Systems Biology & Neurobiology

Homework Report

Simon Chardin, Émile Sabatier, Samuel Diebolt

Understanding the Impact of Combination of Mutations on Rifampicin Resistance

ESPCI PARIS | PSL

Teachers: Philippe Nghe, Andrew Griffiths

1 Abstract

Deep mutational scanning (DMS) makes use of large-scale mutagenesis to reveal intrinsic protein properties, functions and the consequences of genetic variation. Recently, the CRISPR/Cas9-mediated genomic error-prone editing (CREPE) technology was developed as a high-throughput method for mutating essential genes of *Escherichia coli* [1]. Its authors applied the technology to target *rpoB*, the gene encoding the β subunit of bacterial RNA polymerase, and used deep sequencing to study resistance against the antibiotic rifampicin. In particular, the authors studied epistastis effects by comparing fitness of double mutants in *rpoB* with those from the respective single mutations in the presence of rifampicin. In this report, we replicated the aforementioned epistasis study using a simplified dataset provided by A. Choudhury.

2 State of the Art

Studying epistasis—be it in human or within bacteria—raises many challenges, as it can rarely be done using observational studies. However, understanding how combinations of mutations affect protein functions and behaviour within cells could give us insights into a huge number of biological processes, from antibiotic resistance to genetic diseases. In this section, we establish a non-exhaustive review of available methods and technologies used for studying epistasis in bacteria.

Deep mutational scanning (DMS) can provide significant insights into the function of essential genes in bacteria. This method couples genotype to phenotype to assess the activities of as many as 1 million mutant versions of a protein in a single experiment [2]. DMS is capable of scoring comprehensive libraries of genotypes for fitness in given environments in a massively parallel fashion. Essential bacterial genes are often targets of interest as they are key to their evolution, and can lead to phenotypes such as antibiotic resistance when mutated.

When the phenotype of interest is the cell's fitness in a specific environment, the presence of genetic interactions between mutations, i.e. epistasis, can constrain the course of evolution. Given the potential impact of epistasis in a variety of biological processes, recent studies have focused on measuring genome-wide levels of epistasis using the multiplex automated genome engineering (MAGE) technology [3]. MAGE was created for large-scale programming and evolution of cells: it simultaneously targets many locations on a chromosome for modification in a single cell or across a population of cells, thus producing combinatorial genomic diversity. It is based on lambda red-mediated recombination of single-stranded oligos to introduce mutations at specific genomic loci [4].

However, the MAGE technology has some limitations, as it was optimized only for a few cell strains and can lead to the accumulation of numerous off-target modifications. A recent study improved on MAGE by using a dominant-negative mutant protein of the methyl-directed mismatch repair (MMR) system, allowing efficient modification of multiple loci, without any observable off-target mutagenesis and prior modification of the host

genome. This improved technology, termed pORTMAGE, was used to achieve a transient suppression of DNA repair in *Escherichia coli*. In addition, pORTMAGE allows comparison of epistatic effects across a wide range of bacterial species [5].

Antibiotic resistance is one of the main topic of study when it comes to understanding the effects of epistasis in bacteria, as the key factors contributing to resistance are yet to be understood. A team of researchers studied how epistatic effects in *Escherichia coli* could be used to perturb the evolution of bacteria towards antibiotic resistance [6]. In this context, robotic lab-evolution platforms can be used to keep population size and selection pressure constant for hundreds of bacteria populations evolving in parallel. Using this method, specific cellular functions that drastically curtail the evolvability of resistance where identified. Using whole-genome sequencing, the team showed that strong negative epistasis was generally underlying these functions.

The CRISPR gene editing technology, already widely used in genome engineering, also allows for investigating how gene expression governs the adaptive pathways available to bacteria during the evolution of resistance. In that sense, the controlled hindrance of adaptation of organisms (CHAOS) approach was recently developed to induce negative epistasis in *Escherichia coli* to deter adaptation [7]. Using a library of deactivated CRISPR-Cas9 devices, the team perturbed the bacterial gene expression and observed that epistatic effects caused large losses of cell fitness in environment containing ciprofloxacin, a clinically-relevant antibiotic. Another team introduced the homologous sequence integration (HoSeI) method to study bacterial genome-wide epistatic phenomena. HoSeI is a genetic marker-less genome editing approach that introduces base substitutions in the target sequence by screening dead or alive cells. It was demonstrated in a strain of *Escherichia coli* to study the effects of epistasis on regulators of bacteria adaptive growth [8].

