Selection maintains protein interactome resilience in the long-term evolution experiment with *Escherichia coli*

Rohan Maddamsetti[1,*]

[1]Department of Biomedical Engineering, Duke University, Durham NC
*E-mail: rohan.maddamsetti@duke.edu

Abstract

Most cellular functions are carried out by a dynamic network of interacting proteins. An open question is whether the network properties of protein interactomes represent phenotypes under natural selection. One proposal is that protein interactomes have evolved to be resilient, such that they tend to maintain connectivity when proteins are removed from the network. This hypothesis predicts that interactome resilience should be maintained during long-term experimental evolution. I tested this prediction by modeling the evolution of protein-protein interaction (PPI) networks in Lenski's long-term evolution experiment with *Escherichia coli* (LTEE). In this test, I removed proteins affected by nonsense, insertion, deletion, and transposon mutations in evolved LTEE strains, and measured the resilience of the resulting networks. I compared the rate of change of network resilience in each LTEE population to the rate of change of network resilience for corresponding randomized networks. The evolved PPI networks are significantly more resilient than networks in which random proteins have been deleted. Moreover, the evolved networks are generally more resilient than networks in which the random deletion of proteins was restricted to those disrupted in LTEE. These results suggest that evolution in the LTEE has favored PPI networks that are, on average, more resilient than expected from the genetic variation across the evolved populations. My findings therefore support the hypothesis that selection maintains protein interactome resilience over evolutionary time.

Significance Statement

Understanding how protein-protein interaction (PPI) networks evolve is a central goal of evolutionary systems biology. One property that has been hypothesized to be important for PPI network evolution is resilience, which means that networks tend to maintain connectivity even after many nodes (proteins in this case) have been removed. This hypothesis predicts that PPI network resilience should be maintained during long-term experimental evolution. Consistent

with this prediction, I found that the PPI networks that evolved over 50,000 generations of Lenski's long-term evolution experiment with *E. coli* are more resilient than expected by chance.

Introduction

When redundant nodes and connections in a network carry a cost, removing those redundancies may increase efficiency at the expense of reducing resilience to unexpected disruptions. After a critical point, called the percolation threshold, further pruning can cause a catastrophic breakdown of connectivity and function (Callaway, et al. 2000). In the context of protein-protein interaction (PPI) networks, efficiency may refer to the cost of synthesizing the proteins that make up the nodes of the network.

Zitnik and colleagues (2019) formally defined network resilience to measure how quickly a network breaks down as nodes are removed. They then studied the evolution of PPI networks (also called protein interactomes) across the tree of life, concluding with the hypothesis that interactome resilience is favored during evolution. This hypothesis predicts that interactome resilience should be maintained during long-term evolution experiments with large populations of microbes. Here, I use the methods developed by Zitnik et al. (2019) to test this prediction, by examining how protein interactome resilience has evolved in Lenski's long-term evolution experiment with *Escherichia coli* (LTEE).

In the LTEE, 12 initially identical populations of *E. coli* have evolved for more than 50,000 generations (Lenski, et al. 1991; Lenski 2017). The LTEE populations are named by the presence of a neutral phenotypic marker: populations Ara+1 to Ara+6 grow on arabinose, while populations Ara−1 to Ara−6 cannot (Lenski, et al. 1991). Many populations have lost unnecessary metabolic traits (Leiby and Marx 2014; Grant, et al. 2020), and many genes have been disrupted by mutations, in part caused by the evolution of elevated mutation rates in several populations (Tenaillon, et al. 2016; Couce, et al. 2017; Good, et al. 2017; Maddamsetti and Grant 2020a).

