Title:
Paralog interference preserves genetic redundancy

Authors: Angel F. Cisneros†[1-4], Florian Mattenberger†[1-5], Isabelle Gagnon-Arsenault[1-4], Yacine Seffal[2-4], Lou Nielly-Thibault[1-4,6], Alexandre K. Dubé[1-4], María del Mar Varela Vásquez[1-4,7], María Camila Muñoz Vega[1-4,8], Christian R. Landry*[1-4,6]

Affiliations:
[1]Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Université Laval, G1V 0A6, Québec, Canada
[2]Institut de biologie intégrative et des systèmes, Université Laval, G1V 0A6, Québec, Canada
[3]PROTEO, Le regroupement québécois de recherche sur la fonction, l'ingénierie et les applications des protéines, Université Laval, G1V 0A6, Québec, Canada
[4]Centre de recherche sur les données massives, Université Laval, G1V 0A6, Québec, Canada
[5]Current address: Gregor Mendel Institute of Molecular Biology, Austrian Academy of Sciences, 1030, Vienna, Austria
[6]Département de biologie, Faculté des sciences et de génie, Université Laval, G1V 0A6, Québec, Canada
[7]Departamento de Ciencias Biológicas, Bioprocesos y Biotecnología, Facultad Barberi de Ingeniería, Diseño y Ciencias Aplicadas, Universidad Icesi, 760008, Cali, Colombia.
[8]Departamento de química, Facultad de ciencias naturales y exactas, Universidad del Valle, 760032, Cali, Colombia

† Indicates equal contributions
*Corresponding author
Correspondence:
Tel: +1 418 656 3954; E-mail: christian.landry@bio.ulaval.ca

Abstract:

Models of gene duplication often assume that loss-of-function mutations neutrally promote the return to the ancestral singleton state. They thus ignore the potential functional interference between duplicated proteins stemming from their physical interactions. Here, we show that for heteromerizing paralogs, such interference potentiates negative selection on loss-of-function mutations. This effect maintains genetic redundancy over longer timescales depending on the rate and severity of loss-of-function mutations. We experimentally estimate that around 6% of deleterious substitutions for a representative tetrameric protein interfere with a second copy. Interfering mutations typically disrupt either catalysis or the final step of protein complex assembly, with varying degrees of severity. Our work shows that paralog interference potentiates

the negative effects of loss-of-function mutations, contributing to the preservation of genetic redundancy.

Keywords
Gene duplication / loss of function / protein interactions/ paralog interference

Main text:

Introduction

Gene duplication is the main source of genetic redundancy (*1–3*). Immediately after a duplication event, the two paralogs are usually functionally redundant as they have identical sequences and regulatory elements. A common assumption is that loss-of-function (LOF) mutations in one copy are rendered neutral by the presence of its paralog. Since LOF mutations occur much more frequently than other types of mutations (*4*, *5*), the most likely outcome is thought to be a return to a single functional copy. Alternative outcomes include the divergence of functions between paralogs leading to new functions (*6*) or to the partitioning of ancestral ones (*7*). These two scenarios happen gradually and require the persistence of the two copies in the genome for long enough so rare function-affecting mutations could occur. Therefore, the timescales of maintenance of redundant duplicates have consequences for the robustness of cells to mutations and for adaptive evolution. More generally, the number of genes in the genome depends on the balance between the rates of duplication and gene loss (*8*). Any biochemical or biophysical factor that decreases the rate of gene loss will lead to the maintenance of a larger number of genes and the expansion of gene families.

Previous studies have identified factors that affect the rate of gene loss. For example, genes encoding self-interacting (homomeric) proteins (*9*, *10*) and those more likely to acquire dominant deleterious mutations (*11*, *12*) appear to be preferentially retained. These two observations could be mechanistically connected, since the duplication of genes encoding homomeric proteins introduces paralogous proteins that physically interact with each other (*13*, *14*). Classical models for the retention of gene duplicates assume that LOF impacts only the gene copy bearing a mutation, i.e. that the genes contribute independently to function. However, because of their physical interactions, duplicated homomeric proteins are not independent. As a result, LOF mutations in one paralogous protein could interfere with the normal function of the other paralog, causing a stronger decrease in function than LOF mutations for non-interacting paralogs (Fig 1A-B). Negative selection against such deleterious LOF mutations could preserve the two interfering paralogs in the initial redundant state for a longer period than paralogs that do not interact. Since this phenomenon would be analogous to dominant negative alleles at the same locus in diploids, whereby LOF mutations in one allele can inactivate the protein produced by the other allele due to their physical assembly (*15*, *16*), we refer to it as negative paralog interference. Considering that at least 30% of genes code for homomeric proteins (*17*, *18*), negative paralog interference could have a profound effect on shaping extant genomes. Here, we use models and experiments to test if and how homomerization provides access to deleterious mutations that ultimately preserve redundancy.

Negative paralog interference increases the time required to lose genetic redundancy

To quantify the impact of interference, we examine transitions of populations from an initial state having two redundant functional paralogs to a single functional one, that is, the loss of redundancy. We consider three groups of mutations that eliminate the functional contribution of a gene encoding an enzyme that is only active as a homodimer: loss-of-function (LOF), loss-of-affinity (LOA), and loss-of-expression (LOE) (Fig. 1A). These three types of mutations would respectively disrupt its catalysis, dimerization, and synthesis. Other mechanisms such as loss-of-translation or loss-of-stability could be considered but their effect on function would be the same as LOE. The transitions from a state with redundant paralogs to one where one paralog is inactivated by LOF, LOA, or LOE mutations depend on the respective mutation rates (μ_F, μ_A, μ_E), the effective population size (N), and the probabilities of fixation (P_{Fix}). Since these transitions result in the loss of redundancy, the time needed to lose redundancy (T) is:

$$T = \frac{1}{\mu_F \, N \, P_{Fix(F)} + \mu_A \, N \, P_{Fix(A)} + \mu_E \, N \, P_{Fix(E)}}$$

To evaluate the effect of interference on T, we compare the expressions for T with and without interference, which differ only in the effects of LOF mutations on protein activity. If mutations in one paralog were to inactivate it without affecting the function of the other, then LOF, LOA, and LOE mutations would only lead to a loss of half of the total biochemical activity produced by the genes (Fig. 1A). Assuming a typical fitness function with diminishing returns from total protein activity (*19–21*) (Fig. 1B), without interference, the probabilities of fixation for LOF, LOA, and LOE mutations would be identical to that of the loss of one copy ($P_{Fix(F)} = P_{Fix(A)} = P_{Fix(E)} = P_{Fix(Loss)}$). Conversely, physical interactions that would produce interference would cause LOF mutations in one paralog to affect the function of the other one, while LOA and LOE mutations would not be expected to, at least in the simple case of a dimeric protein that assembles in one step. As a result, with interference, LOF mutations should cause sharper decreases in biochemical activity than LOA and LOE mutations for a homooligomeric complex and thus, they would have a lower probability of fixation ($P_{Fix(F)} < P_{Fix(A)} = P_{Fix(E)} = P_{Fix(Loss)}$) (Fig. 1A).

Gene duplications are more likely to fix if they are neutral or advantageous (*5*, *22–24*). Accordingly, the subsequent return to one copy can be neutral or deleterious. If the loss of one copy is nearly neutral (Fig. 1B-C), LOA and LOE mutations are neutral while LOF mutations can be selected against because they reduce the concentration of total functional complexes. The time needed to lose redundancy with interference (T_{Int}) is greater than without interference (T_{NoInt}) when LOF mutations are more deleterious (smaller $P_{Fix(F)}$), and this effect is potentiated when the rate of LOF mutations (μ_F) increases relative to the sum of rates of LOA and LOE mutations ($\mu_A + \mu_E$) (Fig. 1C). Even when the loss of one copy is slightly deleterious (Fig. 1D), paralogs stay redundant for longer with interference when $P_{Fix(F)}$ is smaller than $P_{Fix(Loss)}$ (Fig. 1E). This condition is sufficient to extend the time needed to lose redundancy (see Methods). Thus, the rate

and severity of LOF mutations that interfere negatively with the sister copy contribute to the preservation of genetic redundancy.

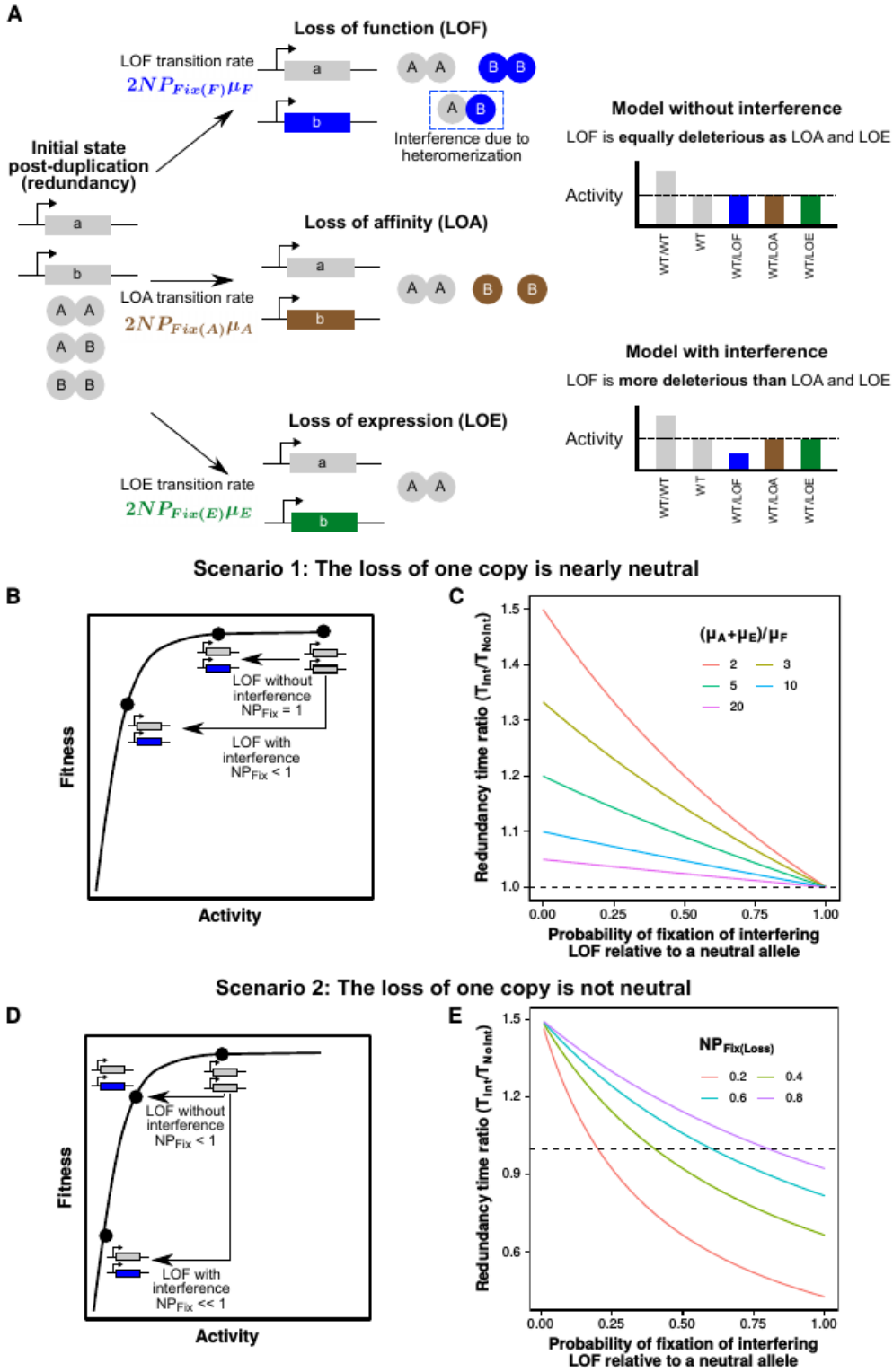

Figure 1. Interference between paralogs leads to longer periods of genetic redundancy. A) Model of the transitions from the initial state to loss of function, affinity, or expression of one copy. These transitions depend on their respective mutation rates (μ) and probabilities of fixation (P_{Fix}). LOF mutations are equally deleterious as LOA

and LOE mutations in the classical system, but more deleterious than LOA and LOE mutations in the system with interference. **B-E)** Two scenarios for the effect of interference on the maintenance of redundant paralogs. The halving in activity due to LOF, LOA, and LOE mutations in the model without interference can be nearly neutral ($NP_{Fix(Loss)} = 1$) (B-C) or deleterious ($NP_{Fix(Loss)} < 1$) (D-E). The ratio of the time needed to lose redundancy in a system with interference (T_{Int}) versus a system without interference (T_{NoInt}) is presented in both scenarios (C, E) for varying probabilities of fixation for LOF mutations and mutation rates. $(\mu_A+\mu_E)/\mu_F$ varies in panel C while in panel E, it is kept constant at 2.

A study system to identify and quantify the impact of interfering mutations

While we have focused on dimers so far, paralogous proteins can also assemble into larger oligomeric complexes. Dimeric paralogs with the same abundance and binding affinity will form a mixture of 25% of each homodimer and 50% of the heterodimer. If interfering mutations inactivate the mutant homodimers and the heterodimer, the fraction of functional complexes would be reduced to 25%. Such interfering mutations would be more deleterious than the loss of one copy, which decreases activity to 50%. The effects of negative interference would be even worse for higher order assemblies. For tetramers, the expected fraction of oligomers containing only subunits without the interfering mutation is around 6% (Fig. 2A) (*25*). In addition, for larger oligomeric complexes, the presence of multiple assembly interfaces could potentially make LOA also contribute to interference as partially assembled complexes could be inactive (*16*).

