Doctoral Dissertation (Censored)
博士論文 (要約)

Deciphering Evolutionary Constraints through Microbial Laboratory Evolution combined with Machine Learning

(大腸菌進化実験と機械学習を用いた
進化的拘束の探究)

A Dissertation Submitted for the Degree of Doctor of Philosophy
December 2020

令和 2 年 12 月 博士 (理学) 申請

Department of Physics, Graduate School of Science
The University of Tokyo
東京大学大学院 理学系研究科 物理学専攻

Junichiro Iwasawa
岩澤 諄一郎

Doctoral Dissertation

Deciphering Evolutionary Constraints through Microbial Laboratory Evolution combined with Machine Learning

Junichiro Iwasawa

A Dissertation Submitted for the Degree of
Doctor of Philosophy
December 2020

Supervisor: Chikara Furusawa

Department of Physics, Graduate School of Science,
The University of Tokyo

Abstract

The prediction and control of evolution is a crucial topic for both evolutionary biology and tackling antibiotic resistance. Although the lack of sufficient data has long hindered the mechanism of evolution, laboratory evolution experiments equipped with high-throughput sequencing/phenotyping are now gradually changing this situation. The emerging data from recent laboratory evolution experiments revealed repeatable features in evolutionary processes, suggesting the existence of constraints which could lead to actual predictions of evolutionary outcomes. These results also paint an upbeat picture of evolution: biologically feasible states and evolutionary trajectories could be distributed on a low-dimensional manifold within the high-dimensional space spanned by biological features. However, previous laboratory evolution experiments were performed on a limited number of environments and we thus lack a systematic investigation of evolutionary constraints leading to the low-dimensional dynamics. In addition, fine-grained approaches for predicting and controlling evolutionary trajectories were out of the reach. This dissertation is dedicated to solve these problems through the utilization of laboratory evolution combined with machine learning based techniques.

First, we study the evolutionary constraints of *Escherichia coli* through the multi-omics data acquired from a large-scale laboratory evolution experiment. In biological data, it is often the case that the dimensionality p is much larger than the number of samples N which makes it difficult to perform statistical analyses such as covariance estimation. We will show how the the utilization of machine learning such as random forest regression and supervised principal component analysis could overcome this $p \gg N$ problem and contribute to probe the low-dimensional manifold of *E. coli*'s phenotypes. Our analyses also identify the genotype-phenotype map, revealing the mutations that lead to the different strategies of stress resistance. We further discuss how our analyses could decipher the evolutionary constraints of *E. coli*.

We next develop a novel method for predicting and controlling stress resistance evolution by inferring an empirical fitness landscape based on phenotypes of *E. coli*. The concept of the fitness landscape has been influential in many areas of research on evolution since they provide information on the predictability of evolution. However, the high-dimensionality of the genotypic space kept us from constructing an empirical genotype-fitness landscape capable of predicting evolution. Focusing on the fact that evolution leads to low-dimensional phenotypes rather than genotypes, we infer the phenotype-fitness landscape based on the stress resistance profiles. To do so, unlike typical laboratory evolution experiments, we monitor the resistance levels to multiple stresses during the course of evolution which allows the dense sampling of phenotypes and the corresponding fitnesses along evolutionary trajectories. We show that the structure of the inferred landscapes corresponds with the resistance acquiring mechanisms of *E. coli* and provide information of the directionality of evolution. We discuss how the inferred landscapes could be utilized for predicting and controlling evolution.

List of Publications

In reverse chronological order of publication:

1. Junichiro Iwasawa, Tomoya Maeda, Masako Kawada, Hazuki Kotani, and Chikara Furusawa, "Dynamical multi-stress data from laboratory evolution reveal the phenotypic landscape for *Escherichia coli*". *Manuscript in preparation.*

2. Junichiro Iwasawa, Daiki Nishiguchi, and Masaki Sano, "Algebraic correlations and anomalous fluctuations in ordered flocks of Janus particles fueled by an AC electric field". *Submitted to Physical Review Research.* arXiv:2011.14548

3. Tomoya Maeda*, Junichiro Iwasawa*, Hazuki Kotani, Natsue Sakata, Masako Kawada, Takaaki Horinouchi, Aki Sakai, Kumi Tanabe, and Chikara Furusawa, "High-throughput laboratory evolution reveals evolutionary constraints in *Escherichia coli*". *Nature Communications* **11**, 5970 (2020). (*co-first authors)

4. Junichiro Iwasawa, Yuichiro Hirano, and Yohei Sugawara, "Label-Efficient Multi-Task Segmentation using Contrastive Learning". Presented at *Brain Lesion Workshop of MICCAI 2020.* To be published in the workshop proceedings: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. arXiv:2009.11160.

5. Daiki Nishiguchi, Junichiro Iwasawa, Hong-Ren Jiang, and Masaki Sano, "Flagellar dynamics of chains of active Janus particles fueled by an AC electric field", *New Journal of Physics* **20**(1):015002, (2018).

6. Tomoyuki Mano, Jean Baptiste Delfau, Junichiro Iwasawa, and Masaki Sano, "Optimal run-and-tumble-based transportation of a Janus particle with active steering", *Proceedings of the National Academy of Science* **114**:E2580 (2017).

Papers 1 and 3 include the main contents of this dissertation described in Chapters 5 and 4, respectively.

Acknowledgements

I would first like to express my deepest gratitude and respect to my research and dissertation supervisor, Chikara Furusawa. The first time I met him was the seminar at the physics department in July 2015 where he discussed the dynamics and low-dimensionality observed in adaptive evolution. This eye-opening seminar inspired and led me to the wonderful field of microbial laboratory evolution. Being in the Furusawa lab, I have been able to broaden my research field from physics to machine learning and microbiology. The interdisciplinary atmosphere of the lab was something that I always enjoyed.

My sincere thanks must also go to the members of the dissertation advisory committee: Akinao Nose, Hideo Higuchi, Sosuke Ito, Sotaro Uemura, and Wataru Iwasaki. Their valuable advices and critical discussions played an important role to improve this dissertation.

I am greatly indebted to Masaki Sano, my supervisor during the Master's course, for showing me the way as a physicist through his broad interests and critical thinking. I am also very grateful to Daiki Nishiguchi, a former member of the Sano lab, for numerous discussions and advices about research and academic life, etc.. This all started from my undergraduate research in Sano lab concerning active Janus particles, where Daiki taught me experimental methods and the basics for analyzing data. Thanks to his encouragement (which was more than just once), I have been able to pursue the research on Janus particles in parallel with my Ph.D. work, resulting in a very memorable paper for me.

