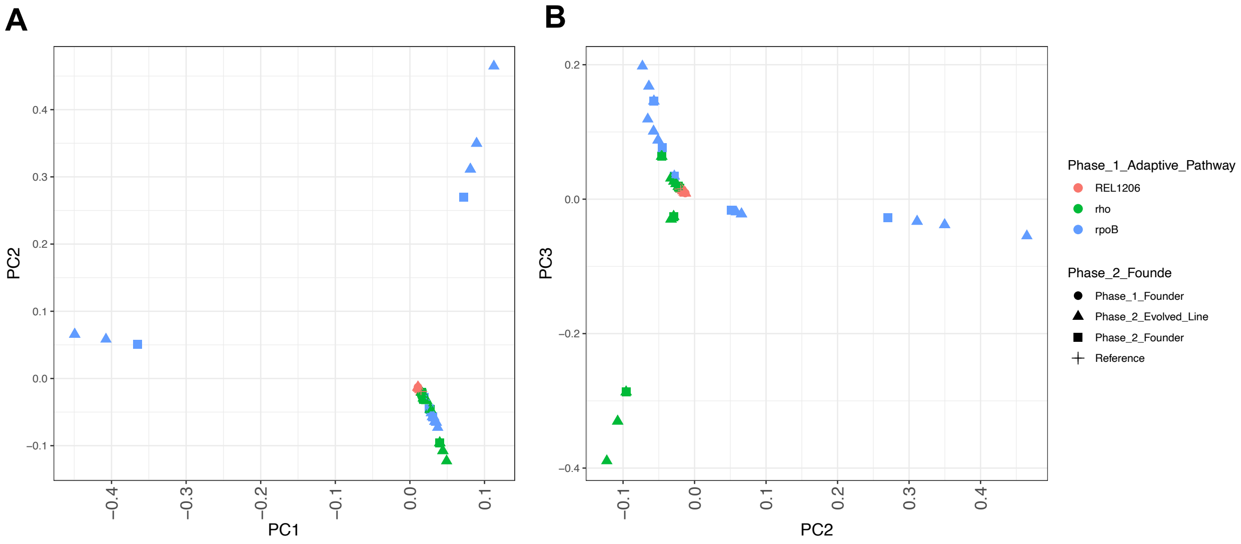

Figure S2.3: Principal components analysis based on genome mutations. A) PC1 and PC2 plotted together and B) PC2 and PC3 plotted together.

References

Agrawal A, Mohanty BK, Kushner SR. 2014. Processing of the seven valine tRNAs in Escherichia coli involves novel features of RNase P. Nucleic Acids Res. 42:11166–11179. doi: 10.1093/nar/gku758.

Andersen CL, Matthey-Dupraz A, Missiakas D, Raina S. 1997. A new Escherichia coli gene, dsbG, encodes a periplasmic protein involved in disulphide bond formation, required for recycling DsbA/DsbB and DsbC redox proteins. Mol. Microbiol. 26:121–132. doi: 10.1046/j.1365-2958.1997.5581925.x.

Barrick JE et al. 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 461:1243–1247. doi: 10.1038/nature08480.

Batarseh TN, Hug SM, Batarseh SN, Gaut BS. 2020. Genetic Mutations That Drive Evolutionary Rescue to Lethal Temperature in Escherichia coli. Genome Biol. Evol. 12:2029–2044. doi: 10.1093/gbe/evaa174.

Bay RA et al. 2017. Predicting Responses to Contemporary Environmental Change Using Evolutionary Response Architectures. Am. Nat. 189:463–473. doi: 10.1086/691233.

Bennett AF, Lenski RE. 1993. EVOLUTIONARY ADAPTATION TO TEMPERATURE II. THERMAL NICHES OF EXPERIMENTAL LINES OF ESCHERICHIA COLI. Evol. Int. J. Org. Evol. 47:1–12. doi: 10.1111/j.1558-5646.1993.tb01194.x.

Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc. Natl. Acad. Sci. 105:7899–7906. doi: 10.1073/pnas.0803151105.

Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life's tape. Science. 362. doi: 10.1126/science.aam5979.

Bobay L-M, Ochman H. 2017. The Evolution of Bacterial Genome Architecture. Front. Genet. 8. https://www.frontiersin.org/articles/10.3389/fgene.2017.00072 (Accessed July 19, 2022).

Card KJ, Thomas MD, Graves JL, Barrick JE, Lenski RE. 2021. Genomic evolution of antibiotic resistance is contingent on genetic background following a long-term experiment with Escherichia coli. Proc. Natl. Acad. Sci. 118:e2016886118. doi: 10.1073/pnas.2016886118.

Chang CC et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4:7. doi: 10.1186/s13742-015-0047-8.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34:i884–i890. doi: 10.1093/bioinformatics/bty560.

Chou H-H, Delaney NF, Draghi JA, Marx CJ. 2014. Mapping the Fitness Landscape of Gene Expression Uncovers the Cause of Antagonism and Sign Epistasis between Adaptive Mutations. PLOS Genet. 10:e1004149. doi: 10.1371/journal.pgen.1004149.

Conrad TM et al. 2010. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. Proc. Natl. Acad. Sci. U. S. A. 107:20500–20505. doi: 10.1073/pnas.0911253107.

Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory evolved microbes from next-generation sequencing data using breseq. Methods Mol. Biol. Clifton NJ. 1151:165–188. doi: 10.1007/978-1-4939-0554-6_12.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32:3047–3048. doi: 10.1093/bioinformatics/btw354.

Fisher RA. 1930. *The genetical theory of natural selection*. Oxford University Press: Oxford, UK.

González-González A, Hug SM, Rodríguez-Verdugo A, Patel JS, Gaut BS. 2017. Adaptive Mutations in RNA Polymerase and the Transcriptional Terminator Rho Have Similar Effects on Escherichia coli Gene Expression. Mol. Biol. Evol. 34:2839–2855. doi: 10.1093/molbev/msx216.

Khademi SMH, Sazinas P, Jelsbak L. 2019. Within-Host Adaptation Mediated by Intergenic Evolution in Pseudomonas aeruginosa. Genome Biol. Evol. 11:1385–1397. doi: 10.1093/gbe/evz083.

Kolbe JJ, Leal M, Schoener TW, Spiller DA, Losos JB. 2012. Founder Effects Persist Despite Adaptive Differentiation: A Field Experiment with Lizards. Science. 335:1086–1089. doi: 10.1126/science.1209566.

Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 344:1519–1522. doi: 10.1126/science.1250939.

Lenski RE. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. ISME J. 11:2181–2194. doi: 10.1038/ismej.2017.69.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. Am. Nat. 138:1315–1341.

Lenski RE, Travisano M. 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. 91:6808–6814. doi: 10.1073/pnas.91.15.6808.

Leray M et al. 2021. Natural experiments and long-term monitoring are critical to understand and predict marine host–microbe ecology and evolution. PLOS Biol. 19:e3001322. doi: 10.1371/journal.pbio.3001322.

Losos JB. 2011. Convergence, Adaptation, and Constraint. Evolution. 65:1827–1840. doi: 10.1111/j.1558-5646.2011.01289.x.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32:268–274. doi: 10.1093/molbev/msu300.

Oksanen J et al. 2020. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528. doi: 10.1093/bioinformatics/bty633.

Plucain J et al. 2016. Contrasting effects of historical contingency on phenotypic and genomic trajectories during a two-step evolution experiment with bacteria. BMC Evol. Biol. 16:86. doi: 10.1186/s12862-016-0662-8.

R Core Team. 2019. R: A language and environment for statistical computing. https://www.R-project.org/.

Reznick DN, Bryga H. 1987. Life-History Evolution in Guppies (poecilia Reticulata): 1. Phenotypic and Genetic Changes in an Introduction Experiment. Evolution. 41:1370–1385. doi: https://doi.org/10.1111/j.1558-5646.1987.tb02474.x.

Rodríguez-Verdugo A, Carrillo-Cisneros D, González-González A, Gaut BS, Bennett AF. 2014. Different tradeoffs result from alternate genetic adaptations to a common environment. Proc. Natl. Acad. Sci. 111:12121–12126. doi: 10.1073/pnas.1406886111.

Ruan L, Pleitner A, Gänzle MG, McMullen LM. 2011. Solute transport proteins and the outer membrane protein NmpC contribute to heat resistance of Escherichia coli AW1.7. Appl. Environ. Microbiol. 77:2961–2967. doi: 10.1128/AEM.01930-10.

Seemann T. 2022. Snippy. https://github.com/tseemann/snippy (Accessed May 16, 2022).

Sukhodolets MV, Cabrera JE, Zhi H, Jin DJ. 2001. RapA, a bacterial homolog of SWI2/SNF2, stimulates RNA polymerase recycling in transcription. Genes Dev. 15:3330–3341. doi: 10.1101/gad.936701.

Tenaillon O et al. 2012. The Molecular Diversity of Adaptive Convergence. Science. 335:457–461. doi: 10.1126/science.1212986.

Trinh V, Langelier M-F, Archambault J, Coulombe B. 2006. Structural Perspective on Mutations Affecting the Function of Multisubunit RNA Polymerases. Microbiol. Mol. Biol. Rev. 70:12–36. doi: 10.1128/MMBR.70.1.12-36.2006.

Vlachostergios PJ, Faltas BM. 2018. Treatment resistance in urothelial carcinoma: an evolutionary perspective. Nat. Rev. Clin. Oncol. 15:495–509. doi: 10.1038/s41571-018-0026-y.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-Term Dynamics of Adaptation in Asexual Populations. Science. https://www.science.org/doi/abs/10.1126/science.1243357 (Accessed September 7, 2021).

CHAPTER 3

Using genomes and evolutionary analyses to screen for host-specificity and positive selection in the plant pathogen *Xylella fastidiosa*

3.1 Abstract

Xylella fastidiosa infects several economically important crops in the Americas, and it also recently emerged in Europe. Here, using a set of *Xylella* genomes reflective of the genus-wide diversity, we performed a pan-genome analysis based on both core and accessory genes, for two purposes: i) to test associations between genetic divergence and plant host species and ii) to identify positively selected genes that are potentially involved in arms-race dynamics. For the former, tests yielded significant evidence for specialization of *X. fastidiosa* to plant host species. This observation contributes to a growing literature suggesting that the phylogenetic history of *X. fastidiosa* lineages affects host range. For the latter, our analyses uncovered evidence of positive selection across codons for 5.3% (67 of 1,257) of core genes and 5.4% (201 of 3,691) of accessory genes; these genes are candidates to encode interacting factors with plant and insect hosts. Most of these genes had unknown functions, but we identified some tractable candidates including *nagZ_2*, which encodes a beta-glucosidase that is important for *Neisseria gonorrhoeae* biofilm formation; *cya*, which modulates gene expression in pathogenic bacteria; and *barA*, a membrane associated histidine kinase that has roles in cell division, metabolism, and pili formation.

3.2 Introduction

Bacteria exhibit extensive intraspecific variation in genome content. This variation is the raw material for evolutionary adaptation, including the evolution of pathogenicity and virulence (Furuya & Lowy 2006; Yacoubi et al. 2007; Juhas 2015; Chen et al. 2018). One example of genome variation comes from an early study of *Escherichia coli* that compared two pathogenic strains and one non-pathogenic laboratory strain (Welch et al. 2002). Of the entire set of protein coding genes annotated by the three genomes, only 39.2% were shared among the three isolates. Intriguingly, the two pathogenic strains each had 1,300 unique genes, while the laboratory strain had only 585, suggesting that genes that vary across accessions (i.e., accessory genes) contribute to virulence. Similar patterns have been illustrated for plant pathogens (Badet & Croll 2020; Kim et al. 2020). In *Xanthomonas*, for example, horizontal gene transfer (HGT) has shuffled virulent accessory genes from pathogenic strains to previously non-pathogenic strains (Chen et al. 2018), facilitating the infection of common bean (*Phaseolus vulgaris* L.). In short, accessory genes contribute to host-pathogen interactions, making them a critical focus for comparative analyses of genome evolution and function.

Here we investigate variation in the genome content of another plant pathogen. *Xylella fastidiosa* is endemic to the Americas and was first identified as the causal agent of Pierce's Disease (PD), an economically devastating disease in grapevines (*Vitis vinifera* ssp. *vinifera*) (Sicard et al. 2018; Burbank & Roper 2021). *X fastidiosa* causes additional economically and ecologically impactful diseases like citrus variegated chlorosis, coffee leaf scorch, oak leaf scorch and elm leaf scorch, among others. Historically, the geographic distribution of *X. fastidiosa* was limited to the Americas, but it was recently introduced to the European continent by anthropogenic transmission, which has further expanded its host range and led to emerging diseases like olive quick decline syndrome (OQDS) in Italy (Schuenzel et al. 2005; Loconsole et al. 2016). *X. fastidiosa* has since been detected in various plants species across locations in Europe including France, Spain, and Portugal (Chatterjee et al. 2008; Rapicavoli et al. 2018). In susceptible hosts, *X. fastidiosa* can lead to significant crop losses, and it continues to threaten crops globally (Tumber et al. 2014; Alston et al. 2015).

For each of these diseases, *X. fastidiosa* is transmitted by xylem-feeding insect vectors into the plant host, where it then utilizes cell wall degrading enzymes to systemically colonize the xylem. In the xylem, it forms biofilms that are thought to be integral to pathogenicity (Koo et al. 2017; Castro et al. 2021). Colonization is also governed, in part, by virulence and pathogenicity factors that influence a wide range of bacterial functions – e.g., biofilm formation, host cell wall degradation, regulatory systems, stress responses and bacterial membrane composition -- although other abiotic factors (like plant drought stress) likely also contribute to disease progression (Rapicavoli et al. 2018). Given its economic impact, the effects and mechanisms of *X. fastidiosa* infection have been studied widely, especially in grapevine (Roper & Lindow 2016). However, many pathogenicity factors likely remain undiscovered, and crucial questions remain unanswered about the genetic factors that govern host-pathogen interactions and potential host specialization (Rapicavoli et al. 2018).

In this context, it is helpful to recognize that *X. fastidiosa* consists of three commonly recognized subspecies that form distinct phylogenetic clades: ssp. *fastidiosa*, *multiplex*, and *pauca.* Each subspecies has unique phenotypic characteristics and DNA markers

(Marcelletti & Scortichini 2016). Two other subspecies, *morus* and *sandyi*, have also been suggested, though they are not recognized as broadly (Burbank & Roper 2021); indeed, *morus* is believed to be a product of a recombination event between *fastidiosa* and *multiplex* isolates (Sicard et al. 2018). The recognition of subspecies is critical, because initial work suggested that subspecies correlate with specific plant hosts (Nunney et al. 2013). While it has long been known that that genetic differences among strains facilitate host-plant specialization (Almeida & Purcell 2003; Hernandez-Martinez et al. 2006; Almeida et al. 2008; Roper & Lindow 2016), there is not a clear one-to-one correspondence between pathogen and host. For example, some strains can infect more than one host species, as demonstrated by a strain that causes PD in grapevines and also leaf scorch in almonds (Almeida & Purcell 2003). Consequently, the questions of the evolution of and determinants of host specificity are still central for understanding the distribution and effects of this pathogen.

In this study, we analyze *X. fastidiosa* genome evolution among isolates from different plant hosts. Our study is not unique in some respects, because numerous comparative genomic studies of *X. fastidiosa* have been published already. Many of these studies have focused on clarifying phylogenetic relationships. For example, Marcelletti and Scortichini (2016) studied 21 genomes to resolve taxonomic relationships among subspecies; Giampetruzzi et al. 2017 extended sampling to 27 genomes, in part to place a novel strain (ST53) in the broader *X. fastidiosa* phylogeny; and Denancé et al. (2019) used kmers from 46 genomes to untangle species and subspecies relationships. Another recent study compared *X. fastidiosa* populations from Central/South America (Costa Rica, Brazil), North America (California, Southeastern US), Europe (Spain, Italy), and Asia (Taiwan) to elucidate the evolutionary origins of the subsp. *fastidiosa* and *pauca* (Castillo et al. 2021). Still other studies have focused on populations. For example, Vanhove et al. (2020) isolated and sequenced *X. fastidiosa* subsp. *fastidiosa* from symptomatic grapevines from five different California locations (Vanhove et al. 2020).

One common theme of genomic studies is that they identify the set of genes that are present in most samples (i.e., core genes) and used those genes as the basis to perform phylogenetic inference. These phylogenies have been used for various purposes. For example, two recent papers have used phylogenies to explore the question of host specificity. In one, Uceda-Campos et al. (2022) found that *X. fastidiosa* isolates grouped on the phylogeny by geography, but not by plant host species, suggesting host specificity is not correlated with phylogenetic relationships and genetic divergence (Uceda-Campos et al. 2022). In contrast, Kahn and Almeida (2022) used the phylogeny to infer the ancestral character states of plant hosts and found that the ancestral host plant could be inferred for most ancestral nodes (Kahn & Almeida 2022). They concluded that genetic history affects host range and also identified ~30 genes whose presence/absence correlated with specific plant hosts.

In this study, we combined 20 new *X. fastidiosa* genomes with publicly available data to build a dataset for molecular evolutionary analysis and to investigate patterns of host specificity in a phylogenetic context. For host-specificity analyses, we focused on core genes, but we also assessed the phylogenetic signal, pattern of gene gain and loss, and potential host associations of accessory (i.e., non-core) genes. Our goals for these analyses were to add to the growing literature about genetic correlations between phylogenetic history and host specificity but also to further consider the dynamic evolution of accessory genes in this context (Kahn & Almeida 2022). In addition, we performed extensive analyses of the ratio of nonsynonymous to synonymous (dN/dS or ω) substitutions to identify genes under positive selection (ω > 1.0). Genes under positive selection may be involved in arms-race (or Red-Queen) dynamics between pathogens and hosts (Daugherty & Malik 2012; Aleru & Barber 2020). In other systems, ω analyses have identified genes with functions that contribute to host defense and also discovered entirely new sets of genes and pathways involved in pathogen-host interactions (Mitchell et al. 2012; Ng et al. 2015; Daugherty et al. 2016). Here we apply tests for positive selection in the hope of gaining insights into the sets of genes that may affects host-pathogen interactions.

3.3 Methods

Novel *X. fastidiosa* genomes. Fully extracted DNA from 20 *X. fastidiosa* isolates were provided by the French Collection of Plant-Associated Bacteria (CIRM-CFBP; http://www6.inra.fr/cirm_eng/CFBP-Plant-Associated-Bacteria) and from the University of California, Riverside. Genomic DNA was prepared for Illumina sequencing using the Illumina Nextera DNA Flex Library Prep kit, following the manufacturer's recommendations and for Pacific Biosciences (PacBio) sequencing with the SMRTbell Express Template Prep Kit 2.0. SMRTbell libraries had 10kb DNA target insert size (Pacific BioSciences, Menlo Park, CA) using 360ng of sheared DNA as input. DNA libraries were sequenced with both Illumina and PacBio technologies at the University of California, Irvine Genomics High Throughput Facility (https://ghtf.biochem.uci.edu). The Illumina sequencing reads were quality assessed using FastQC, and reads were trimmed using Trimmomatic v. 0.32 (Andrews 2010; Bolger et al. 2014) using default options. PacBio sequencing reads were corrected and trimmed using Canu v. 1.5 (Koren et al. 2017). The long and short reads were used for genome assembly with Unicycler v. 0.4.8 in hybrid assembly mode (Wick et al. 2017). Genome assembly statistics were calculated using Quast v. 5.0.2 (Gurevich et al. 2013). As is common practice (Chase et al. 2018), short contigs (<500 bp) were removed from the assemblies using Seqkit v. 0.13.2 (Shen et al. 2016).

Genome Assembly of public data and sample set curation. We complemented our set of novel genomes with publicly available data. To do so, we downloaded all available whole genome assemblies of *X. fastidiosa* and *X. taiwanensis* (as an outgroup) from the National Center for Biotechnology Information (NCBI) and Sequence Read Archive (SRA) databases on July 9, 2020 (Supplementary Table S3.1). In addition, we downloaded the raw, short-read sequences for an additional 20 isolates (Castillo et al. 2020; Vanhove et al. 2020). For each isolate, we gathered information about its geographic origin and its host plant from NCBI and from the Pathosystems Resource Integration Center (PATRIC) database. To assemble the raw reads from the 20 unassembled accessions into genomes, we assessed quality and trimmed the reads and applied SPAdes v. 3.14.0 (Bankevich et al. 2012) with the --*careful* option, following Vanhove et al. (Supplementary Table S3.2; 2020). If long reads were also available, as they were for 5 isolates from the work of Castillo et al. 2020, then whole genome assembly was performed with Unicycler v. 0.4.8 in hybrid assembly mode (Wick et al. 2017).

In total, we gathered and generated 148 *Xylella* genome assemblies. From this set, we removed isolates that did not have information about their host isolation source or were lab-derived recombinant strains. The remaining 129 genomes were re-annotated by the same method - based on Prokka v. 1.14.6 analysis – that we applied to the new genomes, to ensure homogeneity. The Prokka analyses were then input into Roary v. 3.13.0 with options *-i 80 -cd 100 -e -n -z* to obtain a core gene alignment for initial comparisons among isolates (Seemann 2014; Page et al. 2015); we defined core genes as those that were detectable in 100% of the samples. This core set was aligned with MAFFT and polished using gBlocks v. 0.91b (Castresana 2000; Katoh et al. 2002; Katoh & Standley 2013). The polished alignment was used as input for RAxML v. 8.2.12 to build a preliminary phylogenetic tree (Stamatakis 2014), which we used to evaluate and curate the isolates (Supplementary Figure S3.1).

