Full Title: The idiosyncratic drivers of correlated evolution

Short Title: Identifying correlated evolution and epistasis

Jonathan Dench[1*], Aaron Hinz[1], Stéphane Aris-Brosou[1,2‡], Rees Kassen[1‡],

1 Department of Biology, University of Ottawa, Ottawa, Ontario, Canada

2 Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

‡These authors contributed equally to this work.

* jdenc017@gmail.com

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.

Author summary

As we shift focus from the detection of independent, strong drivers of evolution, and consider the interacting effects (e.g. epistasis) of multiple factors, we see a growth of high-throughput computational approaches but a dearth of empirical validation. As the breadth and depth of sequencing data continues to grow, it has become important to ensure our growing repertoire of approaches is sufficiently validated prior to widespread usage. In this study, we apply a computational approach recently described to predict correlated evolution within whole-genome sequence data from *Pseudomonas aeruginosa* in the context of fluoroquinolone antibiotic selection. We then performed direct experimental analyses to identify the causes of correlated evolution. We find that very little of the most strongly correlated evolution is actually epistatic, and that instead, most correlated evolution relates to physical linkage likely driven by horizontal gene transfer.

Introduction

Adaptive evolution often involves changes to multiple characters that share a common genetic basis, 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 [1–3], molecular biology [4, 5], life history evolution [6–8], and comparative biology [9, 10]. 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 [11]. Although the original approach focused on identifying correlated evolution among sites within genes [11–13], 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 [9], and extends it to consider nucleotide positions by taking genome sequence data as traits to be compared. The output is a list of pairs of sites and an associated probability that each pair has co-evolved 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 epistasis as the cause of correlated evolution.

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. $gyrA$, $parC$) [14–17]. 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 which 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 tied to antibiotic resistance, whereas correlated evolution among synonymous sites is best explained by idiosyncratic processes.

Results

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 [17]. We then used this alignment to construct a phylogenetic tree rooted by PA7, an outlier strain [18] (S1 Fig.). The subclade containing this strain was removed for downstream analyses.

We subsequently employed AEGIS to identify pairs of nucleotide positions that evolve in a correlated manner. AEGIS makes use of Pagel's maximum likelihood model for predicting correlated evolution between pairs of traits, which in our analysis are the state of nucleotide positions [9]. As this approach can be computationally expensive, requiring on the order of n^2 comparisons in a genome of length n (with $n \sim 6.7 \times 10^6$), we focused our analyses on a subset of 12 genes against all other positions in the exome. Six of these 12 genes were chosen either because they were previously shown to evolve during fluoroquinolone selection [14, 15] ($gyrA$, $gyrB$, $nfxB$, $parC$, $parE$) or are linked to biofilm formation [19], which can contribute to antibiotic resistance [20] ($morA$). The six other genes ($dnaA$, $dnaN$, $lpd3$, $ribD$, $rpoB$, $serC$) were randomly selected from among the other 5,971 genes in our alignment.

AEGIS calculated the probability that each position within these 12 genes evolved in a correlated manner with any other nucleotide position in the entire *P. aeruginosa* exome. Among the more than 144 million pairwise comparisons, we found that 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) (S1 File). At medium

($P \leq 10^{-7}$) or high ($P \leq 10^{-11}$) significance levels, only 178 and nine pairs, respectively, showed evidence for correlated evolution (Fig. 1A). 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 [21–23, but see], save for one pair of nonsynonymous sites, *gyrA* (c248t)-*parC* (c260g/t), that presumably impacts the level of resistance (Fig. 1A).

Experimental validation of the nonsynonymous pair

To test if epistasis is involved in the correlated evolution of the substitutions identified in $gyrA - parC$, we first used a modified allelic replacement protocol [24] 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.

Our results, shown in Fig. 1B-C, reveal that the pattern of changes in resistance for single and double mutants was similar for both genotypes. 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 *parC* mutation 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 genetic background.