To generate the libraries of variants that are used in methods involving recombineering, error-prone PCR (epPCR) was developed to perform random mutagenesis. Error-prone PCR protocols are modifications of standard PCR methods, designed to alter and enhance the natural error rate of the polymerase. Taq polymerase is commonly used because of its naturally high error rate. Creating high-quality libraries of random sequences is an important step in this process as it allows variants of individual molecules to be generated from a single-parent sequence. Combined with the synonymous PAM-inactivating mutation (SPM), precise genome manipulation with high efficiency can be achieve in a few steps using CRISPR gene editing methods [9].

Unfortunately, these methods are still failing in scalability, as the vast number of possible epistatic interactions erodes statistical power. Systematically testing interactions is challenging both from a computational and from a statistical point of view, given the large number of possible interactions to consider [10]. Thus, it is of interest to build models to predict genetic traits based on epistasis. Algorithms such as MINED were developed to detect significant pairwise epistasis effects that contribute the most to cell fitness using machine learning approaches [11]. Another study proposed a reinforcement learning approach, EpiRL, where epistasis is modeled as a one-step Markov Decision Process [12]. The use of machine learning in trying to find highly interacted genes could help tackle the challenges raised by the high-dimensionality of epistasis data.

3 Motivations & Hypotheses

Improvements in DNA synthesis and sequencing have underpinned comprehensive assessment of gene function in bacteria. And currently, the genome mutagenisis techniques and study are using genetic transfer networks to make better predictions of the sequence or word sequence of an entire genome. But low-editing efficiencies and mutational biasing, is a downfall that needs to be attended to. It impacts greatly the quality of the fitness data. In recent years a small number of approaches have also achieved a high degree of effectiveness without mutational annotation.

The aim is to mesure the non-synonymous mutations and not the deletions. In this paper, we propose a novel model that achieves the goal. The idea is to investigate the functional basis of epistasis, and because rpoB plays a central role in transcription, we measured the effects of common rpoB mutations on transcriptional efficiency. Because mutations using the CREPE technology allows the study of combination of mutations. Two mutations are considered to be purely additive if the effect of the double mutation is the amount of the consequences associated with the particular single variations. This occurs whenever genetics are not linked with each other. Simple, component qualities were researched in early stages in the particular background of genes, they are usually fairly rare, along with many genes showing a minimum of some degree of association with epistatic connection.

We address this concept by measuring the particular fitness effect associated with rifampicin resistance mutations in the β subunit of RNA polymerase (rpoB) of Escherichia coli. Epistasis for fitness means that the selective effect of a mutation is conditional on the genetic background in which it appears.Epistasis can be easily seen in nature, the process in which it is involved is still not well known. Furthermore, its consequence on evolution, and its role in natural selection is still incomplete to our knowledge. The mutational path to high fitness genotypes is linked and even supposedly dependent of the genetic background in which novel mutations appear. This background effect is independent of the population as well as any loci of any gene. Sign epistasis has been defined in that the sign of the fitness effect of a mutation is under epistatic control [13].

Thus, using the CREPE method we can assess the fitness effects of individual mutations on the same loci, as well as on diverse location. All this in correlation with the stress to which the bacteria are exposed. We want to explore the theorical and empirical consideration implying the strong genetic constraint on the selective accessibility to high fitness genotype mutation path.

4 Methods & Results of the Supporting Article

4.1 CREPE Protocol

The CRISPR/Cas9-mediated genomic Error-Prone Editing (CREPE) is a method used to induce mutations from an error-prone PCR in a targeted region using CRISPR/cas9 method. First of all, a single functionnal gRNA must be found for the target. Then a Synonymous

Point accepted Mutation (SPM) is introduced in the target. The next step is to amplify and clone the target region with the SPM and unmutated end-homology (for the next step) into a plasmid, inducing error-prone PCR libraries.