To curate the dataset, we created a distance matrix from the RAxML phylogenetic tree, using the Tree and reticulogram REConstruction (T-REX) server (Boc et al. 2012). Many of the genomes – most of which were gathered for population genomic analyses - were sampled from the same plant host and were nearly identical genetically. To limit sampling biases for our species-wide study, we removed clones and near-clones based on the distance matrix. That is, if two or more isolates had a pairwise distance ≤ 0.0001 and came from the same host, we retained the isolate with the more contiguous assembly. We also used CheckM (Parks et al. 2015) to assess genome completeness based on a set of conserved single copy genes (Supplementary Table S3.3). After applying these filters, our final dataset consisted of 63 *X. fastidiosa* genomes and one outgroup genome (*X. taiwanensis* PLS229) that were isolated from 23 distinct plant host species (Supplementary Table S3.1).

Pan-genome analysis. To perform a pan-genome analysis, we applied Roary to the 64 *Xylella* genomes using *gff* files from Prokka as input. Roary was applied with the option *-i 80,* as used in previous microbial studies (Chase et al. 2018; Rodriguez & Martiny 2020), to lower the blastp sequence identity to 80% from the default 95%. We defined a core gene as a gene present in 95% of the isolates used in the analysis (i.e., a core gene was present in at least 60 of 63 *X. fastidiosa* accessions). From the Roary output, we extracted a representative nucleotide sequence of each core and accessory gene using cdbfasta (https://github.com/gpertea/cdbfasta) and translated the nucleotide sequence to amino acids using the transeq command from EMBL-EBI (Madeira et al. 2019). The representative sequences were the basis for functional categorization -- using the eggNOG-mapper v. 2 (Huerta-Cepas et al. 2017, 2019) -- of both core and accessory genes. Gene Ontology (GO) enrichment analyses were performed online at (http://geneontology.org) using *Xanthomonas campestris* as the reference list (Ashburner et al. 2000). To explore function further, we also used the Conserved Domain Database online tool (https://www.ncbi.nlm.nih.gov/cdd/) to identify protein domains.

Phylogenetic Tree Construction. We used the core gene alignment from Roary to build a phylogenetic tree, based on a subset of genes that were present in all 63 *X. fastidiosa* samples and the *X. taiwenensis* outgroup. To do so, we curated the alignments with gBlocks v. 0.91b (Castresana 2000), used the polished alignment as input for IQtree v. 2.0.3, and selected the best nucleotide model for phylogenetic tree construction (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). We ultimately constructed an unrooted tree using the GTR+F+R8 model with RAxML (Stamatakis 2014), using the 'best tree' option. Phylogenetic trees were visualized and annotated using the ape package v. 5.5 in R v. 4.0.2 (Paradis & Schliep 2019; R Core Team 2019). We used the most likely phylogeny to test associations between phylogenetic relatedness, geography, and host isolation source (plant taxonomic order information taken from https://www.itis.gov/) with ANOSIM implemented in the vegan package v. 2.5-7 in R (Oksanen et al. 2020).

X. fastidiosa is naturally transformant and undergoes homologous recombination (Burbank & Roper 2021; Kung & Almeida 2011), but recombined genomic regions can obscure vertical phylogenetic relationships. To account for potential recombination among *X. fastidiosa* genomes, we applied Gubbins v. 3.2.1 (Croucher et al. 2015; Shikov et al. 2022), using again the subset of genes that were found in all 64 samples. From this input, Gubbins identified regions that were likely to have undergone recombination and removed them from the alignment. We then built a phylogeny from this recombination-adjusted core gene alignment using RAxML, as described above. We assessed the congruence between the two phylogenetic trees (i.e., with and without removal of potentially recombining regions) using phytools v. 1.0-1 in R (Revell 2012).

Finally, we also built a Neighbor-Joining (NJ) tree based on the presence-absence matrix of accessory genes. We first calculated the Euclidean distances from the presence-absence matrix of the accessory genes using the *dist* function in R (Mateo-Estrada et al. 2019). We then built an NJ tree from the Euclidean distances using the ape package in R (Paradis & Schliep 2019). We also utilized the ANOSIM and Mantel test (in the *vegan* package) to measure the correlation between accessory gene content and phylogenetic relatedness. The Mantel test required two distance matrices, which were the Euclidean distances estimated from the accessory gene presence-absence matrix and the distances from the RAxML core gene phylogeny generated by the reticulogram REConstruction (T-REX) server (Boc et al. 2012).

Gain and Loss of Accessory Genes. We utilized GLOOME to investigate gene gain and loss dynamics along the core phylogenetic tree of *X. fastidiosa* (Cohen et al. 2010). GLOOME uses a mixture-model approach, coupled with maximum-likelihood inference, to infer rates of gain and loss of genes along the branches of a phylogeny. It takes as input the phylogenetic topology, in this case the phylogenetic topology based on core genes, and a presence-absence matrix of genes. The pattern of genic presence and absence was obtained through M1CR0B1AL1Z3R, as recommended by the GLOOME authors, and then directly input into GLOOME using default settings (Avram et al. 2019). The default settings included a fixed rate of gene gains and losses with gamma distributed rates across genes (or sites). Among the output, GLOOME returned two phylogenetic trees with branch lengths representing either the number of expected gain events or the number loss events on each branch. As recommended (Cohen et al. 2010), branch lengths representing relative gain and loss rates were extracted from the phylogenetic trees using FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). To normalize expected gain (or loss) events with sequence divergence, we calculated the ratio of inferred gain (or loss) against the branch lengths of the sequence-based core phylogeny. Outlier branches with excess normalized gains or losses were identified using the interquartile range criterion.

Positive selection analyses. We employed codeml from PAML v. 4.9 to calculate ω, the ratio of nonsynonymous to synonymous rates (Yang 1997, 2007). We performed *codeml* analysis on nucleotide alignments of the single-copy core genes, single-copy accessory genes, and multicopy genes (defined as genes with two or more copies in a single accession). For all tests, we required at least four sequences, as the minimum number suggested for *codeml* analysis (http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf). For each gene and sequence set, we ran analyses by generating an unrooted maximum-likelihood tree for each gene based on the DNA alignment, using RAxML v. 8.2.12. This approach recognizes that the phylogeny of a single gene may not follow the consensus phylogeny due to a history of recombination. For completeness, however, we also performed codeml analyses by assuming the global phylogeny for the subset of genes that were present in all 64 samples. The outcomes of the two approaches were highly correlated (Supplementary Figure S3.2), and so for simplicity we focused on results based on phylogenies inferred separately for each gene.

Given the input phylogenies, we performed *codeml* analyses that relied on calculating likelihood ratios (LRs) under various models (Yang 2007). Briefly, we used the models to test the null hypothesis that ω = 1.0 against the alternative of positive selection (ω > 1.0) in two different ways. The first was a global test across the entirely phylogeny of a gene – i.e., across all branches and all sites. This test requires the comparison of two models: one (Model = 0 with Fix_omega = 1 and Omega = 1 in the *codeml* control file) that estimates a single ω from the data and another that sets ω=1.0 (Model = 0 with Fix_omega = 0 in the *codeml* control file). The two models yielded evidence for positive selection when the initial ω estimate was >1.0 and when the likelihood of the two models differed significantly, based on P < 0.01 after FDR correction. The second set of analyses was across sites – i.e., testing for genes with variable selection pressure across sites. For each gene, we first compared models M0 and M3 to test for heterogeneity in evolutionary rates across codons. If that test was significant, we then compared sites models M1a and M2a from *codeml* to test for specific codons with evidence of positive selection (ω > 1.0). For all summary statistics of ω, we excluded estimates of ω that were greater than 10 as potentially unreliable due either to low d_s or poorly resolved optimization. Individual codon residues under positive selection were identified using the Empirical Bayes analysis in codeml.

Data availability statement. All high-throughput sequence data generated in this study have been submitted to the NCBI Sequence Read Archive database at https://www.ncbi.nlm.nih.gov/sra and can be accessed with project number PRJNA833428.

3.4 Results

Core and Accessory Genes in *Xylella*. To investigate genome evolution in *X. fastidiosa*, we sequenced 20 novel *X. fastidiosa* genomes using hybrid approaches and retrieved publicly available genomes and raw sequencing data (Supplementary Tables S3.1 and S3.2). After filtering for isolation source and genetic distance, we retained a sample of 63 genomes that were broadly distributed among the subspecies. All our analyses were performed on this final set of 63 *X. fastidiosa* genomes with the *X. taiwanensis* outgroup. The *X. fastidiosa* genomes ranged in size from 2.42 Mb to 2.96 Mb, with an average length of 2.61 Mb (Figure 3.1A) and an average of 2,478 predicted genes (Figure 3.1B). The samples were extracted from 22 plant hosts representing 12 botanical orders (Figure 3.1C).

We categorized each gene as either core (present in 95% or more of *X. fastidiosa* samples) or accessory (Page et al. 2015). Across all 64 genomes, we identified 10,477 genes within the pan-genome; of those, 1,257 were core genes and 9,220 were accessory genes, with nearly 4,000 genes found in only a single isolate (Figure 3.1D). We performed functional analyses on both the core and accessory gene sets by grouping protein coding sequences into clusters of orthologous gene (COG) (Figure 3.1E-F). We compared COG category rankings between the core and accessory gene sets, with significant differences (Wilcoxon rank sum test, paired; P = 0.0001). After excluding genes with unknown function, the largest COG categories in the core gene set were 'translation, ribosomal structure, and biogenesis' (123 genes), 'cell wall/membrane/envelope biogenesis' (119 genes), and 'amino acid transport and metabolism' (92 genes). In contrast, the largest categories for accessory genes were 'replication, recombination and repair' (547 genes), 'intracellular trafficking, secretion, and vesicular transport' (364 genes), and 'transcription' (298 genes). Additionally, we investigated the core and accessory gene lists for significant GO based enrichment of specific biological processes (Supplementary Tables S3.4 and S3.5).

Phylogenetic Patterns of Core Genes, Accessory Genes and Hosts. We constructed a maximum likelihood phylogeny based on a subset of 1,024 genes that were present in all 64 isolates. The topology was highly supported; it had a mean bootstrap support of 93.75% across all nodes, with a median of 100% (Figure 3.2). The lowest bootstrap supports were primarily found at nodes separating *X. fastidiosa* strains isolated predominantly from grapevines, reflecting relatively low evolutionary divergence among these samples. As expected (Denancé et al. 2019), isolates clustered into three distinct clades representing the three main subspecies (ssps. *fastidiosa*, *multiplex*, and *pauca*), with 27, 23 and 13 isolates in each clade, respectively. To account for the possibility that homologous recombination impacted the resolution of the core phylogeny, we extracted regions of the core gene alignment that had an apparent history of recombination (Croucher et al. 2015), ultimately removing 85.1% of the alignment. The phylogeny inferred from this alignment was nonetheless highly congruent with the phylogeny that did not consider recombination. Only five accessions had altered positions between the recombination-adjusted and non-adjusted trees (Supplementary Figure S3.3).

To investigate general evolutionary patterns of the accessory gene complement, we compared the core gene phylogeny against a phylogeny based on accessory gene composition (Figure 3.3). Both the core gene and the accessory gene phylogenies clustered into three groups, and all members of the groups were consistent between phylogenetic treatments. This pattern broadly suggests that accessory genes, while defined by their inconstancy, are not exchanged *en masse* to a sufficient degree to alter phylogenetic signal among subspecies. Within subspecies, however, relationships at the tips of the phylogeny often differed between core and accessory trees. As an example, the cluster corresponding to *multiplex* displayed the most discordance between the core and accessory trees, with all OTUs contributing to phylogenetic incongruence (Figure 3.3). Interestingly, our *multiplex* sample also had more plant host species than our *fastidiosa* and *pauca* samples, suggesting the possibility (but by no means proving) that host factors may affect or moderate genome content (Kahn & Almeida 2022). Nonetheless, we found a significant correlation between distance matrices based on the core and accessory phylogenies (Mantel test, R = 0.1144, P =

0.019), which is consistent with the fact that the two trees had the same three major clades. The overarching impression of these analyses is that accessory gene composition does not turn-over so rapidly, due to HGT or other mechanisms, to erase phylogenetic and historical signals of subspecies diversification within *X. fastidiosa*.

We used both species phylogenies (based on alignments with and without putative recombinant regions) to test for associations between *X. fastidiosa* and their isolation sources (i.e., geographic location or host plant information) using ANOSIM (see Methods). There was a weakly significant phylogenetic association (ANOSIM R = 0.08178, P = 0.042) between the geographic location and the phylogeny built from the full core gene alignment (ANOSIM R = 0.08178, P = 0.042) but not with the phylogeny built from non-recombinant regions (ANOSIM R = -0.004147, P = 0.4895). Applying the same approach to host species revealed a significant phylogenetic signal for both phylogenies (ANOSIM R = 0.1381, P = 0.047; non-recombining regions only, ANOSIM R = 0.6698, $P < 1 \times 10^{-4}$). Since *X. fastidiosa* infects a wide range of plants, we also retrieved the taxonomic order of each plant host to test for a phylogenetic signal at a deeper taxonomic level, recapitulating the significant association with both phylogenies (ANOSIM R = 0.3152, P < 0.0001; non-recombining regions only, ANOSIM R = 0.1226, P = 0.0198). In other words, strains isolated from plants within the same taxonomic order were more phylogenetically similar to one another than isolates taken from unrelated plants.

We hypothesized that accessory genes are crucial in pathogen-host interaction and therefore repeated ANOSIM analyses with a distance matrix based on the presence and absence of accessory genes (Figure 3.3). We found a significant association between accessory gene content and geographic isolation source (ANOSIM R = 0.4553, P = 0.5307) and a weakly significant association between accessory gene content and host species (ANOSIM R = 0.1503, P = 0.0372). The association was lost, however, at the level of plant order (ANOSIM R = 0.02367, P = 0.3033). Overall, associations were less evident based on accessory gene content vs. the core-gene phylogeny.

Gene Gain and Loss. The sheer number of accessory genes indicate that the genome content of *X. fastidiosa* is, like other microbes (Bolotin & Hershberg 2015; Iranzo et al. 2019), shaped by extensive gene gain and loss events that are probably mediated by HGT (Firrao et al. 2021). We were interested in assessing the pattern of gene gain and loss across the phylogenetic tree, hypothesizing that both could be enhanced on branches that lead to host shifts. We used GLOOME to estimate the number of gain and losses of accessory genes across the *X. fastidiosa* phylogeny and represented those estimates phylogenetically (Figure 3.4). Ignoring the branch leading to the *X. taiwanensis* outgroup (PLS229), the internal branches discriminating the *X. fastidiosa* subspecies were estimated to average ~550 separate gene gain and gene loss events. The remainder of the tip and ingroup branches averaged ~100 gene gain and loss events (average gains/branch = 92.8 genes; average losses/branch = 100.0 genes; Figure 3.4A&B).

While it is useful to estimate the number of gains and losses on each branch, we thought it more helpful to normalize the number of estimated gain and loss events by branch lengths, estimated from the sequence analysis of core genes. This normalization by branch length converted the number of gene gains and losses to *rates* of gene gain (or loss) relative to sequence divergence. We then sought to identify branches with aberrantly high rates of gene gain or loss (Figure 3.4C&D), which reflect branches with especially notable turnover of accessory genes. Like a previous microbial study (76), we found that most of the phylogenetic lineages with outlier rates were located at the tips of the phylogenetic tree. For example, of the 21 branches with high rates of gene gain, 19 were at the tips of the phylogeny (Figure 3.4A). Similarly, 18 of 21 branches with high rates of gene loss were external branches. These observations suggest features about the evolutionary dynamics of genic turnover (see Discussion).

Characterizing selection with ω. We characterized selection on individual genes by estimating the dN/dS ratio (ω); we especially sought to identify genes that experienced positive selection (i.e., $\omega > 1.0$), as a potential signal of genes that contribute to dynamics between the pathogen and its hosts. To do so, we applied a series of nucleotide substitution models to individual genes, ultimately resulting in tests for positive selection on two levels: globally across a phylogeny and across codon sites (see Methods). For these tests, we examined the full complement of 1,257 core genes, a subset of 3,691 accessory genes, and a set of 187 multicopy genes.

Testing selection globally for each gene: We first estimated a single ω value for each gene, using a method that assumes ω is constant across all branches of the entire gene tree and across all codons in the nucleotide alignment. Applied to the core genes, ω estimates ($\widehat{\omega}$) ranged from 0.01048 to 2.92803 with an average of 0.21973 (Figure 3.5A). Nineteen core genes had $\widehat{\omega}$ higher than 1.0, but none of these were significantly > 1.0 (P > 0.01, FDR correction). In fact, the vast majority (1,144 of 1,257) of core genes had $\widehat{\omega}$ significantly < 1.0 (P < 0.01, FDR correction; Figure 3.5A), reflecting pervasive purifying selection. The range of $\widehat{\omega}$ was substantially broader for accessory genes, from $\widehat{\omega} = 0.0001$ to 9.60069, with an average of 0.51443 (Figure 3.5B). Among accessory genes, 367 (9.9%) had a global estimate of $\omega > 1.0$, but only eight had statistically significant evidence for positive selection. These eight genes were candidates to encode proteins involved in host-pathogen interactions, but seven of eight were annotated as hypothetical genes. Overall, the average $\widehat{\omega}$ was significantly higher in the accessory gene set compared to the core genes (Welch's T-test, $P < 2.2 \times 10^{-16}$), reflecting either lower purifying selection against these genes, more positive selection, or both.

We also identified 187 genes that had 2 or more copies within a single accession in a syntenic context, but that were single copy in other accessions. We performed *codeml* analysis to estimate ω for each multicopy gene; $\widehat{\omega}$ ranged from 0.02272 to 4.26800, with an average of 0.51129 (Figure 3.5D). Over half of the genes had $\widehat{\omega}$ significantly < 1.0 (59.4%; $P < 0.01$, FDR correction); only one, a hypothetical gene (*group_1109*) had $\widehat{\omega}$ significantly higher than 1.0 ($\widehat{\omega} = 1.85845$, $P < 0.01$, FDR correction).

Positive selection in codon sites: The global test is a conservative criterion to search of positive selection, perhaps overly so. Accordingly, we turned to an alternative method that tests for variation in ω among codon sites and identifies whether sites are under positive selection. To do so, we ran the sites models in *codeml*, which are a group of nested models. For completeness, we first compared sites model M0, which represents the null hypothesis that there is a single ω value for all sites, against sites model M3, which permits ω to vary among sites. In the core genes, the likelihood ratio test was significant for 501 genes ($P < 0.01$, FDR correction). We then took this set to compare to test for positive selection using sites models. A total of 67 core genes had evidence of positive selection among sites ($P < 0.01$, FDR correction). We also tested for positive selection on codon sets within the 3,691 accessory genes using the same approach. Of the total, 895 displayed evidence of variable ω among sites (P < 0.01, FDR correction) and 201 yielded evidence of positive selection (P < 0.01, FDR correction). Finally, we applied the sites models to the set of 187 multicopy genes, yielding another 33 genes with evidence for positive selection. To sum, 5.3% (i.e., 67 of 1,257) of core genes, 5.4% (201/3,691) of accessory genes and 17.6% of multicopy genes had significant evidence of at least one codon with an apparent history of positive selection. Among the 201 accessory genes, four (*cya, group_454, group_1057*, and *group_3542*) also had evidence for positive selection via the global test.

3.5 Discussion

Host-pathogen interactions can drive rapid evolution of pathogenic bacteria, particularly for genes involved in arms-race dynamics (Daugherty & Malik 2012; Sironi et al. 2015). Here, we investigated the genomic evolution of the plant pathogen, *X. fastidiosa*, through comparative genomic analysis of genomes that represent diversity across the species, based on a sample set of 64 genomes. The sample was isolated from 23 different plant hosts (Figure 1C) from throughout the world (Supplementary Figure S3.1). With these data, we constructed a pangenome that contained 1,257 core genes and 9,220 accessory genes, similar to previous studies (Giampetruzzi et al. 2017; Castillo et al. 2020). Of the core genes, the majority were, as expected (Tettelin et al. 2008), involved in essential cellular processes -- such as translation, cell wall biogenesis, and amino acid metabolism (Figure 3.1E). We used the set of core genes to infer a maximum likelihood phylogeny, either with or without adjusting to putatively recombining regions of the genome (Figure 3.2; Supplementary Figure S3.3). As with previous systematic treatments of *X. fastidiosa* (Yuan et al. 2010; Marcelletti & Scortichini 2016; Denancé et al. 2019), both phylogenies identified three clades corresponding to the three main subspecies (*fastidiosa, multiplex*, and *pauca*).

We employed both phylogenies to investigate the relationship between the *X. fastidiosa* phylogeny and the plant host. The question of host specialization was first addressed using phylogenetic approaches with multilocus sequencing typing (MLST) data. In this work, Sicard et al. (2018) generated MLST data from 7 housekeeping genes from 50 *X. fastidiosa* genotypes. After building a phylogeny, they tested coevolutionary relationships between host species and *X. fastidiosa* MLST types but found no significant evidence of coevolution, implying a lack of host specialization. This topic was recently revisited with full genome data (Kahn & Almeida 2022; Uceda-Campos et al. 2022), but the results were inconsistent between studies. Uceda-Campos et al. (2022) found no evidence that plant host species clustered on their *X. fastidiosa* phylogeny, but the samples did cluster by geography. In contrast, Kahn and Almeida (2022) inferred ancestral character states of plant hosts on the *X. fastidiosa* phylogeny and were able to resolve the character state of some deep nodes. They inferred, for example, that coffee plants were the ancestral host species for the node separating *X. fastidiosa* ssp. *fastidiosa* from other subspecies. These patterns suggest that phylogenetic history is associated with specific plant hosts and host ranges.