Despite evidence for genomic and phenotypic streamlining, it is unknown how PPI network resilience has evolved during the LTEE. To examine this question, I compared the rate of change of protein interactome resilience in each LTEE population to the expected rate of change in corresponding sets of randomized networks (Methods). I compare rates of change of interactome resilience between real and simulated networks, because this approach is simple and

accounts for phylogenetic correlations among genomes from the same population; it treats the independent LTEE populations as the appropriate unit of statistical replication. Importantly, the randomized networks have no such phylogenetic correlations, for the sake of sampling large numbers of independent replicates given computational time constraints. I applied this approach to two curated datasets of protein-protein interactions in *E. coli*, which I refer to as Zitnik interactome (Zitnik, et al. 2019) and the Cong interactome (Cong, et al. 2019). Overall, protein interactome resilience is higher in the LTEE than in the randomized networks, indicating that this system-level property is being maintained over long-term experimental evolution.

Results

LTEE PPI networks are more resilient than PPI networks with random proteins deleted

I calculated the resilience of randomized counterparts of the evolved LTEE PPI networks, in which proteins to deleted from the network were sampled at random (Figure 1). In this randomization scheme, a protein may be removed from the PPI network, regardless of its essentiality in *E. coli*. For this reason, I expected that the evolved LTEE PPI networks would be more resilient than these randomized PPI networks, and this was indeed the case (Zitnik interactome: Wilcoxon signed-rank exact test $p = 0.00024$; Cong interactome: Wilcoxon signed-rank exact test $p = 0.00024$).

LTEE PPI networks are more resilient than PPI networks with gene disruptions sampled across LTEE populations

In part, the previous finding could be caused by removing essential proteins, which are known to be highly connected and consequential for network resilience (Zitnik et al. 2019), from the randomized networks. I therefore conducted a second test, in which I restricted the proteins sampled for deletion to those that were disrupted in at least one LTEE population. Here, the probability of sampling proteins for deletion was weighted by the frequency of observed disruptions across LTEE populations (Figure 1). This resampling procedure takes parallel genetic evolution into account (Woods, et al. 2006; Ostrowski, et al. 2008; Tenaillon, et al. 2016). I

found that that evolved LTEE PPI networks, on the whole, are more resilient than randomized PPI networks generated from gene disruptions tabulated across all 12 LTEE populations (Zitnik interactome: Wilcoxon signed-rank exact test $p = 0.0061$; Cong interactome: Wilcoxon signed-rank exact test $p = 0.0012$).

Genes in large deletions in the LTEE neither show physical modularity nor fewer interactions

The resampling procedures operates on the level of individual gene disruptions and losses, such that large deletions are replaced by a sample of individual gene disruptions. Therefore, the resampling procedures could bias the results by breaking up the block structure of multi-gene deletions. This would matter if disrupting a block of x genes has less of an effect on PPI network resilience than disrupting x genes across the genome.

I examined two PPI properties through which the absence of large deletions in the randomized genomes could have an effect. First, systematic bias could be introduced if genes within large deletions tend to have fewer interactions than genes affected by small indels or nonsense mutations. Second, systematic bias could be introduced if interactions within large deletions show physical modularity, such that genes within large deletions preferentially interact with each other, but not with genes elsewhere on the chromosome.

I find no difference in PPI degree between genes knocked out by multi-gene deletions, and those disrupted by single-gene mutations in the 50,000 generation LTEE clones (Zitnik interactome: Wilcoxon rank sum test $p = 0.46$; Cong interactome: Wilcoxon rank sum test $p = 0.26$). Furthermore, I find that interactions removed by multi-gene deletions in 50,000 generation LTEE clones are *further apart* in the genome, on average, than interactions removed by single-gene disruption mutations. For the case of the Zitnik interactome, interactions removed by multi-gene deletions have a mean genomic distance of 1,071,063 base-pairs, compared to a mean genomic distance of 1,054,403 base-pairs for interactions removed by single-gene disruptions (Wilcoxon rank sum test: $p = 0.0364$). For the case of the Cong interactome, interactions removed by multi-gene deletions have a mean genomic distance of 714,316 base-pairs, compared to a mean genomic distance of 583,624 base-pairs for interactions removed by single-gene disruptions (Wilcoxon rank sum test: $p < 10^{-4}$). Therefore, the differences between realized and randomized PPI network resilience in the LTEE do not seem to be an artifact of the resampling

procedure, at least with regard to PPI degree and the aspects of physical modularity that I examined.