Notably, mutants carrying the $gyrA\,c248t$ substitution had the expected 8-32 fold increase in ciprofloxacin resistance [14]. However, the $parC\,c260g/t$ substitutions did not affect resistance on their own but did show a $\sim 16\times$ additional increase for double mutants (Fig. 1 B-C; Table 1). This strong epistasis for resistance may well explain why $gyrA\,c248t$ and $parC\,c260g/t$ double mutants are common among fluoroquinolone resistant clinical isolates of *P. aeruginosa* [14, 17] and is consistent with previous work showing that the fluoroquinolone resistance conferred by substitutions in $parC$ depended upon mutations in $gyrA$ [25]. We note that Akasaka*et al.* [14] 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. This result suggests a possible Bateson-like molecular mechanism of the observed epistasis for resistance [26]. 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. 1 D-E). 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 [16]. By contrast, all the single and double mutations in PA01 lead to significant fitness costs, and the double mutant including the *parC c*260*g* mutation, exhibits significant positive epistasis (Table 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 [16]. Combinations of mutations in *gyrA* and *parC* other than those tested here have been shown to be cost-free in *Streptococcus pneumoniae* [28], 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, but not 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 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 [27]. 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

Six of the genes in our 12 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?

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 [17, 29]. 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

Table 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 [27]. 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

correlated evolution [11], 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 [30]. 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 (S3 Fig.). 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 (S3 Fig.).

Second, we tested the association between fluoroquinolone resistance and nonsynonymous mutations at known fluoroquinolone resistance determining sites [14, 17]. 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: $gyrA\,248$, $X^2 = 231.941$, $df = 1$, $P = 2.250 \times 10^{-52}$; $parC\,260$, $X^2 = 177.21$, $df = 1$, $P = 1.98 \times 10^{-40}$, S1 Table). 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)%;S1 Table). Previous work showed that the signal of correlated evolution (i.e. the sensitivity of the AEGIS algorithm) is maximized when roughly half the strains are mutants, and when these mutations are evenly distributed throughout the phylogeny [11], 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 [15]. 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 [31, 32], and hence detect sites potentially involved in drug resistance, our alignment contains too few taxa and too little evolutionary depth for proper statistical analysis [33]. 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.

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 physically linked. While it has been known for some time that synonymous sites can experience purifying selection [34, 35] and a number of recent reports demonstrate that they can sometimes contribute to adaptation as well [22, 23], we find no signal of epistasis-driven selection acting on pairs of synonymous sites here. Estimates of the change in stability of mRNA transcripts and index of translation efficiency, which are proxies for translation rate [36] and efficiency [37], respectively, relative to the wild type are shown in Fig. 2 and S2,3 Tables. 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 S4 Table, S1 Appendix Fig. 1 C-F). 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 $P \leq 10^{-4}$; $t = -5.1784$, $df = 7,528$, $P = 2.29 \times 10^{-7}$ S4 Fig.).

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, which is near the quinolone-resistance determining region $gyrB$, which 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 positions along the gene that could have functional effects on regulating or sensing internal oxygen levels (positions 292, 293, 306) [38] or signal transduction (position 1356) [39]. While these results point to hitchhiking as a plausible explanation for the correlation between sites, we failed to detect additional functional links with nonsynomous mutations in the other genes. Moreover, AEGIS did not detect correlations between any of the correlated synonymous mutations and additional nonsynonymous mutations, even at $P \leq 10^{-4}$ (S1 File), suggesting that hitchhiking is not a general explanation for the correlated evolution of synonymous mutations in this data set.

A second possible mechanism is horizontal gene transfer (HGT). Examination of the similarity between the species tree and the gene trees for each of the four genes containing a pair of synonymous mutations revealed little correspondence between the two for all but *parC* (S5 Table), 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 tRNA genes are involved in information processing, and so interact with a large number of other gene products (transcripts and/or proteins). A successful HGT of such genes should therefore involve the co-transfer of most of their interacting partners as well, an unlikely event [40, 41]. As before, all tRNA gene trees differed from each other and the species tree (S5 Table). Taken together, 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. 1), it would seem that they have repeatedly evolved as a driverless cohort [42]. What would cause the repeated evolution of these four mutations remains unclear.

Conclusions

Our results show that correlated evolution can be widespread ($\sim 127,000$ interactions at $P \leq 0.01$) and that, in the context of antibiotic resistance in *P. aeruginosa*, it can be underlain by epistasis. This result is consistent with experimental work showing that multiple mutations can contribute to the evolution of antibiotic resistance [27, 43, 44], both within and between genes [45], 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 under antibiotic selection. Furthermore, our results lend support to the idea that high-throughput techniques for identifying correlated evolution *can* be underlain by epistasis. Similar interpretations have been made previously, for instance in the case of the evolution of RNA genes where base pairing is critical [46].

