Effect of spatial distance between mutations on the fitness of mutated E. coli bacteria.

Guterman Julie, Janati Idrissi Ali, Mas Pablo

Abstract

Resistance to rifampicin in bacteria have been largely characterized over the past decades to study bacterial evolution and the emergence of antibiotic resistance. Some single amino acid mutations have already been shown to conf er resistance by damaging the binding pocket of the rpoB gene, - coding for bacterial ß-subunit of RNA polymerase - on which rifampicin binds and thus inhibits elongation of DNA. Nevertheless, the combination of some of these mutations confer a lowered fitness compared to the sum of the separate mutations. Here we used the library of mutations applied to rpoB generated by Cas9-mediated recombineering by A. Choudhury & al. to study the effect of distance between mutations on their epistasy. As expected, most of the combinations occur between two close amino acids, since they are both in rifampicin binding pocket. Here we show that the amplitude of epistasy increases when residues are closer and that close epistatic residues are more likely to have a negative epistasy. Besides, if we focus on one single residue, it does not have a predefined effect on epistasy : it can induce either more or less resistant phenotypes depending on the other mutation it is combined with.

Keywords: CRISPR-Cas9; deep mutational scanning; recombineering; fitness; mutations, distance

Motivations

Understanding of antibiotic resistance is of prime importance to design more efficient treatments against bacterial infection diseases. Bacteria can acquire resistance to antibiotics by many different ways: mutation or overexpression of the antibiotic target, modification of the antibiotic molecule, reduction of membrane permeability, pumping out of the antibiotic, creation of biofilms… This new capacity is then vertically or horizontally transferred to other bacteria and the genotype coding for this phenotype is selected. Here we focus on mutations of the rifampicin target by inducing random mutations on its gene. Rifampicin is known to bind to the rpoB, the beta subunit of RNA polymerase, to inhibit elongation of DNA and thus blocks protein production and causes the death of the cell. Bacteria can acquire resistance to rifampicin by mutating one or more amino acids crucial for rifampicin fixation on rpoB. This mechanism is not linear: when several mutations occur, their effect on bacteria fitness (which compares the number of cells after antibiotic treatment to the number of cells before antibiotic treatment thanks to the logarithm of their ratio) cannot yet be computed knowing the fitness induced by the individual mutations. Epistasis is the occurrence of several mutations having various phenotypic results depending on the genetic context in which they occur. Epistatic interactions between mutations are significant because they affect evolution and can show how biological functions are built from their genetic composition and give us a better understanding of bacterial resistance. Especially, epistasis can turn beneficial mutations into deleterious on different genetic backgrounds or on the contrary strongly increase the resistance conferred. Nevertheless, understanding relationships between genetic mutations is not easy, notably due to the three-dimensionality of proteins once translated from their one-dimension gene.

In order to have a better understanding of the phenomenon, we studied the impact of the spatial distance between two amino acids on their epistasis to see if only geographically close mutations can interact or if more complex interactions are at stake.

State of the art

The Synthetic Theory of Evolution (or Modern Synthesis) emerged between the 1930s and 1950s and consists of a reformulation of Darwinian evolutionism in the terms of Mendelian genetics. To date, it provides the most complete and satisfactory description of how populations of living organisms change over time, in other words, natural evolution. One of the pillars of the Modern Synthesis is the idea that all genetic mutation is "random" or the result of "chance" in relation to the adaptation of organisms and their species However, the notion of chance invoked in this framework has rarely been subjected to a thorough epistemological analysis. For the last twenty years, a debate has been going on between supporters and detractors of the Modern Synthesis concerning genetic mutations: some support the "random" or "at random" character of any mutation in relation to adaptation, others question it and suggest the possibility of "directed" or partly "Lamarckian" mutations More specifically, the discovery of a number of molecular mechanisms of mutation, known as "mutators", which produce an increase in the rate of mutation in response to changes in the environment, has led a number of biologists, historians and philosophers of biology to challenge the Modern Synthesis idea that all genetic mutation is the result of chance versus adaptation. It is therefore interesting to look at the different parameters that could stimulate the appearance of mutations in proteins, elements that control all cellular processes.

The study of the effects of a few mutations on protein stability and behavior helped lay the groundwork for protein research in the past (Fersht *et al*, 1985). Since then, proteins have been the target of numerous mutagenesis methods that aim to define their roles as key players in the cell's machinery. Despite this, the ability to quantify the effects of protein mutations has been limited to a small number of mutations. Recent developments known as "deep mutational scanning" (Fowler & Fiels, 2014) have made large-scale mutagenesis studies possible. Deep mutational scanning addresses the issue of being unable to predict the most informative mutations in a protein to study. Changes to amino acids far from binding or active sites may have a big impact on a protein's thermodynamic stability or enzymatic activity. Highly conservative mutations may be neutral, deleterious, or hyperactivating, with unpredictable consequences. Multiple mutations combined can result in significant increases or decreases in activity that are unexpected. Deep mutational scanning can expose the unexpected by allowing the effect of mutations to be studied in an unbiased manner. It can also help in cases that are otherwise difficult to solve.

A new method using deep mutational scanning has been developed recently (Choudhury *et al*, 2020) to track mutation frequency. It revealed biases in the position and the number of introduced mutations in rpoB, the ß-subunit of RNA polymerase, when placed in the presence of rifampicin. Our work is a continuation of this research, trying to see whether or not there is an influence of the spatial distance between two mutations on the fitness. Even though biologists came up with many definitions of fitness, there is still a gray area regarding this idea. To sum up, fitness involves the ability of organisms to survive and reproduce in the environment in which they are. Therefore, it is interesting to dig into the subject of fitness as there are many applications, in particular to understand and address the antibiotic resistance issue.

Methods

1. Dataset

The error-prone PCR library was obtained via A. Choudhury. It was constructed using the CREPE protocol using a wildtype E. coli subjected to three different concentrations of rifampicin (10, 50 and 100 µg/mL).