The disagreement among previous studies, and the fact that all such analyses are properties of the sampled isolates, makes the issue worthy of further assessment. In our study, we have found a significant, non-random association between phylogenetic relationships and both the species and taxonomic order of plant hosts (P < 0.0001) based on core phylogenies. These results are consistent with some level of specialization of *X.*

fastidiosa to plant hosts and with the recent analysis of Kahn and Almeida (2022). Moreover, these results were robust to phylogenetic treatment – i.e., the inclusion or exclusion of genomic regions inferred to have histories of recombination. Although it is difficult to quantitatively compare ANOSIM results across studies, it is worth noting that the association of *X. fastidiosa* to plant order is similar in magnitude to the association between a gut colonizing bacterium (*Bifidobacterium*) and the host species from which it was isolated (Rodriguez & Martiny 2020).

Given some evidence for host specialization, we hypothesized that it is driven in part by accessory gene content. Under this hypothesis, we predicted an association between genes and hosts should be as (or more) pronounced for accessory genes than for core genes. Instead, we found no significant association between accessory gene complement and taxonomic order and only a weak association with plant species. Our results are unlike, for example, the case of bifidobacteria, where the association with host species was nearly as strong for accessory genes as for host genes (Rodriguez & Martiny 2020). We cannot be sure why we do not detect a signal for host specialization of accessory genes, but we can think of three explanations. One is that that host associations, to the extent they exist, are not driven by accessory genes but rather by evolutionary divergence in core genes. Another is statistical power: because there are many more sequence changes among the core genes than there are changes in accessory gene content, the distance matrix for core genes likely has a higher signal-to-noise ratio than accessory gene content. Finally, if accessory genes do mediate host shifts, it is possible – and even likely - that only a subset of accessory genes drive these shifts. Under this scenario, there may be significant associations for a small subset of accessory genes, but the signal of this association is weak across the entire accessory gene set. This conjecture seems reasonable given that Kahn and Almeida (2002) found that the presence/absence of a subset of only ~30 accessory genes correlated with the plant host. In addition, it is worth emphasizing that *X. fastidiosa* interacts not only with plants but also insect vectors and microbial communities, so that some subset of accessory genes likely contribute to these interactions instead of those with plant hosts.

The pattern of gene gain and loss events. Another potential tool to study adaptation to specific hosts is by examining shifts in gene composition through gene duplication, deletion, or HGT events (Hurles 2004; Arnold et al. 2022). We estimated the number of gene loss and gain events along the core-gene phylogeny, and then normalized those numbers relative to sequence divergence. Using this approach, we found that most branches followed a consistent rate of gene gain or loss relative to sequence divergence. The fact that the accessory gene phylogeny recapitulates the three subspecies (Figure 3) suggests, along with previous evidence, that *X. fastidiosa* evolves predominantly through vertical inheritance and intraspecific recombination, rather than HGT from other bacterial species (Nunney et al. 2013; Castillo & Almeida 2021).

We have, however, identified 19 and 18 lineages with enriched gain or loss events, respectively, and most of these branches were at the tip of the phylogeny. Again, a potential explanation for these gain and loss dynamics is that they reflect host shifts. There are some isolated examples that are consistent with this hypothesis. For example, isolates XF6c, Pr8x, RAAR17, and OLS0478 in *pauca* have branches with enriched gene gains (Figure 3.4A). Two of these (OLS0478 and Pr8x) were isolated from oleander and plum, respectively, and are the only isolates associated with those plant hosts in their clades, suggesting a host shift. More globally, however, the evidence for this hypothesis is unconvincing. When we, for example, contrast gene gains between pairs of sister taxa with the same plant host, three of the 16 sister pairs had enriched rates of gene gain. This proportion of enriched branches was not significantly lower than the reminder of the tree ($P > 0.05$; Fisher's Exact Test), despite the fact that the sister taxa did not experience a host shift. All of these inferences are of course dependent on our sample and ignore the vector component of the *X. fastidiosa* lifecycle, so there are limitations to our conclusions. At present, however, the evidence for an association between host shifts and enhanced gene gain and loss events is weak.

This leaves unexplained the pattern of enriched rates of gene gain and loss at the tips of the tree. We suspect this pattern is analogous to patterns of mutations in populations, as suggested previously (Graña-Miraglia et al. 2017). New mutations begin as rare, low frequency variants in single individuals. Eventually most of these mutations are removed by the processes of genetic drift and natural selection, so that there are more new mutations in populations than old mutations. In a phylogenetic context, these new mutations would be evident at the tip of the trees, so it may be reasonable to expect higher effective rates of gene gain and loss in the 'newest' phylogenetic branches. This explanation only has credence, however, if the observed gain and loss events are both recent – i.e., newer than the sequence mutations that define the tip branches – and frequent.

The identification of positively selected genes. Many previous studies have implicated genes and their protein products in ongoing arms-races between pathogens and their hosts

(Anderson et al. 2010; Schulte et al. 2010). One way to approach this question is agnostic to function, which is to screen for genes with a history of positive selection. Ours is not the first attempt to detect selection in *X. fastidiosa* genomes. Previous studies have searched for selection by comparing levels of polymorphism or rates of synonymous and nonsynonymous mutations in the core genome using Tajima's D and the McDonald-Kreitman test (Castillo et al. 2021, 2020). Other work has measured ω in core genes but without statistically testing for positive selection (Castillo & Almeida 2021) or by applying the global test for w for > 1.0 (Vanhove et al. 2020). To our knowledge no other study of *X. fastidiosa* has tested for positive selection in accessory genes nor applied codon sites models. The set of positively selected *X. fastidiosa* genes represents candidate pathogenicity factors to mediate interactions with the environment, including plant host, insect vectors, or members of the microbial community.

To study positive selection, we estimated ω, or the ratio of nonsynonymous to synonymous mutations, for each core gene and for each accessory gene found in four or more isolates. In total, this exercise encompassed 5,135 genes: 1,257 core genes, 3,691 accessory genes, and 187 multicopy genes. We began by applying a global test that estimates ω over all sites and phylogenetic lineages. This approach can be overly conservative, because a significant test of $\omega > 1.0$ requires that positive selection is very strong, acts across many sites in a gene, is present in most of the branches of the phylogeny, or all of the above. As expected, we found only a few genes (eight accessory genes in total) that were significant for positive selection with this test. Unfortunately, the annotations of 7 of 8 of these genes yielded few insights into their functions. To explore gene function further, we identified protein domains using the Conserved Domain Database. We found, for example, that the gene *group_7848* contains a VirB3 protein domain, which is part of the Type IV secretory pathway and is commonly associated with the membranes of the bacterial cell. The gene *cya* was also implicated using this test, which encodes adenylate cyclase and plays essential roles in regulation of cellular metabolism (Danchin et al. 1984). Interestingly, the *cya* protein is involved in the cyclic AMP system, which is a global regulator in gram-negative bacteria and has been shown to modulate gene expression in pathogenic bacteria (Smith et al. 2004; Kim et al. 2005).

The global test did allow, however, for two broad generalizations about patterns of selection in *X. fastidiosa*. First, as a group the core genes are under strong purifying selection with most (>90%) having ω estimates significantly < 1.0. Second, accessory genes generally have lower levels of purifying selection, as evidenced by a lower proportion (45%) of significant tests for $\omega < 1.0$ and by much higher average $\widehat{\omega}$ values (0.21973 vs. 0.51443; Figure 3.5A). The proportion of significant tests must be compared between genic sets with caution, because the smaller sample sizes (*n*=4 to 59) for accessory genes likely reduce statistical power relative to the minimum of 60 samples for all core genes, as do any differences in gene lengths. Nonetheless, the contrasting pattern of ω is consistent with the ideas that core genes have conserved biological functions and that accessory genes are more amenable to evolutionary change due to their nonessential, but still potentially biological relevant, cellular roles (Horesh et al. 2021). Accessory genes may also experience higher variation in their selection dynamics because recombination affects them more than core genes (Castillo & Almeida 2021).

Given few signals of positive selection with the global test, we turned to codon site models. To our surprise, the proportion of positively selected genes was similar for core genes (5.3%) and accessory genes (5.4%). The salient question is whether these genes give some clue to function. Of the 67 core genes with evidence for positive selection at the codon level, 40% were unannotated. We performed a functional analysis by grouping the protein coding sequences of these 67 core genes into COG categories to infer cellular functions. Excluding the category of unknown function, the largest category was 'cell wall/membrane/envelope biogenesis,' followed by the 'amino acid metabolism and transport,' 'carbohydrate metabolism and transport,' 'translation,' and 'intracellular trafficking and secretion' (Supplementary Figure S3.4A).

Of the 201 accessory genes with evidence for positive selection at the codon level, 82% were not annotated for function. The remaining set of 36 genes was enriched for GO categories related to protein secretion by the type IV secretion system (Supplementary Table S3.6). To better infer function, we performed a COG analysis and found that the largest categories (excluding the category of unknown function) were 'intracellular trafficking and secretion,' 'replication, recombination and repair,' and 'secondary metabolites biosynthesis, transport and catabolism' (Supplementary Figure S3.4B). Intriguingly, of this set of 201 genes, 50 overlapped with the set of 367 genes that had a gene-wide estimate of $\widehat{\omega} > 1$. While these are especially strong candidates for having a history of positive selection, a disappointing 94% of them were unannotated for function. The three genes with annotations were: *cya*, *nagZ_2*, and *bacterial adaptive response A* (*barA)*. The gene *nagZ_2* encodes a beta-glucosidase that is important for biofilm formation in *Neisseria gonorrhoeae*, suggesting it could play a similar role in *X. fastidiosa*. It merits further functional analysis, since biofilms are important to the infection cycle (Bhoopalan et al. 2016). *barA* encodes a membrane associated histidine kinase that has a regulatory role in cell division, metabolism, and pili formation, and it has been implicated in regulating the virulence response of uropathogenic *E. coli* (Palaniyandi et al. 2012; Sahu et al. 2003). Finally, the multicopy genes also yielded evidence of positive selection, including *cdiA1*, which is part of the secretory contact-dependent growth inhibition (CDI) system that modulates biofilm formation in *Acinetobacter baumannii* (Roussin et al. 2019).

As a final exercise, we cataloged the incidence of positive selection in a set of 35 genes that have been listed as virulence and pathogenicity factors in *X. fastidiosa* (Rapicavoli et al. 2018). Of the 35, we could identify 29 in our database based on the PD number annotation and reference sequence (http://www.microbesonline.org/operons/gnc183190.html; Table 1). We expected that this set of 29 genes would be enriched for evidence of positive selection relative to the genomic background, because these genes are putatively involved in arms-race interactions. The trend for these genes was in the expected direction, because 4 of 29 (= 13.9%) were significant vs. 301 of 5,135 (= 5.8%) in the rest of the genome. However, the difference in proportions was not significant (Fisher Exact Test, P = 0.1091). Nonetheless, this set of experimental genes is interesting. All four genes with evidence of positive selection encode proteins associated with the membrane of gram-negative bacteria and are involved in membrane transport or adhesin. Specifically, the genes *fimF, xadA* and *xatA* encode proteins involved in fimbrial adhesion, non-fimbrial adhesion, and biofilm formation, respectively, and the gene PD1311 encodes a protein involved in membrane transport (Ma Rodriguez et al. 1993; Sun et al. 2005; Abbas et al. 2007; Das et al. 2009; Zeiner et al. 2012). Because there is a resolved protein structure for fimF (Gossert et al. 2008), we investigated the location of positively selected codons. Of the four positively selected codons (N80, D87, F137, and D142), one (D87) was in a flexible loop and a second (D142) comprised part of the second β-sheet of the protein (99). Together this suggests that changes in the amino acid sequence of *fimF* may be impacting its function

We must caution that positive selection analyses are subject to false positives, and they are also dependent on specific analysis features, like the sample set, the criteria for determining homology, and the sequence alignments. We have nonetheless found several genes with some evidence of positive selection that may also contribute to functions relevant to infection. We believe they represent suitable candidates for further functional analyses to elucidate their role in host-pathogen interactions and perhaps even host specificity.

Figures

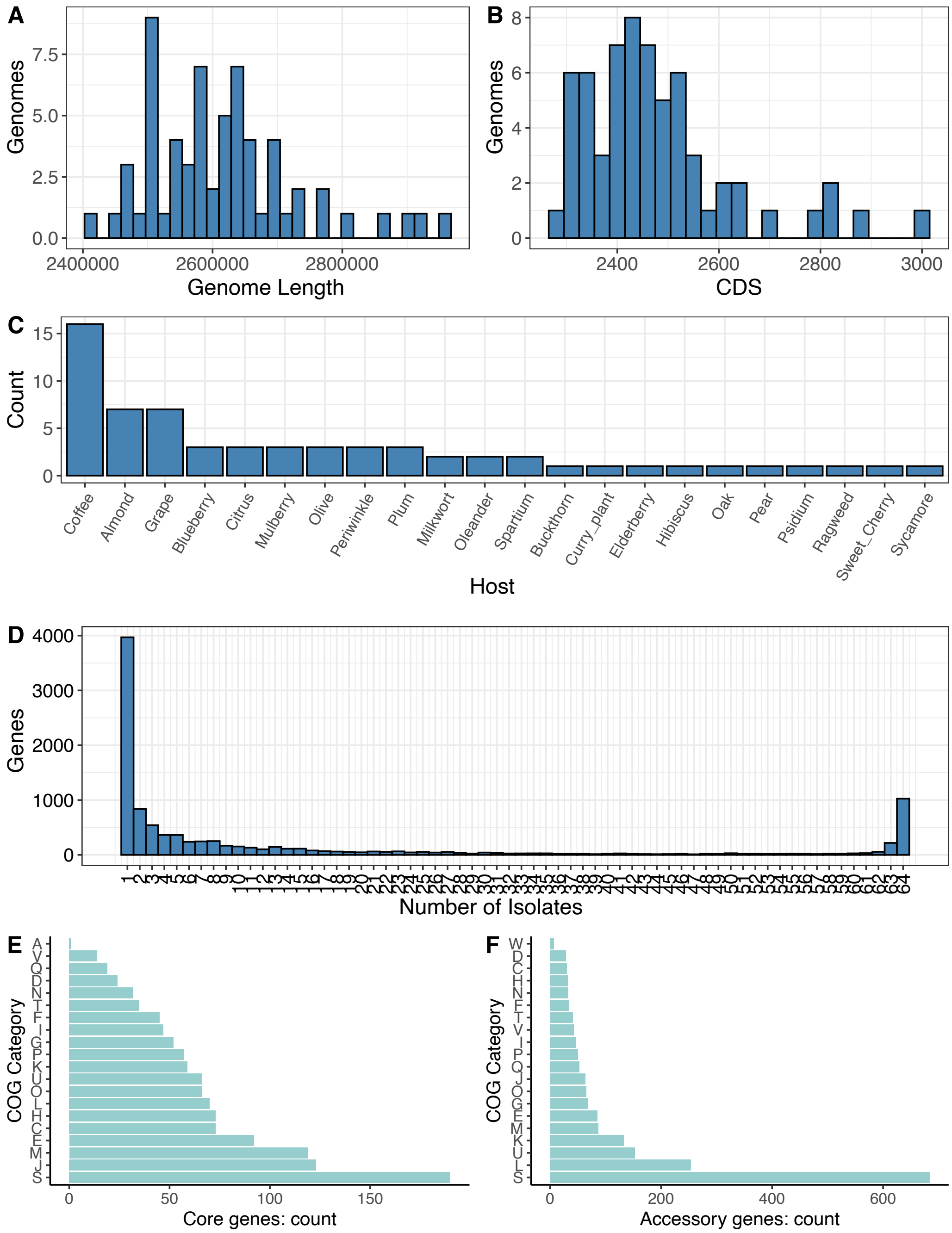

Figure 3.1: Histograms reporting the characteristics of the 64 *Xylella* genomes. A) Genome lengths, exhibited in base pairs. B) The number of genes within a genome. C) A histogram of the plant species from which genomes were isolated. D) A histogram of the number of genes found in *x* number of genomes; this histogram shows, for example, that nearly 4,000 genes are found in only of one the genomes out of the entire sample of 64 genomes and that 1,024 genes are found in all 64 genomes; E) the distribution of functional categories for the set of 1,257 core genes and F) the distribution of functional categories for the set of 9,220 accessory genes. A key to the COG categories for panels E) and F) is in Figure S4.

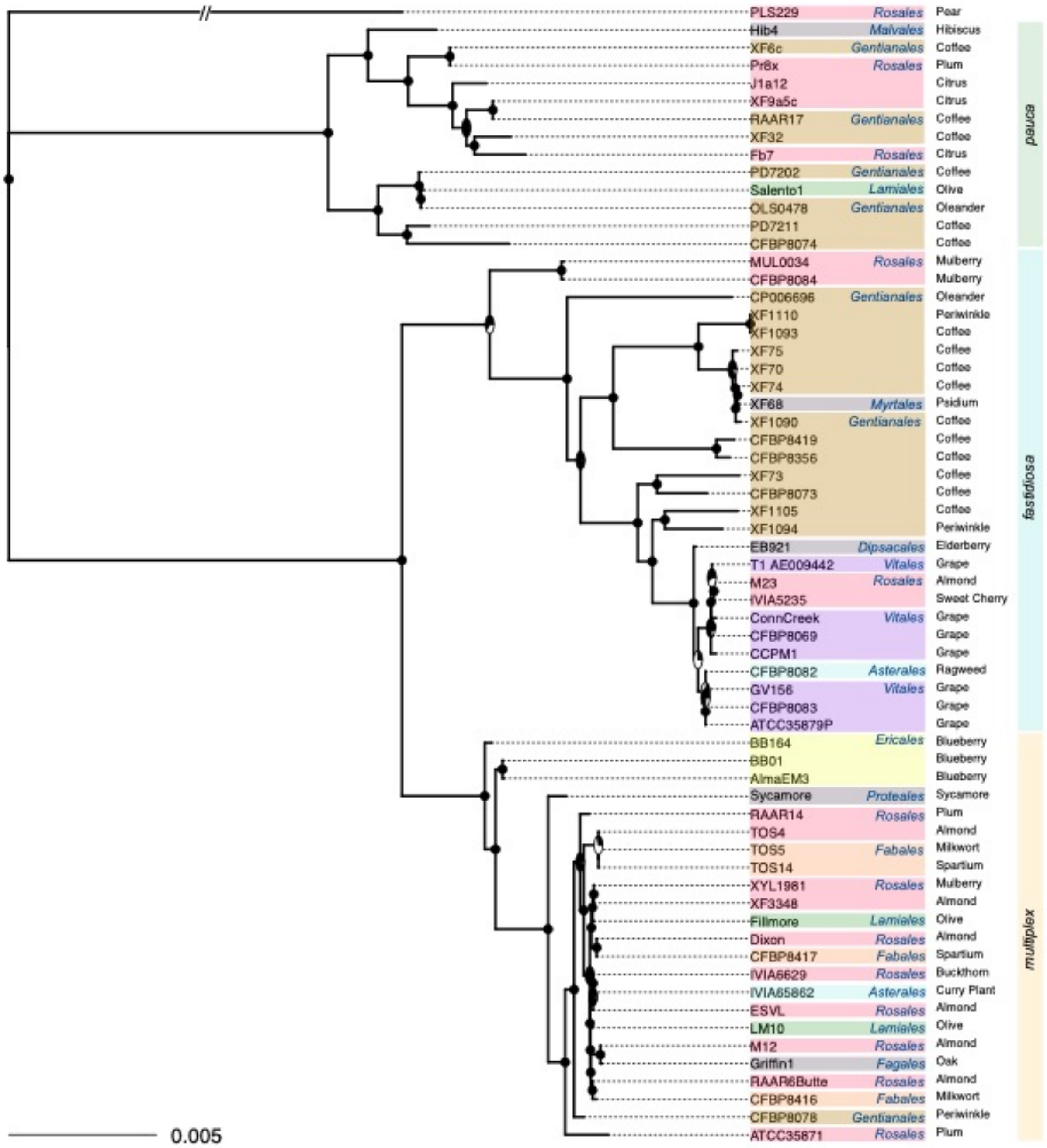

Figure 3.2: The inferred phylogeny of 64 *Xylella* genomes, based on maximum likelihood inference of core gene alignments. Each isolate is labelled at the tips and is colored according to the order of the plant isolation source (host). The common name of the host is provided to the right of order information. The three *X. fastidiosa* subspecies are indicated, as are bootstrap values at each node. The bootstrap values are pie charts, where black represents the percent of bootstrap support, and the scale bar reflects the magnitude of sequence divergence per nucleotide site.