However correlated evolution is necessary but not sufficient evidence for epistasis [11], as it can also arise through physical linkage. 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 [21–23, 47–49], we have little evidence that synonymous substitutions, either on their own or in combination, impact fitness in our data set. The fact that many of the synonymous mutations we recovered here are GC→AT mutations 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, and mutational biases related to GC content, underlying correlated evolution between nonsynonymous substitutions [5]. Taken together, these results suggest that physical linkage may often be an important driver 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, some instances of correlated evolution in our data set appear to be unconnected to fitness or epistasis in these selective conditions. We therefore urge caution in interpreting instances of correlation between sites from high throughput computational analyses as solely due to epistasis.

Materials and methods

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` [**?**,]accessed Dec 4, 2014]Winsor:2016aa, as well as 390 draft *P. aeruginosa* genomes previously published [17]. These 390 draft genomes were selected because they were published along with information detailing which of the strains were resistant to the fluoroquinolone antibiotic levofloxacin. Due to the dissimilarity between contigs across the draft genomes, we were unable to construct a global alignment using whole genome alignment algorithms as implemented in MAUVE [50] or MUGSY [51]. This led us to develop a custom algorithm to reconstruct a concatenated alignment, using PA14 as a reference. For this we obtained an annotated reference database for PA14 from `www.pseudomonas.com` [**?**,]accessed Dec 4, 2014]Winsor:2016aa. From this database, we obtained the nucleotide sequence for the 5,977 genes in PA14. Each of them was used as a query for a BLASTn [52] search against a local database containing the 390 draft genomes, as well as the three complete genomes of PA01, PA7 and PA14. Results with less than 90% identity were discarded. These BLASTn results were then used to construct scaffolds for each of the 390 draft genomes, making sure that the same genomic information was not used more than once. Scaffolds were first built using genomic information which resulted in $\geq 90\%$ identity with a single reference gene. Then, the remaining results were inserted into scaffolds so that genomic information was used where it had highest identity. Following this scaffold assembly, the sequences of each gene were aligned with MUSCLE [53]. After alignment, we discarded all genes that were not conserved. We considered a gene not conserved if less than 50% of strains had $\geq 90\%$ alignment (non-gap characters) with the PA14 reference. The alignments of conserved genes were concatenated into a single exome alignment using the perl script *catfasta2phyml.pl* [54].

To estimate the species tree of these bacterial strains, we focused on 'core genes', *viz.*, genes that are unlikely to be subject to horizontal gene transfer. According to the

complexity hypothesis [40, 41], these are genes whose products are involved in complex multiprotein interactions such as information processing genes (e.g. rRNA genes, transcription and translation polymerases, etc...). Based on the functional annotations of PA14 genes from the Clusters of Orthologous Groups database [55, 56], we extracted 1,290 information genes which we identified with the COG terms A, B, J, K, and L. These were concatenated into a single alignment, used to estimate a phylogeny by maximum likelihood. For this we employed FastTree [57] compiled with *DUSE DOUBLE* as recommended for large sequence alignments. A GTR model of evolution was assumed [58]. This tree was rooted with the subclade containing the taxonomic outliers PA7 [18], and AZPAE14941, AZ-PAE14901 and AZPAE15042 [17]. After rooting the tree, we removed the subclade from both the phylogeny and the alignment, leading to our analysis of correlated evolution being based on 389 *P. aeruginosa* genomes.

Analysis of correlated evolution

We used AEGIS [11] to test for signals of correlated evolution among sites in our alignment. For this, AEGIS performed pairwise comparisons between nucleotide positions testing if a model of dependent evolution was significantly better than an independent model at explaining the observed phylogenetic distribution of nucleotide states. The analysis relied on Pagel's algorithm to detect correlated evolution [9], as implemented in BayesTraits [59], using the maximum likelihood and binary state options. To recode the nucleotide states into binary characters, we assigned '0' to nucleotide positions that were identical in at least 80% of levofloxacin sensitive strains, and '1' to all others. We chose this method of recoding so that signals of correlated evolution would more likely reflect adaptation to fluoroquinolone antibiotic selection.

After recoding the data, there were 410,665 informative (polymorphic) sites within

the alignment. We estimated that a full pairwise analysis would have required more than one year's time even when parallelized across twelve computational nodes with four processors each. Hence, we tested the signal of correlated evolution between polymorphic sites in twelve genes and the rest of the alignment.