As shown in Fig. 1, a key factor affecting the maintenance of paralogs is the fraction of LOF mutations that can interfere with the other copy. We set out to estimate this fraction using a small homomeric protein for which we can estimate the functional impact of mutations by measuring antibiotic resistance (*26*). The *E. coli* plasmid-encoded DfrB1 enzyme uses NADPH to catalyze the reduction of dihydrofolate (DHF), the same reaction as the *E. coli* dihydrofolate reductase (FolA), although the two proteins are not homologous. DfrB1 forms a homotetramer whose active site is located in the central pore (Fig. 2B). While FolA is inhibited by the antibiotic trimethoprim (TMP), DfrB1 is insensitive to this molecule (*27*). We monitor the impact of mutations on DfrB1 through its complementation of the growth defect caused by TMP (Fig. 2C). Using an inducible promoter, we tune the expression level of DfrB1 to be limiting for growth such that both adaptive and deleterious mutations can be detected.

We experimentally simulated a gene duplication by expressing two copies of the gene on separate plasmids while preserving their promoter and thus, their transcriptional regulation. To selectively amplify the mutated paralog for subsequent high-throughput library sequencing (see Methods, fig. S1A), one copy of DfrB1 bears synonymous mutations in codons 4-9 (Syn4-9). We first confirmed that these mutations do not affect growth in the presence of TMP (fig. S1B) and that protein copies from the different plasmids form heteromers (fig. S1C). Then, we measured the effect of all amino acid substitutions of the enzyme downstream of codon 9 while expressing a singleton mutated gene (experiment 1) or when duplicated alongside a WT copy (experiment 2) (Fig. 2D, fig. S2A).

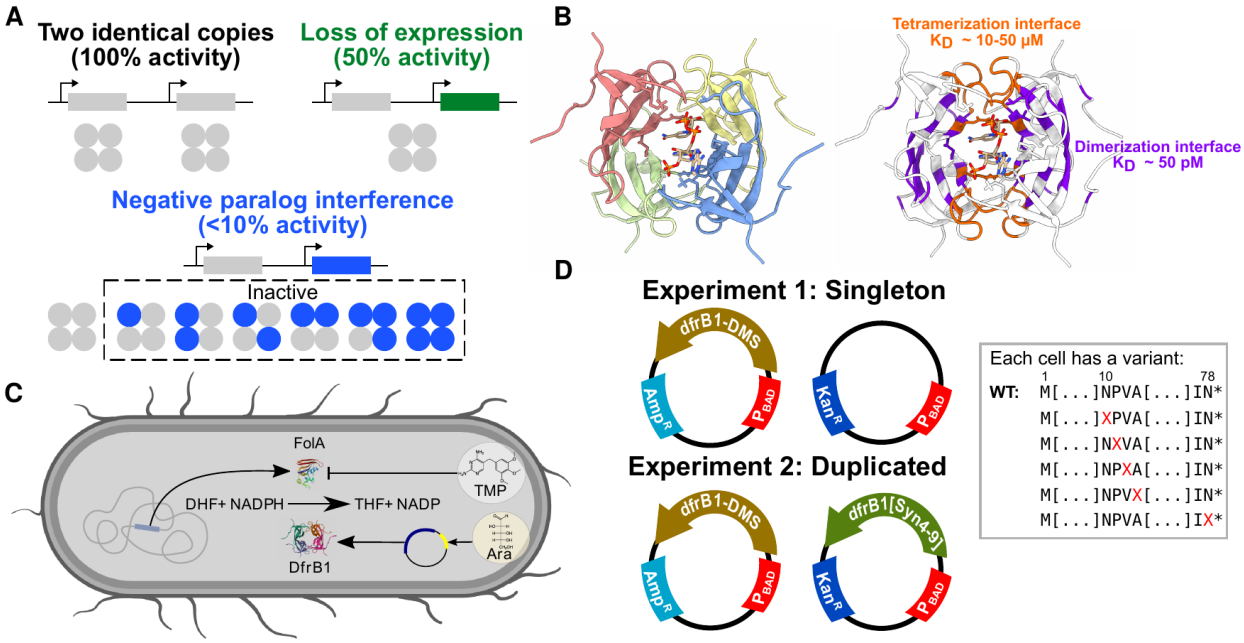

Figure 2: Study model to identify and measure the impact of negative interference mutations. **A)** The loss of expression of one copy halves the total concentration of complexes, while negative paralog interference produces a high fraction (>90%) of inactive tetramers. **B)** Structure of DfrB1. *Left:* The four identical subunits are shown in red, yellow, green, and orange. The two substrates (dihydrofolate and NADPH) are shown in the active site (central pore). *Right:* Residues in direct contact at the dimerization and tetramerization interface are colored in orange and purple, respectively, along with their known dissociation constants (see methods). Coordinates were taken from PDB: 2RK1 (*28*) and visualized with ChimeraX (*29*). **C)** Study system in which DfrB1 complements the inhibition of FolA by trimethoprim (TMP) in *E. coli*. **D)** Experimental simulation of gene duplication using two plasmids. One plasmid contains an ampicillin resistance marker (Amp^R) and a complete library of *dfrB1* mutations (*dfrB1-DMS*) from codon 10 onwards. The second plasmid contains a kanamycin resistance marker (Kan^R) along with either an empty insertion site or a WT copy of *dfrB1* bearing synonymous mutations in codons 4-9 (*dfrB1[Syn4-9]*). The two copies are controlled by the same arabinose-inducible promoter..

Mutations causing negative paralog interference have mild destabilizing effects and cluster in structural hotspots

Globally, combining the DMS library with a WT copy led to a decrease in the magnitude of fitness effects (Fig 3A, tables S1-S2). Most mutations that were deleterious when expressed as a singleton became milder when co-expressed with a second WT copy. We note that on the other end of the spectrum, some mutations were beneficial only in the singleton. In line with a previous study (*26*), we found that some of these mutations lead to an increase in DfrB1 abundance (fig. S3, table S3).

Mutations causing negative paralog interference have a more deleterious fitness effect than stop codons. We initially defined high-confidence negative paralog interference candidates as mutations with fitness effects below the mean minus 2.5 standard deviations of stop codons that

are in the same competing pool (see Methods). In the experiment expressing the singleton mutant library, 986 out of 1310 possible missense mutations (75.3%) were deleterious (worse than the 2.5 percentile of WT synonymous codons) and 766 (58.5%) were within 2.5 standard deviations of the mean of stop codons (Fig. 3A). Only two mutations were below the lower limit of this interval, which could happen for instance if some mutants were toxic or due to experimental noise. When the mutant library was expressed alongside the WT copy, 967 (73.8%) missense mutations were deleterious, 554 (53.7%) behaved like stop codons, and 63 had more deleterious effects than stop codons (representing 4.8% of available missense mutations and 6.5%of deleterious mutations) (Fig. 3A, table S4). As a result, these 63 mutations are strongly selected against despite having a redundant WT copy of the gene present, which results from negative interference.

We set out to identify the general properties of interfering mutations. We computationally estimated the effects of all mutations on the folding, dimerization, and tetramerization energies of DfrB1, as well as on the binding affinities of the DfrB1 tetramer for its substrates DHF and NADPH (see Methods, table S2) (*30*, *31*). Since the catalytic core of the protein is formed by the contribution of the four chains, a single destabilizing mutation in one chain of the complex could be sufficient to reduce or prevent any activity. Compared to the rest of mutations, those causing negative paralog interference are predicted to have milder destabilizing effects on protein folding, slightly more stabilizing effects on tetramerization affinity, and more disruptive effects on NADPH binding (Fig. 3B). We did not observe significant differences in terms of dimerization affinity and DHF binding (Fig. 3B). NADPH is the first substrate to bind, which is consistent with its disruption being more important than that of DHF binding (*32*). These results hold when comparing the interfering mutations against others occurring at the tetramerization interface and active site (fig. S4).

Considering that negative interference mutations have different effects on the two interaction interfaces (Fig. 3B), we asked if those mutations were clustered together in the protein structure. We calculated the relative enrichment of negative paralog interference mutations across protein regions including the two interfaces, the active site, and buried residues (see Methods). Mutations are most strongly enriched in residues interacting with the substrates, followed by the tetramerization interface (Fig. 3C). In turn, they are underrepresented in the dimerization interface, confirming the differences between the two interfaces. The DfrB1 active site and its tetramerization interface are thus hotspots of negatively interfering mutations.

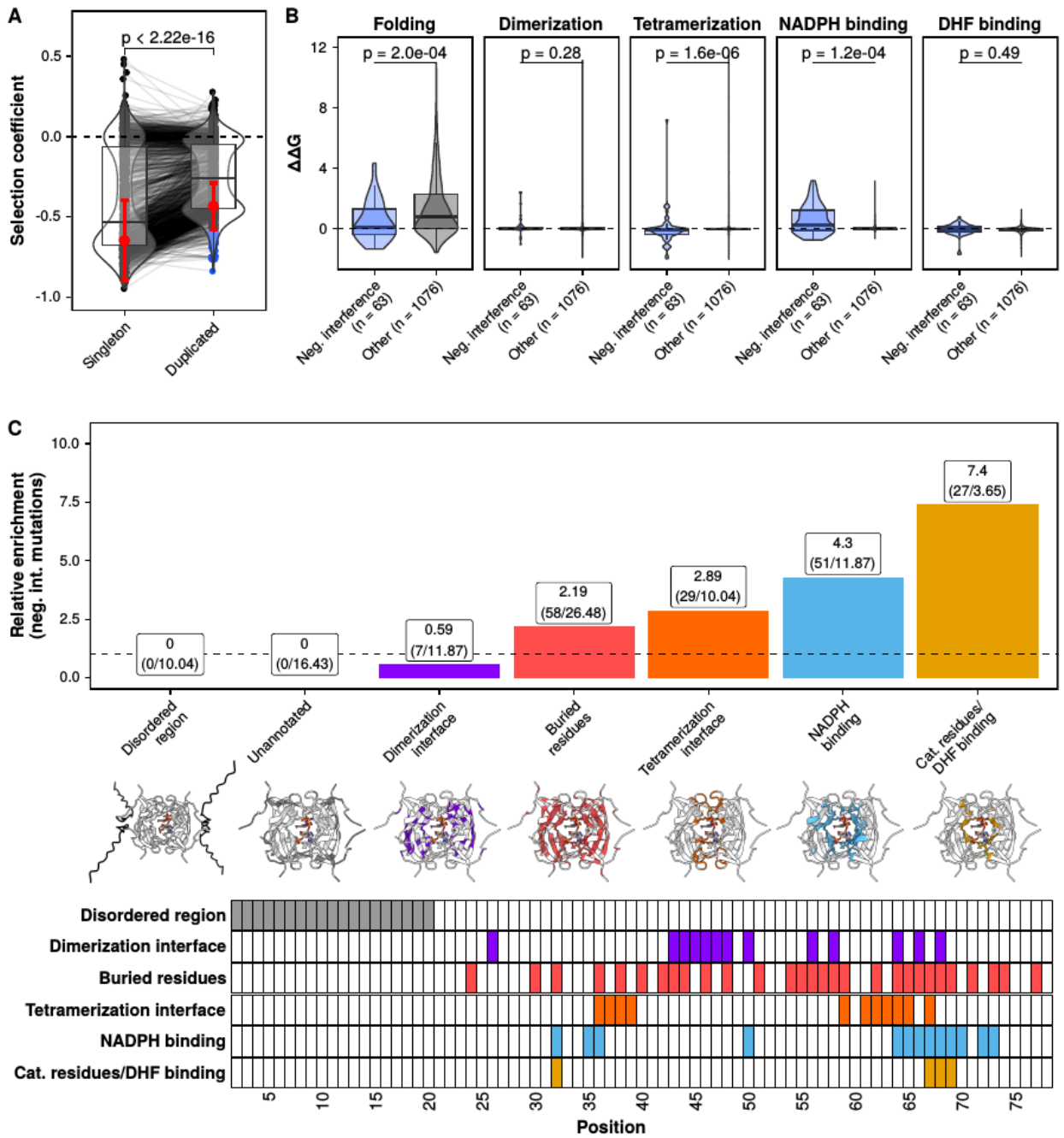

Figure 3: General properties of negative interference mutations in an oligomeric complex. A) Distributions of measured selection coefficients for the mutant library in the singleton (n = 1443) and duplicated (n =1448) backgrounds. Red dots and error bars represent the mean +/- 2.5 standard deviations of the selection coefficient of stop codons. Negative interference mutations (blue) are conservatively defined as those that are more deleterious than the mean minus 2.5 standard deviations of stop codons in the duplicated background. The p-value was calculated using a paired Mann-Whitney U test. **B)** FoldX-predicted effects of negative interference mutations on DfrB1 folding, dimerization, tetramerization, and NADPH and DHF binding compared to the effects of other mutations. p-values were calculated using Mann-Whitney U tests. **C)** Relative enrichment of negative interference mutations across protein regions, calculated as a ratio of observed over expected counts. The dashed line indicates a relative enrichment of 1, in which the observed counts of negative interference mutations are identical to the

expected counts. Cartoons indicate the different protein regions tested for the enrichment of negative interference mutations. Positions along the DfrB1 sequence belonging to each region are shown at the bottom.

The magnitude of deleterious effects depends on the interference mechanism

Since interfering mutations cluster at the catalytic site and the tetramerization interface, we set out to dissect the contributions of LOF and LOA. We selected two representative mutations: the active-site mutation Q67C due to its severe effect on catalytic activity (*33*) and the S59Y substitution since it is predicted to be highly destabilizing for tetramerization (predicted $\Delta\Delta G$ = 29.4 kcal/mol, table S2). Expressed as a single gene copy, the LOA caused by S59Y is as deleterious as the LOF mediated by Q67C (fig. S5, table S5). Conversely, when expressed with a WT copy, the interference caused by Q67C results in more deleterious effects than stop codons in both bulk competition (Fig. 4A) and pairwise competition (Fig. 4B) assays. The same is observed for the interfering effect of S59Y although it was only statistically significant in the more sensitive pairwise competition assay (Fig. 4A-B). Thus, LOA mutations at the tetramerization interface can also produce interference, albeit with slightly less deleterious effects. These results suggest that our estimate derived from the bulk competition experiment that 6.1% of deleterious mutations cause negative interference is likely a lower bound, dominated by LOF effects.