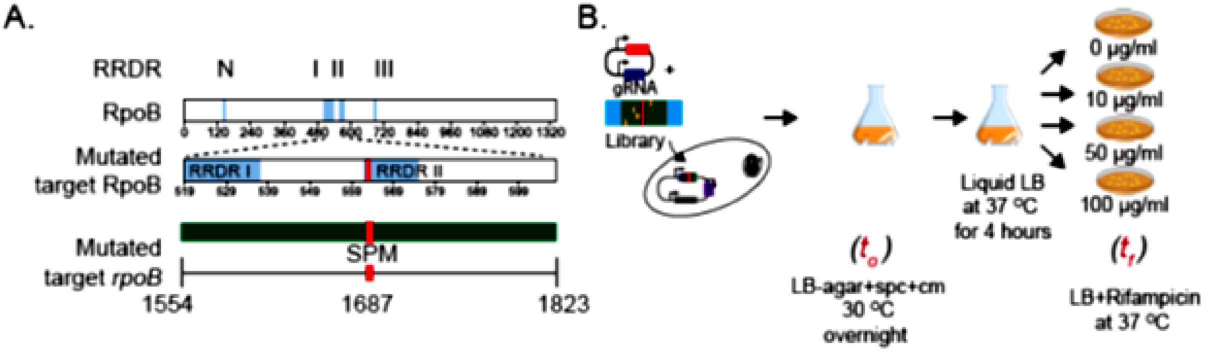


Figure 1. *CREPE-mediated mutagenesis of rpoB for resistance to rifampicin.*

- ***A.*** *Four distinct regions within rpoB (N, I, II, and III) are rifampicin resistance-determining regions (RRDRs). They used CREPE to make a library covering a 270-bp-long*
- ***B.*** *Experimental setup for studying resistance to 3 different concentrations of rifampicin*

2. Data processing

2.1. Fitness calculations

Fitness calculations for resistance of rifampicin with each replicate were estimated using the two time point enrichment score calculation algorithm. The fitness for each variant was estimated as follows:

$$fitness, f = \log\left(\frac{C_{i,post-sel}}{C_{wt,post-sel}}\right) - \log\left(\frac{C_{i,pre-sel}}{C_{wt,pre-sel}}\right)$$

where $\boldsymbol{C_i}$ is the counts associated with a variant post- and pre- selection and $\boldsymbol{C_{wt}}$ is the count associated with a wildtype reference with no mutations.

Fitness is calculated for all three concentrations of rifampicin.

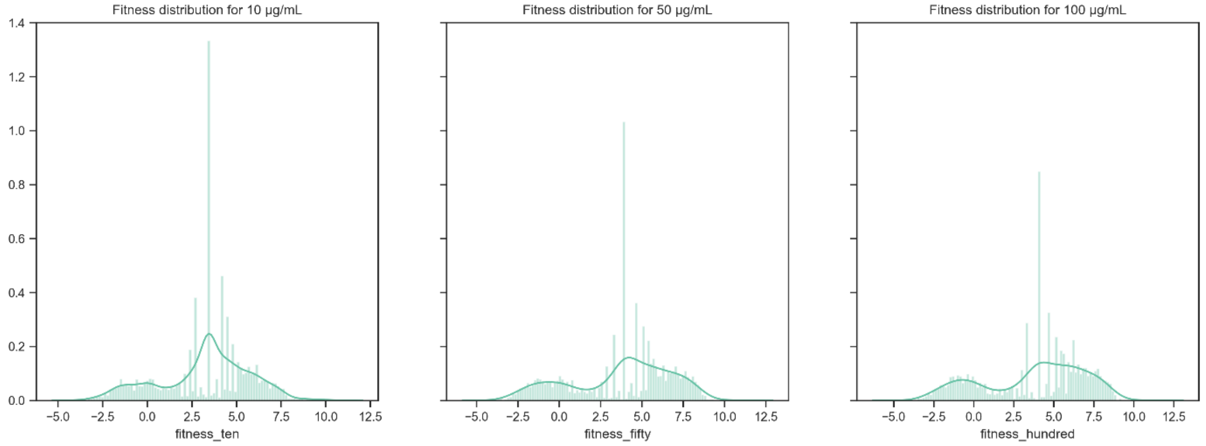


Figure 2. Fitness distribution for all rifampicin concentrations (raw)

2.2. Filtering

The dataset we get is noisy due to errors introduced during PCR and sequencing. Usually, the counts associated with PCR errors and sequencing errors are very low, what we do is identify thresholds below which we can eliminate the data. Since rpoB is an essential gene, stop codons cannot occur. Therefore, if any rows have stop codons that are most likely errors. Such stop codon mutations in our dataset are denoted by "*", they are removed. The number of rows in our dataset goes from 187 222 to 791. We obtain a new fitness distribution with two different populations, one resistant to rifampicin (positive fitness) and one not resistant (negative fitness).

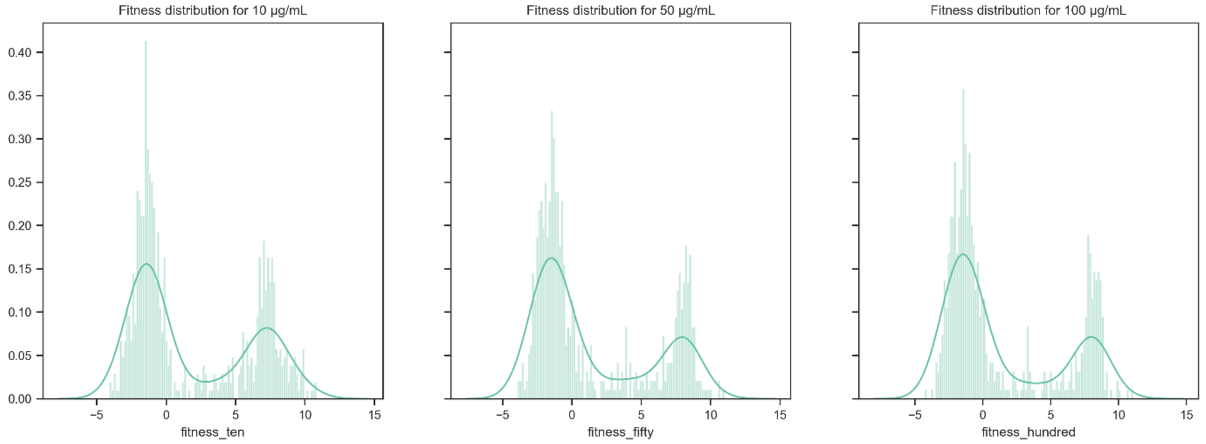


Figure 3. Fitness distribution for all rifampicin concentrations (filtered)

3. Distance and calculations

3.1. Filtering double mutants

We are only interested in bacteria with two mutations to look at the distance between these two mutations. We first isolated the double mutants for which we had fitness scores for the single mutants. We then calculated the sum of fitness scores for all three concentrations ($fitness_{1+2}$).

$$fitness_{1+2} = fitness_1 + fitness_2$$

We create two datasets, one containing only the bacteria with 1 mutation (df_1mutation) and the other containing the bacteria with two mutations existing in df_1mutation (df_2mutations).