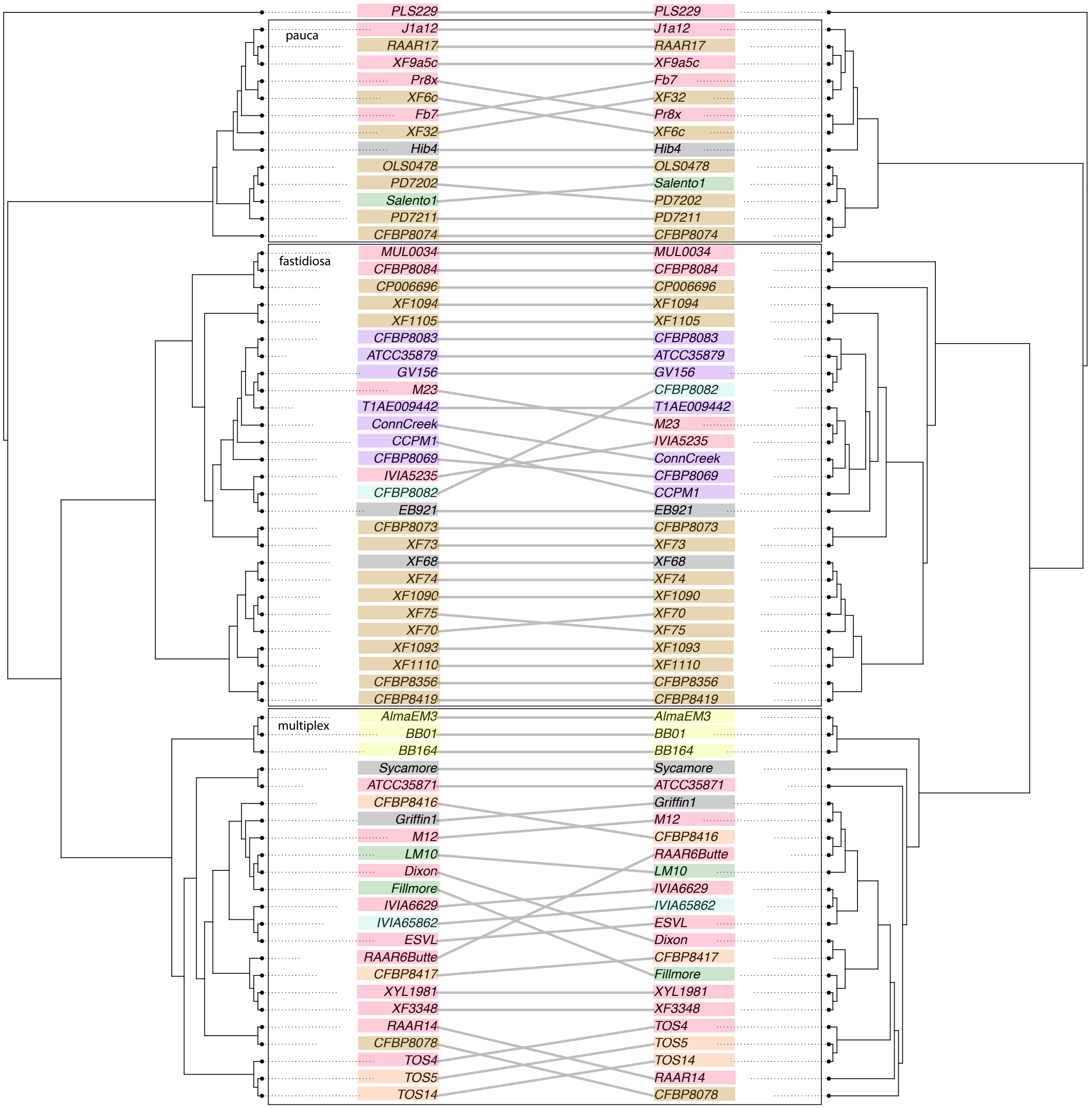

Figure 3.3: A comparison of a NJ tree based on distances due to gene presence / absence (on the left) to the likelihood tree based on the core gene alignments (from Figure 2, on the right). As in Figure 2, the isolates are labelled at the tips of trees, with the colors representing plant order. Both phylogenies contain three main *X. fastidiosa* clades, representing the three subspecies. Lines connect the same isolate between the two trees, with angled lines representing topological discordance between phylogenies. The three *Xylella* subspecies are outlined in a black box and labelled.

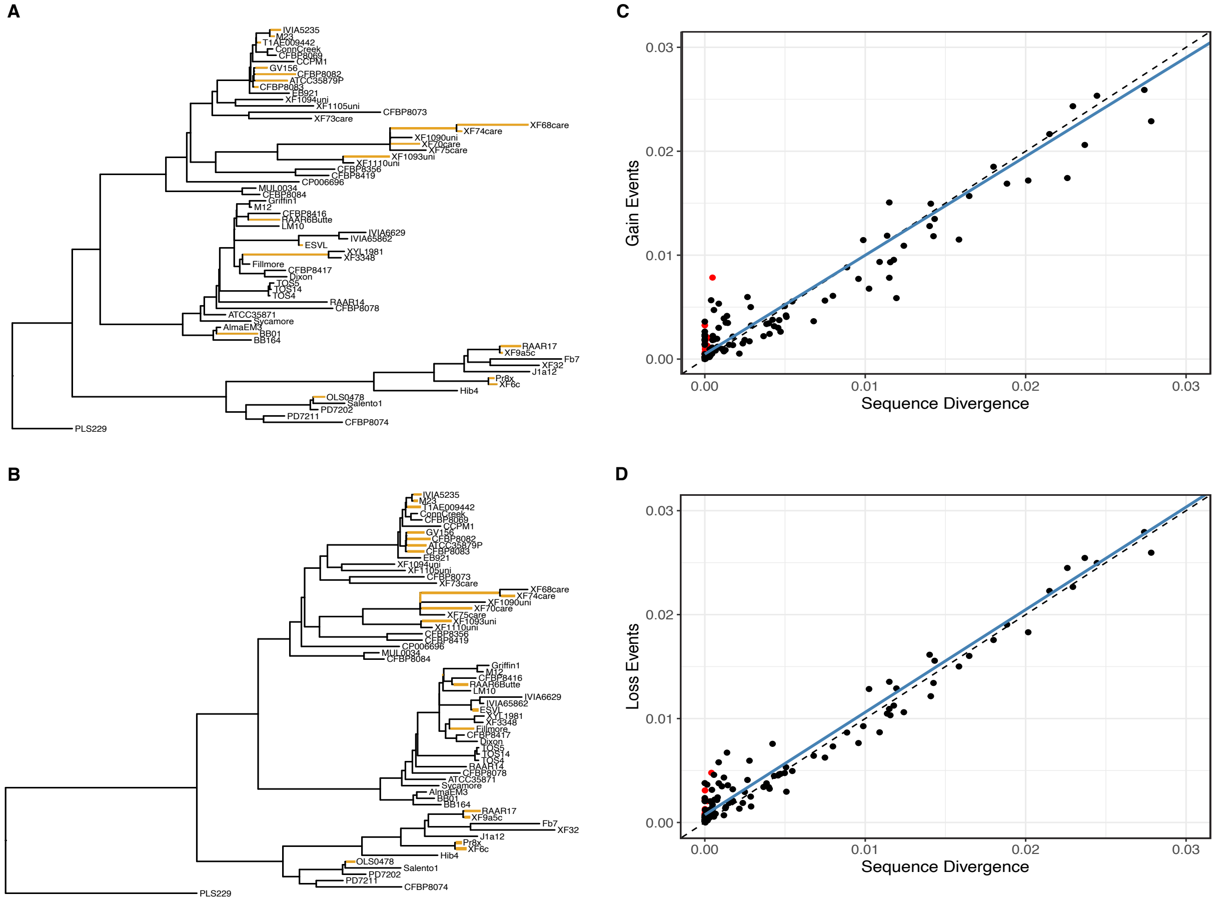

Figure 3.4: The result of gene gain and loss analyses. A) The phylogeny of the isolates, with branch lengths proportional to the number of gene gain events. The colored branches are branches with outlier gene gain rates. B) The phylogeny of the isolates, with branch lengths proportional to the number of gene loss events. The colored branches are branches with outlier gene loss rates. C) A plot of the gene gains against sequence divergence; in the plot each dot represents one of the 125 branches on the phylogeny. Outlier dots are colored red. D) As in C, with gene losses plotted again sequence divergence.

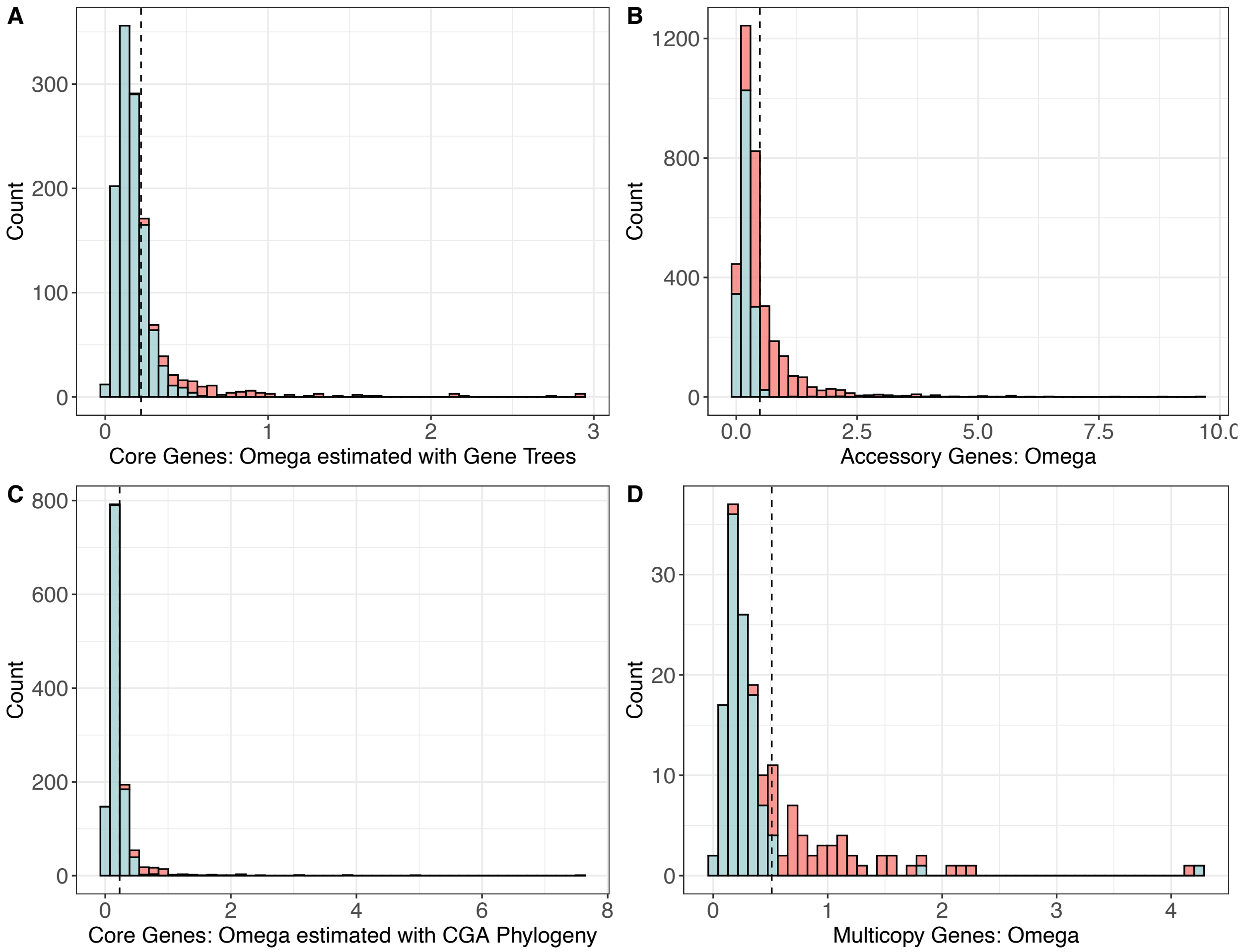

Figure 3.5: Estimated values of ω under M0 (the one-ratio model) in the core and accessory genes. The distribution of $\widehat{\omega}$ values is plotted for A) core genes estimated with gene trees, B) accessory genes with gene trees C) core genes estimated with the core gene alignment (CGA) phylogeny, and D) multicopy genes with gene trees. Histogram bars are shaded to reflect the outcome of the likelihood ratio test (non-significant tests are colored red and significant tests are colored blue) between a model that estimated $\widehat{\omega}$ and a model with ω fixed to 1.0. The horizontal dashed line denotes $\widehat{\omega}$ for each gene set.

Tables

Table 3.1 *Codeml* results for experimentally identified virulence and pathogenicity genes, as listed (Rapicavoli et al. 2018)

PD Number	Gene Name	Pan-genome classification	No. Genomes[1]	(M0)[2]	M2a vs. M1a p-value[3]
PD0058	*fimF*	Accessory	41	0.31555	**3.25E-08**
PD0062	*fimA*	Accessory	26	0.81255	0.247
PD0233	*rpfB*	Accessory	57	0.16832	1
PD0279	*cgsA*	Core	64	0.14404	1
PD0406	*rpfC*	Accessory	44	0.34502	1
PD0528	*xatA*	Core	64	0.43097	**1.38E-41**
PD0731	*xadA*	Accessory	58	0.39196	**0.004**
PD0732	*xpsE*	Core	64	0.05825	1
PD0814	*wzy*	Accessory	43	0.17675	1
PD0843	*tonB1*	Core	64	0.11374	0.534
PD0848	*pilL*	Core	64	0.18195	1
PD0986		Core	64	0.10828	1
PD1099	*dinJ/relE*	Accessory	25	0.10271	1
PD1100		Accessory	15	0.20708	0.731
PD1284	*algU*	Core	64	0.19261	1
PD1311		Accessory	33	0.42541	**3.47E-05**
PD1380	*csp1*	Core	64	0.15702	1
PD1391	*gumH*	Accessory	46	0.12964	1
PD1394	*gumD*	Core	63	0.11504	1
PD1485	*pglA*	Accessory	59	0.28401	0.114
PD1678	*phoQ*	Core	64	0.1086	1
PD1679	*phoP*	Core	64	0.03272	1

PD1703	*lesA/lipA*	Core	64	0.06614	1
PD1792	*hxfB*	Core	64	0.10828	1
PD1826	*chiA*	Core	64	0.11424	1
PD1856	*engXCA1*	Core	63	0.24034	1
PD1964	*tolC*	Core	64	0.10051	1
PD1984	*gacA*	Core	64	0.13444	1
PD2118	*hxfA*	Core	64	0.10828	1

[1] the number of genomes, out of 64, in which the gene was detected.

[2] M0 estimates a single w across the entire phylogeny of sequences.

[3] The p-value of tests after FDR correction. Bolded values are significant at p < 0.01. The notation E refers to the power of ten.

Supplemental Information

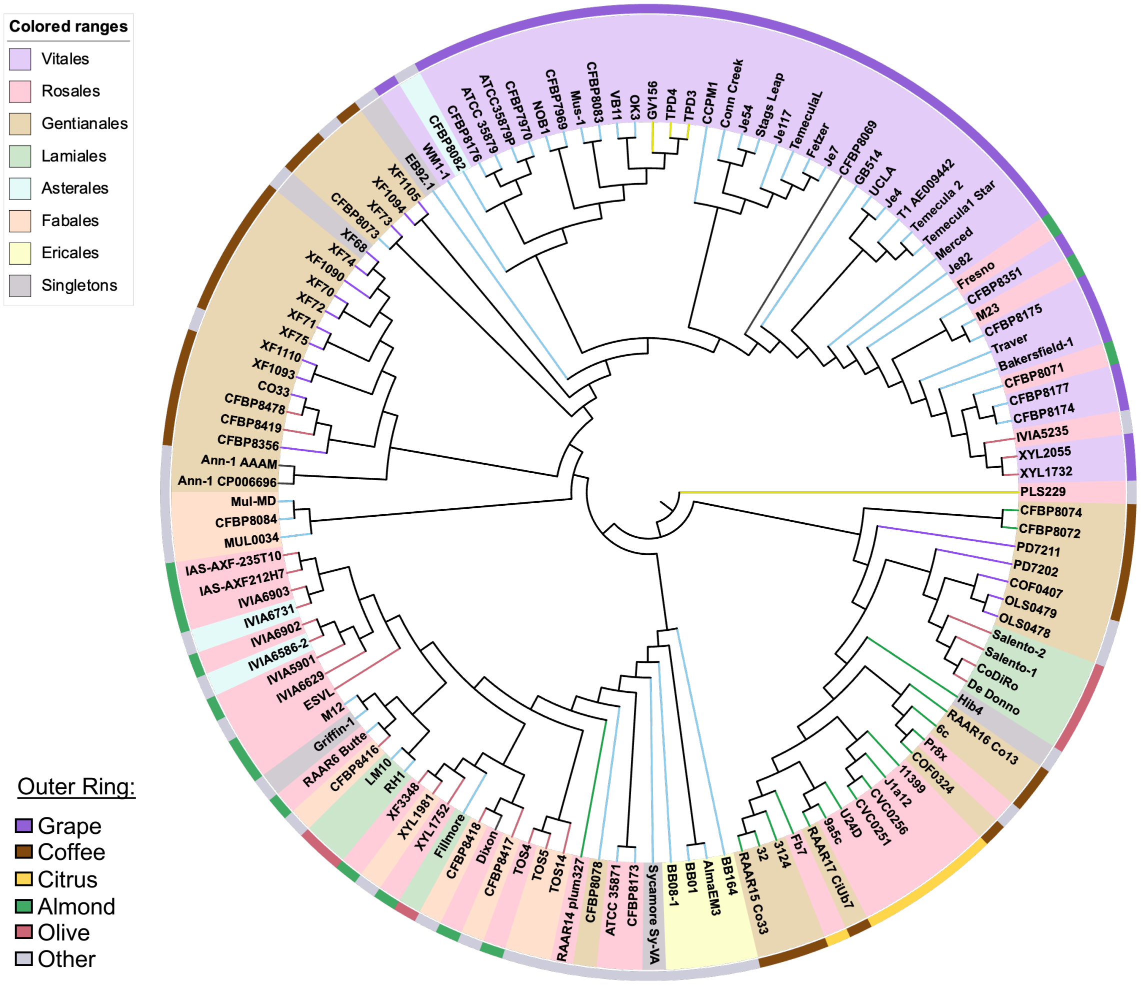

Figure S3.1: Phylogenetic relationships of the set of 129 publicly available and novel *Xylella fastidiosa* and *X. taiwanensis* genomes gathered to develop this study. The outer ring denotes the specific plant host from which the bacteria was isolated, and the shaded ranges denote the host plant's taxonomic order. The branches are colored to denote the continent of isolation: North America (Blue), Central America (Purple), South America (Green), Europe (Red), Asia (Gold).

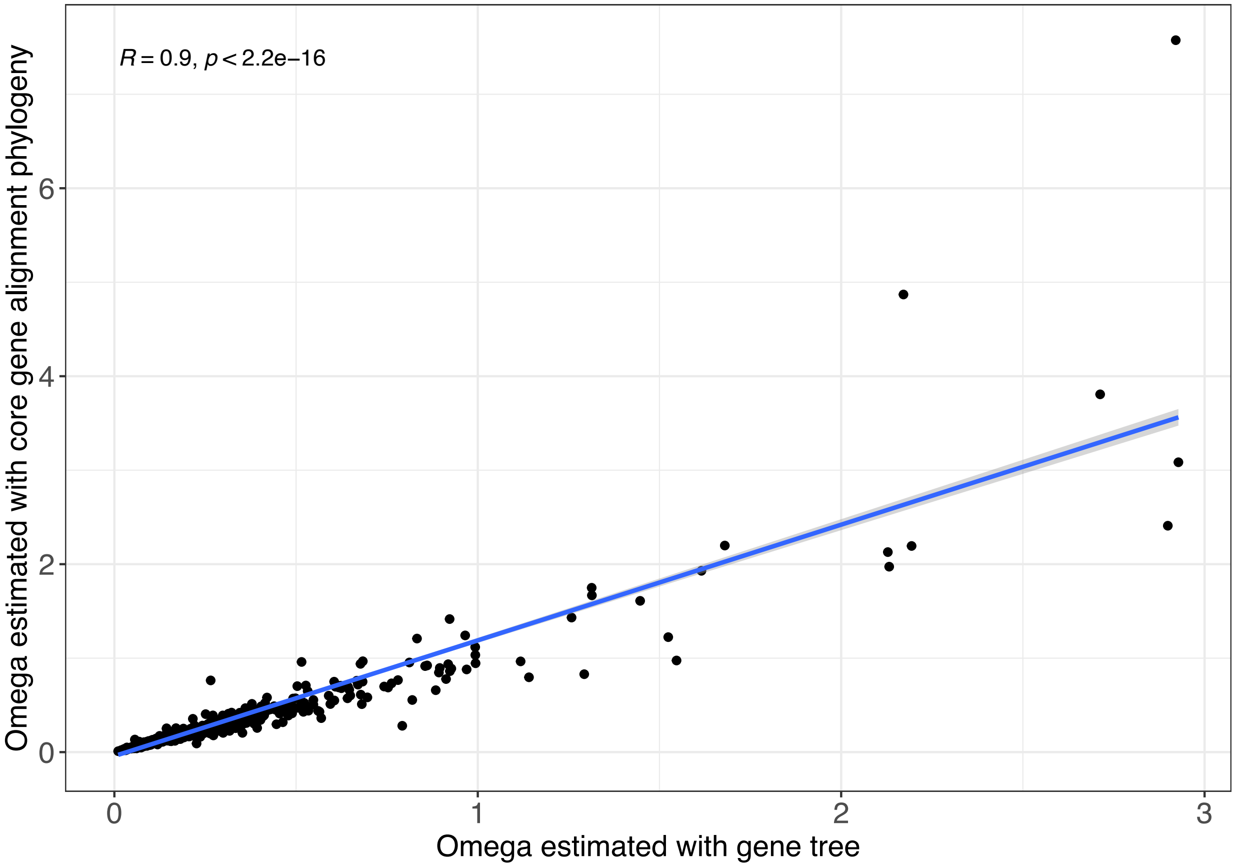

Figure S3.2: Correlation plot between the values of omega estimated with two different methods: building an unrooted maximum-likelihood gene tree plotted on the x-axis and with the global phylogeny built from the core gene alignment on the y-axis.

