New computational approaches to study the evolution of asexual haploids

Jonathan Dench

A thesis submitted in partial fulfillment of the requirements for the

Doctorate in Philosophy degree in Biology

Department of Biology

Faculty of Science

University of Ottawa

Abstract

Numerous factors can influence the evolutionary fate of mutations. Despite this, we tend to study strong evolutionary drivers, or evolution under simple contexts, in part because they are the conditions we have a means to study. My thesis evaluates novel computational approaches to advance detection, and study, of factors that influence a mutation's evolutionary outcome. First, I present the novel computational tool AEGIS that I use to detect phylogenetic signals of correlated evolution followed by an experimental approach to evaluate the role of epistasis as a potential cause of correlated evolution among sites associated with antibiotic resistance in *Pseudomonas aeruginosa*. Second, I developed rSHAPE, a novel *in silico* approach for experimental evolution with asexual haploids, to complement empirical work by providing a common framework in which to test various evolutionary scenarios. After demonstrating that rSHAPE replicates the expected evolutionary dynamics of *de novo* mutations, I provide evidence that the common laboratory practice of serial passaging may increase stochasticity of evolutionary outcome. Through my work, I have demonstrated that a marriage of computational and experimental approaches will offer new opportunities to understand how the interaction of evolutionary factors influence the fate of mutations.

Résumé

De nombreux facteurs peuvent influencer le destin évolutif des mutations. Malgré cela, nous avons tendance à étudier les facteurs evolutifs les plus puissants, ou l'évolution dans des contextes simples, en partie parce que ce sont les conditions que nous avons le moyen d'étudier. Ma thèse évalue de nouvelles approches informatiques pour améliorer la détection et l'étude des facteurs qui influencent le destin évolutif des mutations. Tout d'abord, je présente le nouvel outil informatique AEGIS que j'utilise pour détecter les signaux phylogénétiques d'évolution corrélée, suivi d'une approche expérimentale visant à évaluer le rôle de l'épistasie en tant que cause potentielle de l'évolution corrélée entre sites associés à la résistance aux antibiotiques dans les souches de *Pseudomonas aeruginosa*. Deuxièmement, j'ai développé rSHAPE, une nouvelle approche informatique pour simuler l'évolution expérimentale avec des haploïdes asexués, afin de complémenter les travaux empiriques en fournissant un cadre commun permettant de tester différents scénarios d'évolution. Après avoir démontré que rSHAPE reproduisait des dynamiques évolutives attendues des mutations, je prouve que la pratique courante en laboratoire du passage en série peut augmenter la stochasticité des résultats évolutifs. Au travers de mes travaux, j'ai démontré qu'un mélange d'approches informatiques et expérimentales offrira de nouvelles opportunités pour comprendre comment l'interaction de facteurs évolutifs influence le sort des mutations.

Acknowledgments

I find it difficult to explain both how difficult and rewarding this experience has been. Before I thank the countless individuals who have made a meaningful impact in my life over the past years, I want to "set the scene" for the readers interested in this section.

Though it feels tacky to say my time completing this thesis has been hell, but worth it, I do not have a more succinct way to put it. Since I started my thesis work I have (in this order): overcome a decade long mental health related challenge, had my wife ask for a divorce, lost my house as well as the majority of all my worldly possessions, relapsed with mental health related difficulties, found meaning in my life for the first time in more than a decade, found new love, lost my father after watching him wither away for roughly one year, got married, started a new career, bought a house, and as of the time of this writing have very recently welcomed my daughter into this world. And while my thesis is not responsible for a single one of the aforementioned events, it has been a constant in my life that offered me the framework and focus I needed to reshape quite literally who and what I am.

So, my biggest thanks go out to my supervisors who took a chance on someone who had been "out of the game" of academia for five years. While my current life path makes it seem like I only stepped in long enough to accomplish some skills

training, I wish to acknowledge what they have both offered me. For whatever it is worth, I thank you Rees and Stéphane. Also, my thanks go out to my thesis advisory committee members Alex Wong and Nicolas Rodrigue whom have offered me valuable alternate viewpoints through their supportive criticism.

Now, for all of my peers in the Kassen and Aris-Brosou labs, thank you Éléonore Lebeuf-Taylor, Leah-Clarke, Alanna Leane, Alana Shick, Anita Melnyk, Nicolas Rode, Jeremy Dettman, Aaron Hinz, Felipe Dargent, Sonal Shewaramani, Gabriel Perron, Jean-Claude Nshogozabahizi, Neke Ibeh, Matti Ruuskanen, Graham Colby, Samuel Long, Mohammed Hussen, Louis Parent, and Jayson Azzi. Each of you have left a positive mark in my life, and some of you have even been good friends, I wish you all the greatest success in whatever it is that makes you happy.

I also want to thank the myriad of biology graduate students that I have worked along side between the 2014-15 and 2017-18 scholastic years. For just being exceptional human beings, and people I wish I could spend more time with, thanks go out to Kim Mitchel, Kevin Kwok, Chris Angel, Micheal Country, Eric Chen, Ethan Hermer, Khang Hua, Patrick Chen, Peter Soroye, Véronique Boucher-Lalonde, Marie-Bé Leduc, Caitlyn Mary Catherine, Jordan Silke, Cassandra Robillard, Shannon Whelan, Emina Alic, and Malcolm Hughes.

For help understanding the work that came before mine own and comments related to rSHAPE, special thanks to Dr. Lindi Wahl. Further, for providing the resources to complete my studies, my thanks go out to Compute Canada and the Center for Advanced Computing.

For those exceptional people I met while traveling at conferences, know that you have always got a place to sleep if you come to visit my corner of the world. Many thanks go out to Claudia Bank, Inês Fragata, Hermina Ghenu, Arjan de Visser, Thomas Bataillon, and Ariana Longley.

Lastly, I must thank my family whom have accepted me despite my being effectively absent for nearly six years. Without all of you, I would not have gotten to start this nor found a way through it all. Thanks to my mother Hélène Chartier, my step-mother Margaret George-Dench, my father Philip Dench, and my sisters Roxanne Dench and Jaydee Swarbrick. Finaly the most important member of my family, my wife Elysabeth Théberge - the mother of our child. I met you at one of the blandest points in my existence and you have brought light and colour into my life, today marks a monumental step into our continuing adventures.

Table of Contents

List of Abbreviations

Abbreviation	Meaning
AA	Amino Acid
AEGIS	Analysis of Epistasis & Genomic Interacting Sites
aLRT	Approximate Likelihood Ratio Test
AU	Approximately Unbiased
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BLAST	Basic Local Alignment Search Tool
CAC	Center for Advanced Computing
CF	Cystic Fibrosis
COG	Cluster of Orthologous Groups
CRAN	Comprehensive R Archive Network
D	Dilution factor
DNA	Deoxyribonucleic Acid
ϵ	Epistasis
FDR	False Discovery Rate
ΔG	Change in Folding Free Energy
Γ	Gamma
GNU	GNU's Not Unix!
GTR	General Time Reversible
H or HA	Hemagglutinin
HEC	Haploid Evolutionary Constructor
HGT	Horizontal Gene Transfer
HoC	House of Cards
I_{TE}	Index of Translation Efficiency

Continued on next page

List of Abbreviations – *Continued from previous page*

Abbreviation	**Meaning**
K	Population Capacity
L	Length
λ	Shape parameter for exponential probability generating function
LB	Lysogeny Broth
MIC	Minimum Inhibitory Concentration
mRNA	Messenger Ribonucleic Acid
MUSCLE	Multiple Sequence Comparison by Log-Expectation
N or NA	Neuraminidase
NP	Nucleoprotein
PA	*Pseudomonas aeruginosa*
PAML	Phylogenetic Analysis by Maximum Likelihood
PCM	Phylogenetic Comparative Method
PCR	Polymerase Chain Reaction
PGLS	Phylogenetic Generalised Least Squares
r	Growth rate
RAM	Random Access Memory
RMF	Rough Mount Fuji
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcription Polymerase Chain Reaction
rSHAPE	R-Package for Simulating Haploid Asexual Population Evolution
s	Selection-Coefficient
SH test	Shimodaira–Hasegawa test
SQL	Structured Query Language
t	Time
T	Generation
τ	Tree-shape
tRNA	Transfer Ribonucleic Acid
WAG	Whelan and Goldman (a nucleic acid model)
WT	Wild Type
χ^2	Chi-squared

List of Figures

List of Tables

Chapter 1

Introduction

"No man is an island"

-John Donne (1624) *Devotions Upon Emergent Occasions*, Meditation 17

It is a population, not individuals, that evolves. Though individuals carry the genetic variation upon which selection acts, it is the sum of every individuals' interactions that drive evolution by ultimately affecting how each contributes to the gene pool. Evolutionary outcome is affected by a myriad of interacting factors including the environment [Mira *et al.* 2015; Ogbunugafor *et al.* 2016], different species [Lawrence *et al.* 2012], a species' traits [Laland *et al.* 2017], as well as mutations both within [Bank *et al.* 2016; Li *et al.* 2016; Weinreich *et al.* 2006] and between genes [Lehner 2011]. And so to paraphrase, "No evolutionary factor is an island" where at the extreme even neutral factors (*e.g.* synonymous mutations) may be selective in the right context [Bailey *et al.* 2014]. But which factors are important, and when do their interactions matter? Recent work shows that the interplay among mutations [Bank *et al.* 2016], as well as sympatric populations [Lieberman *et al.* 2014], affect evolutionary outcome such that specific studies provide limited information about the evolutionary picture as a whole. Thus, my

thesis aimed to study interactions between specific evolutionary factors while considering the evolutionary picture. In order to understand the picture as a whole, I worked with systems where many of the specifics are already known. The common pathogens influenza A [Hayden and de Jong 2011; Hurt 2014; van der Vries *et al.* 2013] and *Pseudomonas aeruginosa* [Kos *et al.* 2015; Lieberman *et al.* 2014; Marvig *et al.* 2015] are well studied in part because they are becoming increasingly resistant to our management practices [Organization *et al.* 2014]. As a result, there exists a wealth of published research, experimental methodologies, and genomic information for both organisms. Focusing on mutations in epidemic influenza A viruses, and antibiotic resistant *P. aeruginosa* strains, my research was primarily focused on the detection of correlated evolution among mutations. This thesis represents part of my work driven to better understand the number of evolutionary interactions that exist, how often they matter, and what predictive patterns could be understood.

1.1 An era of antiviral and antibiotic resistance

The World Health Organization lists both global influenza pandemic [Organization 2019a], and antimicrobial resistance [Organization 2015], among the top ten threats to global health in 2019 [Organization 2019b]. Adamantane (*i.e.* M2 protein) and neuraminidase inhibitor classes of antivirals are the primary means of managing the influenza virus [van der Vries *et al.* 2013]. It was long believed that any antiviral/antimicrobial resistances would fail to persist among wild populations due to fitness costs associated with resistance mutations [Fisher 1958; Orr 2005; Oz *et al.* 2014]. However, adamantane resistance is now widespread [van der Vries *et al.* 2013] and the growing persistence of neuraminidase resistant strains results from "permissive" mutations that overcome the costs of resistance [Hurt 2014]. When

genotypes carry no pleiotropic fitness costs outside of the selective context, possibly due to "permissive" mutations, we call these instances of cost free resistance. These "permissive" mutations have been found to accumulate in populations due to the sublethal doses associated with prophylactic use of the antivirals [McKimm-Breschkin 2013]. The prophylactic use of antibiotics to manage bacterial pathogens has also been linked to an increase in resistant strains [Cabello 2006; Singer *et al.* 2003] that may similarly persist between bouts of antibiotic selection due to compensatory mutations [Melnyk *et al.* 2017]. While specific studies warn that the interactions among resistance mutations [Hall and MacLean 2011], or between resistance and "permissive" mutations, can lead to epidemic and resistant pathogens, it remains to be seen how common these interactions are and what evolutionary understanding we can infer from their study.

1.2 Background on my chosen asexual haploid study systems

Owing to a lack of recombination during reproduction, the comparative study of evolution for asexual haploids is relatively simpler than that of sexual organisms [but see Gogarten and Townsend 2005; Ochman *et al.* 2000]. Further, it is easier to detect the independent effect of mutations because there are no masking effects caused by other alleles on additional copies of the genome [Mable and Otto 2001; Otto and Gerstein 2008][but see Zhang *et al.* 2013]. As a result, evolutionary models require fewer assumptions resulting in higher confidence of the correlations between phenotype and genotype data. It is also fortuitous that most asexual haploids are microscopic and have generation times on the order of hours, thus facilitating the

experimental replication and evolutionary time required for manipulative studies. While there are numerous approaches for studying asexual haploids, I will review some of the background pertinent for my study organisms influenza A and *Pseudomonas aeruginosa*.

1.2.1 Influenza A

The influenza A virus is a single stranded, segmented RNA virus associated with more than 200,000 hospital visits [Thompson *et al.* 2004] and 50,000 deaths [Thompson *et al.* 2003] each year in the United States of America. Infleunza A strains are generally identified by their hemagglutinin (H) and neuraminidase (N) proteins found on the viral envelope (*e.g.* H1N1, H3N2). The first isolation of an influenza virus was in 1933 [Smith *et al.* 1933] when nasal filtrates (*i.e.* bacterial free materials) were shown to transmit influenza among ferrets [Smith *et al.* 1933]. The use of animal models was not generally replaced until the 1950's when cultured cells became both the preferred system in which to grow viruses for study and the gold standard against which other methods were compared [Hsiung 1984; Leland and Ginocchio 2007]. Since the start of the millennium, viral research has increasingly relied on nucleic acid amplification techniques (*e.g.* reverse-transcription polymerase chain reaction – RT-PCR) [Leland and Ginocchio 2007] that have higher sensitivity and strain specificity than culturing techniques. Owing to the progression of molecular methods, whole-genome sequencing can now be used to track the evolution of infectious viral communities [Depledge *et al.* 2011] as well as perform high-throughput detection of unknown viruses [Cheval *et al.* 2011]. Thanks to the growing databases of viral sequence data, computational studies can perform comparative analyses that infer evolutionary events by fitting sequence data with

evolutionary models. Some examples include reconstruction of the phylogeny for novel epidemic strains [Liu *et al.* 2013], determining different modes of evolution for the H and N proteins [Sandie and Aris-Brosou 2014], and identification of global transmission networks [Aris-Brosou 2014; Nelson *et al.* 2007]. While computational studies have identified mutations associated with existing epidemic/antiviral resistant strains, it remains to be seen if computational approaches can predict novel sets of correlated mutations that interact during the evolution of influenza A viruses.

1.2.2 *Pseudomonas aeruginosa*

The bacterium *P. aeruginosa* is a rod shaped, Gram-negative, opportunistic pathogen responsible for 11–13.8% of all hospital acquired infections [Driscoll *et al.* 2007]. *Pseudomonas aeruginosa* is especially dangerous for immune compromised individuals [Sadikot *et al.* 2005] and is the predominant infectious threat to the health of cystic fibrosis patients [Burns *et al.* 1998]. There are several common research strains of *P. aeruginosa*, generally isolated from hospital patients, the first of which was PAO1 isolated in Australia in 1954 [Holloway 1955]. The bacterium grows well in shaken liquid lysogeny-broth (LB) [Bertani 1951] at 37°C but can survive a range of temperatures, utilise various carbon sources, and may grow anaerobically on certain of those sources when also given nitrates [LaBauve and Wargo 2012]. The reproducible cultivation of bacteria with growth media can be attributed to the work of Louis Pasteur in 1860 [Sandle 2010]. Bacteria can be isolated by serial passaging among different selective liquid media as well as streaking on structured surfaces (*e.g.* agar plates). Though the types and uses of growth media have advanced over the past 150 years, use of culture media in bacteriology remains a fundamental tool [Sandle 2010]. While culture media

remains a predominant experimental environment, molecular techniques have advanced to allow rapid screening of genetic variants [Panagea *et al.* 2003], detection of genomic diversity in populations [McElroy *et al.* 2014], as well as recombinant engineering to design specific genotypes for study [Engler *et al.* 2008; Hoang and Schweizer 1999; Zheng *et al.* 2004]. There also exist extensive resources, including sequence information, for *P. aeruginosa* at the Pseudomonas Genome Database `http://www.pseudomonas.com` [Winsor *et al.* 2011]. By sequencing experimentally evolved populations of *P. aeruginosa*, studies have found parallel evolution resulting from host-patient interactions [Wong *et al.* 2012], have better understood the dynamics of adaptation during early colonisation of hosts [Schick and Kassen 2018], and have shown how fluctuating antibiotic selection promotes the evolution of correlated sets of mutations [Melnyk *et al.* 2017]. Through computational analysis of the whole genomes of clinical isolates, studies have been able to identify different evolutionary trends among epidemic and nonepidemic strains of *P. aeruginosa* [Dettman *et al.* 2013], convergent evolution in response to antibiotic therapies [Marvig *et al.* 2015], and better understand resistance determinants [Kos *et al.* 2015]. By comparing which mutations interact during the evolution of antibiotic resistance, could we better predict sets of mutations that would lead to - or better yet prevent - cost free antibiotic resistance?

1.3 A primer on evolutionary interactions

Though evolutionary studies, such as in the context of antiviral or antibiotic resistance, may identify specific mutations that drive phenotypic changes (*e.g.* bacterial β-lactam resistance [Medeiros 1997], amantadine and oseltamivir resistance mutations in influenza A [Zaraket *et al.* 2010], aminoglycoside resistance

in *P. aeruginosa* [Lau *et al.* 2014]), these mutations do not evolve in isolation. The correlated evolution of mutations/substitutions occurs when their interactions affect evolutionary outcome.

While the evolutionary history of a genome shapes which substitutions are present, prior substitutions can affect the evolutionary outcome of mutations via interactions we term historically contingent [Travisano *et al.* 1995]. Studies have shown that historical contingency can affect the antibiotic resistance conferred by mutations [Card *et al.* 2019; Nyerges *et al.* 2018]. Historical contingency is one example of correlated evolution, a term used to describe when the evolutionary fate of mutations influence one another. When a set of mutations is correlated because an adaptive mutation drives the fixation of other neutral or deleterious mutations, we call this hitchhiking [Hedrick 1982; Maynard Smith and Haigh 1974]. Yet mutations need not interact despite being under selection and co-occurring in a single genome. Though a double mutant carrying two independent beneficial mutations would be more fit than a mutant carrying only one of the pair, these mutations are not said to interact if their deterministic evolutionary fates are simply the result of their independent contributions. When mutations are correlated, and the resultant phenotype is not a sum (or product [see Trindade *et al.* 2009]) of the independent parts, we call this epistasis. Bateson first coined the term epistasis when he found that offspring did not present expected Mendelian ratios and he reasoned that one evolutionary factor had modified another [Bateson 1909]. Consider a scenario of two mutations, one for eyelessness and the other for red eye colour; these two mutations will necessarily be epistatic because no matter the allele for eye colour an eyeless mutant will not express colour in its eyes. Though an extreme case of epistasis, this example highlights the extent to which epistatic interactions may influence evolutionary outcome. When epistasis affects the fitness

of an organism, the probability that an individual will pass on the correlated mutations will change proportional to the epistatic effect on fitness. When considering evolutionary fate among a set of mutations, epistasis can alter which permutations are viable [Weinreich *et al.* 2006].

Mutations may interact not only with respect to presence or absence, but also due to their timing. While Gillespie's seminal evolutionary model for strong selection and weak mutation (SSWM) [Gillespie 1983, 1984] assumes that beneficial mutations arise and sweep to fixation before another beneficial mutation arises, recent studies suggest that this assumption is not generally met [Bailey *et al.* 2016; Desai *et al.* 2007]. When multiple mutations existing in different individuals of the same asexual species compete for fixation, we call this clonal interference. The fixation rate of mutations slows as a result of clonal interference [Gerrish and Lenski 1998]. Under slower rates of fixation, we expect more mutations will arise and clonally compete prior to any one sweeping to fixation. The evolutionary fate of a mutation begins in a stochastic phase when it is relatively rare but becomes deterministic upon its selection coefficient once the mutation is "established" [Desai and Fisher 2007]. The longer time to fixation caused by clonal interference provides more opportunity for beneficial mutations to compete deterministically and is expected to result in the fixation of mutations with higher selective coefficients [Perfeito *et al.* 2007; Rozen *et al.* 2002]. Clonal competition highlights how the interaction among mutations - even in separate genomes - can affect evolutionary dynamics.

1.4 Studying evolutionary interactions

Visualising data is a powerful tool to support study and when studying the interactions among evolutionary traits we can visualise their history of shared descent using a phylogeny. Darwin inspired phylogenies with his drawing of a tree of life [Darwin 1859] though he did not use the term phylogeny and simply grouped organisms by shared sets of morphological traits regardless of actual evolutionary relationships. Phylogenetic trees present a form of hierarchical clustering that attempts to map evolutionary relationships and may often rely on molecular information (*e.g.* sequence data). With a phylogeny, biologists can compare measured traits and attempt to understand their evolutionary history by testing alternative hypotheses [Harvey and Pagel 1991; Pagel 1999]. Though biologists have a long history of using comparative methods to test evolutionary hypotheses, both with and without phylogenies [Harvey and Pagel 1991], the field of phylogenetic comparative methods (PCM) is relatively young having emerged in the 1980's [Cooper *et al.* 2016]. PCMs combine trait data and information about their evolutionary history to quantify significant trait changes while accounting for the effect of shared descent. Felsenstein [1985] was one of the earliest to remark that correlations among traits measured in different organisms are not independent and so traditional statistical methods were not appropriate. Felsenstein proposed the method of independent contrasts to account for the non-independence of phylogenetic traits. The method of independent contrasts was quickly improved by Cheverud *et al.* [1985] who proposed the use of phylogenetic generalised least squares (PGLS) [Pagel 1997]. With PCMs, it is possible to test for significant correlations in trait evolution, though correlation is not necessarily causation. The seminal work of Pagel [1994] provided a PCM to test for correlated evolution among

discrete characters, that he applied to a data set of social and morphological traits. Despite the fact that genetic information is a set of discrete characters, and that Pagel's algorithm has been implemented in software available since 2006 [Pagel and Meade 2006], I knew of no previous work that had ever applied Pagel's algorithm to genetic information. It has been noted that one barrier in the uptake of novel analytical algorithms is the rigidity of proprietary software [Freckleton 2009]. Regardless of the rate of uptake, PCM's continue to be improved by including additional trait data (*e.g.* fossil record) to help distinguish between alternative models of trait evolution [Slater *et al.* 2012].

Despite improvements to comparative methods, we can only infer evolutionary processes through comparative study of extant evolutionary diversity. It is only with direct and manipulative study that we can confidently assess the interplay of evolutionary factors [Losos 2011]. Comparative approaches may fail to explain an evolutionary story because multiple scenarios may select for the same phenotype. As examples, consider two populations of otherwise distinct microbes sharing a phenotype of antibiotic resistance; i) we may presume the trait repeatedly evolved due to similar selective pressures but resistance can evolve from direct selection as well as an off-target effect of sublethal selection by different antibiotics [Kohanski *et al.* 2010], and ii) the mechanisms underlying antibiotic resistance may derive from different specific adaptive mutations, or general resistance mechanisms [Poole 2005], that may evolve *de novo* or be in the gene pool prior to selection. So while the compared evolutionary outcome may be similar (*i.e.* resistance) confidently understanding the details of how traits evolve requires consideration of the interplay between intrinsic (*e.g.* standing variation, mutation rate, genomic context) and extrinsic (*e.g.* abiotic selection, biological interactions) factors. It is worth noting this may be precisely why Slater *et al.* [2012] found that including additional

information in PCMs improved the ability to distinguish between evolutionary models.

However, fitting models to data only provides a means to infer processes whereas experimental evolution, such as with microbes, is a powerful tool for the direct manipulative study of interacting evolutionary factors [McDonald 2019]. Yet, there is an infinite number of possible interactions to test and experimental evolution can be labour intensive. So, how do we prioritise which experiments are run? Traditionally, studies have focused on well known mutations of large effect but this biased approach limits our understanding of evolution [Bank *et al.* 2016; de Visser and Krug 2014]. Further, such focused approaches are not practical for all evolutionary hypotheses and so alternative methods must be adopted [Maynard Smith 1978]. More recent empirical methods allow for a less biased assessment of all point mutations within a few sequential codons in a genome [Hietpas *et al.* 2011; Nyerges *et al.* 2018] and have shown how the distribution of fitness effects, both nonsynonymous and synonymous, can interact with the selective environment [Bank *et al.* 2014; Fragata *et al.* 2018]. However, the effects of mutations can be influenced not only by the external environment but also the internal environment (*i.e.* presence of other mutations, *a.k.a.* genomic context). It is only with computational approaches that we can practicably study whole genomes, and more importantly which pairs of mutations interact, to identify candidates for more direct experimental assessment.

It remains an open question as to whether Pagel's PCM for correlated evolution of discrete states [Pagel 1994] could be used to scan whole genomes and identify pairs of interacting mutations under a given selective context. A computational approach involving Pagel's algorithm would allow us to study how much of genomic evolution is correlated and when paired with manipulative study we can measure

when correlated evolution is epistatic. By matching a comparative computational approach with additional analyses (*e.g.* direct experimentation), we may gain more detailed understanding of the underpinnings of correlated evolution and so better understand the interaction of genomic factors.

1.5 Thesis rationale

1.5.1 Chapter 2

In my first study, I helped develop the novel pipeline AEGIS (Analysis of Epistasis & Genomic Interacting Sites) that implements Pagel's algorithm of correlated evolution. Pagel's algorithm was published with an analysis of phenotypic traits and had never been tested with genomic sequence information. I assessed the statistical performance of Pagel's algorithm when estimating the signals of correlated evolution from genomic information. I then helped refine an approach used to predict epistasis among amino acid positions within different segments of the influenza A virus genome.

This first study assumed signals of correlated evolution would represent epistasis because there is relatively strong purifying selection in the influenza A virus genome [Hughes and Hughes 2007] that should result in primarily adaptive substitutions [Illingworth and Mustonen 2012]. But not all correlated evolution results from adaptive interactions. When the presence of an adaptive mutation drives the inheritance of other non-adaptive mutations, we call this genomic hitchhiking [Hedrick 1982; Maynard Smith and Haigh 1974]. Genomic hitchhiking can persist when recombination fails to separate the inheritance pattern of mutations and hitchhiking can lead to signals of correlated evolution. In asexual organisms, such as

many bacteria, we expect higher rates of hitchhiking due to relatively lower recombination rates than in sexual organisms. Muller's ratchet [Felsenstein 1974; Muller 1964] represents a worst case scenario of genomic hitchhiking whereby deleterious mutations accrue in the genetic background of otherwise adaptive asexuals. A comparative analysis of correlated evolution could find a positive association between adaptive drivers of evolution and deleterious hitchhikers but similarly detect pairs of "permissive" and beneficial mutations. Are there patterns among pairs of correlated mutations that offer insight to how infleunza A strains evolved?

1.5.2 Chapter 3

For my second study, I used AEGIS to identify signals of correlated evolution among nucleotide positions in the genome of *Pseudomonas aeruginosa*. Using a mixture of clinical isolates that were sensitive, or resistant, to fluoroquinolone antibiotics, I aimed to identify patterns of correlated evolution in the context of antibiotic resistance. This effort also represented the first attempt to identify pairwise correlated evolution among sites distributed throughout a whole genome and so the first opportunity to identify prevalence, and patterns of, correlated evolution. Using additional computational and direct experimental methods, I assessed the mechanisms underlying signals of correlated evolution.

To directly measure fitness and phenotypic effects of mutations identified in my second study, I generated mutant variants of *Pseudomonas aeruginosa* and performed competition experiments. The use of microbes for the manipulative study of evolution, and specifically the practice of competitive fitness assays [Lenski *et al.* 1991], are common. Since the best experimental measurements include the least

noise, I optimised various steps in my protocol to achieve a sensitivity for epistatic effects $\approx 2\%$. Fitness effects are weakly advantageous so long as $N_e s \approx 1$ [Keightley and Eyre-Walker 2010] (where N_e is effective population size, and s is the selective coefficient) meaning that my optimised protocol would not be sensitive to all advantageous epistatic effects in populations where $N_e > 50$. Since microbial experiments are generally run with populations much larger than 50, and involve more manipulation (*i.e.* one source of error) than a simple competitive fitness assay, I became interested in understanding what combination of factors may increase noise in data from microbial experiments. Many microbial experimental evolution studies will record which genomic changes result from selection. It has been observed that the repeatability of genomic evolution can be quite low in microbial experimental evolution despite the evolved mutations being strongly adaptive [Lässig *et al.* 2017]. With mutation rates on the order of 10^{-3} per genome per generation [Drake *et al.* 1998], and populations of at least millions, there is a non-trivial potential for hitchhiking mutations to arise, as well as standing variation to be generated, in relatively few generations of any microbial study. While my optimised competitive fitness protocol required roughly twenty generations of growth, many microbial evolution experiments will involve hundreds or thousands of generations. A factor we can control, and thus worth studying, is our experimental design which can influence the types of growth (*e.g.* exponential *v.s.* logistic) experienced by our studied microbes.

1.5.3 Chapter 4

Chemostat growth chambers represent an environment in which fresh liquid growth medium can be replaced at the rate at which it is exhausted by microbial

communities. While a chemostat setup can allow microbes to divide constantly throughout an experiment, the systems can be both expensive and challenging to setup and maintain. It is thus more common for experimental microbes to be grown in batch cultures wherein they undergo rounds of logistic growth punctuated by regular population bottlenecks introduced by the common practice of serial passaging. Serial passaging is the act of repeatedly transferring a small proportion of a large population to fresh growth media in order to prolong the generation time of an experiment. Theoretical work has previously highlighted that this practice is expected to cause loss of rare mutations [Wahl *et al.* 2002], and potentially bias the types of mutations that persist [Wahl and Zhu 2015]. Yet, theory also suggests that when growth is purely exponential, not logistic, serial transfer does not favour the time during growth phases at which mutations arise [Wahl *et al.* 2002]. I wanted to test if the practice of serial passaging, and the logistic growth form of microbes, interacted to bias which mutants will fix during microbial experimental evolution.

Thus, for my third study I developed the novel *in silico* experimental system rSHAPE (R-package for Simulated Haploid Asexual Evolution). I designed this system not only for the needs of my immediate question, but also as a singular framework to support the study of alternative growth dynamic, and fitness landscape, hypotheses giving results that could be measured, and thus compared, from consistent output. After ensuring that rSHAPE replicated expected fitness dynamics and evolutionary outcome for *de novo* mutants, I applied rSHAPE to assess how the practice of serial passaging, under conditions of exponential or logistic growth, affected the probability of novel mutation persisting within a population.

1.5.4 My thesis in a nutshell

The goal of my thesis was to develop novel computational tools for the detection and study of factors that interact to affect the evolution of asexual haploids. To advance our collective understanding of evolution, I have studied correlated evolution among amino acid (nucleotide) positions throughout viral (microbial) genomes as well as developed an *in silico* system for evaluating experimental conditions outside the realm of *in vitro* control. This thesis does not provide a complete analysis of whole genome correlated evolution, nor does it quantify the extent of all possible interactions among evolutionary factors. However, it serves to establish novel computational methods for addressing these questions and shares the insights of some initial forays into an underexplored realm of our understanding.

Chapter 2

Widespread historical contingency in influenza viruses

This chapter is published in:

Nshogozabahizi, J.C., Dench, J. and Aris-Brosou, S. 2017 "Widespread Historical Contingency in Influenza Viruses" Genetics 205(1):409-420, https://doi.org/10.1534/genetics.116.193979

Collaborator Contributions:

J.C. Nshogozabahizi and J. Dench are co-first authors having contributed equally. J. Dench shared in the writing of the manuscript and development of AEGIS' code, performed the simulation studies, and compared AEGIS to the work of Li *et al.* [2016]. J.C. Nshogozabahizi led writting of the manuscript, began development of AEGIS's code, analysed correlated evolution within influenza A using AEGIS and compared results to the literature.

2.1 Abstract

In systems biology and genomics, epistasis characterises the impact that a substitution at a particular location in a genome can have on a substitution at another location. This phenomenon is often implicated in the evolution of drug resistance or to explain why particular 'disease-causing' mutations do not have the same outcome in all individuals. Hence, uncovering these mutations and their locations in a genome is a central question in biology. However, epistasis is notoriously difficult to uncover, especially in fast-evolving organisms. Here, we present a novel statistical approach that relies on a model developed in ecology and that we adapt to analyze genetic data in fast-evolving systems such as the influenza A virus. We validate the approach using a two-pronged strategy: extensive simulations demonstrate a low-to-moderate sensitivity with excellent specificity and precision, while analyses of experimentally-validated data recover known interactions, including in a eukaryotic system. We further evaluate the ability of our approach to detect correlated evolution during antigenic shifts or at the emergence of drug resistance. We show that in all cases, correlated evolution is prevalent in influenza A viruses, involving many pairs of sites linked together in chains, a hallmark of historical contingency. Strikingly, interacting sites are separated by large physical distances, which entails either long-range conformational changes or functional tradeoffs, for which we find support with the emergence of drug resistance. Our work paves a new way for the unbiased detection of epistasis in a wide range of organisms by performing whole-genome scans.

Keywords: Correlated evolution; Epistasis; Networks; Influenza

2.2 Introduction

One of the most fundamental questions in biology concerns the emergence of new structures and new functions, in particular at the molecular and genetic level [Lynch 2007]. As such, a large body of experimental work has accumulated over the past decade to unravel the mutational history at the origin of simple phenotypes. For instance, one particular bacterial drug resistance is conferred by five mutations, but out of the $5! = 120$ possible ways in which these mutations can accumulate, only a handful of mutational trajectories are evolutionarily accessible [Weinreich *et al.* 2006]. This line of work suggests that some mutations are required in order for subsequent mutations to occur. Further work demonstrates that such permissive mutations are not limited to bacteria, as they are also found in vertebrate [Ortlund *et al.* 2007], yeasts [Sorrells *et al.* 2015] and viral systems [Gong *et al.* 2013]. However, while such chains of dependent or conditional substitutions —called historical contingency [Harms and Thornton 2014; Mohrig *et al.* 1995]— are expected to lead to mutational trajectories, their shape and ramifications are not completely elucidated.

As the experimental determination of these trajectories can be tedious, taking > 20 years in the case of the Long-Term Evolution Experiment [Blount *et al.* 2008], computational solutions were sought to reconstruct historical contingencies and the mutational correlations they imply. Initial solutions relied on protein sequence alignments to compute sitewise vectors of amino acid frequencies, from which pairs of co-evolving residues could be identified [Neher 1994; Taylor and Hatrick 1994]. While this general approach is still used in statistical physics to predict protein folds [Shindyalov *et al.* 1994; Sutto *et al.* 2015], numerous refinements were brought either through the use of metrics such as mutual information [Atchley *et al.* 2000; Gloor

et al. 2005; Korber *et al.* 1993] or by correcting for shared evolutionary history. One of the first methods to detect correlated evolution while accounting for phylogeny was given in the general context of the evolution of discrete morphological characters [Pagel 1994]. Further leveraging on Schöniger and von Haeseler [1994], the method was quickly extended to analyze RNA molecules by modeling dinucleotides [Muse 1995; Rzhetsky 1995] and to map the correlated residues hence detected onto a three-dimensional protein structure [Pollock *et al.* 1999; Poon *et al.* 2007a,b]. More recently, a full evolutionary model was proposed to detect epistatic sites in a Bayesian framework [Nasrallah and Huelsenbeck 2013]. However, such approaches rely on complex models and Bayesian computing tools, may scale up poorly with increasingly large data sets [Aris-Brosou and Rodrigue 2012; Poon *et al.* 2008a], and are also generally geared towards detecting positive epistasis (when two mutations increase fitness values).

Here we build on these developments to describe a novel, yet intuitive, statistical method for detecting correlated evolution among pairs of amino acids (AAs) in the typically fast-evolving influenza A virus [Worobey *et al.* 2014]. Our method takes inspiration from an approach developed in ecology and aimed at detecting correlated evolution among phenotypic traits [Pagel 1994; Pagel and Meade 2006]. We validate our approach using a two-pronged procedure based on both extensive simulations and analyses of experimentally-validated data sets (both in viral and in eukaryotic systems). The analysis of several large data sets leads us to reconsider the nature of epistasis in influenza viruses. We find evidence that interacting AAs form networks of sites undergoing substitutions that are most likely to be permissive (*i.e.* contigent), as they occur in a temporal sequence. These networks cover large physical distances among interacting AAs suggesting long-range structural and/or functional effects.

2.3 Materials and methods

2.3.1 General approach to detect epistasis

Repurposing of BayesTraits. In order to model correlated evolution at the molecular level, we employed the maximum likelihood model implemented in BayesTraits [Pagel 1994; Pagel and Meade 2006]. See Talavera *et al.* [2015] for a similar model. This model was originally developed as a time-homogeneous Markov process with discrete states in continuous time in order to investigate the coevolution of discrete binary traits on phylogenetic trees. This is achieved by testing whether a dependent model of trait evolution fits the data better than an independent model. The dependent model allows two traits to coevolve as the rate of change at one trait depends on the state at the other trait, while the independent model does not place any restriction on rates of change [Pagel 1994] (see Fig. A.2.1). In both cases, the likelihood function is optimised by summing (integrating) over all unobserved pairs of character states at internal nodes. As these two models are nested, a likelihood ratio test can be employed for model selection. The test statistic, twice the log-likelihood difference, is assumed to follow a χ^2 distribution with four degrees of freedom under the null hypothesis (independence).

This general framework further assumes a phylogenetic tree, with a known topology and branch lengths proportional to the amount of evolution separating each node. Both assumptions can be addressed as described below, either by a bootstrap analysis, or by resorting to trees sampled from their posterior distribution.

Data recoding. The model described above was implemented for binary traits. Here however, our goal is to analyze the coevolution of pairs of sites (DNA or AA) along a sequence alignment. In the case of proteins, each site has twenty possible

AAs, which represent the states of our system. To avoid resorting to tensor kernels, data recoding is therefore necessary to reconcile the data with the approach. In this context, we evaluated two strategies. First, AAs were partitioned according to their physicochemical properties. Binary properties (side-chain group types) were naturally recoded "0" / "1". For those with $k > 2$ states, we compared each of the k_i states against the other ones (k_j where $i \neq j$). For instance, in the case of charge, we first assigned state "0" to negative and state "1" to non-negative AAs, and circled through the two other states (Fig. A.2.2). This recoding was based on the physicochemical properties as implemented in the R package `protr` ver. 0.2-1 [Xiao *et al.* 2014]. However, results from this first method failed to find any significantly correlated pairs of mutations when applied to our first dataset. Hence a second recoding method was devised and used thereafter.

In our second recoding method, AAs were classified as either being in the outgroup or ingroup consensus state: at each position of an alignment, the outgroup state was defined as the majority-rule consensus AA present in a clade used to root the tree. Here, this rooting clade was defined either as the one containing the oldest sequences and this clade was removed for downstream analysis, or as the basal clade in a relaxed molecular clock analysis (see below). The ingroup state was then defined as any AA that differs from the consensus AA in the outgroup. Within each column of the alignment, sites sharing the outgroup state were all recoded as "0", while those sharing the ingroup state were recoded as "1".

Code optimisation. To discover pairs of AAs that are potentially interacting, we need to run the above model on all $\frac{n(n-1)}{2}$ pairs of sites in an alignment of length n. This computation can be prohibitively long even with an alignment of modest size. For instance, the influenza H3N2 nucleoprotein has 498 AAs, which leads to

analyzing 123,753 pairs of sites under both the dependent and the independent model (*i.e.* 247,506 models need to be run). To decrease the computational cost, we first compressed the alignment into site patterns [Yang 2006] (p. 105). Then, as pairs of sites can be independently compared, we parallelised the code using R's `foreach` (ver. 1.4.2) and `doMC` (ver. 1.3.3) [Analytics and Weston 2013] packages to take advantage of multicore / multiprocessor architectures (Fig. A.2.3). To account for multiple testing, we computed the False Discovery Rate (FDR) according to the Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg 1995]. We used R (v3.0.2) for all analyses [Team *et al.* 2013]. In the analyses presented below, the pairs of AAs identified to be interacting were subsequently mapped on three-dimensional protein models predicted by homology modeling with SWISS-MODEL [Biasini *et al.* 2014], and plotted using KiNG [Chen *et al.* 2009].

2.3.2 Validation based on simulations

To validate our method for detecting correlated pairs of sites, we followed two approaches: an extensive simulation study and the analysis of data sets in which epistasis was experimentally confirmed. We first present our simulation strategies that, in order to avoid biasing our results, were based on two different frameworks of simulating correlated evolution. The Supplementary Text presents additional simulation results.

Simulations with PHASE. First, sequences were simulated with PHASE 2.0 [Gowri-Shankar and Jow 2006], which allowed us to simulate two categories of sites: those that evolve independently and those that evolve in a correlated manner. Sites that evolve independently (main sequence of length l_i) were simulated using the General Time-Reversible (GTR) model of nucleotide substitution [Tavaré 1986] [see

also Aris-Brosou and Rodrigue 2012]. Both nucleotide frequencies and transition probabilities were arbitrarily set to reflect published values for *Pseudomonas aeruginosa* [Garrity *et al.* 2004], although this particular setting has no impact on downstream analyses. With PHASE, sites that evolve in a correlated manner (epistatic sites l_e) were simulated under the RNA7D model of dinucleotide substitution [Tillier and Collins 1998]. We modified the transition probabilities such that double substitutions occurred for 95% of changes, single substitutions for 5%, while mismatch substitutions were not permitted. In the results presented below, equilibrium dinucleotide frequencies were all assumed to be equal (set to $\frac{1}{16}$). Because influenza A viruses typically exhibit dinucleotide bias [Greenbaum *et al.* 2008], which could lead to false positives, we also performed simulations under the RNA16 model with a dinucleotide bias matching that of influenza A viruses (see Supplementary Text). As total sequence length was determined ahead of time (part of our factorial design), when simulating correlated evolution, we first simulated a number l_e of pairs of epistatic sites, which were then concatenated to the nucleotide sites simulated under the independent model of length l_i to obtain a final sequence of length $l_s = 2l_e + l_i$. Where correlated evolution was not simulated $l_s = l_i$.

Simulations with Coev. Our second simulation strategy employed Coev [Dib *et al.* 2014], which produces correlated evolution by randomly defining a dinucleotide profile (two-letter nucleotide states, *e.g.* AA, AT, CG, *etc.*) and simulating evolution under the Jukes-Cantor substitution model [Jukes and Cantor 1969]. The transition rates for each nucleotide of the dinucleotide may be set independently; however, as we have no *a priori* reason to assume either evolves with a different rate, we left these rates equal to their default setting. The strength of selection for the defined profile is governed by the ratio of the parameters d and s (*i.e.* d/s), which represent

the likelihood of a dinucleotide evolving toward ($d/s > 1$) or away ($d/s < 1$) from the defined dinucleotide profile. When $d/s = 1$, there is no selection and sites evolve independently; we varied the strength of selection from independent ($d/s = 1$) to the default setting offered by Coev-web [Dib *et al.* 2015] ($d/s = 100$), with intermediate strengths ($d/s = \{2, 33, 66\}$). PHASE was still used to simulate sites that evolved independently (l_i) to be consistent in our use of a GTR substitution model.

Sensitivity and specificity. Sensitivity (Specificity), *a.k.a.* the true positive (negative) rate, measures the proportion of actual positives (negatives) correctly identified. In both sets of simulations, we assessed the impact of branch lengths, tree shape, and number of sequences, both in the presence and absence of epistasis (Fig A.2.4B). With PHASE, branch lengths (b) were varied across a $\log_2$ scale for $b \in (-12, -1)$. To reduce possible factorial combinations, all branches of a single simulated tree were the same length. Two tree shapes were used, simulated tree topologies (τ) being either symmetrically bifurcating or pectinate (Fig. A.2.4). All simulated trees were ultrametric. Each tree contained a number of sequences (n_s) equal to 16, 32, 64 or 128. This led to a full factorial design containing 192 simulation conditions ($12b \times 2\tau \times 4n_s \times 2$ for with or without epistasis). Each simulation condition was replicated 100 times and l_s was set to 100 bp. When epistasis was simulated, the number of epistatic pairs was set to 3. An ANOVA was used to assess the significance of each of these factors. All tests were conducted at the $\alpha = 0.01$ (1%) significance threshold.

With Coev, branch lengths (b) were varied across a $\log_2$ scale for $b \in \{-12, -10, -8, -6 - 4\}$. Each tree contained a number of sequences (n_s) equal to 32, or 128. The same two tree shapes (τ) were used, being either symmetrically bifurcating or pectinate. This led to a full factorial design containing 100 simulation

conditions ($5b \times 2\tau \times 2n_s \times 5d/s$). Each simulation condition was replicated 100 times and l_s was set to 100 bp. When epistasis was simulated, the number of epistatic pairs was set to 3, as under PHASE. All tests were here again conducted at the $\alpha = 0.01$ (1%) significance threshold unless otherwise stated.

2.3.3 Validations based on previous evidence

Previous computational analyses. As a first validation of our approach on actual data, we reanalyzed those studied in a previous computational study [Kryazhimskiy *et al.* 2011]. The four data sets obtained from S. Kryazhimskiy (*pers. comm.*) consist of: 1,219 HA and 1,836 NA sequences from H1N1 viruses, as well as 2,149 HA and 2,339 NA sequences from H3N2 viruses. These data sets are here denoted KDBP11-H1 (HA in H1N1), KDBP11-N1 (NA in H1N1), KDBP11-H3 (HA in H3N2) and KDBP11-N2 (NA in H3N2), respectively.

Experimental evidence. As computational studies make predictions that are not always tested or validated, we reanalyzed data sets in which epistasis was experimentally confirmed. A number of recent studies reported evidence for epistasis and we present results on three of these.

First, we reanalyzed data published by Gong and collaborators [Gong *et al.* 2013] which comprised 424 H3N2 human influenza A nucleoprotein (NP) sequences spanning 42 years between 1968-2010. This gene was originally chosen because it evolves relatively slowly and is hence amenable to experimental validation as all substitutions can be easily tested by site-directed mutagenesis [Gong *et al.* 2013]. The viral sequences that they used were downloaded from the IVR database [Bao *et al.* 2008]. This data set is here denoted Gong13NP.

We also retrieved a second data set as analyzed by Duan and collaborators [Duan

et al. 2014], who used an alignment that contained 1366 human influenza A H1N1 neuraminidase (NA) collected between 1999 and 2009; as in their analysis, the 2009 H1N1 pandemic sequences were excluded. This data set is denoted Duan14NA.

Finally, a recent study performed a combinatorial analysis based on mutations of the eukaryotic $\text{tRNA}^{Arg_{CCT}}$ gene to detect epistasis [Li *et al.* 2016]. We ran our computational analysis on an alignment of the eukaryotic tRNA gene constructed in a manner similar to Li and collaborators: we downloaded all available eukaryotic $\text{tRNA}^{Arg_{CCT}}$ genes from GtRNAdb [Chan and Lowe 2016], aligned the sequences with the cmalign tool from Infernal [Nawrocki and Eddy 2013], and kept the 75 nucleotide region that spanned the conserved portion within *Saccharomyces sp.* As our method requires a phylogenetic tree, we used a phylogeny of eukaryotes[Hedges *et al.* 2015], keeping only those tips corresponding to the taxa in our alignment. Lastly, as the branch lengths of this tree were in millions of years, we rescaled these to expected number of substitutions using the best available eukaryotic tRNA molecular clock [Soares *et al.* 2009]. While Li *et al.* [2016] did not perform a fully exhaustive analysis of all pairs of sites, and our analysis requires sites to be polymorphic, we present our comparison for those pairs of sites tested in both of the analyses. For this analysis only, we used a statistical threshold of $\alpha = 0.05$ as in Li *et al.* [2016]. This data set is denoted Li2016.

2.3.4 Nature of correlations in influenza evolution

After performing these validations, both on simulations and on experimentally-validated data, we set out to investigate the nature of these interactions by testing two hypotheses. First, we revisited the work by Koel and collaborators [Koel *et al.* 2013] that experimentally validated the existence of AA

substitutions involved in changes of antigenic clusters. For this, we used 877 H3N2 human influenza hemagglutinin (HA) sequences as in Koel *et al.* [2013], collected between 1968 and 2003, to test if these substitutions responsible for antigenic changes also showed evidence for correlated evolution. This data set is denoted Koel13HA.

Second, we tested if our statistical approach could detect some evidence for correlated evolution at pairs of sites involving the S31N substitution, which is responsible for conferring resistance to the anti-influenza drug Adamantane [Abed *et al.* 2005]. For this, we retrieved 668 H3N2 human influenza A matrix protein 2 (M2) sequences that were collected between 1968 and 2003. This data set is denoted Adam03M2.

2.3.5 Phylogenetic analyses

Maximum likelihood was employed to estimate phylogenetic trees for the KDBP11-H1, KDBP11-N1, KDBP11-H3, KDBP11-N2, Gong13NP, Duan14NA and Adam03M2, alignments with FastTree (ver. 2.1.7) [Price *et al.* 2010b] under the WAG $+\Gamma_4$ model to account for among-site rate variation. Trees were rooted using the earliest sequences, which were AAD17229/USA/1918, CAA24269/Japan/1968, AAF77036/USA/1918, ABI92283/Australia/1968, Aichi68-NPA/Aichi/2/1968, A/Victoria/JY2/1968 and A/Albany/17/1968, respectively. The R package `APE` was used to visualise the trees [Paradis 2006]. For the Koel13HA data set, a rooted tree was reconstructed using BEAST (ver. 1.8.0) under a relaxed molecular clock assuming an uncorrelated lognormal prior [Drummond *et al.* 2006] and a constant-size coalescent prior under the FLU $+\Gamma_4$ substitution model. Analyses were run in duplicate to check convergence, for a total of 100 million steps with a

thinning of 5,000; log files were combined with LogCombiner after conservatively removing the first 10% of each chain as a burn-in period, as checked with Tracer ver. 1.5 (tree.bio.ed.ac.uk/software/tracer).

Phylogenetic uncertainty was taken into account by running our algorithm on bootstrapped trees (or trees sampled from the posterior distribution). Because only the Gong13NP data set showed a large proportion of SH-like aLRT [Anisimova and Gascuel 2006] node support values in the low $(0.0, 0.8)$ range (Fig. A.2.5-A.2.12), the results of the bootstrap analyses are only shown for this case. All data sets and the AEGIS script (Analysis of Epistasis & Genomic Interacting Sites) used in this work are available at github.com/sarisbro/AEGIS.

2.4 Results and discussion

2.4.1 Simulation studies

Excellent specificity but mediocre sensitivity. As a first means to validating our approach, we conducted a fully factorial simulation study. In order to avoid biasing our results, the simulation models differ from the analysis models. Our simulation results with PHASE demonstrate that alignments containing fewer than 32 sequences have a poor ability to detect epistasis, with a sensitivity generally $\leq 20\%$ (Fig. A.2.13). For this reason, we henceforth focus on alignments of at least 32 sequences.

The specificity (Sp) of our approach is never below 99.9999%, which occurs for the symmetric tree shape with the largest number of sequences and rather short branch length of 2^{-6} substitutions per site (Fig. 2.4.1A). Variation in specificity was so low that we report the value on a $-\log_{10}(1 - Sp)$ scale in order to emphasize the

performance of our approach against thresholds set by the minimum, mean and maximum number of pairwise comparisons among replicates. These thresholds represent the number of true negatives calculable in our analyses (*i.e.* (# pairwise comparisons - # epistatic pairs)/# pairwise comparisons) and thus performance above these thresholds demonstrates that high specificity results from low false positive detection rate, and not simply from a large dataset with unduly large numbers of true negatives. In spite of the low variation in specificity, the number of sequences (n_s) explains most of the variance (Table A.3.1). While we find a two-way interaction with tree shape (τ) and branch length, there was no direct interaction between n_s and either branch length or τ (Table A.3.1). We find a three way interaction between branch length, τ, and n_s, which reinforces the idea that while the number of sequences dominates the specificity of our method, the amount of evolutionary time (sum of branch lengths over topology) remains an important factor (see Fig. 2.4.1A).

While specificity is excellent, we find that sensitivity is mediocre (Fig. 2.4.1B), which is consistent with previous studies [Poon *et al.* 2007b]. Our approach's sensitivity is a function of branch lengths and the number of sequences (Table A.3.1), with the existence of an optimal branch length where sensitivity could become excellent, reaching close to 100%, before degrading quickly again (on a log scale). This optimal response reflects that short branch lengths carry no information while long ones have random site patterns and thus convey no information [Yang 1998]. Pectinate trees show an optimum for shorter branch lengths than symmetric trees, and an ANOVA confirms an interaction between tree shape and branch length with respect to sensitivity (Table A.3.1). Lastly, we note that larger alignments appear to have better sensitivity (Fig. 2.4.1A) and that this does have an interaction

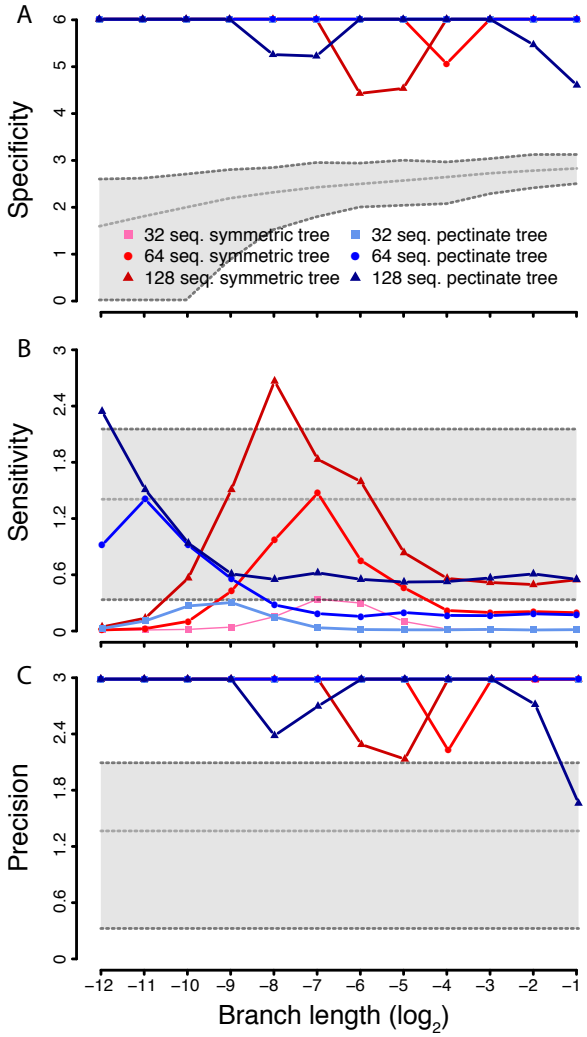

Figure 2.4.1: Specificity, sensitivity and precision results, from simulated data, of our novel epistasis detection method. Results are shown for alignments with 32 (squares), 64 (circles) and 128 (triangles) sequences. Tree shapes are colour-coded (symmetric in red; pectinate in blue). Branch lengths were varied on a $\log_2$ scale. All y-axes show the mean of $-log_{10}(1 - summary\ statistic)$ to highlight performance of our method as a $summary\ statistic$ approaches 1; when this value is 1, we arbitrarily assigned it a value 10% larger than the largest finite value within the set. Panel (A) shows specificity (true negative rate) with a grey shaded polygon which illustrates thresholds for excellent specificity in parameter space. The thresholds were established by subtracting three (the number of epistatic pairs simulated) from the minimum (lower dashed line), mean (middle dashed line) and maximum (upper dashed line) number of pairwise comparisons performed across all simulations with the same branch length. These thresholds represent the number of calculable true negatives and allow us to demonstrate our method's excellent specificity. Panels (B) and (C) show sensitivity (true positive rate) and precision (positive predictive value) respectively. Each panel includes a grey shaded polygon which illustrates thresholds for 50% (lower dashed line), 95% (middle dashed line) and 99% (upper dashed line) detection. These thresholds were arbitrarily chosen to demonstrate idealised benchmarks of performance.

with tree shape. We note that there exists large variance in sensitivity, due to our method of simulating epistasis, but have omitted error bars from Fig. 2.4.1 for reasons of clarity. Due to the low sensitivity of our method, we further investigated the prevalence of false positives by plotting precision (Fig. 2.4.1C). From this figure we find that less than 1% of positives are false except when branches are long.

Because dinucleotide bias, as typically found in influenza A viruses, could lead to false positives, we also assessed the impact of this parameter on the performance of our approach. We mimicked bias as found in the four KDBP11 data sets, and employed the full RNA16 model in PHASE to perform simulations similar to those under GTR as above. Our results show that both specificity and precision are unchanged, while sensitivity appears to be decreased. As we did not change the method simulating dependent sites, we did not expect sensitivity to be affected. After thorough review of our data we can only conclude that this difference results from the large variance in sensitivity resulting from the method by which epistasis is simulated. Altogether, dinucleotide bias does not lead to an increase in false positives (Fig. A.2.14).

To further confirm that our simulation results are not biased by our method of simulating epistasis, we compared our detection method to simulated evolution generated by Coev [Dib *et al.* 2014]. These results (see Supplementary Text) show that both specificity and precision are comparable with the PHASE simulations, while sensitivity is reduced (Fig. A.2.15). These findings reflect the particularities of how epistasis was simulated and confirm the excellent specificity and precision of our method. Importantly, these properties are found even when epistasis is simulated at very weak levels (Fig. A.2.15).

Note that unlike the real data sets analyzed in the next subsection, all these simulations are based on a state space of size four (DNA) and not 20 (AAs): the

effect of this reduction of the state space in the simulations is to lead to fewer unique site patterns, and hence larger proportions of false positives in DNA data than in AA data. Our DNA-based simulations show however that false positives are already well-controlled.

2.4.2 Real data analyses

Limited overlap with previous computational results. As a first evaluation of our algorithm on real data, we reanalyzed four large data sets (> 2000 sequences) previously analyzed with another computational method designed to detect epistasis [Kryazhimskiy *et al.* 2011]. Briefly, this method consists in mapping mutations on the phylogeny by parsimony, in order to then infer pairs of sites that are candidates for correlated evolution, before determining if mutations at each pair are temporally "close enough" to be considered as interacting. Unlike our approach, no data recoding is necessary; conversely, no mapping is involved in our method as the likelihood function describing changes of states at pairs of sites plays this role.

Here, the AA data were recoded as outgroup / ingroup character states to match the above simulations: at each position of the alignment, the "outgroup" state represents the majority-rule consensus AA in the outgroup sequences and the "ingroup" state represents all the AAs types that differ from the outgroup consensus AA. Overall, our approach not only detects fewer epistatic sites than the previous method, but the two approaches also show limited overlap (Fig. A.2.16). This marginal overlap is even found before FDR correction (Fig. A.2.17, Table A.3.4-A.3.7), so that lack of power is an unlikely explanation of the difference. Our extensive simulations suggest that this difference may be the result of the low-to-average sensitivity and of the excellent specificity / precision of our method,

so that the AAs pairs that we detect may actually be coevolving sites among the truly epistatic pairs. This result begs the question as to whether the few pairs of sites we detect would have any experimental evidence supporting epistasis.

Detected pairs are almost all experimentally-validated. Because the previous data sets were only examined from a computational point of view, we turned to additional data that have been experimentally validated.

First, we analyzed the Gong13NP data set of 424 influenza NP protein sequences [Gong *et al.* 2013]. As this is the smallest data set in our study, we assessed two ways of recoding the data into binary character states. We first partitioned AAs according to their physicochemical properties. With this recoding strategy, four pairs of sites were detected before FDR correction (Table A.3.8), but none of them matching the pairs detected by Gong and collaborators. Note that after FDR, no interactions were significant (Table A.3.9).

The estimated phylogenetic tree for this data set shows a pectinate-like (asymmetric) shape and short branch lengths (Fig. 2.4.2) – even if H3N2 viruses exhibit a more punctuated mode of evolution than H1N1 viruses [Sandie and Aris-Brosou 2014]. Our simulation results show that, under these conditions, with > 100 sequences, we can expect a sensitivity $\geq 80\%$ (Fig. 2.4.1A), so that the Gong13NP data set fulfills all the conditions for detecting true interactions. This suggests that even if the physicochemical recoding seemed *a priori* to be a good idea, capitalizing on the chemistry of life, it wastes statistical power on multiple three-way tests (Fig. A.2.2).

To better mimic our simulation conditions, we then recoded AAs as outgroup / ingroup states. With this recoding strategy, seven pairs of sites were detected:

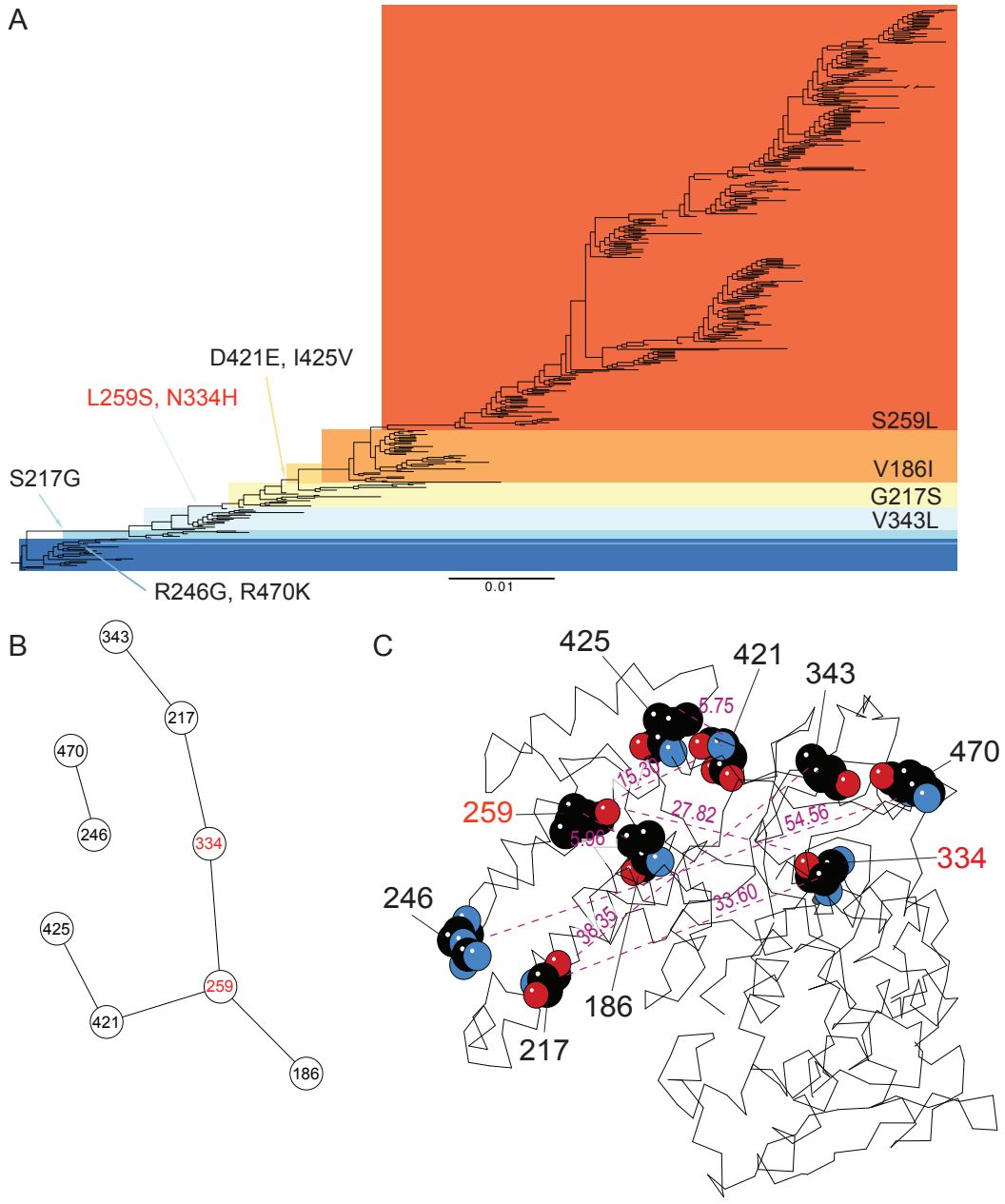

Figure 2.4.2: Epistatic pairs of amino acids detected in the Gong13NP data set with the outgroup / ingroup recoding. (A) The epistatic mutations that we detected are plotted on the NP phylogenetic tree. The substitutions in red were experimentally validated [Gong *et al.* 2013]. (B) Chained epistasis of interacting AAs. (C) The epistatic sites are mapped on three dimensional NP protein structure (based on template 3ZDP). The numbers show the AA positions experimentally validated (in red) and those detected only in this study (in black). The numbers in purple show the physical distance between epistatic sites (in Å).