3.2. RNA Polymerase rpoB structure

We use the 5UAQ Escherichia coli RNA polymerase RpoB as our reference for the position of amino acids. The .pdb file corresponding to the structure is imported via the BioPython python library. We can visualize the structure with the NGL View python library.

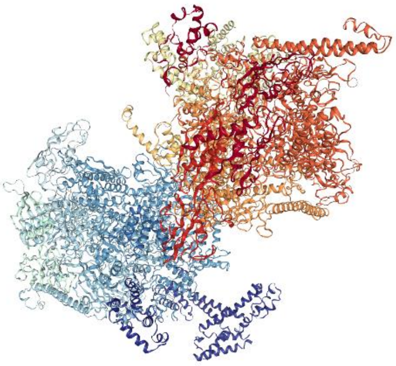


Figure 4. *3D Ribbon structure of 5UAQ E. coli RNA polymerase*

3.3. Parsing the structure file

Structure objects are organized in a specific hierarchy of objects. We just focused on the core elements which are model/chain/residue/atom. The 5UAQ Escherichia coli RNA polymerase contains only 1 model but 12 chains (A to L). The chain that interests us is the C chain which corresponds to RpoB. Within this chain there are 1339 residues each corresponding to an amino acid. Each residue is composed of atoms, and in particular of a carbon in alpha (CA). It is the position of this alpha carbon that we will use as reference for the position of each amino acid. We can extract the position of each alpha carbon, we obtain 3 coordinates (x, y and z).

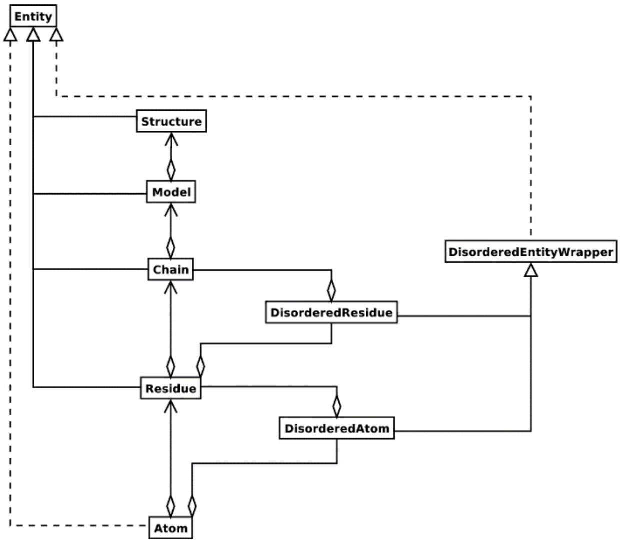
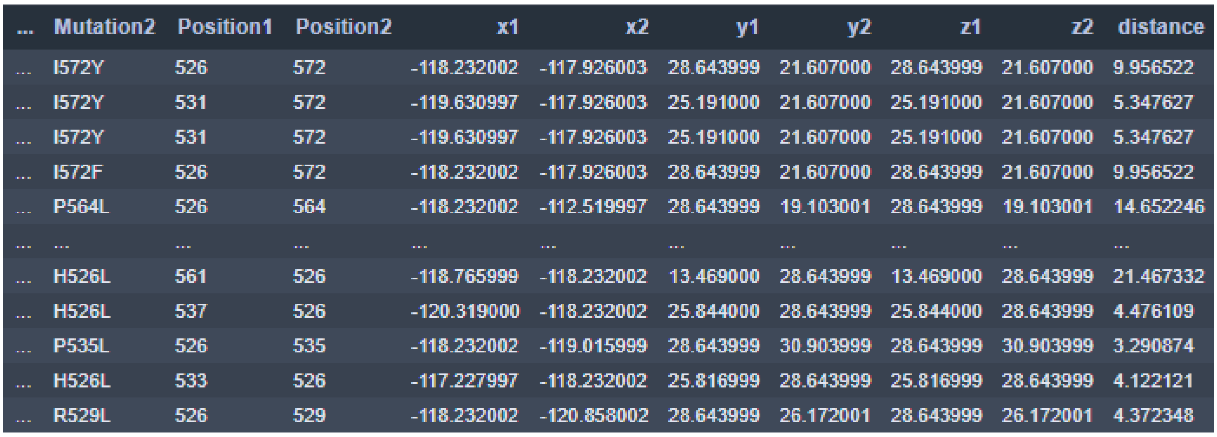


Figure 5. *Structure design of the .pdb file*

3.4. Distance calculation

We create 6 columns corresponding to the x, y and z coordinates of each of the two mutations. The distance between the two mutations is then calculated using the formula:

$$distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$



...	Mutation2	Position1	Position2	x1	x2	y1	y2	z1	z2	distance
...	I572Y	526	572	-118.232002	-117.926003	28.643999	21.607000	28.643999	21.607000	9.956522
...	I572Y	531	572	-119.630997	-117.926003	25.191000	21.607000	25.191000	21.607000	5.347627
...	I572Y	531	572	-119.630997	-117.926003	25.191000	21.607000	25.191000	21.607000	5.347627
...	I572F	526	572	-118.232002	-117.926003	28.643999	21.607000	28.643999	21.607000	9.956522
...	P564L	526	564	-118.232002	-112.519997	28.643999	19.103001	28.643999	19.103001	14.652246
...	...	...	...	...	...	...	...	...	...	...
...	H526L	561	526	-118.765999	-118.232002	13.469000	28.643999	13.469000	28.643999	21.467332
...	H526L	537	526	-120.319000	-118.232002	25.844000	28.643999	25.844000	28.643999	4.476109
...	P535L	526	535	-118.232002	-119.015999	28.643999	30.903999	28.643999	30.903999	3.290874
...	H526L	533	526	-117.227997	-118.232002	25.816999	28.643999	25.816999	28.643999	4.122121
...	R529L	526	529	-118.232002	-120.858002	28.643999	26.172001	28.643999	26.172001	4.372348

Figure 6. Dataframe illustrating the coordinates for two mutations

3.5. Fitness difference and fitness ratio

The fitness difference is defined as:

$$difference = fitness_{12} - fitness_{1+2}$$

where $fitness_{12}$ is the calculated fitness using the equation in paragraph 2.1. and $\boldsymbol{fitness_{1+2}}$ is the sum of fitness of mutation 1 and mutation 2.

The fitness ratio is defined as:

$$ratio = \frac{fitness_{12}}{fitness_{1+2}}$$

Fitness difference and fitness ratio are calculated for all three concentrations, but we will mainly focus on the 50 µg/mL one for our analysis.