We initially assumed that LOA mutations do not cause interference because we considered a simple dimer (Fig. 1A). However, since DfrB1 assembles as a dimer of dimers (*32*), protein copies harboring substitution S59Y could sequester WT proteins into dimers that cannot tetramerize. While LOF substitutions would produce a majority of inactive heterotetramers by binding with WT copies, LOA substitutions disrupting the tetramerization interface would capture a smaller but significant proportion of WT copies into inactive dimers (Fig. 4C). The magnitude of deleterious effects for each interference mechanism could depend on the fraction of WT copies sequestered into inactive complexes (Fig. 4D). On the other hand, LOA substitutions at the dimerization interface would lose the capacity to bind the WT protein and could therefore not cause interference.

To validate our model, we first confirmed that the S59Y leads to a significant decrease in the stability of the tetramer using calorimetry (Fig. 4E, fig. S6A, table S6) and that the oligomeric form is lost using native gel electrophoresis (fig. S6B). If our model is correct, combining S59Y and Q67C on the same subunit should relieve the interfering impact of Q67C on activity by reducing its assembly with WT subunits. Indeed, the double mutant S59Y+Q67C was less deleterious than Q67C alone when expressed with the WT (Fig. 4F, table S7). In fact, the S59Y+Q67C mutant also has a higher fitness than S59Y alone when expressed with the WT (Fig. 4F), potentially because the two combined mutations further destabilize the protein (Fig. 4E), reducing its potential for interference. We conclude that the hierarchical nature of different mechanisms of negative interference could produce varying deleterious effects due to the ways

they impact protein activity or protein complex assembly.

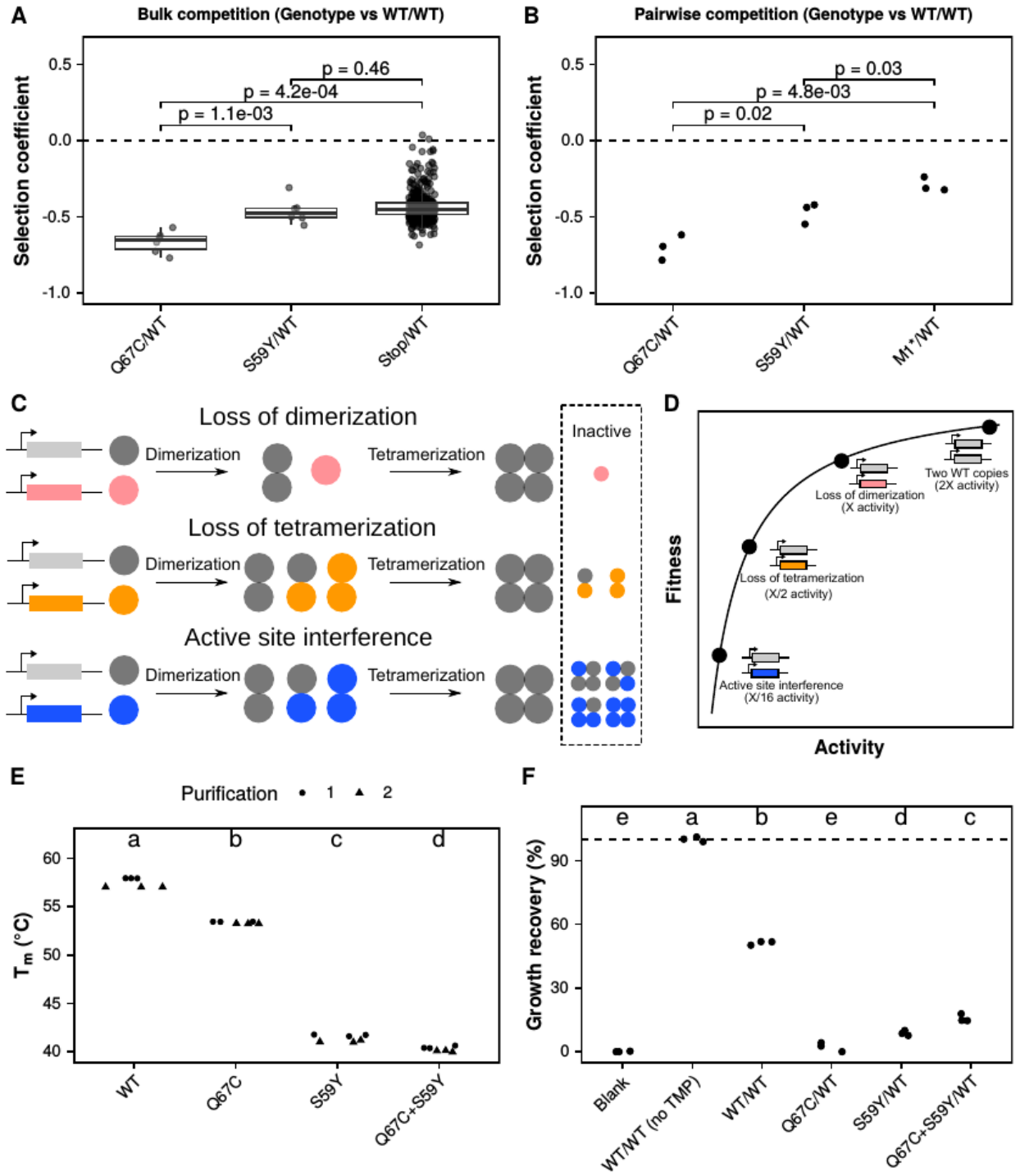

Figure 4. Mechanisms mediating the magnitude of negative interference LOF and LOA mutations. A-B) Selection coefficients for Q67C and S59Y mutations in duplication with a WT gene using two WT copies as a reference in DMS bulk competition experiment (A) and direct pairwise competition assays (B). Stop codons from position 10 onwards (Stop/WT) and at position 1 (M1*/WT) are shown for comparison. p-values in panels A-B were

calculated using t-tests. **C)** Model of paralog interference for tetramers. The loss of dimerization eliminates physical interactions and potential for interference. Mutant protein copies with disrupted tetramerization interfaces do not tetramerize and instead sequester WT copies in heterodimers. Mutations in catalytic sites do not disrupt tetramerization, sequestering a large fraction of WT copies in inactive tetramers. **D)** Relative positions of each genotype along a fitness curve based on the concentration of active tetramers. **E)** Temperature at which tetramers harboring different mutations dissociate (T_m). Three replicates are shown for each of two separate purifications. **F)** Comparison of growth recovery at 0.0015% arabinose and 10 µg/mL TMP for different genotypes relative to a WT/WT reference growing in the absence of TMP. Labels at the top of panels E-F indicate significantly different groups as determined by an analysis of variance (ANOVA) and Tukey's post hoc honestly significant difference test ($p < 0.05$).

Discussion

Gene duplication has been shown to occur at a rate roughly comparable to point mutations (*1*, *8*, *34*). While previous work has suggested that homomeric proteins and genes with a high access to dominant deleterious mutations could be preferentially retained (*9–12*), the underlying mechanisms remain unclear. Here, we explore how homomeric paralogs can interfere with each other, producing deleterious mutations that would impact the product of both duplicates, just as dominant mutations do within a locus. We demonstrate that selection against such interfering mutations favors the maintenance of functional redundant paralogs. Using an experimental model, we find that 6% of deleterious mutations for a small homotetrameric protein (DfrB1) are more deleterious when expressed alongside a WT copy than entirely losing one copy.

We identify two broad classes of interference mechanisms based on the detailed study of a tetrameric enzyme. First, active sites formed at the interface between subunits can be inactivated by mutations in one subunit. Around 20% of eukaryotic protein complexes have been reported to have multi-chain ligand binding sites (*35*), with some having known cases of dominant negative effects (*36*). Second, negative interference mutations are also prevalent at the tetramerization interface but depleted at the dimerization interface of DfrB1. Since DfrB1 assembles sequentially into dimers and then into tetramers (*32*), disrupting the dimerization interface eliminates all assemblies, whereas destabilizing the tetramerization interface allows mutant alleles to sequester WT copies into non-productive heterodimers. Dimers of dimers, such as human hemoglobin, are among the most common oligomeric protein complexes (*37*), suggesting these negative interference mechanisms could apply to the retention of many such duplicates. Indeed, it could explain why a large fraction of dominant negative mutations are found at interaction interfaces (*16*). In the case of hemoglobin, since its dimerization interface is strongly heterospecific and necessary for the assembly of tetramers (*38*), mutations disrupting tetramerization have a high potential to sequester the other subunit in non-productive dimers.

The number of genes in a genome depends on the balance between gene birth, which is dominated by gene duplication, and gene loss. The mechanisms described here by which paralogs can interfere with their sister copy alter this balance, with several implications. First, genes that are prone to negative interference, such as those encoding oligomeric proteins, should be maintained by purifying selection, even if they are redundant. As a result, their gene families could increase in size. Second, the longer retention time may help acquire sub- or neofunctionalization mutations that will eventually further enhance their retention time (*7*, *39–41*). Our work therefore illustrates how mutations and biochemistry interact to shape genome content.

Acknowledgments
We thank the Landry lab members for their discussions during the project. We thank Pascale Lemieux, François D. Rouleau, Simon Aubé, and Pavithra Venkataraman for their comments on the manuscript. We also thank Alicia Pageau and François D. Rouleau for bioinformatics support, Claudèle Lemay St. Denis and Joelle Pelletier for help with DfrB1 experiments, and Sarah Otto for comments on our model.

Funding
This work was funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to CRL RGPIN-2020-04844 and by a seed grant from PROTEO. CRL holds the Canada Research Chair in Cellular Systems and Synthetic Biology. AFC was supported by a PROTEO graduate fellowship, the Ministère de l'enseignement supérieur du Québec, the Agencia mexicana para la cooperación y el desarrollo internacional. FM was supported by a Fonds de Recherche du Québec - Santé (FRQS) postdoctoral fellowship 2022-2023-BF15-315411 (https://doi.org/10.69777/315411). LNT was supported by an Alexander Graham Bell Ph.D. fellowship from Natural Sciences and Engineering Research Council of Canada. MMVV and MCMV were supported by Mitacs Globalink Research Internship Fellowships.

Attributions

Conceptualization: AFC, FM, YS, LNT, CRL
Methodology: AFC, FM, YS, LNT, CRL
Investigation: AFC, FM, IGA, YS, LNT, MMV, MCMV
Experiments: FM, IGA, AKD, MMV, MCMV, AFC
Visualization: AFC, FM, YS
Funding acquisition: CRL
Project administration: CRL
Supervision: CRL
Writing – original draft: AFC, CRL, FM, IGA
Writing – review & editing: AFC, CRL, FM, IGA

Competing interests
The authors have no conflict of interest to declare

Data and materials availability
All raw sequencing data are available at SRA BioProject PRJNA1097537 (accession numbers SRR28586753- SRR28586760 and SRR28586807- SRR28586813). Scripts for data analyses and figure generation are available at https://github.com/Landrylab/Paralog_interference_DfrB1.

All software and materials used are listed in the key resources table (table S8).

Materials and Methods

Model of transitions between states

We built a model assuming transitions between four states in a haploid population: the initial state after duplication (two active copies) and the states after the acquisition of a loss-of-function (LOF), loss-of-affinity (LOA), or loss-of-expression mutation (LOE) (Fig. 1). Transitions between these states are considered to be irreversible and to only occur toward the loss of properties. The system can leave the initial state of redundancy by having one of the copies acquire LOF, LOA, or LOE mutations. Transitions between states are given by their respective mutation rates (μ_F, μ_A, and μ_E) and probabilities of fixation, which vary depending on whether the model allows for paralog interference or not. The expected time the system maintains redundancy is inversely proportional to the rates of these transitions, as shown in equation 1:

$$T = \frac{1}{2\,N\left(P_{Fix(F)}\,\mu_F + P_{Fix(A)}\,\mu_A + P_{Fix(E)}\,\mu_E\right)} \qquad (1)$$

where T is the time needed to lose redundancy; N is the effective population size; $P_{Fix(F)}$, $P_{Fix(A)}$, and $P_{Fix(E)}$ are the probabilities of fixation of LOF, LOA, and LOE mutations; and μ_F, μ_A, and μ_E are the rates of LOF, LOA, and LOE mutations.