Discussion

I find that evolved networks in the LTEE lose protein interactome resilience more slowly than expected, based on comparisons with networks with random gene losses and networks with gene losses weighted by their occurrence across LTEE populations (Figure 1). These results are consistent with the hypothesis proposed by Zitnik and colleagues (2019) that protein interactome resilience is favored during evolution.

The fact that evolved PPI networks in the LTEE are more resilient than networks with random deletions implies purifying selection on highly interacting genes in the LTEE. Indeed, evidence for purifying selection on essential genes in the LTEE has been found (Maddamsetti and Grant 2020b) as has evidence of purifying selection on highly interacting genes (Maddamsetti 2020). That said, it is unclear whether purifying selection on essential genes maintains network resilience as a side-effect, or whether the maintenance of network resilience causes selection on essential genes, or both. I thus conducted a second analysis that controlled for purifying selection on essential genes by resampling gene disruptions based on their frequency of occurrence across the LTEE populations. Overall, the evolved LTEE PPI networks still showed more resilience than the randomized networks, supporting the hypothesis that resilience is maintained during evolution in the LTEE.

This work has important limitations. First, the resampling procedure used to generate randomized genomes does not maintain the block structure of large deletion mutations. For this reason, I analyzed the protein interaction degree and genomic interaction distance distribution of genes affected by large deletions in the LTEE. This analysis did not uncover any systematic biases that would affect the broad import of my findings. Second, the resampling procedure does not preserve the phylogenetic structure within each population (i.e. randomized networks at later time points are not subnetworks of the randomized networks at earlier timepoints), for the sake of computational tractability. Duplication and amplification mutations are also ignored, owing to their rarity in these data (Tenaillon, et al. 2016), and the evolution of new interactions is ignored due to a lack of data.

Several questions remain. It is unclear if protein interactome resilience is under direct selection, or is a byproduct of selection on other phenotypes. Fitness increases in each LTEE population (Wiser, et al. 2013) while resilience decreases (Figure 1). Therefore, interactome resilience negatively correlates with fitness gains in the LTEE (Figure 2). In this vein, it would be interesting to ask whether variation in interactome resilience correlates with evolvability in the sense of the distribution of fitness effects (DFE) for beneficial mutations (Mustonen and Lässig 2010; Woods, et al. 2011; Łuksza and Lässig 2014; Levy, et al. 2015; Ba, et al. 2019), or with mutational robustness in the sense of the DFE for deleterious mutations (Johnson, et al. 2019).

There is also an intriguing connection between network resilience and the deterministic mutation hypothesis for the evolution of sex (Kondrashov 1988; Azevedo, et al. 2006). Loss-of-function mutations may have little effect on network integrity when they occur in a genome with high network resilience. By contrast, they may have catastrophic effects on network integrity when they occur in a genome with low network resilience. The deterministic mutation hypothesis states that synergistic epistasis between deleterious mutations—such that those that together cause network fragmentation— confers a selective advantage to sex. Near a critical threshold of network resilience, additional loss-of-function mutations are more likely to fragment biological networks, which could contribute to the synergistic epistasis required by the deterministic mutation hypothesis for the evolution of sex. Gene disruptions continue to accumulate over time in each LTEE population, suggesting that might be worthwhile to test for such synergistic interactions at a later point (Elena and Lenski 1997), especially in the context of PPI network resilience.

Methods

Datasets

I downloaded a table of nonsense SNPs, small indels, mobile element insertions, and large deletions affecting protein-coding regions, which I operationally define as loss-of-function mutations, from: barricklab.org/shiny/LTEE-Ecoli. In the analyses restricted to 50,000 generation LTEE clones, I used single clones from each of the 12 populations (the LTEE 50,000 generation ‘A’ clones) to maximize independence.