I would like to express my appreciation to all members of the Furusawa lab for the wonderful and exciting three years during my Ph.D. course. I would especially like to thank Yui Uchida, Masato Tsutsumi and Masayoshi Hiranaka for the stimulating chats at *Morikawa, Sesamitei* and also online during the quarantine, as they gave me important insights to advance this research. Saburo Tsuru taught me the very basics of microbial experiments which widened my view of research. I also appreciate the dense discussions which enriched chapter 5 on fitness landscapes / mechanisms for drug resistance. Nen Saito showed me how to play around with theoretical concepts and how to blend them with biological systems. I also thank him for his nice advice to use the phrase "Deciphering" in the title of this dissertation. I would like to thank all the past and current members of the lab for the fruitful time: Ryudo Ohbayashi, Yuki Aoki, Yukitaka Isaka, Kazuaki Nakamura, Humza Hemani, Yuma Shimizu, Yuji Omachi, Yuki Kanai, Shun Hasegawa, Naoki Hatanaka, Toyohiro Azuma, Shota Nagao, Yoshiyuki Nakamura, Naomi Yamaguchi, Keiko Nagase, Naomi Kobayashi, Masayo Saito, and Chiyoko Kobayashi.

I would like to offer my special thanks to Kunihiko Kaneko and the members of the Kaneko lab for valuable comments on this work through several seminars. Actually, the idea of inferring a fitness landscape from resistance time series data came up to me while I was preparing for the seminar in Kaneko lab, July 2020.

I would also like to thank all members of the Multiscale Biosystem Dynamics team at RIKEN BDR, Osaka for helping me out during my two-months research intern and at other occasions: Takaaki Horinouchi, Tomoya Maeda, Taisuke Seike, Atsushi Shibai, Kumi Tanabe, Hazuki Kotani, Masako Kawada, Natsue Sakata, Naomi Yokoi, Yumeko Tarusawa and Eri Noda. I am especially grateful to Tomoya Maeda for the wonderful collaborations. The biological insights given in chapter 4 owe a lot to the numerous discussions with him. Tomoya has also taught me much about microbiology and laboratory evolution experiments ever since our first contact at the RIKEN QBiC spring school in March 2016.

Finally, I would like to thank my parents, Seiichiro Iwasawa and Tomoko Iwasawa for their love and encouragement. I always appreciate the inspiring discussions with them on economics, philosophy and other miscellaneous topics. I wouldn't have been able to accomplish my academic life so far without their kind support.

Contents

Chapter 1

Introduction

The nontrivial and eye-catching features of Life have always stimulated scientists to investigate the underlying mechanisms and theories at the interface of physics and biology [1–3]. Hypothesizing that life is driven by electricity, Luigi Galvani performed the renowned 'frog leg experiment' which later led to the invention of batteries and opened the door to what is now called electrophysiology [4]. Based on the rules of chemical bonding and X-ray diffraction, James Watson and Francis Crick presented the theory of base pairing [5] and the theoretical hypothesis which led to the well known central dogma of molecular biology [6]. Our passion to understand biological systems have expanded the horizons of physics, and vice versa, theoretical ideas based on physical intuitions have provided new avenues in biology. However, it is also true that we are still far from a general theory of biological systems which has the level of predictive power that has been realized in other fields in physics.

How can we construct a predictive theory of biological systems given their overall complexity and large degrees of freedom? To survive in nature, biological systems have to solve an immense number of tasks including and not limited to sensing/reacting to external stimuli [7–13], metabolizing intracellular resources [14–17], transmitting information from one to another (including their offsprings) [18–22], etc. Although it is not easy to elucidate all of these relevant processes, we strongly expect that they are subject to the laws of physics and thus constrained in a certain manner. In other words, biological systems should be distributed on a low-dimensional manifold which lies within the immense parameter space [23–27]. Therefore, if we could find out the constraints that separate this low-dimensional manifold from the other parts of the parameter space, we could be able to build a predictive theory of biological systems. For example, general principles such as conservation laws and symmetries have led to theoretical predictions of flocking which has recently been observed experimentally in the long-wavelength behavior of living (and non-living) motile matter[1] [36–44].

In this dissertation, we focus on one of the most fascinating features of biological systems: evolution [45–50]. We especially explore the dynamics of stress resistance evolution of microbial systems [51–53]. Evolution, adaptation and learning work on different but overlapping time scales, and they all aim to tune the parameters of a biological system to solve the problems living things face when trying to survive and reproduce. Of course, the dynamics of evolution should also be subject to (evolutionary) constraints, keeping the relevant parameters on a low-dimensional manifold. Thus, our ultimate goal here is to construct a framework for predicting and controlling microbial evolution by elucidating these relevant constraints. Predicting evolution has long remained a difficult task due to the lack of experimental data which kept us from directly comparing theory and experiments. Recently, however, the situation is changing where laboratory evolution experiments combined with high-throughput sequencing/phenotyp-

[1] Although it is out of the scope of this dissertation, the field of Active Matter which is the framework for mechanical and statistical properties of motile living/non-living matter, has grown rapidly within the past 25 years and is still actively expanding [28–35]

ing has started to provide us an unprecedented amount of data on evolutionary dynamics [46, 50, 54–60]. In addition, recent progress in machine learning methods [61–69] and theoretical/empirical modelling including the works on fitness landscapes [47, 70–77] have allowed us to investigate the relevant constraints which underlie the massive and high-dimensional data. Given these recent developments in experiment, theory and data analysis, it is fair to say that the time is ripe for building a framework for predicting and controlling evolution [49, 78].

We mainly discuss the following questions in this dissertation:

- What are the evolutionary constraints which underlie the dynamics of stress resistance evolution of *Escherichia coli*?
- How can we elucidate the relevant constraints from the high-dimensional data from laboratory evolution and how could these constraints be understood in the two different high-dimensional spaces spanned by gene expression and stress resistance profiles?
- How can we utilize the evolutionary constraints to predict and control the dynamics (trajectories) of stress resistance evolution?

In Chapter 2, we review the recent progress in laboratory evolution experiments and empirical fitness landscape modelling. Recent laboratory evolution experiments have revealed repeatable patterns in especially the phenotypic changes through evolutionary processes. Together with the recent attempts to construct fitness landscapes from experimental data, we discuss the evolutionary constraints underlying the data, and their implications for the predictability in evolution.