Direct experimental tests for epistasis

Unless otherwise stated all growth and incubation conditions were performed in lysogeny broth (LB) [60] 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, we created mutant constructs using a modified allelic replacement protocol [24]. We modified the protocol such that replacement alleles were constructed from the wild-type (WT) template rather than coming from an evolved strain. This resulted in only the mutation of interest being inserted instead of all differences contained in an evolved strain's allele. The replacement protocol is based on a selection counter-selection method [61], whereby a vector-borne mutant allele recombines with homologous chromosomal sequences before the delivery vector is itself removed from the targeted genome. As the effect of a mutation may be contingent upon the genetic background (i.e. WT) in which it arose [62–64], constructs were created using two distinct genetic backgrounds: PA01, and PA14 [65].

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. An overlapping primer 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 [66], which

permitted scarless ligation of the PCR products (see Table 3 for details). The vector, derived from pAH79 [24] 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 pir). The mutant alleles were transferred into *P. aeruginosa* strains as previously described [24] via tri-parental conjugation, involving a helper *E. coli* that carries pRK2013 [67]. Through a round of selection (LB agar with 100 g/mL Nitrofurantoin and 80 g/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. Then 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 those with very similar fitness measures for downstream analyses.

Minimum inhibitory concentration assays

We performed minimum inhibitory concentration assays to identify if our mutant constructs had increased resistance to the fluoroquinolone antibiotic ciprofloxacin.

Table 3. 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

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 ($i.e.\, 2^{(5,4,\ldots,-6)}$) g/mL. We inoculated each well with $5\mu L$ 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 (S6 Table). 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 [68]. 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.

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 [27] such that $\epsilon = W_{WT}W_{AB} - W_A W_B$ where W stand 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.

Quantifying importance of sites in predicting resistance

Using an adaptive boosting machine learning algorithm, implemented in the R package *adaBoost* [69], we quantified the importance of polymorphic sites in our alignment for predicting resistance to a fluoroquinolone antibiotic (levofloxacin). Adaptive boosting is a machine learning 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 [70]. The importance of a site is quantified by 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 was obtained from the same study which published the 390 draft genomes used to construct our alignment [17]. 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 the six, previously identified, focal *P. aeruginosa* genes which evolve in response to fluoroquinolone selection [14, 15, 17].

Phylogenetic test for horizontal gene transfer

Like our analysis for correlated evolution, we used a phylogenetic method to test for horizontal gene transfer (HGT). By constructing the phylogeny for each gene containing our most strongly correlated synonymous substitutions, we were able to test for HGT using the Approximately Unbiased (AU) test [71] implemented in CONSEL [72] using *baseml* (REV + Γ_5, no clock) output from PAML [73]. For each of the genes, we used the AU test to calculate the probability that evolution along any of the gene trees, and the species tree used in our original analysis for correlated evolution, could have resulted in the gene's alignment. A significant difference between the compared trees this is evidence for horizontal gene transfer [74].

Testing for biological effects of synonymous substitutions

All the most significantly correlated pairs of synonymous substitutions were pairs of substitutions in the same gene. To test if these paired substitutions could be correlated due to translational effects, we estimated the free energy (ΔG) of folded mRNA transcripts for both WT and mutant transcripts, which we normalized as relative values through division of the WT estimate (ΔG_{rel}; Fig. 2A). The ΔG values were calculated with the mFold server [75], based on the QuickFold service with default settings (`http://unafold.rna.albany.edu/?q=DINAMelt/Quickfold`). Similar to previous studies [76], 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.

For pairs of codons containing synonymous substitutions, we estimated effects of translation elongation rate via the index of translation elongation [**?**, I_{TE}:]]Xia:2014aa, calculated with default settings in DAMBE ver. [77]. These calculations used the codon frequency of genes that are highly expressed in *P. aeruginosa* [78], 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 `https://github.com/JDench/Pseudomonas_DAMBE_ITE`.

References

References

1. Falconer DS. Introduction to quantitative genetics. Oliver And Boyd; Edinburgh; London; 1960.

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

3. Lynch M, Walsh B, et al. Genetics and analysis of quantitative traits. vol. 1. Sinauer Sunderland, MA; 1998.

4. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. Journal of Molecular Biology. 2000;299(2):283 – 293.

5. Callahan B, Neher RA, Bachtrog D, Andolfatto P, Shraiman BI. Correlated Evolution of Nearby Residues in Drosophilid Proteins. Plos Genetics. 2011;7(2):e1001315–e1001315.