After that, a Cas9-mediated lambda red recombineering technique has been set up. And finally, before the selection, plasmids are cured from the cells at 37°C (Phillips 1999).

4.2 Antibiotic selection

Using the CREPE protocol, the authors succeeded to construct mutant libraries. The idea is to study the resistance against an antibiotic, here the rifampicin. Without any mutation on the rpoB gene, the rifampicin can stop the protein production in the cell, by binding itself near the fork in the β sub-unit. The mutations due to the error-prone PCR libraries are covering partly the rifampicin resistance-determining regions. This is why it's a good marker for the fitness scoring.

They decided to compare 4 different cultures. The first one is the wild-type E. coli MG1655, without any change. The second one is a single colony of E. coli MG 1655 + SPM in order to ensure that the SPM does not affect the fitness of the cells. The third one is targeting the rpoB regions : gRNA + SPM only (SPM in RRDR II) . And finally with mutations in addition : gRNA + error-prone PCR synonymous PAM mutation (this adds mutations in the rifampicin resistance-determining regions.

They have made them grow at 37°C for 4 hours. After the growth, the Optical Density (OD) at 600 nm have been measured to have a reference. Then, they divided the cultures into 3 samples, each with a different rifampicin concentration. The concentrations are 10, 50 and 100 $\mu g/mL$.

At the end, a part of the cultures are taken for next-gen sequencing, in order to know the mutations linked to the measured fitness.

4.3 Data acquisition

After the selection, the libraries must be sequenced. The DNA was extracted. Each library is, after being amplified, sequenced using Nextera Next-generation sequencing MiSeq 2X300 kit with Illumina. (DNA sequencing manufacturer).

The raw data have been manipulated to extract information wanted, such as the amino acid changes.

4.4 Fitness calculations

For each variant, the fitness was estimated as

$$f = \log\left(\frac{C_{i,\text{post}} + 0.5}{C_{\text{wt,post}} + 0.5}\right) - \log\left(\frac{C_{i,\text{pre}} + 0.5}{C_{\text{wt,pre}} + 0.5}\right), \tag{1}$$

where $C_{i,\text{post}}$ and $C_{i,\text{pre}}$ are the variant read counts, respectively post- and pre-selection, for condition $i \in \{10\,\mu\text{g mL}^{-1}, 50\,\mu\text{g mL}^{-1}, 100\,\mu\text{g mL}^{-1}\}$, and $C_{\text{wt,post}}$ and $C_{\text{wt,pre}}$ are the wild-type read counts, respectively post- and pre-selection [14]. The $\frac{1}{2}$ constant was added

to each count to assist with very small counts. The standard error of this estimate was computed as

$$\mathrm{SE}(f) = \sqrt{\frac{1}{C_{i,\mathrm{post}}} + \frac{1}{C_{i,\mathrm{pre}}} + \frac{1}{C_{\mathrm{wt,post}}} + \frac{1}{C_{\mathrm{wt,pre}}}}. \tag{2}$$

A filter is necessary in the data. Not all the reads are relevant. Because they have targeted an essential gene, they consider that stop codons could not repeat. They used the following filter:

$$C_i \geq C_{\text{max-stopcodon}} + 2.56 \times \sqrt{C_{\text{max-stopcodon}}}. \tag{3}$$

Only the reads following this condition are kept. For each replicate, the fitness have been combined using Fisher score iterations (Rubin et al. 2017). Only the synonymous mutations have been selected in the dataset. The criteria to determine if a mutant is resistant against the rifampicin is its fitness greater than 2.96 standard deviations than the mean fitness of synonymous mutations.

We used the same formulas in our personal analysis.

4.5 Epistasis measurement

They wanted to compare the fitness of a double mutant with the sum of the fitness of the two single mutants. Only the double mutants where the two single mutants went through the filter are selected. The epistasis is defined as:

$$\varepsilon = f_{AB} - (f_A + f_B). \tag{4}$$

The idea is to compare the epistasis for different concentrations of rifampicin, in order to see if we gain or not combining two mutations, and to know which mutations are concerned.