In Chapter 3, we give the prerequisites for machine learning techniques used in high-dimensional data analysis. Not limited to laboratory evolution experiments, recent biological experiments are often associated with high-dimensional data such as transcriptome data. However, it is often the case that the number of samples N are far less than the dimensions of the data p which makes statistical analysis difficult. We review the recent advances in the field of machine learning to solve such $p \gg N$ problems.

In Chapter 4, we elucidate evolutionary constraints from a multi-omics dataset consisting genotypes, gene expression levels and stress resistance profiles which was acquired via a large-scale laboratory evolution experiment. To do so, we utilize a method called Supervised PCA which is based on random forest regression and principal component analysis. We will show that supervised PCA provides a near optimal representation in the gene expression space which corresponds with the constraints given in the high-dimensional space spanned by stress resistance profiles.

In Chapter 5, we propose a methodology to predict and control stress resistance evolution through the utilization of a empirical fitness landscape based on the data from laboratory evolution experiments. Building on the fact that phenotypes show better convergence than genotype changes, we construct a fitness landscape based on the stress resistance profiles to eight different stresses. We will discuss how our empirical fitness landscape based on stress resistance phenotypes could contribute to the prediction and control of resistance evolution.

In Chapter 6, we conclude this dissertation and discuss possible future directions of research.

Throughout the dissertation, we will show how the interplay of microbial laboratory evolution and machine learning techniques could decipher the evolutionary constraints underlying stress resistance evolution. We believe this work adds substantially towards the prediction and control of evolution.

Chapter 2

Searching for principles of stress resistance evolution

The evolution and spread of antibiotic resistance in bacterial pathogens is an increasing public health concern [51, 52, 79]. Although the development of novel antibiotics might help in the short run, history suggests the high likelihood of pathogens evolving resistance to novel antibiotics as they have done to existing antibiotics [79, 80]. Recently, alternative strategies have been carried out to combat antibiotic resistance through the investigation of mechanisms/predictability of resistance evolution [52, 78, 81]. Specifically, laboratory evolution combined with high-throughput sequencing/phenotyping has changed the game of understanding evolution by allowing us to repeatedly replay life's tape of evolution and investigate underlying mechanisms [45, 46, 50, 82]. In addition, considerable effort has been devoted in understanding the properties of empirical fitness landscapes which provide information on the predictability of resistance evolution [47]. Building on the increasingly accumulating knowledge in the field, we are entering the age where it is possible to discuss the predictability of evolution in a data-driven manner [49, 78].

Here in this chapter, we review the findings of laboratory evolution experiments / empirical landscapes mainly concerning antibiotic/stress resistance evolution. In Sect. 2.1, we review recent advances in laboratory evolution specifically focusing on the observed genotypic/phenotypic changes and their repeatability. Next in Sect. 2.2, we discuss experiments that reconstructed small regions of fitness landscapes and their implications to the predictability of evolution. Finally in Sect. 2.3, tracing the recent efforts in laboratory evolution and empirical fitness landscapes, we discuss possible directions for constructing a framework of predicting stress resistance evolution.

2.1 The state of the art in microbial laboratory evolution

2.1.1 Basic protocols for laboratory evolution

Laboratory evolution combined with high-throughput sequencing/phenotyping provides an unprecedented amount of evolutionary information of bacteria through the direct observation of evolution. Serial transfer is a popular method where cultures are grown in flasks / microwell plates, and a small portion of the grown cells are transferred to fresh medium at a fixed time interval. A pioneering work of microbial laboratory evolution is the long-term evolution experiment (LTEE) performed by the Lenski group where they have evolved *Escherichia coli* with limited nutrients for more than 70,000 generations [82–84]. The LTEE provided rich information of microbial evolution and their genetic bases. For example, it has been

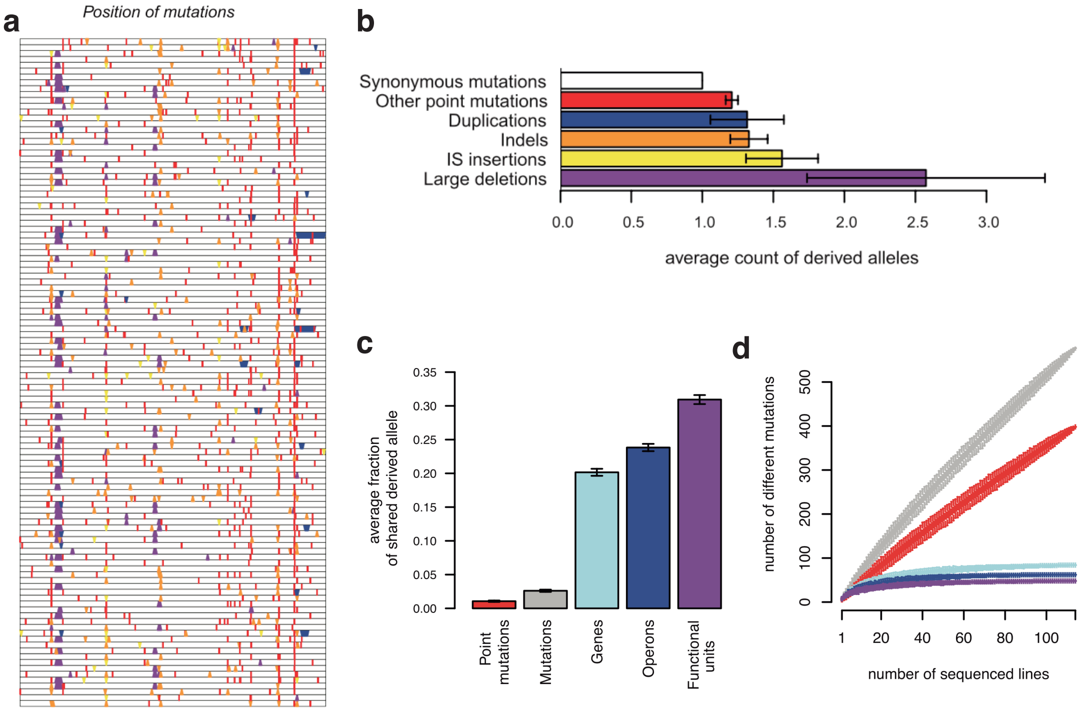

Fig. 2.1 **a**, The spectra of mutations, mapped along the *E. coli* B chromosome, in 114 independently evolved clones of *E. coli* under heat stress (42.2°C) for 2,000 generations (left panel). One strain was excluded from the analysis due to its increased mutation rate caused by the *mutL* gene. The colors of the spectra correspond to the histograms on the right. **b**, The number of lines sharing the same mutational types. **c**, The average fraction of shared derived allele among the 114 clones. **d**, The number of different mutations for subsets of the clones. The colors represent the levels of organization, as defined in (c). Figures reproduced from [54].

shown that *E. coli* could digest citrate[1] with the contribution of several potentiating mutations that did not themselves have a benefit in fitness [50, 85]. Starting from 1988, the LTEE is still ongoing.