I used two different PPI datasets for *E. coli*. First, I used the curated PPI set published by Zitnik et al. (2019). This set of interactions corresponds to the interactome for species 511145 in their cross-species interactome dataset. For robustness, I also used a second *E. coli* protein interactome that has been published by Cong and colleagues (2019). This set of interactions includes those from protein structures in the Protein Data Bank (their Table S4), the Ecocyc database (their Table S5), yeast two-hybrid experiments (their Table S6), affinity purification and mass spectrometry (their Table S7), and those supported by known and high-confidence novel prediction coevolutionary information (their Tables S8 and S10, respectively). The Zitnik and Cong interactomes were filtered based on the genes they share with the *E. coli* B str. REL606 genome (the ancestral LTEE clone).

Network resilience analysis

The snap.py python module interface to the Stanford Network Analysis Platform (Leskovec and Sosič 2016) was used to generate a graph representing the ancestral REL606 interactome. Then, a protein interactome network was generated for each LTEE genome, by pruning the REL606 interactome of nodes (proteins) and edges (interactions) affected by loss of function mutations in the given genome. Network resilience was calculated using the method described in Zitnik et al. (2019). In brief, resilience measures how much a network resists fragmentation into many small, isolated components as increasing fractions of nodes are randomly removed. Specifically, the entropy of the distribution of component sizes in the network is calculated, normalized by the logarithm of the total number of nodes in the original network. A fraction f of the nodes of the graph are randomly removed, ranging from 0.01 to 1.0, and the entropy of the network component distribution is calculated for each value of f. The entropy of the network component distribution increases monotonically as f increases. The area under the curve (AUC) of the mapping from f to the entropy of the network component distribution is calculated numerically, and network resilience is defined as 1 – AUC. This resilience calculation was conducted 100 times per genome. The average network resilience out of these 100 samples was used as the estimate for a genome's network resilience for a given simulation run. 100 simulations of this entire algorithm were conducted, such that network resilience was calculated 10,000 times for each genome. The trend over time in each population was then plotted, and a linear regression was fit to estimate the change in network resilience over

time in each LTEE population. The *y*-intercept was fixed as the mean estimate for the resilience of the ancestral LTEE clone REL606 over all 100 simulation runs, so that the linear regression only fits one parameter (a slope representing network resilience over time). I then constructed two null distributions for comparison to the actual data as follows:

Randomization over all genes in REL606. For each of the 264 LTEE genomes, I constructed a corresponding randomized genome in which the number of disrupted genes was held fixed, and a selection of genes in REL606 was drawn at random. Then, an interactome network was constructed with those genes disrupted. The resilience of these randomized networks was calculated as above.

Randomization over genes disrupted across LTEE populations. For each genome, I constructed a corresponding randomized genome in which the number of disrupted genes was held fixed, and a selection of genes was drawn from genes that were disrupted across all LTEE populations. Then, an interactome network was constructed with those nodes (genes) disrupted. The resilience of this randomized network was calculated as above. This procedure weighted each gene based on the multiplicity of observed disruptions of that gene across LTEE populations. Specifically, the set of disrupted genes for each LTEE population was calculated, then the multiplicity of those genes across populations was calculated, such that a gene that was disrupted in 5 populations would be represented 5 times for sampling, while a gene that was never disrupted in any LTEE genomes would be omitted.

The differences between the slopes for linear regressions for the actual data and the randomized data for each population were tabulated, and a Wilcoxon signed rank test was used to test whether the paired difference in slope, over all LTEE populations, was significantly greater than zero.

Data Availability Statement

All data used is publicly available, as described in the Methods. The data and analysis codes underlying this article are available on the Dryad Digital Repository (DOI pending publication). Analysis codes are also available at: https://github.com/rohanmaddamsetti/LTEE-network-analysis.

Acknowledgements

I thank Richard Lenski, Jeffrey Barrick, and Lingchong You for guidance and advice, Tom Milledge and Duke Research Computing for research support, Nkrumah Grant for valuable discussions, and Zachary Blount for detailed comments and feedback on an earlier version of this manuscript. The LTEE is supported by a grant from the National Science Foundation (currently DEB-1951307) to Richard Lenski and Jeffrey Barrick.

References

Azevedo RB, Lohaus R, Srinivasan S, Dang KK, Burch CL. 2006. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. Nature 440:87-90.