259:334, 421:425, 246:470, 217:334, 217:343, 259:421 and 186:259 (Fig. 2.4.2A, Table A.3.9). Gong *et al.* [2013] detected 259:334 as their strongest signal, as well two additional pairs (384:65 and 280:312), with a weaker signal. However, we detected six additional pairs of sites that were never shown to be epistatic. These could either be false positives —but again our simulations suggest that our approach is extremely specific (Fig. 2.4.1B)— or simply coevolving pairs of sites that are not epistatic (coevolution is necessary but not sufficient for epistasis to exist), or that are evolving under negative epistasis.

While we find multiple pairs of correlated sites, almost all these pairs are linked to the experimentally confirmed L259S and N334H substitutions (Fig. 2.4.2B), hereby forming a network of interacting sites. Can we say anything about the nature of these interactions? If they were physical, we would expect that interacting sites would be in close spatial proximity on the folded protein, as in the case of compensatory mutations in RNA molecules [Chen *et al.* 1999; Kimura 1985]. However, the spatial distribution of these epistatic pairs of sites on the protein structure does not conform to this prediction: while two residues are considered physically linked when their distance is ≤ 8.5 Å [Atilgan *et al.* 2004], we find that the average distance between interacting AAs is 25.9 Å ($sd = 18.1$; Fig. 2.4.2C). This distribution strongly suggests that chained epistasis is not linked by spatially-close physical interactions (Fig. A.2.18), so that thermodynamic [Thomas *et al.* 2010] or compensatory changes of the 3D structure [Weinreich *et al.* 2006] act at very long spatial ranges.

To further validate our approach, we also analyzed the Duan14NA data set, comprising 1366 NA sequences of pre-pandemic H1N1 viruses [Duan *et al.* 2014]. We identify the epistatic pair 275:354 (Fig. 2.4.3, Table A.3.10), which was experimentally proven to confer oseltamivir resistance and that dominated the

population in 2008-2009 [Duan *et al.* 2014]. In their study, Duan and collaborators showed that D354G was the main mutation responsible for maintaining the function of NA after alteration of enzyme activity by H275Y (H274Y in N2 numbering), so that this is potentially the strongest existing interaction. However, our approach fails to identify the five other mutations (V234M, R222Q, K329E, D344N and D354G) that were further identified by Duan and collaborators to be interacting with H275Y. The estimated H1N1 tree is more symmetrical in shape than the one estimated for the Gong13NP data and has shorter average branch lengths (see scale bar in Fig. 2.4.3). Our simulation results suggest that in this case, the sensitivity of our approach can be very small (Fig. 2.4.1A). This low sensitivity might explain why we fail to detect the five additional sites interacting with position 275.

In spite of this negative result, we find that the interacting pair is, again, a long-range interaction (23.7 Å; Fig. 2.4.3, inset). The objective of the next two sections is to explore more systematically the nature of epistasis in influenza A viruses, focusing more specifically on (i) the prevalence of chained long-range epistasis and (ii) the potential nature of these long-range interactions.

Lastly, we ran our approach on the Li2016 data set. While our analysis found a signal for correlated evolution in 25.7% of the pairs of sites tested by Li *et al.* [2016], 90.0% of our significant pairs of sites were identified as epistatic by Li and collaborators (Fig. A.2.19A). We also found a small number of site pairs (37) that were not detected to be epistatic by Li and collaborators. We note however that (i) four of these 37 site pairs could not be statistically tested as the original study had only one biological replicate, (ii) all the site pairs we identify as correlated have epistasis measures that fall within the range of those site pairs deemed significant by Li and collaborators (Fig. A.2.19B) and (iii) one site pair we identified forms a

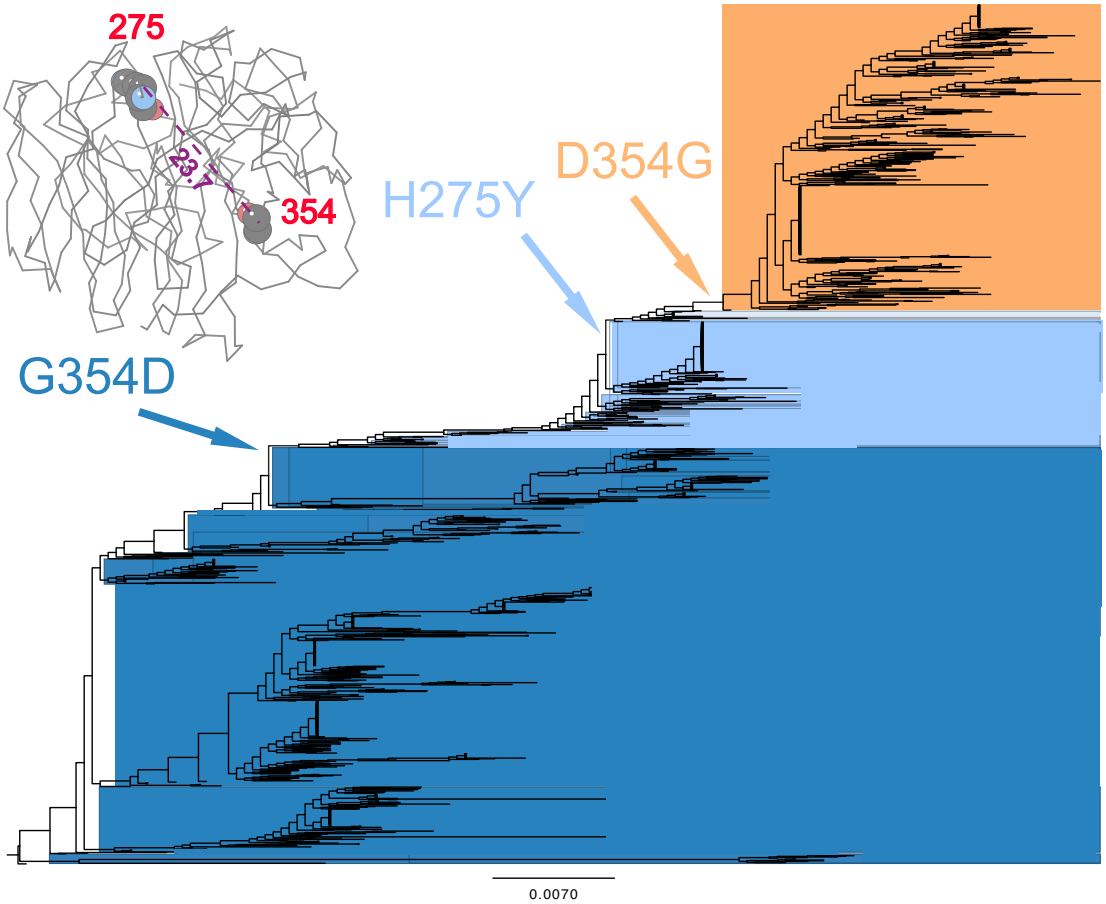

Figure 2.4.3: Epistatic pairs of amino acids detected in the Duan14NA data set. The epistatic mutations that we detected are plotted on the NA phylogenetic tree. Inset: the epistatic sites are mapped on three dimensional NA protein structure (based on template 1HA0). The numbers show the AA positions experimentally validated (in red). The numbers in purple show the physical distance between epistatic sites (in Å).

Watson-Crick base pair in the folded tRNA molecule. It is therefore possible that the variability in fitness measures in the Li et al.'s high throughput experiment be at least partially responsible for these discrepancies. In any case, this comparison supports the results of our simulation studies that show that our method has excellent specificity and mediocre sensitivity.

Correlated networks follow a temporal pattern. In a fifth study, the mutations involved in changes of antigenic clusters were investigated, as it was found that double mutations could suffice to explain such cluster changes [Koel *et al.* 2013]. The high mutation rate of this virus' antigen however does not explain why these cluster changes do not occur more often than the observed 3.3 years, which led Koel and collaborators to postulate that "co-mutations" may be required to maintain viral fitness, and hence accelerate evolutionary trajectories when one of the two mutations is neutral or even deleterious [Drake 2007]. In light of our study, the immediate interpretation of these results would be that correlated evolution is involved during such cluster changes. We re-analyzed these data in order to test this hypothesis.

By doing so, we find that some of the substitutions previously (and experimentally) implicated in cluster change are indeed involved in epistasis (Fig. 2.4.4A-B, Table A.3.11). In particular, as in the Gong13NP data set, we find evidence for networks of correlations, either as short chains such as position 155 interacting with both 158 and 146, which are involved in two consecutive cluster changes, but we also find a much larger network of interactions involving 20 sites, some of which are also involved in the last four cluster changes (Fig. 2.4.4B). Again, as in the Gong13NP data set, a temporal sequence of substitutions along this network can be found: G124D, found at the SI87/BE89 transition, interacts with

K299R, G172D and E82K; G172D interacts with sites involved in the next transition, BE89/BE92, such as G135K, which is again involved in the next transition, BE92/WU95, where G172E interacts with N262S and V196A, which is itself interacting with K156Q, involved in the WU95/SY97 transition; finally, K156Q interacts with T192I, involved in the SY97/FU02 transition. It is tempting to propose that such chained interactions reflect permissive substitutions and hence may provide an explanation, as an evolutionary constraint, to the paradox of high mutation rate and slow antigenic evolution. Yet, can we delve further into the nature of these constraints?

At first inspection, Fig. 2.4.4C suggests that all these interactions are located in the head of the HA protein and hence might respond to steric constraints, *i.e.* physically-mediated. However, Fig. A.2.20 shows that the strength of these associations is not related to physical distance between pairs of interacting sites. Again, the average distance between pairs of interacting sites is 23.7 Å ($sd = 10.4$), which is much larger than the canonical 8.5 Å for close proximity. Can we obtain some evidence about the nature of such long-range interactions?

Long-range interactions can be functionally-mediated. The analysis of a sixth data set, Adam03M2, sheds some light on this question. Influenza A viruses are resistant to M2 inhibitors, such as adamantane, and this resistance is associated with the S31N amino acid substitution, which is found in more than 95% of the currently circulating viruses [Garcia and Aris-Brosou 2014; Wang *et al.* 2013]. There is evidence supporting that the spread of S31N may be unrelated to drug selective pressure, but instead results from its interaction with advantageous mutations located elsewhere in the viral genome [Simonsen *et al.* 2007] – or maybe just in the

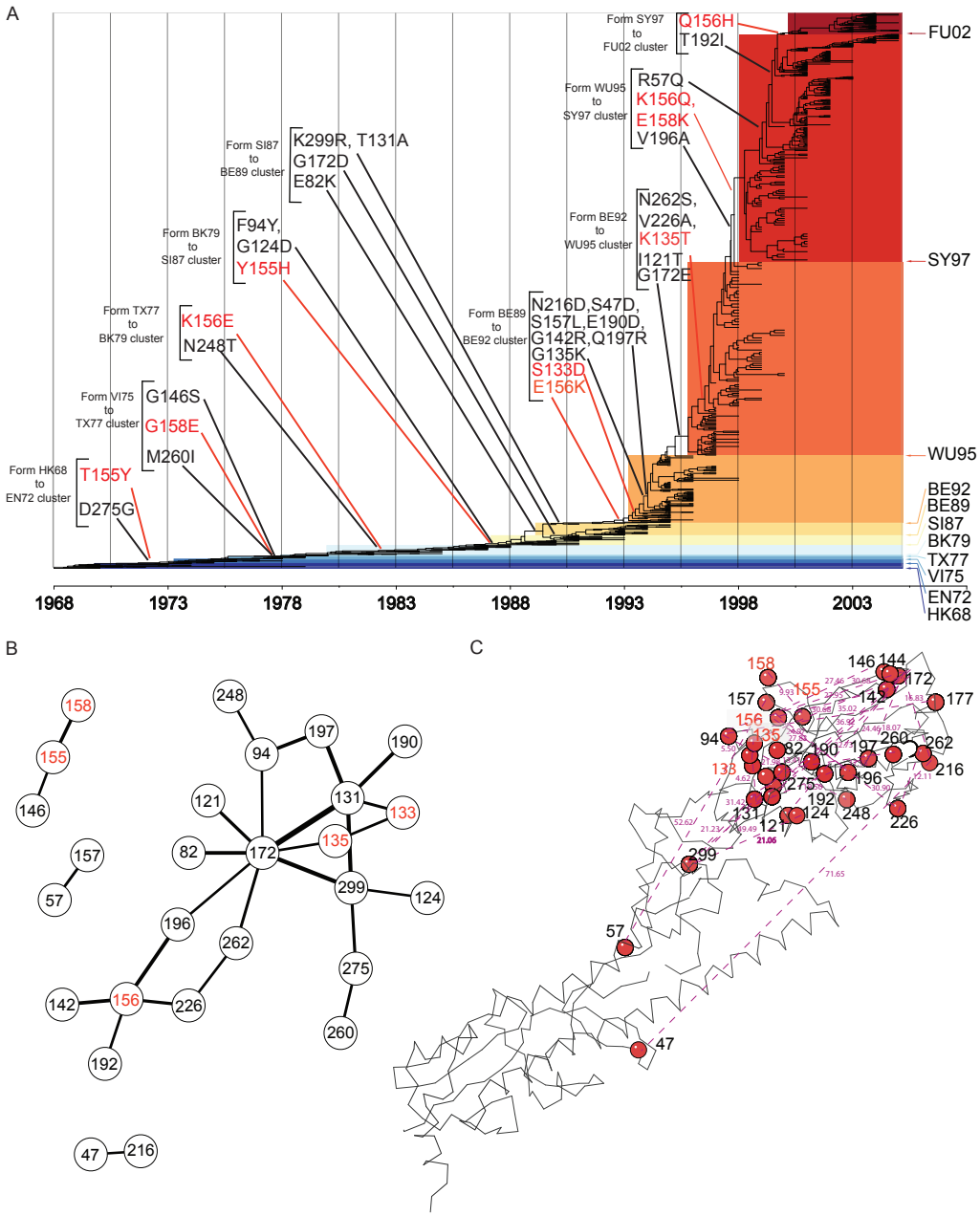

Figure 2.4.4: Epistatic pairs of amino acids detected in the Koel13HA data set. (A) The epistatic mutations that we detected are plotted on the HA phylogenetic tree. The substitutions in red were experimentally validated to be responsible for cluster change [Koel *et al.* 2013]. The antigenic clusters are named after the first vaccine strain in the cluster, with letters and digits referring to location and year of isolation (HK, Hong Kong; EN, England; VI, Victoria; TX, Texas; BK, Bangkok; SI, Sichuan; BE, Beijing; WU, Wuhan; SY, Sydney; FU, Fujian). (B) The thickness of the links is proportional to $-\log_{10} P$-value, the strength of evidence supporting the interaction. (C) The epistatic sites are mapped on three dimensional HA protein structure (based on template 3WHE). The numbers show the AA positions experimentally validated (in red) and those detected only in this study (in black). The numbers in purple show the physical distance between epistatic sites (in Å).

M2 gene. To test this epistatic hypothesis, we used our approach to analyze a data set of M2 sequences.

The results show that only one pair of epistatic sites (S31N×V51I) is detected (Fig. 2.4.5, Table A.3.12). Again, it is a long-range interaction (35.96 Å), but the reason why this data set is illuminating is that the mutation at position 51 has been shown to play a role in virus replication by stabilizing the amphipathic helixes of the M2 protein [Stewart and Pekosz 2011]. Thus, V51I may enhance the fitness of M2 protein to increase the frequency of adamantane resistance associated with S31N mutation. The reversion I51V that appeared in few sequences in 2000 (red clade in Fig. 2.4.5) was apparently quickly lost, which supports the hypothesis that V51I mutation is permissive of S31N. This reversion also supports that, even in the face of high mutation rates, (i) our algorithm still maintains high specificity (Fig. 2.4.1) and (ii) epistasis can be a very powerful force.

2.4.3 Conclusions

With historical contingency, the accumulation of epistatic substitutions can be seen as a coevolutionary process, where what happens at one AA site is conditional on what happened at another site. Here, it is this very idea of coevolution that we harnessed by co-opting a method developed in ecology to test for the correlated evolution of phenotypic traits [Pagel 1994; Pagel and Meade 2006]: instead of treating pairs of phenotypic traits as such, we repurposed the method to deal with pairs of AA sites. Because the original method was developed to handle binary traits, we explored two ways of recoding data and showed that treating AAs as outgroup/ingroup consensus states was a more sensible (albeit less intuitive) option than using physicochemical properties. Extensive simulations demonstrated

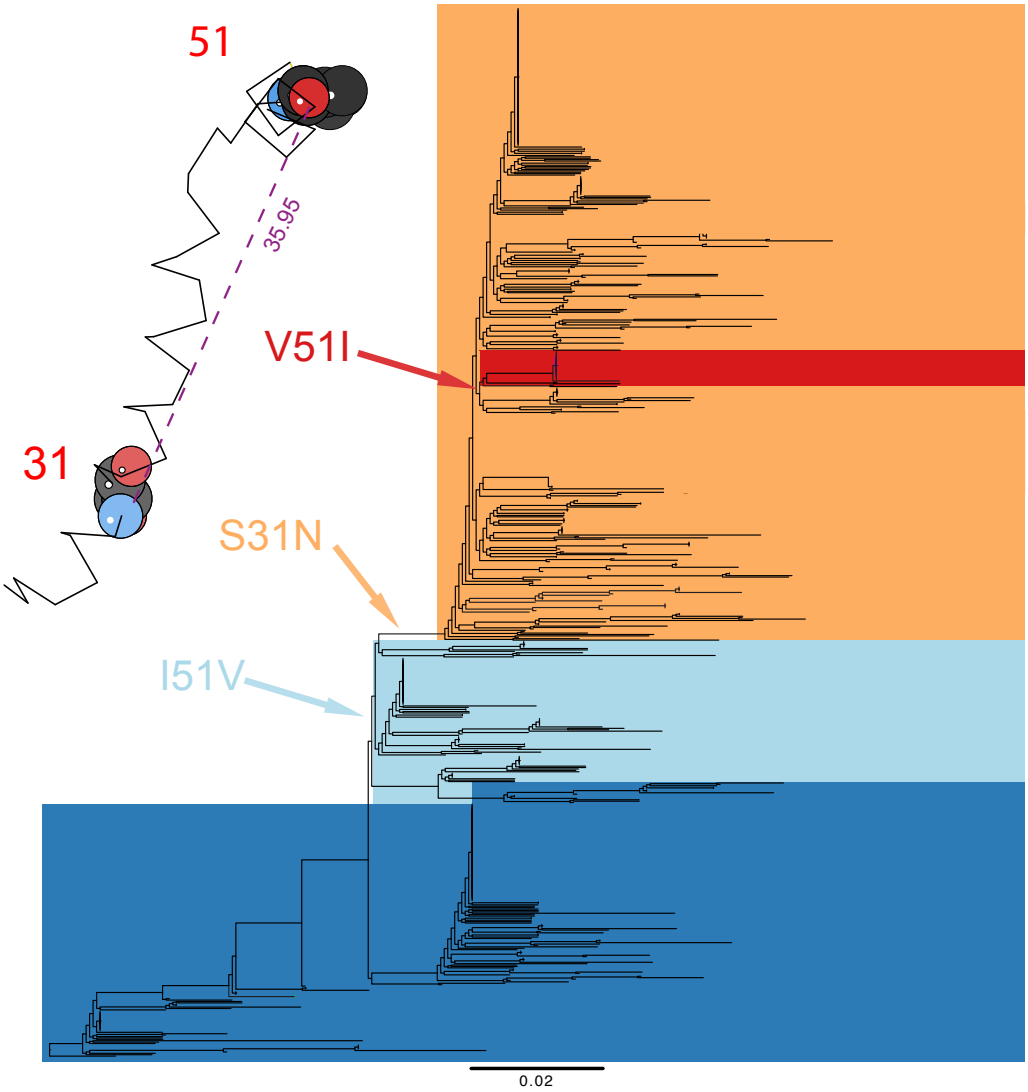

Figure 2.4.5: Epistatic pairs of amino acids detected in the Adam03M2 data set. The epistatic mutations that we detected are plotted on the M2 phylogenetic tree. Inset: the epistatic sites are mapped on three dimensional M2 protein structure (based on template 2KIH). The numbers show the AA positions detected (in red). The numbers in purple show the physical distance between epistatic sites (in Å).

mediocre sensitivity, but excellent specificity and precision, even in the face of dinucleotide bias, so that pairs of AAs that are detected can be assumed to be actually coevolving. We then validated our approach against other computational results, showing marginal overlap, and against experimentally-validated results, showing extensive overlap, hereby suggesting that detected pairs of AAs are genuinely interacting.

With this good statistical behavior, further analyses of independent influenza data sets showed a consistent pattern: (i) many pairs of correlated sites are involved in epistatic interactions, (ii) these pairs of sites form extensive networks of sites, consistent with Poon *et al.* [2007b], but that are also affected by substitutions that occur sequentially – showing evidence for historical contingency in fast-evolving organisms – and (iii), more intriguingly, that these epistatic pairs of sites form long-range spatial interactions. This latter point precludes the idea of a close physical link as in the case of tRNA molecules [Chen *et al.* 1999; Kimura 1985], so that these long-range interactions must bring about stability [Thomas *et al.* 2010] and/or conformational [Harms and Thornton 2014; Mitraki *et al.* 1991; Newcomb *et al.* 1997] and/or functional changes, as in the case of the M2 data set or in the case of the Ebola virus [Ibeh *et al.* 2016]. While there is evidence that epistasis can be prevalent in RNA viruses (at least 31% in [Shapiro *et al.* 2006]) and in bacteria (15% in [Weinreich *et al.* 2006]), it is not impossible that epistasis reflects an evolutionary constraint stronger in RNA viruses than in organisms with larger and more redundant genomes: because these viruses have a small genome, mutations are expected to have large fitness effects, which can be alleviated by compensatory mutations [Sanjuán and Elena 2006].

Our choice of focusing on a segmented RNA virus such as influenza may be problematic, in particular in the case of H3N2 viruses, which show a pattern of

punctuated evolution that can be interpreted as the result of clonal interference [Illingworth and Mustonen 2012; Strelkowa and Lässig 2012]. In the absence of recombination, each segment of the virus evolves as a clone, and each clone accumulates different beneficial mutations, only one of which becomes fixed, alongside hitchhiking deleterious mutations that, in our context, would show a pattern of correlated evolution. While (i) this process has to date only be found in H3N2 viruses, and (ii) we also find evidence for chained epistasis in H1N1 viruses (KDBP11-H1, KDBP11-N1, Duan14NA) as well as in eukaryotic tRNA gene (Li2016), clonal interference remains a problem for our approach. On the other hand, the use of influenza has allowed us to limit our analysis to searching for evidence of epistasis within genes, *contra* among genes. This way of analyzing data, intra-genically, might make sense when different segments code for proteins involved in relatively different functions. However, many RNA viruses have overlapping reading frames (such as the M and NS segments in influenza A), so that we can also expect coevolution across viral genes [Neverov *et al.* 2015]. Furthermore, experimental evidence in other organisms, such as yeasts, shows that epistatic interactions can involve multiple genes and hence be *inter*-genic [Sorrells *et al.* 2015]. Although a method to detect epistasis in segmented genomes was proposed [Neverov *et al.* 2015], the computational costs of whole-genome scans can seem prohibitive. Assessing the prevalence of epistasis over entire genomes remains an unexplored area – but one that we are currently investigating.

In this context, how can we explain the existence of large networks of interacting sites? One possibility would be that a changing environment creates new adaptive landscapes, and that natural populations (*contra* those from *in vitro* studies) do not climb peaks on the landscape but rather chase moving targets [Gavrilets 2004, p. 36]. While it is not clear whether such landscapes are robust to changing

environments [Hartl 2014], they are certainly a reality in the world of viruses, where vaccination regularly alters the adaptive landscape, hereby leading to chained networks of epistatic interactions – as a mere by-product of evolution. But then, one can wonder if the metaphor of a landscape itself is appropriate when all mutational trajectories are not accessible from specific genomic backgrounds [Sorrells *et al.* 2015; Weinreich 2010]. This may be one of the reasons why evolution is so difficult to predict [Natarajan *et al.* 2016; Sandie and Aris-Brosou 2014; Weinreich 2010].

Chapter 3

Identifying the drivers of computationally detected correlated evolution among sites under antibiotic selection

This chapter is published in:

Dench, J., Hinz, A., Aris-Brosou, S., and Kassen R., 2020 "Identifying the drivers of computationally detected correlated evolution among sites under antibiotic selection" Evolutionary Applications 13(4):782-794,
https://doi.org/10.1111/eva.12900

Collaborator Contributions:

J. Dench performed the experiment, analysis, and led writing of the manuscript.

A. Hinz designed the site directed mutagenesis protocol used to generate mutants and helped writing the manuscript.

S. Aris-Brosou and R. Kassen proposed the study, supported development of the experimental approach, and helped write the manuscript.

3.1 Abstract

The ultimate causes of correlated evolution among sites in a genome remain difficult to tease apart. To address this problem directly, we performed a high-throughput search for correlated evolution among sites associated with resistance to a fluoroquinolone antibiotic using whole genome data from clinical strains of *Pseudomonas aeruginosa*, before validating our computational predictions experimentally. We show that for at least two sites, this correlation is underlain by epistasis. Our analysis also revealed eight additional pairs of synonymous substitutions displaying correlated evolution underlain by physical linkage, rather than selection associated with antibiotic resistance. Our results provide direct evidence that both epistasis and physical linkage among sites can drive the correlated evolution identified by high throughput computational tools. In other words, the observation of correlated evolution is not by itself sufficient evidence to guarantee that the sites in question are epistatic; such a claim requires additional evidence, ideally coming from direct estimates of epistasis, based on experimental evidence.

Keywords: Correlated evolution, epistasis, antibiotic resistance, *Pseudomonas aeruginosa*, computational methods, experimental validation

3.2 Introduction

Adaptive evolution often involves changes to multiple characters (or traits) in concert, a process called correlated evolution. Such coordinated changes arise either because of physical linkage in the genome, or from strong selection that generates an association between distinct genomic sites. The analysis of correlated evolution among traits has a long history in quantitative genetics [Cheverud 1984; Falconer 1960; Lynch and Walsh 1998], molecular biology [Callahan *et al.* 2011; Goh *et al.* 2000], life history evolution [Baer and Lynch 2003; Kelley *et al.* 2013; Sexton *et al.* 2009], and comparative biology [Pagel 1994; Shimizu *et al.* 2014]. A comparable effort at the genomic level remains a daunting task because the number of potentially interacting sites (*i.e.* nucleotides) can be overwhelmingly large, making it difficult to distinguish genuine instances of correlated evolution arising from linkage or selection from spurious correlations arising due to chance.

In an effort to fill this gap, we have extended a computational approach, implemented in a software called AEGIS (Analysis of Epistasis and Genomic Interacting Sites), designed to detect both positively and negatively correlated pairs of mutations with very high specificity [Nshogozabahizi *et al.* 2017]. Although the original approach focused on identifying correlated evolution among sites within genes [Aris-Brosou *et al.* 2017; Ibeh *et al.* 2016; Nshogozabahizi *et al.* 2017], nothing prevents it from being used to perform the same task across entire genomes. AEGIS makes use of Pagel's phylogenetically informed maximum likelihood model for predicting correlated evolution between pairs of traits based on the co-distribution of their values [Pagel 1994], and was extensively tested through both simulations and analyses of real data in the context of detecting correlated evolution at the molecular level, between pairs of nucleotide positions in a particular genome

alignment [Nshogozabahizi *et al.* 2017]. The output of AEGIS is a list of pairs of sites and their associated probabilities of co-evolving by chance. While this approach uses phylogenetic information to reduce the detection of spurious associations, it is agnostic to the underlying mechanism responsible for correlation among sites, and thus cannot distinguish between linkage and selection leading to nonadditive fitness effects, or epistasis, as the cause of correlated evolution. This approach, like most statistical comparative methods, are now known to detect correlated evolution even in the case of a single unreplicated co-distribution of traits [Maddison and FitzJohn 2014; Uyeda *et al.* 2018].

Consequently, identifying the mechanisms driving correlated evolution still requires direct experimental measurements of epistasis between sites, preferably under conditions similar to the selective setting in which the pairs of sites evolved. To do this, we focused on the evolution of resistance to fluoroquinolone antibiotics in the opportunistic pathogen *Pseudomonas aeruginosa*. This pathogen is a ubiquitous, Gram-negative bacterium that causes both acute opportunistic infections and chronic respiratory tract infections in cystic fibrosis patients. Treatment with fluoroquinolone antibiotics imposes strong selection that regularly leads to the evolution of resistance at a number of well-known target sites such as efflux pump regulators (*e.g. nfxB*) and DNA topoisomerases (*i.e. gyrA*, *parC*) [Akasaka *et al.* 2001; Kos *et al.* 2015; Melnyk *et al.* 2015; Wong *et al.* 2012]. As *P. aeruginosa* is amenable to genetic manipulation, we can introduce putative correlated mutations, either one at a time or in combination, into a range of genetic backgrounds. The level of epistasis can then be measured by comparing the effects of the two substitutions against the expected combined effects of each single substitution on traits relevant to selection, such as antibiotic resistance or fitness in the absence of antibiotics.

Here, we use a unique data set of 393 *P. aeruginosa* exomes to predict correlated evolution among sites and evaluate the mechanistic causes of correlations that evolve during selection by fluoroquinolone antibiotics. We investigate the role of epistasis and linkage, both physical and through hitchhiking, as causes of correlated evolution using a combination of tools including site-directed mutagenesis to reconstruct single and double mutants. For nonsynonymous substitutions that co-evolved in response to antibiotic selection, we test for epistasis with direct estimates of the minimum inhibitory concentration (MIC) of antibiotics and competitive fitness in the absence of drugs. We employ additional computational analyses to infer the mechanism leading to the correlated evolution of synonymous substitutions. Our results demonstrate that epistasis can drive correlated evolution among nonsynonymous sites tied to antibiotic resistance, whereas correlated evolution among synonymous sites is best explained by idiosyncratic processes.

3.3 Materials and methods

3.3.1 Alignment and phylogeny

We assembled a multiple sequence alignment which included the complete genomes of *P. aeruginosa* strains PA01 (PA01), *P. aeruginosa* UCBPP PA14 (PA14), and *P. aeruginosa* PA7 (PA7) downloaded from `www.pseudomonas.com` [Winsor *et al.* 2016, accessed Dec 4, 2014], as well as 390 draft *P. aeruginosa* genomes [Kos *et al.* 2015], used here as they were published alongside information detailing which of the strains were resistant to the fluoroquinolone antibiotic levofloxacin. The genomes used in our study represent a collection of clinical strains isolated from around the world (Table B.3.1). Due to dissimilarities among contigs across the draft genomes,

we were unable to construct a global alignment using whole genome alignment algorithms as implemented in either MAUVE [Darling *et al.* 2004] or MUGSY [Angiuoli and Salzberg 2011]. Using PA14 reference gene sequences detailed in a database downloaded from `www.pseudomonas.com` [Winsor *et al.* 2016, accessed Dec 4, 2014], we assembled an exome alignment (*i.e.* all coding sequences) using a custom algorithm to concatenate individual gene alignments (see *Alignment algorithm* B.1.1).

The species tree we estimated for these bacterial genomes was reconstructed based on 'highly networked' genes that, according to the complexity hypothesis, are unlikely to be horizontally transferred. We identified these highly networked genes from the Cluster of Orthologous Groups database's [Galperin *et al.* 2015; Tatusov *et al.* 1997] information class genes (those with COG terms A, B, J, K, and L). From an alignment of 1,290 highly networked *P. aeruginosa* genes, we estimated the species tree by maximum likelihood with FastTree [Price *et al.* 2010a] (GTR + Γ model of evolution [Aris-Brosou and Rodrigue 2012]) compiled with *DUSE_DOUBLE* as recommended for large sequence alignments. This tree was rooted with the subclade containing the taxonomic outliers PA7 [Roy *et al.* 2010], AZPAE14941, AZPAE14901 and AZPAE15042 [Kos *et al.* 2015]. After rooting the tree, we removed the subclade from both the phylogeny and the above exome alignment, so that downstream analyses were based on 389 genomes. For additional information please see *Alignment algorithm* (B.1.1) and *Complexity Hypothesis* in the supplementary information.

3.3.2 Analysis of correlated evolution

We used AEGIS [Nshogozabahizi *et al.* 2017] to detect pairs of nucleotide positions (sites) that evolved in a correlated manner. For this, AEGIS performed pairwise comparisons between sites, testing if a model of dependent evolution was significantly better than an independent model at explaining the observed phylogenetic distribution of nucleotide states. This maximum likelihood analysis relied on the algorithm [Pagel 1994], computationally validated with AEGIS for use with amino acid data [Nshogozabahizi *et al.* 2017], as implemented in BayesTraits for binary states [Pagel and Meade 2006]. Nucleotide states in our alignment were converted into binary characters where '0' was assigned where there was a consensus of nucleotide character at 80% of the levofloxacin sensitive strains and '1' otherwise. We chose this method of recoding so that signals of correlated evolution would more likely reflect adaptation to fluoroquinolone antibiotic selection. All R scripts and complete results can be found at `github.com/JDench/sigResults_AEGIS_inVivo`. Due to computational limitations, we limited our analysis of correlated evolution to a 12 focal gene by exome analysis. Six of the focal genes were previously shown to evolve during fluoroquinolone selection [Akasaka *et al.* 2001; Wong *et al.* 2012] (*gyrA*, *gyrB*, *nfxB*, *parC*, *parE*) or were linked to biofilm formation [Choy *et al.* 2004], which can contribute to antibiotic resistance [Starkey *et al.* 2009] (*morA*). The six other genes (*dnaA*, *dnaN*, *lpd3*, *ribD*, *rpoB*, *serC*) were randomly selected from among the other 5,971 genes in our alignment. For additional information please see *Analysis of correlated evolution* (B.1.3) in the supplementary information.

3.3.3 Direct experimental assessment of epistasis

All growth and incubation conditions were performed in lysogeny broth (LB) [Bertani 1951] containing half the normal salt (*i.e.* 5g/L NaCl), at 37°C, in an orbital shaker set at 150 rpm.

Generating mutants: To measure the *in vitro* fitness and antibiotic resistance effects of substitutions predicted with AEGIS, we created mutant constructs using a modified selection counter-selection allelic replacement protocol [Melnyk *et al.* 2017]. Here, replacement alleles were constructed from the wild-type (WT) template and contained only the mutations of interest rather than coming from an evolved strain which may carry non-focal mutations. As the effect of a mutation may be contingent upon the genetic background (*i.e.* WT) in which it arose [Blount *et al.* 2008; Kryazhimskiy *et al.* 2014; Sorrells *et al.* 2015], constructs were created using two distinct genetic backgrounds: PA01, and PA14 [Dettman *et al.* 2013]. For additional information please see *Modified selection counter-selection protocol* (B.1.4) in the supplementary information.

Minimum inhibitory concentration assays: We performed minimum inhibitory concentration assays to identify if our mutant constructs had increased resistance to the fluoroquinolone antibiotic ciprofloxacin. Using a 96-well plate, we carried out serial 2-fold dilutions of ciprofloxacin, in a liquid LB broth, such that columns represented a concentration gradient from 32 to 0.015625 (*i.e.* $2^{(5,4,\ldots,-6)}$) μg/mL. We inoculated each well with 5 μL of overnight mutant culture such that each row represented a single type of inoculum. After 24 hours of growth, we read the absorbance values (at 600 nm) using a plate reader (Table B.3.7). Each plate contained at least one blank row, used as reference for reads of the absorbance. We performed four replicates of each genotype, while ensuring that no genotype

appeared on a single plate more than once.

Competitive fitness assays: We measured relative fitness of WT and mutant constructs via competitive fitness assays. Competitions began when we mixed an equal volume of overnight culture, diluted 100 fold, of a focal strain and a marked competitor [Lenski *et al.* 1991]. Similar to previous work, the marked competitor was a WT strain bearing the neutral *lacZ* marker gene. The frequency of focal and marked strains were estimated using direct counts of diluted culture plated immediately after mixing and again after competition during overnight growth to stationary phase. We plated six replicates of each competition, and time point, on minimal media agar plates containing X-Gal. Counting was done after overnight growth at 37°C, followed by 2-4 days of growth at room temperature. For additional information please see *Calculating competitive fitness* (B.1.5) in the supplementary information.

3.3.4 Quantifying importance of sites in predicting resistance

We quantified the importance of polymorphic sites in our alignment for predicting resistance to levofloxacin based on the machine learning classification algorithm 'adaptive boosting', implemented in the *adaBoost* [Alfaro *et al.* 2013] R package. Measurements of importance provided a relative measure of the strength of association between resistance and substitution at a site in our alignment. We compared the machine learning results to substitutions known to be within the six, previously identified, focal *P. aeruginosa* genes of which five evolve in response to fluoroquinolone selection [Akasaka *et al.* 2001; Kos *et al.* 2015; Wong *et al.* 2012] and one is involved in biofilm formation [Choy *et al.* 2004] which can contribute to

antibiotic resistance [Starkey *et al.* 2009]. For additional information please see *adaptive boosting algorithm* (B.1.6) in the supplementary information.

3.3.5 Assessing horizontal gene transfer

Horizontal gene transfer (HGT) was assessed with a phylogenetic approach [Ravenhall *et al.* 2015], where we reconstructed each gene tree as above (maximum likelihood under GTR + Γ), and tested for tree concordance with the Approximately Unbiased (AU) test [Shimodaira 2002]. This was done with CONSEL [Shimodaira and Hasegawa 2001], on the *baseml* (GTR + Γ_5, no clock) output from PAML [Yang 2007].

3.3.6 Testing for biological effects of synonymous substitutions

To test if pairs of synonymous substitutions were correlated due to translational effects, we estimated the free energy (ΔG) of folded mRNA transcripts for both WT and mutant transcripts, which we normalised as relative values through division of the WT estimate (ΔG_{rel}). We also estimated if synonymous mutations affected translation elongation rates via estimates of the index of translation elongation [I_{TE}: Xia 2014]. For additional information please see *Calculating* ΔG *and* I_{TE} (B.1.7, B.1.8) in the supplementary information.

3.4 Results and discussion

3.4.1 Predicting correlated evolution among sites

We identified pairs of sites that evolve in a correlated manner in the context of fluoroquinolone resistance by first assembling a multiple sequence alignment using the entire set of 5,977 protein-coding genes from 393 previously sequenced *P. aeruginosa* strains. This multiple sequence alignment included three reference strains (PA01, PA14, and PA7) and 390 clinical strains for which we had information on the level of levofloxacin (a fluoroquinolone antibiotic) resistance [Kos *et al.* 2015]. We then used this alignment to construct a phylogenetic tree rooted by PA7, an outlier strain [Roy *et al.* 2010] (Fig. B.2.2). The subclade containing this strain was removed for downstream analyses. We acknowledge that the comparative approach implemented in AEGIS may not be ideally suited to analyze highly promiscuous bacteria; however, we constructed our phylogeny from genes unlikely to undergo horizontal transfer so that signals of correlated evolution would represent independent evolutionary events.

We subsequently employed AEGIS to identify pairs of nucleotide positions that show evidence of correlated evolution. As this approach can be computationally expensive, requiring on the order of n^2 comparisons in a genome of length n ($\sim 84 \times 10^9$ comparisons), we focused our analyses on a subset of twelve genes against all other positions in the exome. AEGIS calculated the probability that each position within these twelve genes evolved in a correlated manner with any other nucleotide position in the entire *P. aeruginosa* exome. Among the more than 10 million pairwise comparisons, we found that $\sim 127,000$ pairs of sites showed evidence of significant correlation with $P \leq 0.01$ after a Benjamini-Hochberg correction for false discovery rate (FDR). At medium ($P \leq 10^{-7}$) or high

($P \leq 10^{-11}$) significance levels, as previously determined based on extensive simulations [Nshogozabahizi *et al.* 2017], only 178 and nine pairs, respectively, showed evidence for correlated evolution (Fig. 3.4.1). Among the nine pairs of sites showing the strongest evidence for correlated evolution, we found substitutions in the nucleotide sequence for each of the two DNA gyrases (*gyrA* and *gyrB*) targeted by fluoroquinolones, within the canonical fluoroquinolone resistance determining gene *parC*, in the biofilm-associated gene *morA*, but also in *dnaN* which was among the set of randomly chosen genes. Notably, all highly significant pairs of substitutions were synonymous, and hence unlikely to affect drug resistance or fitness [but see Agashe *et al.* 2016; Bailey *et al.* 2014; Plotkin and Kudla 2011], save for one pair of nonsynonymous sites, *gyrA* (c248t)-*parC* (c260g/t), that presumably impacts the level of resistance (Fig. 3.4.1). The next most strongly correlated pair of nonsynonymous substitutions involved *parC* and an uncharacterised gene at locus tag *PA14_34000*, but its significance was several orders of magnitude lower than any of the nine strongest correlated pairs ($P = 8.018 \times 10^{-7}$ here *vs.* $P \leq 10^{-11}$ above). In total, we found evidence of correlated evolution among 96 paired nonsynonymous substitutions with $P \leq 10^{-4}$, and many more with $P \leq 10^{-2}$. However, most involved uncharacterised genes while the rest involved genes without sufficient biological rationale to justify investigation within our study. It should be noted that the algorithm we employed to detect correlated evolution [Pagel's algorithm; Pagel 1994], just like pretty much any alternative phylogeny-aware methods used in comparative studies [*e.g.* *Huelsenbeck* et al. *2003; Maddison 1990; Maddison* et al. *2007]*, can detect a significant statistical association even when there is a single case of co-distribution of traits [Maddison and FitzJohn 2014; Uyeda *et al.* 2018]. This is why, despite the care taken in our approach, we performed additional experimental analyses to validate the statistical signal

detected between the most strongly correlated pairs.

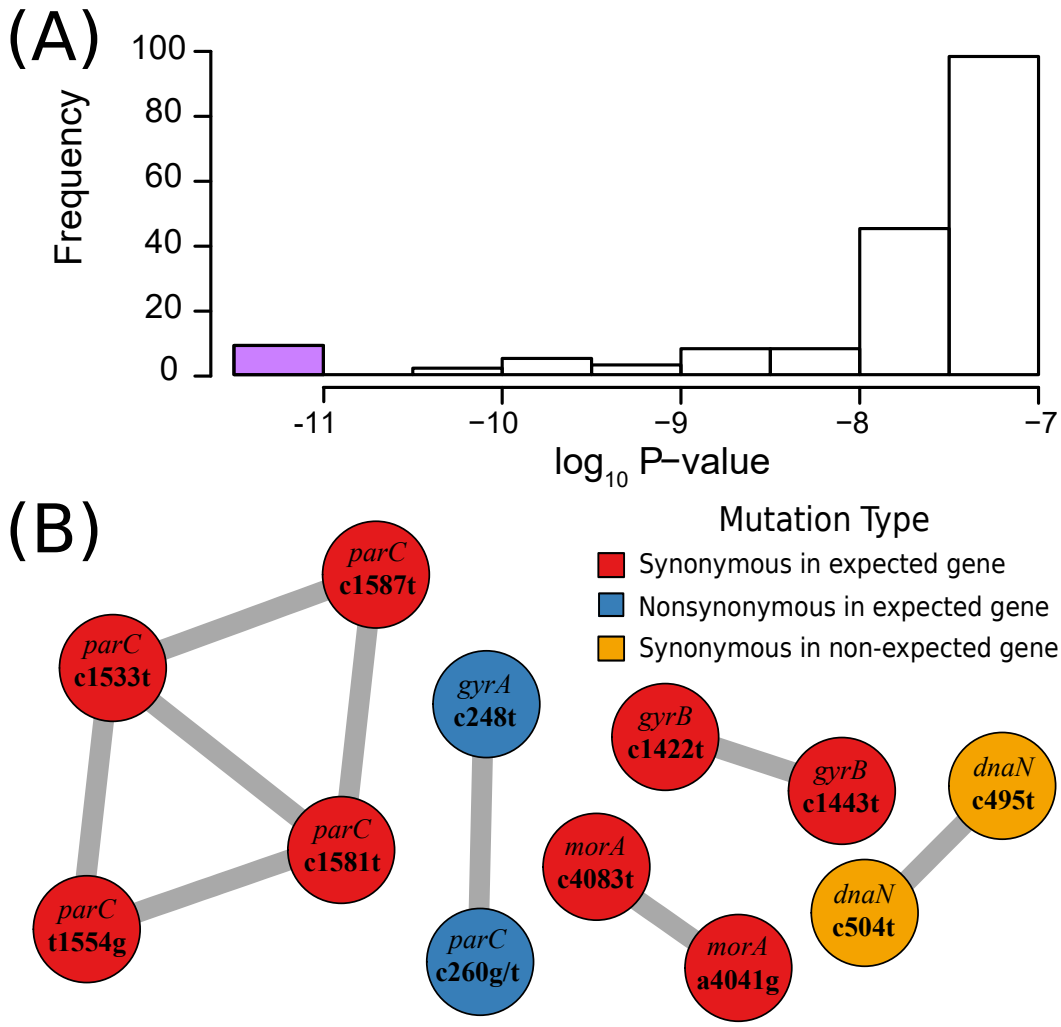

Figure 3.4.1: Correlated pairs of mutations with strongest evidence. (A) a histogram showing the distribution of significance for all correlated pairs with strong ($P < 10^{-7}$) evidence for correlated evolution. The purple bar highlights the strength of evidence for the pairs presented in B. (B) substitutions with the strongest evidence for correlated evolution ($P < 10^{-11}$). Circles represent a substitution in a gene (top text), at a particular nucleotide position (bottom text) on the coding strand. Colours show whether or not the substitution was synonymous, and whether or not it was found in one of the six genes expected to evolve in response to fluoroquinolone selection. Edges connect the predicted pairs of correlated substitutions.

3.4.2 Experimental validation of the nonsynonymous pair

To test if epistasis was driving the correlated evolution of these fluoroquinolone resistance substitutions in *gyrA-parC*, we used a modified allelic replacement protocol [Melnyk *et al.* 2017] to construct mutants bearing all possible combinations of *gyrA-parC* single and double mutations in two genetic backgrounds (PA01 and PA14) that lack the mutations of interest (Methods). We then quantified the MIC of a fluoroquinolone antibiotic (ciprofloxacin) for all genotypes and their competitive fitness against the ancestral (wild-type [WT]) genotypes in the absence of drug. We used the PA01 and PA14 genetic backgrounds because they are phylogenetically distinct [Dettman *et al.* 2013], representing two major clades [Kos *et al.* 2015].

Our results reveal that the pattern of changes in resistance for single and double mutants was similar for both genotypes (Fig. 3.4.2A,B). On its own, the *gyrA* mutation increases resistance, by $\sim$ 8-fold in PA14 and at least twice that in PA01, whereas the *parC* mutations on their own have no effect. In combination, however, the *gyrA* mutation together with either of the *parC* mutations confers between 128-256 fold increases in resistance, depending on genetic background. Support for this interpretation comes from a likelihood ratio test for significant fixed effects in a mixed effects model using biological replicates as a random effect. Results of the tests show that MIC depends on both the genetic background ($X^2 = 46.52$, $df = 1$, $P = 9.059 \times 10^{-12}$) and mutation ($X^2 = 813.95$, $df = 6$, $P < 2.2 \times 10^{-16}$), but also that the magnitude of the fold-increase in MIC was similar for PA01 and PA14 constructs (*i.e.* interaction term, $X^2 = 7.47$, $df = 5$, $P = 0.1878$). This analysis confirms the presence of strong positive epistasis for resistance between these *gyrA-parC* mutants, the magnitude of which is independent of the two genetic backgrounds tested.

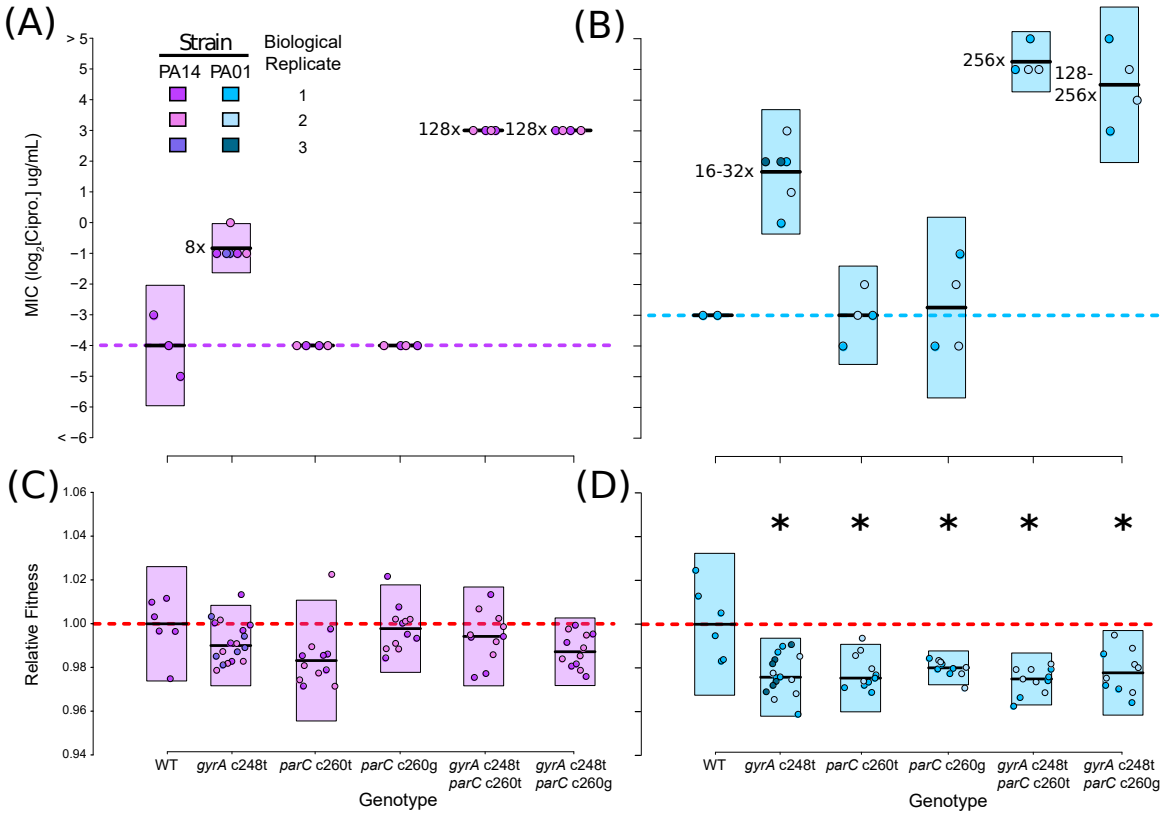

Figure 3.4.2: Empirical tests of epistasis for resistance and fitness. The mean and 95% confidence interval of measurements for each genotype is represented by the dark black lines and coloured rectangles respectively. The colour used reflects the genetic background of genotypes while the hue identifies the biological replicate. (A,B) results of the ciprofloxacin MIC assay of *P. aeruginosa* WT and mutant constructs. Dashed horizontal lines highlight the MIC value for WT strains. Numbers to the left of mean (black) lines represent the MIC fold increase, compared to WT, of mutants. (C,D) results of the competitive fitness assays for WT and mutant construct genotypes. The relative fitness of WT strains is highlighted by the horizontal dashed red line. Mutants with relative fitness significantly different from wild-type are denoted with an asterisk (see Table 3.4.2).

Notably, mutants carrying the *gyrA c248t* substitution had the expected 8-32 fold increase in ciprofloxacin resistance [Akasaka *et al.* 2001]. However, the *parC c260g/t* substitutions did not affect resistance on their own but did show a $\sim$ 16 fold additional increase for double mutants (Fig. 3.4.2 a,b; Table 3.4.1). This strong epistasis for resistance could explain why *gyrA c248t* and *parC c260g/t* double

mutants are common among fluoroquinolone resistant clinical isolates of *P. aeruginosa* [Akasaka *et al.* 2001; Kos *et al.* 2015] and is consistent with previous work showing that the fluoroquinolone resistance conferred by substitutions in *parC* depended upon mutations in *gyrA* [Moon *et al.* 2010]. We note that Akasaka *et al.* [2001] quantified the effect on MIC of both the *gyrA* and *parC* mutations by measuring decatenation activity of purified protein and found that both mutations would increase resistance. The difference between our lack of measured independent effect on resistance for the *parC c260g/t* mutation, and the effect estimated by Akasaka *et al.* [2001], suggests a possible Bateson-like molecular mechanism underlying the observed epistasis for resistance [Bateson 1911]. As ciprofloxacin acts by inhibiting both DNA gyrase and topoisomerase IV (the protein products of *gyrA* and *parC*, respectively), the changed decatenation activity of topoisomerase IV caused by the *parC* substitution may be masked if DNA replication is prevented by non-functional DNA gyrase. Since *gyrA* single mutants can grow under elevated fluoroquinolones levels, but *parC* single mutants cannot, it is reasonable to posit that the *gyrA* mutation compensates for the inhibition of *parC*, or that fluoroquinolones do not entirely inhibit *parC* protein activity. A study of the molecular basis for this epistasis for resistance may provide valuable insight in the development of next generation fluoroquinolones.

The results for fitness costs associated with resistance mutations tell quite a different story (Fig. 3.4.2C,D). We found no evidence for a cost of resistance associated with the *gyrA* and *parC* mutations, either singly or in combination, in the PA14 background, a result that has been observed previously [Melnyk *et al.* 2015]. By contrast, all the single and double mutations in PA01 lead to significant fitness costs, and the double mutant including the *parC c260g* mutation, exhibits

Table 3.4.1: Results from MIC assays of mutant constructs under ciprofloxacin. We measured MIC from four technical replicates for each of two independent constructs (biological replicates), except for *gyrA c248t* that had three independent constructs. The significance of differences in $\log_2$ MIC between wild-type and each mutant construct was assessed with the Dunn test and a Bonferonni correction; $P \leq 0.05$ shown in bold. Epistasis was measured with a multiplicative model, using the MIC determined from reads of the optical density (600 nm) after 24 hours growth, and measurement error was calculated using error propagation [Trindade *et al.* 2009]. There is evidence for epistasis when the absolute value of ϵ is greater than the error of our measures (in bold).

Background	Mutant		X^2	Z	P	ϵ	Error
PA14	*gyrA c248t*		6.480	-2.546	**0.005**		
	parC c260t		0.000	0.000	0.500		
	parC c260g		0.000	0.000	0.500		
	gyrA c248t	*parC c260t*	5.478	-2.341	**0.010**	**0.547**	0.382
	gyrA c248t	*parC c260g*	5.478	-2.341	**0.010**	**0.547**	0.382
PA01	*gyrA c248t*		4.253	-2.062	**0.020**		
	parC c260t		0.000	0.000	0.500		
	parC c260g		0.000	0.000	0.500		
	gyrA c248t	*parC c260t*	4.000	-2.000	**0.023**	**4.461**	2.051
	gyrA c248t	*parC c260g*	3.529	-1.879	**0.030**	**3.271**	3.138

significant positive epistasis (Table 3.4.2). In other words, the resistance mutations compensate for each other, leading to costs that are lower than expected from their additive effects. Our results support the idea that the cost of antibiotic resistance in bacteria can vary substantially across genetic backgrounds [Melnyk *et al.* 2015]. Combinations of mutations in *gyrA* and *parC* other than those tested here have been shown to be cost-free in *Streptococcus pneumoniae* [Gillespie *et al.* 2002], reinforcing the idea that the effect of a mutation, or combination of mutations, depends intimately on genetic background. Moreover, our observation of positive epistasis for resistance in both genetic backgrounds, but not consistently for fitness costs, indicates that the correlated evolution detected in *gyrA* and *parC* is likely driven by epistatic effects on antibiotic resistance, rather than competitive fitness in antibiotic-free environments.

Table 3.4.2: Results from competitive fitness assays of mutant constructs in permissive LB media. We measured fitness from six technical replicates for each of two independent constructs (biological replicates), except for $gyrA\,c248t$ which had three independent constructs. Relative fitness was calculated by dividing the fitness of each mutant construct by that of its wild-type (not shown here). The significance of differences in competitive fitness of wild-type and each mutant construct was assessed with the Dunn test and a Bonferonni correction; $P \leq 0.05$ shown in bold. Epistasis was measured with a multiplicative model and error was calculated using error propagation [Trindade *et al.* 2009]. There is evidence for epistasis when the absolute value of ϵ is greater than the error of our measures (in bold).

Background	Mutant		X^2	Z	P	ϵ	Error
PA14	$gyrA\,c248t$		1.778	1.333	0.456		
	$parC\,c260t$		3.509	1.873	0.153		
	$parC\,c260g$		0.561	0.749	1.000		
	$gyrA\,c248t$	$parC\,c260t$	1.010	1.005	0.787	0.016	0.024
	$gyrA\,c248t$	$parC\,c260g$	3.509	1.873	0.153	-0.003	0.021
PA01	$gyrA\,c248t$		8.266	2.875	**0.010**		
	$parC\,c260t$		7.364	2.714	**0.017**		
	$parC\,c260g$		8.576	2.929	**0.009**		
	$gyrA\,c248t$	$parC\,c260t$	11.000	3.317	**0.002**	0.019	0.021
	$gyrA\,c248t$	$parC\,c260g$	6.224	2.495	**0.032**	**0.022**	0.021

Six of the genes in our twelve gene-by-exome analysis were expected to evolve in response to fluoroquinolone selection. Yet, we only had sufficient evidence to support experimental validation of epistasis between a single pair of sites canonically associated with fluoroquinolone resistance mutations. Two questions remained: (i) how come AEGIS did not identify correlated evolution involving additional nonsynonymous resistance mutations, and (ii) what was/were the mechanism/-s underlying the highly significant correlation identified for the eight pairs of synonymous substitutions?

3.4.3 Resistance is strongly associated with only two substitutions in our alignment

The computational approach employed above identified only two nonsynonymous substitutions likely to be correlated during the evolution of resistance to fluoroquinolone antibiotics [Kos *et al.* 2015; Wibowo 2013]. Our lack of success in identifying more pairs of nonsynonymous substitutions could be because no other resistance mutations in our alignment evolved in a correlated manner, or because AEGIS is only able to detect instances of strong correlated evolution [Nshogozabahizi *et al.* 2017], or perhaps that only a few substitutions in our dataset are strongly associated with drug resistance. We therefore performed three additional analyses to uncover further substitutions associated with drug resistance in our dataset.

First, we quantified the importance of all polymorphic positions in the *P. aeruginosa* exome for predicting fluoroquinolone resistance [Long *et al.* 2019]. We did this by training a supervised machine-learning algorithm, based on adaptive boosting, which measured the strength of correlation between substitutions in our alignment and the resistance phenotype. This approach detected that the *gyrA 248* and *parC 260* nucleotide positions were the only two sites in our alignment that strongly predicted resistance (Fig. B.2.4). These two positions were also the only ones among the 100 most important associations that were located in genes expected to evolve in response to fluoroquinolone selection (Fig. B.2.4).

Second, we tested the association between fluoroquinolone resistance and nonsynonymous mutations at known fluoroquinolone resistance determining sites [Akasaka *et al.* 2001; Kos *et al.* 2015]. From our alignment, the *gyrA 248* (*T83I*) and *parC 260* (*S87L* and *S87W*) mutations were present in 41.4% and 32.9% of all

strains, respectively, and these mutations were correlated to the resistance phenotype (χ^2 tests: *gyrA 248*, $X^2 = 231.94$, $df = 1$, $P = 2.25 \times 10^{-52}$; *parC 260*, $X^2 = 177.21$, $df = 1$, $P = 1.98 \times 10^{-40}$, Table B.3.3). We note that another study has shown evidence for epistasis in the context of drug resistance when mutations in *parC* (*S87L* and *S87W*) co-occur with a mutation in *gyrA*, but at a different position [*S91F* instead of *T83I*; Schubert *et al.* 2019]. In contrast, nonsynonymous substitutions at additional amino acid sites known to confer fluoroquinolone resistance were identified in our alignment, but either did not correlate with resistance (at $P \leq 0.05$), or represented less than 10% of all strains included in our analysis (max. 8.74%, mean (sd) 3.54 (1.81)%; Table B.3.3). Previous work showed that the signal of correlated evolution (*i.e.* the sensitivity of the AEGIS algorithm) is maximised when roughly half the strains are mutants, and when these mutations are evenly distributed throughout the phylogeny [Nshogozabahizi *et al.* 2017], which is not the case here.

Third, we searched for loss of function mutations that are known to occur under fluoroquinolone selection, such as those affecting the *nfxB* or *orfN* genes [Wong *et al.* 2012]. There were no examples of premature stop codons within the sequences of these two genes in our alignment, although the presence of nonsynonymous loss of function mutations cannot be ruled out. Although it is in principle possible to resort to branch-site codon models to identify positions under selection in branches of interest [Yang and Nielsen 2002; Zhang *et al.* 2005], and hence detect sites potentially involved in drug resistance, our alignment contains too few taxa and too little evolutionary depth for proper statistical analysis [Arenas 2015]. Altogether, these three lines of evidence suggest that among the subset of known fluoroquinolone resistance mutations, only the *gyrA c248t* and *parC c260g/t* substitutions are present in enough strains to have been detected as correlated and

show a significant association with fluoroquinolone resistance in our alignment.

3.4.4 Synonymous substitutions: the role of multiple drivers

AEGIS provided strong evidence for correlated evolution among synonymous mutations. This result is surprising because synonymous mutations are often assumed to evolve neutrally, and so we would not *a priori* expect them to exhibit strong correlated evolution unless they were under strong selection or were physically linked. This result could be due to strong phylogenetic uncertainty, which can affect comparative methods [Guigueno *et al.* 2019; Shoji *et al.* 2016]. To account for this we could rerun the analyses on bootstrapped trees [Nshogozabahizi *et al.* 2017] but this is a taxing solution as it would be computationally quite demanding, especially here, possessing statistical properties that would deserve scrutiny at a level beyond the scope of the present work. However, the phylogenetic tree that we reconstructed has internal nodes that are fairly well supported (Fig. B.2.2), making it unlikely that the signal of correlated evolution that we detect between synonymous mutations is due to phylogenetic uncertainty.