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

7. Sexton JP, McIntyre PJ, Angert AL, Rice KJ. Evolution and Ecology of Species Range Limits. Annual Review of Ecology Evolution and Systematics. 2009;40:415–436.

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

9. Pagel M. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. Proceedings: Biological Sciences. 1994;255(1342):37–45.

10. Shimizu A, Dohzono I, Nakaji M, Roff DA, Miller III DG, Osato S, et al. Fine-tuned Bee-Flower Coevolutionary State Hidden within Multiple Pollination Interactions. Scientific Reports. 2014;4(1).

11. Nshogozabahizi JC, Dench J, Aris-Brosou S. Widespread Historical Contingency in Influenza Viruses. Genetics. 2017;205(1):409–420.

12. Ibeh N, Nshogozabahizi JC, Aris-Brosou S. Both Epistasis and Diversifying Selection Drive the Structural Evolution of the Ebola Virus Glycoprotein Mucin-Like Domain. J Virol. 2016;90(11):5475–84.

13. Aris-Brosou S, Ibeh N, Noël J. Viral outbreaks involve destabilized evolutionary networks: evidence from Ebola, Influenza and Zika. Sci Rep. 2017;7(1):11881.

14. Akasaka T, Tanaka M, Yamaguchi A, Sato K. 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. 2001;45(8):2263–8.

15. Wong A, Rodrigue N, Kassen R. Genomics of Adaptation during Experimental Evolution of the Opportunistic Pathogen *Pseudomonas aeruginosa*. PLoS Genet. 2012;8(9):e1002928.

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

17. Kos VN, Déraspe M, McLaughlin RE, Whiteaker JD, Roy PH, Alm RA, et al. The Resistome of *Pseudomonas aeruginosa* in Relationship to Phenotypic Susceptibility. Antimicrobial Agents and Chemotherapy. 2015;59(1):427–436.

18. Roy PH, Tetu SG, Larouche A, Elbourne L, Tremblay S, Ren Q, et al. Complete Genome Sequence of the Multiresistant Taxonomic Outlier *Pseudomonas aeruginosa* PA7. PLoS One. 2010;5(1):e8842.

19. Choy WK, Zhou L, Syn CKC, Zhang LH, Swarup S. MorA Defines a New Class of Regulators Affecting Flagellar Development and Biofilm Formation in Diverse *Pseudomonas* Species. Journal of Bacteriology. 2004;186(21):7221–7228.

20. Starkey M, Hickman JH, Ma L, Zhang N, De Long S, Hinz A, et al. *Pseudomonas aeruginosa* Rugose Small-Colony Variants Have Adaptations That Likely Promote Persistence in the Cystic Fibrosis Lung. Journal of Bacteriology. 2009;191(11):3492–3503.

21. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12.

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

23. Agashe D, Sane M, Phalnikar K, Diwan GD, Habibullah A, Martinez-Gomez NC, et al. Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. Molecular Biology and Evolution. 2016;33(6):1542–1553.

24. Melnyk AH, McCloskey N, Hinz AJ, Dettman J, Kassen R. Evolution of Cost-Free Resistance under Fluctuating Drug Selection in *Pseudomonas aeruginosa*. mSphere. 2017;2(4).

25. Moon DC, Seol SY, Gurung M, Jin JS, Choi CH, Kim J, et al. 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. 2010;35(1):76–79.

26. Bateson W. Review from Cambridge: Mendel's Principles of Heredity. The Edinburgh Review. 1911;213(435):77.

27. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I. Positive epistasis drives the acquisition of multidrug resistance. PLoS Genetics. 2009;5(7):e1000578–e1000578.

28. Gillespie SH, Voelker LL, Dickens A. Evolutionary Barriers to Quinolone Resistance in *Streptococcus pneumoniae*. Microbial Drug Resistance. 2002;8(2):79–84.

29. Wibowo A. Performance standards for antimicrobial susceptibility testing: 23rd informational supplement M100-S23. In: Performance standards for antimicrobial susceptibility testing: 23rd informational supplement M100-S23. vol. 33 of Developments in Environmental Modelling. Clinical and Laboratory Standards Institute; 2013. p. 62–65.

30. Long GS, Hussen M, Dench J, Aris-Brosou S. Identifying the genetic determinants of particular phenotypes in microbial genomes with very small training sets. bioRxiv. 2017;doi:https://doi.org/10.1101/181222.