Classical model - no paralog interference:

In a system without interference (Fig. 1A), mutations in one of the protein copies do not affect the function of the other copy. Thus, the probabilities of fixation in equation 1 for LOF, LOA, and LOE mutations are then identical and can be replaced by a general term ($P_{Fix(Loss)}$) that reflects the loss of activity of one copy, as shown in equation 2:

$$T_{NoInt} = \frac{1}{2\,N\left(P_{Fix(Loss)}\,\mu_F + P_{Fix(Loss)}\,\mu_A + P_{Fix(Loss)}\,\mu_E\right)} \qquad (2)$$

Model with paralog interference:

As opposed to the classical model, mutations in one copy can modify the function of the second copy in a system with paralog interference (Fig. 1B). Since such modifications rely on the physical interaction between the two protein copies, they are only possible for LOF mutations and not for LOA and LOE mutations. As a result, the probabilities of fixation for LOA and LOE mutations are identical ($P_{Fix(Loss)}$), but different from that of LOF mutations ($P_{Fix(F)}$). Equation 1 can then be simplified to obtain the general form of the model with interference, as shown in equation 3:

$$T_{\text{Int}} = \frac{1}{2\,N\left(P_{Fix\,(F)}\,\mu_F + P_{Fix\,(Loss)}\,\mu_A + P_{Fix\,(Loss)}\,\mu_E\right)} \qquad (3)$$

Comparison of time needed to lose redundancy assuming neutrality:

In many cases the presence of a duplicate gene makes the loss of the second copy neutral (*5*, *42*). In such a scenario, one can assume that LOF, LOA, and LOE mutations in one copy are neutral, and thus $P_{Fix(Loss)} = 1/N$. As a result, this assumption allows simplifying the equations of the classical and interference models to obtain equations 4 and 5, respectively:

$$T_{NoInt} = \frac{1}{2\left(\mu_F + \mu_A + \mu_E\right)} \qquad (4)$$

$$T_{\text{Int}} = \frac{1}{2\left(N\;P_{Fix\,(F)}\,\mu_F + \mu_A + \mu_E\right)} \qquad (5)$$

The times needed to lose redundancy for the two models can be compared by using the ratio shown in equation 6:

$$\frac{T_{\text{Int}}}{T_{NoInt}} = \frac{\dfrac{1}{2\left(N\;P_{Fix\,(F)}\,\mu_F + \mu_A + \mu_E\right)}}{\dfrac{1}{2\left(\mu_F + \mu_A + \mu_E\right)}} \qquad (6)$$

After applying the division:

$$\frac{T_{\text{Int}}}{T_{NoInt}} = \frac{2\left(\mu_F + \mu_A + \mu_E\right)}{2\left(N\;P_{Fix\,(F)}\,\mu_F + \mu_A + \mu_E\right)}$$

After cancelling terms and dividing by μ_F:

$$\frac{T_{\text{Int}}}{T_{NoInt}} = \frac{1 + \dfrac{\mu_A + \mu_E}{\mu_F}}{N\;P_{Fix\,(F)} + \dfrac{\mu_A + \mu_E}{\mu_F}}$$

This expression allows comparing the times needed to lose redundancy in both models. To identify the conditions in which paralog interference causes the system to maintain redundancy for longer, we can use the following inequality:

$$1 > \frac{T_{\text{Int}}}{T_{NoInt}} = \frac{1 + \frac{\mu_A + \mu_E}{\mu_F}}{N\ P_{Fix\,(F)} + \frac{\mu_A + \mu_E}{\mu_F}}$$

After multiplying both sides by the denominator of the right-hand side:

$$N\ P_{Fix\,(F)} + \frac{\mu_A + \mu_E}{\mu_F} < 1 + \frac{\mu_A + \mu_E}{\mu_F}$$

$$N\ P_{Fix\,(F)} < 1$$

This expression shows that the time needed to lose redundancy is higher in the model with interference if $NP_{Fix(F)} < 1$. Since $P_{Fix(Loss)} = 1/N$, the requirement needed to preserve redundancy for longer in the system with interference is equivalent to $P_{Fix(F)} < P_{Fix(Loss)}$.

Comparison of times needed to lose redundancy without the assumption of neutrality:

Alternatively, the loss of a gene copy after the duplication may not be neutral. Our model can also be used to compare the times needed to lose redundancy even when this assumption is relaxed, as shown in equation 7:

$$\frac{T_{\text{Int}}}{T_{NoInt}} = \frac{\frac{1}{2\ N\left(P_{Fix\,(F)}\ \mu_F + P_{Fix\,(Loss)}\ \mu_A + P_{Fix\,(Loss)}\ \mu_E\right)}}{\frac{1}{2\ N\left(P_{Fix\,(Loss)}\ \mu_F + P_{Fix\,(Loss)}\ \mu_A + P_{Fix\,(Loss)}\ \mu_E\right)}} \qquad (7)$$

After applying the division:

$$\frac{T_{\text{Int}}}{T_{NoInt}} = \frac{2\ N\left(P_{Fix\,(Loss)}\ \mu_F + P_{Fix\,(Loss)}\ \mu_A + P_{Fix\,(Loss)}\ \mu_E\right)}{2\ N\left(P_{Fix\,(F)}\ \mu_F + P_{Fix\,(Loss)}\ \mu_A + P_{Fix\,(Loss)}\ \mu_E\right)}$$

After cancelling terms:

$$\frac{T_{\text{Int}}}{T_{\text{NoInt}}} = \frac{P_{Fix\,(Loss)}\ \mu_F + P_{Fix\,(Loss)}\left(\mu_A + \mu_E\right)}{P_{Fix\,(F)}\ \mu_F + P_{Fix\,(Loss)}\left(\mu_A + \mu_E\right)}$$

Dividing by μ_F to show the expression in terms of $(\mu_A + \mu_E) / \mu_F$:

$$\frac{T_{\text{Int}}}{T_{\text{NoInt}}} = \frac{P_{Fix\,(Loss)} + P_{Fix\,(Loss)}\left(\frac{\mu_A + \mu_E}{\mu_F}\right)}{P_{Fix\,(F)} + P_{Fix\,(Loss)}\left(\frac{\mu_A + \mu_E}{\mu_F}\right)}$$

This expression allows comparing the times needed to lose redundancy in both models. To identify the conditions in which paralog interference preserves redundancy for longer, we can use the following inequality:

$$1 > \frac{T_{\text{Int}}}{T_{\text{NoInt}}} = \frac{P_{Fix\,(Loss)}\,\mu_F + P_{Fix\,(Loss)}\left(\mu_A + \mu_E\right)}{P_{Fix\,(F)}\,\mu_F + P_{Fix\,(Loss)}\left(\mu_A + \mu_E\right)}$$

After multiplying both sides by the denominator of the right-hand side:

$$P_{Fix\,(F)}\,\mu_F + P_{Fix\,(Loss)}\left(\mu_A + \mu_E\right) < P_{Fix\,(Loss)}\,\mu_F + P_{Fix\,(Loss)}\left(\mu_A + \mu_E\right)$$

After cancelling terms:

$$P_{Fix\,(F)} < P_{Fix\,(Loss)}$$

This expression shows that the time needed to lose redundancy is greater in the model with interference if $P_{Fix(F)} < P_{Fix(Loss)}$. The probability of fixation (P_{Fix}) of a mutation depends on its selection coefficient (s). Assuming small fitness effects, the probability of fixation is on the order of s for haploid cells and $2s$ for diploid cells (*43*). Thus, the condition for redundancy to be preserved for longer in the model with interference becomes:

$$s_F < s_{Loss}$$

Or, in other words, that the interfering LOF mutations are more deleterious than the loss of one copy.

Strains and media

The *Escherichia coli* strains MC1061 and BL21(DE3) were used for cloning and mutagenesis while only BL21 (DE3) was used to conduct all the experiments in this study. All transformations with plasmids were done with in-house protocols for chemocompetent cells. Transformed cells were grown on selective 2xYT + glucose medium (1.0% yeast extract, 1.6% tryptone, 0.2% glucose, 0.5% NaCl, and 2% agar) with either ampicillin (AMP; 100 µg/mL), kanamycin (KAN; 50 µg/mL), or both. For all experiments conducted in liquid medium, bacteria

were grown in Luria-Bertani (LB) medium (0.5% yeast extract, 1.0% tryptone, and 1.0% NaCl) with the corresponding antibiotics (AMP; KAN; or AMP + KAN) and with or without L(+)-arabinose (Ara) and trimethoprim [TMP -10 µg/mL in dimethyl sulfoxide (DMSO)]. Protein induction for purification was achieved in Terrific Broth (TB) medium (2.4% yeast extract, 1.2% tryptone, 0.94% K_2HPO_4, 0.22% KH_2PO_4, and 0.4% glycerol).

Plasmid construction

All of the following PCR amplifications were done using KAPA HiFi HotStart polymerase unless specified otherwise. Oligonucleotide and gBlock sequences and descriptions used in this study can be found in table S9. All plasmids were recovered from bacteria using the Presto™ Mini Plasmid Kit. The integrity of assembled and mutagenized plasmids was confirmed by Sanger sequencing (Plateforme de séquençage et de génotypage des génomes, Centre de recherche du centre hospitalier de Québec–Université Laval (CRCHUL), Canada).

For the bulk competition experiments, *dfrB1* gene was expressed in bacteria from a pBAD vector that allows arabinose-controlled induction (*44*). In addition to the same pBAD-*dfrB1* (that we will hereafter call pBAD-*dfrB1*-AMP to distinguish from the one that has a KAN resistance module) that was used by (*26*), minor modifications of the vector were also made for this study. First, the ampicillin resistance cassette was replaced by a kanamycin resistance cassette to generate pBAD-*dfrB1*-KAN. For this, the pBAD-*dfrB1*-AMP plasmid was amplified by PCR (PCR program: 5 min at 95°C; 35 cycles: 20 s at 98°C, 15 s at 65°C, and 2 min at 72°C; and a final extension of 3 min at 72°C), removing the ampicillin marker. The KANMX resistance cassette was amplified from pMoBY-ACT1 (PCR program: 5 min at 95°C; 5 cycles: 20 s at 98°C, 15 s at 61°C, and 1 min at 72°C; 30 cycles: 20 s at 98°C and 1 min 15 s at 72°C; and a final extension of 3 min at 72°C). Both PCR products were incubated with 20 U of DpnI for 1 hour at 37°C to remove template DNA, and subsequently purified on magnetic beads following the manufacturer's instructions. The plasmid (pBAD-*dfrB1*) and insert (KANMX), with their overlapping regions at each extremity, were then assembled by Gibson DNA cloning (*45*), producing pBAD-*dfrB1*-KAN plasmid.

The pBAD-*dfrB1*-KAN plasmid was then used to generate two additional plasmids, one removing *dfrB1* (pBAD-EMPTY-KAN) as a control for the effect of duplication, and another introducing synonymous mutations from position 4 to 9 in the coding sequence of *dfrB1* (pBAD-*dfrB1[Syn4-9]*-KAN) to selectively amplify each paralogue in subsequent high throughput sequencing library preparation. Plasmid pBAD-EMPTY-KAN was constructed by amplification of pBAD-*dfrB1*-KAN with oligonucleotides that excluded *dfrB1* coding sequence (PCR program: 5 min at 95°C; 35 cycles: 20 s at 98°C, 15 s at 68°C, and 2 min 30 s at 72°C; and a final extension of 8 min at 72°C). The PCR product was incubated with 20 U of DpnI for 1 hour at 37°C to remove template DNA, and directly transformed into BL21 competent cells for re-circularization, producing pBAD-EMPTY-KAN plasmid. Plasmid pBAD-*dfrB1[Syn4-9]*-KAN was constructed by amplifying pBAD-*dfrB1*-KAN with oligonucleotides that allow replacement

of codons 4-9 in *dfrB1* coding sequence (PCR program: 5 min at 95°C; 35 cycles: 20 s at 98°C, 15 s at 61°C, and 2 min 40 s at 72°C; and a final extension of 8 min at 72°C). The PCR product was incubated with 20 U of DpnI for 1 hour at 37°C to remove template DNA, and subsequently purified on magnetic beads. Then, 100 ng of the purified PCR product was incubated with T4 polynucleotide kinase for 30 minutes at 37°C in T4 ligase buffer to perform the 5' phosphorylation, then cooled down on ice for 3 minutes and directly incubated with 10 U of T4 ligase for 1 hour at 37°C. Half of the reaction was then transformed into BL21 cells producing pBAD-*dfrB1[Syn4-9]*-KAN. Often used as a control in many experiments, pBAD-*dfrB1[Syn4-9]*-AMP plasmid was also constructed the same way starting from pBAD-*dfrB1*-AMP.

To perform co-immunoprecipitations of the heteromers, we generated vectors to express 6xHIS and FLAG N-terminal tagged DfrB1 proteins. We amplified the backbone of the desired plasmids (either pBAD-*dfrB1*-AMP or pBAD-*dfrB1[Syn-4-9]*-KAN) with the appropriate primers depending on whether the plasmid was bearing the native or the Syn4-9 sequence *dfrB1*, respectively. The resulting PCR product was digested with 20 U of DpnI and purified using magnetic beads as described previously. Then, 50 ng of the purified products and 10 ng of the desired gBlock containing 6xHIS- or FLAG-tag sequence were assembled by Gibson DNA assembly (*45*) and transformed into BL21 competent cells to retrieve the following plasmids: pBAD-6xHIS-*dfrB1*-AMP, pBAD-FLAG-*dfrB1*-AMP, pBAD-6xHIS-*dfrB1*[Syn-4-9]-KAN and pBAD-FLAG-*dfrB1*[Syn-4-9]-KAN.

In order to generate the plasmids needed to measure protein abundance of different mutants of DfrB1, we performed site-directed mutagenesis based on the QuickChange Site-Directed Mutagenesis System (Stratagene, USA). Briefly, we amplified the pBAD-*dfrB1-sfGFP* plasmid using pairs of primers containing the desired mutation at the center (PCR program: 2 min at 95°C; 22 cycles: 20 s at 98°C, 15 s at 68°C, and 3 min at 72°C; and a final extension of 5 min at 72°C). The PCR products were then incubated for 1 hour at 37°C with 20 U of DpnI enzyme to remove parental DNA, and mutated plasmids were retrieved directly by transformation in MC1061 bacteria. We used this method to generate the following mutants: pBAD-*dfrB1*(N15D)-*sfGFP*, pBAD-*dfrB1*(F18W)-*sfGFP*, and pBAD-*dfrB1*(S20M)*sfGFP*.

For validation of the bulk competition experiment fitness results, some new plasmids were needed. To produce pBAD-*dfrB1[Syn4-9]*(Q67C)-KAN, the mutant pBAD-*dfrB1*(Q67C)-AMP was isolated from the DMS library used in (*26*) and Syn4-9 mutations and KANMX resistance cassette were introduced as described above. Mutants with S59Y mutation or with a stop codon (TGA) at position 1 (M1*) to prevent translation were also generated by site-directed mutagenesis on pBAD-*dfrB1*-KAN (M1*) and pBAD-*dfrB1[Syn4-9]*-KAN (S59Y) as was done to introduce Syn4-9 mutations.