Results

Among all the mutations database we had, we chose the mutations that were both in colonies with one single mutation and in colonies with two mutations including the mutation of interest. Three regions of the rpoB gene are commonly identified as rifampicin resistance-determining regions (RRDRs), with 26 residus in RRDR I, 9 residus RRDR II and one single residue in RRDR III. In the database, mutations were targeted on a 90 amino acid region containing half of RRDRI and full RRDRII. The number of mutations is higher close to these regions than outside the RRDRs (Figure 7). Since mutants are not

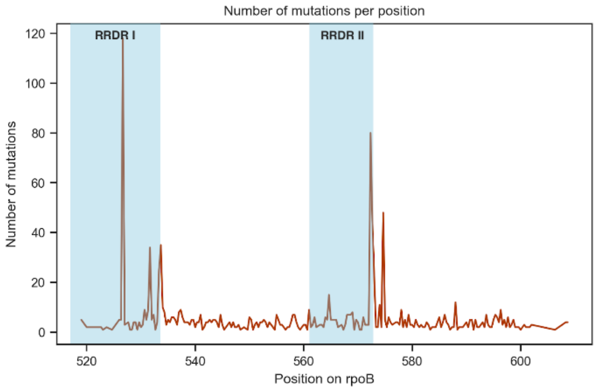


Figure 7. Number of mutations per position (residue on the gene of rpoB). Blue area represent the two targeted resistance-determining regions.

homogeneously distributed along the gene, the distance between residues in two-mutations clones is not uniforme either (Figure 8). Most of the double mutations that occur in one gene are 8 to 14 Å distant.

For the major part of the double mutants, the ratio is close to one, the epistasy is neither positive nor negative (Figure 9). Nevertheless, some combinations of mutations do have significant epistasy effect (Figure 11). In particular L533H+P535Q have a ratio of -7 and L533H+A543T has a ratio of 14.

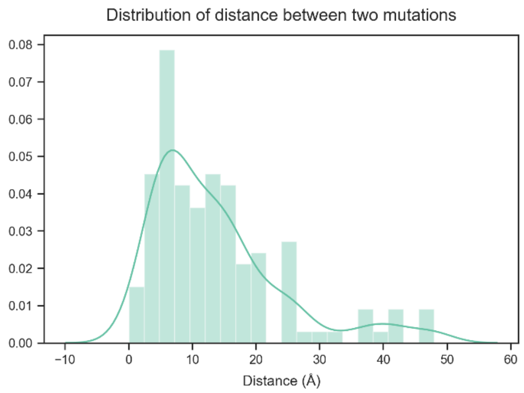


Figure 8. Distribution of distance between two mutations of double mutants. The distance is the spatial distance on the folded rpoB.

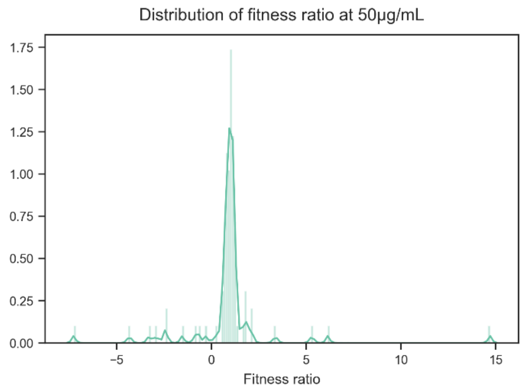


Figure 9. Distribution of fitness ratio at 50µg/mL of rifampicin.

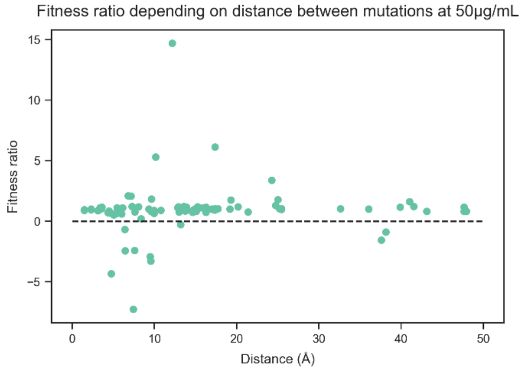


Figure 11. Ratio of the actual fitness over the sum of fitness of the corresponding individual mutations. The dash line corresponds to ratio of 1, no positive nor negative epistasy. The distance is the spatial distance on the folded rpoB.

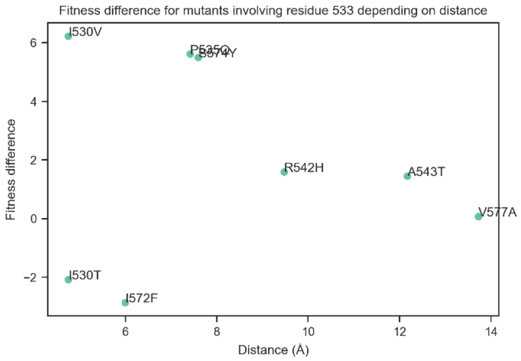


Figure 10. Fitness difference (actual fitness of the double mutant - sum of the fitness of the corresponding individual mutants) for mutants involving the 533th residue depending on the distance between 533th residue and the second residue in the folded rpoB. The residue of the second mutation is labeled on the figure.

Mutations on the residue 533 also happen to be one of the most frequent mutations of the selected double mutants. We can therefore select all the mutations involving residue L533 and see if it always has the same effect on fitness. First of all, it seems that close mutations tend to have higher effects of epistasy than the distant ones (Figure 10). We can see that for one given residue, depending on how it is changed and on which other residue is mutated, the resulting survival rate can improve or decrease compared to the same individual mutations, and that the amplitude of that effect seems to decrease with the distance between the residues.

Interpretation and conclusion

If most of the mutations studied here seem inert to combination with another mutation (fitness ratio close to one), some double mutants do show a different survival rate compared to their individual rates. We showed that epistasy is more frequent when amino acids are close than when they are far away. But this result was quite predictable with the data used given the fact that most mutations are concentrated in small geographic regions (Figure 7). Consequently, we cannot really drive conclusions about the effect of distance on the combination of mutations since these mutations are not uniformly distributed along rpoB gene. Nevertheless, we noticed that close mutations are more likely to have a negative ratio than a positive one. Indeed, one mutation on the catalytic core of rpoB can prevent rifampicin to bind and therefore enhance bacteria fitness. But if two close residues of this essential part of the gene are mutated, it might not give a much bigger advantage on hiding rpoB to rifampicin. In fact, if the catalytic core of rpoB mutates too much, the ß-subunit of RNA polymerase could be less effective/loose its activity. This effect is statistically more probable on close residues since their mutations might be redundant for the binding of rifampicin and therefore it is less likely to give a bigger advantage than one single mutation and therefore compensate the loss of activity of the rpoB protein.