5 Effects of Epistasis in Double Mutants

In this section, we replicated the analyses performed by the authors of the article supporting this report to better understand the impact of epistasis on rifampicin resistance. In particular, we were interested in comparing the fitness of double mutants, compared to the sum of fitness from the respective single mutations.

5.1 Data Preprocessing

The dataset provided by A. Choudhury was obtained by processing the sequencing output from a single biological replicate. Paired Illumina reads were assembled and aligned, and variant counts and amino acid changes were extracted from the qrowdot alignment output. Finally, variants were aggregated by grouping on the mutation positions and summing the read counts. The result is a dataset where each row corresponds to a unique genotype, with columns:

- `aa_change`: list of amino acid changes, with original amino acid, position and new amino acid;
- `pre`: read counts before selection;
- `ten`: read counts after selection on 10 µg mL^{-1} of rifampicin;
- `fifty`: read counts after selection on 50 µg mL^{-1} of rifampicin;
- `hundred`: read counts after selection on 100 µg mL^{-1} of rifampicin.

In this dataset, the wild-type corresponds to the row having the highest pre-selection read counts. The dataset was further processed using code available in the archive attached with this report and on a GitHub repository (https://github.com/sdiebolt/espci-sbn-homework). First, all rows that don't contain the SPM were removed from the dataset. Then, non-synonymous mutations were extracted from the `aa_change` column the dataset was aggregated by grouping on the non-synonymous mutations and summing their counts. This step was performed to ensure that later epistasis analyses have access to a dataset containing a unique observation of each combination of non-synonymous mutations. The fitness score and its standard error were then computed for each variant using equations (1) and (2). Finally, the filter described in equation (3) was used to remove erroneous reads. These steps resulted in a dataset containing 483 unique synonymous and non-synonymous mutations, with read counts, fitness and fitness standard error for each condition (pre-selection and rifampicin at concentrations 10 µg mL^{-1}, 50 µg mL^{-1} and 100 µg mL^{-1})

5.2 Distribution of Fitness in Synonymous vs. All Mutations

Since epistasis is measured using non-synonymous mutations only, We were first interested in the distribution of fitness estimates for all mutations and only synonymous mutations. Figure 1 shows histograms of these distributions for each selection condition.

As observed by the authors of the supporting article, the histograms show bimodal distributions for all mutations and unimodal distributions for synonymous mutations only at each rifampicin concentration. As synonymous mutations are unlikely to cause rifampicin resistance [1], the authors defined a threshold for resitant mutations as

$$t = \mu_{\text{syn},i} + 2.56 \times \sigma_{\text{syn},i}, \tag{5}$$

where $\mu_{\text{syn},i}$ and $\sigma_{\text{syn},i}$ are respectively the mean and standard deviation of the normal probability density function fitted on the fitness of synonymous mutations for rifampicin concentration i. The 2.56 constant is an approximate value of the 99.5% percentile point of the standard normal distribution. Therefore, the second mode of the distribution of all mutations for each condition corresponds to non-synonymous mutations that conferred resistance to rifampicin.

It is interesting to see that while both modes seem symmetric at low rifampicin concentration, this symmetry is broken at higher concentrations, with a frequency decrease of resistant mutations. This observation could mean that the resistance property of some variants is concentration-dependent.