On the other hand, it has been known for some time that synonymous sites can experience purifying selection [Lawrie *et al.* 2013; Zhou *et al.* 2010], and a number of recent reports show that they can sometimes contribute to adaptation as well [Agashe *et al.* 2016; Bailey *et al.* 2014]. However, we found no signal of epistasis-driven selection acting on the most strongly correlated pairs of synonymous sites identified in our study (Fig. 3.4.1). Estimates of the change in stability of mRNA transcripts and index of translation efficiency, which are proxies for translation rate [Kudla *et al.* 2009] and efficiency [Xia 2014], respectively, relative to the wild type are shown in Fig. 3.4.3 and Table B.3.3,B.3.4. While two of the

synonymous mutations in *parC* ($c1581t$ and $c1587t$) could produce more stable and abundant mRNA transcripts, the only evidence for epistasis in either metric we could identify was for translation efficiency in the mutations in *morA*, and the direction of this effect is opposite to what would be expected if it were to result in positive selection.

By contrast, two lines of evidence point to physical linkage between sites as the proximate cause of correlation between synonymous mutations. First, the proportion of synonymous pairs occurring within the same gene was higher than expected by chance among all observed (synonymous and nonsynonymous) pairs ($X^2 = 7246.037$, $df = 2$, $P \leq 2.16 \times 10^{-16}$; see Table B.3.6). Second, the physical distance between pairs of correlated synonymous substitutions, as mapped on the circular genome of PA14, was negatively associated with the strength of correlation (for pairs with $P \leq 10^{-4}$; $t = -5.1784$, $df = 7,528$, $P = 2.29 \times 10^{-7}$ Fig. B.2.5) , though the correlation was quite poor ($R^2 = 0.03$). Physical distance remains significantly associated to strength of correlation even when the 8 most significantly correlated pairs are removed substitutions ($t = -6.922$, $df = 7,520$, $P \approx 4.8 \times 10^{-12}$, $R^2 = 0.006$).

The mechanism responsible for the close physical linkage between synonymous mutations is less obvious. One possibility is that both hitchhiked to high frequency alongside a third, non-synonymous mutation in the same gene that was under strong positive selection. We found a nonsynonymous mutation at nucleotide position 1374 of *gyrB*, which is within the gene's quinolone-resistance determining region [Bruchmann *et al.* 2013], that was always present alongside the synonymous mutations. Moreover 97% of the strains that have substituted the pair of synonymous mutations in *morA* also fixed nonsynonymous mutations at other

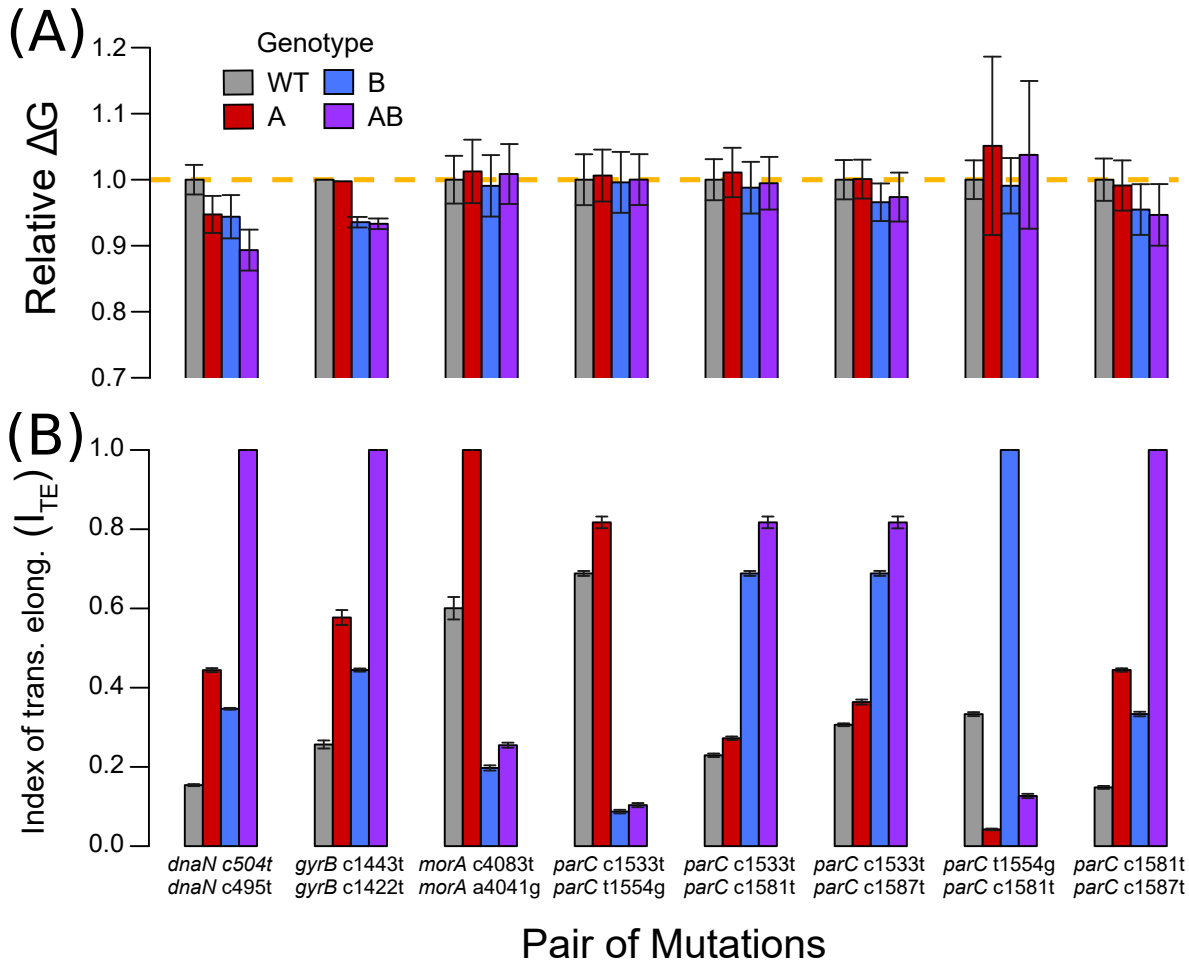

Figure 3.4.3: Computationally predicted biological effects of the strongly correlated pairs of synonymous substitutions. Each set of bars is for a pair of substitutions, where the first bar (grey) is the wild-type value (inferred from PA14 reference genome), the second bar (red) is a mutant carrying the first substitution (A) listed below the bars while the third bar (blue) represents a mutant carrying the second listed substitution (B), and the fourth bar (purple) represents a double mutant. Panel A estimates of the relative change of mRNA folding free energy (ΔG) are relative to the WT value (gold dashed line) and error bars represent the 95% confidence interval. Panel B measures of the mean index of translation elongation (I_{TE}, B) with error bars presenting standard deviation among measures estimated from double mutant strains present in our data set.

positions along the gene that could have functional effects on regulating or sensing internal oxygen levels (positions 292, 293, 306) [Taylor and Zhulin 1999] or signal transduction (position 1356) [The UniProt Consortium 2017]. While these results point to hitchhiking as a plausible explanation for the correlation between sites, we

failed to detect additional functional links with nonsynonymous mutations in the other genes. Moreover, AEGIS did not detect correlations between any of the correlated synonymous mutations and additional nonsynonymous mutations, even at $P \leq 10^{-4}$ (Fig. B.2.1), suggesting that hitchhiking is not a general explanation for the correlated evolution of synonymous mutations in this data set. It should further be noted for the correlated pairs of substitutions found in *gyrB* and *morA*, that unless the pair of mutations repeatedly arose in the context of a third substitutions under selection, hitchhiking would not have produced a signal of correlated evolution.

A second possible mechanism we considered is horizontal gene transfer (HGT) and/or incomplete lineage sorting. Comparing the similarity of the species tree and the gene trees for each of the four genes with a strongly correlated pair of synonymous mutations reveals little correspondence between the two for all but *parC* (Table 3.4.3), suggesting widespread recombination in our strain collection. To further assess the prevalence of recombination, we included tRNA genes located in close proximity of each of our four focal genes in the PA14 assembly. The rationale for this analysis was that both tRNA genes and three of our four focal genes, being involved in information processing, should interact with a large number of other gene products (transcripts and/or proteins), so that a successful HGT of such genes should involve the co-transfer of most of their interacting partners, an unlikely event [Aris-Brosou 2005; Jain *et al.* 1999]. However, we found that all tRNA gene trees differed from each other and the species tree (Table 3.4.3) suggesting that recombination (and thus HGT and/or incomplete lineage sorting) may be widespread in our strain collection. Widespread HGT is consistent with the findings of Kos *et al.* [2015], who found that the *parE* and β-lactamase genes were

Table 3.4.3: Approximately unbiased (AU) tests of significant differences among phylogenetic trees. We built each tree using the multiple sequence alignment for one of the four focal genes that contain the most significantly correlated synonymous substitutions (*dnaN*, *gyrB*, *morA*, *parC*) and four tRNA genes unlikely, according to the complexity hypothesis [Aris-Brosou 2005; Jain *et al.* 1999] to undergo HGT (*arginyl tRNA synthetase, glycyl tRNA synthetase subunit beta, isoleucyl tRNA synthetase, methionyl tRNA formyltransferase*). We tested for significant differences among all the gene trees and the species tree used in the analysis of correlated evolution. Values in each cell represent the probability, calculated via the AU test, that a tree (column names) describes evolution observed in an alignment (row names). In bold are the "best" tree(s) (*i.e.* insignificantly different trees).

Focal Gene	***dnaN***	***gyrB***	***morA***	***parC***	**arginyl tRNA synthetase**	**glycyl tRNA synthetase subunit beta**	**isoleucyl tRNA synthetase**	**methionyl tRNA formyltransferase**	**Species Tree**
dnaN	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
gyrB	0.000	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
morA	0.000	0.001	**1.000**	0.002	0.001	0.000	0.000	0.000	0.001
parC	0.000	0.000	0.000	**0.423**	0.000	0.000	0.000	0.000	**0.577**
arginyl tRNA synthetase	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000
glycyl tRNA synthetase subunit beta	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
isoleucyl tRNA synthetase	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
methionyl tRNA formyltransferase	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000

horizontally transferred in a number of the strains included in our exome alignment.

Importantly, both hitchhiking and HGT represent population-level processes that can lead to a pattern of physical linkage between synonymous sites. However, neither mechanism explains how such strong intra-gene correlations arose in the first place. Functional constraints, for example via intra-molecular pairing that forms strong hairpins in mRNA, and locally-biased mutation that depends on nucleotide context, have been proposed as potential mechanisms [Tsunoyama *et al.* 2001]. A full analysis of the contribution of these processes to the patterns we observe here, while interesting, is beyond the scope of this work, however.

Altogether, these results suggest that the close physical linkage between pairs of synonymous mutations is due in part to widespread HGT, except in the *parC* gene. Without any evidence that either HGT or hitchhiking with a nonsynonymous mutation are driving the correlated evolution of the four synonymous mutations in *parC*, which form a network of five correlated pairs (Fig. 3.4.1), these mutations could in principle have evolved repeatedly as a driverless cohort - that is, mutations that self-assemble by chance, and drift to fixation via linkage [Buskirk *et al.* 2017]. However, both the species tree (results of AEGIS analysis) and the gene trees (Fig. B.2.3) show that these four synonymous *parC* substitutions have evolved independently numerous times. It is therefore unlikely that a cohort of four substitutions repeatedly self-assembled by chance. Whether these mutations co-occur for selective reasons related to the ecology of the CF lung [Poole 2005; Smith *et al.* 2006; Sriramulu *et al.* 2005] and antibiotic treatment [Kos *et al.* 2015] or other processes such as functional constraints associated with transcription and translation or mutational biases remains unclear.

3.5 Conclusions

Our results show many instances of correlated evolution ($\sim 127,000$ interactions at $P \leq 0.01$ from $\sim 10,220,000$ comparisons tested - *i.e.* $\sim 1.2\%$). In the context of antibiotic resistance in *P. aeruginosa*, moreover, some instances of correlated evolution can be underlain by, among other processes, epistasis. This result is consistent with experimental work showing that multiple mutations can contribute to the evolution of antibiotic resistance [Hall and MacLean 2011; MacLean *et al.* 2010; Trindade *et al.* 2009], both within and between genes [Lehner 2011], although typically the number of mutations identified remains quite small. Our results suggest, unsurprisingly perhaps, that a wide range of mutations can evolve in a correlated manner in the context of antibiotic selection – but also that only a limited number of these mutations are easily interpreted in the context of drug resistance (here, those found in *gyrA* and *parC*). Furthermore, our results lend support to the idea that high-throughput techniques for identifying correlated evolution *can* lead to identifying epistasis. Similar interpretations have been made previously, for instance in the case of the evolution of RNA genes where base pairing is critical [Dutheil *et al.* 2010], or in the context of genome-wide association studies, where the detection of interacting residues is used as a first screen for epistasis, before fitting the linear models traditionally employed to identify genetic determinants of drug resistance [Schubert *et al.* 2019]. This latter approach, which notably also revealed epistatic interactions between *gyrA* and *parC*, requires that MIC values of all the genomes included in the analysis be known – data that was not available to us. In spite of this, both approaches should be highly complementary.

However correlated evolution is necessary but not sufficient evidence for epistasis [Nshogozabahizi *et al.* 2017], as it can also arise through other mechanisms such as

physical linkage. Computational approaches like ours detect relationships between phenotype and genomic substitutions though these associations are not evidence alone for a causal relationship. Indeed, we have observed a strong signal that physical linkage has driven the correlated evolution between pairs of synonymous substitutions in our data set, perhaps linked to the wholesale transfer of genes during recombination. Despite recent reports that synonymous mutations may not, in fact, be silent [Agashe *et al.* 2016; Bailey *et al.* 2014; Chamary and Hurst 2005; Cuevas *et al.* 2012; Duan *et al.* 2003; Fragata *et al.* 2018; Plotkin and Kudla 2011], we have little evidence that synonymous substitutions, either on their own or in combination, impact fitness in our data. The fact that many of the most strongly correlated synonymous substitutions we identified here are GC→AT mutations (Fig. 3.4.3) reflects the pronounced GC bias of *P. aeruginosa* and reinforces the notion that these pairs of synonymous mutations have not been selected. It is notable that another computational study found evidence of physical linkage underlying correlated evolution between nonsynonymous but not synonymous substitutions and so attributed their co-occurrence to functional constraints [Callahan *et al.* 2011]. While this does not seem to be the case for the synonymous mutations in our analysis, together these results suggest that physical linkage may often be important in generating signals of correlated evolution. If so, this result suggests using caution when interpreting the results of computational studies that identify strong correlated evolution among traits or sites within a genome.

While computational methods provide us with a high-throughput means of determining candidates for epistasis, we often lack a sufficient understanding of epistasis at the molecular level to have confidence in this interpretation in the absence of additional evidence. Our work represents a first step in this direction: by using a computational approach to predict candidates for correlated evolution, with

follow-up analyses suggesting that only a subset of these candidates are actually under epistasis, we were able to provide direct experimental evidence that epistasis underlies at least some instances of correlated evolution. However, this interpretation relies on our ability to conduct experimental validations under selective conditions that we presume to closely resemble those in which the sites evolved. This may not always be possible. Indeed, most of the strongest instances of correlated evolution in our data set appear to be unconnected to fitness or epistasis in an antibiotic selective context. As densely sampled population genomic data are not always available (such as the > 3000 genomes in Skwark *et al.* [2017]), we urge caution in interpreting instances of correlation between sites from high throughput computational analyses as solely due to epistasis.

Data Accessibility

All R scripts and complete results of the computational analysis of correlated evolution can be found at `github.com/JDench/sigResults_AEGIS_inVivo`.

Chapter 4

The SHAPE of logistic growth shows that serial passaging biases fixation probability

This chapter is published in:

Dench, J., (Pre-print) "The SHAPE of logistic growth shows that serial passaging reduces repeatability" Biorxiv,

https://www.biorxiv.org/content/10.1101/392944v3

4.1 Abstract

The forward time simulation tool rSHAPE (R-package for Simulated Haploid Asexual Population Evolution) was designed to complement the theoretical and empirical study of evolution. Included with rSHAPE are functions to programmatically build, run, and initially process results of an evolutionary experiment defined by the range of experimental conditions. As experimental

evolution often studies both the emergence and fate of *de novo* mutants, I validated rSHAPE by confirming its ability to replicate seminal theoretical expectations concerning changes in fitness through time and the fixation probability of mutants. As an example of how rSHAPE can support both theoretical and empirical research, I applied rSHAPE to study how the laboratory protocol of serial passaging, common in microbial experimental evolution, affects the fixation probability of *de novo* mutants. Unlike related theoretical work which modelled growth as effectively exponential [Wahl *et al.* 2002], this study considered populations experiencing logistic growth which is common to microbes undergoing serial passaging. In contrast to exponential growth, when a population undergoes logistic growth the probability of a mutant arising and eventually fixing depends upon when a mutant arises during a growth phase. Users can download software and documentation for rSHAPE through CRAN at `https://cran.r-project.org/web/packages/rSHAPE/index.html`, or via GitHub at `https://www.github.com/Jdench/SHAPE_library/`.

4.2 Introduction

While certain independent evolutionary drivers have been elucidated, such as selective context or large effect beneficial mutations, researchers are increasingly interested in studying the fate of populations, and the mutations they carry, in a more wholistic context [Losos *et al.* 2013]. Though experimental evolution can be paired with next generation sequencing technologies in order to track the emergence and fate of mutations [Goodwin *et al.* 2016], these *in vivo/in vitro* approaches are limited by availability of physical or financial resources. Even when experimental design is optimised, it is still not commonly feasible for researchers to fully sequence

whole genomes, of whole populations, at multiple time intervals, for each replicate of an experiment. Such an approach is not fanciful idealism as the interaction among mutations can affect evolutionary outcome [Domingo *et al.* 2019; Phillips 2008; Weinreich *et al.* 2006]. Furthermore, at least one well funded research group has sequenced whole replicate populations, at multiple time points, shedding new light on evolutionary dynamics [Good *et al.* 2017]. While development of evolutionary theory with, and analytical analysis of, existing data are much less costly approaches, we need to combine both evolutionary theory and empirical study to improve our understanding of evolution in increasingly complex systems [de Visser and Krug 2014]. Simulations, or *in silico* evolution, can leverage theoretical models based on empirical evidence to support both theoretical and empirical study of evolutionary biology, while being limited only by the available software and computational resources.

Simulation software complement theoretical and empirical study by allowing the comparison of empirical observation to the dynamics expected from various theoretical models and combinations of evolutionary parameters. There exist simulation tools to track changes in population demographics [Guillaume and Rougemont 2006] or specific mutations [Arnold *et al.* 2018; Dalquen *et al.* 2012; Lambert *et al.* 2008; Zanini and Neher 2012], though most of these implement a single theoretical model and provide only limited output. The AVIDA[Beckmann *et al.* 2010] and AEVOL[Knibbe *et al.* 2007] tools are agent-based models for studying genetic changes but their evolutionary systems are genomic abstractions that are practical for their implementation but make it unclear to what extent their results apply to living systems. A powerful tool for simulating microbial experimental evolution is the Haploid Evolutionary Constructor (HEC) [Lashin *et al.* 2014], though it simplifies evolution of the genome by mapping each locus to a

single quantitative trait value and assumes mutations do not interact. Knowing no single framework to simulate experimental evolution while tracking detailed changes in both population demographics and all mutations in large genomes (*i.e.* millions of mutational sites), I found the existing software impractical, or not applicable, for comparing evolution across different models of growth and fitness landscape scenarios. To address this, I developed the R-package for SHAPE (Simulated Haploid Asexual Population Evolution - rSHAPE).

The simulation tool rSHAPE was designed to simulate microbial experimental evolution in a framework that can readily be expanded to different models, and evolutionary scenarios, thanks to a modular programming design. I chose to simulate microbial evolution experiments as they are commonly used for the manipulative study of evolution. These experiments often study how the selective context influences the emergence and persistence of *de novo* mutants arising from an initially isogenic population, but are also used to estimate underlying evolutionary parameters such as the fitness landscape of selective contexts. The design of many microbial evolution experiments requires the practice of serial passaging, a protocol in which a small fraction of stationary phase communities is transferred to fresh growth media to extend the evolutionary time of experiments. When communities are well mixed prior to serial transfer, loss of mutants is proportional to their representation in the community. While serial transfer is expected to result in the loss of rare mutants, theory suggests that the probability of a mutant arising during exponential growth, and subsequently surviving many rounds of serial transfer, is roughly uniform [Wahl *et al.* 2002]. For experimental microbial populations, continuous cell division (similar to exponential growth conditions) will only generally be achieved in chemostats, but chemostat setups may favour biofilm phenotypes [Larsen and Dimmick 1964] because their constant growth vessel can

add unintended niche to the selective context. Further, experiments may avoid chemostats due to their expense and the difficulty in both setup and maintenance. Thus, most experimental microbial communities are not grown in chemostats and so undergo logistic growth between serial transfers. It remains unclear how serial transfer may affect the joint probability of mutants arising and subsequently fixing in evolving populations.

In this paper I begin by explaining the framework of rSHAPE. I then report on rSHAPE's ability to reproduce expected trends in fitness through time as well as the probability of a *de novo* mutant fixing under the conditions of several theoretical models. I then provide one example of how rSHAPE can be used to study evolutionary scenarios outside the scope of existing analytical models. I used rSHAPE to test if serial passaging influences the fixation probability of a *de novo* mutant arising in a population that undergoes logistic growth.

4.3 Design and implementation

4.3.1 Simulating evolution

The software rSHAPE is a forward time, discrete step, *in silico* experimental evolution system that tracks changes in population demographics. Populations are composed of haploid genotypes that are tracked by which of their L binary state sites (*i.e.* genome of length L) carry mutations and where a value of 0 represents the unmutated state and 1 depicts mutation. Note that while it may be practical to consider each site as a nucleotide base, each site could equally be considered as any mutable element which influences genotype fitness (*e.g.* a codon or gene). To define the experimental conditions, rSHAPE includes the function *defineSHAPE* which

allows users to directly set 30 experimental parameters (Appendix C.1, Table C.4.1). Once the parameters are defined, the experiment is started by calling the function *runSHAPE* which will sequentially simulate the evolution of n identical but independent starting populations. In each time step, the population may experience stochastic loss events (*e.g.* a population bottleneck) and each individual within the population may die, reproduce, and/or mutate. Though it may be practical to think of a time step as a biological generation in which each individual of average fitness will produce r offspring (where r is the growth rate), this need not be the case (*e.g.* users can set the average probability of birth such that $P_{birth} \in [0, 1]$). Regardless, after each time step rSHAPE records the evolving populations' demographics as the current number of individuals for each genotype and their numbers of deaths, births, as well as mutants generated including the "parent" genotype from which they arose. By tracking a mutant's "parent" genotype, the entire evolutionary history can be reconstructed permitting detailed evolutionary analyses. To further support analyses, rSHAPE tracks the full fitness landscape of all explored mutational space. When a "parent" individual experiences a mutation, the explored mutational space represents all possible mutant offspring from which the offspring is eventually selected. At the onset of an rSHAPE experiment, the evolving population will consist entirely of wild-type (0 state for all genomic positions) individuals. In its current form, rSHAPE resembles many microbial evolution experiments in that populations are finite and evolve in a single, well mixed (*i.e.* homogeneous) environment. While these are conditions at the launch of rSHAPE, the starting population, evolutionary environment, as well as other implemented evolutionary models (*e.g.* growth forms, fitness landscapes) could readily be changed due rSHAPE's modular programming design. This modular design is achieved through rSHAPE being an iterative implementation of functions

to calculate the outcome of stochastic events, deaths, births, and mutation (Fig. 4.3.1). For a detailed description of the iterative functions called during a run of rSHAPE, please refer to Appendix C.1.

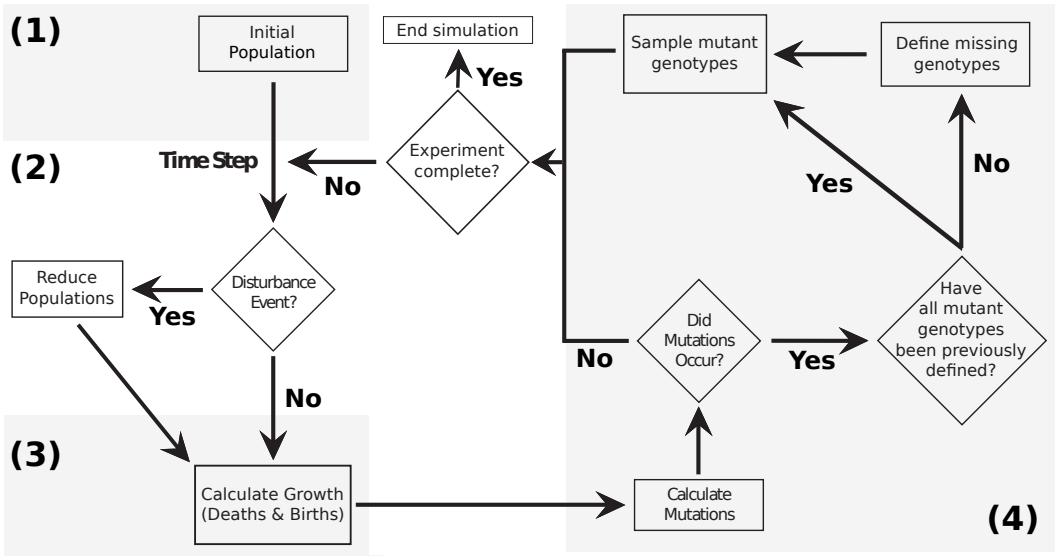

Figure 4.3.1: rSHAPE's flowchart. After an experiment begins, rSHAPE will cycle through discrete time steps which calculate 2) disturbance events, 3) growth (*i.e.* deaths and births), and 4) mutations. Experiments will continue for the pre-determined number of time steps or until there are no individuals living in the population. At the end of the experiment, rSHAPE logs will report the changes occurring in each step and the fitness value of all mutants that could have been produced. The record of mutant fitness values is referred to each time mutation occurs and existing genotype values are retrieved while undefined mutant fitness values are computed and added to the record.

4.3.2 Running experiments

While populations simulated with rSHAPE are finite, I have set no upper bound on either their size or their genome length. It is possible to define an evolving community of individuals with genomes containing trillions of positions and which grows exponentially without death. It is not practicable - if at all possible - to store in active memory the full demographic changes and genotype fitness values for realistic microbial populations containing billions of individuals with genomes

containing millions of sites. To mitigate this challenge, rSHAPE exports the full demographic breakdown, as well as the explored fitness landscape of all genotypes, to SQL databases which act as both the experimental record and a reference when needed during the run. As a result, *in silico* evolution with rSHAPE is less limited by RAM than by disk space and the number of available processors to run independent replicates and parameterisations.

To facilitate rSHAPE's use for simulating a breadth of experimental conditions, I have included the *shapeExperiment* function that uses template files found here: https://github.com/JDench/SHAPE_library/tree/master/SHAPE_templates. With *shapeExperiment*, rSHAPE builds parameter combinations from the range of values (for each parameter) recorded in the control files. Using these parameter combinations, rSHAPE creates an experimental folder within which all files necessary to run the experiment will be written including automated scripts for running the whole experiment. Before ending each independently replicated simulation, rSHAPE writes a summary log to facilitate analysis. Further, once all simulations in an experiment are complete, users can use the *summariseExperiment* function to combine information from all summary logs and calculate additional summary statistics. Experiments built with *shapeExperiment* will have a script to run *summariseExperiment*. At present, *summariseExperiment* will create output to facilitate comparison of population dynamics, evolutionary trajectories, and the repeatability of evolution. Future improvements will expand the breadth of summary analytics and allow users to more easily select from among the set of possible summary output.

4.4 Results and discussion

4.4.1 Validating implementation

To validate the implementation of rSHAPE, I assessed its ability to replicate evolutionary trends expected under various parameterisations of microbial experimental evolution. Since the goal of many microbial evolution experiments is to analyse factors that affect the emergence and maintenance of mutations, I first tested if populations evolving in rSHAPE showed expected fitness trends, and second that the evolutionary dynamics simulated with rSHAPE can replicate the theoretical fixation probability of *de novo* mutants under different growth conditions.

To test rSHAPE's ability to replicate standard evolutionary trends, I initially tested that the fitness of evolving populations increase through time and that population bottlenecks alter the rate of fitness increase [Wahl *et al.* 2002] (Fig. 4.4.1). For this, I simulated large communities (carrying capacity $K = 1{,}000{,}000$) of individuals with 100 genomic positions ($L = 100$), evolving for $T = 10{,}000$ generations, for each of three different regular bottleneck sizes (*i.e.* disturbance event, dilution factor; $D \in \{10, 100, 1000\}$), where mutant genotype fitness was calculated with either an additive model (a smooth fitness landscape defined by the sum of independent mutational fitness contributions - *i.e.* no epistasis; no interaction between mutations) or a House of Cards (HoC) fitness landscape [Kauffman and Levin 1987; Kingman 1978] (the fitness of every genotype is a random value independently drawn from the same distribution - *i.e.* a maximally rugged landscape). The independent effect of all mutations was assumed to be beneficial and so based on extreme value theory [Gillespie 1984], the random component of all genotype fitness calculations was drawn from an exponential distribution (with rate parameter $\lambda = 100$). For statistical comparison, I replicated

each parameter combination 500 times by initialising five independent fitness landscapes within which evolution was replicated 100 times. For the additive fitness landscape models, I expected fitness to increase until all mutations had fixed since all mutational effects were beneficial. However, each genotype simulated with a HoC fitness landscape is equally likely to be the most fit genotype (*i.e.* global optimum) and many genotypes will be the highest among their nearest mutant "neighbours" (*i.e.* local optimum) [Kauffman and Levin 1987], so that fitness increase should generally stop long before all mutations fix. From analysis of the evolutionary trends in these first simulations, I found for the additive fitness landscape models that both fitness and the number of mutations increased throughout because not all L (100) mutations had yet fixed, whereas in the HoC models evolution stopped after only a few mutations had fixed and a local optima had been found (Fig. 4.4.2).

As my second assessment, I tested if rSHAPE could replicate seminal theoretical works related to the fixation probability of *de novo* mutants. The theoretical works compared herein all build upon the mathematical framework of Haldane [1927] who studied fixation probability in the simplest scenario of a constant size population. For all but the simplest scenario studied by Haldane [1927], I compared rSHAPE to the published approximate solutions (requiring simplifying assumptions) because the theoretical models of Ewens [1967]; Otto and Whitlock [1997]; Wahl and Gerrish [2001] would not otherwise have closed form solutions. All approximate solutions included at least the assumptions that selection coefficients (s) were "small" and that effective population sizes (N_e) were "large" (no explicit values/formula were published for either). When comparing rSHAPE to these theoretical works, I simulated large populations with not less than one million individuals – a value that is within the range of population sizes considered by the works compared here.

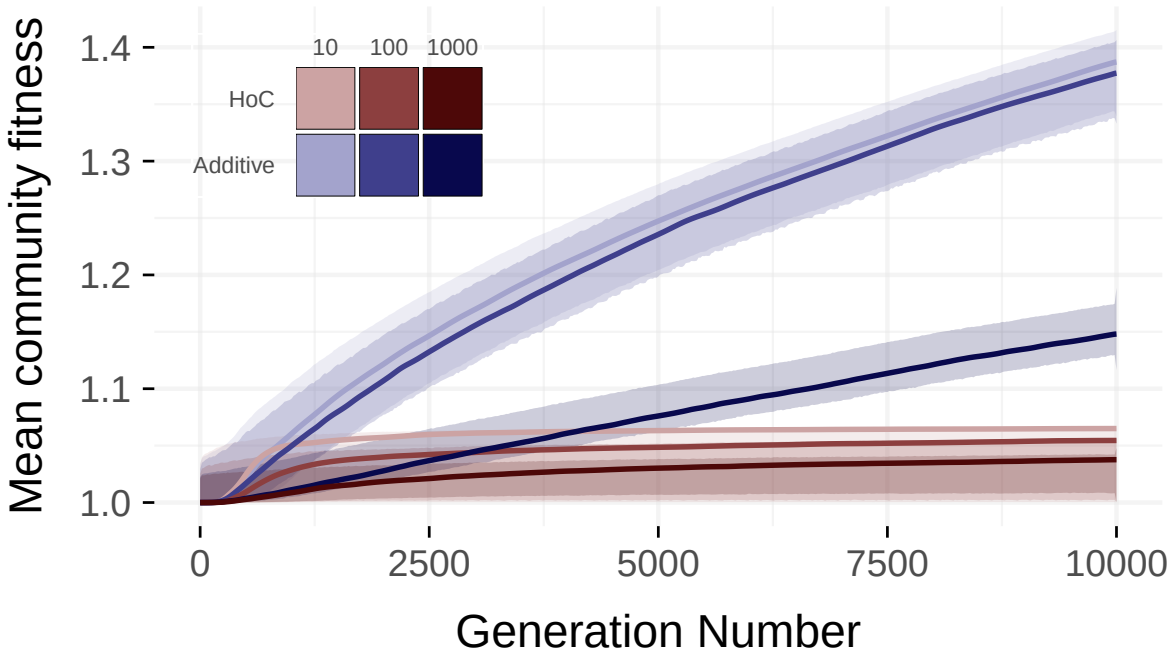

Figure 4.4.1: Community fitness through time simulated with rSHAPE. Colours represent results from different fitness landscape models (red - additive, blue - House of Cards, HoC) and shading reflects intensity of population bottlenecks. Solid lines represent the trend in mean community fitness across all replicates while the surrounding shaded polygon represent the absolute minimum and maximum values observed.

Since the exact fixation probability of Haldane's model can be calculated, and because Haldane's simplifying $2s$ approximation is part of all other theoretical works studied, I compared these two values to benchmark what may be considered a "small" selection coefficient. From this, I find that the approximate solutions may overestimate fixation probability by at least 5% once $s \geq 0.03$ (Fig. 4.4.3B).

As I compared rSHAPE to approximate solutions, I expected it to effectively replicate theoretical fixation probabilities when model assumptions were not violated. From tests, I found that rSHAPE replicated the predicted fixation probability of *de novo* mutants arising in communities of constant size [Haldane 1927] (Fig. 4.4.3, and Appendix C.2.1: "Haldane's $2s$") as well as for those experiencing logistic growth [Ewens 1967; Otto and Whitlock 1997] (Fig. 4.4.4, and

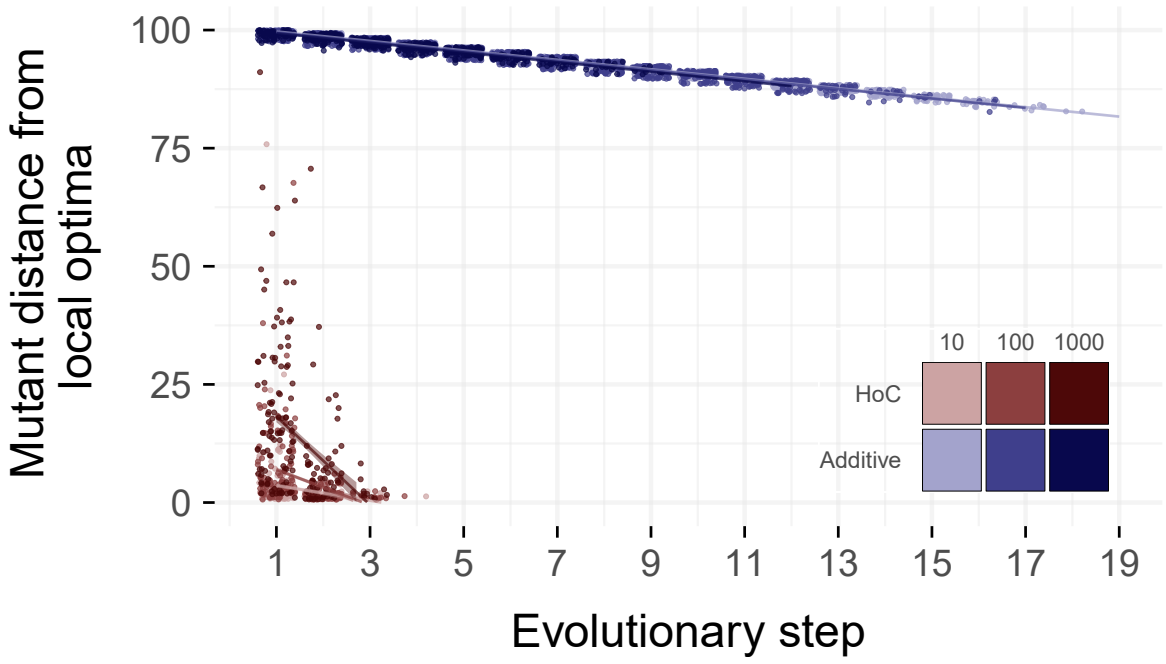

Figure 4.4.2: Distance of successful mutants from local optima as a function of ordered evolutionary step. Colours represent results from different fitness landscape models (red - additive, blue - House of Cards, HoC) and shading reflects intensity of regular population disturbance events (*i.e.* bottlenecks). Solid lines represent the average trend of mutational distance between "next evolutionary step" mutants and a local fitness optimum. Mutational distance represents the number of mutations separating two genotypes.

Appendix C.2.2: "Logistic Growth"). Under parameterisations which violated the underlying assumptions (*i.e.* high growth rate and/or large selection coefficient), rSHAPE underestimates fixation probability by an amount proportional to the difference observed when benchmarking "small" s (Figs. 4.4.3B and 4.4.4B). Thus, I have found that rSHAPE can replicate theoretically expected evolutionary dynamics under proper parameterisations. Further, from testing parameterisations that violated model assumptions, I found that rSHAPE predicts values that deviate from approximate solutions by an amount similar to the expected difference (from benchmarking "small" s) between the exact and approximate solutions of the underlying model. This suggests that rSHAPE provides a means to test evolutionary

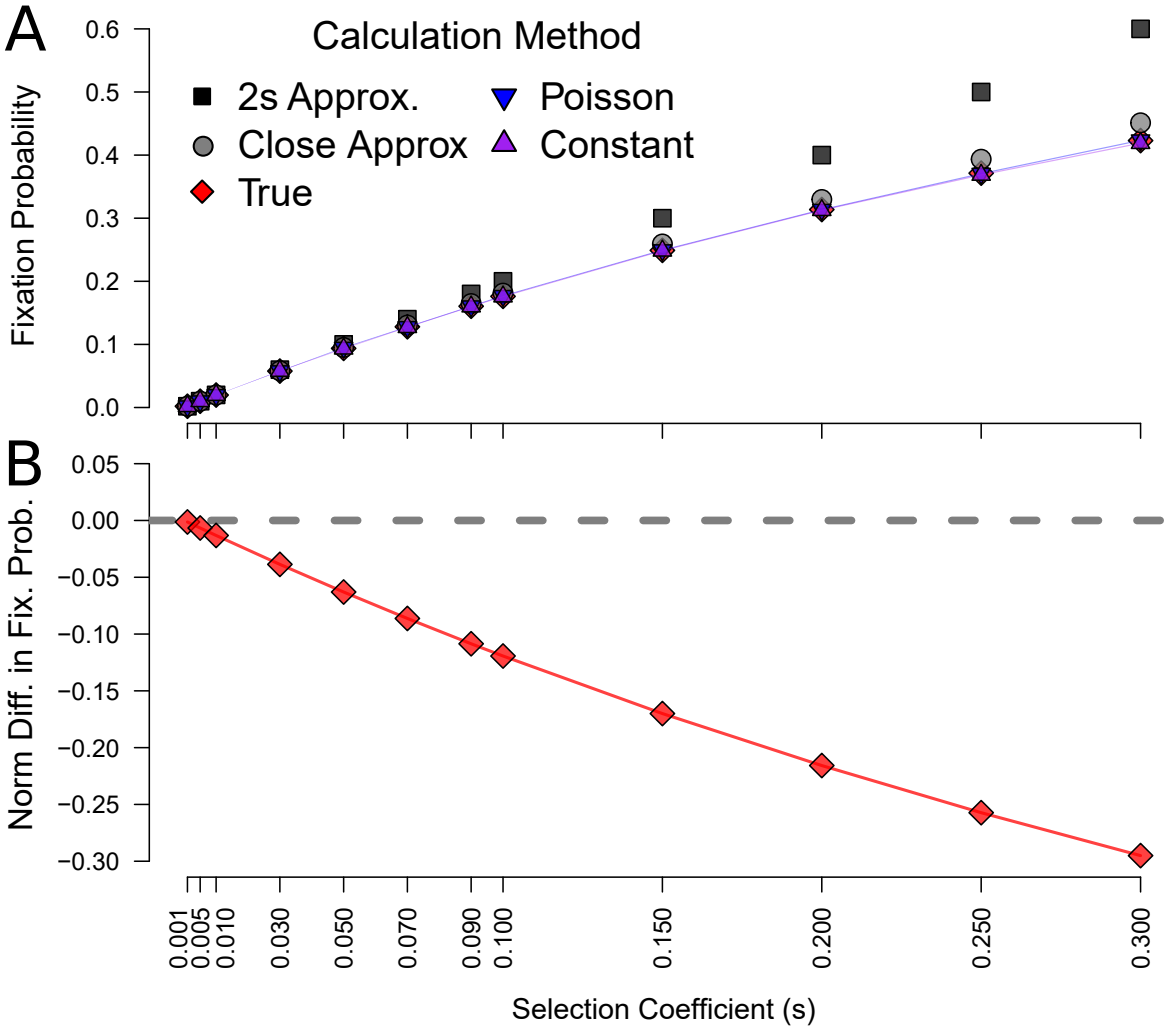

Figure 4.4.3: Theoretical and simulated probability of a single mutant fixing in populations of constant size. A shows the fixation probability, dependent on the selection coefficient s, where colouring of points distinguish between the theoretical calculations or simulated growth model. Theoretical calculations: black squares - *Haldane's 2s* approximation, grey circles - the closer approximation $1 - e^{-2s}$; rSHAPE growth models: blue triangles - *Poisson*, purple triangles - *Constant* and; red diamonds represent the true un-approximated fixation probability. Note, the red diamonds are not always evident as they are overlain by the fixation probabilities estimated with rSHAPE. B, the normalised difference between the true fixation probability and the analytical approximation of $2s$. Negative values reflect that the true fixation probability is lower than the analytical approximation of $2s$.

scenarios under conditions not supported by at least some analytical models.

All simulation scripts, and resulting summary files, can be found at github.com/Jdench/rSHAPE_validation. The full results for this section are

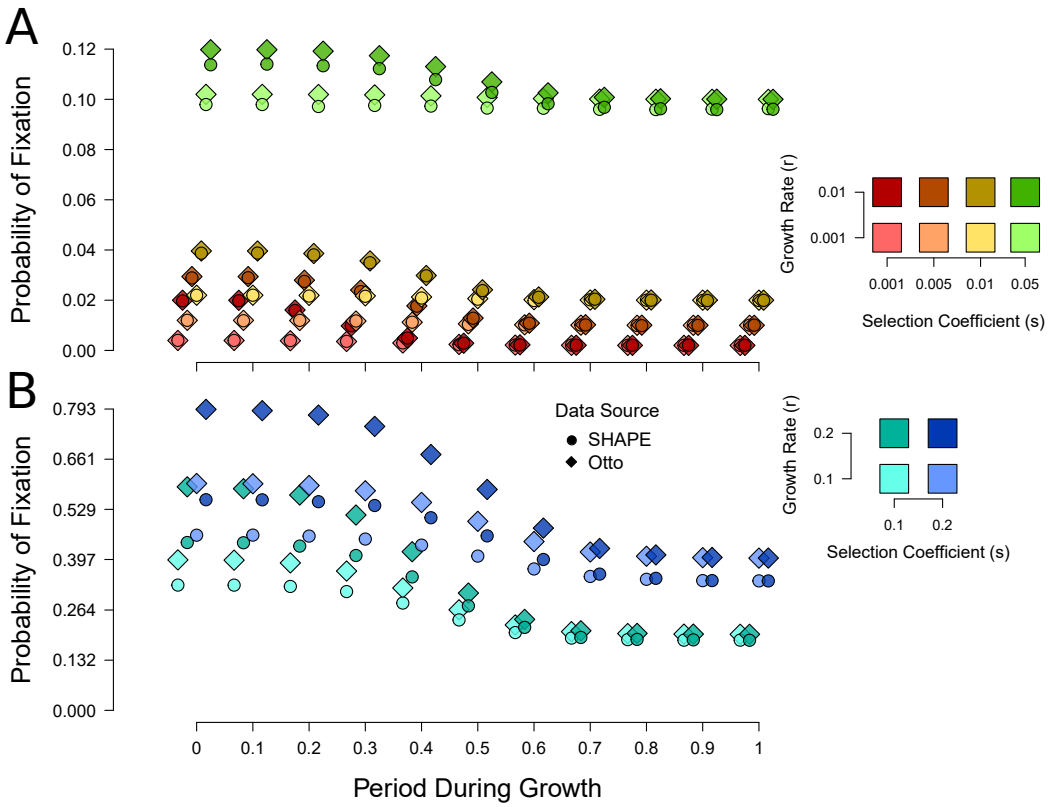

Figure 4.4.4: Theoretical and simulated probability of a single mutant fixing in populations growing to carrying capacity. Diamond shapes represent the analytical approximation of Otto and Whitlock [1997] while circles are for estimates using rSHAPE. The colour reflects the selection coefficient (s) with darker colours represent higher intrinsic growth rates (r). The period during the growth phase is scaled from the start of growth until the point where the number of individuals reaches carrying capacity. A presents the range of parameters originally studied in the theoretical work whereas B shows larger parameter values that implicitly violate assumptions of the analytical approximation.

available upon request.

4.4.2 rSHAPE application

Having validated the implementation of rSHAPE, I applied the simulation framework to assess the fixation probability of *de novo* mutants arising in a population that undergoes logistic growth and periodic population bottlenecks. It is common for microbial experimental evolution to include serial passaging, a practice

that allows continued evolution by transferring a portion of stationary phase communities to fresh growth media. The process of serial passaging introduces periodic bottlenecks which delineate growth phases of the community and is expected to bias the phenotype of successful mutations [Wahl and Zhu 2015]. Earlier theoretical work suggests that the probability of a mutation both arising within the population and surviving repeated rounds of serial passaging is roughly uniform throughout growth phases that are effectively exponential [Wahl *et al.* 2002]. However, there are no estimates of how serial passaging affects fixation probability when growth is logistic in these communities.

To estimate this, I used rSHAPE to simulate the evolution of populations undergoing repeated rounds of serial passaging between growth phases that were modelled as either exponential or logistic growth. To compare the results of rSHAPE with previous theoretical work, I parameterised my simulations following the exponential and nutrient limited growth conditions of Wahl *et al.* [2002]. The nutrient limited growth parameters were applied to my simulations of logistic growth. I found that rSHAPE does estimate a nearly uniform survival probability for *de novo* mutants arising in an exponential growth population (Fig. 4.4.5A). The parameters used in these simulations included a growth rate ($r = 2$) and mutation selection coefficient ($s = 0.1$) that would be deemed as "high" values outside parameter range of the simplifying assumptions in antecedent theoretical works [Haldane 1927; Otto and Whitlock 1997]. Similar to my findings during validation, and because of the "high" selection coefficient and growth rate parameters, I had expected rSHAPE to estimate a lower fixation probability than the analytical approximation of Wahl *et al.* [2002], which it did (Fig. 4.4.5A). For conditions of logistic growth, I found that the joint probability of a mutant arising during the growth phase and subsequently surviving repeated rounds of serial passaging was

not uniform (Fig. 4.4.5B). To better understand what evolutionary factors contributed to this result, I visualised and compared the individual probabilities that the mutant lineage arose during the growth phase, ultimately survived serial passaging, and also measured the expected mutant lineage size by the end of the growth phase in which it arose (Fig. 4.4.5C,D,E respectively). Note that the joint probabilities estimated for exponential and logistic growth differ by two orders of magnitude (Fig. 4.4.5A,B), this reflects the magnitude difference between the product of the initial populations size and mutation rate parameters used ($N_0 \in (10^5, 10^7)$ and $\mu \in (5 \times 10^{-5}, 4 \times 10^{-9})$ for the exponential and logistic growth conditions respectively). This order of magnitude difference is reflected in the estimated probability of the mutant's arrival throughout a growth phase (Fig. 4.4.5C). Under logistic growth, *de novo* mutants have a relatively higher probability of arising at all times up until the environmental carrying capacity is nearly reached (Fig. 4.4.5C). Yet, the probability of surviving repeated rounds of serial passaging is lower when growth is logistic (Fig. 4.4.5D), an observation likely underlain by the relatively smaller mutant lineage population size achieved in the first logistic growth phase (Fig. 4.4.5E).

While a simple application, this example demonstrates how rSHAPE can be used to test hypotheses related to evolutionary scenarios where analytical modelling is not yet practicable. By applying rSHAPE to study the effects of serial passaging on *de novo* mutants arising during logistic growth, I provide evidence that the timing of mutant arrival during a growth phase does matter. If we assume that any, but not all, "next evolutionary step" mutants are equally likely to result from mutation events throughout the growth phase, then a biased fixation probability relative to time of arrival will increase stochasticity of survival for unestablished (*i.e.* newly

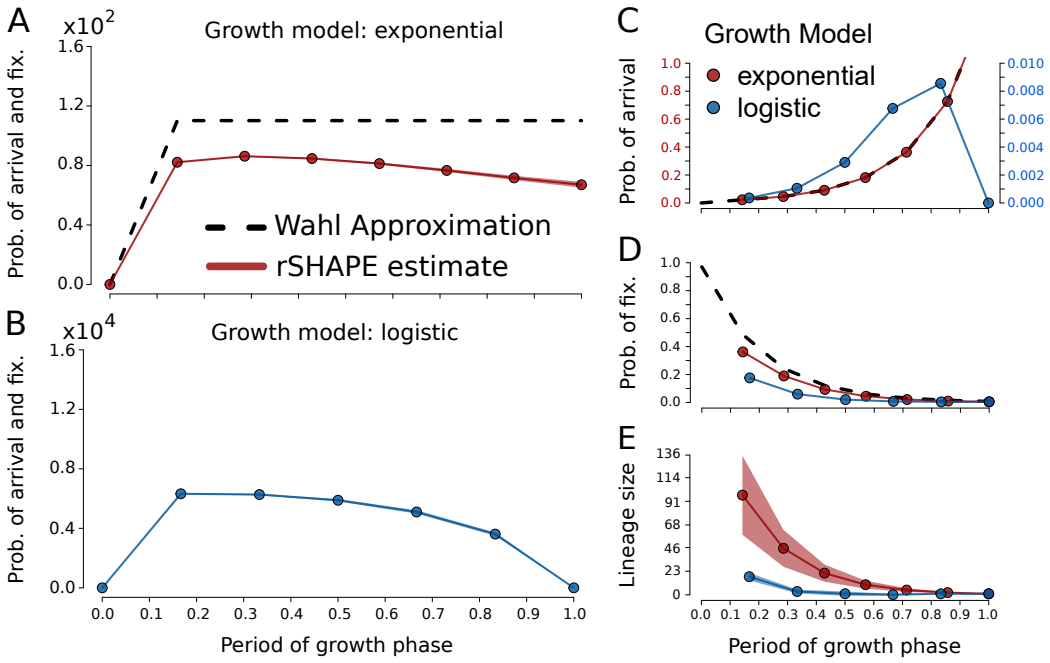

Figure 4.4.5: The joint probability of a single mutant arising during growth and then surviving many population bottlenecks, as a function of the time when a mutant arises during growth phases. Mutant selection coefficient $s = 0.1$, bottleneck strength was set to $D = 100$, and other parameters varied based on the model of growth (see main text). A compares the discrete analytical approximations of Wahl *et al.* [2002] to the estimates of rSHAPE for exponential growth. B shows the estimated probability of fixation when growth follows logistic form. C shows the independent probability of a mutant having arisen at a time during the growth phase whereas D visualises the probability to have ultimately survived repeated rounds of serial passaging. E shows estimates of the mutant lineage size by the end of the first growth phase in which it arose.

arisen, small) mutant genotypes. Previous work has shown that microbial evolution experiments likely represent a case where an intermediate number of all possible mutants compete for fixation - a condition termed intermediate clonal interference [Bailey *et al.* 2016]. Intermediate clonal interference would suggest that only a subset of all mutants arise in any given generation and so serial passaging would introduce a "lottery" effect biasing fixation toward the subset of beneficial mutants arising first during an experiment. Since many microbial evolution experiments include serial passaging within their experimental design, it may be that serial passaging reduces the repeatability of genomic changes observed in those

experiments. While this hypothesis remains to be tested, rSHAPE would be an ideal platform with which to perform such an analysis.

4.5 Availability and future directions

rSHAPE is made available through CRAN or can be installed from GitHub `github.com/JDench/SHAPE_library`. This software is made freely available under the GNU General Public License v3.0.

Future developmental priorities for rSHAPE are threefold. First, I would expand the applicability of rSHAPE by including methods to handle different habitat patches, partitioning of the genome, and temporal heterogeneity to selection. Second, implement additional fitness landscape models such Fisher's Geometric Model[Fisher 1958] and an eggbox model [Ferretti *et al.* 2016]. Lastly, I want to translate rSHAPE from the language of R to C in order to improve runtime through use of POSIX thread libraries and time for database queries.

Chapter 5

General conclusions

While studying the "usual suspects" (*i.e.* large effect, main drivers) affecting evolution has advanced our understanding of general principles, the role of interacting evolutionary factors remains understudied. This knowledge gap persists in part because we need to prioritise the limited time and resources available for studying evolutionary principles. In my thesis I have leveraged the rapidly growing resource of computing power to expand our capacity for studying genomic evolution. With my newly developed software, I studied the underpinnings of correlated evolution in the influenza A and *Pseudomonas aeruginosa* genomes before assessing if the common laboratory practice of serial passaging may reduce the repeatability of experimental evolution involving logistic growth conditions.

With my first two studies, I provide empirical evidence that the software AEGIS can predict epistatic pairs of both amino acid and nucleotide sequence positions. Furthermore, this approach is practical for genome-wide analyses and may be able to identify correlated networks of molecular/genomic sequence substitutions, though this last point requires direct experimental validation.

With my last study, I demonstrate that the *in silico* simulation software rSHAPE

may be well positioned to support both theoretical and empirical study of evolution. Theoretical study may be supported as rSHAPE can provide accurate estimates under evolutionary conditions for which we lack appropriate analytical models. Empirical study of evolution will benefit from rSHAPE because the software can provide a level of detailed observation impractical, or possibly impossible, to obtain from *in vivo/in vitro* systems.

5.1 Review of chapter 2

In my first study I helped develop the novel tool AEGIS for detecting correlated evolution among genomic sites. Specifically, I identified that AEGIS had exceptional specificity, but moderate sensitivity (better than 50% under half of the conditions tested and $\geq 95\%$ under specific conditions for alignments with ≥ 128 sequences), to detect correlated evolution and that the later was most affected by the balance of phenotypes compared and the phylogenetic tree's structure. Through analysis of real datasets and previously published results, identified either computationally or through direct experimental methods, my co-authors and I found that AEGIS identified signals of correlated evolution among most experimentally validated sites but had much less overlap with other purely computational approaches. From analysis of correlated evolution in the NP gene of influenza A, we found evidence of long range interactions between seven pairs of amino acid positions.

More interestingly, the seven correlated pairs appear to be linked and form a putative network of correlated evolution involving long range interactions. Previous work has shown that correlated evolution of amino acid positions form energetically coupled pathways connected through the protein fold [Lockless and Ranganathan 1999] as well as "mutative networks" driven by some of the amino acid sites [Du

et al. 2010]. By mapping the substitution history onto a phylogeny of influenza A, we found that two of the correlated pairs involve amino acid substitutions that revert once additional correlated partners mutate. One of these revertant amino acid substitutions forms a correlated pair within the putative network that was previously validated as epistatic [Gong *et al.* 2013]. Correlated networks may evolve for any number of reasons, but in our case the timing and reversions involving network hub amino acid positions allows us to suggest a possible mechanism. The viability of evolutionary paths may be contingent upon epistatic interactions [Weinreich *et al.* 2006] involving substitutions not otherwise under selection and so correlated networks may simply evolve due to a change in the fitness landscape of mutations caused by epistasis. At the same time, while epistatic pairs should evolve in response to their current selective context, they may well carry pleiotropic fitness costs in other contexts [Fisher 1958; Orr 2005]. When selective pressures change, the previously beneficial substitutions may carry a cost that requires compensation by secondary site mutations [Melnyk *et al.* 2017]. Both a change in genomic and selective environment can result in correlated networks evolving due to the "fitness seascape" of non-equilibrium selective dynamics [Mustonen and Lässig 2009]. Revertant substitutions within correlated networks suggest pleiotropic costs, but it is not obvious if these are due to a change in selective context driven by the genome or environment. However, the influenza A strains studied were isolated across the globe and over 40 years during which the vaccines used to protect from influenza A have changed. This suggests that a change in selective context likely plays an important role in the changing selective context of our putative correlated network. Regardless, AEGIS implements a method to detect pairwise correlated evolution and it remains to be tested if chained pairs of correlated substitutions form genuine correlated networks such as was the case for the pairwise approach of Lockless and

Ranganathan [1999].

Only one of the correlated pairs in our putative network had been previously identified and was experimentally validated to be epistatic. While it may yet be that the six remaining pairs are false positives, it is also possible that these other sites have not been previously identified because they are i) historically contingent upon the main epistatic pair of amino acid positions; ii) their correlation is not due to positive epistasis; or that iii) their evolutionary implications leave a signal that was only detectably by the approach implemented in AEGIS.

Considering the relatively poor overlap between AEGIS and other computational approaches, as well as recommendations that computational predictions be validated by direct experimentation [de Visser and Krug 2014], we should directly measure if there are epistatic interactions among the correlated interactions we have identified in influenza A. At the same time, we should measure the fitness landscape for the ordered set of substitutions - as mapped on the influenza A phylogeny - to asses the validity of the putative correlated network identified in the NP of influenza A. This study would use a cell culture system for competitive growth of the viruses and include RT-PCR to estimate viral titres at various time points. Like other studies considering the interactions among a set of focal mutations [Malcolm *et al.* 1990; O'Maille *et al.* 2008], a further investigation of the putative network of correlated sites within influenza A's NP segment should measure the phenotypic effects of all combinations. Such an analysis would provide direct evidence for whether or not AEGIS' pairwise approach can detect correlated networks of evolution.

Despite the simulations we performed to assess the statistical performance of AEGIS, previous work demonstrates the need for caution when using the phylogeny-based approach that AEGIS implements [Rodrigue 2013]. Caution is recommended by Rodrigue [2013] because a phylogeny-based site-wise likelihood

modelling approach violates the underlying asymptotic conditions of the likelihood ratio test and the predictive ability of such an approach cannot be properly tested. Further, Rodrigue [2013] shows how the results of simulations may provide inappropriate confidence in the use of phylogeny-based site-wise likelihood modelling. From communications with Dr. Rodrigue, an alternative approach could be to identify some general model, or mixture of models, that would represent both independent and correlated evolution. In this alternative case, general model parameters (*e.g.* transition rates) would be estimated for each of the independent and correlated scenarios using all pairs of sites as the data. After fitting the general parameters, an empirical Bayes paired site-wise approach would estimate the evidence of independent *v.s.* correlated evolution. However, Bayesian approaches are computationally demanding [Poon *et al.* 2008b] and as such impractical for genome-wide analyses. While we could try to fit the general model parameters using a sub-sample of all site pairs, we lack empirical evidence to inform what extent of sub-sampling would be appropriate. Despite the limitations of the approach implemented in AEGIS, my co-authors and I have shown - through comparison to experimentally validated pairs of sites - that AEGIS can identify genuine instances of correlated evolution and thus represents a valuable improvement to our research "toolkit". Until alternative computational approaches - such as described above - become feasible, further work should continue to assess AEGIS by experimentally validating novel correlated pairs of sites.

5.2 Review of chapter 3

Having shown that AEGIS could detect correlated evolution between pairs of genomic sites, and that some of these sites are epistatic, I assayed the exome of

Pseudomonas aeruginosa with AEGIS to detect signals of correlated evolution in response to fluroquinolone antibiotic selection. I found significant evidence of correlated evolution among $\sim 1.25\%$ of the ~ 10 million pairwise positions tested in the exome of *P. aeruginosa* but that only a single correlated pair of sites showed evidence of responding to fluoroquinolone antibiotic selection. Between that pair of sites, I directly measured epistasis for antibiotic resistance, and genotype dependent fitness epistasis in the absence of antibiotic selection.

Most of the detected coevolving pairs of sites were both synonymous and I found that their signals of correlated evolution were likely driven by physical linkage. While I have demonstrated that AEGIS can detect correlated evolution among sites distributed across a prokaryotic exome, I found little evidence that these pairs of sites evolved in response to the considered selective context and so performed additional computational analyses that led me to conclude they were unlikely to affect evolutionary outcome.

Though I studied the selective context of fluoroquinolone antibiotic selection, *P. aeruginosa* infections of CF patients may be treated with a cocktail of multiple drugs and rifampicin is known to have synergistic effects with fluoroquinolones [Neu 1993]. The majority of isolates used in my analysis came from CF patients. AEGIS had detected more than 20,000 significantly correlated pairs of sites that involve at least one substitution in *rpoB* a gene wherein substitutions are known to confer rifampicin resistance [Campbell *et al.* 2001]. A follow-up study should focus on the combined fluoroquinolone and rifampicin selective context. One recent study has shown that for *Escherichia coli*, mutations in *rpoB* conferred a competitive advantage while under ciprofloxacin (a fluoroquinolone antibiotic) selection and that they carried significant fitness costs outside the context of antibiotic selection [Pietsch *et al.* 2016]. Furthermore, the ciprofloxacin selected mutations in *rpoB* fixed

after mutation at *gyrA 259*. We did detect correlated evolution between substitutions at *gyrA 259* and *rpoB 864*. This correlated pair is found within one of two putative correlated networks found between substitutions in the *gyrA* and *rpoB* genes (Fig. 5.2.1). The first putative network, involving *gyrA 259*, contains six substitutions that form five correlated pairs between the genes. Most of the substitutions involved in those five correlated intergenic pairs are associated with several other intragenic substitutions forming the putative network.

From my analysis, how many of the significantly correlated pairs of synonymous substitutions may have selective advantage? Other studies have demonstrated that synonymous substitutions may be adaptive [Agashe *et al.* 2016; Bailey *et al.* 2014; Chamary and Hurst 2005; Cuevas *et al.* 2012; Duan *et al.* 2003; Lebeuf-Taylor *et al.* 2019; Plotkin and Kudla 2011; Walsh *et al.* 2020]. In my study, I detected a "driverless cohort'' [Buskirk *et al.* 2017] of four strongly correlated synonymous substitutions in the topoisomerase *parC*. Despite *parC* protein being a target of fluoroquinolone antibiotics, I found no evidence that the detected "cohort" responded to fluoroquinolone antibiotic selection, were hitchhiking with a driver mutation, or that they were horizontally transferred. A recent study demonstrates that synonymous substitutions may reduce fitness by leading to proteins more accessible to degradation despite being functional and properly folded [Walsh *et al.* 2020]. Could this "driverless cohort" in *parC* form a correlated network resulting from compensatory or permissive synonymous substitutions?