Furthermore, the evaluation of epistasy is not fully objective: depending on the quantity observed, the results might change a lot. We characterized epistasy with the ratio and the difference between the actual fitness of the double mutants and the sum of individual fitness of single mutants and obtained two figures very different with different interpretations possible on the same dataset (Figure 12). Studying the ratio makes epistasy more flat but also more sensitive to the sum of the individual fitness if they are of opposite signs. Studying the difference amplifies less the initial mutations with different signs but does not normalize the global fitness. A more legitimate approach would separate cases where the two individual mutations are positive/negative/positive and negative. Here, the dataset is not large enough to have significant results with this approach either.

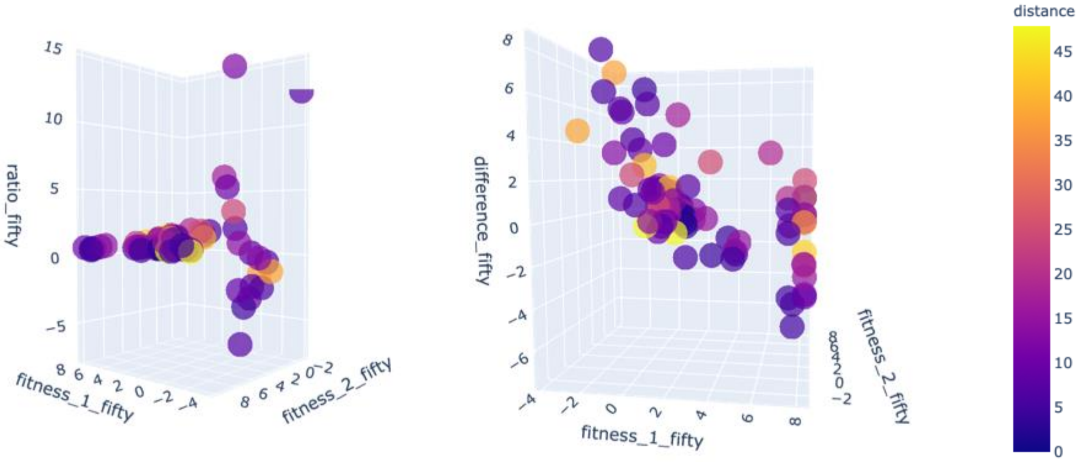


Figure 12. *Comparison of two ways of assessing epistasy in bacteria with two mutations. Abscissa are the fitnesses of the corresponding mutations in single mutants. Left ordinate is the fitness ratio and right ordinate is the fitness difference. The heatmap corresponds to the spatial distance between the residues in the folded rpoB.*

Perspectives

In this study, we used a library of mutants generated by CRISPR/Cas9- mediated genomic error-prone editing (CREPE) technology to link distance between two amino acids to the phenotype acquired when they are both mutated. But since mutations were mainly located in specific areas of the genome, we have not been able to truly study the impact of the distance between two amino acids on the phenotype generated by their mutations. Besides, the area covered is 90 amino acid long, which did not enable us to have a wide playfield to test the influence of distance. The length of this target is particularly determined by the length of the cassette tolerated by Cas9. To increase the genomic region targeted we could do several generations of mutant libraries by using overlapping cassettes that will in the end cover the whole rpoB gene (Figure 13). With this new dataset, we would have enough data to study the epistasy of positive/positive mutations, positive/negative mutations and negative/negative mutations and relate their phenotype to the location of mutations. We could also have more distant mutations to study and try to find new RRDRs.

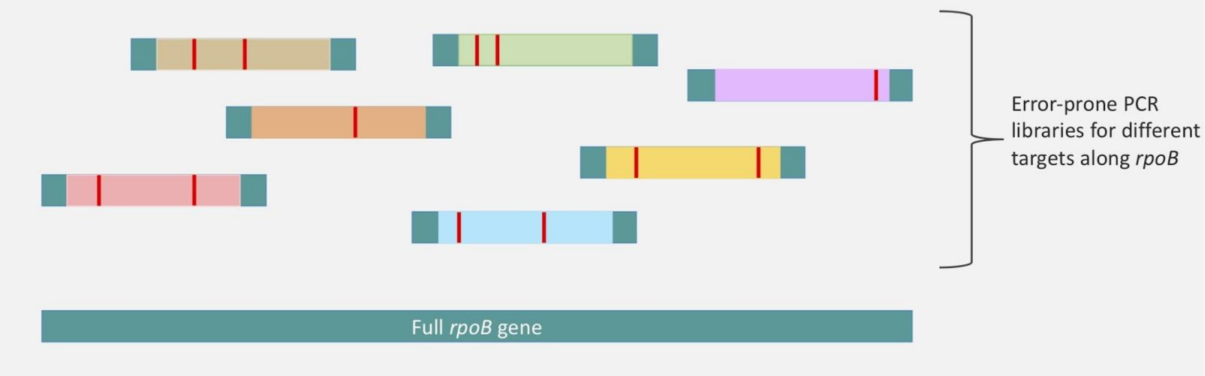

Error-prone PCR libraries for different targets along *rpoB*
Full *rpoB* gene

Figure 13. *Generation of overlapping mutants libraries by designing overlapping CRISPR Cas9 cassettes. The red lines figure random mutations on the error-prone PCR fragment.*

Data availability

The code produced in this study is available in the following Github repository : https://github.com/pablo-mas/sbn-project

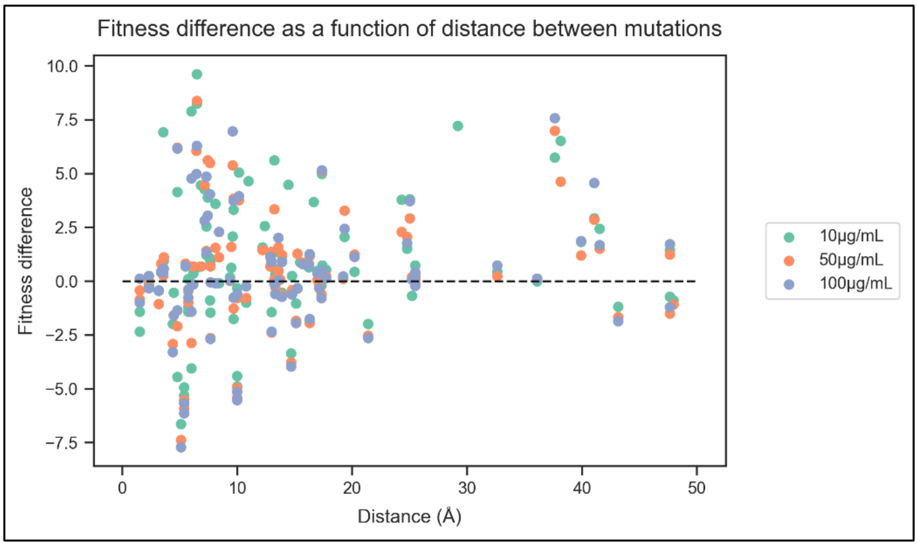
The raw dataset used is available witht his link (expiring 6th of May 2021): https://we.tl/t-3D0liM3gFH or you can contact us.