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

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

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

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

35. Lawrie DS, Messer PW, Hershberg R, Petrov DA. Strong Purifying Selection at Synonymous Sites in *D. melanogaster*. PLoS Genetics. 2013;9(5):1–18.

36. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. Science. 2009;324.

37. Xia X. A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics. 2014;199(2):573–579.

38. Taylor BL, Zhulin IB. PAS Domains: Internal Sensors of Oxygen, Redox Potential, and Light. Microbiology and Molecular Biology Reviews. 1999;63(2):479–506.

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

40. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(7):3801–3806.

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

42. Buskirk SW, Peace RE, Lang GI. Hitchhiking and epistasis give rise to cohort dynamics in adapting populations. Proceedings of the National Academy of Sciences. 2017;114(31):8330–8335.

43. MacLean RC, Perron GG, Gardner A. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in *Pseudomonas aeruginosa*. Genetics. 2010;186(4):1345–54.

44. Hall AR, MacLean RC. Epistasis buffers the fitness effects of rifampicin-resistance mutations in *Pseudomonas aeruginosa*. Evolution. 2011;65(8):2370–9.

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

46. Dutheil JY, Jossinet F, Westhof E. Base Pairing Constraints Drive Structural Epistasis in Ribosomal RNA Sequences. Molecular Biology and Evolution. 2010;27(8):1868–1876.

47. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, et al. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Human Molecular Genetics. 2003;12(3):205–216.

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

49. Cuevas JM, Domingo-Calap P, Sanjuán R. The Fitness Effects of Synonymous Mutations in DNA and RNA Viruses. Molecular Biology and Evolution. 2012;29(1):17–20.

50. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14(7):1394–403.

51. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27(3):334–342.

52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403 – 410.

53. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792–1797.

54. Nylander J. catfasta2phyml; 2015. `https://github.com/nylander/catfasta2phyml`.

55. Tatusov RL, Koonin EV, Lipman DJ. A Genomic Perspective on Protein Families. Science. 1997;278(5338):631–637.

56. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Research. 2015;43(D1):D261–D269.

57. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3).

58. Aris-Brosou S, Rodrigue N. The Essentials of Computational Molecular Evolution. In: Anisimova M, editor. Evolutionary Genomics. vol. 855 of Methods in Molecular Biology. Humana Press; 2012. p. 111–152.

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

60. Bertani G. Studies on Lysogenesis I. : The Mode of Phage Liberation by Lysogenic *Escherichia coli*. Journal of Bacteriology. 1951;62(3):293–300.

61. Schweizer HP. Bacterial genetics: past achievements, present state of the field, and future challenges. BioTechniques. 2008;44:633–641.

62. Sorrells TR, Booth LN, Tuch BB, Johnson AD. Intersecting transcription networks constrain gene regulatory evolution. Nature. 2015;.

63. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 2014;344(6191):1519–1522.

64. Blount ZD, Borland CZ, Lenski RE. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proceedings of the National Academy of Sciences. 2008;105(23):7899–7906.

65. Dettman JR, Rodrigue N, Aaron SD, Kassen R. Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. Proc of the Nat Acad of Sci of the USA. 2013;110(52):21065–21070.

66. Engler C, Kandzia R, Marillonnet S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. PLoS One. 2008;3(11):1–7.

67. Figurski DH, Helinski DR. 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. 1979;76(4):1648–1652.

68. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-Term Experimental Evolution in *Escherichia coli*. I. Adaptation and Divergence During 2,000 Generations. The American Naturalist. 1991;138(6):1315–1341.

69. Alfaro E, Gámez M, García N. adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software. 2013;54(2):1–35.

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

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

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

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

74. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring Horizontal Gene Transfer. PLoS Computational Biology. 2015;11(5):1–16.

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

76. Takanami M, Zubay G. 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. 1964;51(5):834–839.

77. Xia X. DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. Journal of Heredity. 2017;108(4):431–437.

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

Figure Captions

Fig 1. Correlated evolution among genomic sites can be underlain by epistasis. A: substitutions with the strongest evidence for correlated evolution ($P < 10^{-11}$). Inset is a histogram showing the distribution of significance for all correlated pairs with strong ($P < 10^{-7}$) evidence for correlated evolution, where the purple bar highlights the strength of evidence for the pairs presented below. Circles represent a substitution in a gene (top text), at a particular nucleotide position (bottom text) on the coding strand. Colors 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. For Panels B-E, the mean and 95% confidence interval of measurements for each genotype is represented by the dark black lines and colored rectangles respectively. The color used reflects the genetic background of genotypes while the hue identifies the biological replicate. B,C: 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. D,E: 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 2).