The evolution of antibiotic (or stress) resistance could be observed through laboratory evolution experiments with increasing antibiotic concentration. For example, Toprak and his colleagues developed an automated culture device, which they call a morbidostat, that could dynamically adjust the drug concentration to maintain nearly constant growth inhibition [55, 86]. A similar drug concentration adjustment protocol was used by a robotic system in [87]. Another popular way to observe resistance evolution is to expose bacterial populations to varying concentrations of drugs and to transfer the population which grew in the highest concentration to fresh medium[2] [57, 59, 89]. In both methods, an increase in resistance could be observed through the time course of laboratory evolution. Combined with subsequent genotype/phenotype analysis, the evolved strains could provide information of the underlying mechanisms for resistance evolution. In the following subsections, we review the seminal works on resistance evolution that have been driven by laboratory evolution experiments[3].

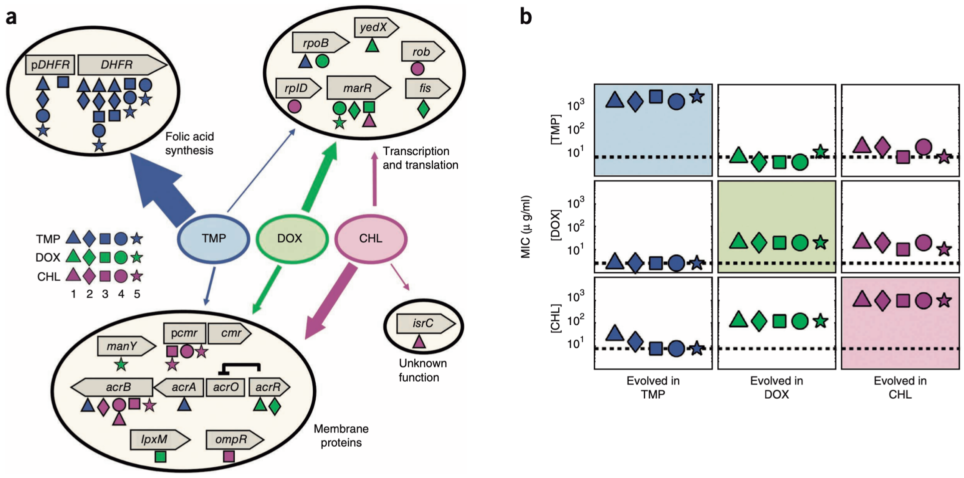

Fig. 2.2 **a**, The identified mutations in [55] and their corresponding functional groups. The thickness of the arrows show the mutation frequency for each functional group. **b**, The resistance profiles of the sequenced clones which evolved in TMP, DOX and CHL, respectively. Figures reproduced from [55].

2.1.2 Convergence and divergence in microbial evolution

Molecular diversity underlying adaptation to heat stress

Tenaillon and his colleagues performed an extensive genotype profiling where they sequenced 115 replicate populations of *E. coli* that evolved under high temperature [54]. These *E. coli* were serially propagated for 2,000 generations in Davis minimal medium with 25 mg/L glucose at a constant temperature of 42.2°C. They have detected 1258 molecular changes in total with an average of 11.0 events per clone where they estimate that $\sim 80\%$ of the intergenic and non-synonymous mutations were beneficial. Through the sequencing results, they found that the convergence of mutations varied by mutational type where 69% of large ($> 30 \sim$bp) deletions were identical between at least two lines while most of the point mutations were not shared (Fig. 2.1b). Although the convergence of exact mutations could not be observed, Fig. 2.1a clearly shows that there exists an extent of convergence in a more coarse grained level such as the gene/operon level or the level of functional units. For example, *rpoB* and other genes which encode the RNA polymerase complex were commonly mutated. Tenaillon *et al* conclude that their extensive experiment of 115 copies was not sufficient to capture all possible mutations, while the discovery of affected genes, operons and functional units could be captured with ~30 replicates (see Fig. 2.1c,d)[4]. These results suggest that while *E. coli* have a large mutational target when evolving against stresses, the diverse mutations show convergence at the level of coarse grained modules.

Specific and shared mutations for antibiotic resistance

Toprak *et al* performed laboratory evolution of *E. coli* under antibiotic concentrations that were dynamically adjusted to maintain growth inhibition [55, 86]. Here, the drug concentration is adjusted adaptively so that the bacterial growth is always inhibited by 50% by observing the time course of optical density (OD). They exposed the bacteria to three antibiotics separately: chloramphenicol (CHL), doxycycline (DOX) and trimethoprim (TMP). The laboratory evolution experiment was performed for

[1] It has been known that *E. coli* could ‘eat’ citrate under anaerobic conditions through the utilization of *citT* which is turned on under the absence of oxygen. Blount *et al* have found that one of their evolved populations could grow in citrate rich environments under the presence of oxygen by landing a duplicated *citT* sequence downstream of the *rnk* gene. We refer the interested readers to [48] which gives an illustrative description of the LTEE experiments for beginners.

[2] Of course, there are also studies that observe resistance evolution under a fixed concentration of drugs [56, 88].

[3] It should be noted that along with mutations, horizontal gene transfer is a major mechanism for bacteria to acquire drug resistance [52, 53]. Mostly because of its simplicity, past laboratory evolution experiments focus on the effects of mutations and these studies will be the main target of this review. However, there exists studies that try to evaluate the impact of horizontal gene transfer on drug resistance evolution [90].

[4] In Fig. 2.1c, we could see that the number of different point mutations / mutations grows linearly against the number of sequenced lines while the genes / operons / functional units grow in a logarithmic-like manner in the range of $1 <$ number of sequenced lines < 115. This scaling is informative when thinking of the sufficient number of replicates for laboratory evolution.