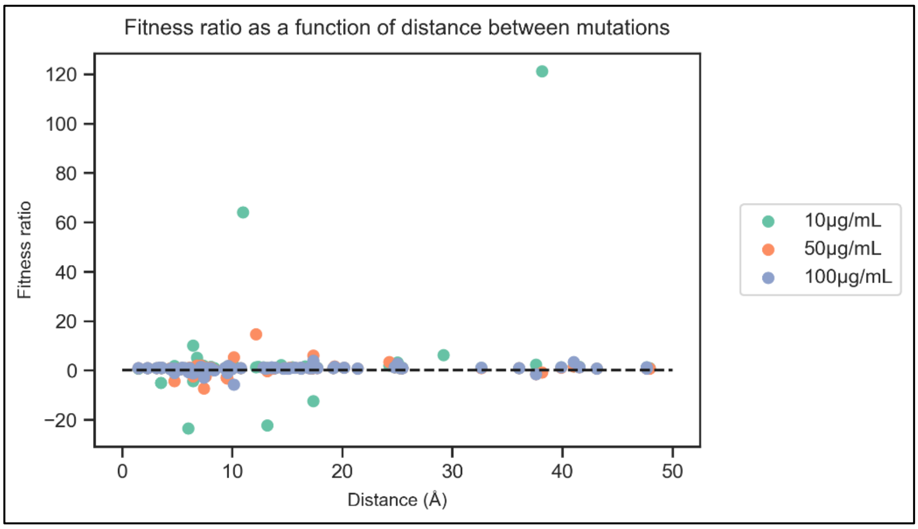
You can contact us to get further information or code/data related questions at mailto:pablo.mas@espci.fr

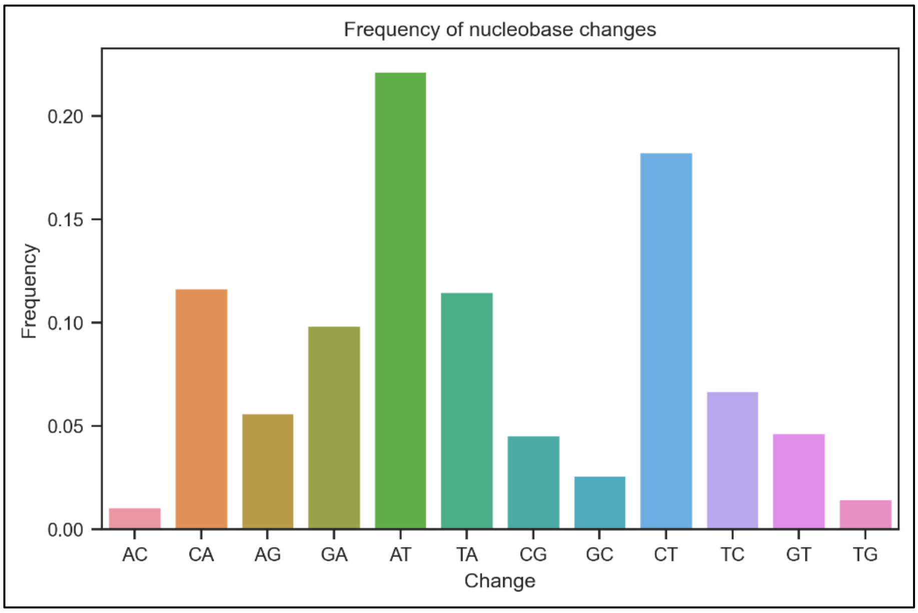
Acknowledgements

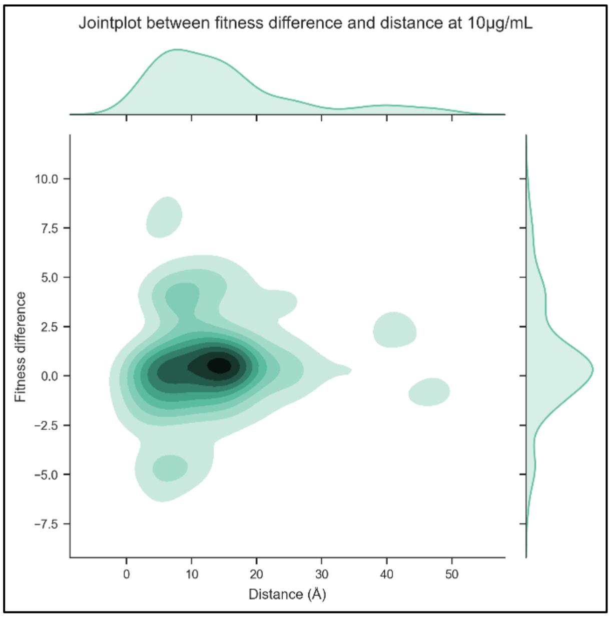
We would like to acknowledge all the teaching staff of the System Biology and Neurobiology course and in particular Alaksh Choudhury for having provided us with the data on which we worked, but especially for his supervision and his help in the realization of this project.