Though I found that correlated evolution in the exome of *P. aeruginosa* was underlain by a pair's proximity in the genome, it remains to be seen how much of this is biologically meaningful rather than simply a product of physical linkage. How many of the numerous correlated pairs of synonymous substitutions identified

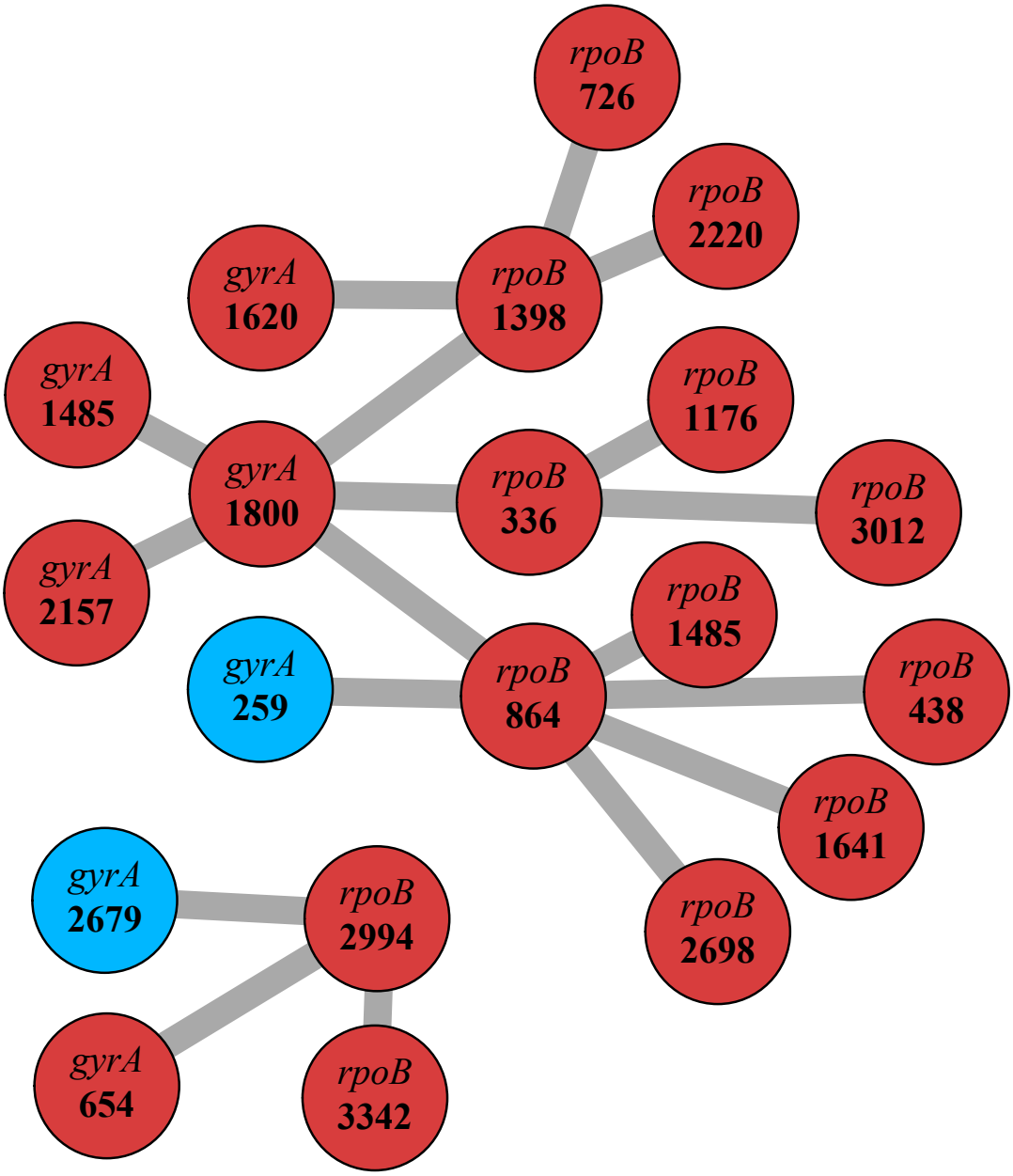

Figure 5.2.1: Putative *gyrA-rpoB* correlated networks in *P. aeruginosa*. Identified from an analysis of correlated evolution in the context of fluoroquinolone antibiotic resistance (see Chap. 3). Circles show the gene name and nucleotide positions identified as significantly correlated - $P \leq 0.0001$ - when connected by grey lines. Blue (red) circles represent sites that contained nonsynonymous (synonymous) substitutions in the analysed data. Correlated evolution was identified between pairs of nucleotide positions, the putative network shown requires validation such as by direct experimental evidence.

in my analysis are instances of compensatory substitutions that maintain the viability of the protein product? Is it possible that correlated pairs of synonymous substitutions between genes affect fitness for proteins that physically interact? As a first step to answer these questions, I propose that we follow up with a study of the *gyrA-rpoB* or *parC* putative networks identified. We should assess potential transcriptional and translational effects of the correlated synonymous pairs using methods similar to Lebeuf-Taylor *et al.* [2019] both within and without fluoroquinolone antibiotic selection.

While it remains to be directly measured if the correlated sites involving *rpoB*, or the "cohort" in *parC*, have any detectable epistatic interactions, my exome-wide analysis of correlated evolution in *P. aeruginosa* has identified three additional putative correlated networks worthy of attention. This demonstrates that AEGIS has practical implications for directing future studies and may prove to be a tool for identifying more than simply pairwise correlated evolution.

5.3 Review of chapter 4

In my last study, I provided evidence that the common laboratory practice of serial passaging biases fixation probability toward mutants arising early in experiments. Since previous work suggests that only some of all mutants arise and compete at one time for fixation [Bailey *et al.* 2016], the biased fixation probability caused by serial passaging introduces a "lottery" effect that would impact repeatability - at the genomic level - provided multiple beneficial mutations exist. I performed this study with the new *in silico* experimental evolution framework rSHAPE which I believe to be best suited for two follow-up studies.

The first would be to study the underlying mechanisms that led Vogwill *et al.*

[2016] to report that weak and strong population bottlenecks result in "greedy" phenotypic steps while intermediate levels allow greater diversity. I suspect that the bottleneck strengths studied by Vogwill *et al.* [2016] led to similar evolutionary outcome underlain by different evolutionary dynamics. At low population bottlenecks, evolutionary steps may be "greedy" due to clonal interference [Gerrish and Lenski 1998; Muller 1932] whereas under high bottlenecks only the largest effect mutations are likely to survive the process of serial passaging. The higher diversity observed at intermediate bottleneck strength may simply result from increased stochasticity due to loss of rare mutants. These suggestions are supported by theoretical work proposing that intermediate bottleneck strengths minimise loss of rare mutants [Wahl *et al.* 2002].

My second proposed study would quantify the proportion of the stochasticity in microbial experimental evolution attributable to the practice of serial passaging. I expect that sharing such a quantity with the research community may add insight when interpreting the results of experimental evolution. Furthermore, knowing what proportion of variance is attributable to the practice of serial passaging would allow for better partition of variance in models that use microbial experimental evolution to estimate repeatability of evolution or derived factors.

Regardless, I believe rSHAPE represents a valuable tool with which to reproduce the selective contexts of published *in vivo*/*in vitro* experiments while allowing us to study the influence of otherwise un-manipulable underlying evolutionary parameters. By comparing results of rSHAPE to previously published observations, we can gain insight into unknowns such as a genome's fitness landscape.

5.4 Final thoughts

Here I have developed two novel computational tools for the study of interacting evolutionary factors and applied them to study the evolution of substitutions in the genome of asexual haploids. I have demonstrated that results may vary between computational comparative approaches but that they can identify pairs of genomic sites worthy of direct experimentation. While my study of correlated evolution led to an appreciation of Levins [1966]'s statement that "Our truth is the intersection of many independent lies", through development of rSHAPE I have provided a tool to study the "analytically insoluble" [Levins 1966] interactions affecting haploid asexual population evolution. Through my thesis I have shown how computational approaches require the support of direct experimentation to confirm mechanisms, but also how we can gain additional insight into the results of experimental approaches thanks to computational tools. This thesis demonstrates that the interaction of evolutionary factors should not be ignored and provides new means for their study. Our interpretation of experiments often looks for a single "smoking gun'' but similar experiments have been known to show rather equivocal, or contradictory, results. I believe we could reconcile our interpretations of evolutionary observations if we better understood the interplay among evolutionary factors.

References

Abed, Y., Goyette, N., and Boivin, G. 2005. Generation and characterization of recombinant influenza A (H1N1) viruses harboring amantadine resistance mutations. *Antimicrob Agents Chemother*, 49(2): 556–9.

Agashe, D., Sane, M., Phalnikar, K., Diwan, G. D., Habibullah, A., Martinez-Gomez, N. C., Sahasrabuddhe, V., Polachek, W., Wang, J., Chubiz, L. M., and Marx, C. J. 2016. Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. *Molecular Biology and Evolution*, 33(6): 1542–1553.

Aita, T., Uchiyama, H., Inaoka, T., Nakajima, M., Kokubo, T., and Husimi, Y. 2000. Analysis of a local fitness landscape with a model of the rough mt. fuji-type landscape: Application to prolyl endopeptidase and thermolysin. *Biopolymers*, 54(1): 64–79.

Akasaka, T., Tanaka, M., Yamaguchi, A., and Sato, K. 2001. Type II topoisomerase mutations in fluoroquinolone-resistant clinical strains of *Pseudomonas aeruginosa* isolated in 1998 and 1999: role of target enzyme in mechanism of fluoroquinolone resistance. *Antimicrob Agents Chemother*, 45(8): 2263–8.

Alfaro, E., Gámez, M., and García, N. 2013. adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, 54(2): 1–35.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3): 403 – 410.

Analytics, R. and Weston, S. 2013. foreach: Foreach looping construct for r. *R package version*, 1(1): 2013.

Angiuoli, S. V. and Salzberg, S. L. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3): 334–342.

Anisimova, M. and Gascuel, O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*, 55(4): 539–52.

Arenas, M. 2015. Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 319(6).

Aris-Brosou, S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol Biol Evol*, 22(2): 200–9.

Aris-Brosou, S. 2014. Inferring influenza global transmission networks without complete phylogenetic information. *Evolutionary Applications*, 7(3): 403–412.

Aris-Brosou, S. and Rodrigue, N. 2012. The essentials of computational molecular evolution. In M. Anisimova, editor, *Evolutionary Genomics*, volume 855 of *Methods in Molecular Biology*, pages 111–152. Humana Press.

Aris-Brosou, S., Ibeh, N., and Noël, J. 2017. Viral outbreaks involve destabilized evolutionary networks: evidence from ebola, influenza and zika. *Sci Rep*, 7(1): 11881.

Arnold, B. J., Gutmann, M. U., Grad, Y. H., Sheppard, S. K., Corander, J., Lipsitch, M., and Hanage, W. P. 2018. Weak epistasis may drive adaptation in recombining bacteria. *Genetics*.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. 2000. Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. *Mol Biol Evol*, 17(1): 164–78.

Atilgan, A. R., Akan, P., and Baysal, C. 2004. Small-world communication of residues and significance for protein dynamics. *Biophysical Journal*, 86(1): 85–91.

Azzalini, A. 2016. *The R package* `sn`*: The Skew-Normal and Skew-t distributions (version 1.4-0)*. Università di Padova, Italia.

Baer, C. F. and Lynch, M. 2003. Correlated evolution of life-history with size at maturity in *Daphnia pulicaria*: patterns within and between populations. *Genetics Research*, 81(2): 123–132.

Bailey, S. F., Hinz, A., and Kassen, R. 2014. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nature Communications*, 5: 4076–4076.

Bailey, S. F., Blanquart, F., Bataillon, T., and Kassen, R. 2016. What drives parallel evolution? *BioEssays*, 39(1): e201600176.

Bank, C., Hietpas, R. T., Wong, A., Bolon, D. N., and Jensen, J. D. 2014. A bayesian mcmc approach to assess the complete distribution of fitness effects of new mutations: Uncovering the potential for adaptive walks in challenging environments. *Genetics*, 196(3): 841–852.

Bank, C., Matuszewski, S., Hietpas, R. T., and Jensen, J. D. 2016. On the (un)predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences*, 113(49): 14085–14090.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. 2008. The influenza virus resource at the national center for biotechnology information. *Journal of Virology*, 82(2): 596–601.

Bateson, W. 1909. *Mendel's principles of heredity*. Cambridge :University Press.

Bateson, W. 1911. Review from cambridge: Mendel's principles of heredity. *The Edinburgh Review*, 213(435): 77.

Beckmann, B. E., Clune, J., and Ofria, C. 2010. Digital evolution with avida. In *Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '10, pages 2917–2926, New York, NY, USA. ACM.

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

Bertani, G. 1951. Studies on lysogenesis I. : The mode of phage liberation by lysogenic *Escherichia coli*. *Journal of Bacteriology*, 62(3): 293–300.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., and Schwede, T. 2014.

Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*, 42(Web Server issue): W252–8.

Blount, Z. D., Borland, C. Z., and Lenski, R. E. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 105(23): 7899–7906.

Bruchmann, S., Dötsch, A., Nouri, B., Chaberny, I. F., and Häussler, S. 2013. Quantitative contributions of target alteration and decreased drug accumulation to pseudomonas aeruginosa fluoroquinolone resistance. *Antimicrobial Agents and Chemotherapy*, 57(3): 1361–1368.

Burns, J. L., Emerson, J., Stapp, J. R., Yim, D. L., Krzewinski, J., Louden, L., Ramsey, B. W., and Clausen, C. R. 1998. Microbiology of sputum from patients at cystic fibrosis centers in the united states. *Clinical Infectious Diseases*, 27(1): 158–163.

Buskirk, S. W., Peace, R. E., and Lang, G. I. 2017. Hitchhiking and epistasis give rise to cohort dynamics in adapting populations. *Proceedings of the National Academy of Sciences*, 114(31): 8330–8335.

Cabello, F. C. 2006. Heavy use of prophylactic antibiotics in aquaculture: a growing problem for human and animal health and for the environment. *Environmental Microbiology*, 8(7): 1137–1144.

Callahan, B., Neher, R. A., Bachtrog, D., Andolfatto, P., and Shraiman, B. I. 2011. Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genetics*, 7(2): e1001315–e1001315.

Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A., and Darst, S. A. 2001. Structural mechanism for rifampicin inhibition of bacterial rna polymerase. *Cell*, 104(6): 901–912.

Card, K. J., LaBar, T., Gomez, J. B., and Lenski, R. E. 2019. Historical contingency in the evolution of antibiotic resistance after decades of relaxed selection. *PLOS Biology*, 17(10): 1–18.

Chamary, J. and Hurst, L. D. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology*, 6(9): R75.

Chan, P. P. and Lowe, T. M. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res*, 44(D1): D184–9.

Chen, V. B., Davis, I. W., and Richardson, D. C. 2009. King (kinemage, next generation): a versatile interactive molecular and scientific visualization program. *Protein Science*, 18(11): 2403–2409.

Chen, Y., Carlini, D. B., Baines, J. F., Parsch, J., Braverman, J. M., Tanda, S., and Stephan, W. 1999. RNA secondary structure and compensatory evolution. *Genes & Genetic Systems*, 74(6): 271–286.

Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., Brisse, S., Moszer, I., Bourhy, H., Manuguerra, C. J., Lecuit, M., Burguiere, A., Caro, V., and Eloit, M. 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. *Journal of Clinical Microbiology*, 49(9): 3268–3275.

Cheverud, J. M. 1984. Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology*, 110(2): 155 – 171.

Cheverud, J. M., Dow, M. M., and Leutenegger, W. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: Sexual dimorphism in body weight among primates. *Evolution*, 39(6): 1335–1351.

Choy, W.-K., Zhou, L., Syn, C. K.-C., Zhang, L.-H., and Swarup, S. 2004. MorA defines a new class of regulators affecting flagellar development and biofilm formation in diverse *Pseudomonas* species. *Journal of Bacteriology*, 186(21): 7221–7228.

Cooper, N., Thomas, G. H., and FitzJohn, R. G. 2016. Shedding light on the 'dark side' of phylogenetic comparative methods. *Methods in Ecology and Evolution*, 7(6): 693–699.

Cuevas, J. M., Domingo-Calap, P., and Sanjuán, R. 2012. The fitness effects of synonymous mutations in DNA and RNA viruses. *Molecular Biology and Evolution*, 29(1): 17–20.

Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. 2012. Alf—a simulation framework for genome evolution. *Molecular Biology and Evolution*, 29(4): 1115–1123.

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*, 14(7): 1394–403.

Darwin, Charles; Wallace, A. R. 1859. *The Origin of Species*. John Murray, John Murray, Albemarle Street, 1st edition.

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

Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y.-C., Gray, E. R., Grant, P., Kanda, R. K., Leproust, E., Kellam, P., and Breuer, J. 2011. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLOS ONE*, 6(11): 1–7.

Desai, M. M. and Fisher, D. S. 2007. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, 176(3): 1759–98.

Desai, M. M., Fisher, D. S., and Murray, A. W. 2007. The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17(5): 386–394.

Dettman, J. R., Rodrigue, N., Aaron, S. D., and Kassen, R. 2013. Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. *Proc. of the Nat. Acad. of Sci. of the U.S.A.*, 110(52): 21065–21070.

Dib, L., Silvestro, D., and Salamin, N. 2014. Evolutionary footprint of coevolving positions in genes. *Bioinformatics*, 30(9): 1241–9.

Dib, L., Meyer, X., Artimo, P., Ioannidis, V., Stockinger, H., and Salamin, N. 2015. Coev-web: a web platform designed to simulate and evaluate coevolving positions along a phylogenetic tree. *BMC Bioinformatics*, 16: 394.

Domingo, J., Baeza-Centurion, P., and Lehner, B. 2019. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, 20(1): 433–460. PMID: 31082279.

Drake, J. W. 2007. Too many mutants with multiple mutations. *Critical Reviews in Biochemistry and Molecular Biology*, 42(4): 247–258.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. 1998. Rates of spontaneous mutation. *Genetics*, 148(4): 1667–1686.

Driscoll, J. A., Brody, S. L., and Kollef, M. H. 2007. The epidemiology, pathogenesis and treatment of pseudomonas aeruginosa infections. *Drugs*, 67(3): 351–368.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5): e88.

Du, Q.-S., Wang, C.-H., Liao, S.-M., and Huang, R.-B. 2010. Correlation analysis for protein evolutionary family based on amino acid position mutations and application in pdz domain. *PLOS ONE*, 5(10): 1–11.

Duan, J., Wainwright, M. S., Comeron, J. M., Saitou, N., Sanders, A. R., Gelernter, J., and Gejman, P. V. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Human Molecular Genetics*, 12(3): 205–216.

Duan, S., Govorkova, E. A., Bahl, J., Zaraket, H., Baranovich, T., Seiler, P., Prevost, K., Webster, R. G., and Webby, R. J. 2014. Epistatic interactions between neuraminidase mutations facilitated the emergence of the oseltamivir-resistant H1N1 influenza viruses. *Nature Communications*, 5.

Dutheil, J. Y., Jossinet, F., and Westhof, E. 2010. Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Molecular Biology and Evolution*, 27(8): 1868–1876.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792–1797.

Engler, C., Kandzia, R., and Marillonnet, S. 2008. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE*, 3(11): 1–7.

Ewens, W. 1967. The probability of survival of a new mutant in a fluctuating environment. *Heredity*, 22(3): 438–443.

Falconer, D. S. 1960. *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London.

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics*, 78(2): 737–756.

Felsenstein, J. 1985. Phylogenies and the comparative method. *The American Naturalist*, 125(1): 1–15.

Ferretti, L., Schmiegelt, B., Weinreich, D., Yamauchi, A., Kobayashi, Y., Tajima, F., and Achaz, G. 2016. Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. *Journal of Theoretical Biology*, 396: 132 – 143.

Figurski, D. H. and Helinski, D. R. 1979. Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proceedings of the National Academy of Sciences of the United States of America*, 76(4): 1648–1652.

Fisher, R. A. 1958. *The Genetical Theory of Natural Selection*. Dover Publications, New York, 2nd edition.

Fragata, I., Matuszewski, S., Schmitz, M. A., Bataillon, T., Jensen, J. D., and Bank, C. 2018. The fitness landscape of the codon space across environments. *Heredity*, 121(5): 422–437.

Freckleton, R. P. 2009. The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, 22(7): 1367–1375.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1): D261–D269.

Garcia, V. and Aris-Brosou, S. 2014. Comparative dynamics and distribution of influenza drug resistance acquisition to protein M2 and neuraminidase inhibitors. *Mol Biol Evol*, 31(2): 355–63.

Garrity, G. M., Bell, J. A., and Lilburn, T. G. 2004. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology. *Springer, New York, Berlin, Heidelberg*.

Gavrilets, S. 2004. *Fitness landscapes and the origin of species*, volume 41. Princeton University Press, Princeton, N.J.

Gerrish, P. and Lenski, R. 1998. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103(0): 127–144.

Gillespie, J. H. 1983. A simple stochastic gene substitution model. *Theoretical Population Biology*, 23(2): 202 – 215.

Gillespie, J. H. 1984. Molecular evolution over the mutational landscape. *Evolution*, 38(5): 1116–1129.

Gillespie, S. H., Voelker, L. L., and Dickens, A. 2002. Evolutionary barriers to quinolone resistance in *Streptococcus pneumoniae*. *Microbial Drug Resistance*, 8(2): 79–84.

Gloor, G. B., Martin, L. C., Wahl, L. M., and Dunn, S. D. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19): 7156–65.

Gogarten, J. P. and Townsend, J. P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9): 679–687.

Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. 2000. Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2): 283 – 293.

Gong, L. I., Suchard, M. A., and Bloom, J. D. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*, 2.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., and Desai, M. M. 2017. The dynamics of molecular evolution over 60,000 generations. *Nature*.

Goodwin, S., McPherson, J. D., and McCombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6): 333–351.

Gowri-Shankar, V. and Jow, H. 2006. *PHASE*: a software package for phylogenetics and sequence evolution. Online - URL: http://www.bioinf.man.ac.uk/resources/phase/.

Greenbaum, B. D., Levine, A. J., Bhanot, G., and Rabadan, R. 2008. Patterns of evolution and host gene mimicry in influenza and other rna viruses. *PLoS Pathogens*, 4(6): e1000079.

Guigueno, M. F., Shoji, A., Elliott, K. H., and Aris-Brosou, S. 2019. Flight costs in volant vertebrates: A phylogenetically-controlled meta-analysis of birds and bats. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, 235: 193 – 201.

Guillaume, F. and Rougemont, J. 2006. Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, 22(20): 2556–2557.

Haldane, J. B. S. 1927. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Proceedings of the Cambridge Philosophical Society*, 23: 838.

Hall, A. R. and MacLean, R. C. 2011. Epistasis buffers the fitness effects of rifampicin-resistance mutations in *Pseudomonas aeruginosa*. *Evolution*, 65(8): 2370–9.

Harms, M. J. and Thornton, J. W. 2014. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513): 203–7.

Hartl, D. L. 2014. What can we learn from fitness landscapes? *Curr Opin Microbiol*, 21: 51–7.

Harvey, P. H. and Pagel, M. 1991. *The comparative method in evolutionary biology*. Oxford series in ecology and evolution ; 2. Oxford University Press, Oxford ; New York.

Hayden, F. G. and de Jong, M. D. 2011. Emerging influenza antiviral resistance threats. *The Journal of Infectious Diseases*, 203(1): 6–10.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol*, 32(4): 835–45.

Hedrick, P. W. 1982. Genetic hitchhiking: A new factor in evolution? *BioScience*, 32(11): 845–853.

Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. 2011. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, 108(19): 7896–7901.

Hilterbrand, A., Saelens, J., and Putonti, C. 2012. CBDB: The codon bias database. *BMC Bioinformatics*, 13(1): 62.

Hoang, T. and Schweizer, H. 1999. Characterization of *Pseudomonas aeruginosa* enoyl-acyl carrier protein reductase (fabi): a target for the antimicrobial triclosan and its role in acylated homoserine lactone synthesis. *Journal of Bacteriology*, 181(17): 5489–5497.

Holloway, B. 1955. Genetic recombination in pseudomonas-aeruginosa. *Journal of general microbiology*, 13(3): 572–581.

Hsiung, G. D. 1984. Diagnostic virology: from animals to automation. *The Yale Journal of Biology and Medicine*, 57(5): 727–733.

Huelsenbeck, J. P., Nielsen, R., and Bollback, J. P. 2003. Stochastic Mapping of Morphological Characters. *Systematic Biology*, 52(2): 131–158.

Hughes, A. L. and Hughes, M. A. K. 2007. More effective purifying selection on rna viruses than in dna viruses. *Gene*, 404(1): 117 – 125.

Hurt, A. C. 2014. The epidemiology and spread of drug resistant human influenza viruses. *Current Opinion in Virology*, 8: 22 – 29. Antivirals and resistance / Virus evolution.

Ibeh, N., Nshogozabahizi, J. C., and Aris-Brosou, S. 2016. Both epistasis and diversifying selection drive the structural evolution of the Ebola virus glycoprotein mucin-like domain. *J Virol*, 90(11): 5475–84.

Illingworth, C. J. R. and Mustonen, V. 2012. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. *PLoS Pathog*, 8(12): e1003091.

Jain, R., Rivera, M. C., and Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7): 3801–3806.

James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An introduction to statistical machine learning with applications in R*. Springer.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21): 132.

Kauffman, S. and Levin, S. 1987. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1): 11 – 45.

Kauffman, S. A. and Weinberger, E. D. 1989. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2): 211 – 245.

Keightley, P. D. and Eyre-Walker, A. 2010. What can we learn about the distribution of fitness effects of new mutations from dna sequence data? *Philosophical Transactions of the Royal Society B*, 365(1544): 1187–1193.

Kelley, J. L., Fitzpatrick, J. L., and Merilaita, S. 2013. Spots and stripes: ecology and colour pattern evolution in butterflyfishes. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1757).

Kimmel, M. and Axelrod, D. E. 2002. *The Galton-Watson Process*, pages 33–63. Springer New York, New York, NY.

Kimura, M. 1985. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, 64(1): 7–19.

Kingman, J. F. C. 1978. A simple model for the balance between selection and mutation. *Journal of Applied Probability*, 15(1): 1–12.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M., and Beslon, G. 2007. A long-term evolutionary pressure on the amount of noncoding DNA. *Mol. Biol. Evol.*, 24(10): 2344–2353.

Koel, B. F., Burke, D. F., Bestebroer, T. M., van der Vliet, S., Zondag, G. C., Vervaet, G., Skepner, E., Lewis, N. S., Spronken, M. I., Russell, C. A., *et al.* 2013. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161): 976–979.

Kohanski, M. A., DePristo, M. A., and Collins, J. J. 2010. Sublethal antibiotic treatment leads to multidrug resistance via radical-induced mutagenesis. *Molecular Cell*, 37(3): 311–320.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A*, 90(15): 7176–80.

Kos, V. N., Déraspe, M., McLaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., Corbeil, J., and Gardner, H. 2015. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrobial Agents and Chemotherapy*, 59(1): 427–436.

Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., and Plotkin, J. B. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS genetics*, 7(2): e1001301.

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

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. 2009. Coding-sequence determinants of gene expression in escherichia coli. *Science*, 324.

Kvitek, D. J. and Sherlock, G. 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet*, 7(4): e1002056.

LaBauve, A. E. and Wargo, M. J. 2012. Growth and laboratory maintenance of pseudomonas aeruginosa. *Current protocols in microbiology*, Chapter 6: Unit–6.E.1.

Laland, K., Odling-Smee, J., and Endler, J. 2017. Niche construction, sources of selection and trait coevolution. *Interface Focus*, 7(5): 20160147.

Lambert, B. W., Terwilliger, J. D., and Weiss, K. M. 2008. Forsim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics*, 24(16): 1821–1822.

Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D., and Desai, M. M. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464): 571–4.

Larsen, D. H. and Dimmick, R. L. 1964. Attachment and growth of bacteria on surfaces of continuous-culture vessels. *Journal of Bacteriology*, 88(5): 1380–1387.

Lashin, S., Matushkin, Y. G., Klimenko, A., Suslov, V. V., and Kolchanov, N. A. 2014. *Evolution, Biodiversity and Ecology in Microbial Communities: Mathematical Modeling and Simulation with the "Haploid Evolutionary Constructor"Software Tool*. IntechOpen.

Lau, C. H. F., Hughes, D., and Poole, K. 2014. MexY-promoted aminoglycoside resistance in *Pseudomonas aeruginosa*: involvement of a putative proximal binding pocket in aminoglycoside recognition. *mBio*, 5(2): e01068–14–e01068–14.

Lawrence, D., Fiegna, F., Behrends, V., Bundy, J. G., Phillimore, A. B., Bell, T., and Barraclough, T. G. 2012. Species interactions alter evolutionary responses to a novel environment. *PLOS Biology*, 10(5): 1–11.

Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. 2013. Strong purifying selection at synonymous sites in d. melanogaster. *PLoS Genetics*, 9(5): 1–18.

Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., and Kassen, R. 2019. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *eLife*, 8: e45952.

Lehner, B. 2011. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8): 323–331.

Leland, D. S. and Ginocchio, C. C. 2007. Role of cell culture for virus detection in the age of technology. *Clinical Microbiology Reviews*, 20(1): 49–78.

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. 1991. Long-term experimental evolution in *Escherichia coli.* i. adaptation and divergence during 2,000 generations. *The American Naturalist*, 138(6): 1315–1341.

Levins, R. 1966. The strategy of model building in population biology. *American Scientist*, 54(4): 421–431.

Li, C., Qian, W., Maclean, C. J., and Zhang, J. 2016. The fitness landscape of a tRNA gene. *Science*, 352(6287): 837–840.

Lieberman, T. D., Flett, K. B., Yelin, I., Martin, T. R., McAdam, A. J., Priebe, G. P., and Kishony, R. 2014. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature Genetics*, 46: 82–87.

Liu, D., Shi, W., Shi, Y., Wang, D., Xiao, H., Li, W., Bi, Y., Wu, Y., Li, X., Yan, J., Liu, W., Zhao, G., Yang, W., Wang, Y., Ma, J., Shu, Y., Lei, F., and Gao, G. F. 2013. Origin and diversity of novel avian influenza a h7n9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. *The Lancet*, 381(9881): 1926 – 1932.

Lockless, S. W. and Ranganathan, R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438): 295–299.

Long, G. S., Hussen, M., Dench, J., and Aris-Brosou, S. 2019. Identifying genetic determinants of complex phenotypes from whole genome sequence data. *BMC Genomics*, 20(1): 470.

Losos, J. B. 2011. Convergence, adaptation, and constraint. *Evolution*, 65(7): pp. 1827–1840.

Losos, J. B., Arnold, S. J., Bejerano, G., Brodie, III, E. D., Hibbett, D., Hoekstra, H. E., Mindell, D. P., Monteiro, A., Moritz, C., Orr, H. A., Petrov, D. A., Renner, S. S., Ricklefs, R. E., Soltis, P. S., and Turner, T. L. 2013. Evolutionary biology for the 21st century. *PLOS Biology*, 11(1): 1–8.

Lässig, M., Mustonen, V., and Walczak, A. M. 2017. Predicting evolution. *Nature Ecology and Evolution*, 1(3): 785–801.

Lynch, M. 2007. *The origins of genome architecture*. Sinauer Associates, Sunderland, Mass.

Lynch, M. and Walsh, B. 1998. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Mable, B. K. and Otto, S. P. 2001. Masking and purging mutations following ems treatment in haploid, diploid and tetraploid yeast (*Saccharomyces cerevisiae*). *Genetical Research*, 77(1): 9–26.

MacLean, R. C., Perron, G. G., and Gardner, A. 2010. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in pseudomonas aeruginosa. *Genetics*, 186(4): 1345–54.

Maddison, W. P. 1990. A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution*, 44(3): 539–557.

Maddison, W. P. and FitzJohn, R. G. 2014. The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters. *Systematic Biology*, 64(1): 127–136.

Maddison, W. P., Midford, P. E., and Otto, S. P. 2007. Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology*, 56(5): 701–710.

Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F., and Wilson, A. C. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, 345: 86–89.

Marvig, R. L., Sommer, L. M., Molin, S., and Johansen, H. K. 2015. Convergent evolution and adaptation of *pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat Genet*, 47(1): 57–64.

Maynard Smith, J. 1978. Optimization theory in evolution. *Annual Review of Ecology and Systematics*, 9: 31–56.

Maynard Smith, J. and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(01): 23–35.

McDonald, M. J. 2019. Microbial experimental evolution – a proving ground for evolutionary theory and a tool for discovery. *EMBO reports*, 20(8): e46992.

McElroy, K. E., Hui, J. G. K., Woo, J. K. K., Luk, A. W. S., Webb, J. S., Kjelleberg, S., Rice, S. A., and Thomas, T. 2014. Strain-specific parallel evolution drives short-term diversification during pseudomonas aeruginosa biofilm formation. *Proceedings of the National Academy of Sciences*, 111(14): E1419–E1427.

McKimm-Breschkin, J. L. 2013. Influenza neuraminidase inhibitors: antiviral action and mechanisms of resistance. *Influenza and Other Respiratory Viruses*, 7(s1): 25–36.

Medeiros, A. A. 1997. Evolution and dissemination of β-lactamases accelerated by generations of β-lactam antibiotics. *Clinical Infectious Diseases*, 24(Supplement 1): S19–S45.

Melnyk, A. H., Wong, A., and Kassen, R. 2015. The fitness costs of antibiotic resistance mutations. *Evol Appl*, 8(3): 273–83.

Melnyk, A. H., McCloskey, N., Hinz, A. J., Dettman, J., and Kassen, R. 2017. Evolution of cost-free resistance under fluctuating drug selection in *Pseudomonas aeruginosa*. *mSphere*, 2(4).

Microsoft and Weston, S. 2019. *foreach: Provides Foreach Looping Construct*. R package version 1.4.7.

Mira, P. M., Meza, J. C., Nandipati, A., and Barlow, M. 2015. Adaptive landscapes of resistance genes change as antibiotic concentrations change. *Mol Biol Evol*, 32(10): 2707–15.

Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J., and King, J. 1991. Global suppression of protein folding defects and inclusion body formation. *Science*, 253(5015): 54–8.

Mohrig, J. R., Moerke, K. A., Cloutier, D. L., Lane, B. D., Person, E. C., and Onasch, T. B. 1995. Importance of historical contingency in the stereochemistry of hydratase-dehydratase enzymes. *Science*, 269(5223): 527–9.

Moon, D. C., Seol, S. Y., Gurung, M., Jin, J. S., Choi, C. H., Kim, J., Lee, Y. C., Cho, D. T., and Lee, J. C. 2010. Emergence of a new mutation and its accumulation in the topoisomerase iv gene confers high levels of resistance to fluoroquinolones in *Escherichia coli* isolates. *International Journal of Antimicrobial Agents*, 35(1): 76–79.

Muller, H. 1964. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1): 2 – 9.

Muller, H. J. 1932. Some genetic aspects of sex. *The American Naturalist*, 66(703): 118–138.

Müller, K., Wickham, H., James, D. A., and Falcon, S. 2017. *RSQLite: 'SQLite' Interface for R*. R package version 2.0.

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics*, 139(3): 1429–39.

Mustonen, V. and Lässig, M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet*, 25(3): 111–9.

Nasrallah, C. A. and Huelsenbeck, J. P. 2013. A phylogenetic model for the detection of epistatic interactions. *Molecular biology and evolution*, 30(9): 2197–2208.

Natarajan, C., Hoffmann, F. G., Weber, R. E., Fago, A., Witt, C. C., and Storz, J. F. 2016. Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science*, 354(6310): 336–339.

Nawrocki, E. P. and Eddy, S. R. 2013. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22): 2933–5.

Neher, E. 1994. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*, 91(1): 98–102.

Neidhart, J., Szendro, I. G., and Krug, J. 2014. Adaptation in tunably rugged fitness landscapes: the rough mount fuji model. *Genetics*, 198(2): 699–721.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., and Holmes, E. C. 2007. Phylogenetic analysis reveals the global migration of seasonal influenza a viruses. *PLOS Pathogens*, 3(9): 1–9.

Neu, H. C. 1993. Synergy and antagonism of fluoroquinolones with other classes of antimicrobial agents. *Drugs*, 45(3): 54–58.

Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B., and Bazykin, G. A. 2015. Coordinated evolution of influenza A surface proteins. *PLoS Genet*, 11(8): e1005404.

Newcomb, R. D., Campbell, P. M., Ollis, D. L., Cheah, E., Russell, R. J., and Oakeshott, J. G. 1997. A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. *Proc Natl Acad Sci U S A*, 94(14): 7464–8.

Nshogozabahizi, J. C., Dench, J., and Aris-Brosou, S. 2017. Widespread historical contingency in influenza viruses. *Genetics*, 205(1): 409–420.

Nyerges, Á., Csörgő, B., Draskovits, G., Kintses, B., Szili, P., Ferenc, G., Révész, T., Ari, E., Nagy, I., Bálint, B., Vásárhelyi, B. M., Bihari, P., Számel, M., Balogh, D., Papp, H., Kalapis, D., Papp, B., and Pál, C. 2018. Directed evolution of multiple genomic loci allows the prediction of antibiotic resistance. *Proceedings of the National Academy of Sciences*, 115(25): E5726–E5735.

Nylander, J. 2015. catfasta2phyml. https://github.com/nylander/catfasta2phyml.

Ochman, H., Lawrence, J. G., and Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784): 299–304.

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

O'Maille, P. E., Malone, A., Dellas, N., Andes Hess, B., Smentek, L., Sheehan, I., Greenhagen, B. T., Chappell, J., Manning, G., and Noel, J. P. 2008. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chemical Biology*, 4(10): 617–623.

Organization, W. H. 2015. Global action plan on antimicrobial resistance. https://www.who.int/emergencies/ten-threats-to-global-health-in-2019.

Organization, W. H. 2019a. Global influenza strategy 2019-2030. https://www.who.int/influenza/global_influenza_strategy_2019_2030/en/.

Organization, W. H. 2019b. Ten health issues who will tackle this year. https://www.who.int/emergencies/ten-threats-to-global-health-in-2019.

Organization, W. H. *et al.* 2014. *Antimicrobial resistance: global report on surveillance*. World Health Organization.

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

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., and Thornton, J. W. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317(5844): 1544–8.

Otto, S. P. and Gerstein, A. C. 2008. The evolution of haploidy and diploidy. *Current Biology*, 18(24): R1121 – R1124.

Otto, S. P. and Whitlock, M. C. 1997. The probability of fixation in populations of changing size. *Genetics*, 146(2): 723–733.

Oz, T., Guvenek, A., Yildiz, S., Karaboga, E., Tamer, Y. T., Mumcuyan, N., Ozan, V. B., Senturk, G. H., Cokol, M., Yeh, P., and Toprak, E. 2014. Strength of Selection Pressure Is an Important Parameter Contributing to the Complexity of Antibiotic Resistance Evolution. *Molecular Biology and Evolution*, 31(9): 2387–2401.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings: Biological Sciences*, 255(1342): 37–45.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26: 331–348.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*, 401(6756): 877–84. Copyright - Copyright Nature Publishing Group Oct 28, 1999; Last updated - 2018-10-16; CODEN - NATUAS.

Pagel, M. and Meade, A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist*, 167(6): 808–825.

Panagea, S., Winstanley, C., Parsons, Y. N., Walshaw, M. J., Ledson, M. J., and Hart, C. 2003. Pcr-based detection of a cystic fibrosis epidemic strain of *Pseudomonas aeruginosa*. *CNS Drugs*, 7(195): 195–200.

Paradis, E. 2006. *Analysis of phylogenetics and evolution with R*. Springer, New York.

Perfeito, L., Fernandes, L., Mota, C., and Gordo, I. 2007. Adaptive mutations in bacteria: High rate and small effects. *Science*, 317(5839): 813–815.

Phillips, P. C. 2008. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9(11): 855–867.

Pietsch, F., Bergman, J. M., Brandis, G., Marcusson, L. L., Zorzet, A., Huseby, D. L., and Hughes, D. 2016. Ciprofloxacin selects for rna polymerase mutations with pleiotropic antibiotic resistance effects. *Journal of Antimicrobial Chemotherapy*, 72(1): 75–84.

Plate, T. and Heiberger, R. 2016. *abind: Combine Multidimensional Arrays*. R package version 1.4-5.

Plotkin, J. B. and Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12.

Pollock, D. D., Taylor, W. R., and Goldman, N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, 287(1): 187–98.

Poole, K. 2005. Efflux-mediated antimicrobial resistance. *Journal of Antimicrobial Chemotherapy*, 56(1): 20–51.

Poon, A. F., Lewis, F. I., Frost, S. D., and Pond, S. L. K. 2008a. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics*, 24(17): 1949–1950.

Poon, A. F. Y., Lewis, F. I., Pond, S. L. K., and Frost, S. D. W. 2007a. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. *PLoS Comput Biol*, 3(1): e11.

Poon, A. F. Y., Lewis, F. I., Pond, S. L. K., and Frost, S. D. W. 2007b. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput Biol*, 3(11): e231.

Poon, A. F. Y., Lewis, F. I., Frost, S. D. W., and Kosakovsky Pond, S. L. 2008b. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics*, 24(17): 1949–1950.

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010a. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3).

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010b. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3): e9490.

R Core Team 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Special Interest Group on Databases (R-SIG-DB), Wickham, H., and Müller, K. 2016. *DBI: R Database Interface*. R package version 0.5-1.

Rahme, L. G., Stevens, E. J., Wolfort, S. F., Shao, J., Tompkins, R. G., and Ausubel, F. M. 1995. Common virulence factors for bacterial pathogenicity in plants and animals. *Science*, 268(5219): 1899–1902.

Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. 2015. Inferring horizontal gene transfer. *PLOS Computational Biology*, 11(5): 1–16.

Rodrigue, N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*, 193(2): 557–564.

Roy, P. H., Tetu, S. G., Larouche, A., Elbourne, L., Tremblay, S., Ren, Q., Dodson, R., Harkins, D., Shay, R., Watkins, K., Mahamoud, Y., and Paulsen, I. T. 2010. Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7. *PLoS ONE*, 5(1): e8842.

Rozen, D. E., de Visser, J. A. G. M., and Gerrish, P. J. 2002. Fitness effects of fixed beneficial mutations in microbial populations. *Curr Biol*, 12(12): 1040–5.

Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. *Genetics*, 141(2): 771–83.

Sackman, A. M. and Rokyta, D. R. 2017. Additive phenotypes underlie epistasis of fitness effects. *Genetics*.

Sadikot, R. T., Blackwell, T. S., Christman, J. W., and Prince, A. S. 2005. Pathogen-host interactions in pseudomonas aeruginosa pneumonia. *American journal of respiratory and critical care medicine*, 171(11): 1209–1223.

Sandie, R. and Aris-Brosou, S. 2014. Predicting the emergence of H3N2 influenza viruses reveals contrasted modes of evolution of HA and NA antigens. *J Mol Evol*, 78(1): 1–12.

Sandle, T. 2010. History and development of microbiological culture media. *institute of Science and Technology Journal*.

Sanjuán, R. and Elena, S. F. 2006. Epistasis correlates to genomic complexity. *Proc Natl Acad Sci U S A*, 103(39): 14402–5.

Schick, A. and Kassen, R. 2018. Rapid diversification of pseudomonas aeruginosa in cystic fibrosis lung-like conditions. *Proceedings of the National Academy of Sciences*, 115(42): 10714–10719.

Schöniger, M. and von Haeseler, A. 1994. A stochastic model for the evolution of autocorrelated dna sequences. *Mol Phylogenet Evol*, 3(3): 240–7.

Schubert, B., Maddamsetti, R., Nyman, J., Farhat, M. R., and Marks, D. S. 2019. Genome-wide discovery of epistatic loci affecting antibiotic resistance in neisseria gonorrhoeae using evolutionary couplings. *Nature Microbiology*, 4(2): 328–338.

Schweizer, H. P. 2008. Bacterial genetics: past achievements, present state of the field, and future challenges. *BioTechniques*, 44: 633–641.

Sexton, J. P., McIntyre, P. J., Angert, A. L., and Rice, K. J. 2009. Evolution and ecology of species range limits. *Annual Review of Ecology Evolution and Systematics*, 40: 415–436.

Shapiro, B., Rambaut, A., Pybus, O. G., and Holmes, E. C. 2006. A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. *Mol Biol Evol*, 23(9): 1724–30.

Shimizu, A., Dohzono, I., Nakaji, M., Roff, D. A., Miller III, D. G., Osato, S., Yajima, T., Niitsu, S., Utsugi, N., Sugawara, T., and et al. 2014. Fine-tuned bee-flower coevolutionary state hidden within multiple pollination interactions. *Scientific Reports*, 4(1).

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, 51(3): 492–508.

Shimodaira, H. and Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinfo*, 17(12): 1246–1247.

Shindyalov, I. N., Kolchanov, N. A., and Sander, C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*, 7(3): 349–58.

Shoji, A., Aris-Brosou, S., and Elliott, K. 2016. Physiological constraints and dive behavior scale in tandem with body mass in auks: A comparative analysis. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, 196: 54 – 60.

Simonsen, L., Viboud, C., Grenfell, B. T., Dushoff, J., Jennings, L., Smit, M., Macken, C., Hata, M., Gog, J., Miller, M. A., *et al.* 2007. The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. *Molecular biology and evolution*, 24(8): 1811–1820.

Singer, R. S., Finch, R., Wegener, H. C., Bywater, R., Walters, J., and Lipsitch, M. 2003. Antibiotic resistance—the interplay between antibiotic use in animals and human beings. *The Lancet Infectious Diseases*, 3(1): 47 – 51.

Skwark, M. J., Croucher, N. J., Puranen, S., Chewapreecha, C., Pesonen, M., Xu, Y. Y., Turner, P., Harris, S. R., Beres, S. B., Musser, J. M., Parkhill, J., Bentley, S. D., Aurell, E., and Corander, J. 2017. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLOS Genetics*, 13(2): 1–24.

Slater, G. J., Harmon, L. J., and Alfaro, M. E. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution*, 66(12): 3931–3944.

Smith, E. E., Buckley, D. G., Wu, Z., Saenphimmachak, C., Hoffman, L. R., D'Argenio, D. A., Miller, S. I., Ramsey, B. W., Speert, D. P., Moskowitz, S. M., Burns, J. L., Kaul, R., and Olson, M. V. 2006. Genetic adaptation by pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences of the United States of America*, 103(22): 8487–8492.

Smith, W., Andrewes, C. H., and Laidlaw, P. P. 1933. A virus obtained from influenza patients. *The Lancet*, 222(5732): 66–68.

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M. B. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet*, 84(6): 740–59.

Sorrells, T. R., Booth, L. N., Tuch, B. B., and Johnson, A. D. 2015. Intersecting transcription networks constrain gene regulatory evolution. *Nature*.

Sriramulu, D. D., Lünsdorf, H., Lam, J. S., and Römling, U. 2005. Microcolony formation: a novel biofilm model of pseudomonas aeruginosa for the cystic fibrosis lung. *Journal of Medical Microbiology*, 54(7): 667–676.

Starkey, M., Hickman, J. H., Ma, L., Zhang, N., De Long, S., Hinz, A., Palacios, S., Manoil, C., Kirisits, M. J., Starner, T. D., Wozniak, D. J., Harwood, C. S., and Parsek, M. R. 2009. *Pseudomonas aeruginosa* rugose small-colony variants have adaptations that likely promote persistence in the cystic fibrosis lung. *Journal of Bacteriology*, 191(11): 3492–3503.

Stephenson, A. G. 2002. evd: Extreme value distributions. *R News*, 2(2): 0.

Stewart, S. M. and Pekosz, A. 2011. Mutations in the membrane-proximal region of the influenza A virus M2 protein cytoplasmic tail have modest effects on virus replication. *Journal of Virology*, 85(23): 12179–12187.

Strelkowa, N. and Lässig, M. 2012. Clonal interference in the evolution of influenza. *Genetics*, 192(2): 671–82.

Sutto, L., Marsili, S., Valencia, A., and Gervasio, F. L. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci U S A*, 112(44): 13567–72.

Takanami, M. and Zubay, G. 1964. An estimate of the size of the ribosomal site for messenger RNA binding. *Proceedings of the National Academy of Sciences of the United States of America*, 51(5): 834–839.

Talavera, D., Lovell, S. C., and Whelan, S. 2015. Covariation is a poor measure of molecular coevolution. *Mol Biol Evol*, 32(9): 2456–68.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. 1997. A genomic perspective on protein families. *Science*, 278(5338): 631–637.

Tavaré, S. 1986. Some probalistic and statistical problems in the analysis of DNA sequences. *In: Lectures on Mathematics in the Life Sciences*, 17: 57–86.

Taylor, B. L. and Zhulin, I. B. 1999. Pas domains: Internal sensors of oxygen, redox potential, and light. *Microbiology and Molecular Biology Reviews*, 63(2): 479–506.

Taylor, W. R. and Hatrick, K. 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng*, 7(3): 341–8.

Team, R. D. C. *et al.* 2013. R: A language and environment for statistical computing.

The UniProt Consortium 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1): D158–D169.

Thomas, V. L., McReynolds, A. C., and Shoichet, B. K. 2010. Structural bases for stability-function tradeoffs in antibiotic resistance. *J Mol Biol*, 396(1): 47–59.

Thompson, W. W., Shay, D. K., Weintraub, E., Brammer, L., Cox, N., Anderson, L. J., and Fukuda, K. 2003. Mortality associated with influenza and respiratory syncytial virus in the united states. *JAMA*, 289(2): 179–186.

Thompson, W. W., Shay, D. K., Weintraub, E., Brammer, L., Bridges, C. B., Cox, N. J., and Fukuda, K. 2004. Influenza-associated hospitalizations in the united states. *JAMA*, 292(11): 1333–1340.

Tillier, E. R. and Collins, R. A. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal rna. *Genetics*, 148(4): 1993–2002.

Travisano, M., Mongold, J., Bennett, A., and Lenski, R. 1995. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science*, 267(5194): 87–90.

Trindade, S., Sousa, A., Xavier, K. B., Dionisio, F., Ferreira, M. G., and Gordo, I. 2009. Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genetics*, 5(7): e1000578–e1000578.

Tsunoyama, K., Bellgard, M. I., and Gojobori, T. 2001. Intragenic variation of synonymous substitution rates is caused by nonrandom mutations at methylated cpg. *Journal of Molecular Evolution*, 53(4): 456–464.

Uyeda, J. C., Zenil-Ferguson, R., and Pennell, M. W. 2018. Rethinking phylogenetic comparative methods. *Systematic Biology*, 67(6): 1091–1109.

van der Vries, E., Schutten, M., Fraaij, P., Boucher, C., and Osterhaus, A. 2013. Chapter six - influenza virus resistance to antiviral therapy. In E. D. Clercq, editor, *Antiviral Agents*, volume 67 of *Advances in Pharmacology*, pages 217 – 246. Academic Press.

Vogwill, T., Kojadinovic, M., and MacLean, R. C. 2016. Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of pseudomonas. *Proceedings of the Royal Society B: Biological Sciences*, 283(1830): 20160151.

Wahl, L. M. and Gerrish, P. J. 2001. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution*, 55(12): 2606–2610.

Wahl, L. M. and Zhu, A. D. 2015. Survival probability of beneficial mutations in bacterial batch culture. *Genetics*, 200(1): 309–320.

Wahl, L. M., Gerrish, P. J., and Saika-Voivod, I. 2002. Evaluating the impact of population bottlenecks in experimental evolution. *Genetics*, 162(2): 961–971.

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., and Clark, P. L. 2020. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences*, 117(7): 3528–3534.

Wang, J., Wu, Y., Ma, C., Fiorin, G., Wang, J., Pinto, L. H., Lamb, R. A., Klein, M. L., and DeGrado, W. F. 2013. Structure and inhibition of the drug-resistant S31N mutant of the M2 ion channel of influenza A virus. *Proceedings of the National Academy of Sciences*, 110(4): 1315–1320.

Weinreich, D. M. 2010. Predicting molecular evolutionary trajectories in principle and in practice. In *Encyclopedia of Life Sciences*, pages 1–9. John Wiley & Sons, Ltd: Chichester.

Weinreich, D. M., Watson, R. A., and Chao, L. 2005. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories.

Weinreich, D. M., Delaney, N. F., DePristo, M. A., and Hartl, D. L. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770): 111–114.

Wibowo, A. 2013. Performance standards for antimicrobial susceptibility testing: 23rd informational supplement m100-s23. In *Performance standards for antimicrobial susceptibility testing: 23rd informational supplement M100-S23*, volume 33 of *Developments in Environmental Modelling*, pages 62–65. Clinical and Laboratory Standards Institute.

Winsor, G. L., Lam, D. K. W., Fleming, L., Lo, R., Whiteside, M. D., Yu, N. Y., Hancock, R. E. W., and Brinkman, F. S. L. 2011. Pseudomonas genome database: improved comparative analysis and population genomics capability for pseudomonas genomes. *Nucleic Acids Res*, 39(Database issue): D596–600.

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. S. L. 2016. Enhanced annotations and features for comparing thousands of pseudomonas genomes in the pseudomonas genome database. *Nucleic Acids Res*, 44(D1): D646–53.

Wong, A., Rodrigue, N., and Kassen, R. 2012. Genomics of adaptation during experimental evolution of the opportunistic pathogen *Pseudomonas aeruginosa*. *PLoS Genet*, 8(9): e1002928.

Worobey, M., Han, G.-Z., and Rambaut, A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495): 254–7.

Xia, X. 2014. A major controversy in codon-anticodon adaptation resolved by a new codon usage index. *Genetics*, 199(2): 573–579.

Xia, X. 2017. DAMBE6: New tools for microbial genomics, phylogenetics, and molecular evolution. *Journal of Heredity*, 108(4): 431–437.

Xiao, N., Xu, Q., and Cao, D. 2014. protr: Protein sequence descriptor calculation and similarity computation with r. *R package version 0.2-1, URL http://CRAN. R-project. org/package= protr*.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Syst Biol*, 47(1): 125–33.

Yang, Z. 2006. *Computational molecular evolution*, volume 21. Oxford University Press Oxford.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8): 1586–91.

Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular biology and evolution*, 19(6): 908—917.

Yee, T. W. and Wild, C. J. 1996. Vector generalized additive models. *Journal of Royal Statistical Society, Series B*, 58(3): 481–493.

Zanini, F. and Neher, R. A. 2012. Ffpopsim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*, 28(24): 3332–3333.

Zaraket, H., Saito, R., Suzuki, Y., Baranovich, T., Dapat, C., Caperig-Dapat, I., and Suzuki, H. 2010. Genetic makeup of amantadine-resistant and oseltamivir-resistant human influenza a/h1n1 viruses. *Journal of Clinical Microbiology*, 48(4): 1085–1092.

Zhang, H., Zeidler, A. F. B., Song, W., Puccia, C. M., Malc, E., Greenwell, P. W., Mieczkowski, P. A., Petes, T. D., and Argueso, J. L. 2013. Gene copy-number variation in haploid and diploid strains of the yeast saccharomyces cerevisiae. *Genetics*, 193(3): 785–801.

Zhang, J., Nielsen, R., and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12): 2472–2479.

Zheng, L., Baumann, U., and Reymond, J. L. 2004. An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic acids research*, 32(14): e115–e115.

Zhou, T., Gu, W., and Wilke, C. O. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Molecular biology and evolution*, 27(8): 1912–1922.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13): 3406–3415.

Appendices

Appendix A

Appendix to Chapter 2

A.1 Supplementary text

A.1.1 Comparison of methods to simulate epistasis.

To ensure that our simulation results were not biased by the method used to simulate epistasis, we ran a second fully factorial simulation study on a representative subset of the parameter space as described in the main text (*Materials and Methods: Validation based on simulations* - Chapter 2.3.2). The following modifications were made: (i) Sites that evolve in a correlated manner were simulated using Coev Dib *et al.* [2014] for five different strengths of correlation. This is controlled by the d/s rate parameter that represents the likelihood of a pair of sites to evolve toward (d) or away (s) from a correlated profile. (ii) Branch lengths (b) were varied across a $\log_2$ scale for all even integer values in $(-12, -4)$. (iii) Each tree contained a number of sequences n_s equal to 32 or 128.

With these Coev simulations, both specificity and precision (Fig. A.2.15) are higher than with the PHASE simulations, even at the lowest strength of correlation.

Sensitivity appears to have an optimal branch length, similar to what we found with PHASE. With Coev however, the decline in sensitivity results from monomorphic site patterns being generated. From our data recoding and analysis, monomorphic sites are not considered since they carry no evolutionary information. The Coev model defines a unique two character profile as the coevolutionary pair to be simulated. Thus, as evolutionary time (or strength (d/s) of correlated evolution) increases, more sequences adopt this unique correlation profile until sites become monomorphic. The Coev method of simulating epistasis reflects correlated evolution when only a single, correlated, adaptive response is observed for adaptation of an entire population; our PHASE simulations did not enforce unique profiles.

That strong selective pressures can lead to unique adaptive profiles reinforces our claim that alignments should be constructed to include taxa that have not undergone the selection, which would generate epistatic response. Comparison of our method's performance while simulating correlated evolution under two different profiles of adaptive response demonstrates that our approach is applicable to detecting epistasis during various adaptive responses.

A.1.2 Signal strength for detecting epistasis.

To determine the optimal phylogenetic distribution of epistatic sites, we generated columns in our alignments with varying proportions of "positively" and "negatively" correlated site pairs. These simulations were done with PHASE. We define positively correlated sites as those that have identical character states (*i.e.* $(1,1)$ and $(0,0)$), and negatively correlated site as those that have opposed character states (*i.e.* $(1,0)$ and $(0,1)$). These negatively correlated sites are equivalent to noise. Note that this designation of positively and negatively correlated pairs should not be

confused with the sign of epistatic interactions (positive or negative epistasis).

Intuitively, this combination leads to entirely symmetric outcomes, so the simulations proceeded by first generating a certain proportion of negatively correlated pairs; this proportion was varied between 1 and 31% of the alignment by increments of 3%. The remaining of each alignment was then filled with "positively" correlated site pairs of type $(0, 0)$ in a proportion ranging from 0 to 100% by increments of 10%. What remained, if any, was filled with the other positive state pair $(1, 1)$ (Fig. A.2.4: panel B5). We did this for sequence alignments comprising 32, 64 and 128 sequences, hereby leading to 604, 726 and 763 unique simulation conditions, respectively. Each of these simulation conditions was assayed for both tree shapes (symmetric and pectinate) with 500 replicates. Each replicate randomly shuffled the assignment of site pairs at tree tips. This resulted in nearly 2.1 million runs and thus sequence length (l_s) was fixed to one single epistatic pair to reduce computational burden. Root depth (tip-to-root distance) of trees was held constant across all conditions to allow for better comparison of the impact of additional sequences.

We find that balanced sequence alignments, that contain an equal proportion of correlated binary states, are optimal and that such alignments permit detection of correlated evolution even with high levels of noise ($\gg 31\%$; Fig. A.2.21). Consistent with our results on sensitivity, only alignments with 64 or more sequences show a high proportion of significant P-values, that is, a high ability to detect epistasis (Fig. A.2.21: panels A, D *vs.* B, C and E, F).

Tree shape, number of sequences, balance of positively correlated pairs as well as the proportion of negatively correlated pairs all affect our ability to detect epistasis (Table A.3.1). While all the terms are significant, and have high, similar AIC values (Table A.3.2), the balance of sequence distribution explains more of the variance

than tree shape, n_s or proportion of negatively correlated pairs (Table A.3.2). We compared the variance explained by our full model and a stepwise reduction of terms. The shape of a tree and n_s have a very weak interaction, while both terms have strong interactions with the balance of positively correlated pairs (Table A.3.2). Notably, tree shape and the proportion of negatively correlated pairs have an equally strong interaction, suggesting that certain phylogenies would be more susceptible to noise.

A.1.3 Assessing the effect of dinucleotide bias.

Dinucleotide bias can be seen as a form of correlated evolution, that can therefore interfere with our algorithm and generate "false positives" – pairs of sites that evolve in a correlated manner but that are not epistatic. As our method was mostly applied to influenza A viruses, which display dinucleotide bias Greenbaum *et al.* [2008], we assessed the impact that dinucleotide bias would have upon the statistical performance of our method. We performed simulations where coevolving sites had dinucleotide bias calculated from the influenza A nucleic acid sequences used in Kryazhimskiy *et al.* [2011]. Similar to our simulations comparing our approach with coevolution simulated by Coev, we performed simulations in a subset of the most interesting parameter space of our larger simulation study: branch lengths (b) varied across a $\log_2$ scale for even integer values in $(-12, -4)$, and the number of sequences n_s was either 32 or 128. As per the simulations presented in the main text, sites evolving epistatically were simulated under the RNA7D model however, independently evolving sites were simulated using the RNA16 model with dinucleotide bias. We set the probability of single mutations to be 95% and double mutations to 5%. These simulations show that specificity was significantly affected

by the inclusion of dinucleotide bias (see Fig. A.2.14: panel A; $F = 21.922$, $df = 1$,$P < 1 \times 10^{-5}$). While sensitivity and precision (see Fig. A.2.14: panel C) are lower when dinucleotide bias is included in simulations, we highlight that specificity remains above the maximum threshold, and precision remains above 95%.

We did not expect sensitivity to be different in this simulation study since epistatic sites are simulated using the same model. However, in Fig. A.2.14: panel B, we observe that sensitivity appears reduced. After carefully verifying this was not generated due an error in the simulation protocols, we performed statistical analysis to test if this difference was significant. Statistical analysis was done using a Type III SS anova of a general linear model with independent model of site evolution as the first term (see Table A.3.3). We find that many parameters are significant despite our having compared identical parameter spaces in our analysis. To explain this difference we note that standard error bars for sensitivity were huge in both these simulations and the larger study in our main text (these bars were not included for reasons of clarity). This information combined with the significance of parameters not varied between simulations suggests that the stochasticity of generating epistatic signal best explains the difference in sensitivity between simulation studies.

Altogether, these simulations provide significant evidence that a dinucleotide bias affects the specificity and possibly precision of our approach. Despite this impact the general performance of exceptional specificity and mediocre sensitivity are consistent.

A.1.4 Detection of chained correlations.

In essence, BayesTraits Pagel [1994]; Pagel and Meade [2006] detects pairs of traits (sites) that evolve in a correlated manner on a phylogenetic tree. However, as we did not simulate chained correlations, one could wonder if our use of BayesTraits can actually detect such patterns and tell a situation where sites 1 and 2 on the one hand and 2 and 3 on the other are correlated, *versus* a situation where 1 and 2, and 1 and 3 are correlated. Considering pairs of states at two sites, the simulations performed above demonstrate that if two sites evolve alongside each other, the chance of detecting this signal of correlation is very high. In other words, correlation is detected if each site undergo mutations together. If pairs of mutations at sites 1 and 2 occur along a different branch than those at sites 1 and 3, then we can clearly reconstruct chained interactions. It is because we find such a temporal pattern (see in particular Fig.,2.4.4 in main text) that we can reconstruct such chained interactions.

A.1.5 Interaction between sequence length and number of sequences.

Our general algorithm employs the BH correction for controlling FDR. As a result, the length of a sequence alignment can be expected to affect the ability to detect epistasis. To test this expectation, we simulated alignments of varying sequence lengths $l_s \in (10, 1280)$ on a $\log_2$ (doubling) scale. As our previous simulations showed little impact of tree shape (Results), we limited this simulation to symmetric trees and used PHASE here again. The number of sequences n_s was set to 32, 64 or 128 sequences. As above, root depth was held constant. For each n_s, the phylogenetic pattern that yielded the strongest signal for correlated evolution was

used in the analysis. In this way, we could establish expectations of the likelihood to detect correlated evolution as a function of l_s and n_s. Each combination was replicated only 10 times, as this simulation does not assess detection performance, but rather establishes a trend in P-value after FDR: the only factor changing is the number of pairwise comparisons, which is a function of sequence length.

Both the sequence length and number of sequences have significant effects on P-values after correcting for the False Discovery Rate (FDR; Table A.3.1; Fig. A.2.22). As expected, FDR-corrected P-values show a power law distribution as a function of sequence length (Fig. A.2.22): doubling sequence length halves our chance of finding a significant result, so that it becomes more difficult to detect any signal for epistasis with long sequences. Epistasis is not recoverable from alignments of only 32 sequences containing > 1000 sites (Fig. A.2.22).