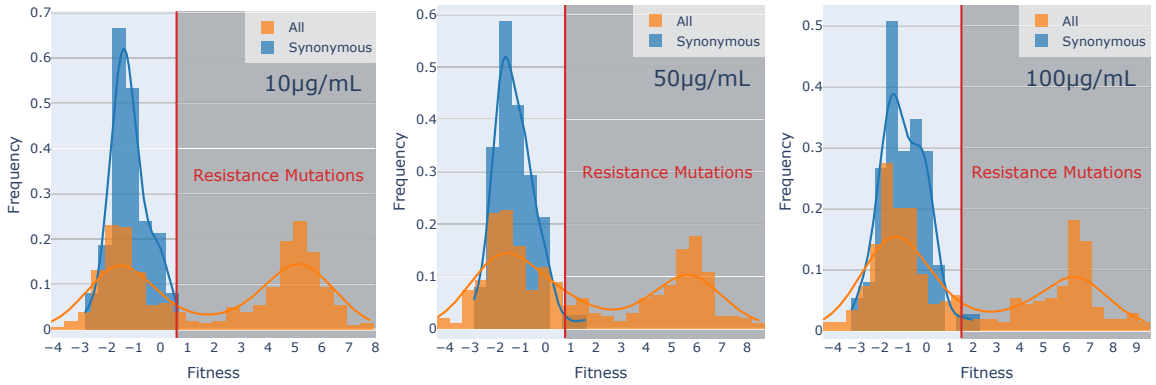

Figure 1: Distribution of fitness estimates for all mutations (orange) and for synonymous mutations only (blue) at different concentrations of rifampicin, 10 µg mL^{-1} (left), 50 µg mL^{-1} (middle), 100 µg mL^{-1} (right). The orange and blue lines are probability densities estimated using kernel density estimation, respectively for all mutations and synonymous mutations only. The red vertical line corresponds to the resistant mutations threshold, defined as $t = \mu_{\text{syn},i} + 2.56 \times \sigma_{\text{syn},i}$, where $\mu_{\text{syn},i}$ and $\sigma_{\text{syn},i}$ are respectively the mean and standard deviation of the normal probability density function fitted on the fitness of synonymous mutations for rifampicin concentration i.

5.3 Epistasis in Double Mutants

After studying how fitness estimates can help understand the impact of rifampicin concentrations on selection—this section isn't studied in this report—, the authors were concerned with the effects of epistasis in double mutants. The actual fitness of double mutants was compared to the sum of fitness from the respective single mutations. Figure 2 shows scatter plots of these comparisons for each condition.

Given the definition of epistasis introduced in equation (4), all double mutants that deviate from the 45° dashed line on Figure 2 are affected by epistasis. The double mutants were categorized as combination of two resistant mutation, one resistant and one sensitive mutation or two sensitive mutations using the same threshold from equation (5). We observe that these three groups show different epistatic behaviour. Not enough double mutations were classified as both sensitive in the 10 µg mL^{-1} condition, but the group shows positive epistasis, meaning that the combination of sensitive mutations is more beneficial than what would be infered by a simple linear relationship. This same phenomenon is observed for the double mutants consisting of one resistant and one sensitive mutation. Interestingly however, the double mutants that are combinations of two resistant mutations show negative epistasis. This effect could be explained by the fact that the cost of adding a resistance mutation to a variant that already is rifampicin-resistant is too detrimental to increase the fitness.