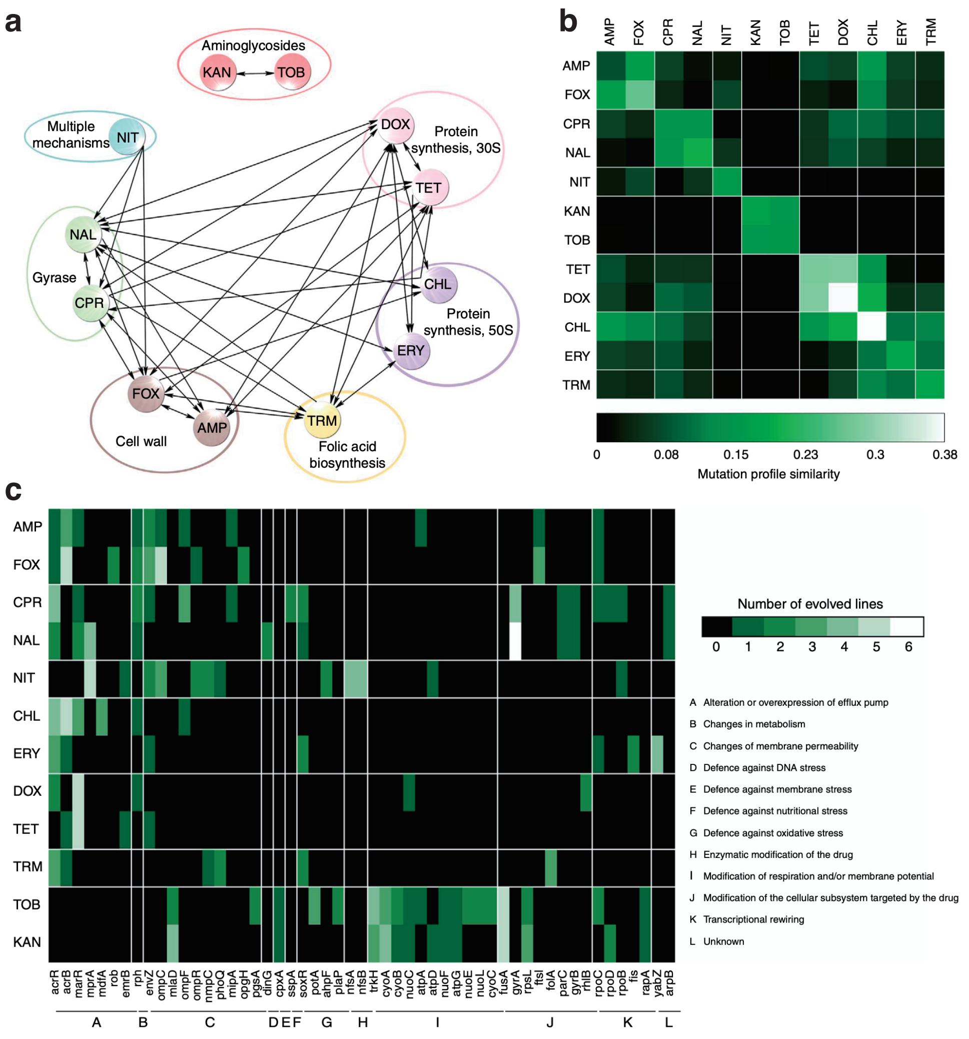

Fig. 2.3 **a**, The network of cross-resistance interactions. An arrow from antibiotic A to B indicates that adaptation to A decreased sensitivity to B in at least 50% of the evolved populations. **b**, Heatmap of the average mutation profile similarity of two strains adapted to different and identical antibiotics. Mutation profile similarity between each pair of evolved lines was measured by the Jaccard's coefficient between their sets of mutated genes. **c**, Mutational profiles of the 12 antibiotic selection regimes. Figures reproduced from [58].

five isogenic populations per antibiotic for 25 days, resulting in a $\sim$ 870, $\sim$ 10 and $\sim$ 1,680 fold increase in resistance for CHL, DOX and TMP, respectively. Isogenic clones were selected from each population of the final day and sequenced to observe their unique and common genetic changes (Fig. 2.2a). The identified single-nucleotide polymorphisms (SNPs) were classified to three functional groups: (i) transcription and translation, (ii) folic acid biosynthesis, and (iii) membrane transport. Interestingly, strains resistant to DOX and CHL involved a wide variety of mutations in transcription/translation and membrane proteins, indicating the large genetic target for adaptation. DOX and CHL evolved strains showed cross resistance to DOX and CHL respectively, which is consistent with the fact that they both have mutations in similar functional units. On the other hand, the mutations in strains resistant to TMP were strictly localized to DHFR and its promoter. In addition, the observed amino acid substitutions were shared among the independent clones. They further sequenced the transient populations of TMP evolved strains and found that the appearance and fixation of the confirmed mutations were mostly sequential. These results suggest the small mutational target for TMP resistance and that the observed sequential fixation of mutations were caused by this small number of beneficial mutations. The constraints for TMP resistance are discussed further in [71].

Genetic determinants underlying cross antibiotic resistance

Lázár and her colleagues have performed laboratory evolution for *E. coli* under 12 different antibiotics: Ampicillin (AMP), Cefoxitin (FOX), Ciprofloxacin (CPR), Nalidixic acid (NAL), Nitrofurantoin