A.2 Figures

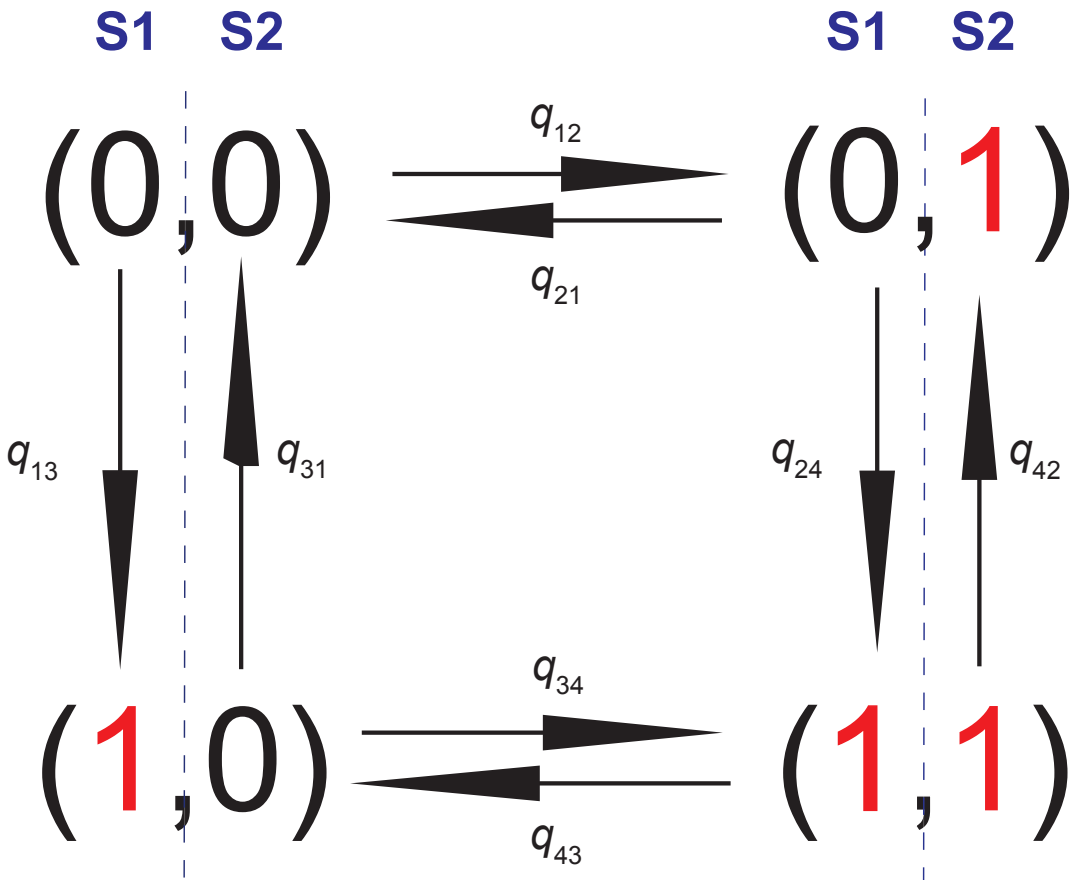

Figure A.2.1: Transitions among the four combinations of states resulting from two binary variables. Pair (i, j) identifies the states of a particular pair of traits, that here are amino acid positions S1 and S2. Transitions are represented by arrows. Instantaneous rates of change are denoted $q_{i,j}$.

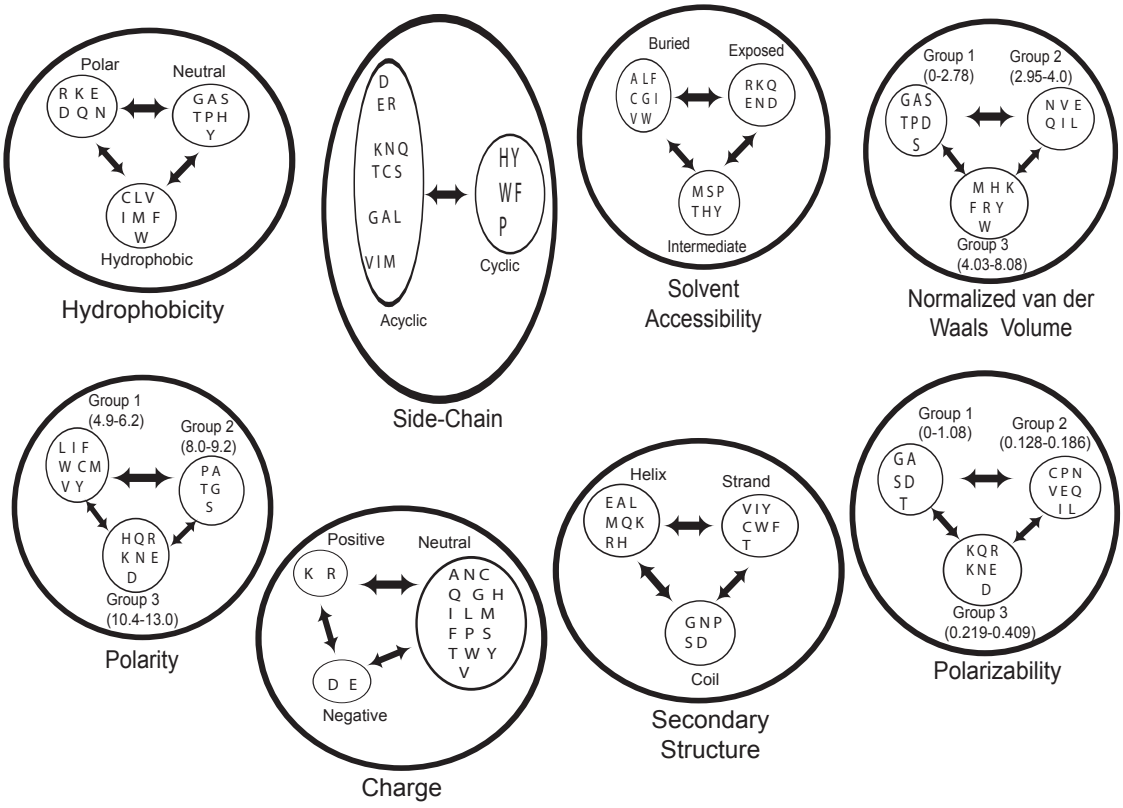

Figure A.2.2: The amino acids properties and their categories (states) used to recode protein alignments. For each category, the arrows show the different transitions among physiochemical properties. For continuous categories, the numbers between bracket indicate the range of values employed in the discretisation process as per the `protr` R package.

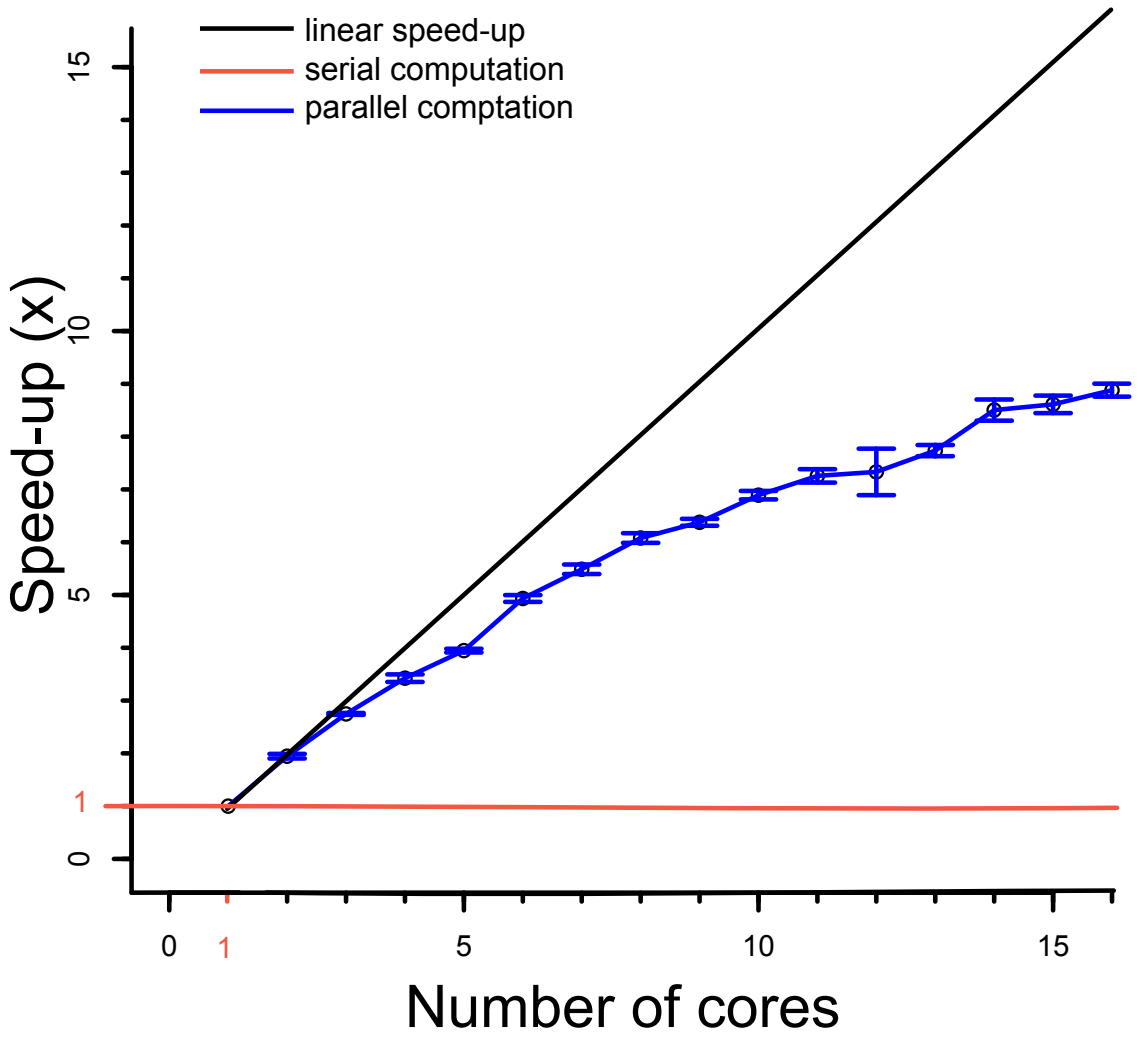

Figure A.2.3: Speed-up as a function of number of cores requested. Speed-up is a measure of "acceleration" (how many times faster is the computation with n cores) compared to a serial computation (on a single core; in red). The black line indicates the ideal situation of linear scaling of performance with resource. The actual speed-up curve is shown in blue. Error bars show one standard error, based on four replicates. Overheads dues to latency (among others) lead to a non-linear speed-up curve.

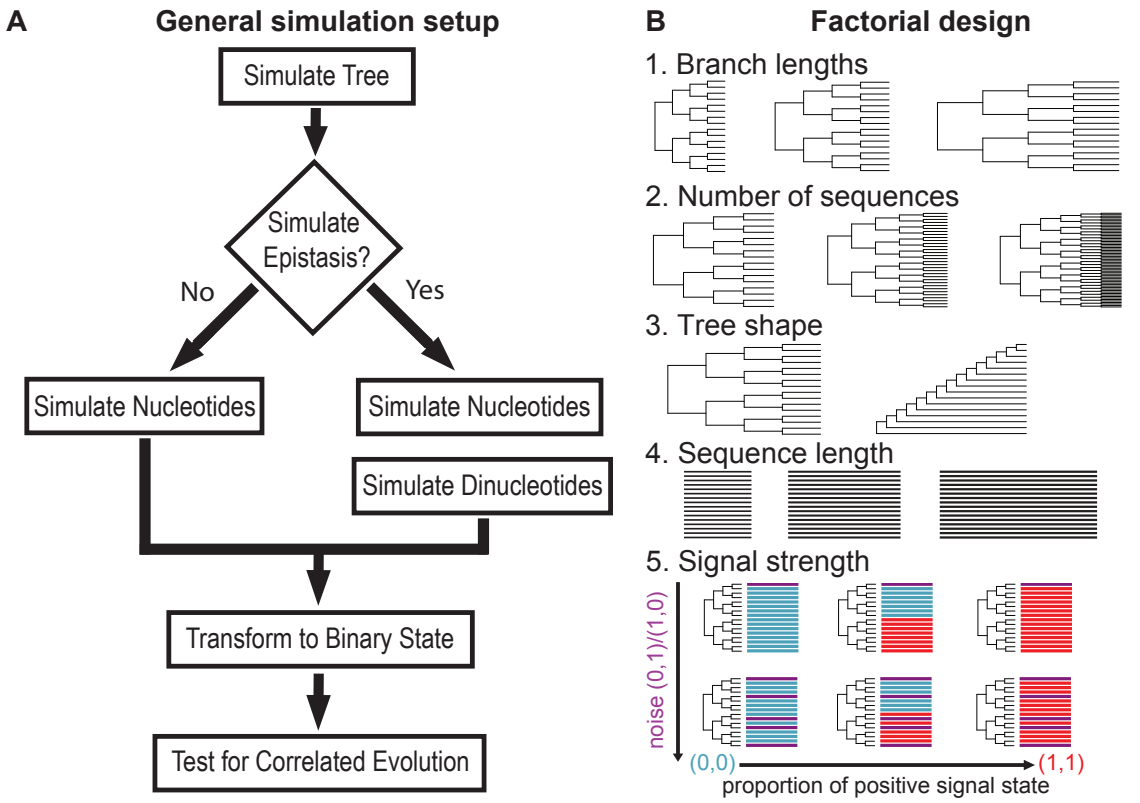

Figure A.2.4: Design of the simulation experiments. (A) The general simulation setup starts by simulating a tree of a particular tree shape (symmetric / pectinate) on which sequences are generated. The program PHASE 2.0 is used to generate either single nucleotides (independent model) or dinucleotides (dependent model). In both cases, nucleotides are recoded as 0's and 1's before being analyzed with BayesTraits, hereby testing for correlated evolution. (B) A fully factorial design was adopted to test for the effect of: 1. branch lengths, 2. the number of sequences, 3. tree shape, 4. sequence length and 5. signal strength. See Methods for details.

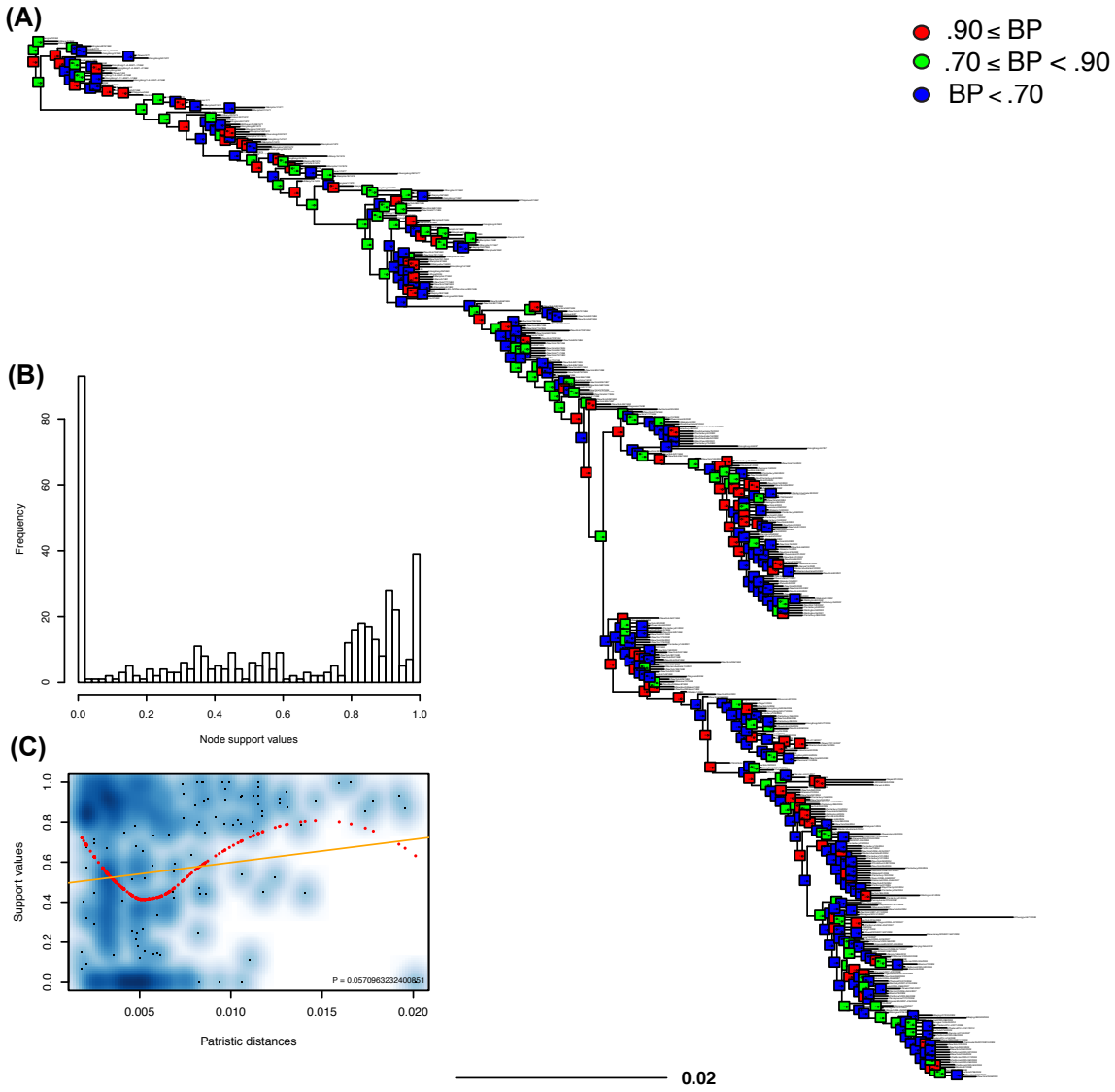

Figure A.2.5: Node support values for the phylogenetic analysis of the Gong13NP data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

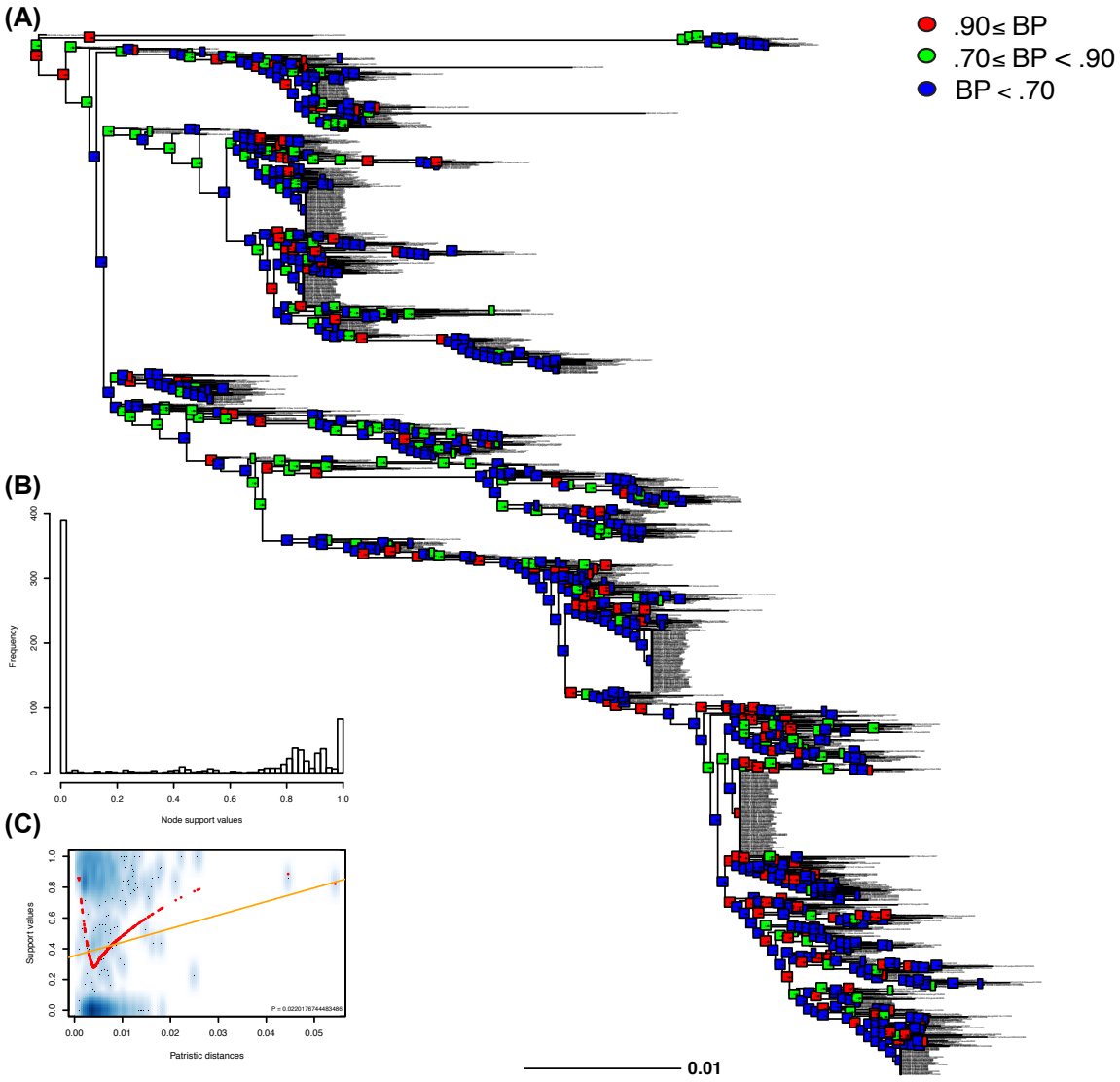

Figure A.2.6: Node support values for the phylogenetic analysis of the Duan14NA data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

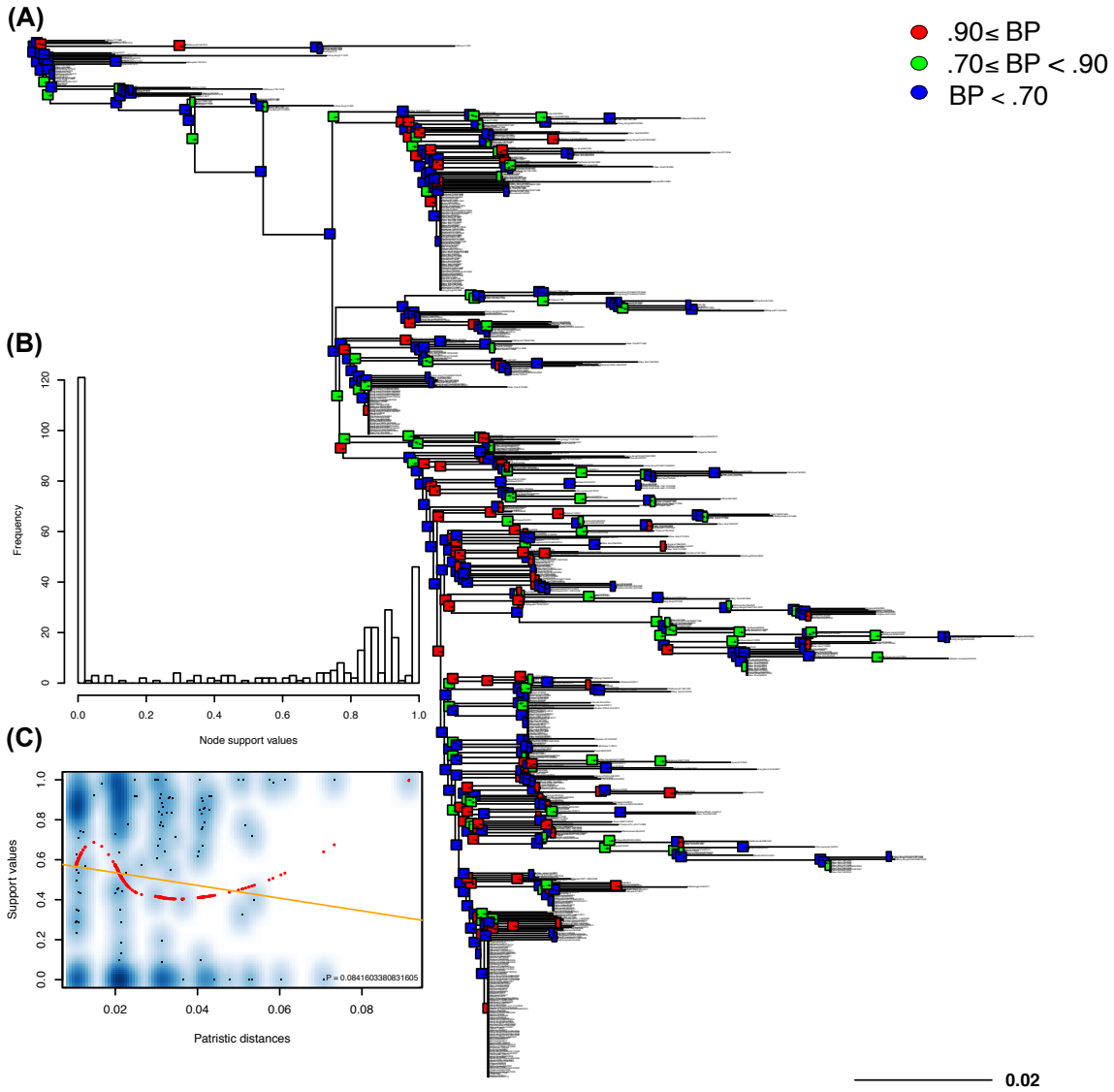

Figure A.2.7: Node support values for the phylogenetic analysis of the Adam03M2 data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

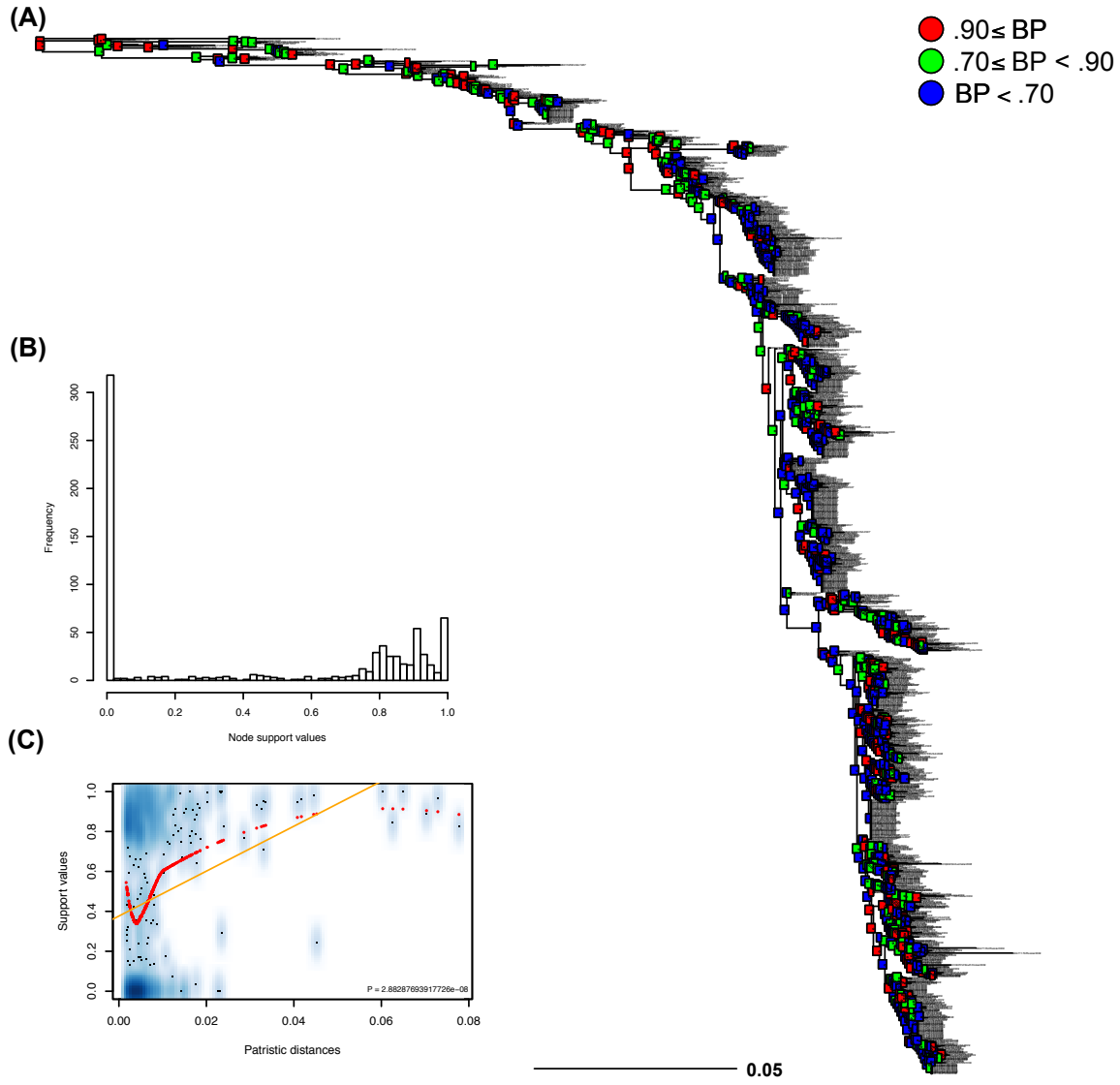

Figure A.2.8: Node support values for the phylogenetic analysis of the KDBP11 H1 data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

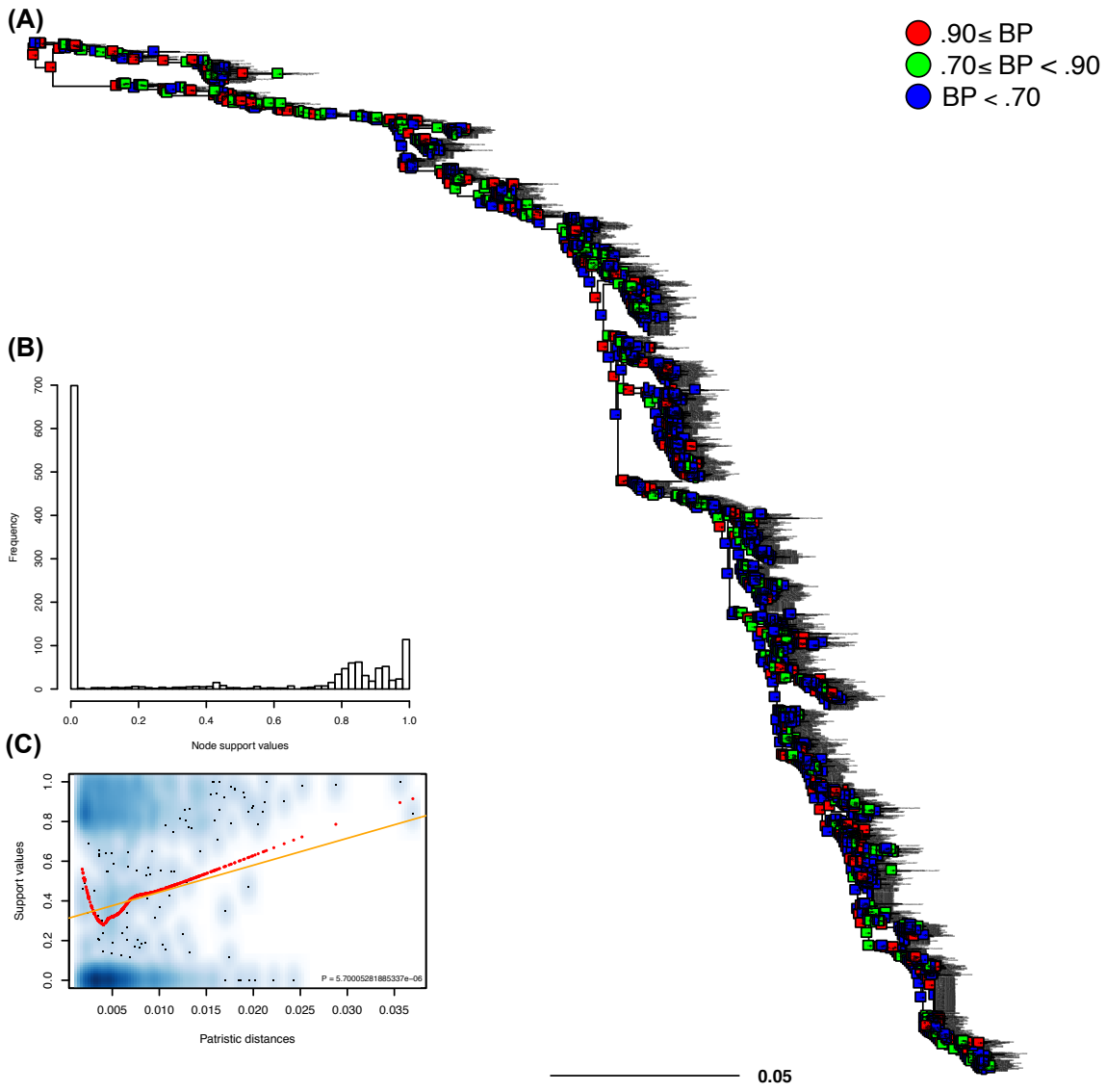

Figure A.2.9: Node support values for the phylogenetic analysis of the KDBP11 H3 data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

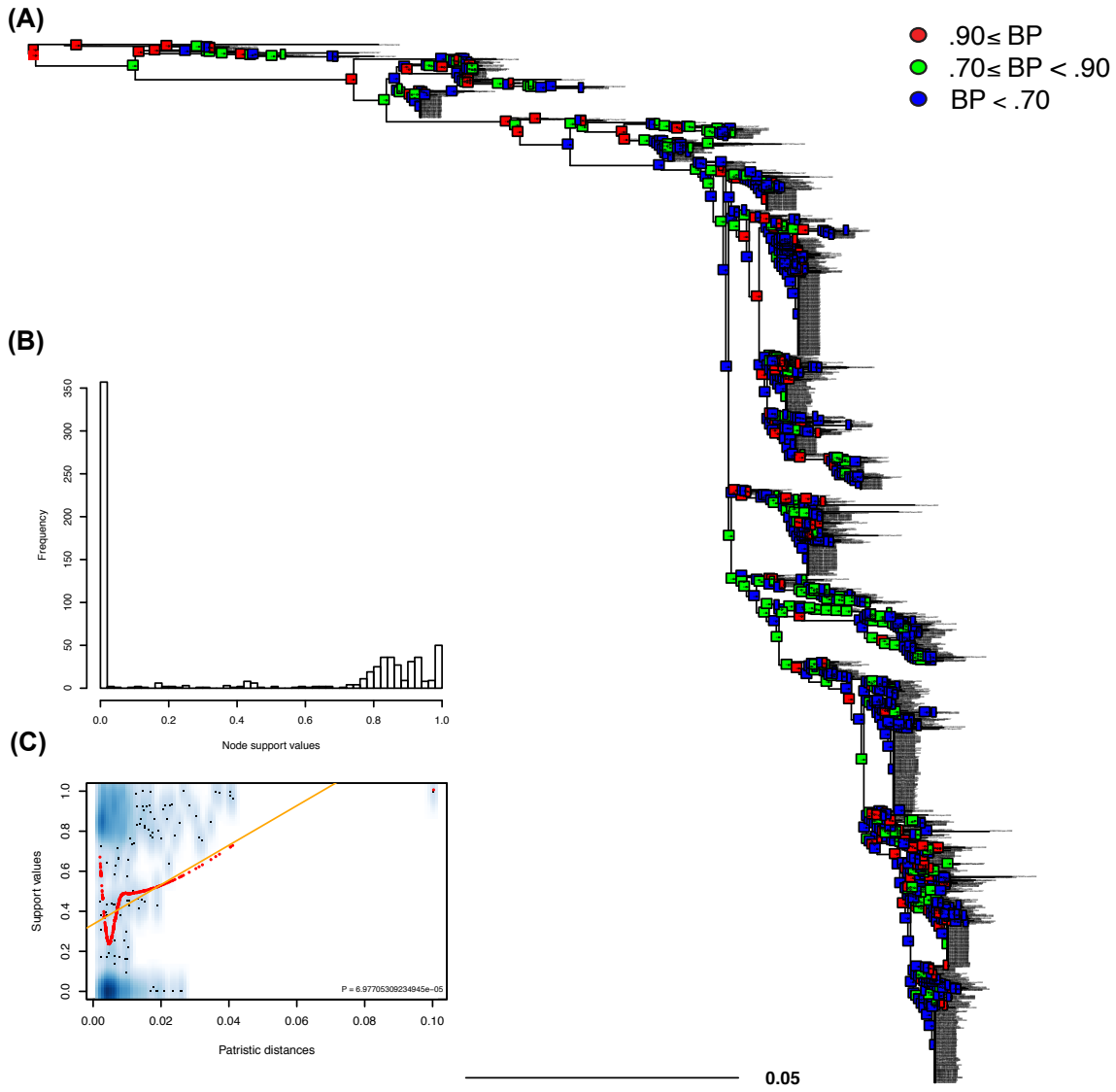

Figure A.2.10: Node support values for the phylogenetic analysis of the KDBP11 N1 data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

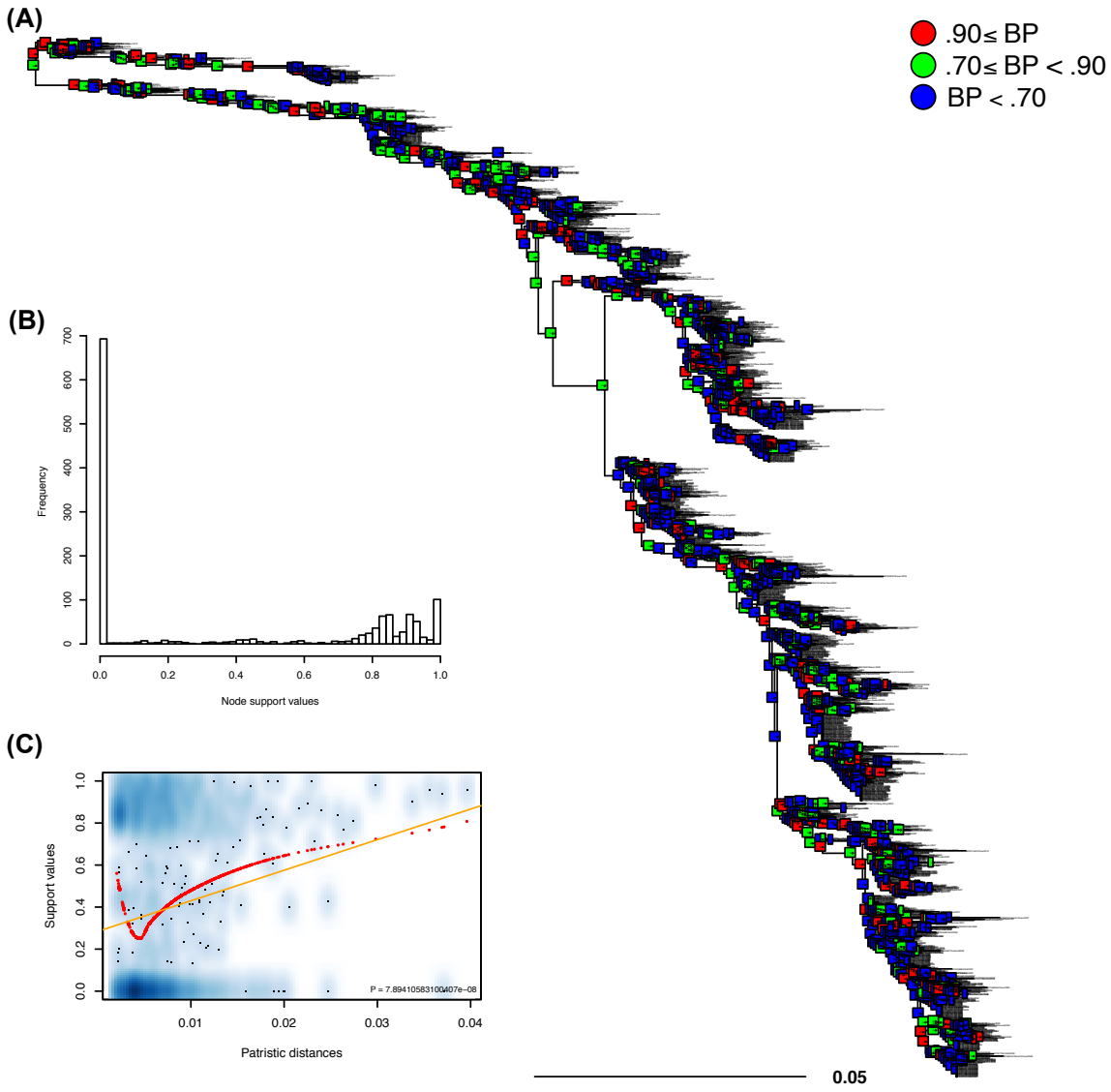

Figure A.2.11: Node support values for the phylogenetic analysis of the KDBP11 N2 data set. (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are colour-coded as shown in the top right corner of the panel. Scale bar is in expected number of substitutions per site. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

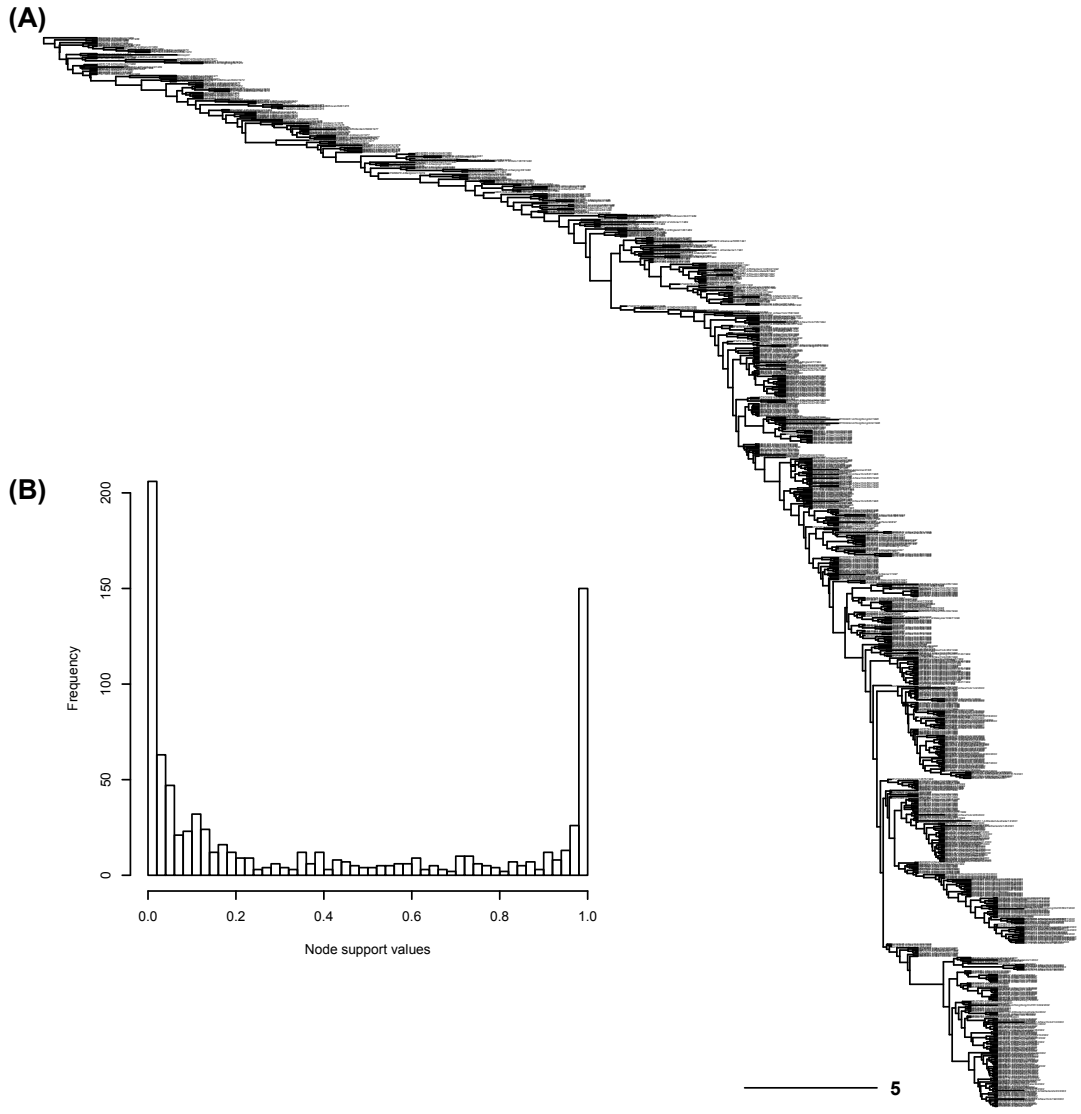

Figure A.2.12: Node support values for the phylogenetic analysis of the Koel13HA data set. (A) The estimated maximum a posteriori phylogenetic tree from the BEAST relaxed clock analysis. Scale bar is in years. (B) The distribution of node support values (posterior distributions) over the entire tree.

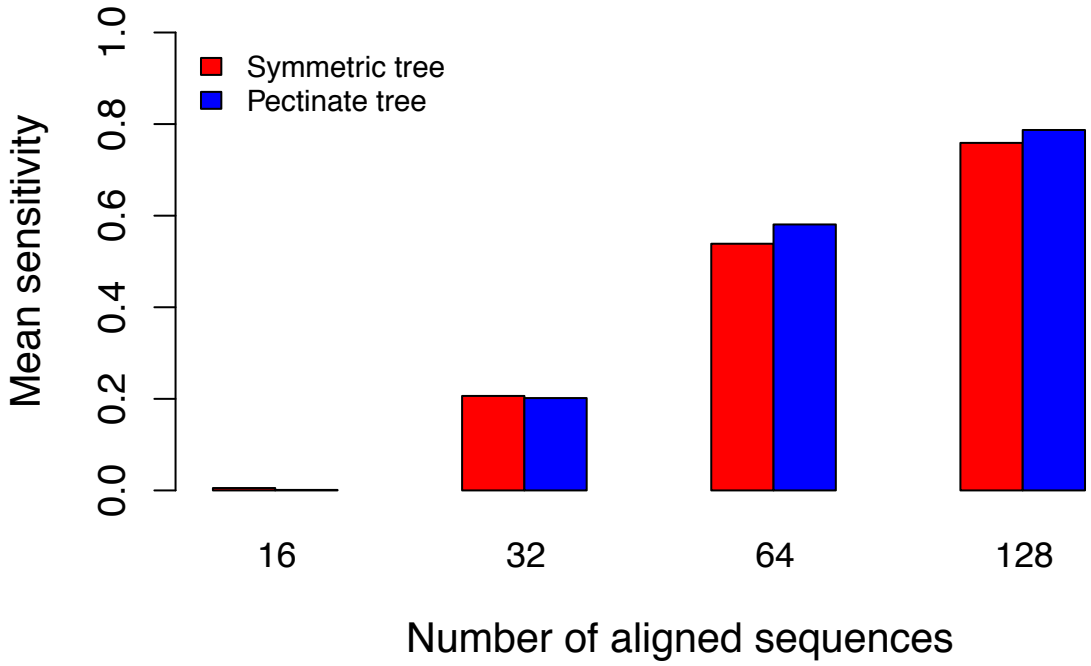

Figure A.2.13: The sensitivity to detect epistasis based on tree shape and number of aligned sequences. The figure presents the mean sensitivity across all analyses branch lengths. Since sensitivity varies for each branch lengths tested, errors bars are not presented as they distract from the trends. From these results we decided to ignore alignments with fewer than 16 sequences.

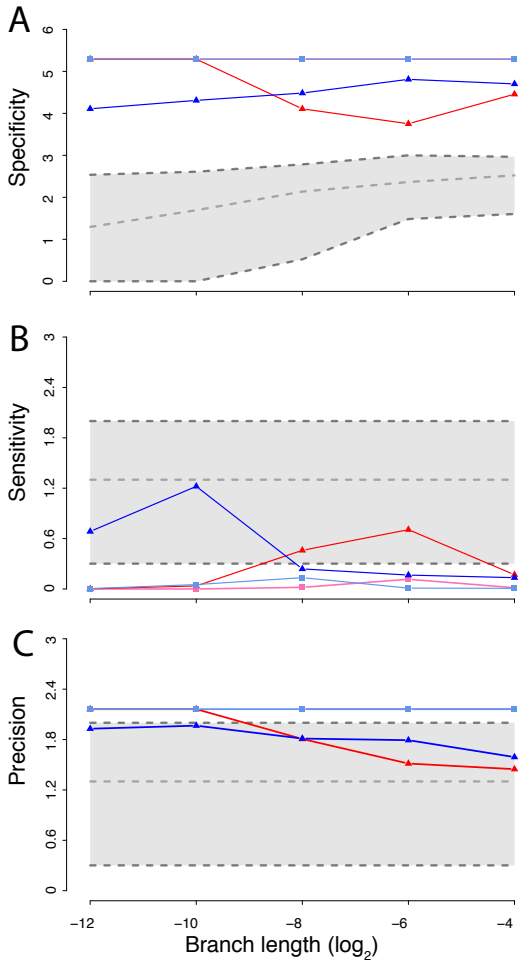

Figure A.2.14: Performance of our novel epistasis detection method, when dinucleotide bias is included in the simulation of independently evolving sites. This figure is to be compared with Fig. 2.4.1 (main text). Two tree shapes, symmetrically bifurcating (red) and pectinate (blue), were used by varying branch lengths (x-axis, on a $\log_2$ scale) and the number of aligned sequences (square: 32, and triangles:128 sequences). All y-axes show the mean of $-log_{10}(1 - summary\ statistic)$ to highlight performance of our method as values approach unity. When the mean of a summary statistic was unity (meaning our $log_{10}(1 - summary\ statistic)$ was infinity), we arbitrarily assigned it a value 10% larger than the largest value in the set. Panel A shows specificity (true negative rate); the polygon shaded in gray shows thresholds for excellent specificity. The thresholds were established by subtracting three (the number of epistatic pairs simulated) from the minimum (lower dashed line), mean (middle dashed line) and maximum (upper dashed line) number of pairwise comparisons performed across all simulations with the same branch length. These thresholds represent the number of calculable true negatives and allow us to demonstrate our method's excellent specificity. Panels B and C show sensitivity (true positive rate) and precision (positive predictive value) respectively. Polygons show the thresholds for 50% (lower dashed line), 95% (middle dashed line) and 99% (upper dashed line) detection.

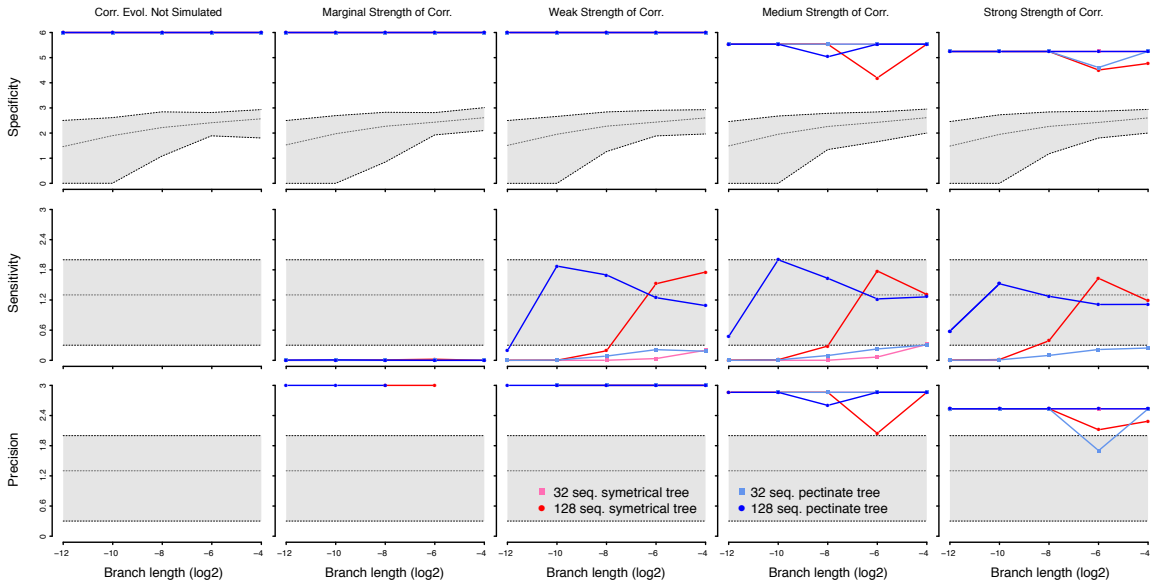

Figure A.2.15: Specificity, sensitivity and precision results of our novel epistasis detection method, when epistasis was simulated with Coev. This figure is to be compared with Fig. 2.4.1 (main text). Results for five levels of correlation are shown with $d/s \in \{1, 2, 33, 66, 100\}$. These levels represent no correlated evolution, marginal, weak, medium, and strong correlations, respectively. Our strong value ($d/s = 100$) was based on the default of Coev-web Dib *et al.* [2015]. Two tree shapes, symmetrically bifurcating (red) and pectinate (blue), were used by varying branch lengths (x-axis, on a $\log_2$ scale) and the number of aligned sequences (square: 32, circles: 64 and triangles:128 sequences). All y-axes show the mean of $-log_{10}(1 - summary\ statistic)$ to highlight performance of our method as values approach unity. When the mean of a summary statistic was unity (meaning our $log_{10}(1 - summary\ statistic)$ was infinity), we arbitrarily assigned it a value 10% larger than the largest finite value in the set. Panel (A) shows specificity (true negative rate); the polygon shaded in gray shows thresholds for excellent specificity. The thresholds were established by subtracting three (the number of epistatic pairs simulated) from the minimum (lower dashed line), mean (middle dashed line) and maximum (upper dashed line) number of pairwise comparisons performed across all simulations with the same branch length. These thresholds represent the number of calculable true negatives and allow us to demonstrate our method's excellent specificity. Panels (B) and (C) show sensitivity (true positive rate) and precision (positive predictive value) respectively. Polygons show the thresholds for 50% (lower dashed line), 95% (middle dashed line) and 99% (upper dashed line) detection.

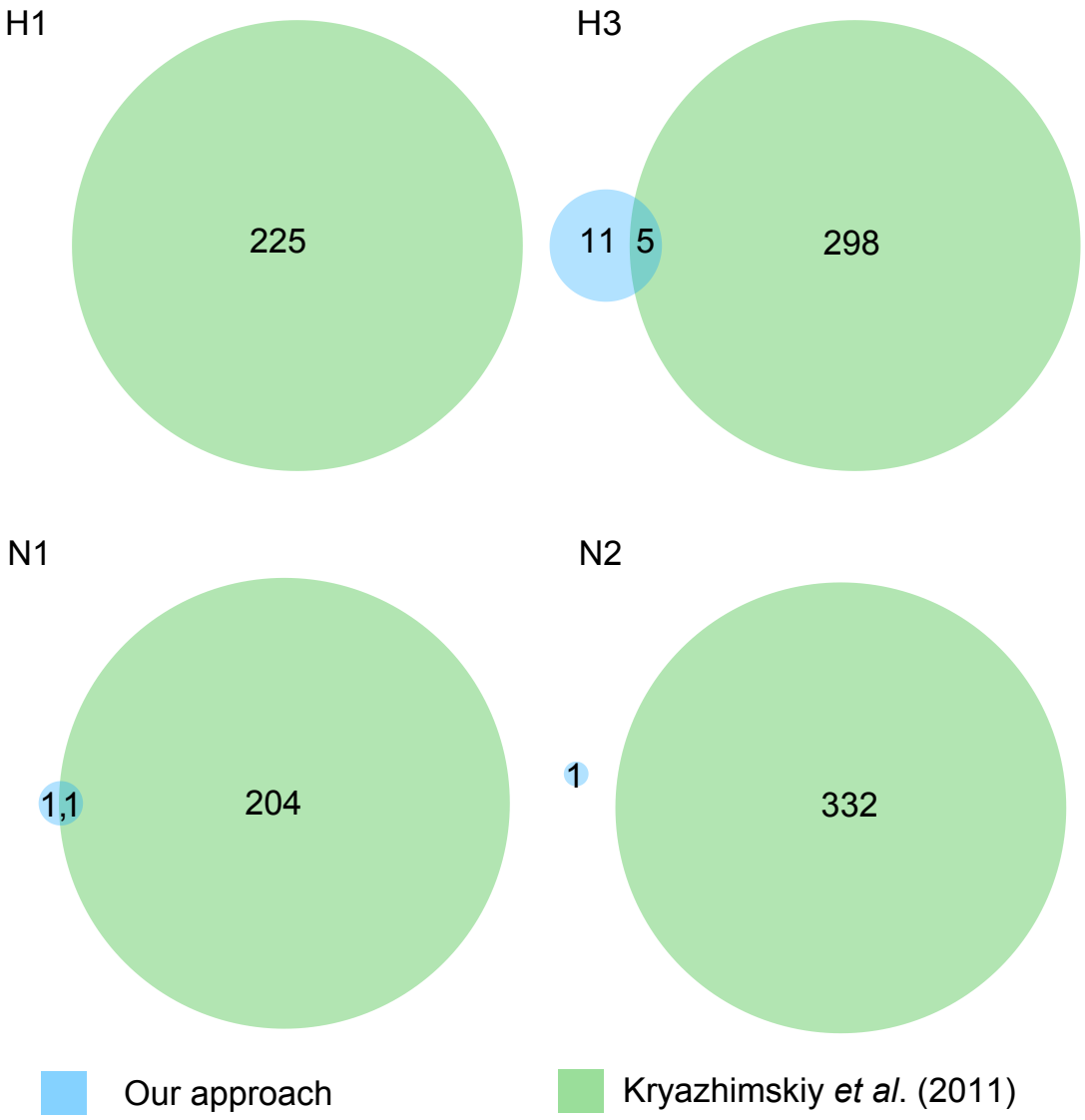

Figure A.2.16: Comparison with the method by Kryazhimskiy *et al.* (2011): after FDR. The numbers in the Venn diagrams show the total of pairs of epistatic sites detected by our method (in blue) and the previous method (in green) using the four large KDBP11 data sets. The diameter of each circle is indicative of the number of pairs of epistatic sites detected by each method.

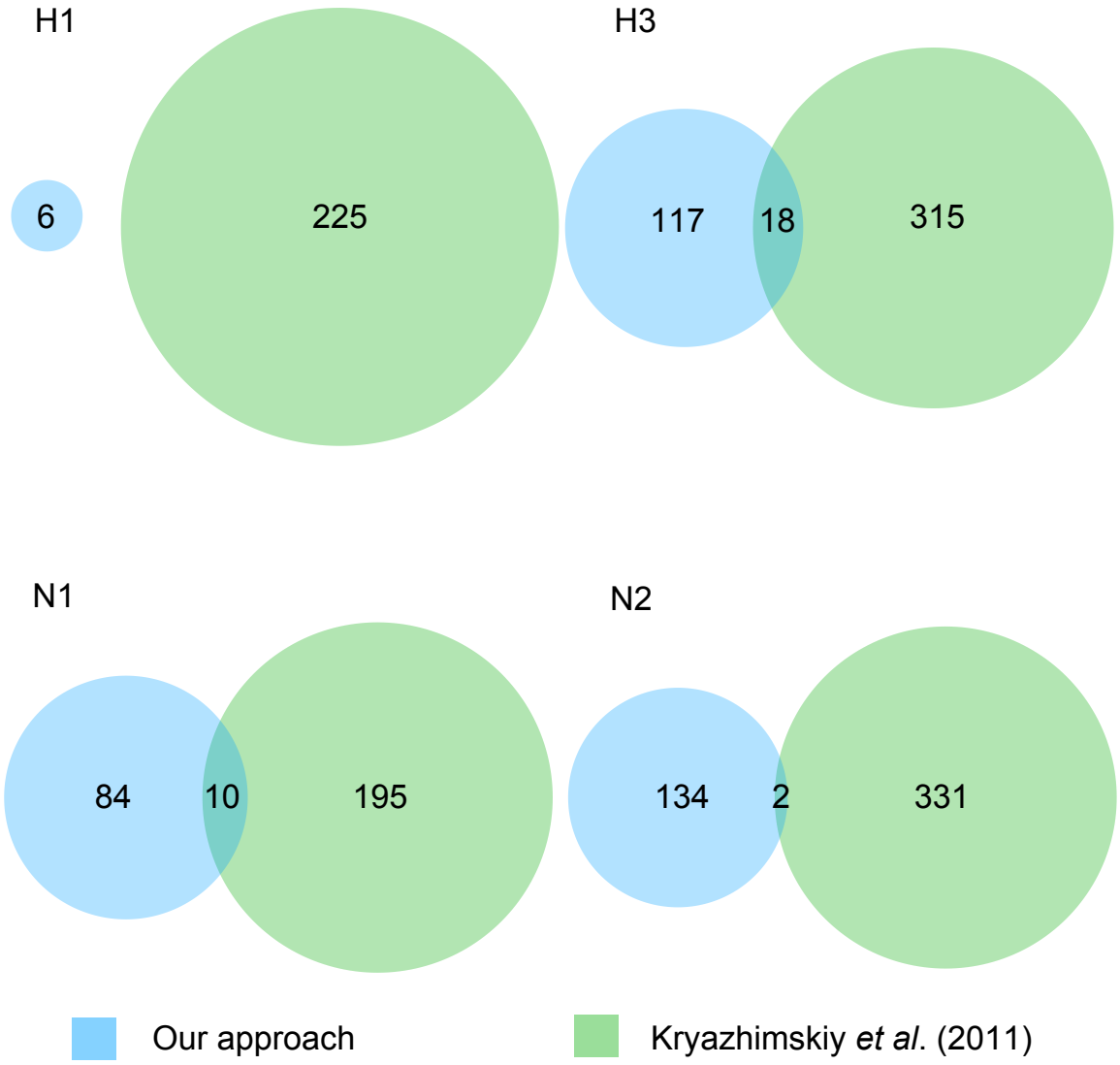

Figure A.2.17: Comparison with the method by Kryazhimskiy *et al.* (2011): before FDR. The numbers in the Venn diagrams show the total of pairs of epistatic sites detected by our method (in blue) and the previous method (in green) using the four large KDBP11 data sets. The diameter of each circle is indicative of the number of pairs of epistatic sites detected by each method.