To evaluate the effect of mutations on DfrB1 stability, proteins were expressed from pQE32 (Qiagen, USA) derived plasmids. Starting from pQE32-*dfrB1*, mutations Q67C and S59Y were introduced by site-directed mutagenesis as explained above for mutations introduced in pBAD-

dfrB1-sfGFP at the exception that the elongation time was 4 min instead of 3 min. A second round of mutagenesis was needed to introduce Q67C in pQE32-*dfrB1*(S59Y). At the end, pQE32-*dfrB1*(Q67C), pQE32-*dfrB1*(S59Y) and pQE32-*dfrB1*(Q67C+S59Y) were generated.

To measure the effect of dominant negative mutations on DfrB1 fitness, we used several constructions made in our arabinose-inducible system (pBAD-*dfrB1*-AMP, pBAD-*dfrB1*(Q67C)-AMP, pBAD-*dfrB1*(S59Y)-AMP and pBAD-*dfrB1*-KAN). A last additional construction was needed to express DfrB1 Q67C and S59Y double mutant. pBAD-*dfrB1*(Q67C+S59Y)-AMP was generated by site-directed mutagenesis as explained above, starting from pBAD-*dfrB1*(S59Y)-AMP and introducing Q67C mutation.

Co-immunoprecipitations of heteromers

The plasmids allowing expression of the 6xHIS- or FLAG-tagged proteins were transformed in various combinations into BL21 competent cells. After protein induction, total proteins were extracted using NEBExpress *E.coli* Lysis Reagent as specified by the manufacturer. The soluble protein fraction was then equally divided for co-immunoprecipitation with either Dynabeads™ His-Tag Isolation and Pulldown or ANTI-FLAG® M2 Affinity Gel following the procedure established by the manufacturers. The purified proteins were migrated on an SDS-PAGE gel and detected by western blotting using THE™ Anti Flag-Tag as a first antibody and an IRDye 800CW Goat anti-Mouse IgG as a secondary antibody (*46*).

Bulk competition experiments and library sequencing

We performed bulk competition experiments as previously described (*26*) with slight modifications. First, we transformed BL21 with either pBAD-EMPTY-KAN or pBAD-*dfrB1[Syn4-9]*-KAN. Then, each of the strains was transformed with 75 ng of each position mutant pool from a single-site mutation library of *dfrB1* (pBAD-*dfrB1*-AMP DMS library) generated previously by a PCR-based saturation mutagenesis method (*26*). Positions 1 to 10 were omitted because they would not be captured by our sequencing approach. These positions are in the N-terminal disordered region of the protein and have been shown to be largely insensitive to mutations. Transformation efficiency was estimated by counting the colonies on each plate. A yield of >640 colonies for each position was needed to ensure a minimum of a 10x coverage for each codon mutant. When the desired transformation efficiency was achieved, all the colonies on the plates were resuspended in 5 mL of 2xYT medium, OD_{600} was measured for each pool, and 15% glycerol was added to the medium to store each pool individually at -80°C for further use. In parallel, pools were equally mixed at a total OD_{600}=25 (OD_{600} of individual pools were adjusted to ~0.36) to generate a starting mutant master pool for the bulk competition assay (fig. S2A).

Bulk competition assays were performed by first inoculating 5 mL of LB AMP+KAN medium with the mutant master pool (in either the background without (pBAD-EMPTY-KAN) or with (pBAD-*dfrB1[Syn4-9]*-KAN) an additional copy of DfrB1) to an OD_{600} = 0.01 and grown

overnight at 37°C under agitation (250 rpm). Cultures were then diluted 1:100 into 5 mL of LB AMP+KAN containing 0.0015% arabinose (LB AMP+KAN+Ara) to induce the expression of *dfrB1* to a level at which the culture reaches 50% of growth recovery in the presence of TMP (see Bacterial growth curves section below, fig. S2B). After 18 hours of incubation at 37°C under agitation (250 rpm), 3 mL of the culture was used to extract plasmid DNA of the initial timepoint (t0). In parallel, three biological replicates were generated by diluting the culture to an OD_{600} = 0.025 into 5 mL of LB AMP+KAN+Ara+TMP. The cells were incubated at 37°C under agitation (250 rpm) until an OD_{600} of 0.8 was reached (roughly five generations). The cultures were diluted 1:40 again into 5 mL of fresh LB AMP+KAN+Ara+TMP and incubated at 37°C under agitation (250 rpm) until an OD_{600} of 0.8 was reached again. At this step, 3 mL of the culture were used to extract plasmid DNA of the final timepoint (t10). At both timepoints (t0 and t10), glycerol stocks were generated by adding glycerol to a final concentration of 15% and kept at -80ºC.

To ensure that only *dfrB1* sequence from the DMS library and not from the duplicated paralog would be amplified, PCR specificity was first tested using primers targeting WT or synonymous DfrB1 sequence at codons 4-9 in PCR reactions with either pBAD-*dfrB1[Syn4-9]*-AMP or pBAD-*dfrB1*-KAN as template (fig. S1A) (PCR program: 5 min at 95°C; 35 cycles: 20 s at 98°C, 15 s at 59°C, and 20 s at 72°C; and a final extension of 2 min at 72°C). Then, plasmid DNA from t0 and t10 was used for high-throughput DNA sequencing library preparation that involved two PCR steps. A first PCR was performed to amplify *dfrB1* from the DMS library and to add the primer binding sites for the Illumina i5 and i7 Nextera indexes (table S10). This PCR was done using as template 1 ng of the plasmid DNA for t0 and t10 obtained from the bulk competition assay using primers with a variable number of Ns in the sequence (from 0 to 3) to generate diversity in the amplicon pool (PCR program: 3 min at 98°C; 20 cycles: 30 s at 98°C, 15 s at 60°C, and 30 s at 72°C; and a final extension of 1 min at 72°C). The PCR product was verified on a 0.8% agarose gel. Then a second PCR was performed using primers to add the i5 and i7 Nextera Illumina indexes for automatic demultiplexing at the end of the sequencing process. This PCR was done using 1 μL of a 1:1000 dilution of the first PCR as a template (PCR program: 3 min at 98°C; 18 cycles: 30 s at 98°C, 15 s at 61°C, and 35 s at 72°C; and a final extension of 1 min at 72°C). Each reaction for the second PCR was performed in quadruplicate and then combined. Each sample was amplified with a different combination of primers (table S10). Finally, the PCR product was verified on a 0.8% agarose gel, purified using magnetic beads, and sent for paired-end 250-bp sequencing on an Illumina NovaSeq 6000 (Plateforme de séquençage et de génotypage des génomes, CRCHUL, Canada).

Sequencing data analysis

The quality of the sequencing data was analyzed using FastQC version 0.11.4 (*47*) and the paired reads were merged with Pandaseq version 1.1.5 (*48*). Cutadapt version 4.2 (*49*) was used to trim the reads by detecting adapters preceding and following positions 10 and 78 respectively using the following argument (*-a AGTAGCAATGAAGTCAGT...GTTTAAACGGTCTCCAGC*) to retain only the range of positions on which DMS was performed. The trimmed identical reads were

aggregated using vsearch 2.15.1 (*50*) and aligned to the reference sequence of *dfrB1* using EMBOSS Needle version 6.6.0.0 (*51*) to identify the mutant codons in each read. Mutations to stop codon TAG were not considered in the data analysis, since this stop codon has been shown to have a lower termination efficiency in *E. coli* (*52*, *53*). Stop codons at positions 15 and 78 were also excluded because they were outliers with high fitness. Variants were counted for each sample and normalized by the total number of reads in the library. The dataset was filtered to remove all the variants with <100 counts at t0 to ensure a high sensitivity and limited bias for our selection coefficient calculations. The resulting filtered read proportions were used to calculate selection coefficients based on changes in frequency for each variant between the starting and the end point of the bulk competition assay normalized to the median of WT. The same equation was used as in Cisneros et al. (*26*):

$$s = \frac{\log_2 \left(\frac{\frac{N_{mut,\,t_{10}}}{median\left(N_{WT,\,t_{10}}\right)}}{\frac{N_{mut,\,t_0}}{median\left(N_{WT,\,t_0}\right)}} \right)}{k}$$

where s is the selection coefficient, N is the number of reads for the corresponding mutant at a specific time point, and k is the number of generations (k = 10). The final selection coefficients were the average of those obtained from three replicates. Selection coefficients for amino acid substitutions were calculated as the median of the selection coefficients of all corresponding synonymous variants (table S2).

Identification of negative interference mutations

Negative interference amino acid substitutions (table S4) were defined as those that had a fitness below the mean of nonsense mutations minus 2.5 standard deviations in the experiment with the duplicated DfrB1 background. Selection coefficients were calculated as indicated above for all missense and nonsense mutations. Mutations were compared to the mean of nonsense mutations and visualized in Fig. 3A-B.

The enrichment of negative interference mutations was calculated as the ratio of observed over expected negative interference mutations. To calculate the expected numbers of negative interference mutations for each protein region, we first calculated the overall percentage of negative interference mutations by dividing their total count over the total number of missense mutations. Then, we multiplied this overall percentage by the number of missense mutations possible in each region (the number of positions times 19 missense mutations). The resulting number was the expected number of negative interference mutations for each protein region.

Protein abundance measurements by flow cytometry

We use the measure of GFP level by cytometry to approximate DfrB1 abundance. A first preculture in 2 mL LB AMP medium was grown with bacteria containing the plasmid of interest (pBAD-*sfGFP* (negative control), pBAD-*dfrB1*-*sfGFP*, pBAD-*dfrB1*(N15D)-*sfGFP*, pBAD-*dfrB1*(F18W)-*sfGFP*, pBAD-*dfrB1*(S20M)-*sfGFP* and pBAD-*dfrB1*(E2R)-*sfGFP* (positive control)). As for the bulk competition assay, after an overnight incubation at 37°C with agitation (250 rpm), cells were diluted 1:100 in 2 mL of fresh medium with the addition or not of 0.2% arabinose. Following an 18-hour incubation at 37°C (250 rpm), cells were diluted again at OD_{600} = 0.025 in a final volume of 2 mL of fresh medium again containing or not arabinose. Cultures were then incubated as above for 2 hours. At this final time point, small aliquots of cells were taken and diluted in sterile filtered water to an OD_{600} = 0.05 in 200 µl. GFP fluorescent measurements and forward scatter (FSC) and side scatter (SSC) data were collected from a Guava easyCyte HT cytometer (Cytek®, USA). From the cytometry data, *E. coli* cells were selected on the basis of FSC and SSC. From the selected data points, the GFP fluorescence signal was measured after excitation with a blue laser (wavelength, 488 nm) and detection in the green channel (525/30 nm). The experiment was run in triplicate.

Direct pairwise competition assays between mutants

For validation of the bulk competition experiment fitness results, we performed a direct competition assay using a modified quantitative PCR method for fitness measurements relative to a reference competitor (*54–57*) (fig. S7A). For all assays, each competitor bears two plasmids, conferring resistance to both ampicillin and kanamycin, and both plasmids had either the Syn4-9 mutations in *dfrB1* or not. In total three genotypes (Q67C/WT; S59Y/WT; and M1*/WT) were tested in pairwise competition against WT/WT duplication. First, all competitors were grown overnight individually in a selective LB AMP or KAN medium with 0.0015% arabinose at 37ºC under agitation (250 rpm). For the grown cultures of each competitor, OD_{600} was measured, adjusted to OD_{600} = 1 and then combined at a 1:1 ratio in 500 µL. A 100 µL aliquot was kept at 4ºC for using directly 1 µL in the downstream qPCR. The combined competitors were then diluted to an initial OD_{600} = 0.025 in LB AMP+KAN medium with 0.0015% arabinose and TMP and incubated for 18 hours at 37ºC under agitation (250 rpm). Final OD_{600} was measured to calculate the number of generations that had passed and 1 µL of the culture was directly used to quantify the replication of each competitor by quantitative PCR (QuantStudio 6 Pro Real-Time System (Thermo Fisher Scientific, USA)) using Fast SYBR™ Green Master Mix and primers specific to pBAD-*dfrB1* or pBAD-*dfrB1[Syn4-9] (58)*. A standard curve was performed using 5-fold dilutions of purified plasmid of pBAD-*dfrB1*-AMP and pBAD-*dfrB1[Syn4-9]*-AMP. (fig. S7B). All direct pairwise competition experiments were performed in triplicate.

Selection coefficients for each competitor relative to the other were calculated using a similar equation to the one used above in the bulk competition experiment:

$$s = \frac{\log_2\left(\frac{R(t)}{R(0)}\right)}{n}$$

where *R(0)* and *R(t)* represent the ratio of the competitor 1 to competitor 2 in the initial ($t = 0$) mixture used for the competition and after 18 hours of competition (t), respectively. n represents the number of generations calculated as n=log2[$OD_{600}(t)/OD_{600}(0)$].

Estimation of biophysical effects of mutations

The biophysical effects of mutations on DfrB1 folding, dimerization, and tetramerization were taken from (*26*). Briefly, we used FoldX version 5.0 (*31*) with the MutateX workflow (*59*) on the biological assembly of DfrB1 (PDB: 2RK1) (*28*). Two residues from different subunits are identified as contacting if they are within a distance smaller or equal to the sum of van der Waals radii of the two closest atoms + 0.5 (Fig. 2B).

The biological assembly for the DfrB1 tetramer structure with DHF and NADPH (PDB: 2RK1) (*28*) was downloaded from the PDB (*60*). Since the reference structure contains an incomplete DHF molecule, the complete DHF molecule (ZINC000004228265) was downloaded from the Zinc20 database (*61*) and superimposed to the coordinates in the reference PDB structure to replace the incomplete molecule. The DHF and NADPH molecules were then parameterized using Obabel version 3.1.0 (*62*) and the Rosetta molfile_to_params.py script (*63*). The effects of selected pairs of mutations on dimerization, tetramerization, and substrate binding affinity were simulated with the Rosetta flexddG protocol (*64*) using the ref2015 energy function (*65*) as described previously (*66*). Briefly, for each pair of mutations 35 structures were generated and relaxed with 35 000 backrub steps following a previously described workflow (*67*). From the resulting distribution of structures, the median $\Delta\Delta G$ values for mutational effects were used for all subsequent analyses.