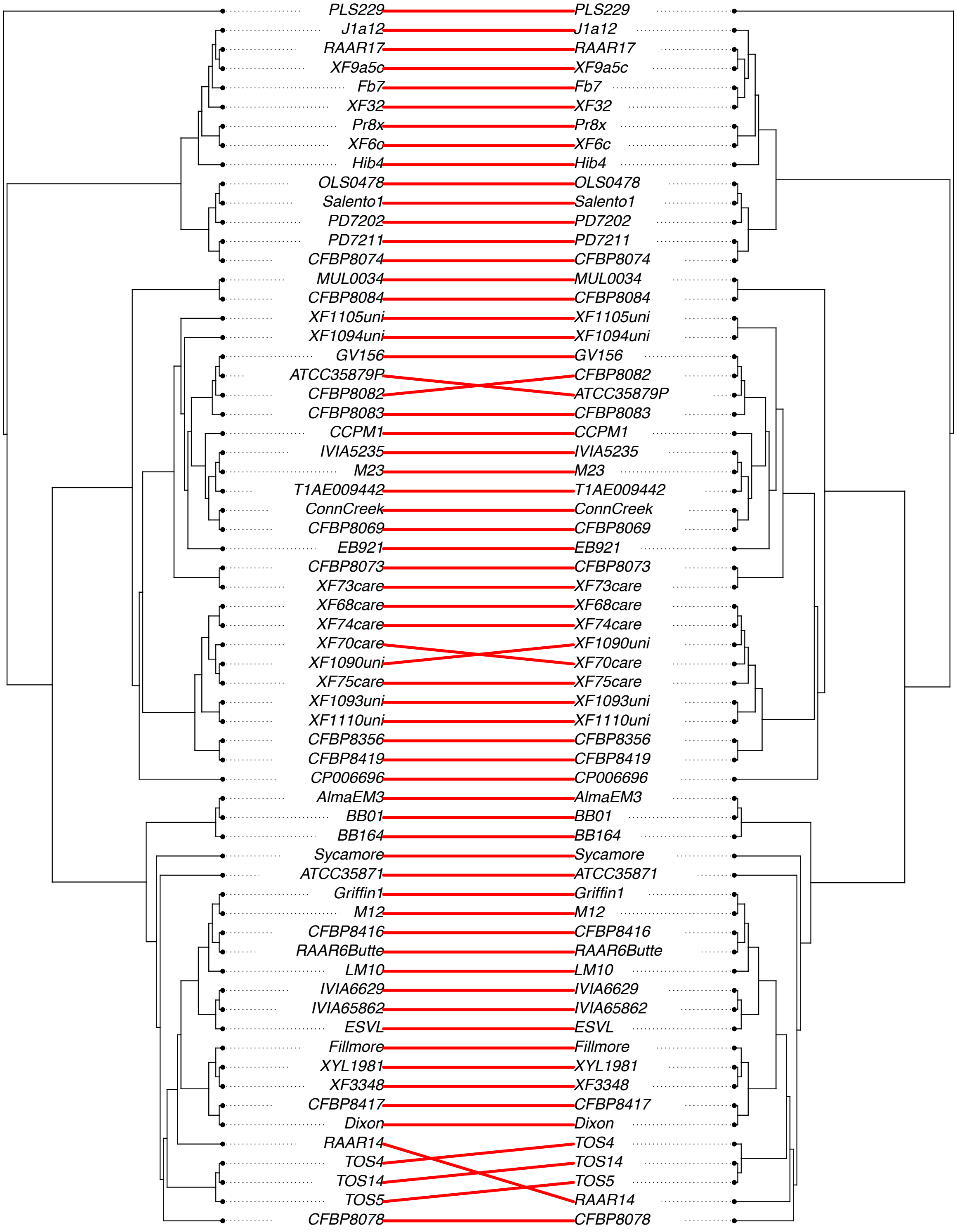

Figure S3.3: Comparison of Gubbins recombination-corrected phylogeny (left) to the phylogeny built from the entire core gene alignment (right).

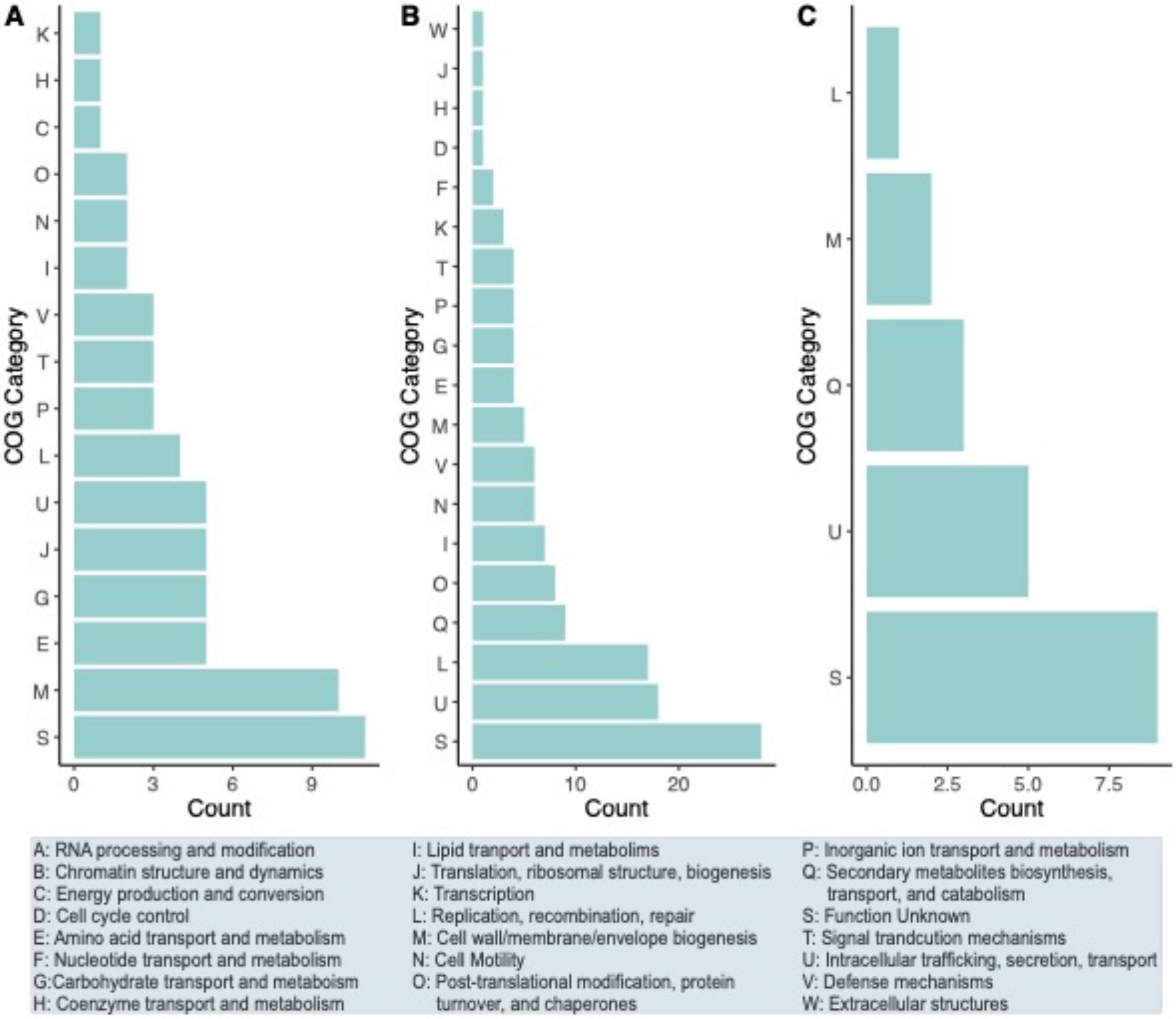

Figure S3.4: The distribution of functional categories for A) the 67 core genes, B) the 201 accessory genes, and C) 33 multicopy genes with evidence for positive selection under the sites models. A key to the COG categories is provided in the bottom panel.

Supplemental Tables

Table S3.1 Accessions gathered for this study

Isolate Name	NCBI Accession	Species	Continent of Isolation	Location of Isolation	Host	Included for Analysis
XF32	AWYH00000000	Xylella fastidiosa	South_America	Brazil	Coffee	Yes
XF3124	CP009829	Xylella fastidiosa	South_America	Brazil: Matao, Sao Paulo	Coffee	No
XF11399	JNBT00000000	Xylella fastidiosa	South_America	Brazil	Citrus	No
XF6c	AXBS00000000	Xylella fastidiosa	South_America	Brazil	Coffee	Yes
XF9a5c	AE003849	Xylella fastidiosa	South_America	Brazil: Sao Paulo, Macaubal	Citrus	Yes
AlmaEM3	PUIY00000000	Xylella fastidiosa	North_America	USA: Georgia	Blueberry	Yes
Ann-1_AAAM	AAAM00000000	Xylella fastidiosa	N/A	N/A	Oleander	No
ATCC_35871	AUAJ00000000	Xylella fastidiosa	North_America	USA:Georgia	Plum	Yes
Ann-1_CP006696	CP006696	Xylella fastidiosa	N/A	USA:California, Palm Springs	Oleander	Yes
ATCC_35879	JQAP00000000	Xylella fastidiosa	North_America	USA: Florida	Grape	No
ATCC_35879_PacBio	CP044352.1	Xylella fastidiosa	North_America	USA: Florida	Grape	Yes
Bakersfield-1	NZ_CP040799.1	Xylella fastidiosa	North_America	USA:Bakersfield	Grape	No
BB01	MPAZ00000000	Xylella fastidiosa	North_America	USA: Georgia	Blueberry	Yes

BB08-1	PUIZ00000000	Xylella fastidiosa	North_America	USA: Florida	Blueberry	No
BBI64	PUJA00000000	Xylella fastidiosa	North_America	USA: Georgia	Blueberry	Yes
CCPM1	PUJB00000000	Xylella fastidiosa	North_America	USA: Georgia	Grape	Yes
CFBP7969	PHFQ00000000.1	Xylella fastidiosa	North_America	USA: North Carolina Horticultural crops research station Castle Hayne	Grape	No
CFBP7970	PHFR00000000	Xylella fastidiosa	North_America	USA: Florida	Grape	No
CFBP8069	PRJNA833428	Xylella fastidiosa	N/A	N/A	Grape	Yes
CFBP8071	PHFP00000000	Xylella fastidiosa	North_America	USA: California	Almond	No
CFBP8072	LKDK01000000	Xylella fastidiosa	South_America	Ecuador	Coffee	No
CFBP8073	LKES01000000	Xylella fastidiosa	North_America	Mexico	Coffee	Yes
CFBP8074	PRJNA833428	Xylella fastidiosa	South_America	Ecuador	Coffee	Yes
CFBP8078	PHFS00000000	Xylella fastidiosa	North_America	USA: Florida	Periwinkle	Yes
CFBP8082	PHFT00000000	Xylella fastidiosa	North_America	USA: Florida	Ragweed	Yes
CFBP8083	PRJNA833428	Xylella fastidiosa	North_America	USA: North Carolina	Grape	Yes
CFBP8084	PRJNA833428	Xylella fastidiosa	North_America	USA: Massachusetts	Mulberry	Yes

CFBP8173	PRJNA833428	Xylella fastidiosa	North_America	USA: Georgia	Plum	No
CFBP8174	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	No
CFBP8175	PRJNA833428	Xylella fastidiosa	North_America	USA: Florida	Grape	No
CFBP8176	PRJNA833428	Xylella fastidiosa	North_America	USA: Florida	Grape	No
CFBP8177	PRJNA833428	Xylella fastidiosa	North_America	USA: Florida	Grape	No
CFBP8351	PHFU01000000	Xylella fastidiosa	North_America	USA: California Fresno county	Grape	No
CFBP8356	PHFV00000000	Xylella fastidiosa	Central_America	Costa Rica	Coffee	Yes
CFBP8416	LUYC00000000	Xylella fastidiosa	Europe	France: Corse, Propriano	Milkwort	Yes
CFBP8417	LUYB00000000	Xylella fastidiosa	Europe	France: Corse, Alata	Spartium	Yes
CFBP8418	LUYA00000000	Xylella fastidiosa	Europe	France: Corse, Alata	Spartium	No
CFBP8419	PRJNA833428	Xylella fastidiosa	Europe	France (intercepted)	Coffee	Yes
CFBP8478	PRJNA833428	Xylella fastidiosa	Europe	France (intercepted)	Coffee	No
CO33	LJZW01000000	Xylella fastidiosa	Central_America	Costa Rica	Coffee	No
CoDiRO	JUJW01000000	Xylella fastidiosa	Europe	Italy: Apulia	Olive	No
COF0324	LRVG00000000	Xylella fastidiosa	South_America	Brazil: Minas Gerais	Coffee	No

				State, Varginha		
COF0407	LRVJ00000000	Xylella fastidiosa	Central_America	Costa Rica: San Jose Province, Curridabat	Coffee	No
Conn_Creek	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	Yes
CVC0251	LRVE00000000	Xylella fastidiosa	South_America	Brazil: Sao Paulo State, Bebedouro	Citrus	No
CVC0256	LRVF00000000	Xylella fastidiosa	South_America	Brazil: Sao Paulo State, Colina	Citrus	No
De_Donno	CP020870	Xylella fastidiosa	Europe	Italy: Apulia (region)	Olive	No
Dixon	AAAL00000000	Xylella fastidiosa	N/A	N/A	Almond	Yes
EB92.1	AFDJ00000000	Xylella fastidiosa	North_America	USA: Florida	Elderberry	Yes
ESVL	QPQV01000000	Xylella fastidiosa	Europe	Spain: Benimantell, Alicante province	Almond	Yes
Fb7	CP010051	Xylella fastidiosa	South_America	Argentina: Corrientes	Citrus	Yes
Fetzer	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	No
Fillmore	CP052855.1	Xylella fastidiosa	North_America	USA:California	Olive	Yes
Fresno	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Almond	No
GB514	CP002165	Xylella fastidiosa	North_America	USA: Texas	Grape	No
Griffin-1	AVGA00000000	Xylella fastidiosa	North_America	USA	Oak	Yes

GV156	VOSD00000000.1	Xylella fastidiosa	Asia	Taiwan: Nantou County	Grape	Yes
Hib4	CP009885	Xylella fastidiosa	South_America	Brazil: Jarinu, Sao Paulo	Hibiscus	Yes
IAS-AXF-235T10	VCQO00000000.1	Xylella fastidiosa	Europe	Spain: El Castell de Guadalest (Alicante)	Almond	No
IAS-AXF212H7	VCPQ00000000.1	Xylella fastidiosa	Europe	Spain: Benimantell (Alicante)	Almond	No
IVIA5235	CP047171	Xylella fastidiosa	Europe	Spain: Mallorca Island	Sweet_Cherry	Yes
IVIA5901	CP047134	Xylella fastidiosa	Europe	Spain: Bolulla, Alicante province	Almond	No
IVIA6586-2	VDCM01000000	Xylella fastidiosa	Europe	Spain: Beniarda, Alicante region	Curry_plant	Yes
IVIA6629	VCPM01000000	Xylella fastidiosa	Europe	Spain: Callosa d'en Sarria (Alicante)	Buckthorn	Yes
IVIA6731	VCPN01000000	Xylella fastidiosa	Europe	Spain: Tarbena (Alicante)	Curry_plant	No
IVIA6902	VCPO01000000	Xylella fastidiosa	Europe	Spain: Castell de Castells (Alicante)	Almond	No
IVIA6903	VCPP00000000.1	Xylella fastidiosa	Europe	Spain: Castell de Castells (Alicante)	Almond	No
J1a12	CP009823	Xylella fastidiosa	South_America	Brazil: Jales, Sao Paulo	Citrus	Yes
Je117	SRR8144172	Xylella fastidiosa	North_America	USA, California, Temecula, Van Roekel	Grape	No

				Vineyard and Winery		
Je4	SRR8144148	Xylella fastidiosa	North_America	USA, California, Santa Barbara, Cebada Canyon road	Grape	No
Je54	SRR8144122	Xylella fastidiosa	North_America	USA, California, Napa valley, Veteran peak (nearby)	Grape	No
Je7	SRR8144115	Xylella fastidiosa	North_America	USA, California, Sonoma, Bradford Mountain	Grape	No
Je82	SRR8144197	Xylella fastidiosa	North_America	USA, California, Bakersfield, General Beale road	Grape	No
LM10	CP052854.1	Xylella fastidiosa	North_America	USA:California	Olive	Yes
M12	CP000941	Xylella fastidiosa	North_America	USA: California, Kern County in the San Joaquin Valley	Almond	Yes
M23	CP001011	Xylella fastidiosa	North_America	USA: California, Kern County in the San Joaquin Valley	Almond	Yes
Merced	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	No

Mul-MD	AXDP00000000	Xylella fastidiosa	North_America	USA: Maryland	Mulberry	No
MUL0034	CP006740	Xylella fastidiosa	North_America	N/A	Mulberry	Yes
Mus-1	AWPK00000000.1	Xylella fastidiosa	North_America	USA	Grape	No
NOB1	JABCJG000000000.1	Xylella fastidiosa	North_America	USA: Stone County, Mississippi	Grape	No
OK3	JABCJH000000000.1	Xylella fastidiosa	North_America	USA: Beaumont, Mississippi	Grape	No
OLS0478	LRVI00000000	Xylella fastidiosa	Central_America	Costa Rica: San Jose Province, Sabanilla	Oleander	Yes
OLS0479	LRVH00000000	Xylella fastidiosa	Central_America	Costa Rica: San Jose Province, Sabanilla	Oleander	No
PD7202	RRUA01000000	Xylella fastidiosa	Central_America	Costa Rica	Coffee	Yes
PD7211	RRTZ00000000	Xylella fastidiosa	Central_America	Costa Rica	Coffee	Yes
PLS229	CP053627.1	Xylella taiwanensis	Asia	Taiwan: Houli District, Taichung City	Pear	Yes
Pr8x	CP009826	Xylella fastidiosa	South_America	Brazil: Jarinu, Sao Paulo	Plum	Yes
RAAR14_plum327	VDDF00000000.1	Xylella fastidiosa	South_America	Brazil: Rio Grande do Sul, Veranopolis	Plum	Yes
RAAR15_Co33	SRR10246966	Xylella fastidiosa	South_America	Brazil: Brasilia DF	Coffee	No

RAAR16_Co13	SRR10246965	Xylella fastidiosa	South_America	Brazil: Sao Paulo, Cordeiropolis	Coffee	No
RAAR17_CiUb7	SRR10246964	Xylella fastidiosa	South_America	Brazil, Minas Gerais, Lavras	Coffee	Yes
RAAR6_Butte	VDDE00000000.1	Xylella fastidiosa	North America	USA: Butte County, California	Almond	Yes
RH1	CP052853.1	Xylella fastidiosa	North_America	USA:California	Olive	No
Salento-1	CP016608	Xylella fastidiosa	Europe	Italy: Taviano, Lecce, Apulia	Olive	Yes
Salento-2	CP016610	Xylella fastidiosa	Europe	Italy: Ugento, Lecce, Apulia	Olive	No
Stag's_Leap	LSMJ00000000	Xylella fastidiosa	North_America	USA: Napa Valley, California	Grape	No
Sycamore_Sy-VA	JMHP01000000	Xylella fastidiosa	North_America	USA: Virginia	Sycamore	Yes
Temecula_2	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	No
Temecula1_AE009442	AE009442	Xylella fastidiosa	North_America	USA: California	Grape	Yes
Temecula1Star	PUJI00000000	Xylella fastidiosa	North_America	USA: California	Grape	No
TemeculaL	PUJJ00000000	Xylella fastidiosa	North_America	USA: California	Grape	No
TOS14	SMTJ00000000	Xylella fastidiosa	Europe	Italy: Tuscany	Spartium	Yes
TOS4	SMTH00000000	Xylella fastidiosa	Europe	Italy: Tuscany	Almond	Yes