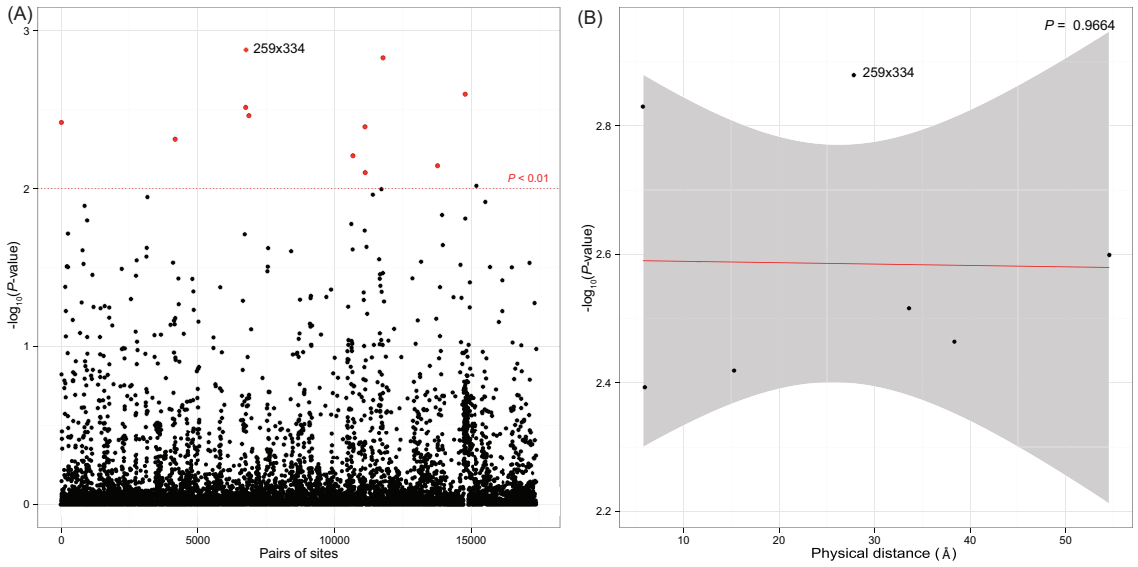

Figure A.2.18: The correlated sites detected in human influenza nucleoprotein and the physical distance among them. (A) Pairs of AA sites analyzed using the Gong13NP data set. Each pair is indexed using an arbitrary ordering. Pair of sites detected as epistatic with a $P < 0.01$ *before* FDR are shown in red. (B) The strength of the association is not related to physical distance (in Å) between sites of each epistatic pair. The significance of the regression (slope) is shown in the top right corner of the panel.

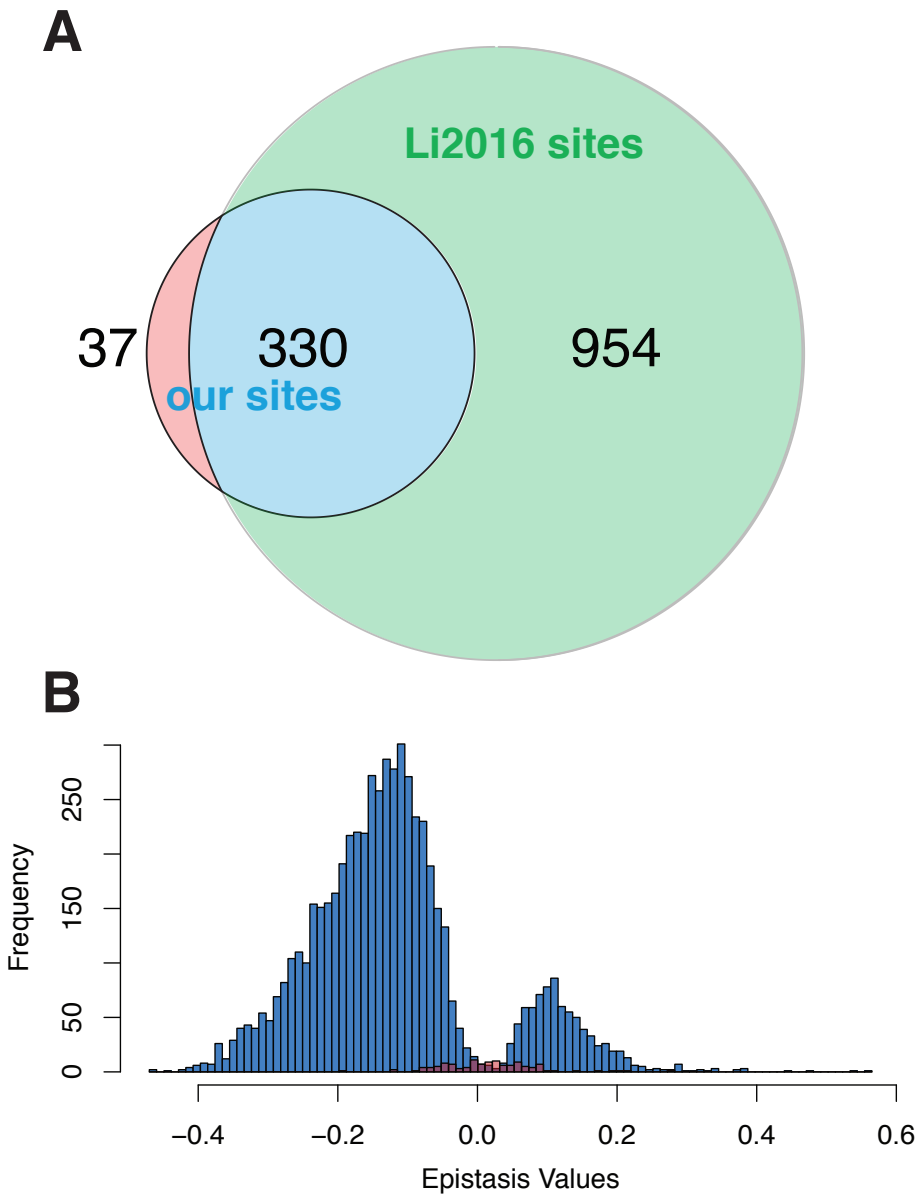

Figure A.2.19: Comparison of significant site pairs identified between our approach and Li *et al.* (2016). We analyzed an alignment of eukaryotic tRNA$^{Arg_{CCT}}$ genes and compared our predictions to the empirical evidence in Li *et al.* [2016]. (A) Comparison of significant site pairs detected by Li *et al.* [2016] and our analysis. (B) The distribution of epistasis values from all biological replicates in Li *et al.* [2016] which resulted in significant epistasis. Bars in red are for biological replicates belonging to the 37 pairs of sites that we (but not Li *et al.* [2016]) found to be significantly correlated. Significance values for both analyses used $\alpha = 0.05$.

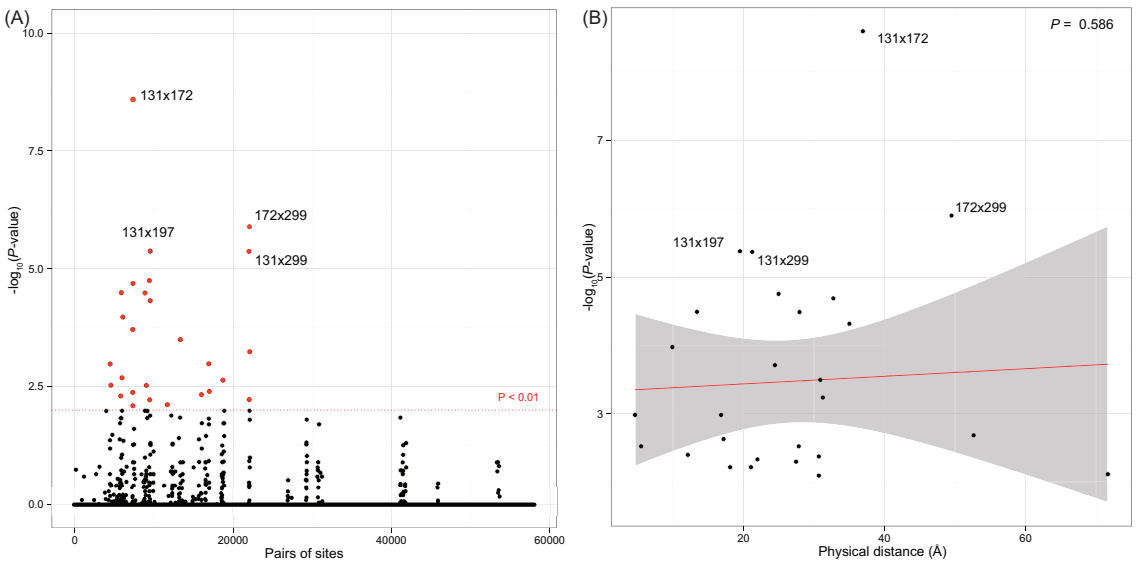

Figure A.2.20: The interacting sites detected in human influenza hemagglutinin and the physical distance among them. (A) Pairs of AA sites analyzed using the Koel13HA data set. Each pair is indexed using an arbitrary ordering. Pair of sites detected as epistatic with a $P < 0.01$ *after* FDR are shown in red. (B) The strength of the association is not related to physical distance (in Å) between sites of each epistatic pair. The significance of the regression (slope) is shown in the top right corner of the panel.

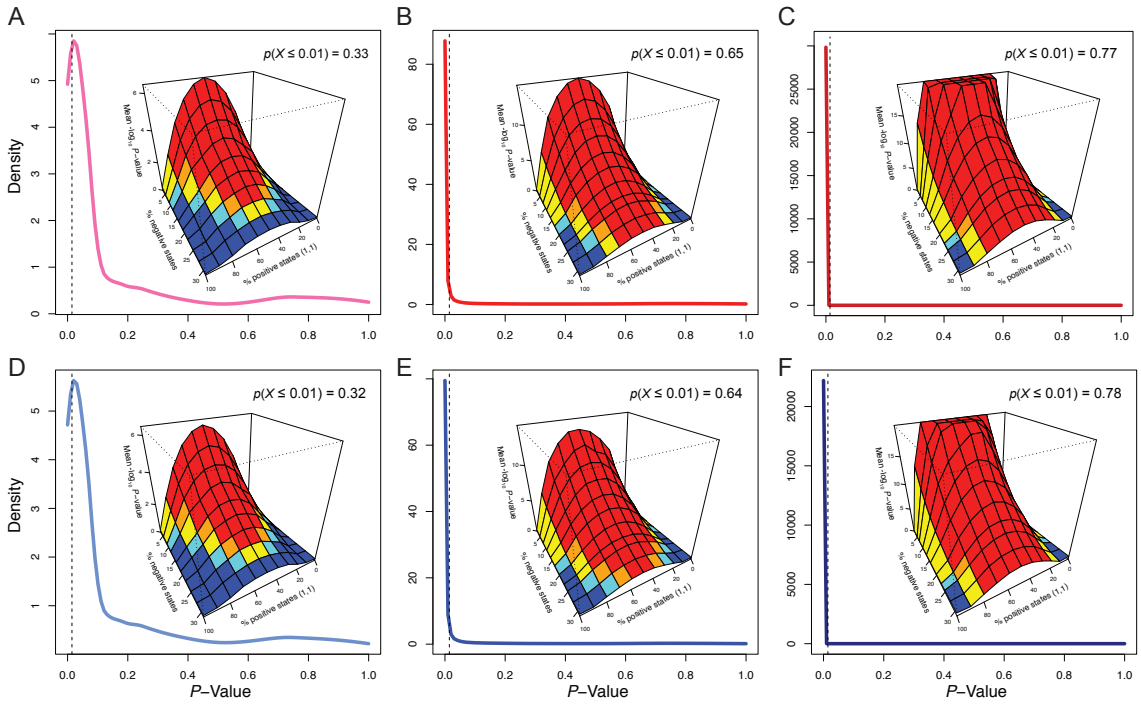

Figure A.2.21: Optimal signal strength for detecting epistasis. Two tree shapes, symmetrically bifurcating (panels A, B, C) and pectinate (panels D, E, F), were studied for alignments of 32 (panels A, D), 64 (panels B, E) and 128 (panels C, F) sequences. In each panel, the main plot shows the density of P-values; the dashed vertical black line shows the significance cutoff; the proportion of significant P-values is shown above each inset. Inset plots show significance ($-\log_{10} P$) plotted as a function of the proportion of identical positively correlated states and the proportion of total sequences that bear negatively correlated states. The proportion of significant results is colour-coded: red: 100%; orange: 75–100%; yellow: 50–75%; light blue: 25–50%; dark blue: 0–25%.

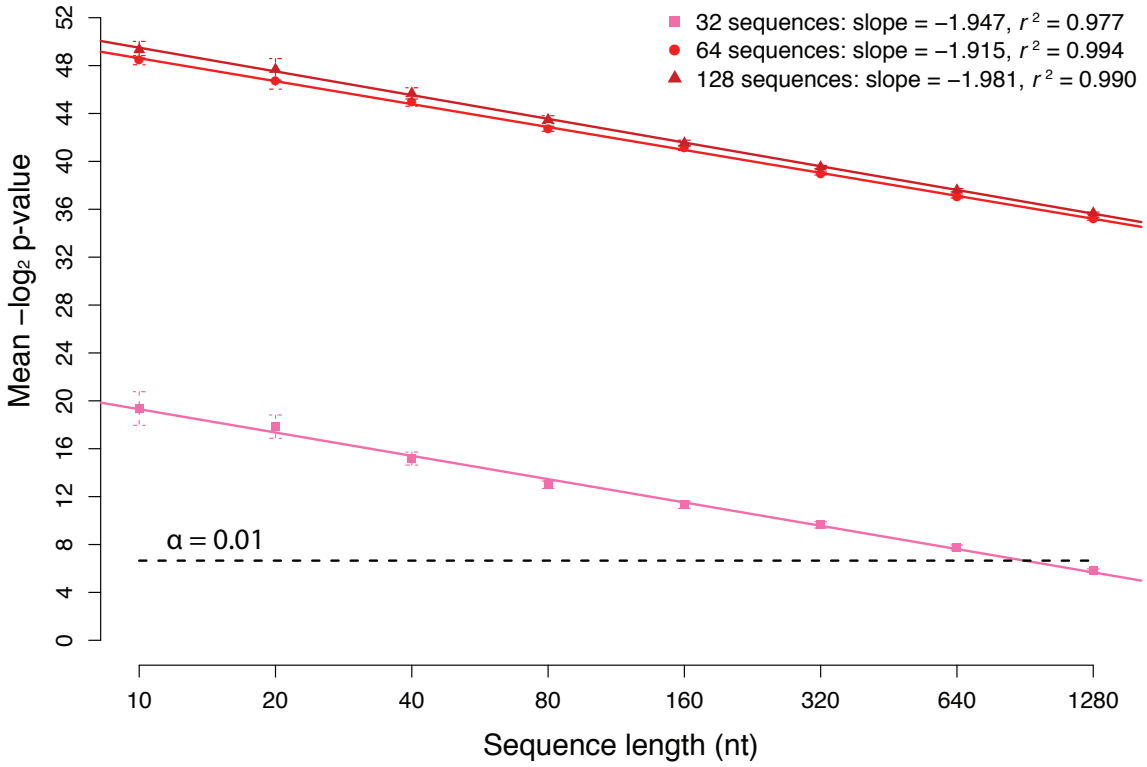

Figure A.2.22: The impact on detection of correlated evolution after false discovery rate (FDR) correction. This plot illustrates the relationship between the q-value (FDR-corrected P-value) and length of the sequence alignment being analyzed with our method. This analysis was done for the symmetrically bifurcating tree shape with 32, 64 and 128 sequences. The simulated epistatic signal was generated to match the most significant sequence distribution previously identified. The black horizontal dashed line represents a significance level of 0.01, below which results are no longer significant. Slopes and r^2 values of linear models are shown; all linear fits are highly significant.

A.3 Tables

Table A.3.1: ANOVA tables for the analysis of specificity, sensitivity, optimal phylogenetic distribution of epistatic pairs and sequence length. P-values in bold are significant. As design was fully factorial all possible interactions (colon notation) among factors were tested and are presented here.

	Df	**Sum Sq**	**Mean Sq**	F **value**	**Pr(>F)**
Branch length specificity					
τ	1	1.000E-09	9.810E-10	1.015	0.31373
n_s	1	2.000E-08	2.003E-08	20.736	**5.30E-06**
b	1	1.000E-09	1.325E-09	1.372	0.24150
τ:n_s	1	1.000E-09	1.185E-09	1.226	0.26814
τ:b	1	7.000E-09	7.434E-09	7.696	**0.00554**
n_s:b	1	4.000E-09	4.002E-09	4.143	0.04183
τ:n_s:b	1	2.100E-08	2.073E-08	21.460	**3.64E-06**
Residuals	18649	1.802E-05	9.660E-01		
Branch length sensitivity					
τ	1	0.3	0.3	4.151	0.041635
n_s	1	733.1	733.1	9224.143	**<2.23E-308**
b	1	17.2	17.2	216.730	**<2.23E-308**
τ:n_s	1	1.7	1.7	21.883	**2.94E-06**
τ:b	1	0.9	0.9	11.830	**0.000585**
n_s:b	1	0.0	0.0	0.000	0.991888
τ:n_s:b	1	0.4	0.4	4.729	0.029681
Residuals	9395	746.7	0.1		
Optimal phylogenetic distribution					

Continued on next page

Table A.3.1 – continued from previous page

Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
τ	1	36.36	36.36	2,208.72	**<2.23E-308**
n_s	1	5,232.70	5,232.70	317,874.80	**<2.23E-308**
+prop	10	130,007.73	13,000.77	789,768.31	**<2.23E-308**
-perc	10	5,672.48	567.25	34,459.10	**<2.23E-308**
τ:n_s	1	0.47	0.47	28.47	**9.50E-08**
τ:+prop	10	284.65	28.46	1,729.18	**<2.23E-308**
n_s:+prop	10	3,511.84	351.18	21,333.64	**<2.23E-308**
τ:-perc	10	122.51	12.25	744.20	**<2.23E-308**
n_s:-perc	10	1,052.12	105.21	6,391.40	**<2.23E-308**
+prop:-perc	100	8,152.99	81.53	4,952.76	**<2.23E-308**
τ:n_s:+prop	10	4.27	0.43	25.95	**5.52E-50**
τ:n_s:-perc	10	0.61	0.06	3.72	**5.24E-05**
τ:+prop:-perc	100	524.02	5.24	318.33	**<2.23E-308**
n_s:+prop:-perc	100	1,741.81	17.42	1,058.11	**<2.23E-308**
τ:n_s:+prop:-perc	100	17.02	0.17	10.34	**5.13E-155**
Residuals	2092516	34,445.96	0.02		
Effect of sequence length					
n_s	1	0.00	0.00	58.99	**4.20E-13**
l_s	1	0.00	0.00	174.58	**3.32E-30**
n_s:ls	1	0.00	0.00	199.52	**3.04E-33**
Residuals	236	0.00	0.00		

Notes—

Continued on next page

Table A.3.1 – continued from previous page

Df	Sum Sq	Mean Sq	F value	Pr($>F$)

τ: tree shape (symmetric or pectinate);

n_s: number of aligned sequences;

b: branch length;

+prop: balance of one positively correlated state;

-perc: percentage of total taxa with negatively correlated state pairs;

l_s: sequence length

Table A.3.2: Type III Sum of Squares and F tests for factors studied in our analysis of the phylogenetic distribution of epistatic pairs. A fully factorial experimental design was used. Likely due to the size of the data set all terms are significant (P-values in bold) so interpretation focuses on the F values to seek out terms that explain more variance.

	Df	Sum of Sq	RSS	AIC	F value	Pr($>F$)
τ	1	34.44	34,480.40	-8,592,813.91	2,092.30	$<$ **2.23E-308**
n_s	1	42.86	34,488.81	-8,592,303.24	2,603.43	$<$ **2.23E-308**
+prop	10	1,114.63	35,560.59	-8,528,269.41	6,771.13	$<$ **2.23E-308**
-perc	10	1.08	34,447.04	-8,594,857.90	6.57	**2.93E-10**
τ:n_s	1	0.29	34,446.25	-8,594,887.78	17.87	**2.36E-05**
τ:+prop	10	54.52	34,500.48	-8,591,613.49	331.20	$<$ **2.23E-308**
n_s:+prop	10	58.45	34,504.41	-8,591,375.04	355.08	$<$ **2.23E-308**
τ:-perc	10	57.70	34,503.66	-8,591,420.69	350.51	$<$ **2.23E-308**
n_s:-perc	10	31.95	34,477.91	-8,592,983.25	194.09	$<$ **2.23E-308**
+prop:-perc	100	284.16	34,730.12	-8,577,908.39	172.62	$<$ **2.23E-308**
τ:n_s:+prop	10	6.96	34,452.92	-8,594,500.91	42.27	**1.42E-84**
τ:n_s:-perc	10	2.13	34,448.09	-8,594,793.99	12.96	**5.56E-23**
τ:+prop:-perc	100	69.43	34,515.39	-8,590,889.13	42.18	$<$ **2.23E-308**
n_s:+prop:-perc	100	147.44	34,593.40	-8,586,163.99	89.57	$<$ **2.23E-308**
τ:n_s:+prop:-perc	100	17.02	34,462.98	-8,594,069.54	10.34	**5.13E-155**

Notes—τ: tree shape (symmetric or pectinate); n_s: number of aligned sequences; +prop: balance of one positively correlated state; -perc: percentage of total taxa with negatively correlated state pairs.

Table A.3.3: Type III Sum of Squares and F tests for differences in sensitivity between our full simulation study (main text), and dinucleotide-biased simulations. Both simulation studies used the same model (RNA7D) to simulated epistatic sites, thus a difference was not expected. Significance was found in a subset of our original simulation study when parameters, other than model for independent site evolution, were identical.

	Sum Sq	Df	F value	$\Pr(>F)$
(Intercept)	0.09	1	1.0948	0.29548
Independent Site Model	29.09	1	336.8009	$<$ **2.2E-16**
τ	0.24	1	2.7449	0.09764
b	0.08	1	0.8956	0.34402
n_s	1.53	1	17.7263	**2.609E-05**
τ:b	0.01	1	0.1693	0.68077
τ:n_s	7.62	1	88.2200	$<$ **2.2E-16**
b:n_s	3.12	1	36.0820	**2.066E-09**
τ:b:n_s	3.54	1	40.9376	**1.759E-10**
Residuals	335.48	3884		

Notes—τ: tree shape (symmetric or pectinate); n_s: number of aligned sequences; b: branch length.

Table A.3.4: Detailed results generated using the KDBP11 data set for HA in H1N1 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	**site-2**	**lkl-indep**	**lkl-dep**	**test-stat**	P**-value**	P**-value (FDR)**
71	73	-38.485092	-29.055355	18.859474	0.000837511178384709	0.984913145780418
12	73	-58.238767	-49.80791	16.861714	0.00205625270630705	0.999999998318065
73	74	-56.071941	-48.042412	16.059058	0.00294093184608624	0.999999998318065
11	73	-35.737211	-28.307707	14.859008	0.00500275725669519	0.999999998318065
41&45&54&84	65	-23.662771	-16.698284	13.928974	0.00752522577487336	0.999999998318065
73	88	-29.379655	-22.670017	13.419276	0.00939886101619647	0.999999998318065

Table A.3.5: Detailed results generated using the KDBP11 data set for NA in H1N1 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
275	354	-151.206402	-122.116619	58.179566	7.00E-12	2.34E-07
461	466	-80.750412	-63.80729	33.886244	7.86E-07	0.0131356962707871
234	275	-140.234912	-128.581571	23.306682	0.00010995716148976	0.9999999999955
50	286	-56.127815	-45.461299	21.333032	0.000271976976207289	0.9999999999955
200	388	-45.296802	-35.134716	20.324172	0.000430930754599101	0.9999999999955
254	270	-64.855428	-54.755728	20.1994	0.000456107550876772	0.9999999999955
249	275	-155.496669	-145.664956	19.663426	0.000581886713368229	0.9999999999955
336	427	-55.900157	-46.137891	19.524532	0.000619734205504541	0.9999999999955
70	466	-51.543637	-41.860559	19.366156	0.000665869398006613	0.9999999999955
435	461	-85.07601	-75.485225	19.18157	0.000723940959468261	0.9999999999955
136	270	-65.039998	-55.451621	19.176754	0.000725521312396338	0.9999999999955
23	274	-71.48771	-61.947853	19.079714	0.000758101690989998	0.9999999999955
34	336	-41.971628	-32.49925	18.944756	0.00080583070915563	0.9999999999955
200	336	-32.754213	-23.344507	18.819412	0.000852815473945556	0.9999999999955
52	466	-49.232332	-40.012794	18.439076	0.00101259734365267	0.9999999999955
105	466	-41.824315	-32.72081	18.20701	0.00112426857912984	0.9999999999955
263	336	-61.402675	-52.397998	18.009354	0.00122891413529491	0.9999999999955
357	383	-36.893395	-27.896693	17.993404	0.00123776646505713	0.9999999999955
200	274	-26.322141	-17.541908	17.560466	0.00150363511573715	0.9999999999955
336	388	-58.414566	-49.684818	17.459496	0.00157333138969318	0.9999999999955
45	275	-140.885071	-132.221152	17.327838	0.00166901818191334	0.9999999999955
220	336	-55.603238	-46.94017	17.326136	0.00167029202285462	0.9999999999955
200	466	-45.927375	-37.268853	17.317044	0.00167711308491059	0.9999999999955
270	396	-86.099665	-77.448379	17.302572	0.00168802709857463	0.9999999999955
23	452	-111.291565	-102.649438	17.284254	0.00170194209259467	0.9999999999955
248	466	-118.425817	-109.808757	17.23412	0.00174060696636791	0.9999999999955
383	466	-46.488327	-37.899067	17.17852	0.00178450131560515	0.9999999999955
93	275	-131.342994	-122.778536	17.128916	0.00182458253941953	0.9999999999955
275	396	-150.982066	-142.567216	16.8297	0.00208588559096545	0.9999999999955
23	173	-72.887506	-64.473463	16.828086	0.00208739064264929	0.9999999999955
334	466	-53.692311	-45.366264	16.652094	0.00225809043765124	0.9999999999955
143	467	-22.125427	-13.864434	16.521986	0.00239306161668063	0.9999999999955
382	466	-40.933061	-32.681373	16.503376	0.00241300606556916	0.9999999999955
332	467	-13.418112	-5.178711	16.478802	0.00243959327709231	0.9999999999955
173	275	-121.008318	-112.792149	16.432338	0.00249065498094114	0.9999999999955
52	461	-57.912071	-49.789049	16.246044	0.00270617009931429	0.9999999999955
390	467	-27.201104	-19.132944	16.13632	0.00284159256104222	0.9999999999955
454	466	-58.950808	-50.92717	16.047276	0.00295637796235626	0.9999999999955
34	287	-36.310054	-28.313782	15.992544	0.00302918482382142	0.9999999999955

Continued on next page

Table A.3.5 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
86	467	-27.507537	-19.537999	15.939076	0.00310201359905926	0.9999999999955
105	461	-50.425315	-42.465231	15.920168	0.00312817848327285	0.9999999999955
23	220	-87.652607	-79.760612	15.78399	0.00332314529606204	0.9999999999955
15	200	-40.968346	-33.088335	15.760022	0.00335867675914581	0.9999999999955
130	331	-68.45982	-60.657796	15.604048	0.00359920495061494	0.9999999999955
383	461	-55.168084	-47.384994	15.56618	0.0036601113151743	0.9999999999955
388	466	-71.514138	-63.817153	15.39397	0.00395012290198515	0.9999999999955
15	274	-47.6502	-39.963419	15.373562	0.00398595416950454	0.9999999999955
17&163	466	-39.08149	-31.412476	15.338028	0.00404910689811722	0.9999999999955
336	366&367	-27.081351	-19.422294	15.318114	0.00408492800602434	0.9999999999955
232	467	-32.514391	-24.881063	15.266656	0.00417893881066467	0.9999999999955
16	466	-58.934043	-51.335155	15.197776	0.00430811378552565	0.9999999999955
334	461	-62.37204	-54.783778	15.176524	0.00434875529997469	0.9999999999955
254	336	-49.567648	-41.996368	15.14256	0.0044144902055856	0.9999999999955
101	336	-35.215942	-27.696958	15.037968	0.00462311335217314	0.9999999999955
57	466	-66.016384	-58.50624	15.020288	0.0046593229258225	0.9999999999955
82	130	-71.473515	-63.99239	14.96225	0.00478015697097867	0.9999999999955
262	274	-49.90798	-42.455595	14.90477	0.00490286108222038	0.9999999999955
307	466	-69.509585	-62.078536	14.862098	0.00499594947834114	0.9999999999955
23	275	-158.091078	-150.690456	14.801244	0.00513171047589567	0.9999999999955
275	352	-137.045128	-129.652756	14.784744	0.00516914075797925	0.9999999999955
14	336	-35.207757	-27.822389	14.770736	0.00520112818749763	0.9999999999955
200	250	-31.51041	-24.131537	14.757746	0.0052309646028319	0.9999999999955
84	287	-51.4864	-44.112225	14.74835	0.00525265069975567	0.9999999999955
95	396	-86.471666	-79.157217	14.628898	0.00553616002222657	0.9999999999955
106	466	-54.834282	-47.543352	14.58186	0.00565187543337498	0.9999999999955
70	461	-60.223376	-52.946299	14.554154	0.00572114045480998	0.9999999999955
15	388	-66.626424	-59.386456	14.479936	0.00591081509238112	0.9999999999955
34	386	-34.010318	-26.778058	14.46452	0.00595097969768998	0.9999999999955
211	274	-28.759868	-21.562656	14.394424	0.00613700829742769	0.9999999999955
270	434	-62.234339	-55.133788	14.201102	0.00668012274939944	0.9999999999955
336	418	-47.091265	-40.052959	14.076612	0.00705451363202247	0.9999999999955
275	418	-127.272988	-120.235474	14.075028	0.00705940740052502	0.9999999999955
15	262	-64.553026	-57.518177	14.069698	0.00707589887701154	0.9999999999955
15	250	-52.840018	-45.970497	13.739042	0.00817623254556632	0.9999999999955
427	466	-68.999729	-62.132703	13.734052	0.0081940589981907	0.9999999999955
52	105	-18.865615	-12.01964	13.69195	0.00834598796723063	0.9999999999955
23	336	-77.920897	-71.082516	13.676762	0.00840146912942918	0.9999999999955
382	461	-49.517328	-42.720275	13.594106	0.00870979383594128	0.9999999999955
173	274	-34.404951	-27.611158	13.587586	0.00873458058217291	0.9999999999955
220	466	-68.70281	-61.910932	13.583756	0.00874917305685707	0.9999999999955
79	201	-56.900371	-50.132123	13.536496	0.00893121045800649	0.9999999999955
27	200	-17.748473	-10.982961	13.531024	0.00895252549151837	0.9999999999955
95	275	-141.70149	-134.940735	13.52151	0.00898970381832009	0.9999999999955
95	270	-76.81909	-70.063706	13.510768	0.00903186238930109	0.9999999999955
287	466	-53.307035	-46.564777	13.484516	0.00913570767452088	0.9999999999955

Continued on next page

Table A.3.5 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
34	250	-40.728247	-33.988081	13.480332	0.00915236589278601	0.9999999999955
220	386	-47.628595	-40.892513	13.472164	0.00918497169373855	0.9999999999955
64	275	-111.303341	-104.594838	13.417006	0.0094081495237065	0.9999999999955
6	45	-54.820441	-48.117229	13.406424	0.00945156809500369	0.9999999999955
93	95	-66.832594	-60.144784	13.37562	0.00957907695899363	0.9999999999955
23	254	-81.617017	-74.937806	13.358422	0.00965099565284477	0.9999999999955
130	344	-88.20565	-81.54131	13.32868	0.00977661863990886	0.9999999999955
101	466	-48.315515	-41.66894	13.29315	0.0099287850007318	0.9999999999955
14	466	-48.307329	-41.663856	13.286946	0.00995559156861858	0.9999999999955

Table A.3.6: Detailed results generated using the KDBP11 data set for HA in H3N2 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
172	212	-95.260701	-66.640486	57.24043	1.10E-11	7.77E-07
172	278	-86.257381	-59.220203	54.074356	5.08E-11	1.79E-06
172	213	-102.53157	-76.376144	52.310852	1.19E-10	2.77E-06
172	242	-90.576939	-64.712052	51.729774	1.57E-10	2.77E-06
172	206	-77.13814	-52.08129	50.1137	3.42E-10	4.82E-06
172	292	-79.869861	-55.371555	48.996612	5.85E-10	6.87E-06
149	172	-142.326561	-118.68048	47.292162	1.33E-09	1.33E-05
137	172	-98.903071	-75.740267	46.325608	2.11E-09	1.86E-05
158	172	-126.00687	-105.252318	41.509104	2.11E-08	0.000165161699520559
172	315	-89.372521	-69.349836	40.04537	4.24E-08	0.00027882614746233
98	172	-82.689173	-62.746316	39.885714	4.57E-08	0.00027882614746233
151	172	-87.87812	-67.97483	39.80658	4.75E-08	0.00027882614746233
172	173	-140.454483	-121.475613	37.95774	1.14E-07	0.000620009254642246
73	172	-103.779205	-85.255579	37.047252	1.76E-07	0.000886978319149443
172	208	-135.618671	-117.172459	36.892424	1.90E-07	0.000890932476549366
110	172	-75.266596	-58.856524	32.820144	1.30E-06	0.00572883065011465
140	172	-107.194851	-92.150417	30.088868	4.69E-06	0.0194692816187225
172	468	-106.731919	-93.090927	27.281984	1.74E-05	0.0682778067287818
172	202	-142.37568	-129.171471	26.408418	2.62E-05	0.0971264383410068
147	236	-133.393355	-122.281267	22.224176	0.000180842506048995	0.637469833822707
236	402	-157.755002	-146.913319	21.683366	0.000231702291790925	0.777857693869534
11	172	-85.524571	-74.819616	21.40991	0.000262583155412055	0.817110082323141
218	236	-131.09927	-120.410813	21.376914	0.000266574920474216	0.817110082323141
149	188	-99.854232	-89.369955	20.968554	0.00032124768836983	0.943665084586376

Continued on next page

Table A.3.6 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
172	241	-86.603477	-76.275782	20.65539	0.000370578823137624	0.9999999999755
236	241	-123.959968	-113.711205	20.497526	0.00039822044147797	0.9999999999755
171	236	-118.15418	-107.907915	20.49253	0.000399127786084086	0.9999999999755
171	245	-157.753608	-147.576446	20.354324	0.000425056094516707	0.9999999999755
154	213	-197.645557	-187.505926	20.279262	0.000439830166879762	0.9999999999755
171	209	-80.154649	-70.265916	19.777466	0.000552529122753942	0.9999999999755
2	377	-132.82316	-123.057211	19.531898	0.000617667224629415	0.9999999999755
236	238	-118.153036	-108.440276	19.42552	0.000648191457816139	0.9999999999755
208	262	-151.535949	-141.930562	19.210774	0.000714430439784697	0.9999999999755
2	91	-138.037954	-128.438014	19.19988	0.000717963619603745	0.9999999999755
243	245	-160.870729	-151.288714	19.16403	0.000729713083036265	0.9999999999755
236	243	-121.367667	-111.865979	19.003376	0.000784744795860992	0.9999999999755
172	402	-120.398298	-110.939689	18.917218	0.000815928741833449	0.9999999999755
153	205	-24.643147	-15.201595	18.883104	0.000828611814868485	0.9999999999755
172	466	-157.397841	-147.982333	18.831016	0.00084835434424424	0.9999999999755
172	238	-80.796828	-71.453865	18.685926	0.000905832076696589	0.9999999999755
15	233	-70.853696	-61.536033	18.635326	0.000926769459038801	0.9999999999755
172	546	-189.734083	-180.459916	18.548334	0.000963889098073767	0.9999999999755
147	209	-95.395213	-86.201372	18.387682	0.00103634281682119	0.9999999999755
91	236	-145.73871	-136.637576	18.202268	0.00112667292298252	0.9999999999755
69	205	-121.873963	-112.777017	18.193892	0.00113093223673189	0.9999999999755
236	546	-227.090789	-218.053187	18.075204	0.00119302430998547	0.9999999999755
41	236	-125.973027	-116.954557	18.03694	0.00121375112605182	0.9999999999755
172	175	-88.807657	-79.810745	17.993824	0.00123753255760162	0.9999999999755
79	153	-33.498754	-24.656956	17.683596	0.00142275580259887	0.9999999999755
15	159	-37.517793	-28.759554	17.516478	0.00153361582691103	0.9999999999755
41	172	-88.616345	-79.881762	17.469166	0.0015665207568073	0.9999999999755
65	122	-102.978119	-94.30307	17.350098	0.00165244594718095	0.9999999999755
172	262	-153.063464	-144.390213	17.346502	0.00165511207631241	0.9999999999755
79	99	-25.345169	-16.69659	17.297158	0.00169212803581087	0.9999999999755
149	154	-237.440547	-228.843112	17.19487	0.00177148141516981	0.9999999999755
73	108	-115.35578	-106.801142	17.109276	0.00184069638482554	0.9999999999755
171	172	-80.797011	-72.251506	17.09101	0.00185580885609293	0.9999999999755
91	209	-107.740568	-99.199889	17.081358	0.00186384383996996	0.9999999999755
15	16&294	-36.476618	-27.980262	16.992712	0.0019392618772972	0.9999999999755
69	79	-130.729512	-122.265429	16.928166	0.00199606187258083	0.9999999999755
172	218	-93.743876	-85.28066	16.926432	0.00199761019179456	0.9999999999755
172	391	-128.171689	-119.717041	16.909296	0.00201297504854159	0.9999999999755
41	209	-87.974885	-79.53564	16.87849	0.00204089059165746	0.9999999999755
16&294	153	-23.713307	-15.294666	16.837282	0.00207882969689865	0.9999999999755
91	245	-184.795639	-176.381198	16.828882	0.0020866482405304	0.9999999999755
241	245	-163.739881	-155.336381	16.807	0.00210715192978539	0.9999999999755
16&294	69	-120.944089	-112.559717	16.768744	0.00214347647400648	0.9999999999755
154	212	-190.374688	-182.016027	16.717322	0.00219327641052169	0.9999999999755
2	209	-166.160652	-157.850034	16.621236	0.00228940680823264	0.9999999999755
41	245	-165.029956	-156.72176	16.616392	0.00229436147719242	0.9999999999755
172	188	-94.673858	-86.429946	16.487824	0.00242979879049932	0.9999999999755

Continued on next page

Table A.3.6 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
245	402	-196.811909	-188.679286	16.265246	0.00268313322906089	0.9999999999755
236	391	-165.528372	-157.46435	16.128044	0.0028520733753099	0.9999999999755
238	245	-158.160035	-150.114325	16.09142	0.00289891315078628	0.9999999999755
236	377	-140.523917	-132.487123	16.073588	0.00292199249379554	0.9999999999755
16&294	79	-24.885007	-16.863386	16.043242	0.00296168483869053	0.9999999999755
2	243	-115.11334	-107.124119	15.978442	0.00304822810272354	0.9999999999755
69	99	-121.404162	-113.41741	15.973504	0.00305492417351449	0.9999999999755
15	79	-46.26147	-38.306894	15.909152	0.00314352243424154	0.9999999999755
91	172	-108.382028	-100.430462	15.903132	0.0031519388392941	0.9999999999755
99	153	-24.173396	-16.232303	15.882186	0.00318139608280876	0.9999999999755
202	209	-141.73422	-133.800234	15.867972	0.0032015399706149	0.9999999999755
15	153	-45.089704	-37.168746	15.841916	0.00323879264171545	0.9999999999755
218	245	-170.86688	-163.033675	15.66641	0.00350106845546394	0.9999999999755
15	363	-58.456346	-50.681616	15.54946	0.00368732505795455	0.9999999999755
158	191	-89.091102	-81.393364	15.395476	0.00394749133007244	0.9999999999755
73	391	-94.804708	-87.148318	15.31278	0.00409457552700854	0.9999999999755
237	546	-195.14541	-187.5139	15.26302	0.00418566143234478	0.9999999999755
147	245	-172.450284	-164.83437	15.231828	0.00424377090065453	0.9999999999755
154	242	-186.387587	-178.823874	15.127426	0.00444409450263472	0.9999999999755
69	142	-122.232786	-114.672781	15.12001	0.00445867239861419	0.9999999999755
65	545	-130.41267	-122.858469	15.108402	0.00448158503491347	0.9999999999755
209	546	-189.092623	-181.581983	15.02128	0.00465728390114206	0.9999999999755
128	313	-82.258413	-74.770147	14.976532	0.00475013965260518	0.9999999999755
142	153	-25.001957	-17.553857	14.8962	0.00492141870197038	0.9999999999755
73	173	-107.087502	-99.705605	14.763794	0.00521705225442526	0.9999999999755
15	205	-37.405868	-30.030795	14.750146	0.00524849869616506	0.9999999999755
172	235	-141.653506	-134.310796	14.68542	0.00540018592006486	0.9999999999755
79	205	-25.814931	-18.491374	14.647114	0.00549197516844835	0.9999999999755
15	264	-36.099708	-28.795231	14.608954	0.00558493640637592	0.9999999999755
15	69	-142.320462	-135.018307	14.60431	0.00559635447213735	0.9999999999755
15	99	-36.936173	-29.656677	14.558992	0.00570898572169087	0.9999999999755
73	188	-61.306877	-54.083756	14.446242	0.0059989478750111	0.9999999999755
188	213	-60.059242	-52.839282	14.43992	0.0060156272313775	0.9999999999755
172	179	-78.89473	-71.708123	14.373214	0.00619441330064496	0.9999999999755
154	217	-231.698276	-224.538646	14.31926	0.00634282405485831	0.9999999999755
212	262	-111.177979	-104.037812	14.280334	0.00645205687136396	0.9999999999755
91	402	-91.63414	-84.524912	14.218456	0.00662950374489812	0.9999999999755
209	218	-93.1015	-86.026803	14.149394	0.00683320368361695	0.9999999999755
151	213	-53.263503	-46.203451	14.120104	0.0069214361698553	0.9999999999755
138	223	-33.911451	-26.881617	14.059668	0.00710703496939846	0.9999999999755
15	229	-43.520539	-36.496346	14.048386	0.00714221807017035	0.9999999999755
360	366	-23.046363	-16.058858	13.97501	0.00737524206945139	0.9999999999755
209	402	-119.756839	-112.786258	13.941162	0.00748522941595586	0.9999999999755
2	238	-112.256468	-105.303071	13.906794	0.00759854979001506	0.9999999999755

Continued on next page

Table A.3.6 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
15	176	-38.773334	-31.830015	13.886638	0.0076657892646832	0.9999999999755
15	179	-39.249057	-32.318012	13.86209	0.00774846751649405	0.9999999999755
554	562&565	-35.843729	-28.949767	13.787924	0.00800360001575418	0.9999999999755
262	468	-122.649197	-115.787454	13.723486	0.00823193099993291	0.9999999999755
160	400	-34.253878	-27.405005	13.697746	0.00832491007154013	0.9999999999755
312	493	-22.524909	-15.678997	13.691824	0.00834644676191187	0.9999999999755
2	171	-112.289281	-105.456631	13.6653	0.00844357817895147	0.9999999999755
209	243	-83.369645	-76.565585	13.60812	0.00865675023049517	0.9999999999755
79	162	-36.750957	-29.966928	13.568058	0.00880923276612045	0.9999999999755
209	495	-101.320696	-94.54879	13.543812	0.00890279004421735	0.9999999999755
79	176	-27.181719	-20.431972	13.499494	0.00907631684629118	0.9999999999755
12	557	-72.257759	-65.510564	13.49439	0.00909651270958733	0.9999999999755
18	79	-34.203613	-27.511071	13.385084	0.00953972412230608	0.9999999999755
391	466	-148.423343	-141.739556	13.367574	0.0096126583070183	0.9999999999755
433&449	514	-22.064224	-15.401217	13.326014	0.00978795692183365	0.9999999999755
79	276	-67.062134	-60.400447	13.323374	0.00979919729328571	0.9999999999755
154	158	-221.120857	-214.461438	13.318838	0.00981853976584857	0.9999999999755
16&294	205	-16.029383	-9.371419	13.315928	0.00983096827222008	0.9999999999755
209	238	-80.155108	-73.505347	13.299522	0.00990132596967297	0.9999999999755
306	403	-22.358672	-15.713962	13.28942	0.00994489331108972	0.9999999999755

Table A.3.7: Detailed results generated using the KDBP11 data set for NA in H3N2 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
4	5	-140.551535	-115.819373	49.464324	4.67E-10	2.52E-05
2	4	-113.963582	-97.896387	32.13439	1.80E-06	0.0348153544795633
30	215	-191.722818	-175.885065	31.675506	2.23E-06	0.0348153544795633
18	215	-141.968795	-126.28716	31.36327	2.58E-06	0.0348153544795633
42	215	-142.317227	-126.887929	30.858596	3.27E-06	0.0353044758154209
215	216	-155.86052	-140.880835	29.95937	4.99E-06	0.0448602999915356
23	215	-153.032804	-138.311892	29.441824	6.36E-06	0.0490014117590225
2	5	-141.640317	-128.095522	27.08959	1.91E-05	0.128596052265248
267	338	-111.370643	-97.965384	26.810518	2.17E-05	0.13016322352628
215	387	-150.643803	-137.556013	26.17558	2.92E-05	0.157372062936907
215	372	-175.876948	-162.935756	25.882384	3.34E-05	0.16323158527533
215	437	-159.054349	-146.202314	25.70407	3.63E-05	0.16323158527533

Continued on next page

Table A.3.7 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
215	310	-181.656122	-169.582247	24.14775	7.46E-05	0.306265388827206
215	265	-170.896434	-158.890888	24.011092	7.95E-05	0.306265388827206
147	310	-168.960636	-157.463052	22.995168	0.000126907935161791	0.456496303305972
23	30	-114.499431	-103.25457	22.489722	0.000160093160150421	0.539874159317256
194	215	-159.70242	-148.579431	22.245978	0.000179042843320709	0.568260920836011
151	208	-298.020159	-286.967044	22.10623	0.000190894074553305	0.572215593699896
147	387	-137.948316	-127.159345	21.577942	0.000243156189340787	0.689103997241034
150	215	-153.709274	-142.974131	21.470286	0.000255431832323016	0.689103997241034
50	208	-128.316646	-117.840993	20.951306	0.000323786781901036	0.803369469548565
221	332	-97.455311	-87.072967	20.764688	0.000352563215472612	0.803369469548565
172	221	-81.930036	-71.559024	20.742024	0.000356226180617258	0.803369469548565
221	399	-106.187484	-95.819908	20.735152	0.000357344266979864	0.803369469548565
267	339	-131.381651	-121.137016	20.48927	0.000399720952128035	0.862693747720811
147	372	-163.181462	-153.039367	20.28419	0.000438844807024497	0.910704246454374
147	194	-147.006933	-136.934545	20.144776	0.000467583059149623	0.934404131091743
208	215	-181.839584	-171.88551	19.908148	0.000520686430556361	0.9999999999875
155	390	-49.067913	-39.428541	19.278744	0.000692771360469524	0.9999999999875
151	437	-275.23493	-265.88771	18.69444	0.000902355388088139	0.9999999999875
223	234	-44.100347	-34.755491	18.689712	0.000904284433107305	0.9999999999875
47	267	-99.701921	-90.428847	18.546148	0.000964840550960244	0.9999999999875
151	216	-272.031728	-262.847282	18.368892	0.00104516099261065	0.9999999999875
208	338	-112.919637	-103.829411	18.180452	0.00113779992965746	0.9999999999875
120	267	-163.302895	-154.224436	18.156918	0.00114992491961707	0.9999999999875
147	215	-217.560704	-208.519712	18.081984	0.00118938836160598	0.9999999999875
82	370	-127.044629	-118.03821	18.012838	0.00122698883474071	0.9999999999875
248	249	-105.019997	-96.015487	18.00902	0.00122909886439504	0.9999999999875
458	459	-50.477343	-41.617396	17.719894	0.00139974359347339	0.9999999999875
2	208	-124.23767	-115.415743	17.643854	0.00144837988963775	0.9999999999875
267	400	-82.484163	-73.810887	17.346552	0.00165507497646944	0.9999999999875
221	338	-96.858553	-88.214593	17.28792	0.00169914824242634	0.9999999999875
42	151	-258.496695	-250.021015	16.95136	0.00197546496346734	0.9999999999875
4	30	-133.032124	-124.591141	16.881966	0.00203772173065686	0.9999999999875
460	462	-57.896555	-49.533401	16.726308	0.00218449237417284	0.9999999999875
151	265	-287.076993	-278.747965	16.658056	0.00225208872620708	0.9999999999875
2	30	-134.120903	-125.81302	16.615766	0.00229500254974246	0.9999999999875
221	267	-115.812898	-107.51043	16.604936	0.00230612129517527	0.9999999999875
367	426	-85.840734	-77.551632	16.578204	0.00233379375398302	0.9999999999875
3	5	-114.796378	-106.517702	16.557352	0.00235560629205034	0.9999999999875
18	151	-258.149354	-249.890235	16.5182380000001	0.00239706526466199	0.9999999999875
30	151	-307.903387	-299.654305	16.498164	0.00241862109504687	0.9999999999875
47	338	-80.747576	-72.549737	16.395678	0.00253168438019746	0.9999999999875
150	151	-269.889838	-261.71149	16.3566959999999	0.00257604302664682	0.9999999999875
328	334	-79.102208	-70.930845	16.342726	0.00259212548843057	0.9999999999875
47	208	-101.250916	-93.09991	16.302012	0.00263956227260953	0.9999999999875
151	338	-277.516803	-269.38117	16.271266	0.00267595081040717	0.9999999999875
467	469	-50.458942	-42.381348	16.155188	0.00281783950547188	0.9999999999875
82	267	-112.040302	-103.962751	16.155102	0.00281794732667073	0.9999999999875
52	370	-112.243004	-104.167152	16.151704	0.0028222107706144	0.9999999999875
4	42	-83.625444	-75.583108	16.084672	0.00290762566973246	0.9999999999875
147	150	-141.013787	-132.990365	16.046844	0.00295694582647521	0.9999999999875

Continued on next page

Table A.3.7 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
2	42	-84.714222	-76.731945	15.964554	0.00306709755514389	0.9999999999875
148	265	-154.588796	-146.624072	15.929448	0.00311530979260433	0.9999999999875
47	339	-100.758584	-92.848993	15.819182	0.00327164401657698	0.9999999999875
82	172	-78.15744	-70.259604	15.795672	0.00330596158678165	0.9999999999875
468	469	-78.861775	-70.974619	15.774312	0.00333744774713474	0.9999999999875
52	267	-97.238677	-89.396752	15.68385	0.00347409805255838	0.9999999999875
23	151	-269.213363	-261.383175	15.660376	0.0035104477183352	0.9999999999875
4	23	-94.34211	-86.512211	15.659798	0.00351134745896731	0.9999999999875
4	216	-97.160477	-89.335684	15.649586	0.00352728137476244	0.9999999999875
127	221	-94.72738	-86.927875	15.59901	0.00360725014316809	0.9999999999875
82	151	-278.186472	-270.400877	15.57119	0.00365199548860917	0.9999999999875
241	287	-70.044727	-62.261923	15.565608	0.00366103903998682	0.9999999999875
331	435	-52.400018	-44.629912	15.540212	0.00370246262268259	0.9999999999875
2	216	-98.249254	-90.486048	15.526412	0.00372516482327112	0.9999999999875
2	23	-95.430889	-87.692935	15.475908	0.0038094215182668	0.9999999999875
155	328	-56.759264	-49.047166	15.424196	0.00389763482251559	0.9999999999875
248	267	-135.975774	-128.313444	15.32466	0.00407311889187079	0.9999999999875
47	249	-68.746144	-61.193065	15.106158	0.00448602771443796	0.9999999999875
127	285	-65.277967	-57.73946	15.077014	0.00454412162507856	0.9999999999875
4	18	-83.278102	-75.767891	15.020422	0.00465904744183665	0.9999999999875
359	419&420	-22.759934	-15.258199	15.00347	0.0046940255441954	0.9999999999875
199	267	-166.807698	-159.307124	15.001148	0.00469883671018834	0.9999999999875
221	437	-94.576657	-87.11821	14.916894	0.00487672489268509	0.9999999999875
2	18	-84.366883	-76.913846	14.906074	0.0049000434090325	0.9999999999875
267	385	-74.1345	-66.768057	14.732886	0.00528853411749663	0.9999999999875
18	30	-103.435422	-96.078339	14.714166	0.00533229497135612	0.9999999999875
334	384	-72.339251	-64.983297	14.711908	0.00533759735197847	0.9999999999875
89	322	-22.493935	-15.217846	14.552178	0.00572611215121643	0.9999999999875
149	390	-39.239054	-32.011958	14.454192	0.00597803773405736	0.9999999999875
194	335	-94.541538	-87.341945	14.399186	0.00612419180716595	0.9999999999875
199	215	-216.7733	-209.575485	14.39563	0.00613375997292154	0.9999999999875
65	381	-85.121412	-77.963863	14.315098	0.00635441619646426	0.9999999999875
267	381	-85.467318	-78.315731	14.303174	0.00638774244647655	0.9999999999875
151	172	-262.588286	-255.441413	14.2937459999999	0.0064142139127632	0.9999999999875
51	431	-150.81514	-143.689073	14.252134	0.0065323406301101	0.9999999999875
159	322	-22.114499	-14.992086	14.244826	0.006553305353589	0.9999999999875
52	199	-133.721387	-126.615921	14.210932	0.00665140378507478	0.9999999999875
199	331	-144.286475	-137.186706	14.199538	0.00668470326733372	0.9999999999875
467	468	-79.489709	-72.425517	14.128384	0.00689638112617863	0.9999999999875
260&452	261	-21.77714	-14.727302	14.099676	0.00698363284295656	0.9999999999875
267	342	-135.50191	-128.473591	14.056638	0.00711646736040761	0.9999999999875
257	282&352	-22.199851	-15.184335	14.031032	0.00719667016713788	0.9999999999875
329	356	-32.774535	-25.765332	14.018406	0.00723654248121752	0.9999999999875
149	370	-105.774035	-98.7817	13.98467	0.00734414381931736	0.9999999999875
52	62	-101.5899	-94.629657	13.920486	0.00755320371228818	0.9999999999875
268	282&352	-21.514353	-14.561359	13.905988	0.00760122744767489	0.9999999999875
267	462	-110.321548	-103.372697	13.897702	0.00762880846715241	0.9999999999875
338	370	-126.37497	-119.429498	13.890944	0.00765137588288978	0.9999999999875
16	21	-65.851742	-58.90748	13.888524	0.00765947302770498	0.9999999999875
52	151	-263.384837	-256.46375	13.842174	0.00781618608652146	0.9999999999875

Continued on next page

Table A.3.7 - Continued from previous page

site-1	**site-2**	**lkl-indep**	**lkl-dep**	**test-stat**	***P*-value**	***P*-value (FDR)**
215	249	-149.334812	-142.465949	13.737726	0.00818093018169386	0.9999999999875
153	368	-19.522221	-12.667663	13.709116	0.00828371284576168	0.9999999999875
155	248	-106.249358	-99.395435	13.707846	0.00828830455355267	0.9999999999875
227	228	-21.372586	-14.519539	13.706094	0.0082946430297719	0.9999999999875
47	344	-55.414002	-48.597998	13.632008	0.00856706076175984	0.9999999999875
197	462	-94.856671	-88.054269	13.604804	0.00866927281861274	0.9999999999875
69	308	-21.946527	-15.156018	13.581018	0.00875961957864446	0.9999999999875
165	274	-28.831479	-22.046053	13.570852	0.00879851364628936	0.9999999999875
11	355	-29.487763	-22.70613	13.563266	0.00882764692170956	0.9999999999875
43	53	-62.18546	-55.418248	13.534424	0.00893927565536456	0.9999999999875
346	390	-34.096673	-27.337333	13.51868	0.00900079186093183	0.9999999999875
47	127	-78.616404	-71.858896	13.515016	0.00901516741845965	0.9999999999875
267	331	-107.803764	-101.056404	13.49472	0.00909520561465604	0.9999999999875
141	151	-254.977731	-248.235514	13.484434	0.00913603386506068	0.9999999999875
106	368	-23.105528	-16.365453	13.48015	0.00915309118315888	0.9999999999875
56	267	-91.562243	-84.838056	13.448374	0.00928058750487426	0.9999999999875
71	127	-86.15186	-79.442996	13.417728	0.00940519423249664	0.9999999999875
356	466	-34.056448	-27.348043	13.41681	0.00940895194781366	0.9999999999875
432	462	-109.814891	-103.10911	13.411562	0.00943046216358712	0.9999999999875
153	344	-33.436465	-26.766585	13.33976	0.00972963369836866	0.9999999999875
151	385	-240.280661	-233.617236	13.32685	0.00978440009898351	0.9999999999875
18	23	-64.745408	-58.082295	13.326226	0.00978705483280917	0.9999999999875

Table A.3.8: Epistatic pairs of sites of amino acid detected before FDR in the Gong13NP data set of NP proteins in H3N2 human influenza A virus. The residues were recoded according to their physicochemical properties as in Fig. A.2.2.

Epistatic pair of Amino acids (years)	Amino acid physicochemical properties groups	Detected by Gong and coworkers	P-value after FDR **
A131S×T373N (2007–1969)	- Hydrophobicity - Normalised van der Waals Volume - Polarity - Solvent Accessibility	No	1
A131S×E375G (2007–1991)	- Hydrophobicity - Normalised van der Waals Volume - Polarity - Solvent Accessibility - Charge - Polarisability	No	1
A131S×R384G* (2007–1991)	- Polarity - Solvent Accessibility - Charge - Polarisability	No	1
A373S×L259S* (1969–1973)	- Polarisability - Normalised van der Waals Volume - Polarity - Secondary structure	No	1

Notes—* Deleterious mutation; **FDR: False Discovery Rate.