Ba ANN, Cvijović I, Echenique JIR, Lawrence KR, Rego-Costa A, Liu X, Levy SF, Desai MM. 2019. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature 575:494-499.

Callaway DS, Newman ME, Strogatz SH, Watts DJ. 2000. Network robustness and fragility: Percolation on random graphs. Physical review letters 85:5468.

Cong Q, Anishchenko I, Ovchinnikov S, Baker D. 2019. Protein interaction networks revealed by proteome coevolution. Science 365:185-189.

Couce A, Caudwell LV, Feinauer C, Hindré T, Feugeas J-P, Weigt M, Lenski RE, Schneider D, Tenaillon O. 2017. Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. Proceedings of the National Academy of Sciences 114:E9026-E9035.

Elena SF, Lenski RE. 1997. Test of synergistic interactions among deleterious mutations in bacteria. Nature 390:395-398.

Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of molecular evolution over 60,000 generations. Nature 551:45-50.

Grant NA, Maddamsetti R, Lenski RE. 2020. Maintenance of metabolic plasticity despite relaxed selection in a long-term evolution experiment with Escherichia coli. bioRxiv.

Johnson MS, Martsul A, Kryazhimskiy S, Desai MM. 2019. Higher-fitness yeast genotypes are less robust to deleterious mutations. Science 366:490-493.

Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435-440.

Leiby N, Marx CJ. 2014. Metabolic erosion primarily through mutation accumulation, and not tradeoffs, drives limited evolution of substrate specificity in Escherichia coli. PLoS Biol 12:e1001789.

Lenski RE. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. The ISME journal 11:2181-2194.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. The American Naturalist 138:1315-1341.

Leskovec J, Sosič R. 2016. Snap: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST) 8:1-20.

Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 519:181-186.

Łuksza M, Lässig M. 2014. A predictive fitness model for influenza. Nature 507:57-61.

Maddamsetti R. 2020. Universal constraints on protein evolution in the long-term evolution experiment with Escherichia coli. bioRxiv.

Maddamsetti R, Grant NA. 2020a. Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with Escherichia coli. Genome Biology and Evolution 12:1591-1603.

Maddamsetti R, Grant NA. 2020b. A simple test to infer mode of selection in metagenomics time series of evolving asexual populations. bioRxiv.

Mustonen V, Lässig M. 2010. Fitness flux and ubiquity of adaptive evolution. Proceedings of the National Academy of Sciences 107:4248-4253.

Ostrowski EA, Woods RJ, Lenski RE. 2008. The genetic basis of parallel and divergent phenotypic responses in evolving populations of Escherichia coli. Proceedings of the Royal Society B: Biological Sciences 275:277-284.

Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536:165-170.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. Science 342:1364-1367.

Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. 2006. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proceedings of the National Academy of Sciences 103:9107-9112.

Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. 2011. Second-order selection for evolvability in a large Escherichia coli population. Science 331:1433-1436.

Zitnik M, Feldman MW, Leskovec J. 2019. Evolution of resilience in protein interactomes across the tree of life. Proceedings of the National Academy of Sciences 116:4426-4433.

Figure 1. Protein-protein interaction (PPI) networks in the LTEE (red) lose network resilience more slowly than randomized networks generated using either all genes in the REL606 genome (yellow) or only those disrupted in the LTEE (blue). A) Analysis based on the *E. coli* PPI network published in Zitnik et al. (2019). B) Analysis based on the *E. coli* PPI network published in Cong et al. (2019).