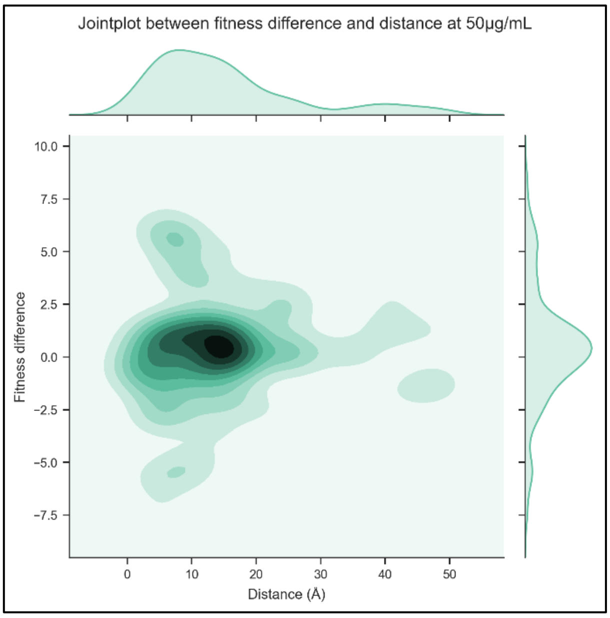
Annex

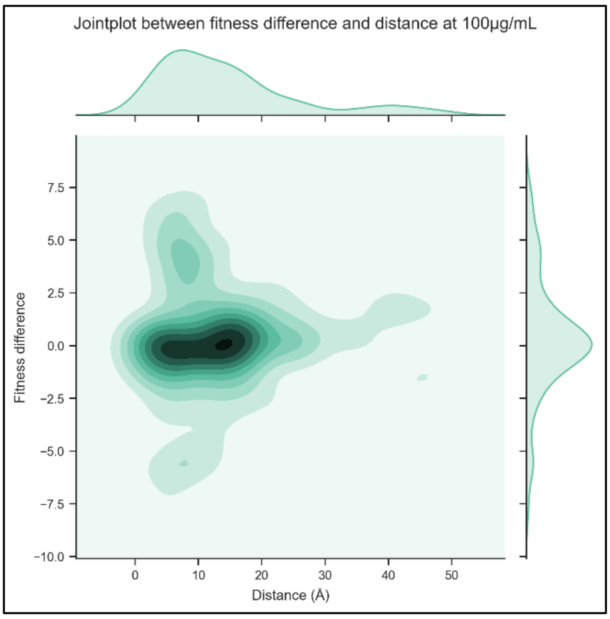

Fitness difference as a function of distance between mutations
Fitness difference
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
0
10
20
30
40
50
Distance (Å)
10µg/mL
50µg/mL
100µg/mL

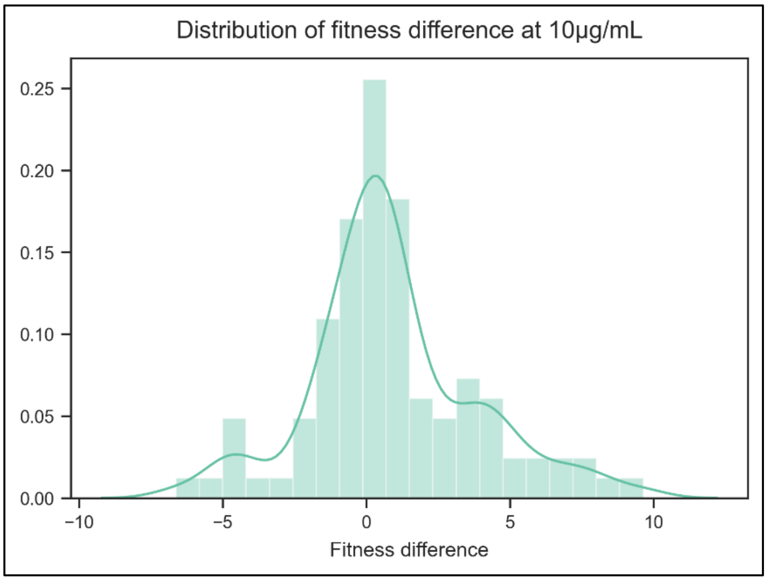

Fitness ratio as a function of distance between mutations
Fitness ratio
120
100
80
60
40
20
0
−20
0
10
20
30
40
50
Distance (Å)
10µg/mL
50µg/mL
100µg/mL

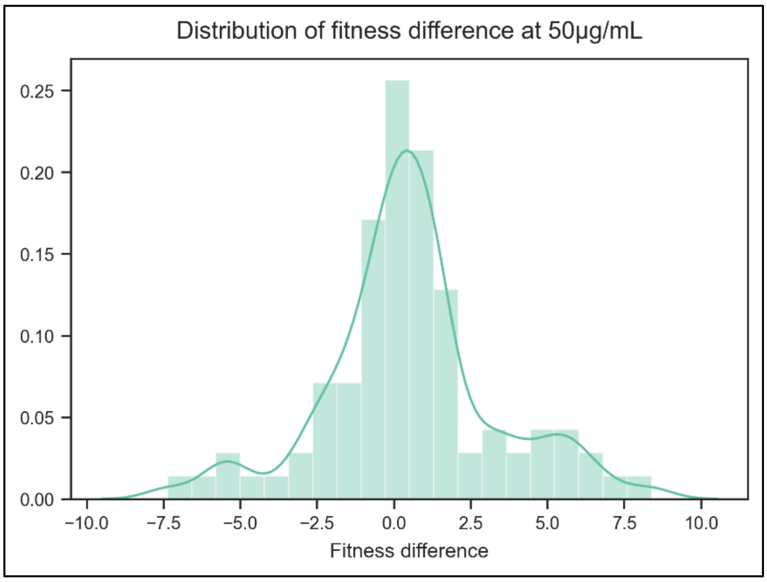

Frequency of nucleobase changes
Frequency
0.20
0.15
0.10
0.05
0.00
AC
CA
AG
GA
AT
TA
CG
GC
CT
TC
GT
TG
Change

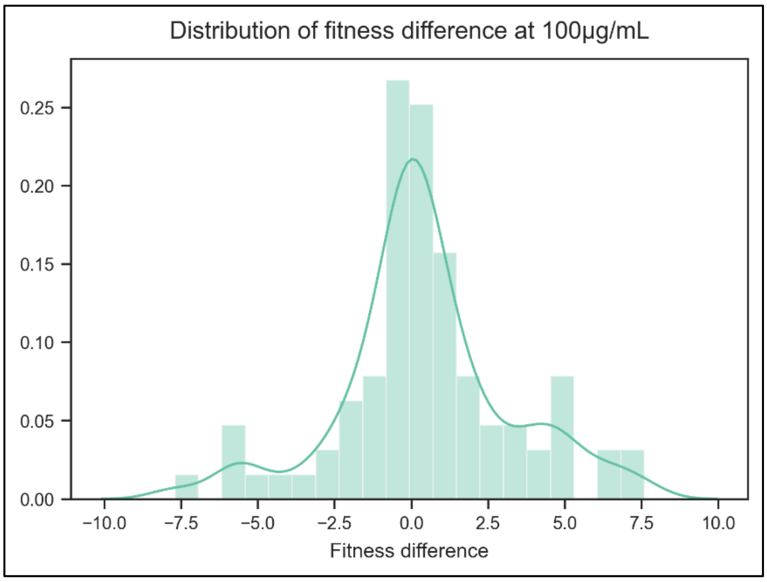

Jointplot between fitness difference and distance at 10µg/mL
Fitness difference
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
0
10
20
30
40
50
Distance (Å)