TOS5	SMTI00000000	Xylella fastidiosa	Europe	Italy: Tuscany	Milkwort	Yes
TPD3	VJWG01000000	Xylella fastidiosa	Asia	Taiwan: Hou-li	Grape	No
TPD4	VJWH01000000	Xylella fastidiosa	Asia	Taiwan: Hou-li	Grape	No
Traver	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	No
U24D	CP009790	Xylella fastidiosa	South_America	Brazil: Ubarana, Sao Paulo	Citrus	No
UCLA	PRJNA833428	Xylella fastidiosa	North_America	USA: California	Grape	No
VB11	JABCJI000000000.1	Xylella fastidiosa	North_America	USA: Beaumont, Mississippi	Grape	No
WM1-1	PUJK00000000	Xylella fastidiosa	North_America	USA: Georgia	Grape	No
XF1090	SRR10246940	Xylella fastidiosa	Central_America	Costa Rica: San Rafael de Montes de Oca\, San Jose	Coffee	Yes
XF1093	SRR10246939	Xylella fastidiosa	Central_America	Costa Rica: San Rafael de Montes de Oca\, San Jose	Coffee	Yes
XF1094	SRR10246938	Xylella fastidiosa	Central_America	Costa Rica: Vargas Araya de San Pedro de Montes de Oca\, San Jose	Periwinkle	Yes
XF1105	SRR10246935	Xylella fastidiosa	Central_America	Costa Rica: San Rafael de Montes de Oca\, San Jose	Coffee	Yes

XF1110	SRR10246946	Xylella fastidiosa	Central_America	Costa Rica: Santiago\, Puriscal\, San Jose	Periwinkle	Yes
XF3348	VDCL00000000.1	Xylella fastidiosa	Europe	Spain: Binissalem, Mallorca	Almond	Yes
XF68	SRR10246937	Xylella fastidiosa	Central_America	Costa Rica: Coronado\, Coronado\, San Jose	Psidium	Yes
XF70	SRR10246936	Xylella fastidiosa	Central_America	Costa Rica: San Rafael de Montes de Oca\, San Jose	Coffee	Yes
XF71	SRR10246945	Xylella fastidiosa	Central_America	Costa Rica: San Rafael de Montes de Oca\, San Jose	Coffee	No
XF72	SRR10246944	Xylella fastidiosa	Central_America	Costa Rica: Granadilla\, Curridabat\, San Jose	Coffee	No
XF73	SRR10246943	Xylella fastidiosa	Central_America	Costa Rica: Santa Rosa\, Santo Domingo\, Heredia	Coffee	Yes
XF74	SRR10246942	Xylella fastidiosa	Central_America	Costa Rica: San Rafael de Montes de Oca\, San Jose	Coffee	Yes
XF75	SRR10246941	Xylella fastidiosa	Central_America	Costa Rica: Granadilla\, Curridabat\, San Jose	Coffee	Yes
XYL1732	QTJT01000000	Xylella fastidiosa	Europe	Spain: Mallorca	Grape	No

XYL1752	VDCK00000000.1	Xylella fastidiosa	Europe	Spain: Ciutadella, Menorca	Almond	No
XYL1981	VDCJ00000000.1	Xylella fastidiosa	Europe	Spain: Campos, Mallorca	Mulberry	Yes
XYL2055	QTJS00000000	Xylella fastidiosa	Europe	Spain: Mallorca	Grape	No

Table S3.2 *X. fastidiosa* genome assemblies for this study

Strain	SRA Number	Host	Assembly_pipeline	Assembly size (bp)	Total Contigs	N50 (bp)	Contigs (>500 bp)	Assembly size of contigs >500 bp	Reference
XF1105	SRR10246935, SRR10246947	Coffee	Unicycler	2584447	363	122384	106	2539589	Castillo et al. 2020
XF70	SRR10246936	Coffee	Spades	2598424	283	99440	115	2554289	Castillo et al. 2020
XF68	SRR10246937	Psidium	Spades	2646695	343	100059	124	2585361	Castillo et al. 2020
XF1094	SRR10246938, SRR10246948	Periwinkle	Unicycler	2590434	539	144534	106	2512534	Castillo et al. 2020
XF1093	SRR10246939, SRR10246949	Coffee	Unicycler	2643497	385	98804	95	2587259	Castillo et al. 2020
XF1090	SRR10246940, SRR10246950	Coffee	Unicycler	2625015	484	97769	119	2562374	Castillo et al. 2020
XF75	SRR10246941	Coffee	Spades	2652878	318	103912	135	2599855	Castillo et al. 2020
XF74	SRR10246942	Coffee	Spades	2543780	314	134662	115	2486909	Castillo et al. 2020
XF73	SRR10246943	Coffee	Spades	2617297	249	110412	103	2575291	Castillo et al. 2020
XF72	SRR10246944	Coffee	Spades	2597649	326	97666	128	2542894	Castillo et al. 2020
XF71	SRR10246945	Coffee	Spades	2667910	373	116569	147	2601030	Castillo et al. 2020
XF1110	SRR10246951, SRR10246946	Periwinkle	Unicycler	2604815	437	98810	109	2543556	Castillo et al. 2020
RAAR15_Co33	SRR10246966	Coffee	Spades	2646151	319	173295	69	2604719	Castillo et al. 2020
RAAR16_Co13	SRR10246965	Coffee	Spades	2722408	557	134428	91	2650941	Castillo et al. 2020
RAAR17_CiUb7	SRR10246964	Coffee	Spades	2663824	296	145178	71	2625892	Castillo et al. 2020
Je117	SRR8144172	Grape	Spades	2549513	219	140934	92	2516893	Vanhove et al. 2020
Je119	SRR8144174	Grape	Spades	2502022	207	116022	95	2472914	Vanhove et al. 2020

Je34	SRR8144142	Grape	Spades	2542080	213	116022	93	2512422	Vanhove et al. 2020
Je4	SRR8144148	Grape	Spades	2516356	209	115997	97	2485913	Vanhove et al. 2020
Je5	SRR8144134	Grape	Spades	2527921	249	103359	109	2487035	Vanhove et al. 2020
Je54	SRR8144122	Grape	Spades	2542091	215	116022	89	2511024	Vanhove et al. 2020
Je7	SRR8144115	Grape	Spades	2510981	177	150327	87	2486735	Vanhove et al. 2020
Je76	SRR8144211	Grape	Spades	2530691	214	117374	94	2496269	Vanhove et al. 2020
Je82	SRR8144197	Grape	Spades	2522635	202	117383	90	2491922	Vanhove et al. 2020
Je93	SRR8144226	Grape	Spades	2510569	183	140443	88	2486976	Vanhove et al. 2020
CFBP8069	PRJNA833428	N/A	Unicycler	2463982	215	215819	96	2437553	This study
CFBP8070	PRJNA833428	Plum	Unicycler	4771158	241	197850	115	4741623	This study
CFBP8074	PRJNA833428	Coffee	Spades	2509164	277	139814	126	2471244	This study
CFBP8075	PRJNA833428	N/A	Unicycler	2373572	143	164556	70	2357337	This study
CFBP8083	PRJNA833428	Grape	Unicycler	2475178	61	618774	33	2468874	This study
CFBP8084	PRJNA833428	Mulberry	Unicycler	2496031	277	94284	113	2456865	This study
CFBP8173	PRJNA833428	Plum	Unicycler	2432302	145	165563	67	2412773	This study
CFBP8174	PRJNA833428	Grape	Unicycler	2456916	214	106132	109	2430570	This study
CFBP8175	PRJNA833428	Grape	Unicycler	2473018	207	126063	97	2447821	This study
CFBP8176	PRJNA833428	Grape	Unicycler	2492699	145	290107	58	2473325	This study
CFBP8177	PRJNA833428	Grape	Unicycler	2481282	133	350837	61	2464348	This study
CFBP8419	PRJNA833428	Grape	Unicycler	2590604	146	222180	85	2577583	This study
CFBP8478	PRJNA833428	Coffee	Unicycler	2589739	173	159384	86	2569985	This study
Conn_Creek	PRJNA833428	Grape	Unicycler	2538857	42	1304995	20	2533765	This study
Fetzer	PRJNA833428	Grape	Unicycler	2511488	78	1266956	33	2501347	This study
Fresno	PRJNA833428	Almond	Unicycler	2552104	58	1628821	25	2544510	This study
Merced	PRJNA833428	Grape	Unicycler	2542264	85	555806	33	2528991	This study

Temecula_2	PRJNA833428	Grape	Unicycler	2487898	79	466303	37	2478574	This study
Traver	PRJNA833428	Grape	Unicycler	2541057	42	623430	27	2537049	This study
UCLA	PRJNA833428	Grape	Unicycler	2630235	109	667954	51	2616493	This study

Table S3.3 CheckM results for *Xylella* genomes

		Number of Copies*						
Bin id	**Included in study**	**0**[a]	**1**[b]	**2**[c]	**3**[d]	**Completeness**	**Contamination**	**Strain heterogeneity**
XF32	Yes	1	480	0	0	99.64	0	0
XF6c	Yes	2	478	1	0	99.28	0.18	0
XF9a5c	Yes	2	478	1	0	99.59	0.18	0
AlmaEM3	Yes	1	480	0	0	99.64	0	0
ATCC35871	Yes	1	480	0	0	99.64	0	0
ATCC35879P	Yes	3	478	0	0	98.91	0	0
BB01	Yes	1	480	0	0	99.64	0	0
BB164	Yes	3	444	29	5	98.91	7.19	100
CCPM1	Yes	2	468	10	1	99.28	3.08	100
CFBP8069	Yes	1	480	0	0	99.64	0	0
CFBP8073	Yes	2	479	0	0	99.63	0	0
CFBP8074	Yes	1	480	0	0	99.64	0	0
CFBP8078	Yes	1	480	0	0	99.64	0	0
CFBP8082	Yes	1	480	0	0	99.64	0	0
CFBP8083	Yes	1	480	0	0	99.64	0	0
CFBP8084	Yes	1	480	0	0	99.64	0	0
CFBP8356	Yes	2	479	0	0	99.28	0	0
CFBP8416	Yes	7	473	1	0	98.31	0.02	100
CFBP8417	Yes	1	480	0	0	99.64	0	0
CFBP8419	Yes	2	478	1	0	99.28	0.36	0
ConnCreek	Yes	1	480	0	0	99.64	0	0
Dixon	Yes	2	479	0	0	99.63	0	0
EB921	Yes	1	479	1	0	99.64	0.18	0
ESVL	Yes	1	479	1	0	99.64	0.36	0
Fb7	Yes	5	476	0	0	98.78	0	0
Fillmore	Yes	1	480	0	0	99.64	0	0
Griffin1	Yes	5	476	0	0	98.46	0	0
GV156	Yes	2	479	0	0	99.46	0	0
Hib4	Yes	1	474	6	0	99.64	1.45	83.33
IVIA5235	Yes	1	480	0	0	99.64	0	0
IVIA65862	Yes	1	478	2	0	99.64	0.72	50

IVIA6629	Yes	1	474	6	0	99.64	1.16	16.67
J1a12	Yes	2	478	1	0	99.59	0.18	0
LM10	Yes	1	480	0	0	99.64	0	0
M12	Yes	1	480	0	0	99.64	0	0
M23	Yes	1	480	0	0	99.64	0	0
MUL0034	Yes	1	480	0	0	99.64	0	0
OLS0478	Yes	1	480	0	0	99.64	0	0
PD7202	Yes	9	472	0	0	98.14	0	0
PD7211	Yes	1	480	0	0	99.64	0	0
PLS229	Yes	50	431	0	0	89.25	0	0
Pr8x	Yes	2	478	1	0	99.59	0.18	0
RAAR14	Yes	1	479	1	0	99.64	0.18	0
RAAR17	Yes	1	479	1	0	99.64	0.18	0
RAAR6Butte	Yes	1	480	0	0	99.64	0	0
Salento1	Yes	4	477	0	0	99.11	0	0
Sycamore	Yes	1	479	1	0	99.64	0.09	100
Temecula1_AE009442	Yes	1	480	0	0	99.64	0	0
TOS14	Yes	1	480	0	0	99.64	0	0
TOS4	Yes	1	479	1	0	99.64	0.36	100
TOS5	Yes	1	480	0	0	99.64	0	0
XF1090uni	Yes	1	480	0	0	99.64	0	0
XF1093uni	Yes	1	479	1	0	99.64	0.36	0
XF1094uni	Yes	2	479	0	0	99.28	0	0
XF1105uni	Yes	2	479	0	0	99.28	0	0
XF1110uni	Yes	2	479	0	0	99.59	0	0
XF68care	Yes	1	479	1	0	99.64	0.36	0
XF70care	Yes	1	480	0	0	99.64	0	0
XF73care	Yes	1	480	0	0	99.64	0	0
XF74care	Yes	1	480	0	0	99.64	0	0
XF75care	Yes	1	480	0	0	99.64	0	0
XYL1981	Yes	1	479	1	0	99.64	0.12	0
CP006696	Yes	1	480	0	0	99.64	0	0
XF3348uni	Yes	1	472	8	0	99.64	1.37	87.5
XF3124	No	1	480	0	0	99.64	0	0
XF11399	No	1	479	1	0	99.64	0.18	0
Ann1AAAM	No	9	469	3	0	98.29	0.74	66.67
ATCC35879	No	3	478	0	0	99.23	0	0

AXF212H7	No	1	476	4	0	99.64	0.41	25
AXF235T10	No	1	473	7	0	99.64	1.1	57.14
BB081	No	2	465	12	2	99.52	5.11	100
Bakersfield-1	No	1	480	0	0	99.64	0	0
CFBP7969	No	1	480	0	0	99.64	0	0
CFBP7970	No	3	478	0	0	99.23	0	0
CFBP8071	No	2	478	1	0	99.59	0.12	0
CFBP8072	No	2	479	0	0	99.59	0	0
CFBP8173	No	1	480	0	0	99.64	0	0
CFBP8174	No	1	480	0	0	99.64	0	0
CFBP8175	No	1	480	0	0	99.64	0	0
CFBP8176	No	2	479	0	0	99.59	0	0
CFBP8177	No	1	477	3	0	99.64	0.91	100
CFBP8351	No	1	480	0	0	99.64	0	0
CFBP8418	No	1	480	0	0	99.64	0	0
CFBP8478	No	2	479	0	0	99.28	0	0
CO33	No	2	479	0	0	99.28	0	0
CoDiRo	No	1	480	0	0	99.64	0	0
COF0324	No	1	479	1	0	99.64	0.18	0
COF0407	No	2	479	0	0	99.28	0	0
CVC0251	No	1	479	1	0	99.64	0.18	0
CVC0256	No	4	476	1	0	98.91	0.18	0
DeDonno	No	2	479	0	0	99.59	0	0
Fetzer	No	1	480	0	0	99.64	0	0
Fresno	No	1	480	0	0	99.64	0	0
GB514	No	3	478	0	0	99.18	0	0
IVIA5901	No	1	480	0	0	99.64	0	0
IVIA6731	No	1	475	5	0	99.64	1.13	40
IVIA6902	No	1	467	13	0	99.64	2.05	23.08
IVIA6903	No	1	474	6	0	99.64	0.97	33.33
Je117	No	3	478	0	0	99.28	0	0
Je4	No	1	480	0	0	99.64	0	0
Je54	No	3	478	0	0	99.28	0	0
Je7	No	3	478	0	0	99.28	0	0
Je82	No	1	480	0	0	99.64	0	0
Merced	No	1	480	0	0	99.64	0	0
MulMD	No	1	480	0	0	99.64	0	0

Mus1	No	4	476	1	0	99.29	0.36	100
NOB1	No	1	480	0	0	99.64	0	0
OK3	No	1	480	0	0	99.64	0	0
OLS0479	No	3	477	1	0	99.23	0.36	0
PLS235	No	54	427	0	0	88.87	0	0
PLS244	No	50	431	0	0	89.25	0	0
RAAR15	No	1	480	0	0	99.64	0	0
RAAR16	No	2	478	1	0	99.59	0.18	0
RH1	No	1	476	4	0	99.64	1.45	100
Salento2	No	2	479	0	0	99.59	0	0
StagsLeap	No	1	480	0	0	99.64	0	0
T1Star	No	1	480	0	0	99.64	0	0
T2	No	1	480	0	0	99.64	0	0
TL	No	1	480	0	0	99.64	0	0
TPD3	No	1	479	1	0	99.64	0.36	100
TPD4	No	1	478	2	0	99.64	0.21	100
Traver	No	1	480	0	0	99.64	0	0
U24D	No	2	478	1	0	99.59	0.18	0
UCLA	No	1	479	1	0	99.64	0.36	0
VB11	No	1	479	1	0	99.64	0.01	100
WM11	No	1	480	0	0	99.64	0	0
XF71care	No	1	480	0	0	99.64	0	0
XF72care	No	1	480	0	0	99.64	0	0
XYL1732	No	1	480	0	0	99.64	0	0
XYL1752	No	1	477	3	0	99.64	0.25	33.33
XYL2055	No	1	477	3	0	99.64	0.91	0

* The reference set of single copy genes included a total of 481 gene markers
a Number of marker genes not found in the genome
b Count of marker genes found with a single copy in the genome
c Count of marker genes found duplicated in the genome
d Count of marker genes found with three copies in the genome

Table S3.4: GO analysis of core genes

GO biological process complete	REFLIST	upload	expected	over/under	fold Enrichment	P-value
pyrimidine-containing compound biosynthetic process (GO:0072528)	20	16	2.99	+	5.35	1.91E-03
translation (GO:0006412)	105	73	15.71	+	4.65	8.14E-19
aromatic amino acid family biosynthetic process (GO:0009073)	22	15	3.29	+	4.56	1.51E-02
regulation of developmental process (GO:0050793)	28	19	4.19	+	4.54	1.32E-03
peptide biosynthetic process (GO:0043043)	113	76	16.91	+	4.5	4.72E-19
regulation of cell shape (GO:0008360)	27	18	4.04	+	4.46	2.93E-03
regulation of cell morphogenesis (GO:0022604)	27	18	4.04	+	4.46	2.93E-03
regulation of anatomical structure morphogenesis (GO:0022603)	27	18	4.04	+	4.46	2.93E-03
ribonucleoside monophosphate metabolic process (GO:0009161)	29	19	4.34	+	4.38	1.93E-03
amide biosynthetic process (GO:0043604)	140	91	20.95	+	4.34	1.53E-22

nucleoside monophosphate biosynthetic process (GO:0009124)	31	20	4.64	+	4.31	1.26E-03
peptide metabolic process (GO:0006518)	133	85	19.9	+	4.27	1.63E-20
nucleoside monophosphate metabolic process (GO:0009123)	36	23	5.39	+	4.27	2.46E-04
RNA methylation (GO:0001510)	22	14	3.29	+	4.25	4.84E-02
ribonucleoside monophosphate biosynthetic process (GO:0009156)	27	17	4.04	+	4.21	9.32E-03
pyrimidine-containing compound metabolic process (GO:0072527)	27	17	4.04	+	4.21	9.32E-03
cell cycle (GO:0007049)	42	26	6.28	+	4.14	6.80E-05
cellular amide metabolic process (GO:0043603)	178	104	26.63	+	3.91	2.08E-23
ribonucleotide biosynthetic process (GO:0009260)	52	30	7.78	+	3.86	2.33E-05
ribose phosphate biosynthetic process (GO:0046390)	53	30	7.93	+	3.78	3.21E-05
ribonucleotide metabolic process (GO:0009259)	78	44	11.67	+	3.77	2.47E-08
glycosaminoglycan biosynthetic process (GO:0006024)	32	18	4.79	+	3.76	1.64E-02

aminoglycan biosynthetic process (GO:0006023)	32	18	4.79	+	3.76	1.64E-02
peptidoglycan biosynthetic process (GO:0009252)	32	18	4.79	+	3.76	1.64E-02
ribosome biogenesis (GO:0042254)	47	26	7.03	+	3.7	3.50E-04
ribonucleoprotein complex biogenesis (GO:0022613)	47	26	7.03	+	3.7	3.50E-04
ribose phosphate metabolic process (GO:0019693)	82	45	12.27	+	3.67	2.94E-08
purine ribonucleotide biosynthetic process (GO:0009152)	42	23	6.28	+	3.66	1.79E-03
monocarboxylic acid biosynthetic process (GO:0072330)	42	23	6.28	+	3.66	1.79E-03
cellular component macromolecule biosynthetic process (GO:0070589)	33	18	4.94	+	3.65	2.23E-02
purine ribonucleotide metabolic process (GO:0009150)	66	36	9.87	+	3.65	3.05E-06
cell wall macromolecule biosynthetic process (GO:0044038)	33	18	4.94	+	3.65	2.23E-02
cellular macromolecule	178	97	26.63	+	3.64	6.22E-20

biosynthetic process (GO:0034645)						
gene expression (GO:0010467)	244	132	36.5	+	3.62	3.41E-28
tRNA metabolic process (GO:0006399)	65	35	9.72	+	3.6	6.45E-06
organonitrogen compound biosynthetic process (GO:1901566)	406	218	60.74	+	3.59	4.74E-51
sulfur compound biosynthetic process (GO:0044272)	56	30	8.38	+	3.58	8.05E-05
nucleotide biosynthetic process (GO:0009165)	71	38	10.62	+	3.58	1.69E-06
peptidoglycan-based cell wall biogenesis (GO:0009273)	36	19	5.39	+	3.53	1.92E-02
nucleoside phosphate biosynthetic process (GO:1901293)	72	38	10.77	+	3.53	2.27E-06
tRNA processing (GO:0008033)	40	21	5.98	+	3.51	7.90E-03
purine nucleotide biosynthetic process (GO:0006164)	44	23	6.58	+	3.49	3.25E-03
vitamin biosynthetic process (GO:0009110)	46	24	6.88	+	3.49	2.08E-03
water-soluble vitamin biosynthetic	46	24	6.88	+	3.49	2.08E-03