Protein expression, purification and stability assessment

To evaluate the effect of mutations on DfrB1 stability, we first purify a 6xHIS-tagged version of DfrB1 and of its mutants (expression from pQE32 plasmid) as is done in (*39*) with the following modifications. *E. coli* BL21 containing plasmid pRep4 was used for expression. Induction was done for 3 hours at 37°C instead of at 16°C and without zinc acetate. Cell pellets equivalent to 50 mL of post-induction cell culture were used for extraction in 2 mL of NEBExpress® E. coli Lysis Reagent. Elution buffer was finally exchanged for potassium phosphate 50 mM pH 8.0. The resulting purified DfrB1 proteins retain their N-terminal 6xHIS-tag, and were used fresh for stability assessment. All of the purification process was done in duplicate.

Protein complex stability was first measured using nano-differential scanning fluorimetry on a Prometheus NT.48 (NanoTemper, Germany). After purification, 10 µL of protein samples were

loaded in high sensitivity capillaries. Measurements were taken along a 20°C to 95°C temperature gradient, increasing at a rate of 1°C/min and at an intensity of 20%. The Tm was measured by finding the temperature at which the first derivative of the 350/330 nm fluorescence ratio is maximized using the PR.ThermControl v2.3.1 (fig. S6A). Each protein variant was tested in triplicate for both purification samples.

DfrB1 being a tetramer, in order to assess the impact of mutations on complex assembly, 6 mg of purified protein samples were also loaded on a NativePAGE™ Bis-Tris Mini Protein Gel, 4 to 16% along with 10 µL of NativeMark™ Unstained Protein Standard. Gel was run for 65 min at 300V on ice according to manufacturer's protocol. Gel was unstained in water for 24 hours (fig. S6B).

Bacterial growth curves

Individual growth curves were used in three different contexts: 1) To make sure there is no fitness difference when expressing *dfrB1* in comparison to *dfrB1[Syn4-9]*, 2) To evaluate the concentration of arabinose needed to express *dfrB1* in duplication such as growth recovery is around 50% in presence of TMP, and 3) To validate the effect on fitness of dominant negative mutations.

In all these cases, to measure fitness for individual WT or mutant DfrB1 in duplication with a WT copy or as a singleton, growth of bacteria was followed with serial OD_{600} measurements in a plate reader. Bacteria were first transformed with 1) pBAD-*dfrB1*-AMP or pBAD-*dfrB1[Syn4-9]*-AMP, 2) pBAD-*dfrB1*-AMP and pBAD-*dfrB1*-KAN, and 3) pBAD-*dfrB1*-AMP, pBAD-*dfrB1*(Q67C)-AMP, pBAD-*dfrB1*(S59Y)-AMP or pBAD-*dfrB1*(Q67C+S59Y)-AMP in combination or not with pBAD-*dfrB1*-KAN. Then, from an overnight preculture grown in LB AMP (also + KAN for the duplication) medium, cells were diluted 1:100 in fresh medium with the addition of different amounts of arabinose, with concentrations ranging from 0 to 0.2%. Following an 18-hour incubation at 37°C (250 rpm), cells were diluted again at OD_{600} = 0.01 in a final volume of 200 µl of fresh medium containing arabinose with the addition of TMP or DMSO. The 96-well plates were incubated at 37°C in a BioTek Synergy HTX Multi-mode Reader (Agilent, USA) or in an Infinite M Nano plate reader (Tecan, Switzerland) for 13 to 16 hours. OD_{600} measurements were taken every 15 min. The plate was agitated at 237 cpm (4 mm) at slow orbital speed (Agilent) or at 200 rpm (Tecan) in between measurements.

The area under each growth curve was calculated using Growthcurver (*68*). Growth recovery at each arabinose concentration with TMP relative to the WT in the absence of TMP was calculated using the following equation:

$$g = 100 * \left(\frac{AUC_{TMP\ ,\ induction}}{AUC_{noTMP\ ,\ induction}} \right)$$

where g refers to the recovered growth percentage and AUC is the area under the curve in the conditions specified by the subscript (with or without TMP and with a given concentration of arabinose).

We used ANOVAs to analyze the differences in growth recovery and tetramer stability across mutants. In each case, the ANOVA was followed by Tukey's honestly significant difference post hoc test computed using the agricolae R package (*69*).

Supplementary figures

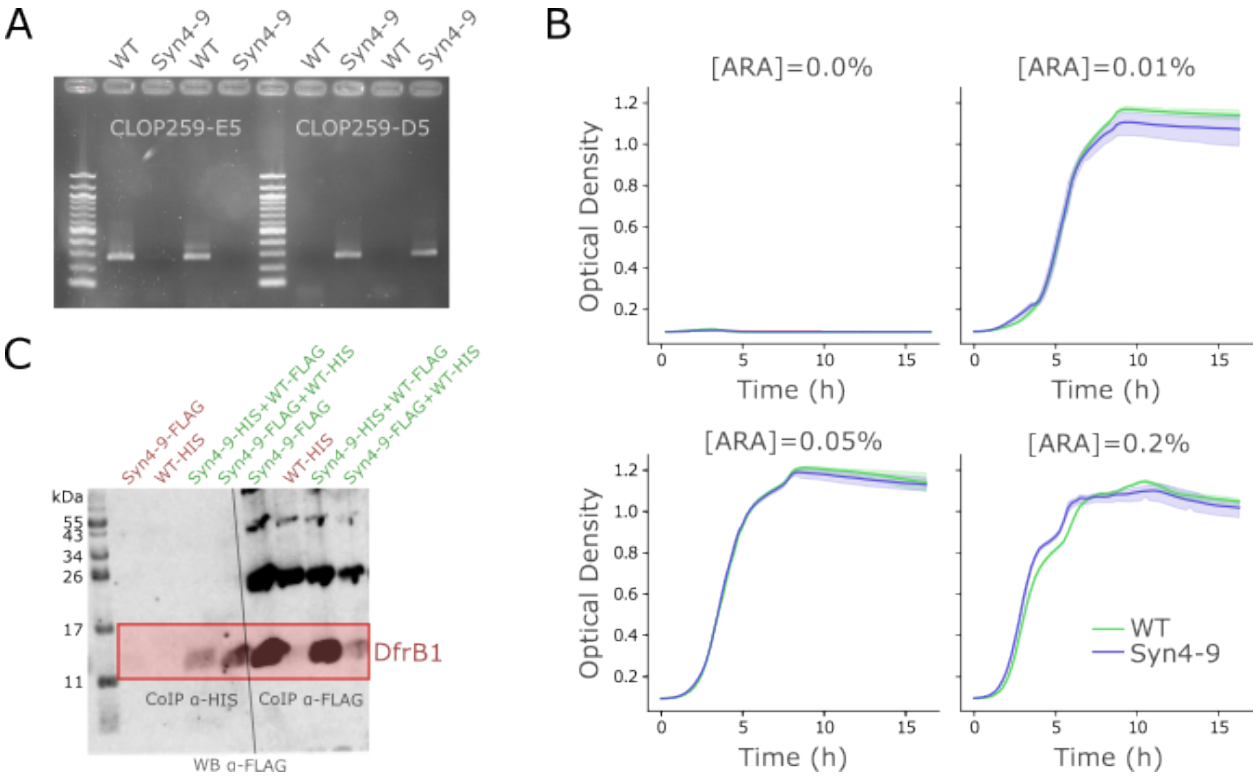

Figure S1: Synonymous mutations in the coding sequence of DfrB1 allows for specific amplification without impacting fitness nor protein complex assembly. A) PCR amplification to validate the differential amplification for the WT and the Syn4-9 variant and distinguish the two paralogs during sequencing. **B)** Growth curves showing there are no differences in the growth of WT and the Syn4-9 variant. **C)** Western blot after Co-IP to show the formation of DfrB1 heteromers when two gene copies are expressed, providing a system to study paralog interference. *E. coli* expressing the WT and the Syn4-9 mutant tagged with epitopes were used for these experiments. Samples were co-immunoprecipitated with α-HIS or α-FLAG as indicated in the two groups and then immunodetected for the presence of α-FLAG. Samples are labeled at the top in green if DfrB1 is expected to be observed and in red otherwise. Bands observed in the Co-IP with α-HIS indicate the presence of heteromers.

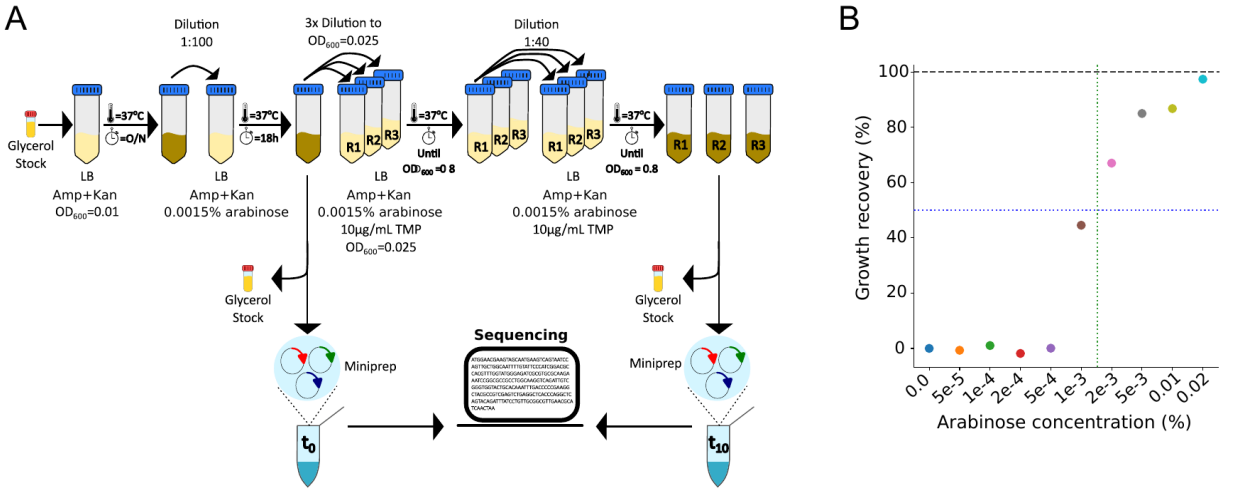

Figure S2: Selection of induction level (arabinose concentration) and design of the screening experiments using the mutant libraries. A) Overview of the bulk competition experiments. A preculture with the desired population is allowed to grow for 18 hours. It is then diluted to an optical density (OD) of 0.025 in three replicate populations, with an aliquot preserved for sequencing to estimate variant frequencies (t = 0). The three replicates are then allowed to grow at 37°C until they reach an OD = 0.8 (five generations). Each replicate is then diluted 1:40 and allowed to grow for five more generations. At t = 10 generations, an aliquot from each replicate is taken for sequencing. Selection coefficients are inferred based on the read counts for each mutation at t = 0 and t = 10, normalized by the read counts for the WT. **B)** The induction of DfrB1 expression with arabinose in media with TMP restores *E. coli* growth. Growth recovery is quantified as the area under the curve (AUC) at a given concentration of arabinose and 10 µg/mL TMP, normalized by the AUC without TMP.

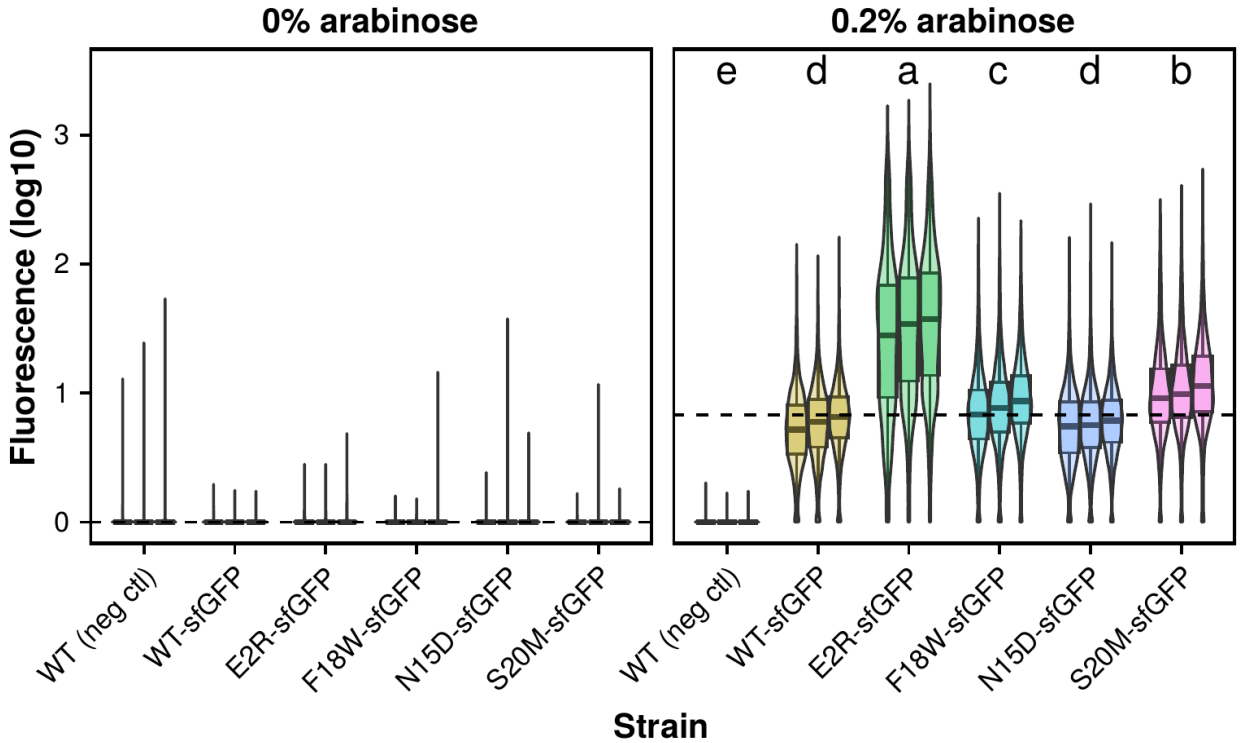

Figure S3: Some substitutions in the disordered region of DfrB1 that increase fitness do so by increasing protein abundance. Mutations from the disordered region with beneficial fitness effects (E2R, N15D, F18W, S20M) were selected and fused to sfGFP. Comparison of fluorescence values observed in flow cytometry for three replicate populations of ~5000 cells harboring each mutant. Dashed lines indicate the median of the WT-sfGFP fluorescence for the pool of cells from the three replicates. Labels at the top indicate groups of fluorescence shown to be significantly different using an analysis of variance (ANOVA) and Tukey's post hoc honestly significant difference test ($p < 0.05$).