Fig 2. 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. In Panel A, estimates of the relative change of mRNA folding free energy (ΔG) are relative to the WT and error bars represent the 95% confidence interval. In Panel B, measures of the mean index of translation elongation ($\mathrm{I_{TE}}$, B) [37] with error bars presenting standard deviation among measures estimated from double mutant strains present in our data set.

Supporting information

S1 Fig. SIFig1_usedTrees_mod: 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 maximum likelihood. 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.

S2 Fig. SIFig2_all_pairedInfo_trees_mod: Phylogeny of the gene(s) which

contain the most significantly correlated pairs of substitutions ($P < 10^{11}$). Each tree was constructed with FastTree [57] (GTR $+\Gamma_5$). The title of each panel indicates the pair of substitutions being considered. For trees on the left hand side, the color of the tips represent the paired nucleotide state of strains. On the right side, the color of tips represent the I_{TE} value of the whole gene(s) sequence.

S3 Fig. SIFig3_adaBoost_importance_levoRes_mod: 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* [69]) on all polymorphic sites in our alignment. Levofloxacin resistance phenotype information was obtained from previously published data [17]. 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.

S4 Fig. SIFig4_distance_vs_pValue_sigPairs_mod: 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.

S1 File. SIFile1_sigResults_AEGIS_inVivo-master.zip: Analysis of correlated evolution using AEGIS. All R scripts and complete results are included in a single .zip file, but can also be found uploaded to GitHub `https://github.com/JDench/sigResults_AEGIS_inVivo`. Site names refer to the gene name, or the genomic island identifier, and the nucleotide position as determined

by mapping to the circular chromosome of PA14.

S1 Appendix.

1 Supplementary analyses

inVitro_PLoS_Genetics_Appendix1: Additional analyses describing the types of correlated pairs detected with AEGIS.

S1 Table. SITable1_siAppend_resCorr.xlsx: 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).

S2 Table. SITable2_siAppend_deltaG.xlsx: 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 [27]. 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.

S3 Table. SITable3_siAppend_ITE.xlsx: Evidence of $\mathbf{I_{TE}}$ epistasis among the strongly correlated pairs of synonymous intragenic substitutions. Epistasis is measured with a multiplicative model and error is calculated using error propagation [27]. 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 the sign of ϵ which has an absolute value greater than estimates of error.

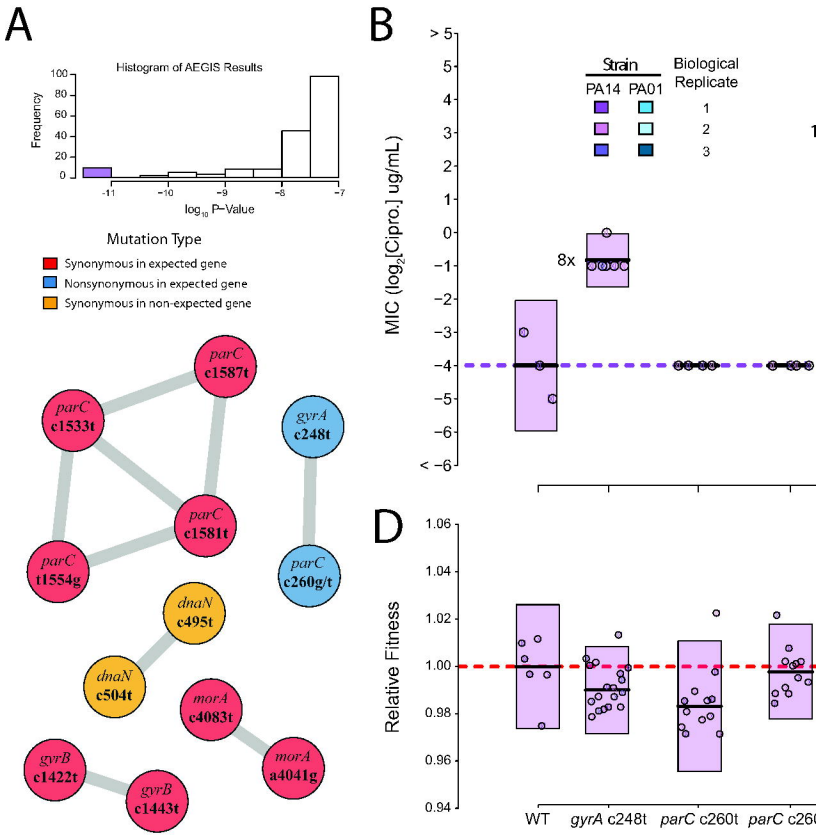
S4 Table. SITable4_siAppend_indTest.xlsx: Test for independence between the relative position of correlated substitutions and the type of pair they form. We used a χ^2 test and present the observed (expected) counts for the number of significantly correlated pairs ($P \leq 10^{-4}$) of substitutions 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 $P \sim 0$.