process (GO:0042364)						
purine nucleotide metabolic process (GO:0006163)	69	36	10.32	+	3.49	7.43E-06
cellular amino acid biosynthetic process (GO:0008652)	100	52	14.96	+	3.48	4.04E-09
cell division (GO:0051301)	54	28	8.08	+	3.47	3.51E-04
cellular nitrogen compound biosynthetic process (GO:0044271)	365	189	54.61	+	3.46	5.09E-41
cell wall biogenesis (GO:0042546)	37	19	5.54	+	3.43	2.56E-02
phospholipid biosynthetic process (GO:0008654)	45	23	6.73	+	3.42	4.32E-03
phospholipid metabolic process (GO:0006644)	45	23	6.73	+	3.42	4.32E-03
organophosphate biosynthetic process (GO:0090407)	135	69	20.2	+	3.42	2.08E-12
purine-containing compound biosynthetic process (GO:0072522)	47	24	7.03	+	3.41	2.77E-03
water-soluble vitamin metabolic process (GO:0006767)	49	25	7.33	+	3.41	1.77E-03
vitamin metabolic process (GO:0006766)	49	25	7.33	+	3.41	1.77E-03

ncRNA metabolic process (GO:0034660)	98	50	14.66	+	3.41	1.78E-08
alpha-amino acid biosynthetic process (GO:1901607)	85	43	12.72	+	3.38	5.63E-07
carboxylic acid biosynthetic process (GO:0046394)	157	79	23.49	+	3.36	3.26E-14
organic acid biosynthetic process (GO:0016053)	159	79	23.79	+	3.32	5.71E-14
RNA processing (GO:0006396)	77	38	11.52	+	3.3	9.26E-06
ncRNA processing (GO:0034470)	73	36	10.92	+	3.3	2.27E-05
external encapsulating structure organization (GO:0045229)	41	20	6.13	+	3.26	2.84E-02
nucleotide metabolic process (GO:0009117)	107	52	16.01	+	3.25	2.83E-08
small molecule biosynthetic process (GO:0044283)	225	109	33.66	+	3.24	1.31E-19
sulfur compound metabolic process (GO:0006790)	93	45	13.91	+	3.23	6.77E-07
carbohydrate derivative biosynthetic process (GO:1901137)	152	73	22.74	+	3.21	3.79E-12
dicarboxylic acid metabolic process (GO:0043648)	50	24	7.48	+	3.21	6.23E-03
nucleoside phosphate	111	53	16.61	+	3.19	3.04E-08

metabolic process (GO:0006753)						
purine-containing compound metabolic process (GO:0072521)	78	37	11.67	+	3.17	3.20E-05
cellular biosynthetic process (GO:0044249)	616	292	92.16	+	3.17	1.39E-64
organic substance biosynthetic process (GO:1901576)	625	294	93.51	+	3.14	1.42E-64
organophosphate metabolic process (GO:0019637)	185	87	27.68	+	3.14	2.42E-14
macromolecule biosynthetic process (GO:0009059)	251	118	37.55	+	3.14	1.49E-20
aromatic compound biosynthetic process (GO:0019438)	235	109	35.16	+	3.1	2.01E-18
organic cyclic compound biosynthetic process (GO:1901362)	266	122	39.8	+	3.07	1.01E-20
biosynthetic process (GO:0009058)	650	296	97.25	+	3.04	1.50E-62
heterocycle biosynthetic process (GO:0018130)	251	113	37.55	+	3.01	2.33E-18
RNA modification (GO:0009451)	56	25	8.38	+	2.98	1.74E-02
RNA metabolic process (GO:0016070)	170	75	25.43	+	2.95	6.31E-11

nucleobase-containing small molecule metabolic process (GO:0055086)	147	64	21.99	+	2.91	7.92E-09
cellular component biogenesis (GO:0044085)	148	64	22.14	+	2.89	9.35E-09
DNA repair (GO:0006281)	72	31	10.77	+	2.88	2.25E-03
nucleobase-containing compound biosynthetic process (GO:0034654)	165	71	24.69	+	2.88	6.94E-10
nucleic acid phosphodiester bond hydrolysis (GO:0090305)	56	24	8.38	+	2.86	3.43E-02
regulation of biological quality (GO:0065008)	73	31	10.92	+	2.84	2.71E-03
carbohydrate derivative metabolic process (GO:1901135)	231	98	34.56	+	2.84	3.48E-14
lipid biosynthetic process (GO:0008610)	95	40	14.21	+	2.81	1.80E-04
cellular amino acid metabolic process (GO:0006520)	190	80	28.43	+	2.81	5.80E-11
cellular nitrogen compound metabolic process (GO:0034641)	703	295	105.18	+	2.8	1.13E-55
alpha-amino acid metabolic process (GO:1901605)	138	57	20.65	+	2.76	6.99E-07
cellular component organization or	197	81	29.47	+	2.75	1.33E-10

biogenesis (GO:0071840)						
carboxylic acid metabolic process (GO:0019752)	327	134	48.92	+	2.74	2.14E-19
oxoacid metabolic process (GO:0043436)	334	136	49.97	+	2.72	1.32E-19
small molecule metabolic process (GO:0044281)	507	203	75.85	+	2.68	2.22E-31
cellular response to DNA damage stimulus (GO:0006974)	80	32	11.97	+	2.67	6.09E-03
organic acid metabolic process (GO:0006082)	341	136	51.02	+	2.67	6.21E-19
cellular component organization (GO:0016043)	152	60	22.74	+	2.64	9.23E-07
monocarboxylic acid metabolic process (GO:0032787)	123	48	18.4	+	2.61	5.03E-05
cellular response to stress (GO:0033554)	107	41	16.01	+	2.56	7.27E-04
organic cyclic compound metabolic process (GO:1901360)	614	233	91.86	+	2.54	2.41E-34
heterocycle metabolic process (GO:0046483)	589	223	88.12	+	2.53	3.20E-32
cellular aromatic compound metabolic process (GO:0006725)	587	220	87.82	+	2.51	4.84E-31
cellular lipid metabolic process (GO:0044255)	127	47	19	+	2.47	2.53E-04

cellular protein metabolic process (GO:0044267)	255	93	38.15	+	2.44	5.67E-10
nucleobase-containing compound metabolic process (GO:0006139)	495	179	74.06	+	2.42	2.37E-22
organonitrogen compound metabolic process (GO:1901564)	781	280	116.85	+	2.4	4.72E-40
cellular metabolic process (GO:0044237)	1323	455	197.93	+	2.3	1.26E-83
nitrogen compound metabolic process (GO:0006807)	1124	384	168.16	+	2.28	2.39E-60
response to stress (GO:0006950)	143	48	21.39	+	2.24	2.57E-03
nucleic acid metabolic process (GO:0090304)	354	118	52.96	+	2.23	3.85E-11
phosphorus metabolic process (GO:0006793)	328	107	49.07	+	2.18	2.48E-09
cellular macromolecule metabolic process (GO:0044260)	567	184	84.83	+	2.17	1.59E-18
phosphate-containing compound metabolic process (GO:0006796)	320	103	47.88	+	2.15	1.26E-08
lipid metabolic process (GO:0006629)	150	48	22.44	+	2.14	6.63E-03
organic substance metabolic process (GO:0071704)	1435	449	214.69	+	2.09	7.90E-69

primary metabolic process (GO:0044238)	1277	394	191.05	+	2.06	6.86E-52
protein metabolic process (GO:0019538)	383	116	57.3	+	2.02	1.81E-08
metabolic process (GO:0008152)	1590	477	237.88	+	2.01	3.77E-72
macromolecule metabolic process (GO:0043170)	821	239	122.83	+	1.95	2.28E-20
cellular process (GO:0009987)	1918	543	286.95	+	1.89	2.91E-90
biological_process (GO:0008150)	2336	582	349.49	+	1.67	3.82E-88
Unclassified (UNCLASSIFIED)	1768	32	264.51	-	0.12	0.00E+00

Table S3.5: GO analysis of accessory genes

GO biological process complete	REFLIST	upload	expected	over/under	fold Enrichment	P-value
glutamine family amino acid biosynthetic process (GO:0009084)	17	8	0.94	+	8.51	1.46E-02
glutamine metabolic process (GO:0006541)	17	8	0.94	+	8.51	1.46E-02
tRNA aminoacylation for protein translation (GO:0006418)	25	10	1.38	+	7.23	4.31E-03
tRNA aminoacylation (GO:0043039)	25	10	1.38	+	7.23	4.31E-03
amino acid activation (GO:0043038)	26	10	1.44	+	6.95	5.69E-03
chromosome organization (GO:0051276)	27	9	1.49	+	6.03	4.03E-02
glutamine family amino acid metabolic process (GO:0009064)	37	11	2.05	+	5.37	1.51E-02
purine nucleotide biosynthetic process (GO:0006164)	44	12	2.43	+	4.93	1.34E-02
tRNA metabolic process (GO:0006399)	65	17	3.6	+	4.73	4.18E-04
purine-containing compound biosynthetic process (GO:0072522)	47	12	2.6	+	4.62	2.33E-02
ribose phosphate biosynthetic	53	13	2.93	+	4.43	1.59E-02

process (GO:0046390)						
ncRNA metabolic process (GO:0034660)	98	24	5.42	+	4.43	5.69E-06
nucleotide biosynthetic process (GO:0009165)	71	15	3.93	+	3.82	1.76E-02
cellular amino acid biosynthetic process (GO:0008652)	100	21	5.53	+	3.8	4.68E-04
translation (GO:0006412)	105	22	5.81	+	3.79	2.61E-04
nucleoside phosphate biosynthetic process (GO:1901293)	72	15	3.98	+	3.77	2.03E-02
ncRNA processing (GO:0034470)	73	15	4.04	+	3.71	2.34E-02
alpha-amino acid biosynthetic process (GO:1901607)	85	17	4.7	+	3.62	9.63E-03
RNA processing (GO:0006396)	77	15	4.26	+	3.52	4.00E-02
peptide biosynthetic process (GO:0043043)	113	22	6.25	+	3.52	7.74E-04
amide biosynthetic process (GO:0043604)	140	27	7.74	+	3.49	5.23E-05
cellular amino acid metabolic process (GO:0006520)	190	35	10.51	+	3.33	1.49E-06
RNA metabolic process (GO:0016070)	170	31	9.4	+	3.3	1.67E-05

peptide metabolic process (GO:0006518)	133	23	7.36	+	3.13	2.63E-03
organonitrogen compound biosynthetic process (GO:1901566)	406	68	22.46	+	3.03	3.08E-13
carboxylic acid biosynthetic process (GO:0046394)	157	26	8.68	+	2.99	1.23E-03
organic acid biosynthetic process (GO:0016053)	159	26	8.79	+	2.96	1.52E-03
gene expression (GO:0010467)	244	39	13.5	+	2.89	5.90E-06
cellular amide metabolic process (GO:0043603)	178	28	9.85	+	2.84	1.20E-03
alpha-amino acid metabolic process (GO:1901605)	138	21	7.63	+	2.75	3.83E-02
cellular macromolecule biosynthetic process (GO:0034645)	178	27	9.85	+	2.74	3.44E-03
small molecule biosynthetic process (GO:0044283)	225	34	12.45	+	2.73	1.99E-04
carboxylic acid metabolic process (GO:0019752)	327	48	18.09	+	2.65	1.45E-06
oxoacid metabolic process (GO:0043436)	334	49	18.47	+	2.65	9.46E-07
cellular nitrogen compound biosynthetic process (GO:0044271)	365	53	20.19	+	2.63	2.03E-07

organic acid metabolic process (GO:0006082)	341	49	18.86	+	2.6	1.49E-06
small molecule metabolic process (GO:0044281)	507	70	28.04	+	2.5	7.71E-10
nucleic acid metabolic process (GO:0090304)	354	47	19.58	+	2.4	3.16E-05
cellular component organization or biogenesis (GO:0071840)	197	26	10.9	+	2.39	4.79E-02
cellular biosynthetic process (GO:0044249)	616	81	34.07	+	2.38	7.21E-11
macromolecule biosynthetic process (GO:0009059)	251	33	13.88	+	2.38	6.22E-03
biosynthetic process (GO:0009058)	650	85	35.95	+	2.36	1.44E-11
organic substance biosynthetic process (GO:1901576)	625	81	34.57	+	2.34	1.28E-10
nucleobase-containing compound metabolic process (GO:0006139)	495	64	27.38	+	2.34	1.73E-07
cellular aromatic compound metabolic process (GO:0006725)	587	74	32.47	+	2.28	9.26E-09
heterocycle metabolic process (GO:0046483)	589	74	32.58	+	2.27	1.04E-08
cellular protein metabolic process (GO:0044267)	255	32	14.1	+	2.27	2.65E-02

organic cyclic compound metabolic process (GO:1901360)	614	77	33.96	+	2.27	3.34E-09
cellular nitrogen compound metabolic process (GO:0034641)	703	86	38.88	+	2.21	3.79E-10
organonitrogen compound metabolic process (GO:1901564)	781	94	43.2	+	2.18	3.34E-11
nitrogen compound metabolic process (GO:0006807)	1124	129	62.17	+	2.07	1.17E-16
cellular macromolecule metabolic process (GO:0044260)	567	64	31.36	+	2.04	3.36E-05
cellular metabolic process (GO:0044237)	1323	144	73.18	+	1.97	7.22E-18
protein metabolic process (GO:0019538)	383	41	21.18	+	1.94	4.45E-02
macromolecule metabolic process (GO:0043170)	821	87	45.41	+	1.92	4.67E-07
primary metabolic process (GO:0044238)	1277	133	70.63	+	1.88	7.50E-14
organic substance metabolic process (GO:0071704)	1435	148	79.37	+	1.86	1.76E-16
metabolic process (GO:0008152)	1590	158	87.95	+	1.8	5.66E-17
cellular process (GO:0009987)	1918	185	106.09	+	1.74	5.42E-23
biological_process (GO:0008150)	2336	198	129.21	+	1.53	3.01E-19
Unclassified (UNCLASSIFIED)	1768	29	97.79	-	0.3	0.00E+00

Table S3.6: List of eight genes with evidence of positive selection based on the global test

Gene	**(M0)[1]**	**Likelihood ratio test P-value[2]**
cya	1.24241	0.00669718
group_454	2.06375	0.00918337
group_1057	2.2049	0.0029329
group_3049	2.49761	0.00248336
group_3542	1.94168	0.00235286
group_3757	6.53607	0.00434296
group_5674	1.58194	0.0079647
group_7848	9.60069	0.00868936

[1] M0 estimates a single w across the entire phylogeny of sequences

[2] The p-value of tests after FDR correction. LRT between the global model which estimates omega for each gene and the model which sets omega to 1 for each gene.

References

Furuya EY, Lowy FD. 2006. Antimicrobial-resistant bacteria in the community setting. 1. Nat Rev Microbiol 4:36–45.

Yacoubi BE, Brunings AM, Yuan Q, Shankar S, Gabriel DW. 2007. In Planta Horizontal Transfer of a Major Pathogenicity Effector Gene. Appl Environ Microbiol 73:1612–1621.

Juhas M. 2015. Horizontal gene transfer in human pathogens. Crit Rev Microbiol 41:101–108.

Chen NWG, Serres-Giardi L, Ruh M, Briand M, Bonneau S, Darrasse A, Barbe V, Gagnevin L, Koebnik R, Jacques M-A. 2018. Horizontal gene transfer plays a major role in the pathological convergence of Xanthomonas lineages on common bean. BMC Genomics 19:606.

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci 99:17020–17024.

Badet T, Croll D. 2020. The rise and fall of genes: origins and functions of plant pathogen pangenomes. Curr Opin Plant Biol 56:65–73.

Kim Y, Gu C, Kim HU, Lee SY. 2020. Current status of pan-genome analysis for pathogenic bacteria. Curr Opin Biotechnol 63:54–62.

Sicard A, Zeilinger AR, Vanhove M, Schartel TE, Beal DJ, Daugherty MP, Almeida RPP. 2018. Xylella fastidiosa: Insights into an Emerging Plant Pathogen. Annu Rev Phytopathol 56:181–202.

Burbank LP, Roper MC 2021. 2021. Microbe Profile: Xylella fastidiosa – a devastating agricultural pathogen with an endophytic lifestyle. Microbiology 167:001091.

Schuenzel EL, Scally M, Stouthamer R, Nunney L. 2005. A Multigene Phylogenetic Study of Clonal Diversity and Divergence in North American Strains of the Plant Pathogen Xylella fastidiosa. Appl Environ Microbiol 71:3832–3839.

Loconsole G, Saponari M, Boscia D, D'Attoma G, Morelli M, Martelli GP, Almeida RPP. 2016. Intercepted isolates of Xylella fastidiosa in Europe reveal novel genetic diversity. Eur J Plant Pathol 146:85–94.

Chatterjee S, Almeida RPP, Lindow S. 2008. Living in two Worlds: The Plant and Insect Lifestyles of Xylella fastidiosa. Annu Rev Phytopathol 46:243–271.

Rapicavoli J, Ingel B, Blanco-Ulate B, Cantu D, Roper C. 2018. Xylella fastidiosa: an examination of a re-emerging plant pathogen. Mol Plant Pathol 19:786–800.

Tumber K, Alston J, Fuller K. 2014. Pierce's disease costs California $104 million per year. Calif Agric 68:20–29.

2015. Assessing the returns to R&D on perennial crops: the costs and benefits of Pierce's disease research in the California winegrape industry. Aust J Agric Resour Econ https://doi.org/10.22004/ag.econ.280230.

Koo H, Allan RN, Howlin RP, Stoodley P, Hall-Stoodley L. 2017. Targeting microbial biofilms: current and prospective therapeutic strategies. 12. Nat Rev Microbiol 15:740–755.

Castro C, DiSalvo B, Roper MC. 2021. Xylella fastidiosa: A reemerging plant pathogen that threatens crops globally. PLOS Pathog 17:e1009813.

Roper C, Lindow SE. 2016. CHAPTER 16: Xylella fastidiosa: Insights into the Lifestyle of a Xylem-Limited Bacterium, p. 307–320. *In* Caroline, R, Steven, EL (eds.), Virulence Mechanisms of Plant-Pathogenic Bacteria. The American Phytopathological Society.

Marcelletti S, Scortichini M. 2016. Genome-wide comparison and taxonomic relatedness of multiple Xylella fastidiosa strains reveal the occurrence of three subspecies and a new Xylella species. Arch Microbiol 198:803–812.

Nunney L, Vickerman DB, Bromley RE, Russell SA, Hartman JR, Morano LD, Stouthamer R. 2013. Recent Evolutionary Radiation and Host Plant Specialization in the Xylella fastidiosa Subspecies Native to the United States. Appl Environ Microbiol 79:2189–2200.

Almeida RPP, Purcell AH. 2003. Biological Traits of Xylella fastidiosa Strains from Grapes and Almonds. Appl Environ Microbiol 69:7447–7452.

Hernandez-Martinez R, Costa HS, Dumenyo CK, Cooksey DA. 2006. Differentiation of Strains of Xylella fastidiosa Infecting Grape, Almonds, and Oleander Using a Multiprimer PCR Assay. Plant Dis 90:1382–1388.

Almeida RPP, Nascimento FE, Chau J, Prado SS, Tsai C-W, Lopes SA, Lopes JRS. 2008. Genetic Structure and Biology of Xylella fastidiosa Strains Causing Disease in Citrus and Coffee in Brazil. Appl Environ Microbiol 74:3690–3701.

Castillo AI, Bojanini I, Chen H, Kandel PP, De La Fuente L, Almeida RPP. 2021. Allopatric Plant Pathogen Population Divergence following Disease Emergence. Appl Environ Microbiol 87:e02095-20.

Vanhove M, Sicard A, Ezennia J, Leviten N, Almeida RPP. 2020. Population structure and adaptation of a bacterial pathogen in California grapevines. Environ Microbiol 22:2625–2638.

Uceda-Campos G, Feitosa-Junior OR, Santiago CRN, Pierry PM, Zaini PA, de Santana WO, Martins-Junior J, Barbosa D, Digiampietri LA, Setubal JC, da Silva AM. 2022. Comparative Genomics of Xylella fastidiosa Explores Candidate Host-Specificity Determinants and Expands the Known Repertoire of Mobile Genetic Elements and Immunity Systems. 5. Microorganisms 10:914.

Kahn AK, Almeida RPP. 2022. Phylogenetics of Historical Host Switches in a Bacterial Plant Pathogen. Appl Environ Microbiol 88:e02356-21.

Daugherty MD, Malik HS. 2012. Rules of Engagement: Molecular Insights from Host-Virus Arms Races. Annu Rev Genet 46:677–700.

Aleru O, Barber MF. 2020. Battlefronts of evolutionary conflict between bacteria and animal hosts. PLOS Pathog 16:e1008797.

Mitchell PS, Patzina C, Emerman M, Haller O, Malik HS, Kochs G. 2012. Evolution-Guided Identification of Antiviral Specificity Determinants in the Broadly Acting Interferon-Induced Innate Immunity Factor MxA. Cell Host Microbe 12:598–604.

Ng M, Ndungo E, Kaczmarek ME, Herbert AS, Binger T, Kuehne AI, Jangra RK, Hawkins JA, Gifford RJ, Biswas R, Demogines A, James RM, Yu M, Brummelkamp TR, Drosten C, Wang L-F, Kuhn JH, Müller MA, Dye JM, Sawyer SL, Chandran K. 2015. Filovirus receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats. eLife 4:e11785.