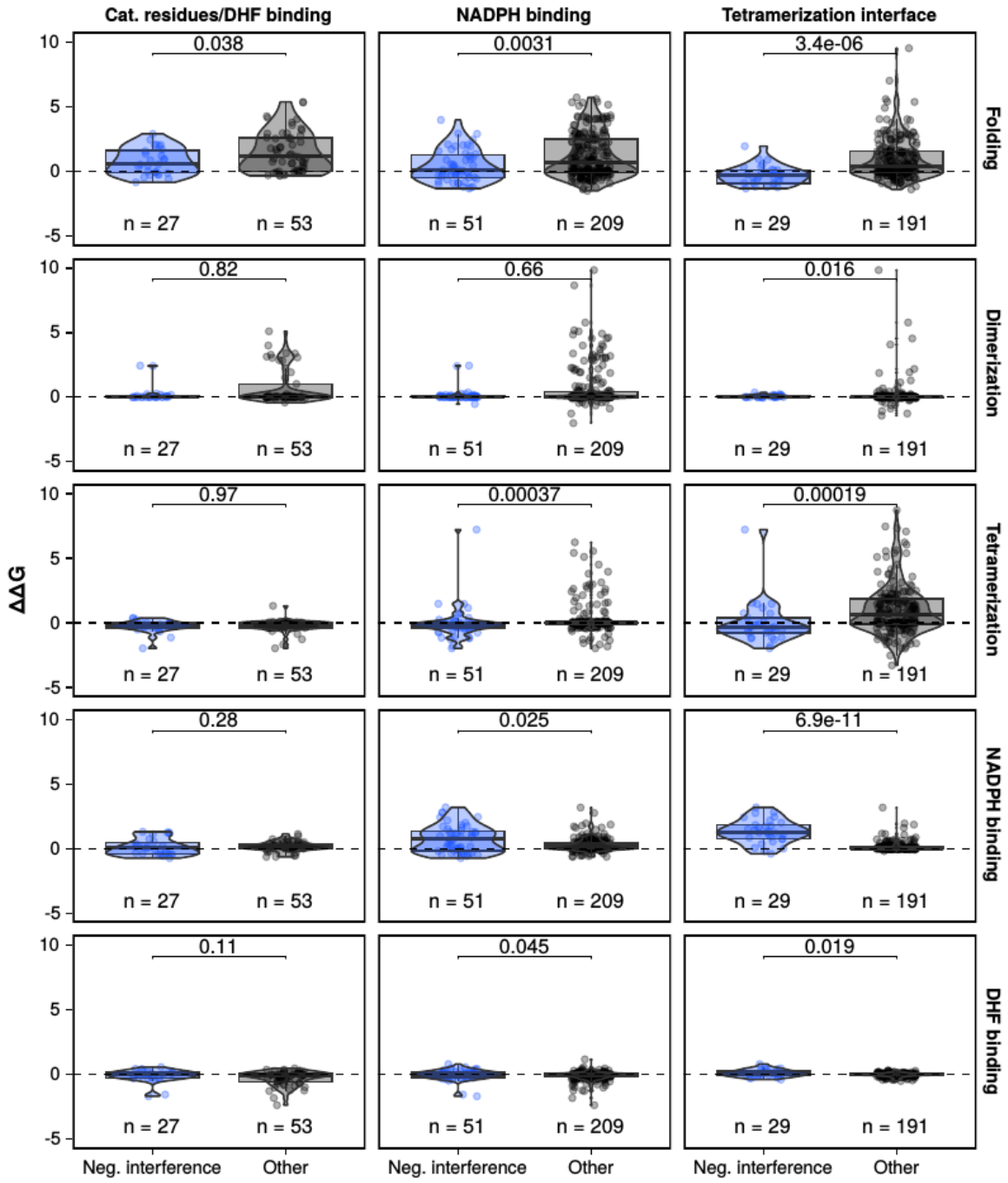

Figure S4: Negative interference mutations in specific protein regions have milder effects than other mutations from the same region. Same data as in Figure 3B, but filtered to show only mutations occurring in the regions with the highest enrichment of negative interference mutations in Figure 3C (columns). Rows indicate the different biophysical effects. p-values were calculated using Mann-Whitney U tests. As shown in Fig. 3C and (*26*), some positions are assigned to more than one region. Effects on NADPH of mutations in the tetramerization interface could be associated with positions that are directly in contact with NADPH or adjacent to residues that are in contact with NADPH.

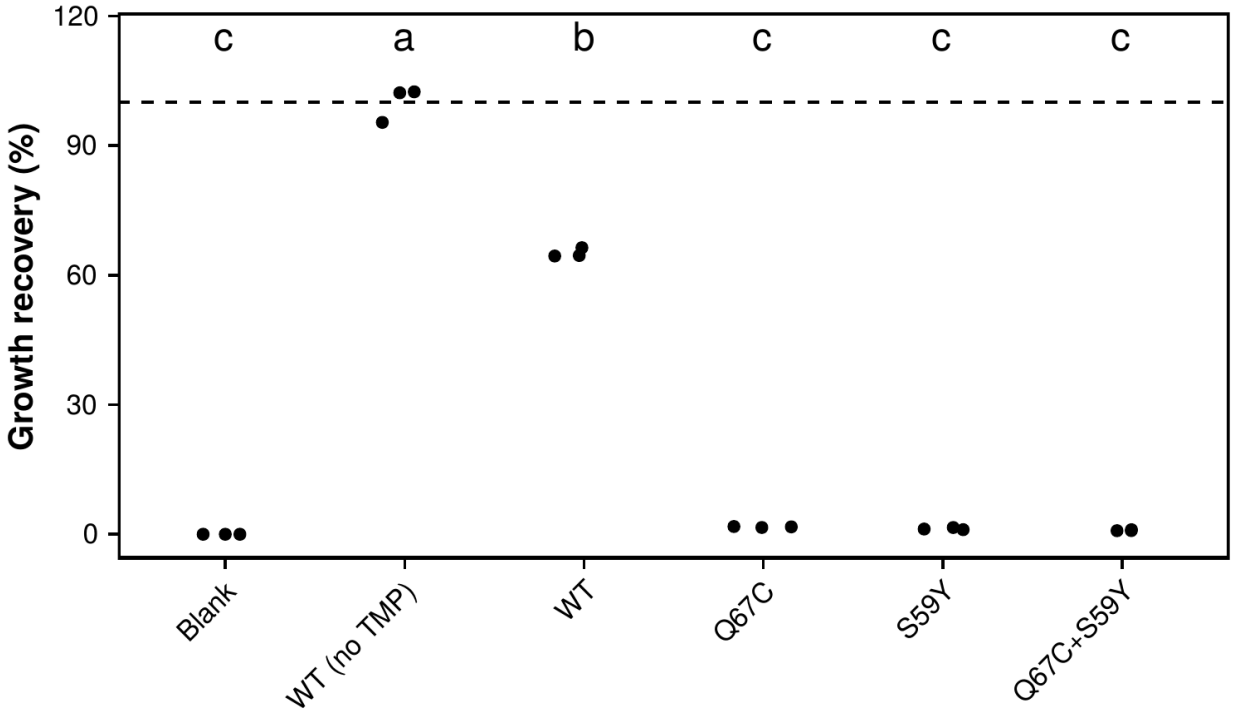

Figure S5: Selected negative interference mutations cannot restore growth as a singleton. Comparison of growth recovery at 0.0015 % arabinose and 10 µg/mL TMP for different genotypes relative to a WT reference growing in the absence of TMP. Labels at the top indicate groups of recovered growth shown to be significantly different using an analysis of variance (ANOVA) and Tukey's post hoc honestly significant difference test ($p < 0.05$).

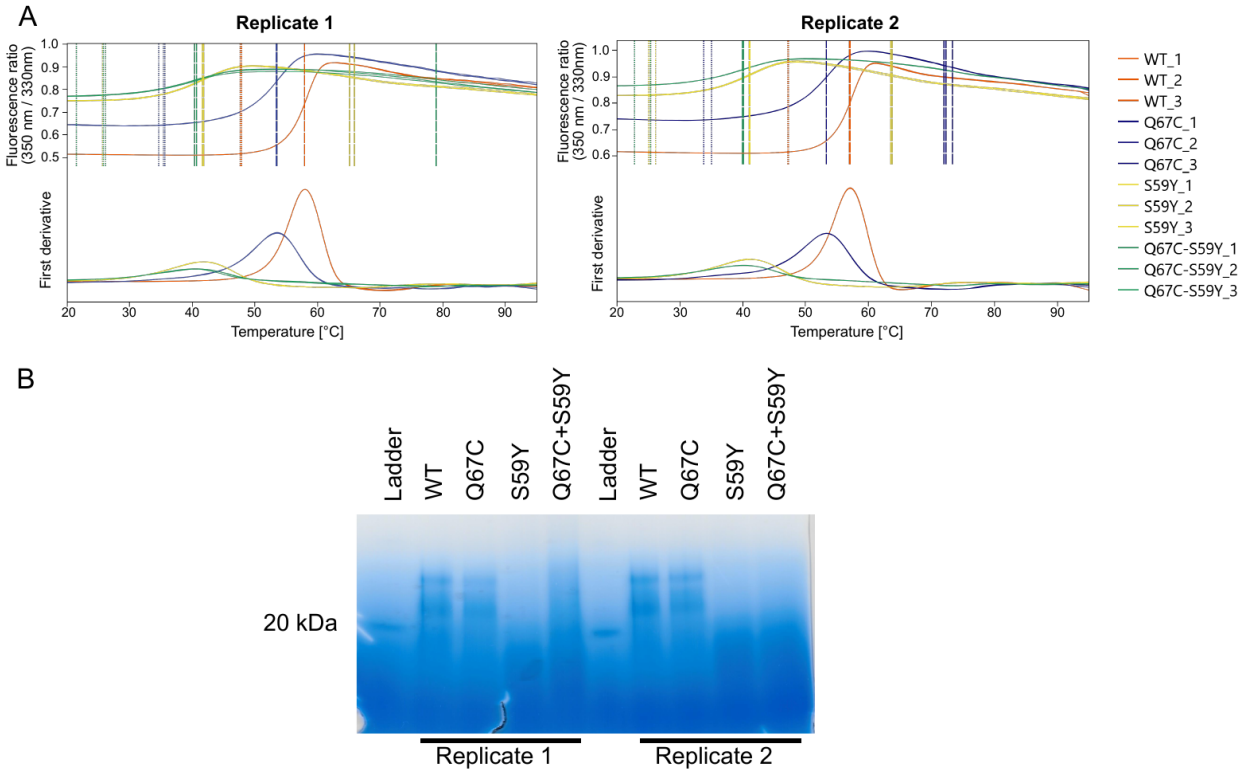

Figure S6: The oligomeric assembly of DfrB1 is disrupted by substitution S59Y but not by Q67C. A) Denaturation curves for different DfrB1 variants. Curves indicate three replicates for each of two separate protein purifications. *Top*: Ratio of emitted fluorescence at 350 nm vs 330 nm as a signal of exposed tryptophan residues. Vertical dashed lines indicate calculated inflection points in the denaturation curves. *Bottom*: First derivative of the ratio of emitted fluorescence **B)** Native gel electrophoresis of purified DfrB1 mutants. Labels indicate samples loaded in the corresponding wells.