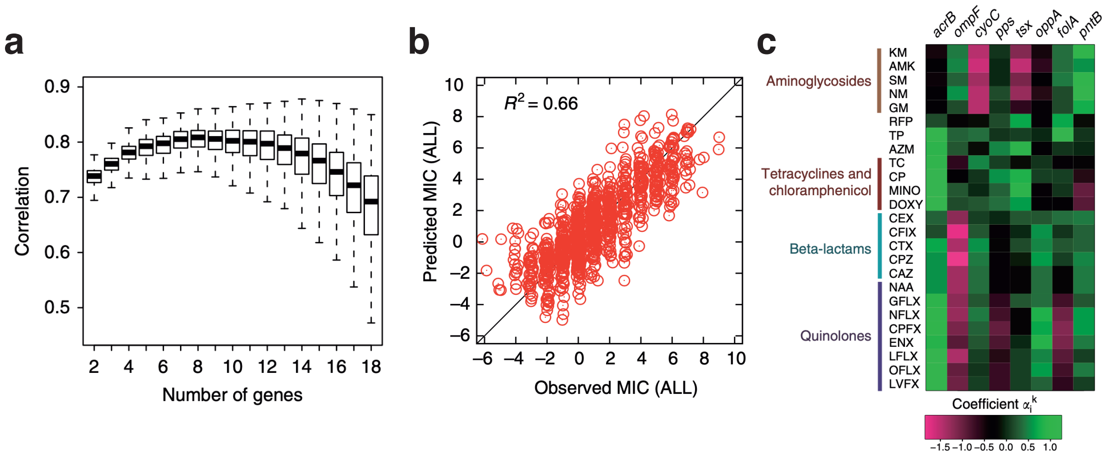

Fig. 2.4 **a**, The correlation between the predicted and observed resistance profiles for the evaluation of the prediction model. The predictions were made by a linear model based on the gene expression levels of N genes selected by a genetic algorithm. The evaluation was done over a test dataset which was not used for parameter optimization. **b**, The comparison between the predicted and observed resistance profiles. The data points have been slightly randomized to avoid the overlapping of points. Here, the resistance was predicted by a linear model based on eight genes shown in c. **c**, The coefficients for gene expression levels in the linear model. Figures reproduced from [59].

(NIT), Kanamycin (KAN),Tobramycin (TOB), Tetracycline (TET), Doxycycline (DOX), Chloramphenicol (CHL), Erythromycin (ERY), and Trimethoprim (TRM) [58]. They evolved 10 independent replicates for ∼240 generations for each antibiotic and performed whole genome sequencing for 5 – 6 strains per antibiotic (63 strains in total). They also measured the frequency of cross resistance, where the resistance acquisition to a certain stress leads to resistance in other stresses, within the 12 antibiotics they used (Fig. 2.3a). Interestingly, they observe asymmetry in the cross resistance network which was especially observed for NIT evolved strains (only 3% of the strains evolved in other environments reached NIT resistance). For the sequenced 63 strains, they measure the similarity in mutation profiles among the strains that evolved in the same environment (Fig. 2.3b). This correlation matrix, together with the individual mutation profiles (Fig. 2.3c) imply that the similarity in mutation profiles differs between different antibiotics. For example, strains evolved in AMP show a wide variety of mutated genes while 6/6 NAL evolved strains had a mutation in *gyrA*. However as a overall trend, the authors observe that several genes and functional modules were mutated repeatedly throughout the resistant strains (e.g. the AcrAB/TolC efflux system). Given that 66% of the shared mutated genes occurred in lines adapted to different antibiotics, Lázár *et al* claim that the ultimate targets of antibiotic selection are relatively limited functional modules. Their study suggests that antibiotic resistant phenotypes, which should be at least partially driven by these functional modules, may not have much variety.

Antibiotic resistance could be predicted by a few number of gene expression profiles

Suzuki, Horinouchi and Furusawa performed laboratory evolution for *E. coli* under 11 antibiotics: cefoperazone (CPZ), cefixime (CFIX), amikacin (AMK), neomycin (NM), doxycycline (DOXY), chloramphenicol (CP), azithromycin (AZM), trimethoprim (TP), enoxacin (ENX), ciprofloxacin (CPFX), colistin (CL). They evolved four independent replicates for 90 days for each antibiotic. For each of the evolved strains, they measured the resistance profiles for 25 antibiotics including the ones used for laboratory evolution, the transcriptome profiles, and the mutation profiles. The unique part of their work is that they measured not only the mutation profiles, but also the gene expression profiles for the evolved strains. In addition, they constructed a linear model to predict the stress resistance profiles based on the gene expression profiles of a limited number of genes. To prevent overfitting, they split the data of the resistance profiles into a training/test set and performed cross validation to optimize the number of genes N to use in the linear model[5] (Fig. 2.4a). As a result, $N = 8$ was optimal for their data with the predictions showing high correlation of $R = 0.66$ (R: Pearson's correlation coefficient) with the observed resistance profiles (Fig. 2.4b,c). The fact that resistance profiles could be predicted by a small

[5] This regression method was also applied for *E. coli* evolved under acids, alcohols, detergents, etc. [91].

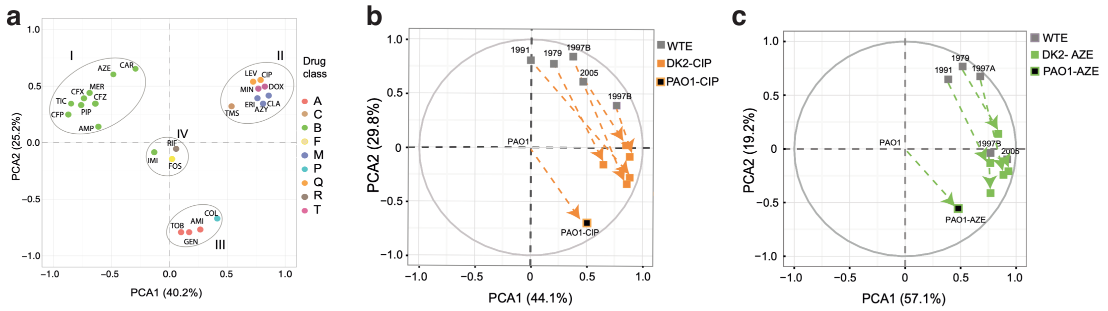

Fig. 2.5 **a**, The results of principal component analysis of the resistance profiles of the evolved *P. aeruginosa* strains, performed on the resistance profiles of 24 antibiotics. **b**,**c**, The distribution of the parent and evolved strains of the PAO1 strain and isolated DK2 strains which evolved in ciprofloxacin (b) and aztreonam (c), respectively. Figures reproduced from [60].

number of gene expression levels suggest that gene expressions changes that are responsible for resistance do not have much degrees of freedom. In other words, their result suggests convergence in the level of gene expression levels. They further investigated the mutations underlying the resistance profiles and found that while mutations in several genes were shared among the evolved strains, there was a certain extent of diversity in the mutation profiles. Consistent with previous laboratory evolution experiments, the work by Suzuki *et al* show the phenotypic convergence and genotypic diversity underlying resistance evolution[6].