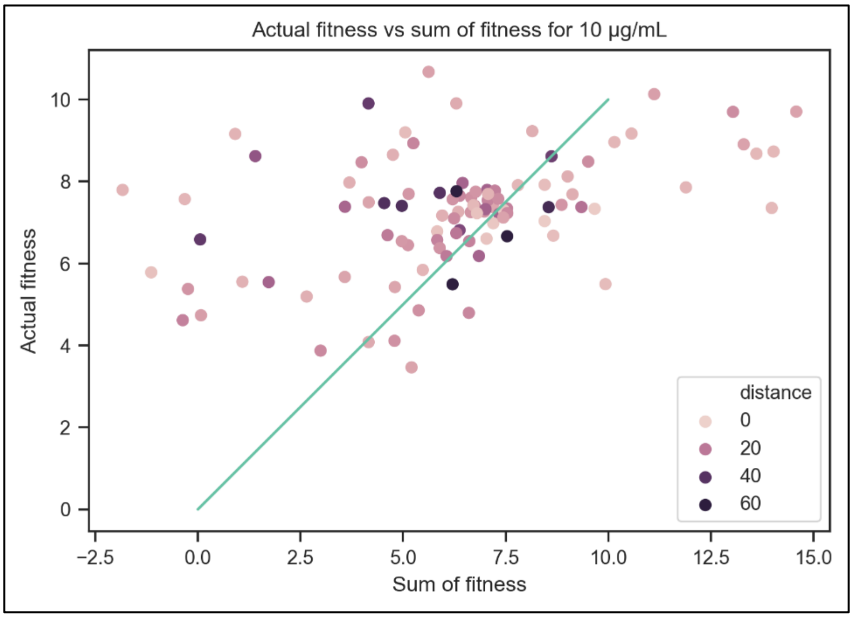

Jointplot between fitness difference and distance at 50µg/mL
Fitness difference
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
0
10
20
30
40
50
Distance (Å)

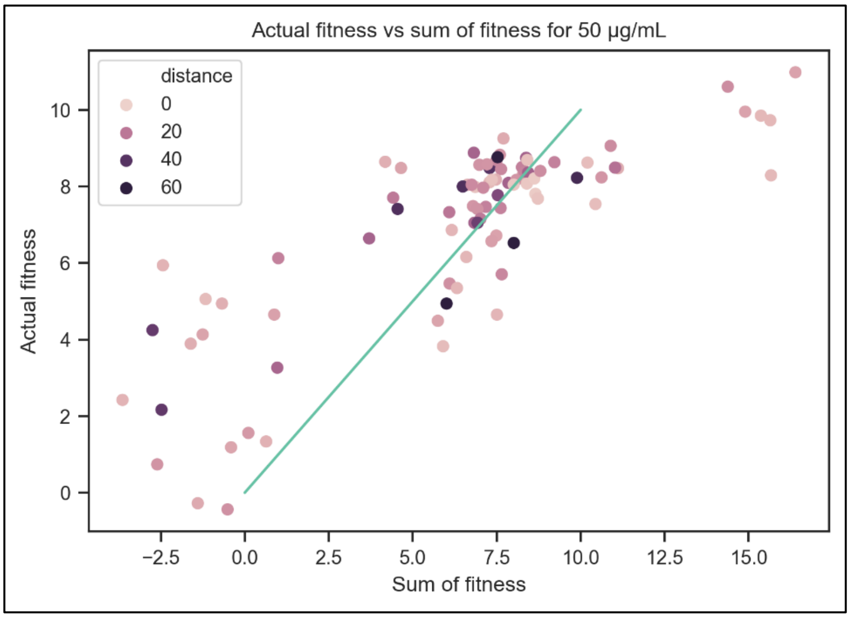

Jointplot between fitness difference and distance at 100µg/mL
Fitness difference
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
0
10
20
30
40
50
Distance (Å)

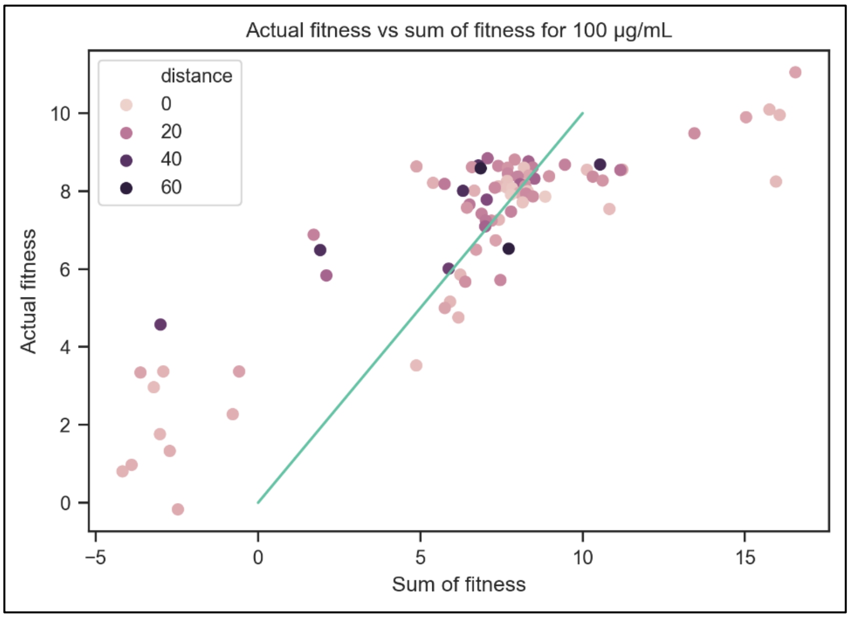

Distribution of fitness difference at 10µg/mL
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
Fitness difference


Distribution of fitness difference at 50µg/mL
0.25
0.20
0.15
0.10
0.05
0.00
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Fitness difference


Distribution of fitness difference at 100µg/mL
0.25
0.20
0.15
0.10
0.05
0.00
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Fitness difference


Actual fitness vs sum of fitness for 10 µg/mL
Actual fitness
0
2
4
6
8
10
Sum of fitness
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
distance
0
20
40
60


Actual fitness vs sum of fitness for 50 µg/mL
Actual fitness
0
2
4
6
8
10
Sum of fitness
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
distance
0
20
40
60


Actual fitness vs sum of fitness for 100 µg/mL
Actual fitness
0
2
4
6
8
10
Sum of fitness
−5
0
5
10
15
distance
0
20
40
60