Daugherty MD, Schaller AM, Geballe AP, Malik HS. 2016. Evolution-guided functional analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. eLife. eLife Sciences Publications Limited. https://elifesciences.org/articles/14228. Retrieved 2 April 2022.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res gr.215087.116.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput Biol 13:e1005595.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinforma Oxf Engl 29:1072–1075.

Chase AB, Gomez-Lunar Z, Lopez AE, Li J, Allison SD, Martiny AC, Martiny JBH. 2018. Emergence of soil bacterial ecotypes along a climate gradient. Environ Microbiol 20:4112–4126.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11:e0163962.

Castillo AI, Chacón-Díaz C, Rodríguez-Murillo N, Coletta-Filho HD, Almeida RPP. 2020. Impacts of local population history and ecology on the evolution of a globally dispersed pathogen. BMC Genomics 21:369.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol Biol Evol 17:540–552.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Boc A, Diallo AB, Makarenkov V. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res 40:W573–W579.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

Rodriguez CI, Martiny JBH. 2020. Evolutionary relationships among bifidobacteria and their hosts and environments. BMC Genomics 21:26.

Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol 34:2115–2122.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32:268–274.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. 6. Nat Methods 14:587–589.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2020. vegan: Community Ecology Package.

Kung SH, Almeida RPP. 2011. Natural Competence and Recombination in the Plant Pathogen Xylella fastidiosa. Appl Environ Microbiol 77:5278–5284.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15.

Shikov AE, Malovichko YV, Nizhnikov AA, Antonets KS. 2022. Current Methods for Recombination Detection in Bacteria. 11. Int J Mol Sci 23:6257.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223.

Mateo-Estrada V, Graña-Miraglia L, López-Leal G, Castillo-Ramírez S. 2019. Phylogenomics Reveals Clear Cases of Misclassification and Genus-Wide Phylogenetic Markers for Acinetobacter. Genome Biol Evol 11:2531–2541.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26:1372–1373.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK, p. 155–170. *In* Russell, DJ (ed.), Multiple Sequence Alignment Methods. Humana Press, Totowa, NJ.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine. Bioinformatics 26:2914–2915.

Avram O, Rapoport D, Portugez S, Pupko T. 2019. M1CR0B1AL1Z3R—a user-friendly web server for the analysis of large-scale microbial genomics data. Nucleic Acids Res 47:W88–W92.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13:555–556.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591.

Denancé N, Briand M, Gaborieau R, Gaillard S, Jacques M-A. 2019. Identification of genetic relationships and subspecies signatures in Xylella fastidiosa. BMC Genomics 20:239.

Bolotin E, Hershberg R. 2015. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol 7:2173–2187.

Iranzo J, Wolf YI, Koonin EV, Sela I. 2019. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. Nat Commun 10:5376.

Firrao G, Scortichini M, Pagliari L. 2021. Orthology-Based Estimate of the Contribution of Horizontal Gene Transfer from Distantly Related Bacteria to the Intraspecific Diversity and Differentiation of Xylella fastidiosa. 1. Pathogens 10:46.

Graña-Miraglia L, Lozano LF, Velázquez C, Volkow-Fernández P, Pérez-Oseguera Á, Cevallos MA, Castillo-Ramírez S. 2017. Rapid Gene Turnover as a Significant Source of Genetic Variation in a Recently Seeded Population of a Healthcare-Associated Pathogen. Front Microbiol 8.

Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host–pathogen interactions from mammalian sequence data. 4. Nat Rev Genet 16:224–236.

Giampetruzzi A, Saponari M, Loconsole G, Boscia D, Savino VN, Almeida RPP, Zicca S, Landa BB, Chacón-Diaz C, Saldarelli P. 2017. Genome-Wide Analysis Provides Evidence on the Genetic Relatedness of the Emergent Xylella fastidiosa Genotype in Italy to Isolates from Central America. Phytopathology® 107:816–827.

Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477.

Yuan X, Morano L, Bromley R, Spring-Pearson S, Stouthamer R, Nunney L. 2010. Multilocus sequence typing of Xylella fastidiosa causing Pierce's disease and oleander leaf scorch in the United States. Phytopathology 100:601–611.

Hurles M. 2004. Gene Duplication: The Genomic Trade in Spare Parts. PLOS Biol 2:e206.

Arnold BJ, Huang I-T, Hanage WP. 2022. Horizontal gene transfer and adaptive evolution in bacteria. 4. Nat Rev Microbiol 20:206–218.

Castillo AI, Almeida RPP. 2021. Evidence of gene nucleotide composition favoring replication and growth in a fastidious plant pathogen. G3 GenesGenomesGenetics 11:jkab076.

Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB, Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB. 2010. Plants versus pathogens: an evolutionary arms race. Funct Plant Biol 37:499–512.

Schulte RD, Makus C, Hasert B, Michiels NK, Schulenburg H. 2010. Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. Proc Natl Acad Sci 107:7359–7364.

Danchin A, Guiso N, Roy A, Ullmann A. 1984. Identification of the Escherichia coli cya gene product as authentic adenylate cyclase. J Mol Biol 175:403–408.

Smith RS, Wolfgang MC, Lory S. 2004. An Adenylate Cyclase-Controlled Signaling Network Regulates Pseudomonas aeruginosa Virulence in a Mouse Model of Acute Pneumonia. Infect Immun 72:1677–1684.

Kim YR, Kim SY, Kim CM, Lee SE, Rhee JH. 2005. Essential role of an adenylate cyclase in regulating Vibrio vulnificus virulence. FEMS Microbiol Lett 243:497–503.

Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, Thomson NR. 2021. Different evolutionary trends form the twilight zone of the bacterial pan-genome. Microb Genomics 7:000670.

Bhoopalan SV, Piekarowicz A, Lenz JD, Dillard JP, Stein DC. 2016. nagZ Triggers Gonococcal Biofilm Disassembly. 1. Sci Rep 6:22372.

Palaniyandi S, Mitra A, Herren CD, Lockatell CV, Johnson DE, Zhu X, Mukhopadhyay S. 2012. BarA-UvrY Two-Component System Regulates Virulence of Uropathogenic E. coli CFT073. PLOS ONE 7:e31348.

Sahu SN, Acharya S, Tuminaro H, Patel I, Dudley K, LeClerc JE, Cebula TA, Mukhopadhyay S. 2003. The bacterial adaptive response gene, barA, encodes a novel conserved histidine kinase regulatory switch for adaptation and modulation of metabolism in Escherichia coli. Mol Cell Biochem 253:167–177.

Roussin M, Rabarioelina S, Cluzeau L, Cayron J, Lesterlin C, Salcedo SP, Bigot S. 2019. Identification of a Contact-Dependent Growth Inhibition (CDI) System That Reduces Biofilm Formation and Host Cell Adhesion of Acinetobacter baumannii DSM30011 Strain. Front Microbiol 10:2450.

Ma Rodriguez A, Olano C, Vilches C, Méndez C, Salas JA. 1993. Streptomyces antibioticus contains at least three oleandomycin-resistance determinants, one of which shows similarity with proteins of the ABC-transporter superfamily. Mol Microbiol 8:571–582.

Sun QH, Hu J, Huang GX, Ge C, Fang RX, He CZ. 2005. Type-II secretion pathway structural gene xpsE, xylanase- and cellulase secretion and virulence in Xanthomonas oryzae pv. oryzae. Plant Pathol 54:15–21.

Abbas A, Adams C, Scully N, Glennon J, O'Gara F. 2007. A role for TonB1 in biofilm formation and quorum sensing in Pseudomonas aeruginosa. FEMS Microbiol Lett 274:269–278.

Das A, Rangaraj N, Sonti RV. 2009. Multiple Adhesin-Like Functions of *Xanthomonas oryzae* pv. *oryzae* Are Involved in Promoting Leaf Attachment, Entry, and Virulence on Rice. Mol Plant-Microbe Interactions® 22:73–85.

Zeiner SA, Dwyer BE, Clegg S. 2012. FimA, FimF, and FimH Are Necessary for Assembly of Type 1 Fimbriae on Salmonella enterica Serovar Typhimurium. Infect Immun 80:3289–3296.

Gossert AD, Bettendorff P, Puorger C, Vetsch M, Herrmann T, Glockshuber R, Wüthrich K. 2008. NMR Structure of the Escherichia coli Type 1 Pilus Subunit FimF and Its Interactions with Other Pilus Subunits. J Mol Biol 375:752–763.

CONCLUSIONS

Adaptive evolution drives species' survival and innovation, and it underlies the vast diversity of life. Our understanding of adaptation has come largely from observations of phenotype. However, recent advances in genome sequencing have facilitated potential links between phenotype and genotype to advance our knowledge of adaptive evolution (Orr 2005). Despite these advances, there are deficits in our understanding of adaptive processes such as characterizing the mechanisms that adaptation utilizes at the genomic and phenotypic levels, the influence of historical contingency on evolutionary change, and the influence of intraspecific genomic variation on adaptive evolution.

In my dissertation, I utilized experimental and bioinformatic methods to examine and characterize adaptive evolution using bacteria as my model organism. Bacteria are ubiquitous and fundamental to ecosystem functioning and often to the health of the eukaryotic organisms with which they associate. Therefore, characterizing and understanding the evolution and adaptation of bacteria is crucial. Additionally, bacteria themselves are convenient agents to study evolution due to their small genomes and their relative ease to grow and maintain in the laboratory. Their growth characteristics allow for experimental replication and for evolution to be monitored over hundreds of generations, because of their short generation times. Advances in genome sequencing and bioinformatic methods allow for comparative genomics, monitoring of genomic evolution, and identifying the causative genetic changes underlying adaptive evolution.

In my first chapter, I studied evolutionary rescue to lethal temperature using *E. coli*. Evolutionary rescue is a phenomenon by which populations can survive lethal environmental conditions due to genetic changes driven by adaptive evolution. To study rescue, a short-term evolution experiment was performed in which hundreds of populations of *E. coli* were maintained for five days at the lethal temperature of 43.0°C. The goals of this experiment were to quantify the frequency of rescue, identify the adaptive mutations that drive rescue, and to discern a mechanism by which rescue occurs to lethal temperature stress. We identified that rescue occurred at an 8.8% frequency in our system, resulting in 26 rescue populations that could be further examined. Through whole genome sequencing of the populations that successfully experienced evolutionary rescue, I identified the causative adaptive mutations underlying rescue to high temperature and characterized the molecular effects of the mutations through mRNA sequencing and gene expression analysis.

Together, the work conducted on evolutionary rescue highlights the importance of evolution as potential agent to rescue populations on the same timescales as ecological change and also to highlight different pathways that adaptive evolution may utilize for rescue. I identified that a single mutation in either the *rpoBC* or *hslVU* operon of *E. coli* was sufficient for evolutionary rescue to lethal temperature and the adaptive mutations likely arose *de novo* during serial transfer (Batarseh et al. 2020). Strikingly, a single nonsynonymous or frameshift mutation in either operon caused significant changes in fitness at two different temperatures and caused hundreds of genes to have altered expression patterns. These findings illustrate that bacterial evolution should be an important factor towards population survival, maintenance, and innovation even in natural populations as a single mutation can be sufficient for adaptive evolution to occur. Additionally, I observed that mutations in different pathways can result in similar rescue outcomes. The *rpoBC* operon encodes the beta subunit of RNA polymerase, which is a global regulator involved in the transcription of all genes, and the *hslVU* operon encodes a heat shock protease with specific functions to degrade misfolded proteins under heat stress. These results suggest that a diversity of adaptive solutions may exist in response to lethal environmental conditions.

In my second chapter, I used experimental evolution to investigate the influence of contingency on future evolutionary change. The goals of this chapter were to examine whether future evolutionary outcomes are significantly influenced by an organism's evolutionary history. The effects of contingency can greatly alter our ability to predict or forecast evolutionary outcomes (Blount et al. 2018). If contingency does significantly influence both genotypic and phenotypic evolution, then evolution would be unpredictable. However, a major goal of evolution and biology is to predict evolutionary outcomes as this has great implications for human health and species survival, especially in the face of climate change. Therefore, by characterizing how and when contingency may influence evolution, we will better understand the factors driving evolutionary change so that we may be able to form predictions.

To study contingency, I utilized a sequential evolution experiment approach using *E. coli* as my model organism. The first phase consisted of an evolution experiment that was previously performed in the Gaut lab and described in Tenaillon et al. (2012). In this evolution experiment, 114 initially identical lines of *E. coli* were evolved at 42.2°C which is a stressful but non-lethal temperature for the ancestral *E. coli*. After 2,000 generations of evolution, whole genome sequencing was performed to identify the adaptive mutations that arose in response to thermal stress. The results illustrated that adaptive evolution occurred through mutation in one of two genes which represented two distinct adaptive pathways each associated with their own sets of mutations in other genes. The first adaptive pathway was characterized by mutations in *rpoB*, encoding the beta subunit of RNA polymerase, and the second was characterized by mutations in *rho*, a transcriptional terminator that works on a subset of genes in the *E. coli* genome (Tenaillon et al. 2012). A second phase of evolution was founded using a subset of the evolved lines from the first phase of evolution, so we could compare and contrast the evolutionary outcomes of the two adaptive pathways in a second environment. The second phase of evolution occurred at 19.0°C, which is towards the lower thermal limit of the ancestral *E. coli*'s thermal niche. This study was unique to study contingency, as we had two separate adaptive pathways that had resulted from evolution to the same selective pressure during Phase 1 that we could then contrast in Phase 2.

Following the second phase of evolution, the evolved populations were assessed for their phenotypic and genotypic changes. I measured the changes in relative fitness of the evolved populations against their ancestral variants (Phase 1 Founder and Phase 2 Founders) and found statistically significant evidence to suggest that the initial genotype at the start of Phase 2 influenced the relative fitness at two temperatures, 19.0°C and 42.2°C, which suggested that contingency due to genotypic differences may influence the trajectory of relative fitness. At 19.0°C, I found that the evolved populations descended from a particular Phase 2 Founder (*rpoB* I966S genotype) experienced the greatest changes in fitness after Phase 2 evolution. Interestingly, the Phase 2 Founder *rpoB* I966S genotype also had the lowest initial relative fitness at the start of Phase 2. This suggests that initial fitness may be indicative of evolutionary change, therefore following Fisher's geometric model, and may be used as a tool to predict fitness outcomes in new environments and can be used in conjunction with genotypic data (Fisher 1930).

I also investigated the genetic changes that occurred during the second phase of evolution at 19.0°C for evidence of contingency. I identified over 1,000 mutations in the Phase 2 evolved populations and identified six regions of the genome that were enriched for mutations in lines descended from either the *rho* or *rpoB* adaptive pathways. Two regions were enriched for mutations in lines descended from *rho* backgrounds, and four regions were enriched for mutations in *rpoB* descended lines. This suggests that contingency influenced the genetic changes that arose during Phase 2 and caused further divergence between lines descended from the two different adaptive pathways. The regions of the genome enriched for mutations in *rho* backgrounds did not have annotation evidence to explore, but the regions enriched for mutations in *rpoB* lines did have annotation and functional evidence which we investigated further. Mutations in genes with large effect, like the genes *rapA* (hepA), *rho*, and *rpoC*, were enriched in *rpoB* lines, suggesting that large effect mutations were necessary for adaptation in *rpoB* lines but not in *rho* lines. Together, the genotypic data suggests that genotypic evolution is significantly influenced by evolutionary history and contingency must be considered when performing evolutionary forecasts.

The results from my second chapter suggest that both phenotypic change (measured by relative fitness) and genotypic change are influenced by contingency due to differences in evolutionary history between bacterial lines. These results are both in contrast and in agreement with other experimental studies that assessed the influence of contingency on bacterial and yeast evolution. Studies using yeast have demonstrated that both phenotypic and genotypic changes are not influenced by differences in evolutionary history. In particular, Kyrazhimskiy et al (2014) found that diverse yeast lines would still accumulate mutations in the same or similar genes despite initial differences in evolutionary history. Instead we found that the lines descended from the two adaptive pathways had enrichment for mutations in different genes. In another study that investigated antibiotic resistance evolution in bacteria, the authors found that both phenotypic and genotypic changes were contingent on evolutionary history, which was similar to our findings (Card et al. 2021). Altogether, this chapter demonstrates that future fitness outcomes and genetic evolution are significantly influenced by the evolutionary history of an organism.

In my third chapter, I utilized a different approach to study adaptive evolution and focused on comparative genomics using bioinformatic methods to study the evolution of the plant pathogen *X. fastidiosa*. By using a comparative genomics approach, I could investigate the phylogenetic relationships and variation in genomic content of this plant pathogen to understand *X. fastidiosa*'s evolutionary history and possibly identify genetic determinants underlying pathogenicity. To investigate *X. fastidiosa* genomes, I gathered publicly available data and generated novel sequences for analysis. I performed a pangenome analysis to characterize the genes in the *Xylella* genus as either a core or accessory gene. Using association tests and phylogenetic methods, I found that both the core gene sequences and the composition of accessory genes were associated with the identity of the plant host the bacteria was isolated from. This suggested that *X. fastidiosa* has signatures of host specificity encoded in both its core and accessory genes, which has previously been debated (Uceda-Campos et al. 2022; Kahn & Almeida 2022).

Additionally, I measured the ratio of nonsynonymous to synonymous mutations (known as dN/dS or ω) in all of the core genes and a subset of accessory genes to identify genes experiencing positive selection. Evolutionary dynamics between host and pathogen can facilitate rapid genomic evolution and would leave signatures in the genome that could be identified by testing for positive selection. Using a global test that quantifies the average value of ω for each gene, I found that the core genes are largely experiencing purifying selection while the accessory genes are more variable in their values of ω. Average ω was significantly higher for accessory genes as well compared to core genes. In addition to the global test, I also performed a test that measured positive selection in codons of genes and I found that 5.3% of core genes and 5.4% of accessory genes have evidence of a history of positive selection in particular codons. The information from the two tests revealed a set of accessory genes that have significant evidence of positive selection and represent genes that are candidates towards pathogenicity and could be further investigated in the laboratory for disease management.

Altogether, my dissertation has revealed various characteristics of adaptive evolution in microbial systems using both experimental and bioinformatic methods and has opened up new lines of research that can be pursued. Using experimental methods, I identified the mutations underlying rescue to lethal temperature and characterized the effects of these mutations by measuring fitness and gene expression. My results suggest that rescue is an important phenomenon serving to rescue populations on the same timescales as ecological change and that evolution to lethal stressors may have qualitative differences compared to evolution to non-lethal stress. Both lines of thought prompt further research efforts. Do the novel changes in gene expression found in rescue mutants persist over a longer period of evolution to lethal stress? Do multiple pathways exist for rescue in other environmental conditions (pH stress, salinity)? Through experimental evolution, I found that evolutionary history may significantly influence both phenotypic and genotypic change which may affect our ability to form evolutionary predictions. With growing evidence to suggest that contingency matters, it would be beneficial to study the factors that influence contingency. Does the intensity of selection affect contingency? Do we still see contingent effects on evolution if the second phase of evolution occurs in a completely different environment (for example, evolving to high temperature stress before evolving to pH stress)? Can we quantify a minimum genetic distance that may serve as a predictive tool to discern if contingency may affect the evolution of similar bacterial strains? Finally, using comparative genomics I found evidence to suggest that *X. fastidiosa* does exhibit host specificity and it is encoded in its genome. The genes identified with sufficient evidence of positive selection could be manipulated in *X. fastidiosa* in the laboratory to see if knocking out the function affects pathogenicity. Additionally, sampling *X. fastidiosa* from a greater geographic range and from diverse host plants (symptomatic and asymptomatic plants) would be beneficial towards better understanding the evolutionary history of *X. fastidiosa*. Are there genetic differences between *X. fastidiosa* found in symptomatic and asymptomatic plants and, if so, are those differences driving pathogenicity? Understanding and characterizing adaptive evolution is imperative and has implications for species survival, evolutionary predictions, and pathogen management.

References

Batarseh TN, Hug SM, Batarseh SN, Gaut BS. 2020. Genetic Mutations That Drive Evolutionary Rescue to Lethal Temperature in Escherichia coli. Genome Biology and Evolution. 12:2029–2044. doi: 10.1093/gbe/evaa174.

Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life's tape. Science. 362. doi: 10.1126/science.aam5979.

Card KJ, Thomas MD, Graves JL, Barrick JE, Lenski RE. 2021. Genomic evolution of antibiotic resistance is contingent on genetic background following a long-term experiment with Escherichia coli. Proceedings of the National Academy of Sciences. 118:e2016886118. doi: 10.1073/pnas.2016886118.

Fisher RA. 1930. *The genetical theory of natural selection*. Oxford University Press: Oxford, UK.

Kahn AK, Almeida RPP. 2022. Phylogenetics of Historical Host Switches in a Bacterial Plant Pathogen. Applied and Environmental Microbiology. 88:e02356-21. doi: 10.1128/aem.02356-21.

Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 344:1519–1522. doi: 10.1126/science.1250939.

Orr HA. 2005. The genetic theory of adaptation: a brief history. Nat Rev Genet. 6:119–127. doi: 10.1038/nrg1523.

Tenaillon O et al. 2012. The Molecular Diversity of Adaptive Convergence. Science. 335:457–461. doi: 10.1126/science.1212986.

Uceda-Campos G et al. 2022. Comparative Genomics of Xylella fastidiosa Explores Candidate Host-Specificity Determinants and Expands the Known Repertoire of Mobile Genetic Elements and Immunity Systems. Microorganisms. 10:914. doi: 10.3390/microorganisms10050914.