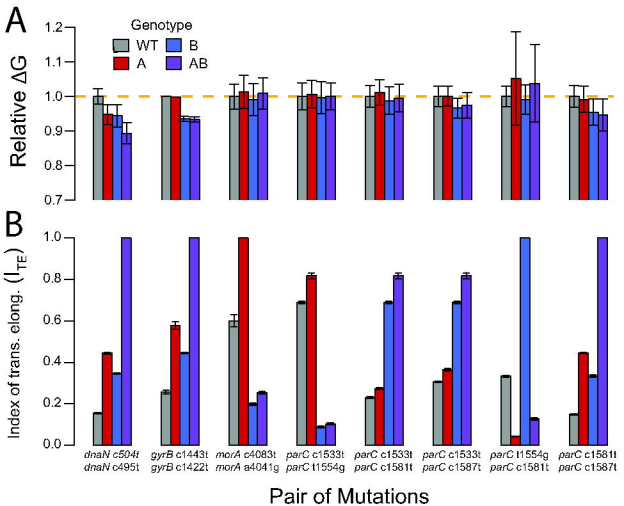
S5 Table. SITable5_siAppend_AUtests.xlsx: Approximately unbiased (AU) tests of significant differences among phylogeny. We built phylogeny using the nucleotide multiple sequence alignment of the four genes which contain the most significantly correlated synonymous substitutions ($dnaN$, $gyrB$, $morA$, $parC$) and four tRNA genes unlikely, according to the complexity hypothesis [40, 41] to undergo HGT ($arginyltRNAsynthetase$, $glycyltRNAsynthetasesubunitbeta$, $isoleucyltRNAsynthetase$, $methionyltRNAformyltransferase$). We tested for significant differences among all the gene trees, a gene tree create from a concatenation of the $gyrA$ and $parc$ genes, and the species tree used in the analysis of correlated evolution. We bolded the best tree and any similar trees.

S6 Table. SITable6_siAppend_statsMIC.xlsx: 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.

Acknowledgments

We would like to thank Dr. Felipe Dargent and Matti Ruuskanen for their constructive feedback on the manuscript. We are grateful for the computational resources of the Centre for Advanced Computing at Queens University.


A
Histogram of AEGIS Results
Frequency
100
80
60
40
20
0
−11
−10
−9
−8
−7
log10 P−Value
Mutation Type
Synonymous in expected gene
Nonsynonymous in expected gene
Synonymous in non-expected gene
parC c1587t
parC c1533t
gyrA c248t
parC c1581t
parC t1554g
parC c260g/t
dnaN c495t
dnaN c504t
morA c4083t
morA a4041g
gyrB c1422t
gyrB c1443t
B
MIC (log2[Cipro.] ug/mL)
> 5
5
4
3
2
1
0
−1
−2
−3
−4
−5
−6
< −6
Strain
PA14 PA01
Biological Replicate
1
2
3
8x
D
Relative Fitness
1.06
1.04
1.02
1.00
0.98
0.96
0.94
WT
gyrA c248t
parC c260t
parC c260
Genotyp


A
Genotype
WT
A
B
AB
Relative ΔG
1.2
1.1
1.0
0.9
0.8
0.7
B
Index of trans. elong. (I_{TE})
1.0
0.8
0.6
0.4
0.2
0.0
dnaN c504t
dnaN c495t
gyrB c1443t
gyrB c1422t
morA c4083t
morA a4041g
parC c1533t
parC t1554g
parC c1533t
parC c1581t
parC c1533t
parC c1587t
parC t1554g
parC c1581t
parC c1581t
parC c1587t
Pair of Mutations