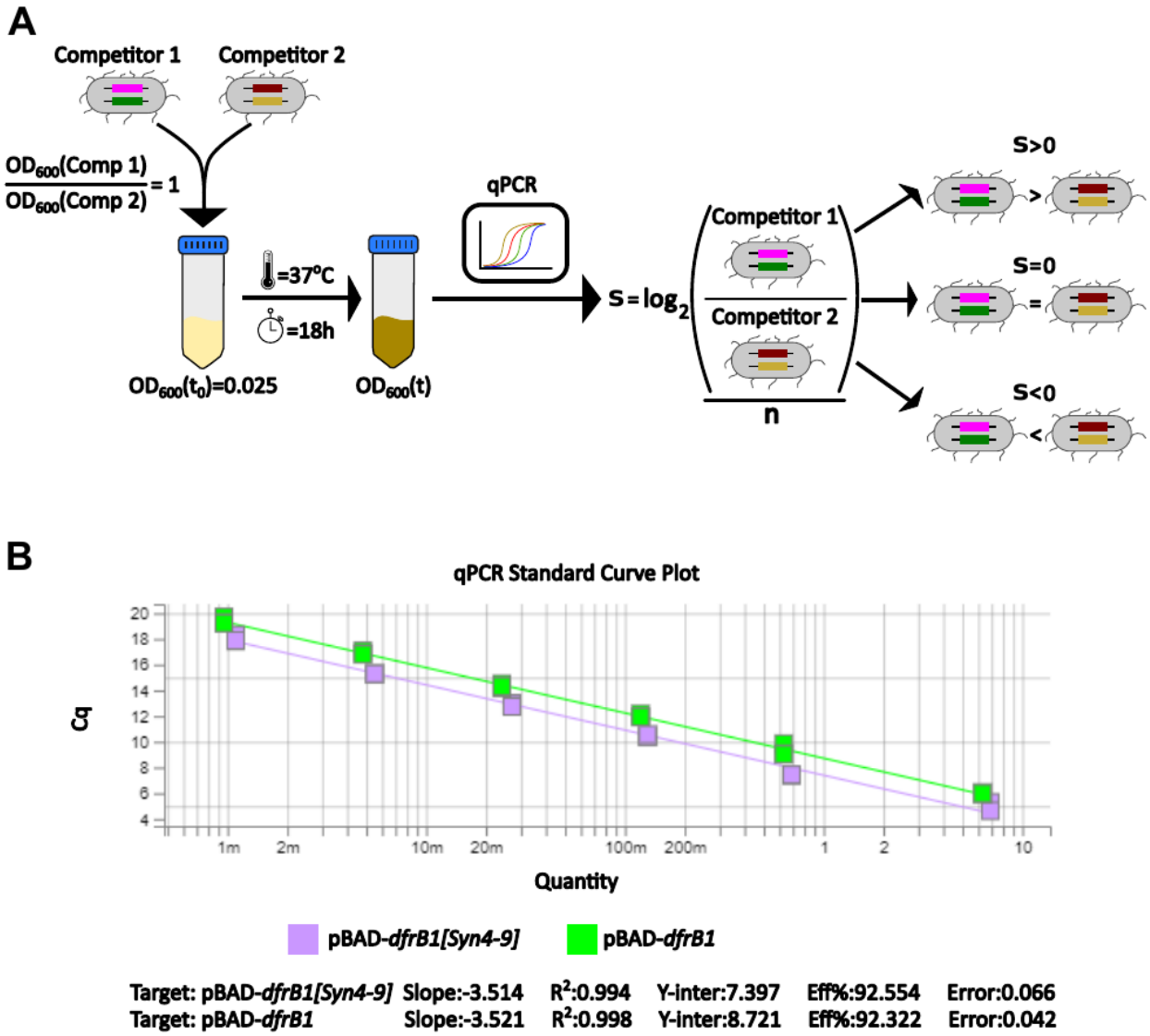

Figure S7: Direct competition experiment and measurement of relative abundance using qPCR. A) In each experiment the two competitors, one with the Syn4-9 mutation and the other without, were inoculated at the same proportion based on the individual OD_{600} and grown together for 18h. The number of generations ($n=\log_2[OD_{600}(t)/OD_{600}(t0)]$) was determined by measuring the OD_{600} of the grown culture. 1 µL was directly used to quantify selection coefficients (s) of one competitor relative to the other targeting the change in proportions for the WT and the Syn4-9 sequence by quantitative PCR. **B)** Standard curve showing the relationship between amplification cycles (Cq) values and log-transformed template concentrations. Both targets showed linearity and similar efficiency (~92%), confirming reliable quantification for all the samples tested across the experiment.

References:

1. K. J. Lipinski, J. C. Farslow, K. A. Fitzpatrick, M. Lynch, V. Katju, U. Bergthorsson, High spontaneous rate of gene duplication in Caenorhabditis elegans. *Curr. Biol.* **21**, 306–310 (2011).

2. M. Lynch, J. S. Conery, "The evolutionary demography of duplicate genes" in *Genome Evolution* (Springer Netherlands, Dordrecht, 2003), pp. 35–44.

3. M. A. Nowak, M. C. Boerlijst, J. Cooke, J. M. Smith, Evolution of genetic redundancy. *Nature* **388**, 167–171 (1997).

4. U. Bergthorsson, D. I. Andersson, J. R. Roth, Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 17004–17009 (2007).

5. L. Mihajlovic, B. R. Iyengar, F. Baier, I. Barbier, J. Iwaszkiewicz, V. Zoete, A. Wagner, Y. Schaerli, A direct experimental test of Ohno's hypothesis. *eLife*, doi: 10.7554/elife.97216.2 (2024).

6. S. Ohno, "Other Mechanisms for Achieving Gene Duplication" in *Evolution by Gene Duplication*, S. Ohno, Ed. (Springer Berlin Heidelberg, Berlin, Heidelberg, 1970), pp. 107–110.

7. A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, J. Postlethwait, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).

8. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).

9. A. Pérez-Bercoff, T. Makino, A. McLysaght, Duplicability of self-interacting human genes. *BMC Evol. Biol.* **10**, 160 (2010).

10. G. Diss, I. Gagnon-Arsenault, A.-M. Dion-Coté, H. Vignaud, D. I. Ascencio, C. M. Berger, C. R. Landry, Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science* **355**, 630–634 (2017).

11. P. P. Singh, S. Affeldt, I. Cascone, R. Selimoglu, J. Camonis, H. Isambert, On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep.* **2**, 1387–1398 (2012).

12. G. Malaguti, P. P. Singh, H. Isambert, On the retention of gene duplicates prone to dominant deleterious mutations. *Theor. Popul. Biol.* **93**, 38–51 (2014).

13. J. B. Pereira-Leal, E. D. Levy, C. Kamp, S. A. Teichmann, Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* **8**, R51 (2007).

14. A. Marchant, A. F. Cisneros, A. K. Dubé, I. Gagnon-Arsenault, D. Ascencio, H. Jain, S.

Aubé, C. Eberlein, D. Evans-Yamamoto, N. Yachie, C. R. Landry, The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *Elife* **8** (2019).

15. I. Herskowitz, Functional inactivation of genes by dominant negative mutations. *Nature* **329**, 219–222 (1987).

16. L. Gerasimavicius, B. J. Livesey, J. A. Marsh, Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* **13**, 3895 (2022).

17. H. Schweke, M. Pacesa, T. Levin, C. A. Goverde, P. Kumar, Y. Duhoo, L. J. Dornfeld, B. Dubreuil, S. Georgeon, S. Ovchinnikov, D. N. Woolfson, B. E. Correia, S. Dey, E. D. Levy, An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999–1010 (2024).

18. E. D. Levy, S. Teichmann, "Structural, evolutionary, and assembly principles of protein oligomerization" in *Progress in Molecular Biology and Translational Science*, J. Giraldo, F. Ciruela, Eds. (Academic Press, 2013)vol. 117, pp. 25–51.

19. J. L. Cherry, Expression level, evolutionary rate, and the cost of expression. *Genome Biol. Evol.* **2**, 757–769 (2010).

20. Z. Wu, X. Cai, X. Zhang, Y. Liu, G.-B. Tian, J.-R. Yang, X. Chen, Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nat Ecol Evol* **6**, 103–115 (2022).

21. H. Kacser, J. A. Burns, The molecular basis of dominance. *Genetics* **97**, 639–666 (1981).

22. D. Ascencio, G. Diss, I. Gagnon-Arsenault, A. K. Dubé, A. DeLuna, C. R. Landry, Expression attenuation as a mechanism of robustness against gene duplication. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2014345118 (2021).

23. R. P. Sugino, H. Innan, Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends Genet.* **22**, 642–644 (2006).

24. B. Papp, C. Pál, L. D. Hurst, Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–197 (2003).

25. L. T. Bergendahl, L. Gerasimavicius, J. Miles, L. Macdonald, J. N. Wells, J. P. I. Welburn, J. A. Marsh, The role of protein complexes in human genetic disease. *Protein Sci.* **28**, 1400–1411 (2019).

26. A. F. Cisneros, I. Gagnon-Arsenault, A. K. Dubé, P. C. Després, P. Kumar, K. Lafontaine, J. N. Pelletier, C. R. Landry, Epistasis between promoter activity and coding mutations shapes gene evolvability. *Sci Adv* **9**, eadd9109 (2023).

27. N. Narayana, D. A. Matthews, E. E. Howell, N.-H. Xuong, A plasmid-encoded

dihydrofolate reductase from trimethoprim-resistant bacteria has a novel D2-symmetric active site. *Nat. Struct. Mol. Biol.* **2**, 1018–1025 (1995).

28. J. M. Krahn, M. R. Jackson, E. F. DeRose, E. E. Howell, R. E. London, Crystal structure of a type II dihydrofolate reductase catalytic ternary complex. *Biochemistry* **46**, 14878–14888 (2007).

29. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

30. R. Guerois, J. E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).

31. J. Delgado, L. G. Radusky, D. Cianferoni, L. Serrano, FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).

32. C. Lemay-St-Denis, J. N. Pelletier, From a binding module to essential catalytic activity: how nature stumbled on a good thing. *Chem. Commun. (Camb.)* **59**, 12560–12572 (2023).

33. M. B. Strader, R. Derike Smiley, L. G. Stinnett, N. C. VerBerkmoes, E. E. Howell, Role of S65, Q67, I68, and Y69 Residues in Homotetrameric R67 Dihydrofolate Reductase. *Biochemistry* **40**, 11344–11352 (2001).

34. M. Lynch, J. S. Conery, The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* **3**, 35–44 (2003).

35. G. Abrusán, J. A. Marsh, Ligand Binding Site Structure Influences the Evolution of Protein Complex Function and Topology. *Cell Rep.* **22**, 3265–3276 (2018).

36. S. R. Wente, H. K. Schachman, Shared active sites in oligomeric enzymes: model studies with defective mutants of aspartate transcarbamoylase produced by site-directed mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 31–35 (1987).

37. S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).

38. C. R. Cortez-Romero, J. Lyu, A. S. Pillai, A. Langanowsky, J. W. Thornton, Symmetry facilitated the evolution of heterospecificity and high-order stoichiometry in vertebrate hemoglobin. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2414756122 (2025).

39. P. C. Després, A. K. Dubé, M.-È. Picard, J. Grenier, R. Shi, C. R. Landry, Compensatory mutations potentiate constructive neutral evolution by gene duplication. *Science* **385**, 770–775 (2024).

40. C. R. Baker, V. Hanson-Smith, A. D. Johnson, Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* **342**, 104–108 (2013).

41. J. T. Bridgham, J. E. Brown, A. Rodríguez-Marí, J. M. Catchen, J. W. Thornton, Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet.* **4**, e1000191 (2008).

42. A. DeLuna, K. Vetsigian, N. Shoresh, M. Hegreness, M. Colón-González, S. Chao, R. Kishony, Exposing the fitness contribution of duplicated genes. *Nat. Genet.* **40**, 676–681 (2008).

43. S. P. Otto, M. C. Whitlock, Eds., "Fixation probabilities and times" in *Encyclopedia of Life Sciences* (Wiley, 2006), pp. 1–4.

44. L. M. Guzman, D. Belin, M. J. Carson, J. Beckwith, Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).

45. D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison 3rd, H. O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

46. D. Bradley, A. Hogrebe, R. Dandage, A. K. Dubé, M. Leutert, U. Dionne, A. Chang, J. Villén, C. R. Landry, The fitness cost of spurious phosphorylation. *EMBO J.* **43**, 4720–4751 (2024).

47. S. Andrews, FastQC: a quality control tool for high throughput sequence data. (2010).

48. A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, J. D. Neufeld, PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).

49. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).

50. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

51. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

52. G. Korkmaz, M. Holm, T. Wiens, S. Sanyal, Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* **289**, 30334–30342 (2014).

53. E. B. Kramer, P. J. Farabaugh, The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* **13**, 87–96 (2007).

54. G. Moratorio, R. Henningsson, C. Barbezange, L. Carrau, A. V. Bordería, H. Blanc, S. Beaucourt, E. Z. Poirier, T. Vallet, J. Boussier, B. C. Mounce, M. Fontes, M. Vignuzzi, Attenuation of RNA viruses by redirecting their evolution in sequence space. *Nat Microbiol* **2**, 17088 (2017).

55. F. Mattenberger, V. Latorre, O. Tirosh, A. Stern, R. Geller, Globally defining the effects of mutations in a picornavirus capsid. *Elife* **10** (2021).

56. P. Carrasco, J. A. Daròs, P. Agudelo-Romero, S. F. Elena, A real-time RT-PCR assay for quantifying the fitness of tobacco etch virus in competition experiments. *J. Virol. Methods* **139**, 181–188 (2007).

57. F. Mattenberger, M. Vila-Nistal, R. Geller, Increased RNA virus population diversity improves adaptability. *Sci. Rep.* **11**, 6824 (2021).

58. K. K. Skarratt, S. J. Fuller, Quantitative real-time PCR eliminates false-positives in colony screening PCR. *J. Microbiol. Methods* **96**, 99–100 (2014).

59. M. Tiberti, T. Terkelsen, K. Degn, L. Beltrame, T. C. Cremers, I. da Piedade, M. Di Marco, E. Maiani, E. Papaleo, MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief. Bioinform.* **23**, bbac074 (2022).

60. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

61. J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, R. A. Sayle, ZINC20-A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).

62. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).

63. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

64. K. A. Barlow, S. Ó Conchúir, S. Thompson, P. Suresh, J. E. Lucas, M. Heinonen, T. Kortemme, Flex ddG: Rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. *J. Phys. Chem. B* **122**, 5389–5399 (2018).

65. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack Jr, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

66. F. D. Rouleau, A. K. Dubé, I. Gagnon-Arsenault, S. Dibyachintan, A. Pageau, P. C. Després, P. Lagüe, C. R. Landry, Deep mutational scanning of Pneumocystis jirovecii dihydrofolate reductase reveals allosteric mechanism of resistance to an antifolate. *PLoS Genet.* **20**, e1011252 (2024).

67. R. C. Eccleston, E. Manko, S. Campino, T. G. Clark, N. Furnham, A computational method for predicting the most likely evolutionary trajectories in the stepwise accumulation of resistance mutations. *Elife* **12**, e84756 (2023).

68. K. Sprouffske, A. Wagner, Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics* **17**, 172 (2016).

69. Felipe de Mendiburu, Muhammad Yaseen, *Agricolae: Statistical Procedures for Agricultural Research* (2020; https://myaseen208.com/agricolae/authors.html).