Phenotype convergence in clinically isolated *P. aeruginosa*

Imamovic *et al* performed laboratory evolution for a *Pseudomonas aeruginosa* strain (PAO1) under 24 different antibiotics for 10 days (e.g. Amikacin (AMI), tobramycin (TOB), ampicillin (AMP), aztreonam (AZE), ciprofloxacin (CIP), etc. See Table 1 in [60] for the full list of antibiotics) [60]. They measured the resistance profiles of the evolved PAO1 strains and further performed principal component analysis (PCA). As a result, 75% of the variance in the drug resistance profiles could be expressed in the first two components of the PCA space, and the 24 evolved PAO1 strains could be classified to four distinct clusters (Fig. 2.5a). This implies phenotypic convergence in the drug resistance profiles even when the strains are exposed to different antibiotics. They further evolved five clinical isolates of *P. aeruginosa* (DK2)[7] from cystic fibrosis patients and evolved them under the 24 antibiotics. Interestingly, the drug resistance profiles of the evolved clinical isolates exhibited high correlation with the PAO1 strains evolved under the same stress (Fig. 2.5b,c, see also Fig.S3A in [60]). These results suggest the phenotypic convergence in the level of drug resistance profiles even for strains with different genetic starting points.

2.1.3 Collateral sensitivities and their applications

Determining collateral sensitivity networks underlying resistance evolution

Using the same evolved strains in [58], Lázár *et al* determined the networks of collateral sensitivity interactions [57]. Here, collateral sensitivity refers to the phenomenon where the resistance to a certain stress leads to sensitivity to another different stress. They especially found that the adaptation to aminoglycosides (KAN, TOB) led to the sensitivity of many other antibiotics (Fig. 2.6a). The authors further investigated the genotypes of the evolved strains and found that all sequenced clones resistant to aminoglycosides had a mutation in genes that influence the membrane electrochemical potential (e.g. *trkH*, *hemA*, *cyoB*, *cyoC*). Building on these observations, they propose that aminoglycoside resistance is achieved by altering the proton-motive force (PMF) across the inner membrane which leads to a reduced uptake of aminoglycosides (Fig. 2.6bc)[8]. On the other side, the majority of the antibiotics showing collateral sensitivity against aminoglycosides were substrates for the AcrAB/TolC and other

[6] Although we focus on the convergence of the end points of evolution here, low dimensional evolutionary trajectories in the phenotype space were observed for laboratory evolution experiments in [92, 93].

[7] DK2 and PAO1 share a common ancestor, but have diverged during the years of isolation.

[8] The authors further measured the membrane potential for the strains adapted to aminoglycosides and confirmed a reduction in membrane potential (see Fig. 4A in [57]).

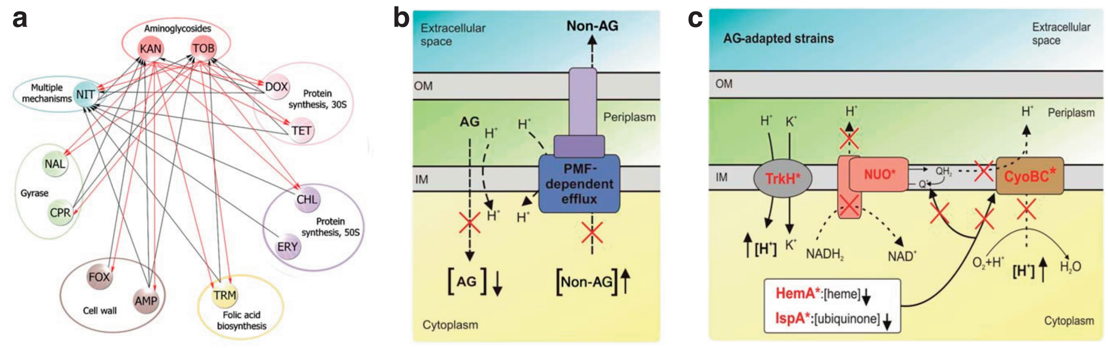

Fig. 2.6 **a**, The determined collateral sensitivity network from laboratory evolved strains. **b**, The suggested molecular mechanisms underlying collateral sensitivities between aminoglycosides and other antibiotics. The alteration of the membrane potential leads to reduced uptake of aminoglycosides and also reduced activity of PMF-dependent efflux pumps such as the AcrAB/TolC efflux pump. **c**, Mutations that lead to the alteration of the membrane potential. Figures reproduced from [57].

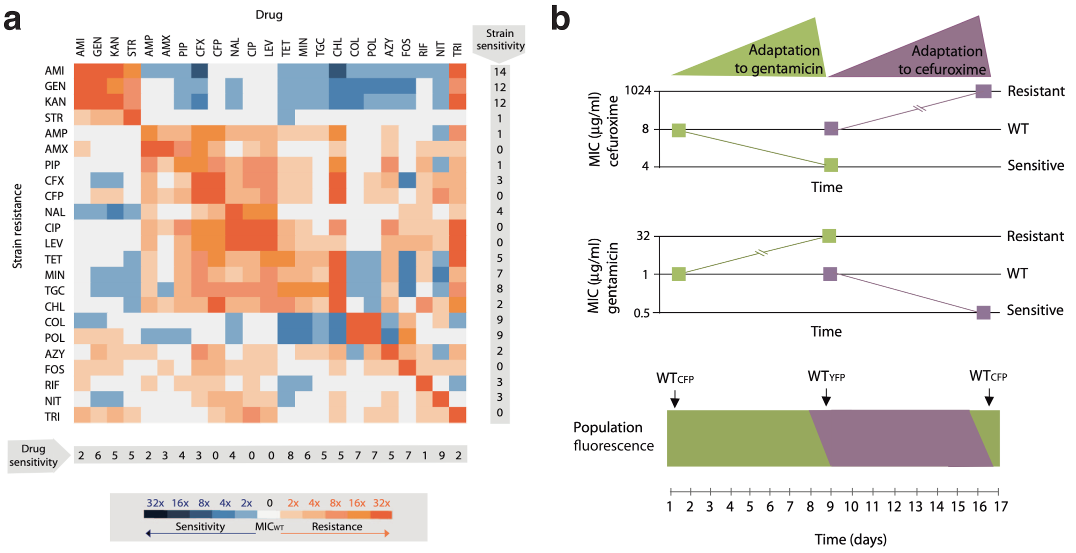

Fig. 2.7 **a**, The collateral sensitivity network elucidated from *E. coli* evolved in 23 antibiotics. **b**, The changes in resistance (minimal inhibitory concentration, MIC) for the CFP-labelled *E. coli* (wild type, WT) and the YFP-labelled WT that was mixed in the population when the drug was switched from gentamicin to cefuroxime. Figures reproduced from [56].

efflux pumps which are dependent on the PMF. These results suggest that the balance of the membrane potential underlies the observed collateral sensitivity between aminoglycosides and other antibiotics that are substrates of PMF-dependent efflux pumps. This collateral sensitivity between aminoglycosides and other stresses was also observed in [56, 59] and other studies as well.