Table A.3.9: Detailed results generated using the Gong13NP data set. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	**site-2**	**lkl-indep**	**lkl-dep**	**test-stat**	P**-value**	P**-value (FDR)**
259	334	-24.112557	-15.110641	18.003832	0.00123197177337009	1
421	425	-41.33507	-32.643448	17.383244	0.00162806970000984	1
259	421	-28.433097	-20.76289	15.340414	0.00404483575372983	1
186	259	-52.01484	-44.549463	14.930754	0.0048470139256368	1
259	411	-43.108741	-35.931834	14.353814	0.00624738040528983	1
286	421	-21.478869	-14.567582	13.822574	0.00788339582234365	1
246	470	-11.518628	-4.755542	13.526172	0.00897146705998841	1

Table A.3.10: Detailed results generated using the Duan14NA data set. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	**site-2**	**lkl-indep**	**lkl-dep**	**test-stat**	P**-value**	P**-value (FDR)**
275	354	-136.541395	-116.877463	39.327864	5.96E-08	0.0020535384925
151	450	-139.615983	-128.751821	21.728324	0.000226982022307198	0.9999999999955
151	332	-159.091073	-148.686575	20.808996	0.000345509589050774	0.9999999999955
69	352	-55.379769	-44.989787	20.779964	0.000350115379225246	0.9999999999955
287	354	-63.891067	-54.806851	18.168432	0.00114397690881296	0.9999999999955
151	173	-137.888164	-128.945172	17.885984	0.00129904744386045	0.9999999999955
82	151	-146.103838	-137.666486	16.874704	0.00204434758921312	0.9999999999955
151	453	-144.681804	-136.271468	16.820672	0.00209431794817483	0.9999999999955
68	332	-95.891365	-87.504276	16.774178	0.00213827945675782	0.9999999999955
23	151	-160.568287	-152.31876	16.499054	0.0024176613631598	0.9999999999955
93	95	-53.766448	-45.618077	16.296742	0.00264576463273458	0.9999999999955
23	452	-93.45056	-85.319271	16.262578	0.00268632246945488	0.9999999999955
130	267	-121.808649	-113.710559	16.19618	0.00276690614373509	0.9999999999955
151	262	-136.41958	-128.732775	15.37361	0.00398586951995639	0.9999999999955
83	151	-154.666562	-147.216912	14.8993	0.00491469796176947	0.9999999999955
151	367	-151.047796	-143.685161	14.72527	0.00530629497982016	0.9999999999955
354	454	-78.108496	-70.771995	14.673002	0.00542977590314364	0.9999999999955
100	151	-130.504031	-123.195851	14.61636	0.00556677477166456	0.9999999999955
68	78	-106.513245	-99.36027	14.30595	0.00637996856547018	0.9999999999955
6	329	-100.48581	-93.444414	14.082792	0.00703545228325753	0.9999999999955
149	151	-156.535571	-149.509767	14.051608	0.00713215283964763	0.9999999999955
151	452	-164.589539	-157.573367	14.032344	0.00719253929803354	0.9999999999955
151	393	-145.958782	-138.989157	13.93925	0.0074914900717713	0.9999999999955
130	239	-53.068511	-46.106562	13.923898	0.00754194493815419	0.9999999999955
6	45	-49.538259	-42.597146	13.882226	0.00768058503537805	0.9999999999955
234	382	-52.245962	-45.500083	13.491758	0.00910694436005499	0.9999999999955

Table A.3.11: Detailed results generated using the Koel13HA data set. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
131	172	-56.325232	-90.641292	68.63212	4.41E-14	2.56E-09
172	299	-63.869614	-91.063367	54.387506	4.37E-11	1.27E-06
131	197	-54.723437	-80.250126	51.053378	2.18E-10	4.20E-06
131	299	-52.796784	-27.588776	50.416016	2.96E-10	4.28E-06
156	196	-149.015007	-125.514329	47.001356	1.52E-09	1.77E-05
82	172	-54.520069	-77.674156	46.308174	2.12E-09	2.05E-05
131	190	-42.767375	-65.287127	45.039504	3.90E-09	3.23E-05
142	156	-144.5229	-122.150763	44.744274	4.49E-09	3.26E-05
94	197	-44.415095	-66.252821	43.675452	7.49E-09	4.83E-05
155	158	-23.289118	-44.194232	41.810228	1.83E-08	0.000105870061375635
121	172	-81.928955	-102.095109	40.332308	3.69E-08	0.000194716683130647
156	226	-182.029994	-162.474344	39.1113	6.61E-08	0.000319184934128395
275	299	-72.503845	-91.350619	37.693548	1.30E-07	0.000577962847914969
172	262	-76.631249	-94.724662	36.186826	2.65E-07	0.00104000518981192
131	133	-89.996826	-108.07345	36.153248	2.69E-07	0.00104000518981192
57	157	-71.313012	-54.028379	34.569266	5.69E-07	0.00206323564386095
260	275	-84.652373	-101.740308	34.17587	6.86E-07	0.00233855447015707
156	192	-159.698581	-142.985686	33.42579	9.77E-07	0.0029905995524521
133	135	-92.222911	-108.932543	33.419264	9.80E-07	0.0029905995524521
226	262	-121.085486	-104.732525	32.705922	1.37E-06	0.00397691373256576
135	172	-56.241943	-72.48434	32.484794	1.52E-06	0.00420337615276976
94	248	-30.942845	-14.856788	32.172114	1.76E-06	0.00464875522733443
146	155	-44.326771	-60.286742	31.919942	1.99E-06	0.00500694975261313
172	196	-85.20475	-100.881214	31.352928	2.59E-06	0.00604869206333314
124	299	-58.586432	-74.256782	31.3407	2.61E-06	0.00604869206333314
47	216	-57.331129	-41.954512	30.753234	3.44E-06	0.00766433062728581
94	172	-46.654893	-61.943841	30.577896	3.73E-06	0.00801358453394216
226	276	-110.892793	-95.935017	29.915552	5.09E-06	0.0103080793574517
124	190	-55.926525	-70.837009	29.820968	5.32E-06	0.0103080793574517
133	157	-101.076796	-115.984826	29.81606	5.34E-06	0.0103080793574517
197	299	-60.894938	-75.768088	29.7463	5.51E-06	0.0103080793574517
190	193	-67.907935	-53.10205	29.61177	5.87E-06	0.0104351974002796
82	124	-52.888221	-67.681553	29.586664	5.94E-06	0.0104351974002796
146	219	-93.45458	-107.94919	28.98922	7.86E-06	0.0130615143536782
82	197	-56.427429	-70.918084	28.98131	7.89E-06	0.0130615143536782
157	226	-110.350463	-95.997502	28.705922	8.97E-06	0.0144137159332662
124	450	-84.747798	-99.073679	28.651762	9.20E-06	0.0144137159332662

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
133	157	-101.076796	-115.984826	29.81606	5.34E-06	0.0103080793574517
197	299	-60.894938	-75.768088	29.7463	5.51E-06	0.0103080793574517
190	193	-67.907935	-53.10205	29.61177	5.87E-06	0.0104351974002796
82	124	-52.888221	-67.681553	29.586664	5.94E-06	0.0104351974002796
146	219	-93.45458	-107.94919	28.98922	7.86E-06	0.0130615143536782
82	197	-56.427429	-70.918084	28.98131	7.89E-06	0.0130615143536782
157	226	-110.350463	-95.997502	28.705922	8.97E-06	0.0144137159332662
124	450	-84.747798	-99.073679	28.651762	9.20E-06	0.0144137159332662
156	193	-141.906891	-127.689876	28.43403	1.02E-05	0.0147925885126242
131	156	-115.085158	-129.301989	28.433662	1.02E-05	0.0147925885126242
57	156	-132.329673	-118.114981	28.429384	1.02E-05	0.0147925885126242
131	262	-53.772767	-67.932639	28.319744	1.07E-05	0.0151901024859121
196	375	-88.241385	-102.325965	28.16916	1.15E-05	0.015908596206199
219	299	-84.39992	-98.45473	28.10962	1.19E-05	0.0159765755133827
54	155	-44.92534	-58.755838	27.660996	1.46E-05	0.0189908627484213
197	276	-64.598716	-78.409744	27.622056	1.49E-05	0.0189908627484213
124	276	-64.414062	-78.209269	27.590414	1.51E-05	0.0189908627484213
172	248	-47.379421	-61.153635	27.548428	1.54E-05	0.0189908627484213
156	262	-136.522344	-122.848716	27.347256	1.69E-05	0.0200587546743246
347	384	-45.407296	-59.078106	27.34162	1.70E-05	0.0200587546743246
155	173&323	-21.437778	-34.852155	26.828754	2.15E-05	0.0249596943834507
133	138	-233.09576	-220.018628	26.154264	2.95E-05	0.0334827438576916
146	260	-68.844774	-55.848314	25.99292	3.17E-05	0.0353938089077092
219	276	-90.87916	-103.849445	25.94057	3.25E-05	0.0355802907411882
155	225	-35.075461	-47.921275	25.691628	3.65E-05	0.0391978185272695
156	246	-221.264234	-208.527253	25.473962	4.04E-05	0.0418124876197024
94	133	-81.269174	-93.938962	25.339576	4.30E-05	0.0437190995686334
54	159	-50.291102	-62.915365	25.248526	4.48E-05	0.0448167713614273
156	375	-139.427401	-126.910084	25.034634	4.95E-05	0.0486452103978388
156	197	-135.830135	-123.368725	24.92282	5.21E-05	0.0500389568285764
375	453	-75.964483	-63.513684	24.901598	5.27E-05	0.0500389568285764
196	220	-190.197331	-177.794496	24.80567	5.50E-05	0.0514666003308989
124	197	-74.605806	-86.927028	24.642444	5.94E-05	0.0534856947229171
135	173&323	-32.68001	-20.37102	24.61798	6.00E-05	0.0534856947229171
142	220	-188.033821	-175.725054	24.617534	6.00E-05	0.0534856947229171
124	375	-77.707519	-90.001173	24.587308	6.09E-05	0.0534856947229171
163	172	-48.55901	-60.814908	24.511796	6.31E-05	0.054559298682351
190	197	-59.452012	-71.688215	24.472406	6.42E-05	0.054744735840454
229	452	-226.236426	-214.036553	24.399746	6.64E-05	0.0557938638811007
146	275	-78.195493	-90.366425	24.341864	6.82E-05	0.0564876048789984
159	248	-28.837736	-40.832342	23.989212	8.03E-05	0.0654365820654832
121	133	-118.085097	-130.066287	23.96238	8.13E-05	0.0654365820654832
124	201	-89.206528	-101.002581	23.592106	9.64E-05	0.0755854137686689
156	276	-126.358671	-114.563495	23.590352	9.65E-05	0.0755854137686689
50	275	-110.836472	-122.471673	23.270402	0.000111809463839929	0.0864212615840093
124	157	-66.812128	-78.413177	23.202098	0.000115381222295596	0.0879987775898886

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
135	197	-59.645916	-71.232886	23.17394	0.000116886421846152	0.0879987775898886
220	452	-172.908979	-161.347677	23.122604	0.000119680940907085	0.0889474890305608
229	375	-236.730719	-225.246181	22.969076	0.00012844020645264	0.0942490983298678
121	190	-62.812756	-74.191632	22.757752	0.000141546180741137	0.10174518241206
261	299	-49.605782	-38.231659	22.748246	0.000142165944029271	0.10174518241206
121	156	-142.863567	-131.514102	22.69893	0.000145424681587181	0.10280815599523
124	262	-77.39017	-88.543916	22.307492	0.000174060335175197	0.120835928031344
142	299	-76.920183	-65.772891	22.294584	0.000175094323868086	0.120835928031344
135	276	-45.598181	-56.70918	22.221998	0.000181023273510994	0.123457872534498
192	246	-181.232002	-170.157115	22.149774	0.000187120238953398	0.126132095954982
172	384	-51.709551	-62.747294	22.075486	0.000193603939211195	0.12778228350828
192	229	-254.692613	-243.665571	22.054084	0.000195512862955383	0.12778228350828
159	172	-56.782493	-67.781472	21.997958	0.000200608073363862	0.12778228350828
220	529	-182.672498	-171.678073	21.98885	0.000201447241126806	0.12778228350828
220	375	-183.67905	-172.702215	21.95367	0.000204721378502803	0.12778228350828
50	260	-94.689275	-83.712958	21.952634	0.000204818592827949	0.12778228350828
135	275	-70.650232	-81.625592	21.95072	0.000204998315788685	0.12778228350828
133	452	-111.861384	-100.91372	21.895328	0.000210267879833248	0.128754390559956
159	173&323	-24.001045	-34.937096	21.872102	0.000212517188936223	0.128754390559956
190	226	-102.042375	-91.120654	21.843442	0.000215325661845345	0.128754390559956
49	529	-83.631148	-72.719162	21.823972	0.000217254543595446	0.128754390559956
63	260	-53.36382	-42.453885	21.81987	0.000217663106345967	0.128754390559956
453	529	-75.705336	-65.007195	21.396282	0.000264224599794471	0.154718182324096
155	163	-25.280744	-35.935942	21.310396	0.000274805841863901	0.159304946528503
49	106	-57.463109	-46.821678	21.282862	0.000278286158348906	0.159725233658278
106	220	-156.810892	-146.1897	21.242384	0.000283482072796493	0.161112311372673
121	196	-91.506957	-102.10514	21.196366	0.000289505995278105	0.162938471323027
196	242	-69.271627	-58.710905	21.121444	0.000299586106181193	0.164549176608459
384	453	-41.724796	-52.280959	21.112326	0.000300836354965583	0.164549176608459
222	299	-51.144795	-40.592114	21.105362	0.000301794731485505	0.164549176608459
54	384	-45.764409	-56.310121	21.091424	0.000303721957859326	0.164549176608459
172	173&323	-46.653453	-57.104388	20.90187	0.000331175002466932	0.176130411862459
133	156	-172.823132	-162.484654	20.676956	0.00036695331471015	0.190004515112401
172	450	-88.325033	-98.655943	20.66182	0.000369494164581097	0.190004515112401
133	384	-88.779026	-99.107331	20.65661	0.000370372782606543	0.190004515112401
172	192	-97.567101	-87.261729	20.610744	0.000378197508699829	0.192316750695869
190	452	-57.611482	-47.322632	20.5777	0.000383936076347235	0.193182273008617
190	375	-64.673245	-54.391881	20.562728	0.000386564493168873	0.193182273008617
106	529	-57.730217	-47.499146	20.462142	0.000404690950940867	0.200258011824123
94	219	-71.216443	-81.439564	20.446242	0.000407632316633544	0.200258011824123
202	222	-54.493339	-44.291284	20.40411	0.000415529163839023	0.20119445225455
155	159	-32.086038	-42.28304	20.394004	0.000417445760909296	0.20119445225455
57	220	-176.494371	-166.303938	20.380866	0.000419950469601527	0.20119445225455
131	276	-42.673108	-52.800566	20.254916	0.000444730422040629	0.211319857095863
216	385	-51.12756	-41.018044	20.219032	0.000452051719206814	0.213052342783894
186	220	-232.61022	-222.575661	20.069118	0.000483950986062465	0.225227730365345
156	157	-125.887003	-115.856311	20.061384	0.000485655792576645	0.225227730365345

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
156	452	-128.94249	-118.936546	20.011888	0.000496707861006196	0.22852503732166
133	229	-270.049502	-260.056591	19.985822	0.000502627916866838	0.228899715244312
82	133	-89.803667	-99.772844	19.938354	0.000513588728797698	0.228899715244312
163	173&323	-16.360794	-26.326642	19.931696	0.000515144927566547	0.228899715244312
94	201	-60.913487	-70.875467	19.92396	0.000516958954721258	0.228899715244312
82	94	-31.19643	-21.238934	19.914992	0.000519069791021454	0.228899715244312
190	192	-83.70289	-73.76272	19.88034	0.000527306458202315	0.228899715244312
82	275	-66.816087	-76.753238	19.874302	0.000528754848769131	0.228899715244312
138	197	-196.176423	-186.240013	19.87282	0.000529110951228873	0.228899715244312
172	375	-80.616478	-90.527502	19.822048	0.000541455136314095	0.231027861632655
49	186	-133.668109	-123.758194	19.81983	0.000542000848405055	0.231027861632655
248	275	-60.168155	-70.032178	19.728046	0.000565066964782535	0.239101693054332
121	261	-62.405206	-72.25909	19.707768	0.000570292590757138	0.239564213668053
193	226	-125.692221	-115.873978	19.636486	0.00058904432396012	0.245661147194015
94	275	-60.533342	-70.31553	19.564376	0.000608634590062218	0.252018194185048
33	54	-54.146174	-44.396762	19.498824	0.000627001926607584	0.257782281457033
157	386	-92.544435	-82.813833	19.461204	0.000637789452684889	0.258428360570166
92	192	-114.465663	-124.189099	19.446872	0.000641947281733724	0.258428360570166
213	275	-71.171593	-80.799194	19.255202	0.000700198623615145	0.27993458076531
63	146	-45.974074	-36.384591	19.178966	0.000724795029350189	0.287783341448154
164	174	-56.108263	-46.53173	19.153066	0.000733344249998602	0.289197048791966
157	192	-90.895589	-81.382307	19.026564	0.000776555386258315	0.304168349604017
57	375	-74.320637	-64.815217	19.01084	0.000782099423448535	0.304283916626252
57	190	-56.288048	-46.791331	18.993434	0.000788282162159093	0.304644779602417
54	260	-68.566102	-59.081965	18.968274	0.000797304697197632	0.304852908491239
121	155	-58.531093	-68.012414	18.962642	0.000799338314484532	0.304852908491239
131	375	-58.396569	-67.869367	18.945596	0.000805524635998256	0.305204334306006
213	347	-51.981455	-61.436208	18.909506	0.000818779048583518	0.308211827573939
260	384	-49.10544	-58.545712	18.880544	0.000829571409825203	0.310259707274626
157	375	-68.291231	-58.86465	18.853162	0.000839904310122042	0.312110595242146
142	229	-241.330907	-231.919572	18.82267	0.000851560607108226	0.314426550280662
124	163	-51.157226	-60.491179	18.667906	0.000913234268536511	0.335064497133301
262	275	-88.295839	-97.621699	18.65172	0.000919934053339233	0.335399855799216
53	189	-86.800404	-77.48348	18.633848	0.000927388166094945	0.336004324928275
5	190	-65.780774	-56.490441	18.580666	0.000949924463000906	0.342031808199767
155	196	-62.001647	-71.268394	18.533494	0.00097036645025228	0.346443758690093
63	201	-69.798372	-60.539949	18.516846	0.000977684168894877	0.346443758690093
192	262	-100.991766	-91.739234	18.505064	0.000982895942460371	0.346443758690093
83	155	-19.979145	-29.22809	18.49789	0.000986082804620758	0.346443758690093
94	299	-34.225275	-24.98394	18.48267	0.000992877745520149	0.346729656071103
197	260	-78.239601	-69.033342	18.412518	0.00102479996611049	0.355734455301947
135	260	-57.909944	-48.723937	18.372014	0.00104369071405264	0.35971801330216
220	226	-226.004872	-216.82416	18.361424	0.0010486862902892	0.35971801330216
213	479	-67.793672	-76.960265	18.333186	0.00106212267358552	0.362183831692662
192	196	-109.842098	-100.684069	18.316058	0.00107035536081723	0.362856726705117
155	156	-112.859521	-121.987746	18.25645	0.00109950091122679	0.369693737437346
124	133	-113.111748	-122.236168	18.24884	0.00110327784330966	0.369693737437346
156	450	-146.622227	-137.504625	18.235204	0.00111007763935278	0.36983448708782
173&323	453	-35.626106	-26.537061	18.17809	0.001139011135188	0.377305574324846

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
124	260	-73.026933	-63.964393	18.12508	0.00116653159437441	0.384226343897071
133	450	-122.819352	-131.850334	18.061964	0.00120015628095649	0.39306813337315
133	299	-95.015347	-104.009823	17.988952	0.00124024857017624	0.402372830676089
138	196	-206.507396	-197.515936	17.98292	0.00124361941014584	0.402372830676089
131	144	-40.354927	-31.368611	17.972632	0.00124938950356557	0.402372830676089
133	246	-196.588891	-187.611804	17.954174	0.00125980814428805	0.402712445219168
53	137	-108.100392	-99.133508	17.933768	0.00127142621777521	0.402712445219168
196	276	-74.911702	-83.872319	17.921234	0.00127861478978042	0.402712445219168
160	213	-33.435388	-42.39487	17.918964	0.00127992097941532	0.402712445219168
121	220	-186.736444	-177.781519	17.90985	0.00128517858143085	0.402712445219168
63	173&323	-23.277442	-14.337132	17.88062	0.00130218482225264	0.404442993617244
142	222	-71.580004	-62.641801	17.876406	0.00130465481812014	0.404442993617244
57	155	-47.248426	-56.170924	17.844996	0.00132321182656014	0.408013774391976
75	157	-66.618	-57.708256	17.819488	0.00133847351204264	0.410536029064083
131	453	-51.268371	-42.37988	17.776982	0.00136429230267776	0.416252762032788
135	299	-42.898676	-34.04359	17.710172	0.00140587086861943	0.426692849496693
143	189	-28.579923	-19.735735	17.688376	0.00141970415627557	0.428647135100494
146	299	-53.891888	-45.069154	17.645468	0.00144733042595702	0.434724066283566
135	155	-32.919557	-41.736097	17.63308	0.00145540468664551	0.434895926210517
131	193	-60.007467	-51.202484	17.609966	0.00147058900589381	0.43505936742826
25	172	-58.717814	-49.913113	17.609402	0.00147096146310055	0.43505936742826
197	384	-51.281896	-60.064081	17.56437	0.001501002402438	0.441592177612016
143	275	-65.141732	-73.918526	17.553588	0.0015082844776122	0.441592177612016
7	479	-64.977813	-56.212322	17.530982	0.00152366577224905	0.443853793051645
92	375	-103.791523	-95.032004	17.519038	0.00153185496336772	0.444008161132133
222	225	-42.124864	-33.396397	17.456934	0.00157514071132425	0.453956964945181
260	262	-74.779231	-66.055495	17.447472	0.0015818407265642	0.453956964945181
126	155	-23.389175	-32.0967	17.41505	0.00160501264800983	0.456748998247154
57	229	-229.585554	-220.879636	17.411836	0.00160732785306916	0.456748998247154
158	275	-59.902418	-68.590982	17.377128	0.0016325406786849	0.460698883376356
155	375	-57.672848	-66.358287	17.370878	0.00163712213171519	0.460698883376356
156	220	-241.394328	-232.731327	17.326002	0.00167039235423028	0.467790554467291
229	529	-235.990236	-227.416409	17.147654	0.00180933848399223	0.504266114985719
163	201	-63.964852	-72.494195	17.058686	0.00188285276345668	0.522243898074563
186	453	-122.060406	-113.554376	17.01206	0.00192254842967854	0.528417986456766
121	197	-81.257522	-89.762545	17.010046	0.00192428154500135	0.528417986456766
33	192	-80.528773	-72.034745	16.988056	0.00194330522641417	0.528889220540984
33	197	-59.591614	-51.105486	16.972256	0.0019570882852008	0.53015143875276
192	347	-94.65143	-86.174938	16.952984	0.00197403068322821	0.532253761426695
222	452	-56.415417	-47.962534	16.905766	0.0020161546394134	0.541094835401829
82	122	-27.678802	-19.279528	16.798548	0.00211512461679475	0.56504043334374
142	529	-89.436843	-81.047975	16.777736	0.00213488335633827	0.566818104947187
157	260	-65.310951	-56.92546	16.770982	0.0021413345693192	0.566818104947187
155	452	-44.859465	-53.222155	16.72538	0.00218539790898842	0.574152125687997
63	248	-26.197581	-17.836658	16.721846	0.00218884974602462	0.574152125687997
192	220	-201.376513	-193.029727	16.693572	0.00221666057942971	0.578827990042974
201	226	-135.545613	-127.21645	16.658326	0.00225181730023905	0.582758253994899
173&323	248	-19.387097	-11.072093	16.630008	0.00228046122106784	0.587548164379124

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
193	452	-79.057276	-70.74903	16.616492	0.00229425908578562	0.588487607092887
159	220	-161.136734	-152.863073	16.547322	0.00236616975457205	0.604259298116923
275	450	-98.261946	-106.518177	16.512462	0.00240324816724158	0.608368105917006
63	142	-63.156418	-54.927657	16.457522	0.00246284967085508	0.620745197475953
57	159	-57.333136	-49.112254	16.441764	0.0024802118342877	0.622415065080772
7	155	-31.195166	-22.993921	16.40249	0.00252401043749195	0.630676228712968
173&323	452	-42.941814	-34.75868	16.366268	0.00256508039109282	0.636613946642521
160	163	-26.946379	-18.771702	16.349354	0.00258448294827329	0.636613946642521
121	450	-94.003495	-102.172221	16.337452	0.00259822267386667	0.636613946642521
173&323	275	-55.955738	-64.123671	16.335866	0.00260005898393745	0.636613946642521
163	192	-73.318484	-65.151683	16.333602	0.00260268251430529	0.636613946642521
124	135	-55.953698	-64.075774	16.244152	0.00270845046910417	0.659701149974659
196	246	-169.998866	-161.881753	16.234226	0.00272044501650903	0.659850199192588
163	220	-153.555007	-145.443602	16.22281	0.00273430476553704	0.660448530242427
172	217	-66.506317	-58.402986	16.206662	0.0027540282814581	0.662204794132106
190	196	-70.953145	-79.052244	16.198198	0.00276442229049456	0.662204794132106
157	347	-63.957638	-55.865698	16.18388	0.00278209317697564	0.66369523238386
452	529	-73.13894	-65.082841	16.112198	0.00287224723577861	0.679608866359536
135	242	-36.093561	-28.052785	16.081552	0.00291166269367737	0.686134497367794
106	146	-48.26613	-40.234117	16.064026	0.00293444267257603	0.688703002952359
142	172	-86.1186	-94.111434	15.985668	0.00303845546913251	0.705678222040266
131	275	-65.896907	-73.889338	15.984862	0.00303954399487061	0.705678222040266
260	347	-69.098607	-61.107562	15.98209	0.00304329059013397	0.705678222040266
124	384	-51.382023	-59.367177	15.970308	0.00305926575782245	0.706556318649272
10	275	-97.113217	-89.146958	15.932518	0.00311106406638095	0.713800871972872
163	375	-55.088634	-47.123892	15.929484	0.00311525997255713	0.713800871972872
57	226	-117.124914	-109.166056	15.917716	0.00313158743115161	0.714717021196294
209	356	-63.999716	-56.060599	15.878234	0.00318698425077535	0.722812995418078
135	160	-34.795395	-26.858048	15.874694	0.00319199804773207	0.722812995418078
124	146	-67.947634	-60.030045	15.835178	0.00324849527740056	0.732744246034672
121	157	-71.215349	-79.122179	15.81366	0.00327967285186859	0.736909438848147
142	246	-167.870281	-160.005454	15.729654	0.00340423315700045	0.761581606145444
133	190	-97.999831	-105.86085	15.722038	0.0034157532792447	0.761581606145444
173&323	197	-47.027883	-54.875769	15.695772	0.00345577878323144	0.76636948952601
33	260	-55.754481	-47.911097	15.686768	0.00346960546209574	0.76636948952601
121	124	-81.04241	-88.88343	15.68204	0.0034768876271406	0.76636948952601
155	244	-30.361867	-38.194334	15.664934	0.00350336047307531	0.769279570546121
157	248	-42.982537	-35.198709	15.567656	0.00365771846383844	0.80014316735364
82	131	-34.458742	-26.688049	15.541386	0.00370053758236977	0.800743716843867
121	131	-59.62614	-67.390754	15.529228	0.003720521168737	0.800743716843867
137	159	-47.792996	-55.556355	15.526718	0.00372465994570903	0.800743716843867
57	92	-88.344692	-80.582992	15.5234	0.00373013799529642	0.800743716843867
82	375	-57.992648	-50.234937	15.515422	0.00374334219880434	0.800743716843867
83	275	-56.443374	-64.190091	15.493434	0.00377997240351158	0.805606618498405
172	260	-75.107703	-67.368664	15.478078	0.00380576304390134	0.808132174560297
146	347	-61.728124	-53.999149	15.45795	0.00383983052194214	0.812390421010898
137	145	-72.056847	-64.335375	15.442944	0.00386542373527898	0.81483132339681

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
138	156	-257.771099	-250.056935	15.428328	0.0038905130048531	0.817148691635269
94	159	-30.908048	-23.204889	15.406318	0.00392859691159841	0.817506550518829
121	229	-240.06565	-232.367633	15.396034	0.00394651672317781	0.817506550518829
94	479	-55.115062	-62.812782	15.39544	0.00394755421606141	0.817506550518829
202	299	-54.730142	-47.033058	15.394168	0.00394977681998065	0.817506550518829
142	175	-74.83521	-67.141825	15.38677	0.00396272797474195	0.817506550518829
94	242	-25.239887	-17.563186	15.353402	0.00402166367405121	0.826722848172868
145	189	-50.169128	-42.522699	15.292858	0.00413080656592058	0.846158503980268
82	479	-63.230761	-70.866305	15.271088	0.00417075880973317	0.847728038740991
54	242	-45.197408	-37.563457	15.267902	0.00417663751623298	0.847728038740991
219	450	-110.984447	-118.616855	15.264816	0.00418233947006941	0.847728038740991
304	452	-79.56105	-71.934924	15.252252	0.00420563286092079	0.849479222813862
160	190	-32.237537	-39.858996	15.242918	0.00422302040104172	0.85002948836246
131	142	-62.348153	-69.959288	15.22227	0.00426173524646933	0.851919733793854
202	386	-94.210581	-86.599464	15.222234	0.00426180304985713	0.851919733793854
135	189	-31.296099	-23.698522	15.195154	0.00431310774307248	0.859212563113098
137	275	-79.92529	-87.51611	15.18164	0.00433893736023183	0.860330350979068
106	173&323	-25.962985	-18.376294	15.173382	0.00435479582101495	0.860330350979068
163	275	-61.425488	-69.009986	15.168996	0.00436324173172065	0.860330350979068
25	155	-28.866135	-21.287199	15.157872	0.00438473493320102	0.860910419108941
124	226	-120.400477	-127.976538	15.152122	0.00439588552796699	0.860910419108941
226	231	-108.30332	-100.737435	15.13177	0.00443557710712461	0.8657397383858
201	452	-87.528315	-79.967121	15.122388	0.00445399276918035	0.8657397383858
131	163	-31.985212	-24.427338	15.115748	0.00446707161298188	0.8657397383858
201	225	-79.175126	-71.623801	15.10265	0.00449298153648703	0.8657397383858
92	196	-103.242272	-110.793034	15.101524	0.00449521582291057	0.8657397383858
82	156	-114.448471	-106.905559	15.085824	0.00452648279179213	0.868874859073476
51	140	-27.99493	-20.456644	15.076572	0.00454500834659843	0.869551596872314
163	197	-51.216915	-58.746908	15.059986	0.00457840562108636	0.873059782415711
146	172	-69.695959	-62.172157	15.047604	0.00460349484717526	0.87436753389091
226	229	-279.335161	-271.816238	15.0378460000001	0.00462336226496685	0.87436753389091
193	202	-74.95527	-67.438098	15.034344	0.00463051290157857	0.87436753389091
54	450	-86.146326	-78.658884	14.974884	0.00475359387992746	0.894694276686348
83	190	-26.728921	-34.208945	14.960048	0.00478480158961803	0.897653553884003
155	453	-42.586713	-50.05972	14.946014	0.0048145074367405	0.900312890670474
275	384	-64.530687	-71.990203	14.919032	0.00487213013360921	0.908065314243136
124	219	-100.050685	-107.506677	14.911984	0.00488729304888491	0.908065314243136
57	223	-84.14084	-76.698188	14.885304	0.00494511270361342	0.912955998179841
172	276	-65.09222	-72.5239	14.86336	0.00499317170162905	0.918902106487099
63	299	-37.507042	-30.080465	14.853154	0.00501567956499505	0.920123241717604
57	186	-124.452017	-131.833345	14.762656	0.00521966724385314	0.95452400670715
82	106	-35.67608	-28.314866	14.722428	0.00531293764369445	0.968525142153985
33	271	-35.925134	-28.574097	14.702074	0.0053607505954899	0.972459371626231
220	248	-152.637438	-145.287951	14.698974	0.00536806967259607	0.972459371626231
50	172	-102.095461	-94.753884	14.683154	0.00540557358263227	0.973502643058585
146	197	-70.425473	-63.084283	14.68238	0.00540741506063247	0.973502643058585
299	452	-57.488652	-50.15792	14.661464	0.00545741157288859	0.979461761239479
173&323	213	-30.960142	-23.641103	14.638078	0.00551384952353073	0.986536595305792
157	163	-42.781998	-35.474937	14.614122	0.00557225687010077	0.993919171568435
83	122	-16.342897	-9.042956	14.599882	0.00560726279606694	0.997095166527608

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
157	453	-59.019873	-51.728437	14.582872	0.00564936107038549	0.998627784066229
155	160	-25.028745	-32.319985	14.58248	0.00565033488310718	0.998627784066229
33	92	-72.218032	-64.932801	14.570462	0.00568027025204132	0.9999999999955
143	248	-27.267079	-19.987242	14.559674	0.00570727434003271	0.9999999999955
220	453	-169.875305	-162.59974	14.55113	0.00572875068182932	0.9999999999955
193	219	-105.948183	-113.213606	14.530846	0.00578005494456268	0.9999999999955
196	529	-90.932077	-83.678513	14.507128	0.00584061667350455	0.9999999999955
275	278	-60.252977	-67.505748	14.505542	0.00584468850953412	0.9999999999955
164	172	-77.033793	-69.812251	14.443084	0.00600727396046807	0.9999999999955
196	347	-82.136695	-74.922972	14.427446	0.00604867095490935	0.9999999999955
135	541	-28.532892	-21.328924	14.407936	0.00610071037092097	0.9999999999955
75	146	-67.529937	-60.33342	14.393034	0.00614075431316519	0.9999999999955
124	220	-179.802458	-172.608846	14.387224	0.00615643640992691	0.9999999999955
133	163	-85.137387	-92.326747	14.37872	0.00617946086947774	0.9999999999955
219	426	-99.656838	-92.480819	14.352038	0.00625225148398223	0.9999999999955
155	172	-54.02349	-61.190159	14.333338	0.00630376701491686	0.9999999999955
54	146	-60.15157	-52.998617	14.305906	0.00638009171022702	0.9999999999955
63	160	-26.006799	-18.858222	14.297154	0.00640463271205716	0.9999999999955
135	219	-82.619265	-89.739675	14.24082	0.00656482548319859	0.9999999999955
122	384	-19.61421	-26.730605	14.23279	0.00658797731585026	0.9999999999955
220	384	-156.740571	-149.626297	14.228548	0.0066002399998899	0.9999999999955
138	142	-204.381318	-197.280048	14.20254	0.0066759139649003	0.9999999999955
197	242	-56.047905	-48.964596	14.166618	0.00678183420708611	0.9999999999955
122	160	-21.372697	-14.294871	14.155652	0.0068144956456776	0.9999999999955
172	347	-71.159305	-64.088703	14.141204	0.00685776341484157	0.9999999999955
190	248	-28.417982	-35.486246	14.136528	0.00687182432451994	0.9999999999955
244	276	-48.214431	-41.146367	14.136128	0.0068730284506795	0.9999999999955
53	278	-89.373057	-82.309062	14.12799	0.0068975713423286	0.9999999999955
196	453	-77.860993	-70.799972	14.122042	0.00691556385341208	0.9999999999955
172	226	-120.953241	-128.005738	14.104994	0.00696738871790314	0.9999999999955
131	304	-65.466472	-58.417651	14.097642	0.00698985561474619	0.9999999999955
57	121	-77.922398	-70.874583	14.09563	0.00699601642551784	0.9999999999955
106	135	-36.089592	-29.047467	14.08425	0.00703096264802816	0.9999999999955
160	219	-74.502493	-81.542656	14.080326	0.00704305225809942	0.9999999999955
155	384	-27.9318	-34.962247	14.060894	0.00710322191274571	0.9999999999955
83	145	-49.581557	-42.556249	14.050616	0.00713525027446926	0.9999999999955
172	529	-81.075068	-74.052081	14.045974	0.0071497620426999	0.9999999999955
2	155	-37.541461	-44.560506	14.03809	0.00717447520251169	0.9999999999955
242	331	-23.549119	-30.546877	13.995516	0.00730938055947039	0.9999999999955
126	193	-52.890505	-45.899921	13.981168	0.00735540286939074	0.9999999999955
106	155	-33.576133	-26.591511	13.969244	0.00739386587593449	0.9999999999955
172	219	-101.114381	-108.09349	13.958218	0.00742960749346167	0.9999999999955
331	384	-28.597205	-35.570754	13.947098	0.0074658252008365	0.9999999999955
156	384	-112.828685	-119.796767	13.936164	0.00750160569561331	0.9999999999955
160	201	-63.752392	-70.700896	13.897008	0.00763112297862256	0.9999999999955
82	242	-32.522742	-25.591032	13.86342	0.00774396575789293	0.9999999999955

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
25	193	-59.454494	-52.53825	13.832488	0.0078493296237222	0.9999999999955
275	509	-57.4627	-50.551978	13.821444	0.00788728787322224	0.9999999999955
146	384	-41.72992	-48.639894	13.819948	0.0078924434293498	0.9999999999955
199	247	-65.406967	-58.507465	13.799004	0.00796496857918394	0.9999999999955
82	226	-99.63263	-92.734582	13.796096	0.00797508988390194	0.9999999999955
138	226	-242.381419	-235.484437	13.793964	0.00798251831619945	0.9999999999955
142	452	-81.076979	-74.185702	13.782554	0.00802238894678764	0.9999999999955
194	226	-145.435549	-138.544347	13.782404	0.00802291439669367	0.9999999999955
126	244	-27.433497	-34.32111	13.775226	0.00804809834599707	0.9999999999955
160	384	-25.484145	-32.36505	13.76181	0.00809537591497678	0.9999999999955
220	450	-190.460282	-183.585038	13.750488	0.0081354857282594	0.9999999999955
106	142	-65.317115	-58.454194	13.725842	0.00822347151126335	0.9999999999955
138	248	-168.733679	-161.872729	13.7219	0.00823763051129656	0.9999999999955
159	453	-46.472956	-53.324344	13.702776	0.00830666006438074	0.9999999999955
160	347	-44.814802	-37.964372	13.70086	0.00831360713550122	0.9999999999955
155	189	-21.071808	-27.91832	13.693024	0.00834207829043065	0.9999999999955
144	158	-27.55264	-34.385354	13.665428	0.00844310679467353	0.9999999999955
426	427	-61.807167	-54.976437	13.66146	0.00845773168901576	0.9999999999955
94	260	-44.074939	-37.244767	13.660344	0.00846184940515771	0.9999999999955
83	375	-51.343718	-44.527575	13.632286	0.00856602235870729	0.9999999999955
347	479	-79.971252	-86.782867	13.62323	0.00859991231801649	0.9999999999955
207	534	-49.252066	-42.451926	13.60028	0.00868638590255322	0.9999999999955
63	472	-21.409695	-14.612078	13.595234	0.00870551256512819	0.9999999999955
57	312	-68.943744	-75.736721	13.585954	0.00874079566293218	0.9999999999955
83	278	-20.432775	-13.655385	13.55478	0.00886034838818517	0.9999999999955
94	155	-25.260075	-32.034964	13.549778	0.00887967946810242	0.9999999999955
275	347	-82.205447	-88.976397	13.5419	0.00891020902684159	0.9999999999955
131	450	-65.606709	-72.374406	13.535394	0.00893549907317859	0.9999999999955
138	375	-199.776984	-193.016337	13.521294	0.00899054964298773	0.9999999999955
49	328	-57.073112	-50.313366	13.519492	0.00899760904776725	0.9999999999955
54	193	-74.387307	-81.14572	13.516826	0.00900806315130598	0.9999999999955
271	299	-46.499555	-39.741173	13.516764	0.00900830641129602	0.9999999999955
94	192	-69.05857	-62.303105	13.51093	0.00903122516243138	0.9999999999955
194	220	-204.799787	-198.048616	13.502342	0.00906506675038055	0.9999999999955
202	452	-62.267793	-55.521928	13.49173	0.00910705539793133	0.9999999999955
196	229	-243.459521	-236.717326	13.48439	0.00913620889882683	0.9999999999955
190	229	-214.765556	-208.041104	13.448904	0.00927844678736212	0.9999999999955
146	450	-78.511167	-71.791849	13.438636	0.00932000622995632	0.9999999999955
57	217	-65.101674	-58.383247	13.436854	0.00932723733819962	0.9999999999955
156	347	-130.002894	-123.288134	13.42952	0.00935705548519916	0.9999999999955
271	529	-61.85077	-55.1434	13.41474	0.00941743061537137	0.9999999999955
260	452	-69.189044	-62.482096	13.413896	0.00942088976152655	0.9999999999955

Continued on next page

Table A.3.11 - Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
142	201	-99.514968	-92.836301	13.357334	0.0096555631683235	0.9999999999955
157	222	-54.129263	-47.451169	13.356188	0.00966037645785334	0.9999999999955
197	386	-103.343478	-96.674705	13.337546	0.00973900452533127	0.9999999999955
53	384	-92.442923	-99.11064	13.335434	0.00974795185568822	0.9999999999955
10	34	-61.255058	-54.591963	13.32619	0.00978720801200927	0.9999999999955
131	529	-59.485435	-52.834357	13.302156	0.00988999689741887	0.9999999999955
260	299	-58.561035	-51.915353	13.291364	0.00993649480352199	0.9999999999955
63	124	-51.664336	-45.019507	13.289658	0.00994386472616582	0.9999999999955

Table A.3.12: Detailed results generated using the Adam03M2 data set. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimisation under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P-values are computed using the standard χ^2 approximation. FDR-corrected P-values follow BH. Only pairs significant at the nominal 1% level (P-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
31	51	-215.103387	-197.277061	35.652652	3.41E-07	0.00113281616036631
12	53	-77.799957	-65.876321	23.847272	8.57E-05	0.142316180090207
13	56	-46.536317	-40.493356	12.085922	0.0167236344812858	0.999999964628273
11	58	-29.888689	-23.892009	11.99336	0.0174007103450287	0.999999964628273
13	54	-71.386993	-66.032614	10.708758	0.0300397186352879	0.999999964628273
6	70	-17.476614	-12.309403	10.334422	0.0351558484101363	0.999999964628273
5	8	-12.900464	-7.964634	9.87166	0.0426459429703552	0.999999964628273
13	51	-168.765715	-164.127949	9.275532	0.0545696117650336	0.999999964628273
63	75	-16.839233	-12.370825	8.936816	0.0626983299054582	0.999999964628273
40	63	-17.108515	-12.661735	8.89356	0.0638157412109532	0.999999964628273
24	52	-58.933899	-54.596959	8.67388	0.0697883949162066	0.999999964628273
4	90	-32.302657	-27.98877	8.62777400000001	0.0711074320946067	0.999999964628273
75	96	-17.939049	-13.701758	8.474582	0.0756614405570145	0.999999964628273
5	10	-24.371314	-20.161276	8.420076	0.0773472039538335	0.999999964628273
16	17	-39.158191	-34.956543	8.403296	0.0778732712198347	0.999999964628273
75	77	-24.809162	-20.607664	8.402996	0.0778827071170962	0.999999964628273
50	88	-56.048288	-51.902973	8.29062999999999	0.0814937940632067	0.999999964628273
14	68	-47.801338	-43.671019	8.260638	0.0824839822106254	0.999999964628273
12	51	-173.532531	-169.559639	7.945784	0.0935845065252358	0.999999964628273
15	94	-27.233707	-23.320416	7.82658199999999	0.0981412984366585	0.999999964628273
9	55	-22.540834	-18.64306	7.795548	0.0993612397204912	0.999999964628273

Appendix B

Appendix to Chapter 3

B.1 Additional methods details

B.1.1 Alignment algorithm

We developed a custom algorithm to reconstruct an genome alignment, made of concatenated genes (*i.e.* an exome alignment), using as reference a PA14 genome database obtained from `www.pseudomonas.com` [Winsor *et al.* 2016, accessed Dec 4, 2014]. From this database, we extracted the nucleotide sequences for the 5,977 genes, and used each of them as a query for BLASTn [Altschul *et al.* 1990] searches against a local database containing the 390 draft genomes, as well as the three complete genomes of PA01, PA7 and PA14. Results with more than 90% identity were then used to construct scaffolds, which were subsequently extended, for each of the 390 draft genomes, making sure that the same genomic information was not used more than once. The extended scaffold sequences of each gene were aligned with MUSCLE [Edgar 2004]. We discarded genes present in less than 50% of strains with $\geq 90\%$ of the length of the PA14 query (non-gap characters). The remaining

genes were concatenated into a single exome alignment with *catfasta2phyml.pl* [Nylander 2015].

B.1.2 Complexity Hypothesis

The species tree for our exome alignment was reconstructed based on highly networked genes that, according to the complexity hypothesis, are unlikely to be horizontally transferred. According to the complexity hypothesis [Aris-Brosou 2005; Jain *et al.* 1999], highly networked genes are those whose products are involved in complex multiprotein interactions such as information processing genes (*e.g.* rRNA genes, tRNA genes, transcription and translation polymerases). Based on the functional annotations of PA14 genes obtained from the Clusters of Orthologous Groups database [Galperin *et al.* 2015; Tatusov *et al.* 1997], we extracted 1,290 information genes (those with COG terms A, B, J, K, and L), and concatenated them into a single alignment used to estimate the species tree.

B.1.3 Analysis of correlated evolution

Recoding our concatenated exome into binary states resulted in 410,665 polymorphic sites. We estimated that a full pairwise comparison of the polymorphic sites would have required more than one year of computing even when parallelised across twelve computational nodes each with four processors. It is for this reason that we performed a 12 gene by exome analysis.

B.1.4 Modified selection counter-selection protocol

Our modified WT allele replacement protocol is based on a selection counter-selection method [Schweizer 2008], whereby a vector-borne mutant allele

recombines with homologous chromosomal sequences before the delivery vector is itself removed from the targeted genome. Mutant alleles of *gyrA* and *parC* were generated from WT chromosomal DNA by amplification of each locus in paired PCR reactions that overlapped at sequences adjacent to the introduced substitution. One of the overlapping primers coded for the substitution of interest through a mismatch at the target site. The paired amplicons were ligated to an allelic replacement vector using Golden Gate assembly [Engler *et al.* 2008], which permitted scarless ligation of the PCR products (see Table B.3.2 for details). The vector, derived from pAH79 [Melnyk *et al.* 2017] and modified for Golden Gate cloning, includes the *TetA* selectable marker and *sacB* counter-selection gene. After ligation, vectors were transformed into chemically competent *Escherichia coli* (DH5 λpir). The mutant alleles were transferred into *P. aeruginosa* strains as previously described [Melnyk *et al.* 2017] via tri-parental conjugation, involving a helper *E. coli* that carries pRK2013 [Figurski and Helinski 1979]. Through a round of selection (LB agar with 100μg/mL Nitrofurantoin and 80μg/mL Tetracycline), we isolated colonies whose genome had recombined with the mutant allele. The recombinants were subsequently counter-selected (LB agar with 5% sucrose) for loss of the plasmid sequences. Mutant constructs were first confirmed by sequencing the region targeted for replacement. To identify constructs least likely to contain secondary mutations outside the sequenced region, we performed competitive fitness assays with a minimum of four independent constructs, and accepted constructs with relatively similar fitness measures for downstream analyses.

B.1.5 Calculating competitive fitness

Competitive fitness was calculated as $\omega = (f_{final} - f_{initial})^{(1/generations)}$, where $f_{initial}$ and f_{final} are the initial and final frequency of focal strains, and the number of generations was based on the dilution factor, calculated as $\log_2(100) \sim 6.64$. The evidence for epistasis (ϵ) was calculated with a multiplicative fitness model [Trindade *et al.* 2009] such that $\epsilon = W_{WT}W_{AB} - W_AW_B$ where W stands for fitness and the subscripts (WT, A, B, AB) represent the wild-type, single and double mutant genotypes. There is evidence for epistasis when the ϵ value is greater than our measurement error estimated via error propagation.

B.1.6 Adaptive boosting algorithm

Adaptive boosting is a supervised machine learning algorithm that uses the weighted sum of many sequentially fitted classifiers, and is considered the "best out-of-the-box" classifier in part because it yields lower classification error rates while being less susceptible to overfitting [James *et al.* 2013]. Nucleotide sites in our exome alignment were quantified for their importance in predicting levofloxacin resistance by computing the number of fitted classification trees which retain a predictor, in this case nucleotide site. Classification was based on the association with the discrete phenotype of sensitivity or resistance to levofloxacin. Phenotype data were obtained from the original study that published the 390 draft genomes used to construct our alignment [Kos *et al.* 2015].

B.1.7 Calculating ΔG

To estimate biological effects of synonymous mutations, ΔG values were calculated with the mFold server [Zuker 2003], based on the QuickFold service with default

settings (unafold.rna.albany.edu/?q=DINAMelt/Quickfold). Similar to previous studies [Takanami and Zubay 1964], we calculated the free energy of each mRNA transcript by considering 50 nucleotides upstream and downstream of each substitution. Mean and error values of ΔG_{rel} were calculated using weighted values for all the predicted folding structures returned by QuickFold.

B.1.8 Calculating I_{TE}

Also, to estimate biological effects of synonymous mutations, we calculated the index of translation elongation [I_{TE}: Xia 2014] with default settings in DAMBE ver. 6 [Xia 2017]. These calculations used the codon frequency of genes that are highly expressed in *P. aeruginosa* [Hilterbrand *et al.* 2012], and lowly expressed genes calculated using all of the other genes in our alignment. Files containing the codon frequency counts for all strains in our alignment are available from github.com/JDench/Pseudomonas_DAMBE_ITE.

B.1.9 Supplementary analyses

While a previous study has shown the excellent specificity of AEGIS at identifying correlated pairs of substitutions from simulated evolution [Nshogozabahizi *et al.* 2017], the $\approx 127,000$ significantly correlated pairs ($P \leq 0.01$) was much higher than expected and we deemed it beneficial to review the results prior to *in vitro* study. As this study is the first to have tried to identify correlated evolution among pairs of sites in the whole genome of *P. aeruginosa* there was no general dataset against which we could compare our results, instead we performed summary analyses to describe trends in the results. Recall that for reasons of computational time, this study analysed the evidence of correlated evolution between the sites in 12 genes

and the rest of the genome. If the results of AEGIS were largely false positives, we would have expected the number of polymorphic sites (*i.e.* sites in our alignment with different nucleotide characters across genotypes) in a gene to be positively correlated with the number of correlated pairs involving substitutions in that gene. We found no evidence that the number of polymorphic sites in the 12 genes was correlated to the number of associated significant pairs (Pearson's ρ, $t = 0.5530$, $df = 10$, $P = 0.5924$). Further, we found no evidence that the proportion of polymorphic sites in a gene (*i.e.* number of sites divided by gene length) differed between the six focal genes (*gyrA*, *gyrB*, *morA*, *nfxB*, *parC*, *parE*), the additional six genes chosen at random (*dnaA*, *dnaN*, *lpd*3, *ribD*, *rpoB*, *serC*), and any of the other genes in our alignment (ANOVA, $F = 0.1227$, $df = 2$, $P = 0.8845$). This led us to conclude that the results of AEGIS reflected some true evolutionary signal.

While we did not have specific information concerning the *in vitro* effects of most substitutions identified in the results, we could define several descriptive traits. For this, we compared trends among our significantly correlated paired sites to what we would expect by chance (*i.e.* from a null model). If chance alone drove our detection of correlated site pairs, we would expect to detect a paired traits according to their frequency in our exome alignment. We found that the frequency of observed paired traits differs from our null expectation (Fig. B.2.1). As a first assessment we leveraged our *a priori* assumption that mutations in the six focal genes were more likely to show an adaptive response to fluoroquinolone selection compared to the other six genes. If our assumption was correct and the results of AEGIS reflect correlated evolution in response to selection, we would expect more pairs to include mutations in the six focal genes. When comparing the number of correlated pairs which include zero, one, or two sites in the six focal genes, we found that pairs with at least one site in an expected gene are consistently higher than by chance

(compare panels a,b in Fig. B.2.1). We next wanted to assess if correlated pairs of potentially adaptive substitutions (*i.e.* nonsynonymous) were detected more often than by chance. While synonymous substitutions may be selected for [Agashe *et al.* 2016; Bailey *et al.* 2014], in response to antibiotic selection we assumed that only nonsynonymous substitutions were likely to be adaptive. We looked for differences in the expected and observed number of correlated pairs where zero, one or two sites were nonsynonymous (compare Fig. B.2.1 C,D). We found that nonsynonymous pairs are dramatically underrepresented comprising only 2 pairs with at least medium (*parC* 786 and PA14_34000 967 - hypothetical type VI secretion protein - $10^{-7} \leq P \leq 10^{-6}$) and strong (*gyrA* 248 and *parC* 260, $P \leq 10^{-11}$) evidence for correlated evolution respectively. We interpret the higher than expected number of synonymous pairs to suggest that hitchhiking [Maynard Smith and Haigh 1974], possibly as "cohorts" [Lang *et al.* 2013], explains the majority of correlated pairs identified by AEGIS.

B.2 Figures

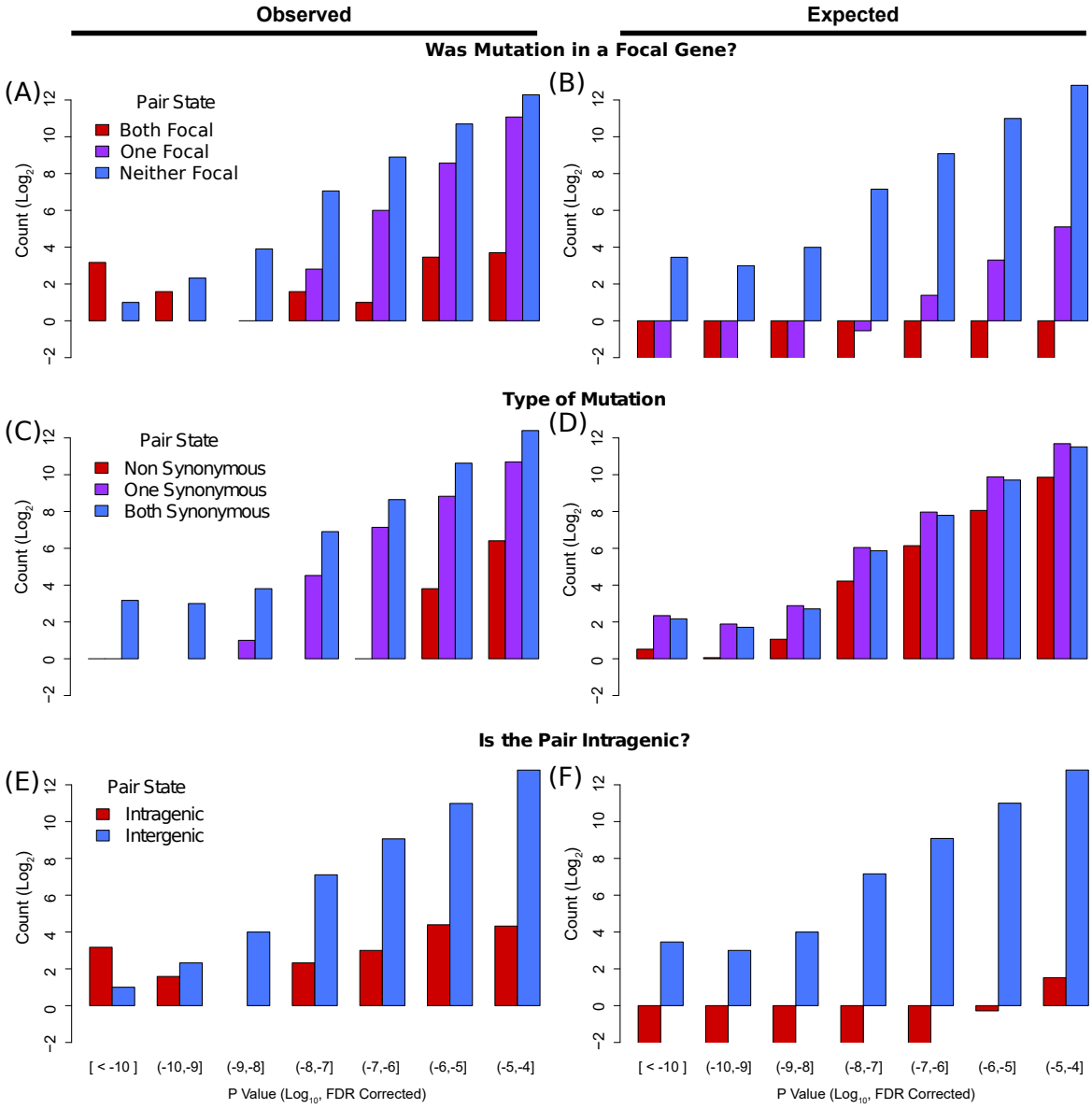

Figure B.2.1: Comparisons of observed and expected types of correlated substitutions with at least weak ($P \leq 10^{-4}$) support. Note that all y-axes are based on $\log_2$ counts and so when there is no bar the count was 0. Panels A,C,E show the observed distributions of when correlated pairs were: (A) in genes we expected to show correlation, (C) nonsynonymous, or (E) intragenic. Whereas B,D,F show the null expected distributions, of their left-most counterpart, if a same number of correlated pairs had been randomly drawn from our alignment.

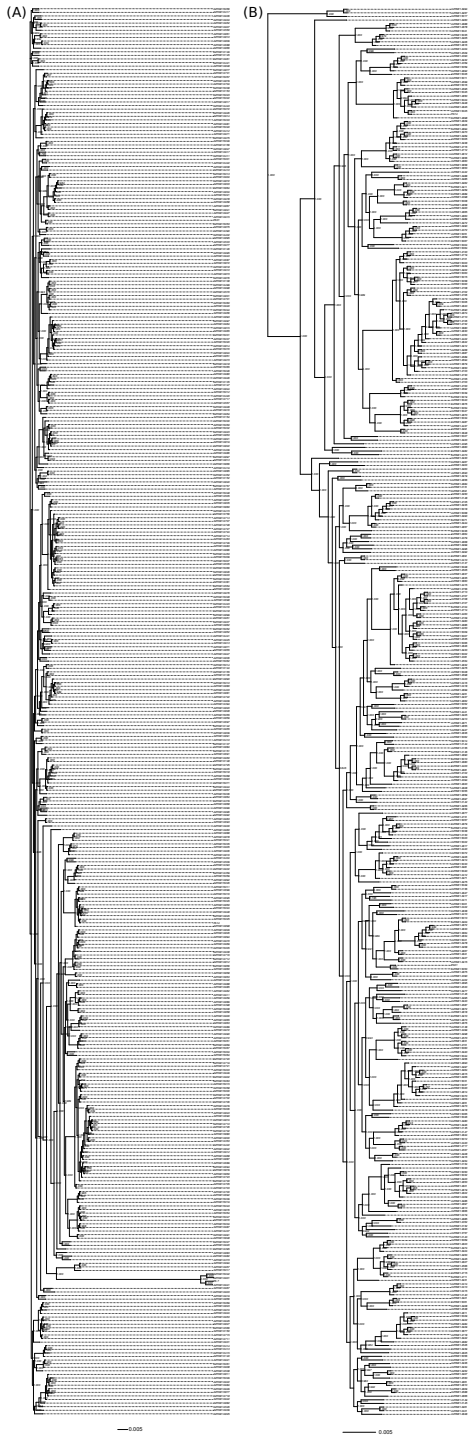

Figure B.2.2: Phylogenetic trees estimated from the concatenated alignment of all conserved Information class, determined from COG terms, genes in our dataset. Support for bifurcations are printed at nodes and were estimated using FastTree's default SH test for local support values. (A) shows the unrooted tree which includes the taxonomic outliers identified by being part of the long branch sub-clade including PA7. (B) shows the rooted tree once the PA7 sub-clade was removed. Analysis of correlated evolution, with AEGIS, was performed with the tree presented in (B).

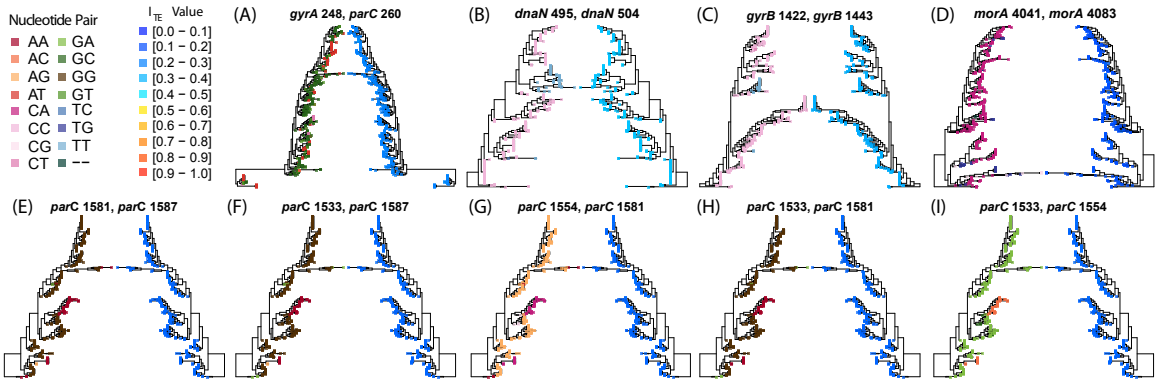

Figure B.2.3: Phylogeny of the gene(s) which contain the most significantly correlated pairs of substitutions ($P < 10^{11}$). Each tree was constructed with FastTree [Price *et al.* 2010a] (GTR $+\Gamma_5$). The title of each panel indicates the pair of substitutions being considered. For trees on the left hand side, the colour of the tips represent the paired nucleotide state of strains. On the right side, the colour of tips represent the I_{TE} value of the whole gene(s) sequence.

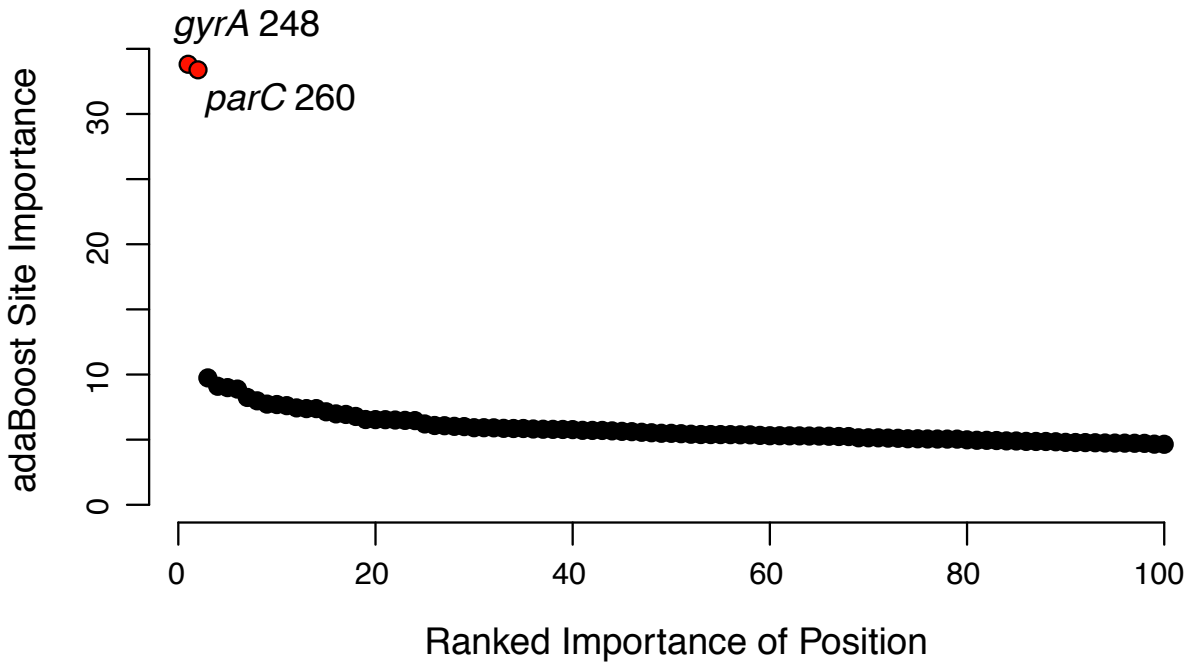

Figure B.2.4: Top 100 ranked importance values in predicting levofloxacin (a fluoroquinolone drug) resistance for nucleotide positions in our alignment. The importance values were obtained by running the adaptive boosting machine learning algorithm *boosting* (implemented in the R package *adabag* [Alfaro *et al.* 2013]) on all polymorphic sites in our alignment. Levofloxacin resistance phenotype information was obtained from previously published data [Kos *et al.* 2015]. Importance values reflect the strength of correlation between genomic sites and the resistance phenotype. Red circles denote genomic positions which have been previously reported in the literature to correlate with fluoroquinolone resistance.

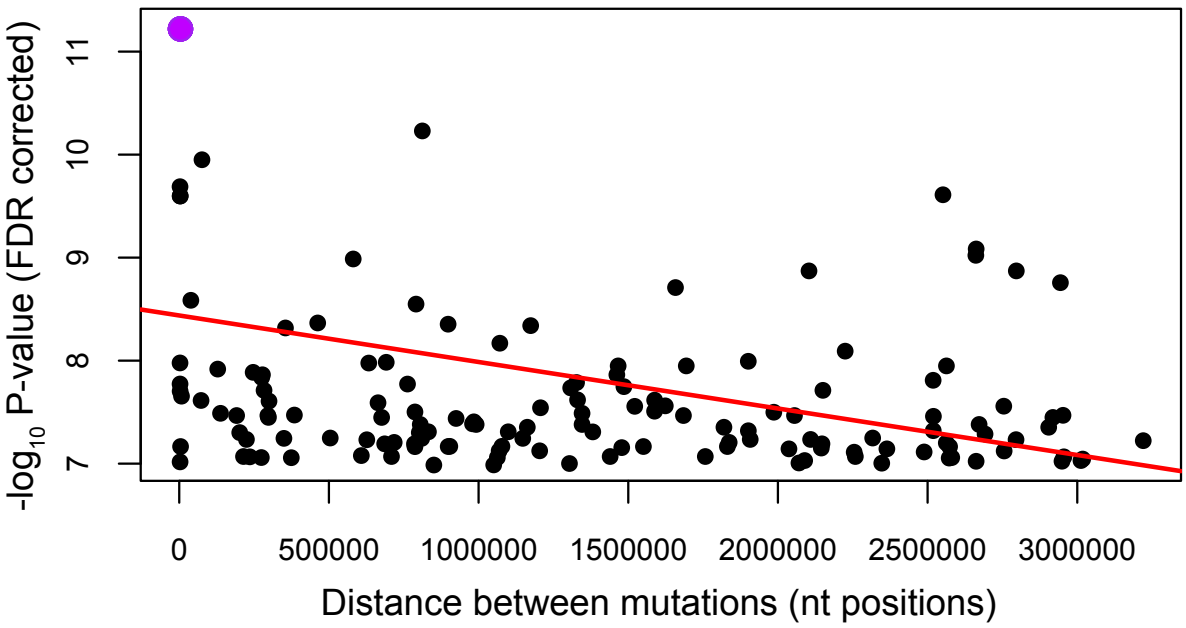

Figure B.2.5: The signal of correlated evolution for paired synonymous substitutions as a function of physical distance. Physical distance is measured as the number of base pairs, using the circular reference chromosome of PA14, separating two mutations. The red line shows the linear regression for all significantly correlated pairs with at least medium support ($P \leq 10^{-7}$). Purple highlights the most significantly correlated pairs of substitutions.

B.3 Tables

Table B.3.1: Table of *P. aeruginosa* strains included in whole exome alignment. Where known we provide the name of isolates used in our study along with the year, country and origin as per the cited reference paper.

