# Systems Biology & Neurobiology

Homework Report

**Simon Chardin, Émile Sabatier, Samuel Diebolt**


**Teachers:** Philippe Nghe, Andrew Griffiths

# 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 $\beta$ 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.**


# 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 existence 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].

Progress toward the aim to curb the tendency of bacteria to evolve as antibiotic resistant, requires a comprehensive understanding of the key factors that contribute to resistance. To study the evolution of bacteria population, and stimulate or generate resistant inducing mutations, different research team use robotic lab-evolution platform that keeps population size and selection pressure under tight control for hundreds of Escherichia coli populations evolving in parallel. Using these methods, it was discovered that membrane transport, LPS biosynthesis, and chaperones curtail the evolvability of resistance. Perturbations of efflux pumps prevented resistance evolution completely or forced evolution on inferior mutational paths [6]. These observations are very favorable for the exploration of key actionable plans to establish in the future. But the difficulty relies in the scalability of those information, the vast number of possible epistatic interactions erodes statistical power. As the size of genome-wide association studies (GWAS) increases, detecting interactions among single nucleotide polymorphisms (SNP) or genes associated to particular phenotypes is garnering more and more interest as a means to decipher the full genetic basis of complex diseases.

Systematically testing interactions is however challenging both from a computational and from a statistical point of view, given the large number of possible interactions to consider [7]. Thus it is of interest to build models to predict genetic traits based on epistasis. The MINED algorithm, like other later, detect very efficiently significant pairwise epistasis effects that contribute most to fitness. And its prediction accuracy is to be even better using faster and different mathematical models [8]. But this comes with computational challenges. In fact, epistasis detection modeling use the one-step Markov Decision Process where the state is genome data, the actions are the interacted genes, and the reward is an interaction measurement for the selected actions [9]. But those limits could be overcome with the help of machine learning.

On the other hand, CRISPR-Cas9 technology also allows for investigating how gene expression governs the adaptive pathways available to bacteria during the evolution of resistance. In that sense, CRISPR EnAbled Trackable genome Engineering (CREATE), the combinaison of CRISPR–Cas9 gene editing and a massively parallel oligomer synthesis allows for a trackable editing on a genome-wide scale. CREATE has been used for site saturation mutagenesis for protein engineering, and for the study of antibiotic resistance genes in bacteria. Thus, it could be of great use for the study of epistasis in E. Coli, but the use of CRISPRCas9 in these bacteria has been linked to cell death. Indeed, because bacteria mainly rely on homologous recombination, to repair double strand breaks, with the simultaneous cleavage of all copies of the Escherichia coli chromosome at the same position cannot be repaired. But inefficient cleavage can be repaired, thus creating a random chance of survival of the bacteria population. Linked with other downfall, the CREATE methods, is only usable for the study of mutants with a very noticeable impact on the fitness. As the use of gRNAs generate variable outcome, is linked to diminished editing efficiency, inadvertent non repairable mutations, and inducing errors with unedited cell

lines.

Another method for studying mutation and epistasis with CRISPRcas9 is the recently developed, HoSeI method. It is a genetic marker-less genome editing approach and introduces base substitutions in the target sequence on the original genome by screening dead or alive cells. It was performed on E. coli K-12 genome. It is prone to be used for the knockout of multiple genes and artificial introduction of mutations, which are useful experimental demonstrations of bacterial genome-wide epistatic phenomena [10]. It uses the combinaison of the construction of sgRNA expression plasmid and transformation of E. Coli strain harbouring pCas by the psgRNA-target and DNA fragment to recover the digested site by CRISPR-Cas9 [11].

In addition to that, the method most often used to generate variants with random mutations is error-prone PCR. 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. Combine with the synonymous PAM-inactivating mutation (SPM), precise genome manipulation with high efficiency can be achieve in a few steps with the CRISPRCas9 method [12].

Using all these method advantages and drawbacks, we aim to study the epistasis of E. Coli using the newly designed CRISPR/Cas9-mediated genomic error-prone editing (CREPE) technology.

## 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 $\beta$ 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.

# Methods & Results of the Supporting Article

# 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.

## 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 $\mu g\,mL^{-1}$ of rifampicin;
- `fifty`: read counts after selection on 50 $\mu g\,mL^{-1}$ of rifampicin;
- `hundred`: read counts after selection on 100 $\mu g\,mL^{-1}$ 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, non-synonymous mutations were extracted from the `aa_change` column.

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 $i \in \{10\,\mu\text{g}\,\text{mL}^{-1}, 50\,\mu\text{g}\,\text{mL}^{-1}, 100\,\mu\text{g}\,\text{mL}^{-1}\}$, 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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] K. Xie, J. Zhang, and Y. Yang, "Genome-wide prediction of highly specific guide rna spacers for crispr–cas9-mediated genome editing in model plants and major crops," *Molecular plant*, vol. 7, no. 5, pp. 923–926, 2014.

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

[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.