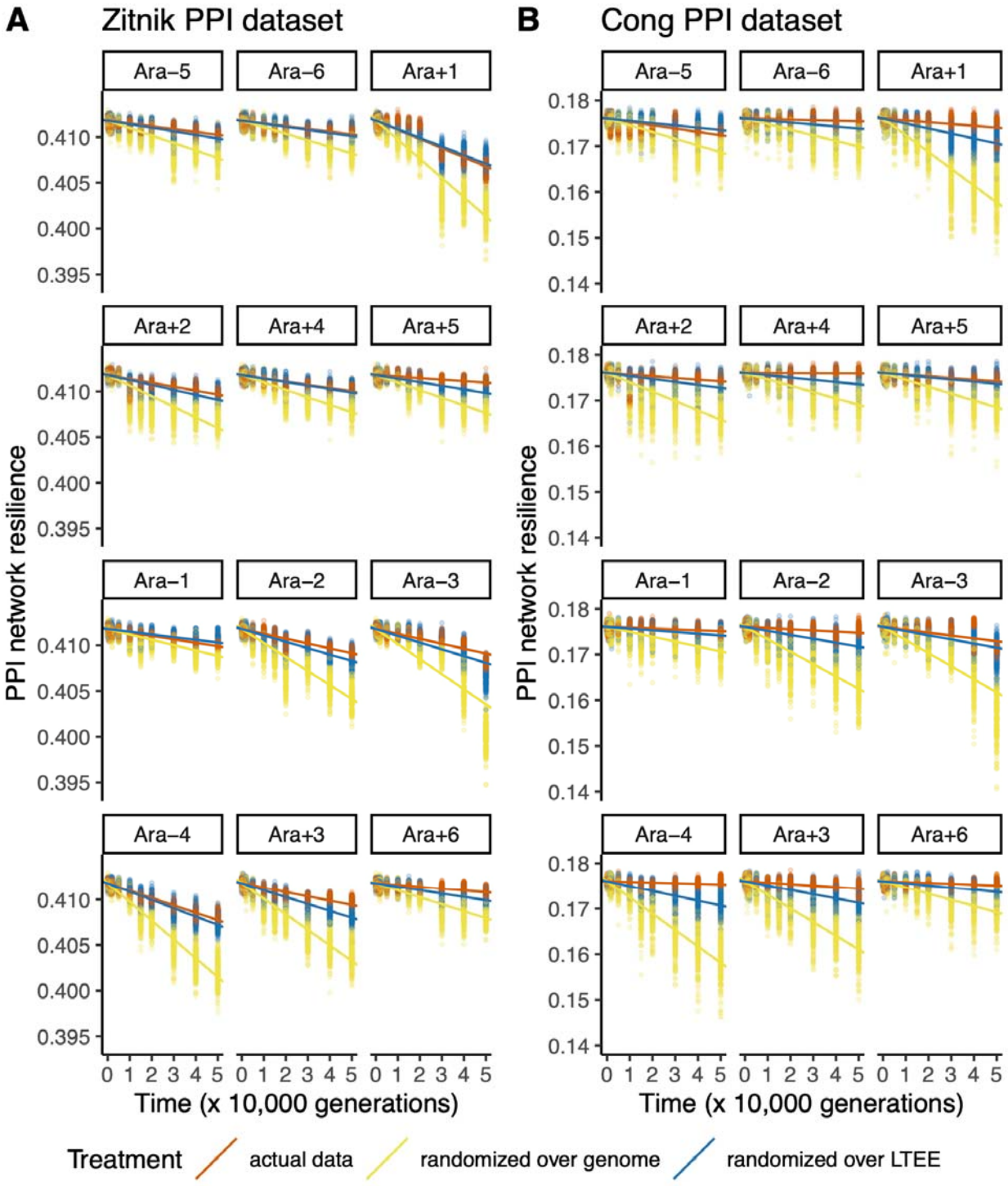
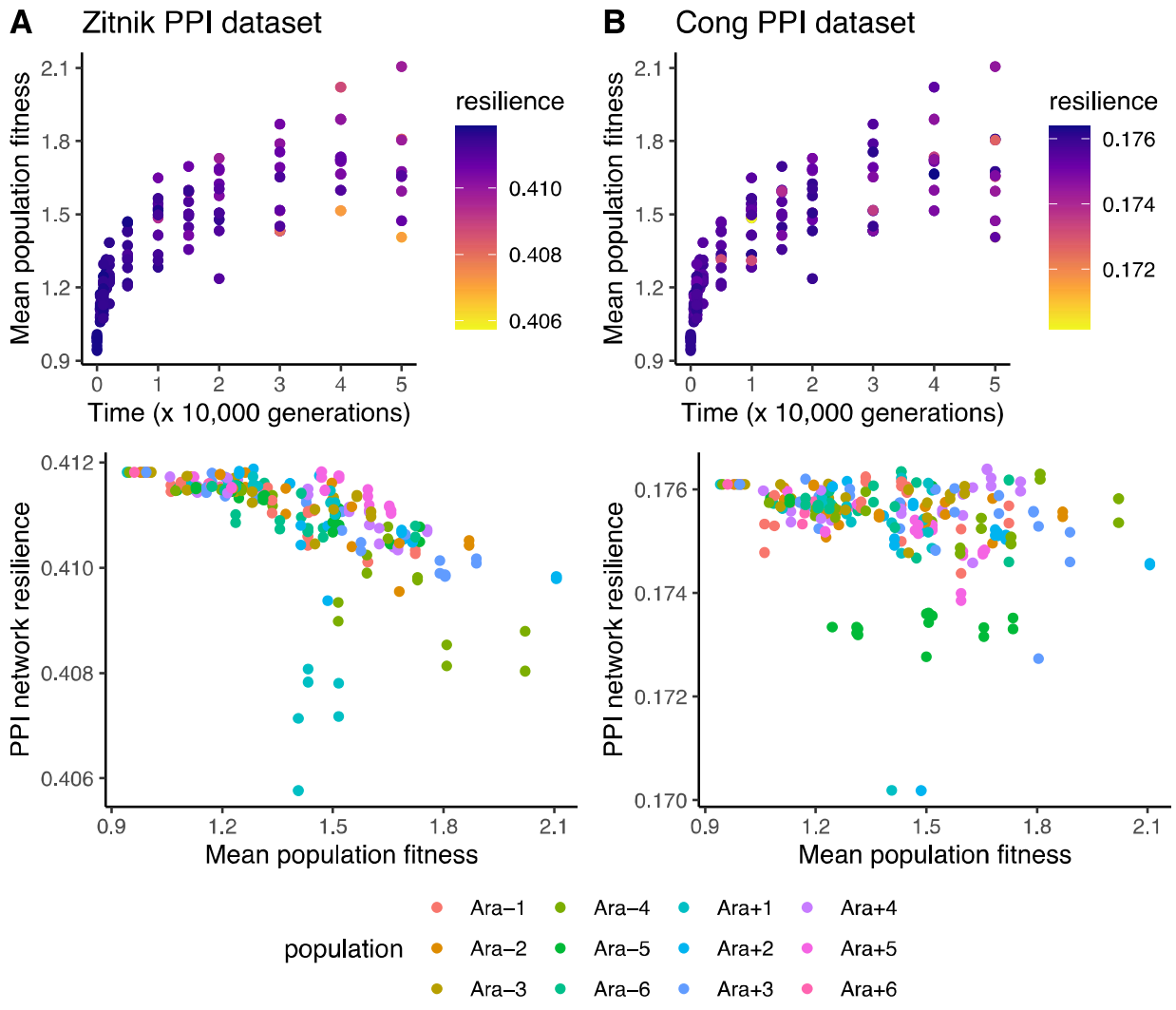
Figure 2. PPI network resilience negatively correlates with mean population fitness, relative to the ancestral LTEE clone, at the times at which the given genomes were isolated from their populations. The top panels show the population fitness data from Wiser et al. (2013). Colors correspond to the PPI network resilience of genomes sampled from a given population at a given time. The bottom panels show the negative correlations between mean population fitness and PPI network resilience. A) Analysis based on the *E. coli* PPI network published in Zitnik et al. (2019). The negative correlation between fitness and resilience is significant (Pearson's product-moment correlation: $r = -0.59$, $p < 10^{-16}$). B) Analysis based on the *E. coli* PPI network published in Cong et al. (2019). The negative correlation between fitness and resilience is significant (Pearson's product-moment correlation: $r = -0.27$, $p < 10^{-4}$).