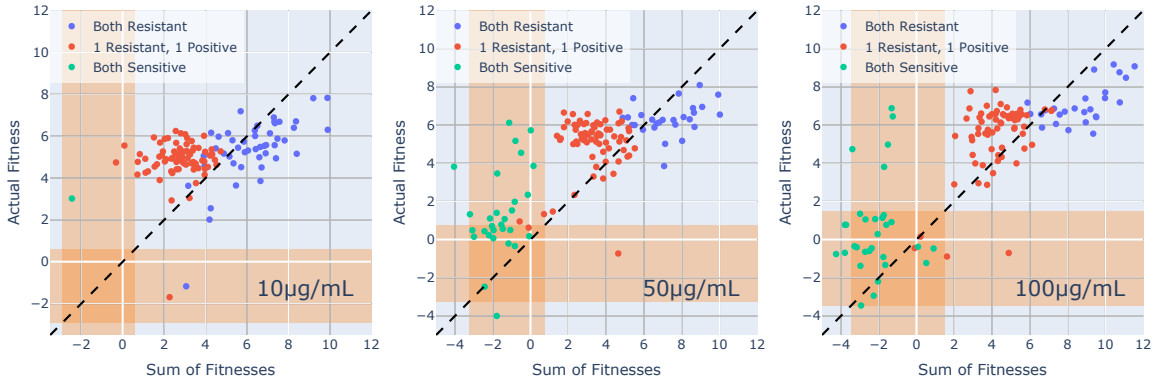

Figure 2: Scatter plots of actual fitness from double mutants vs. sum of fitness from the respective single mutation, at different concentrations of rifampicin, 10 µg mL^{-1} (left), 50 µg mL^{-1} (middle), 100 µg mL^{-1} (right). Double mutants were categorized as combination of two resistant mutations (blue), one resistant and one sensitive mutations (red) or two sensitive mutations (green). The orange shaded areas correspond to the ±3 standard deviations interval for the normal approximation of the synonymous mutations fitness.

5.4 Distribution of Epistasis Effects

6 Next Steps

References

[1] A. Choudhury, J. A. Fenster, R. G. Fankhauser, J. L. Kaar, O. Tenaillon, and R. T. Gill, "Crispr/cas9 recombineering-mediated deep mutational scanning of essential genes in escherichia coli," *Molecular Systems Biology*, vol. 16, no. 3, p. e9265, 2020.

[2] D. M. Fowler and S. Fields, "Deep mutational scanning: a new style of protein science," *Nature methods*, vol. 11, no. 8, p. 801, 2014.

[3] H. H. Wang, H. Kim, L. Cong, J. Jeong, D. Bang, and G. M. Church, "Genome-scale promoter engineering by coselection MAGE," *Nature Methods*, vol. 9, pp. 591–593, June 2012.

[4] H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest, and G. M. Church, "Programming cells by multiplex genome engineering and accelerated evolution," *Nature*, vol. 460, no. 7257, pp. 894–898, 2009.

[5] Á. Nyerges, B. Csörgő, I. Nagy, B. Bálint, P. Bihari, V. Lázár, G. Apjok, K. Umenhoffer, B. Bogos, G. Pósfai, *et al.*, "A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species," *Proceedings of the National Academy of Sciences*, vol. 113, no. 9, pp. 2502–2507, 2016.

[6] M. Lukacisinova, B. Fernando, and T. Bollenbach, "Exploiting epistasis to perturb the evolution of antibiotic resistance," *bioRxiv*, p. 738252, 2019.

[7] P. B. Otoupal, W. T. Cordell, V. Bachu, M. J. Sitton, and A. Chatterjee, "Multiplexed deactivated CRISPR-Cas9 gene expression perturbations deter bacterial adaptation by inducing negative epistasis," *Communications Biology*, vol. 1, pp. 1–13, Sept. 2018.

[8] Y. Miyake and K. Yamamoto, "Epistatic effect of regulators to the adaptive growth of escherichia coli," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[9] F. H. Arnold and G. Georgiou, "Directed evolution library creation," *Methods in molecular biology*, vol. 231, p. 231, 2003.

[10] L. Slim, C. Chatelain, C.-A. Azencott, and J.-P. Vert, "Novel methods for epistasis detection in genome-wide association studies," 2019.

[11] D. He, Z. Wang, and L. Parada, "Mined: an efficient mutual information based epistasis detection method to improve quantitative genetic trait prediction," in *International Symposium on Bioinformatics Research and Applications*, pp. 108–124, Springer, 2015.

[12] K. Huang and R. Nogueira, "Epirl: A reinforcement learning agent to facilitate epistasis detection," in *International Workshop on Health Intelligence*, pp. 187–191, Springer, 2019.

[13] A. D. Garst, M. C. Bassalo, G. Pines, S. A. Lynch, A. L. Halweg-Edwards, R. Liu, L. Liang, Z. Wang, R. Zeitoun, W. G. Alexander, *et al.*, "Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering," *Nature biotechnology*, vol. 35, no. 1, p. 48, 2017.

[14] A. F. Rubin, H. Gelman, N. Lucas, S. M. Bajjalieh, A. T. Papenfuss, T. P. Speed, and D. M. Fowler, "A statistical framework for analyzing deep mutational scanning data," *Genome Biology*, vol. 18, no. 150, 2017.