Isolate	Year of isolation	Country	City	Reference
AZPAE12135	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12136	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12137	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12138	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12140	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12142	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12143	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12144	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12145	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12146	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12147	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12148	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12149	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12150	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12151	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12152	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12153	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12154	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12155	2005	United States	New York	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE12156	2005	United States	New York	[Kos *et al.* 2015]
AZPAE12409	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12410	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12411	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12412	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12413	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12414	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12415	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12416	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12417	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12418	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12419	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12420	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12421	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12422	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE12423	2007	United States	Cleveland	[Kos *et al.* 2015]
AZPAE13756	2009	Canada	unknown	[Kos *et al.* 2015]
AZPAE13757	2009	Canada	unknown	[Kos *et al.* 2015]
AZPAE13848	2010	India	unknown	[Kos *et al.* 2015]
AZPAE13850	2010	India	unknown	[Kos *et al.* 2015]
AZPAE13853	2010	India	unknown	[Kos *et al.* 2015]
AZPAE13856	2010	India	unknown	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE13858	2010	India	unknown	[Kos *et al.* 2015]
AZPAE13860	2010	India	unknown	[Kos *et al.* 2015]
AZPAE13864	2010	India	unknown	[Kos *et al.* 2015]
AZPAE13866	2010	China	unknown	[Kos *et al.* 2015]
AZPAE13872	2010	Mexico	unknown	[Kos *et al.* 2015]
AZPAE13876	2010	Portugal	unknown	[Kos *et al.* 2015]
AZPAE13877	2010	Romania	unknown	[Kos *et al.* 2015]
AZPAE13879	2010	Argentina	unknown	[Kos *et al.* 2015]
AZPAE13880	2010	Mexico	unknown	[Kos *et al.* 2015]
AZPAE14352	2010	France	unknown	[Kos *et al.* 2015]
AZPAE14353	2010	France	unknown	[Kos *et al.* 2015]
AZPAE14359	2010	China	Shatin	[Kos *et al.* 2015]
AZPAE14372	2010	China	Hong Kong	[Kos *et al.* 2015]
AZPAE14373	2010	Germany	München	[Kos *et al.* 2015]
AZPAE14379	2010	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE14381	2010	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14390	2011	China	Hong Kong	[Kos *et al.* 2015]
AZPAE14393	2011	Spain	Madrid	[Kos *et al.* 2015]
AZPAE14394	2011	Spain	Madrid	[Kos *et al.* 2015]
AZPAE14395	2011	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14398	2011	Germany	München	[Kos *et al.* 2015]
AZPAE14402	2011	France	Rouen	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14403	2011	France	Rouen	[Kos *et al.* 2015]
AZPAE14404	2012	China	Shatin	[Kos *et al.* 2015]
AZPAE14410	2012	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE14415	2009	Portugal	unknown	[Kos *et al.* 2015]
AZPAE14422	2009	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14437	2010	Canada	unknown	[Kos *et al.* 2015]
AZPAE14441	2010	Taiwan	unknown	[Kos *et al.* 2015]
AZPAE14442	2010	Taiwan	unknown	[Kos *et al.* 2015]
AZPAE14443	2010	United States	unknown	[Kos *et al.* 2015]
AZPAE14453	2011	United States	Detroit	[Kos *et al.* 2015]
AZPAE14463	2011	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE14499	2011	Spain	Santander	[Kos *et al.* 2015]
AZPAE14505	2011	France	Paris	[Kos *et al.* 2015]
AZPAE14509	2011	France	Nantes	[Kos *et al.* 2015]
AZPAE14526	2010	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14533	2011	Germany	Koln	[Kos *et al.* 2015]
AZPAE14535	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE14538	2010	China	Shatin	[Kos *et al.* 2015]
AZPAE14550	2010	China	Beijing	[Kos *et al.* 2015]
AZPAE14554	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE14557	2010	Germany	Koln	[Kos *et al.* 2015]
AZPAE14566	2011	China	Shatin	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14570	2010	Germany	unknown	[Kos *et al.* 2015]
AZPAE14687	2012	Mexico	unknown	[Kos *et al.* 2015]
AZPAE14688	2012	Mexico	unknown	[Kos *et al.* 2015]
AZPAE14689	2012	Mexico	unknown	[Kos *et al.* 2015]
AZPAE14690	2012	Romania	unknown	[Kos *et al.* 2015]
AZPAE14691	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14692	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14693	2012	Romania	unknown	[Kos *et al.* 2015]
AZPAE14694	2012	Romania	unknown	[Kos *et al.* 2015]
AZPAE14695	2012	Israel	unknown	[Kos *et al.* 2015]
AZPAE14697	2012	Israel	unknown	[Kos *et al.* 2015]
AZPAE14698	2012	Israel	unknown	[Kos *et al.* 2015]
AZPAE14699	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14700	2012	Philippines	unknown	[Kos *et al.* 2015]
AZPAE14701	2012	Philippines	unknown	[Kos *et al.* 2015]
AZPAE14702	2012	Philippines	unknown	[Kos *et al.* 2015]
AZPAE14703	2012	Philippines	unknown	[Kos *et al.* 2015]
AZPAE14704	2012	Greece	unknown	[Kos *et al.* 2015]
AZPAE14705	2012	Greece	unknown	[Kos *et al.* 2015]
AZPAE14706	2012	Greece	unknown	[Kos *et al.* 2015]
AZPAE14707	2012	Greece	unknown	[Kos *et al.* 2015]
AZPAE14708	2012	Greece	unknown	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14710	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14711	2012	Venezuela	unknown	[Kos *et al.* 2015]
AZPAE14712	2012	Venezuela	unknown	[Kos *et al.* 2015]
AZPAE14713	2012	Venezuela	unknown	[Kos *et al.* 2015]
AZPAE14714	2012	Venezuela	unknown	[Kos *et al.* 2015]
AZPAE14715	2012	Venezuela	unknown	[Kos *et al.* 2015]
AZPAE14716	2012	Venezuela	unknown	[Kos *et al.* 2015]
AZPAE14717	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14718	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14719	2012	Colombia	unknown	[Kos *et al.* 2015]
AZPAE14720	2012	Colombia	unknown	[Kos *et al.* 2015]
AZPAE14721	2012	Colombia	unknown	[Kos *et al.* 2015]
AZPAE14722	2012	Italy	unknown	[Kos *et al.* 2015]
AZPAE14723	2012	Italy	unknown	[Kos *et al.* 2015]
AZPAE14724	2012	Italy	unknown	[Kos *et al.* 2015]
AZPAE14725	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14726	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14727	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14728	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14729	2012	Italy	unknown	[Kos *et al.* 2015]
AZPAE14730	2012	Italy	unknown	[Kos *et al.* 2015]
AZPAE14731	2012	Italy	unknown	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14732	2012	United States	unknown	[Kos *et al.* 2015]
AZPAE14809	2004	India	Mumbai	[Kos *et al.* 2015]
AZPAE14810	2004	India	Mumbai	[Kos *et al.* 2015]
AZPAE14811	2004	India	Mumbai	[Kos *et al.* 2015]
AZPAE14812	2004	India	Mumbai	[Kos *et al.* 2015]
AZPAE14813	2004	India	Mumbai	[Kos *et al.* 2015]
AZPAE14814	2004	France	Besancon	[Kos *et al.* 2015]
AZPAE14815	2004	France	Besancon	[Kos *et al.* 2015]
AZPAE14816	2004	France	Besancon	[Kos *et al.* 2015]
AZPAE14817	2004	France	Besancon	[Kos *et al.* 2015]
AZPAE14818	2004	France	Besancon	[Kos *et al.* 2015]
AZPAE14819	2004	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14820	2004	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14821	2004	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14822	2004	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14823	2005	Germany	Koln	[Kos *et al.* 2015]
AZPAE14824	2005	Germany	Koln	[Kos *et al.* 2015]
AZPAE14825	2005	Germany	Koln	[Kos *et al.* 2015]
AZPAE14826	2005	United States	Detroit	[Kos *et al.* 2015]
AZPAE14827	2005	United States	Detroit	[Kos *et al.* 2015]
AZPAE14828	2005	United States	Detroit	[Kos *et al.* 2015]
AZPAE14829	2005	United States	Detroit	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14830	2005	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14831	2005	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14832	2005	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14833	2005	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14834	2005	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14835	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14836	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14837	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14838	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14839	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14840	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14841	2006	China	Beijing	[Kos *et al.* 2015]
AZPAE14842	2006	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14843	2006	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14844	2006	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14845	2006	Germany	Koln	[Kos *et al.* 2015]
AZPAE14846	2006	France	Nantes	[Kos *et al.* 2015]
AZPAE14847	2006	France	Nantes	[Kos *et al.* 2015]
AZPAE14848	2006	France	Nantes	[Kos *et al.* 2015]
AZPAE14850	2006	France	Nantes	[Kos *et al.* 2015]
AZPAE14851	2006	France	Nantes	[Kos *et al.* 2015]
AZPAE14852	2005	Brazil	Sao Paulo	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14853	2007	Brazil	Curitiba	[Kos *et al.* 2015]
AZPAE14855	2007	France	Paris	[Kos *et al.* 2015]
AZPAE14856	2007	France	Paris	[Kos *et al.* 2015]
AZPAE14857	2007	France	Paris	[Kos *et al.* 2015]
AZPAE14858	2007	France	Paris	[Kos *et al.* 2015]
AZPAE14859	2007	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14860	2007	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14861	2007	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14862	2007	India	Chennai	[Kos *et al.* 2015]
AZPAE14863	2007	India	Chennai	[Kos *et al.* 2015]
AZPAE14864	2007	India	Chennai	[Kos *et al.* 2015]
AZPAE14865	2007	India	Chennai	[Kos *et al.* 2015]
AZPAE14866	2007	China	Shatin	[Kos *et al.* 2015]
AZPAE14867	2007	China	Shatin	[Kos *et al.* 2015]
AZPAE14868	2007	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14869	2007	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14870	2007	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14871	2007	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14872	2007	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14873	2007	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14874	2007	United States	Detroit	[Kos *et al.* 2015]
AZPAE14875	2007	United States	Detroit	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14876	2007	United States	Detroit	[Kos *et al.* 2015]
AZPAE14877	2007	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14878	2007	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14879	2007	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14880	2007	Spain	Santander	[Kos *et al.* 2015]
AZPAE14881	2007	Spain	Santander	[Kos *et al.* 2015]
AZPAE14882	2007	Spain	Santander	[Kos *et al.* 2015]
AZPAE14883	2007	Croatia	Split	[Kos *et al.* 2015]
AZPAE14884	2007	Croatia	Split	[Kos *et al.* 2015]
AZPAE14885	2007	Croatia	Split	[Kos *et al.* 2015]
AZPAE14886	2007	Croatia	Split	[Kos *et al.* 2015]
AZPAE14887	2007	Croatia	Split	[Kos *et al.* 2015]
AZPAE14888	2007	United States	Detroit	[Kos *et al.* 2015]
AZPAE14889	2008	China	Shatin	[Kos *et al.* 2015]
AZPAE14890	2008	France	Besancon	[Kos *et al.* 2015]
AZPAE14891	2008	France	Besancon	[Kos *et al.* 2015]
AZPAE14892	2008	France	Besancon	[Kos *et al.* 2015]
AZPAE14893	2008	France	Besancon	[Kos *et al.* 2015]
AZPAE14894	2008	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE14895	2008	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE14897	2008	India	Chennai	[Kos *et al.* 2015]
AZPAE14898	2008	India	Chennai	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14899	2008	India	Chennai	[Kos *et al.* 2015]
AZPAE14900	2008	India	Chennai	[Kos *et al.* 2015]
AZPAE14901	2008	India	Chennai	[Kos *et al.* 2015]
AZPAE14902	2008	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14903	2008	Spain	Madrid	[Kos *et al.* 2015]
AZPAE14904	2008	Spain	Madrid	[Kos *et al.* 2015]
AZPAE14905	2008	Germany	Koln	[Kos *et al.* 2015]
AZPAE14906	2008	Germany	Koln	[Kos *et al.* 2015]
AZPAE14907	2008	China	Shatin	[Kos *et al.* 2015]
AZPAE14908	2008	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14909	2008	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14910	2008	India	Mumbai	[Kos *et al.* 2015]
AZPAE14911	2008	India	Mumbai	[Kos *et al.* 2015]
AZPAE14912	2008	Croatia	Split	[Kos *et al.* 2015]
AZPAE14913	2008	Croatia	Split	[Kos *et al.* 2015]
AZPAE14914	2008	Spain	Santander	[Kos *et al.* 2015]
AZPAE14915	2008	Spain	Santander	[Kos *et al.* 2015]
AZPAE14916	2008	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14917	2008	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14918	2008	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14919	2008	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE14920	2008	Spain	Bilbao	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14921	2009	France	Paris	[Kos *et al.* 2015]
AZPAE14922	2009	France	Paris	[Kos *et al.* 2015]
AZPAE14923	2008	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14924	2008	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14925	2008	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14926	2008	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14927	2008	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14928	2008	Brazil	Sao Paulo	[Kos *et al.* 2015]
AZPAE14929	2009	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14930	2009	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14931	2009	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14932	2009	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14933	2009	France	Rouen	[Kos *et al.* 2015]
AZPAE14934	2009	France	Rouen	[Kos *et al.* 2015]
AZPAE14935	2008	France	Rouen	[Kos *et al.* 2015]
AZPAE14936	2008	Brazil	Curitiba	[Kos *et al.* 2015]
AZPAE14937	2008	France	Rouen	[Kos *et al.* 2015]
AZPAE14938	2008	France	Rouen	[Kos *et al.* 2015]
AZPAE14939	2009	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE14940	2009	France	Rouen	[Kos *et al.* 2015]
AZPAE14941	2009	China	Hong Kong	[Kos *et al.* 2015]
AZPAE14942	2009	China	Hong Kong	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14943	2009	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE14944	2009	United States	Detroit	[Kos *et al.* 2015]
AZPAE14945	2009	United States	Detroit	[Kos *et al.* 2015]
AZPAE14946	2009	United States	Detroit	[Kos *et al.* 2015]
AZPAE14947	2009	China	Beijing	[Kos *et al.* 2015]
AZPAE14948	2009	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14949	2009	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14950	2009	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14951	2009	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE14952	2009	China	Shatin	[Kos *et al.* 2015]
AZPAE14953	2009	China	Shatin	[Kos *et al.* 2015]
AZPAE14954	2009	France	Nantes	[Kos *et al.* 2015]
AZPAE14955	2009	France	Nantes	[Kos *et al.* 2015]
AZPAE14956	2009	Germany	München	[Kos *et al.* 2015]
AZPAE14957	2009	Germany	München	[Kos *et al.* 2015]
AZPAE14958	2009	India	Mumbai	[Kos *et al.* 2015]
AZPAE14959	2009	India	Mumbai	[Kos *et al.* 2015]
AZPAE14960	2009	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14961	2009	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14962	2009	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14963	2009	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14964	2009	France	Nantes	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14965	2009	France	Nantes	[Kos *et al.* 2015]
AZPAE14967	2009	Croatia	Split	[Kos *et al.* 2015]
AZPAE14968	2009	Croatia	Split	[Kos *et al.* 2015]
AZPAE14969	2010	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14970	2010	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14971	2010	China	Shatin	[Kos *et al.* 2015]
AZPAE14972	2010	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14973	2010	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14974	2010	Germany	Aachen	[Kos *et al.* 2015]
AZPAE14975	2010	China	Beijing	[Kos *et al.* 2015]
AZPAE14976	2010	China	Beijing	[Kos *et al.* 2015]
AZPAE14977	2010	China	Beijing	[Kos *et al.* 2015]
AZPAE14978	2010	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14979	2010	United States	Roseburg	[Kos *et al.* 2015]
AZPAE14980	2010	United States	Detroit	[Kos *et al.* 2015]
AZPAE14981	2010	France	Paris	[Kos *et al.* 2015]
AZPAE14982	2010	Croatia	Split	[Kos *et al.* 2015]
AZPAE14983	2010	Croatia	Split	[Kos *et al.* 2015]
AZPAE14984	2010	France	Paris	[Kos *et al.* 2015]
AZPAE14985	2010	China	Shatin	[Kos *et al.* 2015]
AZPAE14986	2010	China	Shatin	[Kos *et al.* 2015]
AZPAE14987	2010	Germany	Koln	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE14988	2010	China	Shatin	[Kos *et al.* 2015]
AZPAE14989	2010	China	Hong Kong	[Kos *et al.* 2015]
AZPAE14990	2010	China	Hong Kong	[Kos *et al.* 2015]
AZPAE14991	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE14992	2010	Germany	München	[Kos *et al.* 2015]
AZPAE14993	2010	Spain	Madrid	[Kos *et al.* 2015]
AZPAE14994	2010	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE14995	2010	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14996	2010	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14997	2010	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14998	2010	Spain	Palma de Mallorca	[Kos *et al.* 2015]
AZPAE14999	2010	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE15000	2010	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE15001	2011	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE15002	2010	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE15003	2011	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE15004	2011	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE15005	2011	United States	Roseburg	[Kos *et al.* 2015]
AZPAE15006	2011	United States	Roseburg	[Kos *et al.* 2015]
AZPAE15007	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE15008	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE15009	2010	Spain	Santander	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE15010	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE15011	2010	Spain	Santander	[Kos *et al.* 2015]
AZPAE15012	2011	Germany	Koln	[Kos *et al.* 2015]
AZPAE15013	2011	Germany	Koln	[Kos *et al.* 2015]
AZPAE15014	2011	Germany	Koln	[Kos *et al.* 2015]
AZPAE15015	2011	Germany	Koln	[Kos *et al.* 2015]
AZPAE15016	2011	United States	Detroit	[Kos *et al.* 2015]
AZPAE15017	2011	United States	Detroit	[Kos *et al.* 2015]
AZPAE15018	2011	United States	Detroit	[Kos *et al.* 2015]
AZPAE15019	2010	France	Besancon	[Kos *et al.* 2015]
AZPAE15020	2010	France	Besancon	[Kos *et al.* 2015]
AZPAE15021	2011	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE15022	2011	France	Nantes	[Kos *et al.* 2015]
AZPAE15023	2011	Spain	Madrid	[Kos *et al.* 2015]
AZPAE15024	2011	Spain	Madrid	[Kos *et al.* 2015]
AZPAE15025	2011	Spain	Madrid	[Kos *et al.* 2015]
AZPAE15026	2011	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE15027	2011	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE15028	2011	France	Paris	[Kos *et al.* 2015]
AZPAE15029	2011	France	Paris	[Kos *et al.* 2015]
AZPAE15030	2011	Germany	München	[Kos *et al.* 2015]
AZPAE15031	2011	Germany	Aachen	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE15032	2011	France	Rouen	[Kos *et al.* 2015]
AZPAE15033	2011	France	Rouen	[Kos *et al.* 2015]
AZPAE15034	2011	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE15035	2011	Spain	Bilbao	[Kos *et al.* 2015]
AZPAE15036	2012	China	Shatin	[Kos *et al.* 2015]
AZPAE15037	2012	France	Paris	[Kos *et al.* 2015]
AZPAE15038	2012	France	Paris	[Kos *et al.* 2015]
AZPAE15039	2012	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15040	2012	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15041	2012	Germany	Koln	[Kos *et al.* 2015]
AZPAE15042	2012	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15043	2012	France	Nantes	[Kos *et al.* 2015]
AZPAE15044	2012	France	Nantes	[Kos *et al.* 2015]
AZPAE15045	2011	France	Nantes	[Kos *et al.* 2015]
AZPAE15046	2012	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE15047	2012	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE15048	2012	Germany	München	[Kos *et al.* 2015]
AZPAE15049	2012	Germany	München	[Kos *et al.* 2015]
AZPAE15050	2012	China	Hong Kong	[Kos *et al.* 2015]
AZPAE15051	2012	China	Hong Kong	[Kos *et al.* 2015]
AZPAE15052	2012	Argentina	Victoria	[Kos *et al.* 2015]
AZPAE15053	2012	Argentina	Victoria	[Kos *et al.* 2015]

Continued on next page

Table B.3.1 – continued from previous page

Isolat	Year of isolation	Country	City	Reference
AZPAE15054	2012	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE15055	2012	Colombia	Bogota	[Kos *et al.* 2015]
AZPAE15056	2012	China	Beijing	[Kos *et al.* 2015]
AZPAE15057	2012	China	Beijing	[Kos *et al.* 2015]
AZPAE15058	2012	France	Nantes	[Kos *et al.* 2015]
AZPAE15059	2012	France	Nantes	[Kos *et al.* 2015]
AZPAE15060	2012	France	Nantes	[Kos *et al.* 2015]
AZPAE15061	2012	France	Nantes	[Kos *et al.* 2015]
AZPAE15062	2012	Brazil	Curitiba	[Kos *et al.* 2015]
AZPAE15063	2012	Brazil	Curitiba	[Kos *et al.* 2015]
AZPAE15064	2012	Brazil	Curitiba	[Kos *et al.* 2015]
AZPAE15065	2012	Brazil	Curitiba	[Kos *et al.* 2015]
AZPAE15066	2003	Croatia	Split	[Kos *et al.* 2015]
AZPAE15067	2004	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15068	2004	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15069	2004	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15070	2004	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15071	2004	Germany	Heidelberg	[Kos *et al.* 2015]
AZPAE15072	2004	Germany	Heidelberg	[Kos *et al.* 2015]
PAO1	1954	Australia	Melbourne	[Holloway 1955]
PA14	1990	United States	Unknown	[Rahme *et al.* 1995]
PA7	Unknown	Argentina	Unknown	[Roy *et al.* 2010]

Table B.3.2: Primer sequences used to construct mutant genotypes. The forward inner primers (those named with inner For) include a single capital letter (in bold) which coded for the site directed mutation of interest. Lower case letters represent sequence homology or spacers, the un-bolded capital letters represent restriction enzyme (Res. Enz.) recognition sequence.

Name	Primer Sequence	Res. Enz.
gyrA_inner_For	gcgGGTCTCgcgaca**T**cgcggtctacgacaccatc	BsaI
gyrA_outer_Rev	gcgGGTCTCAGCAAgccaccacgttgatgccg	BsaI
gyrA_inner_Rev	gcgGGTCTCtgtcgccgtgcgggtgg	BsaI
gyrA_outer_For	gcgGGTCTCAGTCGgcgaggacatcccgatcgaag	BsaI
parC_outer_For	gcgGAAGACaCAATTGACTAGTatcatcccctaaccagcgcc	BbsI, MfeI, SpeI
parC_inner_Rev	gcgGAAGACcggcctgctacgaggcc	BbsI
parC_S87L_inner_For	gcgGAAGACgcaggcc**A**agtcgccgtgcgggtgg	BbsI
parC_S87W_inner_For	gcgGAAGACgcaggcc**C**agtcgccgtgcgggtgg	BbsI
parC_outer_Rev	gcgGAAGACGGCGCGCCtactacgccctcgacgaagc	BbsI, AscI

Table B.3.3: Correlation between know resistance determining loci and resistance phenotype. With knowledge of the resistance phenotype for strains in our alignment, we counted which strains had mutations in the codon of amino acid (AA) positions known to confer fluoroquinolone resistance. We used a chi-squared test for the independence between mutation at these positions and the resistance phenotype and highlight significant ($P \leq 0.05$) values in bold. Recall there were 389 strains in our alignment of which 192 (197) were susceptible (resistant).

Gene	AA	Mutants	Resistant	X^2	df	P
$gyrA$	83	161	156	231.941	1	$\mathbf{2.250} \times 10^{-52}$
	87	34	33	30.110	1	$\mathbf{4.083} \times 10^{-08}$
$gyrB$	466	16	14	7.596	1	**0.006**
	468	9	9	7.072	1	**0.008**
$parC$	87	128	127	177.208	1	$\mathbf{1.979} \times 10^{-40}$
	91	6	6	4.103	1	**0.043**
$parE$	457	8	7	3.061	1	0.080
	459	4	3	0.227	1	0.634
	473	11	7	0.323	1	0.570
$morA$	563	14	7	0.000	1	1.000
	975	15	6	0.333	1	0.564
	1056	15	6	0.333	1	0.564
	1109	15	7	0.003	1	0.960
	1155	16	9	0.041	1	0.839
	1162	15	8	0.000	1	1.000
	1213	15	7	0.003	1	0.960

Table B.3.4: Evidence of relative ΔG epistasis among the strongly correlated pairs of synonymous intragenic substitutions. Epistasis is measured with a multiplicative model and error is calculated using error propagation [Trindade *et al.* 2009]. There is evidence for epistasis when the absolute value of is greater than the error of our measures. We find no evidence for epistasis as ϵ is never greater than estimates of error.

Mutant Pair		ϵ	Error
dnaN c495t	*dnaN* c504t	-9.263 $\times 10^{-4}$	2.387 $\times 10^{-2}$
gyrB c1422t	*gyrB* c1443t	-1.481 $\times 10^{-4}$	4.768 $\times 10^{-3}$
morA a4041g	*morA* c4083t	1.113 $\times 10^{-2}$	3.394 $\times 10^{-2}$
parC c1533t	*parC* c1581t	-7.982 $\times 10^{-3}$	2.794 $\times 10^{-2}$
parC c1533t	*parC* c1587t	8.039 $\times 10^{-4}$	2.514 $\times 10^{-2}$
parC c1533t	*parC* t1554g	-1.798 $\times 10^{-2}$	3.166 $\times 10^{-2}$
parC c1581t	*parC* c1587t	4.633 $\times 10^{-4}$	2.635 $\times 10^{-2}$
parC t1554g	*parC* c1581t	-6.985 $\times 10^{-3}$	3.529 $\times 10^{-2}$

Table B.3.5: Evidence of ITE epistasis among the strongly correlated pairs of synonymous intragenic substitutions. Epistasis is measured with a multiplicative model and error is calculated using error propagation [Trindade *et al.* 2009]. There is evidence for epistasis when the absolute value of is greater than the error of our measures. We find negative epistasis for the sites in $morA$ based on ϵ having an absolute value greater than the estimates of error.

Mutational Pair		ϵ	Error
dnaN c504t	*dnaN* c495t	1.4113×10^{-5}	3.234×10^{-3}
gyrB c1443t	*gyrB* c1422t	5.330×10^{-5}	1.334×10^{-2}
morA c4083t	*morA* a4041g	$\mathbf{-4.437} \times 10^{-2}$	$\mathbf{1.053} \times 10^{-2}$
parC c1581t	*parC* c1587t	1.390×10^{-5}	4.606×10^{-3}
parC c1533t	*parC* c1587t	1.362×10^{-5}	7.304×10^{-3}
parC t1554g	*parC* c1581t	1.769×10^{-6}	2.770×10^{-3}
parC c1533t	*parC* c1581t	3.148×10^{-5}	6.041×10^{-3}
parC c1533t	*parC* t1554g	-1.025×10^{-5}	5.360×10^{-3}

Table B.3.6: Test for independence between the relative position of correlated substitutions and the type of pair they form. We used a χ^2 test and present the observed (expected) counts for the number of significantly correlated pairs ($P \leq 10^{-4}$) of substitutions based on if they were in the same gene and how many of the substitutions were synonymous. Analysis led to a test statistic of 7246.04 with 2 degrees of freedom and $P \sim 0$.

Within the same gene	**Type of Pair**		
	Non Synonymous	**One Synonymous**	**Synonymous**
TRUE	0 (0.52)	6 (1.83)	60 (1.63)
FALSE	101 (1289.31)	2274 (4569.33)	7470 (4048.38)

Table B.3.7: Absorbance (600 nm) readings from ciprofloxacin MIC assay of *P. aeruginosa* WT and mutant constructs. Readings were taken after 24 hours of growth on a 96 well plate where wells contained half salt LB growth medium with ciprofloxacin concentration as denoted by the column heading. The values presented are the mean (standard deviation) values for a minimum of 4 replicates. No genotype appeared more than once on a single 96 well plate.

		Concentration of Ciprofloxacin ($Log_2\,\mu g/mL$)											
Background	Mutation	5	4	3	2	1	0	-1	-2	-3	-4	-5	-6
PA01	WT	0.01	0.021	0.02	0.012	0.127	0.021	0.04	0.326	0.417	0.536	0.675	0.916
		(0.006)	(0.004)	(0.002)	(0.001)	(0.156)	(0.004)	(0.02)	(0.034)	(0.001)	(0.074)	(0.273)	(0.006)
	gyrA c248t	0.013	0.017	0.1	0.398	0.521	0.512	0.754	0.764	0.772	0.779	0.814	0.912
		(0.004)	(0.007)	(0.042)	(0.149)	(0.087)	(0.029)	(0.134)	(0.087)	(0.078)	(0.091)	(0.068)	(0.09)
	parC c260t	0.016	0.031	0.036	0.014	0.053	0.034	0.059	0.447	0.468	0.504	0.639	0.894
		(0.003)	(0.017)	(0.034)	(0.006)	(0.057)	(0.025)	(0.019)	(0.03)	(0.037)	(0.052)	(0.06)	(0.122)
	parC c260g	0.015	0.034	0.022	0.018	0.014	0.029	0.098	0.422	0.463	0.499	0.698	0.859
		(0.005)	(0.018)	(0.006)	(0.001)	(0.002)	(0.019)	(0.114)	(0.057)	(0.053)	(0.06)	(0.063)	(0.129)
	gyrA c248t,	0.418	0.76	0.843	0.803	0.845	0.805	0.754	0.748	0.686	0.807	0.774	0.854
	parC c260t	(0.116)	(0.094)	(0.138)	(0.113)	(0.172)	(0.13)	(0.145)	(0.138)	(0.138)	(0.085)	(0.091)	(0.139)
	gyrA c248t,	0.376	0.598	0.629	0.768	0.814	0.836	0.918	0.81	0.826	0.854	0.881	0.982
	parC c260g	(0.386)	(0.179)	(0.116)	(0.132)	(0.057)	(0.07)	(0.059)	(0.099)	(0.111)	(0.038)	(0.033)	(0.071)
PA14	WT	0.044	0.054	0.063	0.084	0.094	0.1	0.02	0.02	0.04	0.285	0.615	0.936
		(0.006)	(0.005)	(0.01)	(0.012)	(0.017)	(0.093)	(0.004)	(0.003)	(0.019)	(0.22)	(0.163)	(0.086)
	gyrA c248t	0.07	0.079	0.087	0.314	0.25	0.133	0.264	0.864	0.895	0.952	0.899	1.096
		(0.013)	(0.005)	(0.036)	(0.071)	(0.034)	(0.05)	(0.145)	(0.122)	(0.054)	(0.059)	(0.065)	(0.035)
	parC c260t	0.04	0.064	0.096	0.12	0.114	0.07	0.022	0.033	0.025	0.226	0.659	0.819
		(0.007)	(0.004)	(0.008)	(0.028)	(0.018)	(0.013)	(0.003)	(0.013)	(0.003)	(0.085)	(0.119)	(0.169)
	parC c260g	0.045	0.095	0.1	0.13	0.121	0.07	0.028	0.04	0.03	0.249	0.649	0.961
		(0.004)	(0.025)	(0.058)	(0.06)	(0.047)	(0.04)	(0.006)	(0.003)	(0.004)	(0.086)	(0.1)	(0.145)
	gyrA c248t,	0.279	0.294	0.237	0.769	0.996	0.957	0.941	0.956	0.933	0.948	0.927	0.941
	parC c260t	(0.012)	(0.029)	(0.038)	(0.184)	(0.077)	(0.056)	(0.019)	(0.018)	(0.027)	(0.028)	(0.018)	(0.012)
	gyrA c248t,	0.195	0.25	0.268	0.809	1.054	0.956	0.92	0.926	0.921	0.921	0.932	0.987
	parC c260g	(0.078)	(0.041)	(0.096)	(0.064)	(0.073)	(0.048)	(0.056)	(0.022)	(0.035)	(0.025)	(0.038)	(0.125)

Appendix C

Appendix to Chapter 4

C.1 Detailed overview of parameters and main functions

The framework of rSHAPE was written for, and implemented in, R R Core Team [2016]. Nearly all functions imported by rSHAPE are part of the base R package. The exceptions being the SQL interface provided by *DBI* R Special Interest Group on Databases (R-SIG-DB) *et al.* [2016] and *RSQLite* Müller *et al.* [2017], the probability generating functions provided by *evd* Stephenson [2002], *VGAM* Yee and Wild [1996] and *sn* Azzalini [2016], the parallelisation functions of *foreach* Microsoft and Weston [2019], and the data concatenation function of *abind* Plate and Heiberger [2016].

C.1.1 General Parameters

The parameters described here are not an exhaustive list but instead those most likely to be of general interest for an rSHAPE user. It is further worth noting that

some parameter combinations are intentionally redundant. This overparameterisation was intentional to facilitate the implementation of different models in rSHAPE.

A key parameter for any evolution experiment is its duration which rSHAPE controls through the number of generations T to be simulated. In each generation, rSHAPE sequentially calls the stochastic events function, and the functions for deaths, births, and mutation events. Each of the deaths, births, and mutation functions has an associated probability parameter ($P_{d,b,m}$ respectively) which controls the per generation probability of any individual having an associated event. If P_b is anything but 1, the term "generation" loses the traditional biological meaning of an average period of time between the birth and reproduction of an individual with average fitness. For convenience I herein use the term generation and time step interchangeably to mean a discrete unit of time simulated by rSHAPE.

The deaths and births of a population sum to represent their growth and the current growth models implemented in rSHAPE can simulate constant, exponential or logistic growth models. In rSHAPE, growth of evolving populations is calculated with the growth function that is a wrapper for the death and birth functions. Prior to running rSHAPE, users will have defined a focal number of individuals N_f which is interpreted differently by the context of the growth model. In rSHAPE there are two constant number of individual growth models, named *Poisson* and *Constant*, wherein the focal number of individuals defines the starting number of individuals. The *Poisson* model comes from the Galton-Watson branching process Kimmel and Axelrod [2002] and assumes $P_d = P_b$ and that the intrinsic growth rate r is two. This model will result in the population having a roughly constant size over many replicates, but in any one replicate this number is likely to deviate throughout time (T). I developed the *Constant* model to strictly enforce a constant number of

individuals. The *Constant* model calculates the per genotype proportional number of births using the *Poisson* model and then scales these values to sum to the number of deaths. Note that for constant growth models the stochastic events function is ignored. For exponential growth, N_f is the starting number of individuals and is used as the target number of individuals resulting from stochastic loss events. Under logistic growth, N_f represents the environmental carrying capacity (*i.e.* K).

Throughout a run of rSHAPE, mutants arise with probability P_m which is multiplied by either the number of births in a time step or, as one theoretical paper suggests, the total number of individuals Desai and Fisher [2007]. An individual's genome has a constant length L of binary state sites where 0 is the wild-type (WT) state and 1 the mutant. Sites in the genome have no explicit meaning other than that each are similarly mutable genomic regions which may affect genotype fitness. Users may define if revertant mutations are permitted (*i.e.* if 1 can mutate to 0). The fitness w_i of a genotype i (where $i \in \{1, 2, 3, ..., x\}$ and x is the number of unique living genotypes) is calculated given the fitness landscape model parameter that often depends upon a parameterised random effect distribution. The fitness landscape models currently implemented are: Additive, Fixed, House of Cards (HoC) Kingman [1978], Kauffman's NK (NK) Kauffman and Weinberger [1989], or Rough Mount Fuji (RMF) Aita *et al.* [2000]; Neidhart *et al.* [2014]. The additive model draws the effect (*i.e.* selective coefficient) of mutations from the random effect distribution and calculates w_i as the sum of the mutation effects (*i.e.* no epistasis). The fixed model is only practical for small genotypes as it requires that the user supply a matrix defining the fitness value for each genotype, but it does allows users experiment with defined (*i.e.* fixed) fitness landscapes. A full description of the HoC, NK and RMF models are beyond the scope of this work but in brief each calculates w_i using values drawn from the random effect distribution, and some

require additional constants that can be defined in rSHAPE. At present, rSHAPE implements the following probability generating functions for use as the random effect distribution: Beta, χ^2, Exponential, Frétchet, Gamma, General Extreme Value Distribution Stephenson [2002], Normal, Reverse Weibull, Skew Normal Azzalini [2016] and Uniform.

Once initial parameters are defined and a run starts, each generation will begin with a call to the population disturbance function that implements stochastic loss events.

C.1.2 Population Disturbance

This function is used to simulate disturbance events that stochastically reduce the total number of individuals. Disturbance events occur based on a schedule defined either by a fixed number of generations between events or where subsequent events are scheduled so that the expected number of births prior to the next event balances the sum of lost individuals. The expected number of individuals lost for each genotype i is proportional to their frequency within the evolving population. Using the disturbance factor D, the new expected number of individuals of each genotype i is calculated as $\frac{N_i}{D}$. This process is stochastic because the actual number of remaining individuals is calculated as random draw from a Poisson distribution with location parameter $\frac{N_i}{D}$. For simulations using exponential growth, the value D is controlled by N_f such that:

$$D = \sum_{i=1}^{x} N_i / N_f \tag{C.1}$$

where N_i is the number of individuals of genotype i. This prevents exponential population growth from growing indefinitely provided there are population disturbance events. As constant growth models ignore population disturbance

events, the user can only meaningfully set D when growth is logistic. When population growth is logistic, D can either be a constant value or be randomly drawn from a normal distribution parameterised by the user. Any value $D < 1$, either as a constant or resulting from random draws, will automatically be set to one since this function should never increase population sizes.

Next rSHAPE will calculate growth of the evolving population.

C.1.3 Growth

In rSHAPE, deaths and births happen simultaneously meaning that in a single time step an individual may both die and produce offspring. However, deaths are calculated first in order to inform certain growth model calculations.

Death Events

The probability of death P_d controls the proportion of individuals that die in a generation. The user can choose if this probability is a constant value or if it is applied in a density dependent manner. When P_d is constant, the number of deaths for a genotype i is given by:

$$deaths_i = N_i \, P_d \tag{C.2}$$

where N_i is the number of individuals of genotype i. If deaths are density dependent the number is calculated by:

$$deaths_i = N_i \, P_d \, (\frac{\sum_{i=1}^{x} N_i}{K_d})^{c_d} \tag{C.3}$$

where K_d is the population size at which P_d is 100% its defined value. The exponent c_d is used to scale the product of P_D and the ratio of population size and K_d, where larger c_d causes P_d to have little impact until the population is close to K_d. Please

note that K_d is defined separately from the logistic growth carrying capacity K (*i.e.* N_f).

This function is called prior to births but does not directly affect the population sizes (N_i) used in birth calculations. The number of deaths is first calculated so that births can be scaled to deaths such as under conditions of the *Constant* growth model.

Birth Events

The per generation probability of any individual giving birth is controlled by P_b. The number of offspring generated by an individual is calculated given the growth model parameterised by the intrinsic growth rate r, and the genotype's fitness w_i. For the *Constant* growth model, the number of births for each genotype is proportional to their size N_i and is calculated in two steps. The first is to calculate the absolute birth potential of each population by

$$births_{potential_i} = N_i \left(1 + w_i - \bar{w}\right) P_b \tag{C.4}$$

where the term $(1 + w_i - \bar{w})$ calculates relative fitness centred around 1 using the mean fitness $\bar{w}$. This method of calculating relative fitness handles instances when $\bar{w} = 0$ but is sensitive to the magnitude of $\bar{w}$. The user can choose relative fitness to be calculated using the more traditional $\frac{w_i}{\bar{w}}$, but this will be overridden if $\bar{w} = 0$. The second step in calculating the *Constant* growth uses the potential births ($births_{potential_i}$) calculated in eq. C.4 as weightings in the final calculation:

$$births_i = \frac{births_{potential_i}}{\sum_{i=1}^{x} births_{potential_i}} \sum_{i=1}^{x} deaths_i \tag{C.5}$$

where the actual number of births is $births_{potential_i}$ scaled to the sum of deaths. If no potential births occurred, or there were no deaths, then the second step is

skipped and $births_i$ is returned as a vector of zeroes. If the *Poisson* growth model was chosen, then $births_i$ is obtained through draws from a Poisson distribution where the location parameter is given by the product of N_i, w_i, and P_b. This approach was derived from the theoretical work of *Haldane* (1927) and assumes a large population and that $P_b = P_d$.

To calculate births when growth is either exponential or logistic growth, then the intrinsic growth rate r is an additional parameter controlling the number of births for each population. For exponential growth the expected number of births for each genotype are calculated as

$$births_i = N_i \, (e^{ln(r) \, w_i \, P_b} - 1) \tag{C.6}$$

where $ln(r)$ is the natural logarithm of the intrinsic growth rate and w_i is the fitness of a genotype. This equation is derived from the exponential growth model but I subtract 1 from the growth term to result in a calculation for births. Note that if $e^{ln(r) \, w_i \, P_b} 1$ then the result it is set to zero to prevent population decrease being calculated by the birth function. When growth is logistic, the expected number of births for each genotype is calculated in two steps. First, a density dependent growth term dg_i for each genotype i is calculated using the logistic equation

$$dg_i = N_i + (w_i \, r \, P_b) \frac{K - \sum_i^x N_i}{K} \tag{C.7}$$

where K represents the carrying capacity. This density dependent term dg_i represents the amount of growth expected for genotype i given the current total number of individuals. Using dg_i, the number of births for each genotype is

calculated as

$$births_i = N_i \left(\frac{dg_i}{\sum_i^x N_i} - 1\right) \tag{C.8}$$

where similar as to with exponential growth, I subtract 1 to calculate births rather than growth. Recall that while deaths are calculated prior to births they do not affect N_i used in these calculations. The exponential and logistic growth model birth calculations are deterministic and so to make growth calculated by rSHAPE a stochastic process, users can toggle that "drift" be considered. "Drift" causes the number of births to become random draws from a Poisson distribution using location parameters set by the deterministic birth values. Also, since births and deaths are calculated separately, but because users may want growth to perfectly simulate deterministic growth models, users may set the number of births calculated to be scaled by deaths. In this case, any deviations will be adjusted by generating additional births, or deaths, as required for each population using a nested call to the growth function with the *Constant* method.

Once growth has been calculated, rSHAPE will determine if mutants are generated.

C.1.4 Mutation Events

The last step of each generation is to calculate if there have been mutation events. The number of mutations is controlled principally by the per genome, per generation, mutation rate μ. Classically the number of mutants is a product of μ and the number of replication (*i.e.* birth) events, but more recent theoretical work has suggested using all living individuals Desai and Fisher [2007]. Either choice may be simulated with rSHAPE by selection of a logical toggle parameter. The number

of mutants generated from replication events is calculated as:

$$mutants_i = \mu \, births_i \, \frac{r}{r-1} \tag{C.9}$$

where recall r is the number of offspring expected from a single birth event and the term $\frac{r}{r-1}$ ensures that the number of mutants considers not just offspring but also the parental individuals since either may mutate. If mutants can arise from any individual in the population, rSHAPE adds the following vector:

$$\mu \, (N_i - deaths_i - \frac{births_i}{r-1}) \tag{C.10}$$

to the values calculated in eq. C.9. The first time any genotype generates mutants, rSHAPE will calculate, and permanently record, the fitness value of all genotypes in the unexplored neighbouring mutational space. The fitness for a genotype is calculated based on the chosen fitness landscape model (see *General Parameters* above). For each genotype i with at least one mutant, rSHAPE draws $mutant_i$ times (with replacement) from the list of genotypes in the neighbouring mutational space.

C.2 Detailed comparisson to theoretical work

C.2.1 Haldane's $2s$

The seminal theoretical work of Haldane Haldane [1927] approximates the probability of fixation for a single mutant as being $\approx 2s$ (where s is the selection coefficient). Haldane's work makes the assumption that the WT population is very large, that the number of individuals is constant, that s is small ($s \ll 1$), that generations do not overlap, and that a successful mutant lineage must grow from a

single progenitor mutant. I replicated these conditions in rSHAPE with each of the two appropriate growth models: *Poisson* and *Constant*. The first model is identical to the method used by Haldane Haldane [1927] and is called the *Poisson* form because it is based on a Galton-Watson branching process Kimmel and Axelrod [2002] and estimates birth events as draws from a Poisson distribution. This method of simulating births causes a populations size to be approximately constant over many replicates but allows population size to vary within any one replicate. In microbial experimental evolution, growth conditions can force a population to be roughly constant such as in a chemostat where there is a carrying capacity and nutrient flow is limited. The *Constant* growth form better simulates these conditions by scaling the births of the *Poisson* form to be exactly the number of deaths in a generation.

Using a range of selective coefficients,
$s \in \{0.001, 0.005, 0.01, 0.03, ..., 0.09, 0.1, 0.15, ..., 0.3\}$, I calculated the fixation probability from 1,000,000 replicate runs of rSHAPE. Haldane's $2s$ approximation comes from eq. C.11 (eq. 1 of Haldane's theoretical work Haldane [1927])

$$s = \sum_{i=2}^{\infty} \frac{prob_{fix}^{i-1}}{i} \tag{C.11}$$

where $prob_{fix}$ is the probability of fixation. The classic approximation $prob_{fix} \approx 2s$ considers only the first term of eq. C.11 however, the analytical expression of $1 - e^{-2s}$ provides a closer approximation (personal communications with Dr. Lindi Wahl). We can calculate the exact value for $prob_{fix}$ related to s (up to finite level of precision) by exploring the space of $prob_{fix}$ values and plugging

them into the rearranged, and un-simplified form, of eq. 1 from Haldane's work:

$$s = \frac{-ln(1 - prob_{fix})}{prob_{fix}} - 1 \tag{C.12}$$

the search space should be centered around $2s$ and can continue to adjust by $prob_{fix} \pm \delta x$ until s is found. I found that both the *Poisson* and *Constant* growth forms implemented in rSHAPE accurately predict exact fixation probabilities calculated in this way (Fig. 4.4.3A, main text). As s increases, both the close approximation and $2s$ would suggest higher fixation probabilities. One of the assumptions of Haldane's work is that s is small. As "small" is a subjective term, I compared $2s$ against the true fixation probability finding that $2s$ overestimated fixation probability by $\sim 5\%$ when $s = 0.05$ and that this difference increased with s (Fig. 4.4.3B, main text). In fact only when $s < 0.01$ does the $prob_{fix}$ approximation of $2s$ differ by less than 1.3% from the exact value. This result suggests that s be considered small, for theory building upon Haldane's work, only when $s \leq 0.01$.

The work described here was for a large population ($N_f = 10^8$), but I did also run simulations with a smaller population ($N_f = 10^4$) and found that the absolute probability of fixation values were still quite in line with the exact values (Fig. C.3.1A). However, the normalised difference between the exact and estimated probability of fixation values differed significantly, for the small population only, when $s \leq 0.03$ (Fig. C.3.1B).

C.2.2 Logistic growth

When grown outside of a chemostat, microbes grown experimentally tend to follow a logistic growth pattern with an upper bound (*i.e.* carrying capacity K) caused either by density dependent death or reduction in birth rate such as when nutrients

are depleted. Theory suggests that the fixation probability of a single *de novo* mutant is greater than $2s$ relative to the growth rate of the population and becomes $\sim 2s$ once K is reached Ewens [1967]; Otto and Whitlock [1997]. Implicitly, this assumes that while the community is at K births continue but are balanced by deaths. In my simulations replicating the work of Otto and Whitlock [1997], the assumptions are all similar to those of Haldane's $2s$ but, the growth rate r is assumed to be small. I performed simulations with the same range of s as previously described and that range was also used for r. Populations were initiated at $\frac{K}{100}$ and I simulated growth for two carrying capacities $K \in 10^{(4,8)}$ but only discuss the results for the larger as both are similar (Fig. C.3.2). To calculate the theoretical fixation probability, I used the analytical approximation shown in eq. C.13 (eq. 11 from Otto and Whitlock [1997])

$$prob_{fix}(t) \approx \frac{2sK(s+r)}{sK + rN_{WT}(t)} \tag{C.13}$$

where $prob_{fix}(t)$ is the fixation probability given the mutant arose at time t when there were N wild-type individuals in an environment with carrying capacity K. Estimates with rSHAPE were calculated from 1,000,000 replicates.

When both s and r are within the range of values presented in the work of Otto and Whitlock Otto and Whitlock [1997], rSHAPE accurately reproduces the analytical approximation except once either is equal to or greater than 0.05 where rSHAPE estimates visibly lower fixation probabilities (Fig. SC.3.2). Similar to my findings when comparing $2s$, I find that rSHAPE underestimates the probability of fixation by an amount proportional to the value of either r or s when they are not "small" (*i.e.* the assumptions are violated).

C.2.3 Serial Passaging

The laboratory practice of serial passaging, common to microbial experimental evolution, induces regular population bottlenecks as a small proportion of a community grown in liquid media is transferred to new media for continued growth. The ratio of new to old volume is commonly expressed as the dilution factor D and the period between transfers is known as the growth phase. Prior to transfer, the old media should be well mixed so that the existing genotypes are proportionally transferred. However, these repeated bottlenecks result in rare genotypes being lost Wahl *et al.* [2002] because in practice sampling is still a random process and even when samples are perfectly mixed we are not likely to transfer any genotype with less than D individuals. Through the stochastic events function, rSHAPE can simulate serial passaging which permitted this last comparison. Current theory has suggested that the probability of a mutant arising and eventually fixing is roughly uniform throughout the growth phase when a population grows exponentially Wahl *et al.* [2002]. After validating that rSHAPE accurately simulates serial passaging (Fig. SC.3.3), I used it to estimate the combined probability that a mutant will arise at some point throughout a growth phase and then eventually fix.

To make the maths tractable, the work of Wahl *et al.* [2002] assumed that growth was effectively exponential, that there was no competition and that deaths could be ignored. I ran simulations using these conditions and the same parameters from the exponential and nutrient limited growth models of Wahl *et al.* [2002]. The later conditions were applied to logistic growth in rSHAPE.

For the exponential (logistic growth) models, I began independent simulations by seeding a single mutant ($s = 0.1$) into a population of $10^5(10^7)$ individuals at each of the $\tau = 7$ (6) generations in the growth phase. The basal growth rate was $r = ln(2)$

($r = 2$) and each run ended when the mutant was lost, had fixed, or one million serial passaging events of $D = 100$ had occurred at which point fixation was assumed imminent. To estimate the fixation probability, each combination of parameters was replicated 1,000,000 times. I compared estimates against the analytical approximation for survival (eq. C.15) which depends upon the extinction probability ($P_{extinction}$) given by eq. C.14 Wahl and Gerrish [2001]; Wahl *et al.* [2002]

$$P_{extinction} \approx 1 - 2se^{-rt}r\tau \tag{C.14}$$

where t is the generation of the growth phase during which the mutant arises. The probability of a mutant with selective coefficient $s = 0.1$ having been born at time t was calculated as the product between the mutation rate $\mu = 5 \times 10^{-5}(4 \times 10^{-9})$, the number of births in a generation, and the probability density of an exponential distribution with rate $\alpha = 100$ (value was not published in Wahl *et al.* [2002] and is as per personal communications with Dr. Wahl) and subsequently survives. So, the analytical approximation of a mutant arising during the growth phase and ultimately surviving multiple serial passaging events was calculated with eq. C.15 Wahl *et al.* [2002]

$$P_{survival} \approx births(t)\,\mu\alpha e^{-\alpha s}(1 - P_{extinction}) \tag{C.15}$$

where the number of births is the difference in population size between generations t and $t - 1$ (because deaths are ignored). When growth is exponential, rSHAPE estimates a roughly uniform joint probability throughout the growth phase though the exact value is lower than the analytical approximation. This lower estimate is not surprising as the theoretical work being compared is built upon Haldane's $2s$ approximation.

When growth was logistic, rSHAPE estimated a joint probability that is not uniform. This result suggests, similar to other theoretical work Wahl and Zhu [2015], that the protocol of serial passaging can affect which mutations are fixed during an experiment. Because microbial populations used in experimental evolution tend to grow logistically, the practice of serial passaging will bias the fixation of mutations arising early during growth. Early in the growth phase there are fewer mutants being generated which means the mutational space will be less well explored and so of those mutants which arise and fix may be lottery winners instead of mutationally optimal alternatives. It could be argued that if the time to fixation were long then we might expect the dynamics to replicate maximal clonal interference as all mutants eventually arise multiple times and compete for fixation. Under maximal clonal interference we expect mutationally optimal outcomes as selection coefficients drive evolution. However, previous work suggests that the dynamics of microbial experimental evolution reflect intermediate clonal interference Bailey *et al.* [2016] whereby some, but not all, possible mutants compete for fixation and so the order in which they appear matters (*e.g.* timing during growth phase). Other studies have shown how the order in which mutations appear can affect evolutionary outcome Kvitek and Sherlock [2011]; Sackman and Rokyta [2017]; Weinreich *et al.* [2005]. With rSHAPE I have shown that the practice of serial passaging introduces a jackpot scenario where those few mutants arising early during growth phases are most likely to survive. This finding suggests that the stochasticity in outcome between replicate microbial experimental evolution populations is at least somewhat due to the practice of serial passaging. Future studies should use rSHAPE to quantify what proportion of stochasticity is attributable to serial passaging.

C.3 Figures

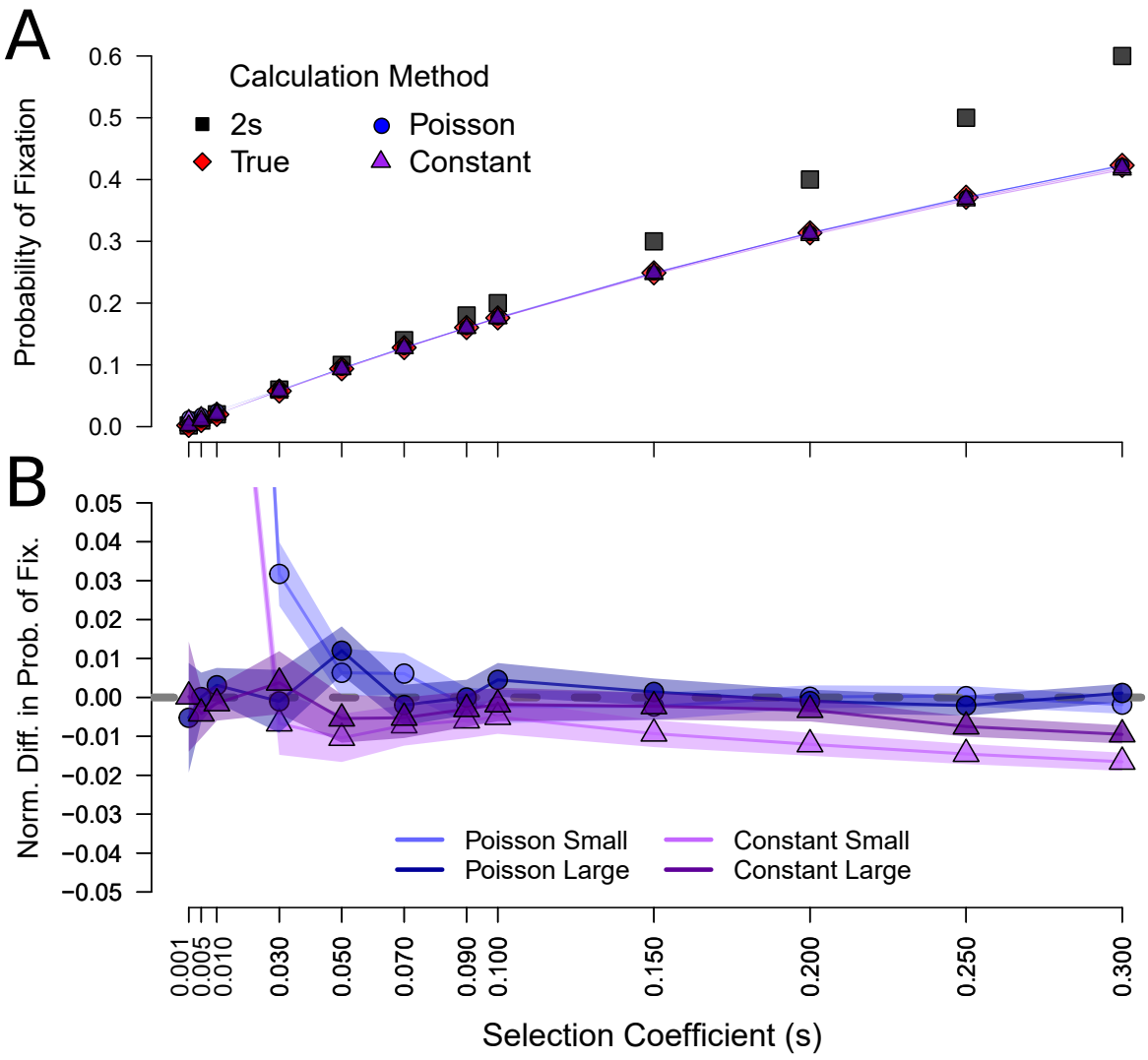

Figure C.3.1: Comparison between the theoretical and estimated probability of a single mutant fixing, in an environment with a constant number of individuals. The colour and shape of points identify the means for calculating/estimating fixation probability whereas the shading identifies if simulated estimates were performed with smaller (10^4) or larger (10^8) population sizes. Panel A shows the fixation probability, dependent on the selection coefficient s, and the red diamonds show the un-approximated theoretical expectation (*i.e.* true fixation probability). Panel B presents the normalised difference between the true fixation probability and the values calculated with rSHAPE. Negative values reflect when the values calculated with rSHAPE are lower than the true fixation probability.

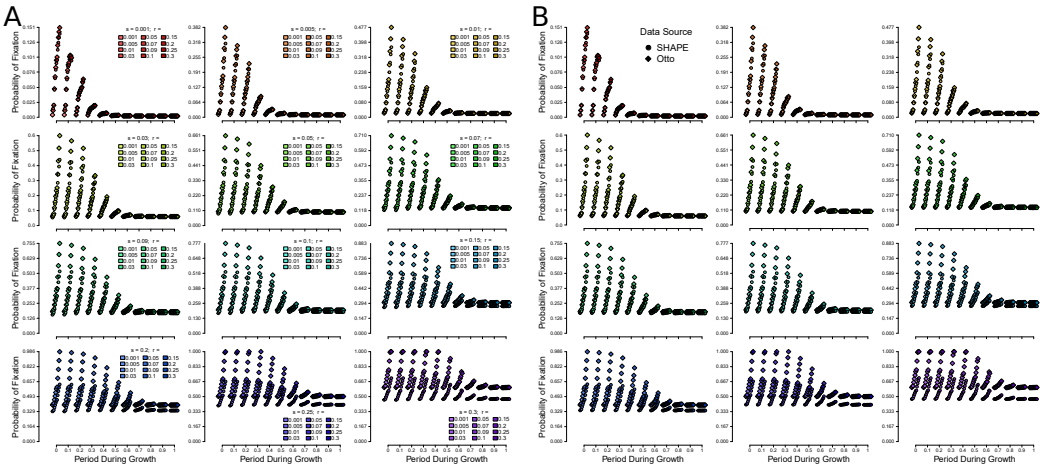

Figure C.3.2: Comparison of the theoretical and estimated fixation probability for a mutant growing logistically. Diamond shapes represent the analytical approximation of Otto and Whitlock [1997] while circles are for estimates using rSHAPE. The colour used to fill points reflects the selection coefficient s and darker colours represent higher intrinsic growth rates r. The period during the growth phase is scaled from the start of growth until the point where the number of individuals reaches carrying capacity. The range of parameters is similar between panels but A is for when $K = 10^4$ whereas B is for $K = 10^8$.

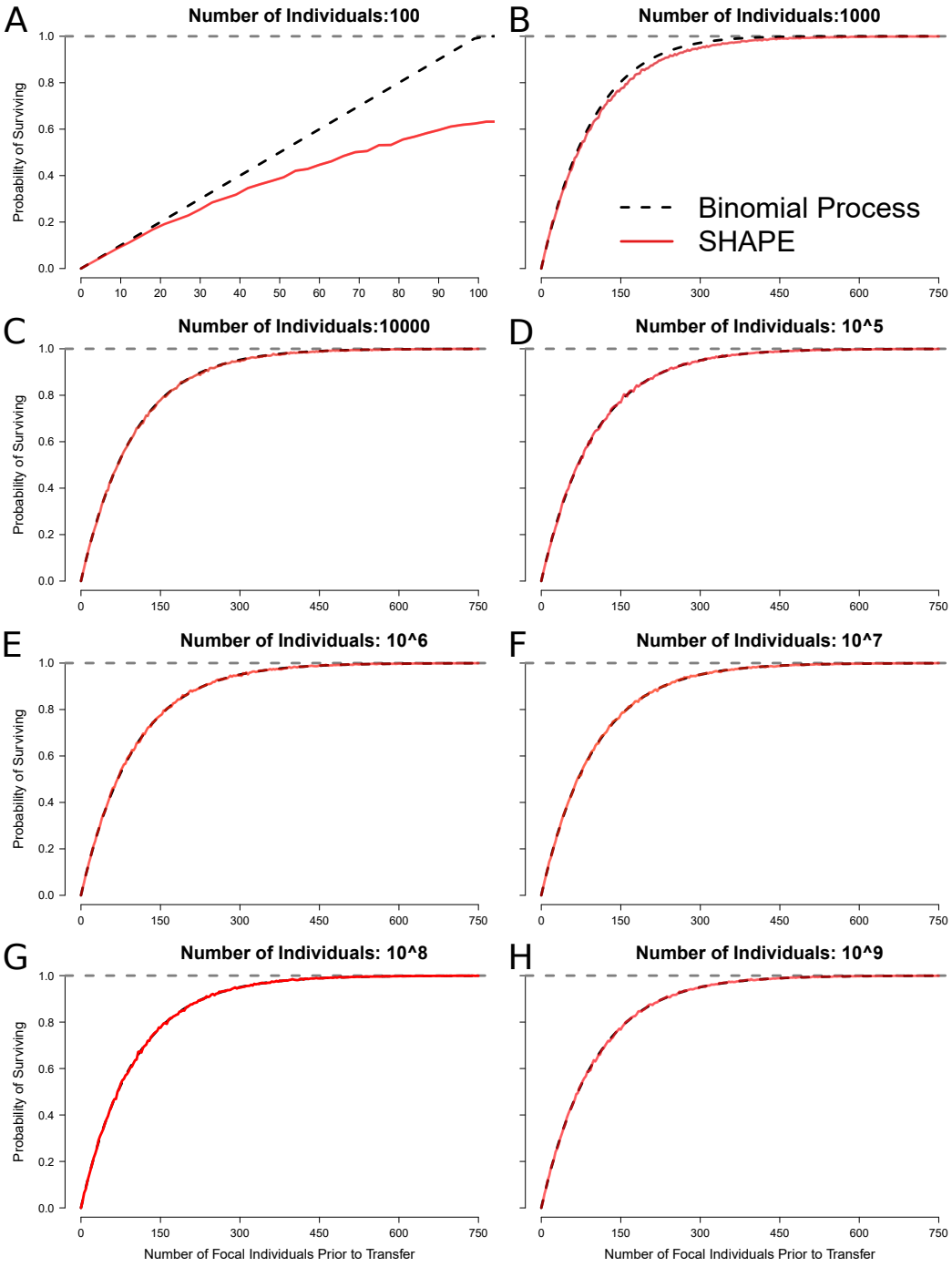

Figure C.3.3: The probability that at least one individual of a focal genotype survives a disturbance event (serial passaging) reducing the total number of individuals 100 fold ($D = 100$). The red polygon (appears as a line) shows the 95% CI of survival probability calculated with rSHAPE, while the black dashed line shows the expectation calculated from a binomial process. The dashed grey line highlights when probability reaches unity. Panels differ in the number of focal individuals, and total population size, prior to transfer.

C.4 Tables

Table C.4.1

List of definable evolutionary parameters in rSHAPE. While rSHAPE offers a variety of modifiable evolutionary parameters, all have default values and many interact with others as part of calculations (*e.g.* The probability of birth (P_b) and intrinsic growth rate (r) will both affect the number of offspring calculated). Where possible, default values were chosen to best reflect standard parameterisation of evolutionary models involving haploid asexuals. For more details, please refer to the reference material included with rSHAPE.

Context	Symbol	Parameter & Meaning
Individual	L	The number of genome positions.
Disturbance		Type of disturbance; either fixed or random bottlenecks.
Events	D	Dilution factor of the first disturbance event.
		Value(s) used to calculate size of disturbance events.
		Number of time steps between disturbance events.
Birth	P_b	Probability that an individual produces offspring in a time step.
		Growth model to be used.
	r	Number of individuals after a birth event (*i.e.*: parent + offspring)
		Logical toggle to add randomness to birth calculations.
	N_f , K	Focal population size used conditionally on growth model.
Death	P_d	Probability that an individual dies in a time step.
		Logical toggle controlling if deaths density dependent.
	c_d	Parameter controlling shape of density dependence.
	K_d	Population size at which 100% of density dependent deaths occur.
		Are births scaled to replace all deaths? Enforces adherence to growth model calculations.
Mutation	P_m	Probability of a mutant arising during a time step.
		Logical toggle: Is mutation probability only applied to individuals undergoing birth?
		Can mutations revert to wild-type state?
Fitness		Fitness landscape model used to calculate genotype fitness.
Landscape		Distribution used to draw random component of fitness calculations.
		Parameterisation of distribution used to draw random fitness components.
		Fitness value of wild-type genotype.
		Logical toggle: are genotype fitness calculations to be used as selection coefficients or relative fitness?
		For RMF fitness landscape model: number of mutations separating the wild-type and optimal genotypes.
		For RMF fitness landscape model: weighting of the independent fitness component.
		For NK fitness landscape model: Number of sites interacting.
Experiment	T	Number of discrete time steps to simulate.
	n	Number of replicates for a parameter combination.
		Proportion of a general time step calculated in each realised time step.
		Logical toggle: are individuals tracked as integer values or allow decimal tracking.