Drug cycling utilizing collateral sensitivities

Imamovic *et al* performed a 10 day laboratory evolution of *E. coli* under 23 antibiotics using LB plates with drug gradients [56]. They measured the drug resistance profiles for the evolved strains and revealed the collateral sensitivity network between these 23 antibiotics (Fig. 2.7a). One of the main claims of their work was that collateral sensitivity relations could be utilized for clinical drug cycling strategies. To show this, the authors performed a drug cycle two antibiotics, gentamicin (aminoglycoside) and cefuroxime (β-lactam) which exhibit collateral sensitivity. They initially evolved CFP-labelled *E. coli* under gentamicin for eight days and then switched the drug environment to cefuroxime. When switching drugs, unadapted *E. coli* (wild type, WT) labeled with YFP were mixed by $1:1$. As a result, the gentamicin adapted *E. coli* were dominated by the WT *E. coli* which adapted to cefuroxime (Fig. 2.7b). Importantly, the authors claim that this drug cycling was not caused by fitness costs since both resistant strains could be recovered from the mixed populations under the absence of drugs. The results of Imamovic *et al* suggest clinical

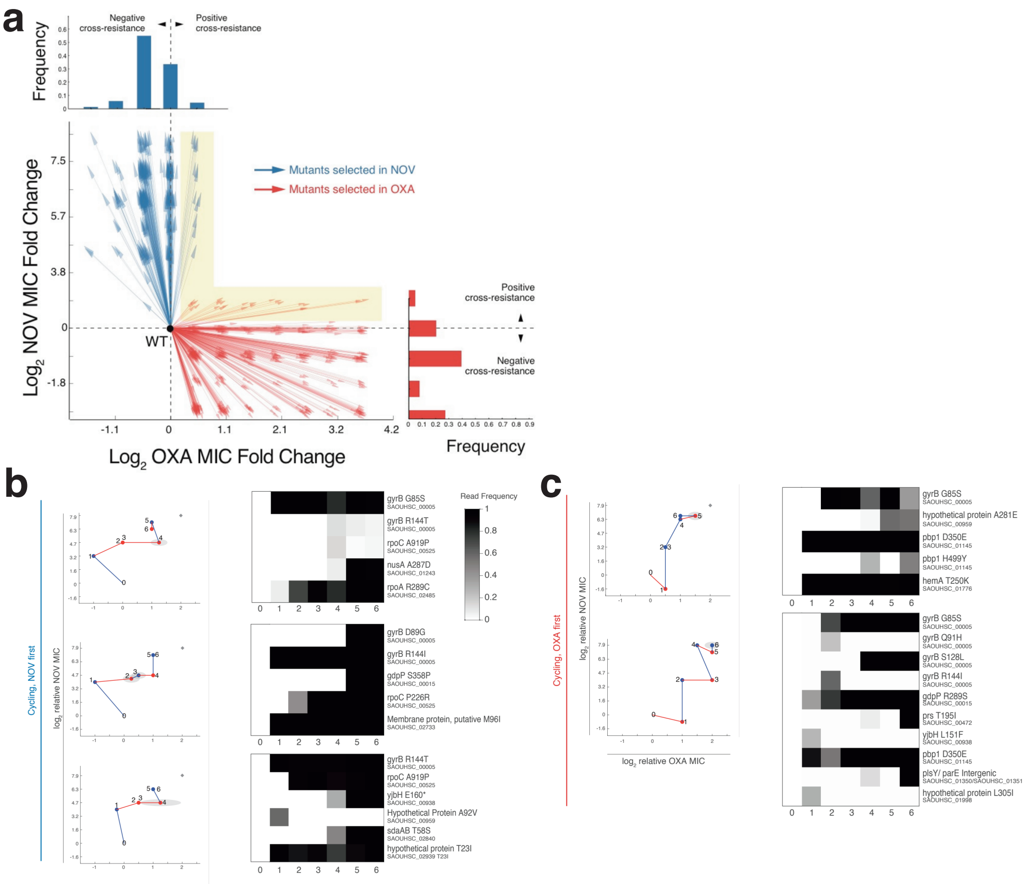

Fig. 2.8 **a**, The distribution of resistance for the mutants selected by OXA (red) and NOV (blue), respectively. **b**, **c**, The phenotype changes of five replicates under a OXA-NOV drug cycle starting from NVO (b) and OXA (c), respectively. The grayscale shading on the right shows the genotypic changes observed in each population. Figures reproduced from [94].

applications of the collateral sensitivity relations elucidated from laboratory evolution experiments.

2.1.4 The influences of adaptation history on resistance phenotypes

Drug cycling jeopardized by population diversity

Jiao and her colleagues investigated a mutant library of 3317 *Staphylococcus aureus* strains resistant to oxacillin (OXA), novobiocin (NOV), and four other drugs, respectively. The authors first observed the resistance profiles of strains resistant to OXA and NOV and find that most mutants resistant to OXA show sensitivity to NOV and vice versa (Fig. 2.8a). Since the results above suggest that mutants showing resistance to both OXA and NOV are rare, the authors hypothesized that drug cycling with OXA and NOV would be successful (would not lead to cross resistance to the two drugs) if the population size was kept small (tight bottleneck) at each step of propagation. To test this hypothesis, they performed a daily cycle of OXA and NOV, transferring 10^6 or 10^7 cells at each selection. Intriguingly, although the cells showed collateral sensitivity at the first step, they gradually found a path to acquire resistance to both OXA and NOV (Fig. 2.8b,c). In other words, the drug cycling method based on collateral sensitivities had been jeopardized by the diversity of resistance acquiring mechanisms of bacteria (this scenario has also been pointed out in [53]). The authors concluded that small population size is necessary, but not sufficient to maintain the collateral sensitivity relations during drug cycling. This is one of the seminal examples how the direction of drug resistance evolution (here in the two dimensional resistance space of OXA and NOV) could be altered by the history of adaptation (i.e. the genotypic background).

Drug-order specific effects in adaptation dynamics

Yen and Papin performed laboratory evolution for *Pseudomonas aeruginosa* under piperacillin (PIP), tobramycin (TOB) and ciprofloxacin (CIP). The unique part of this work was that the authors conducted a drug cycle of 20 days each, which allowed them to investigate how the adaptation to a certain drug

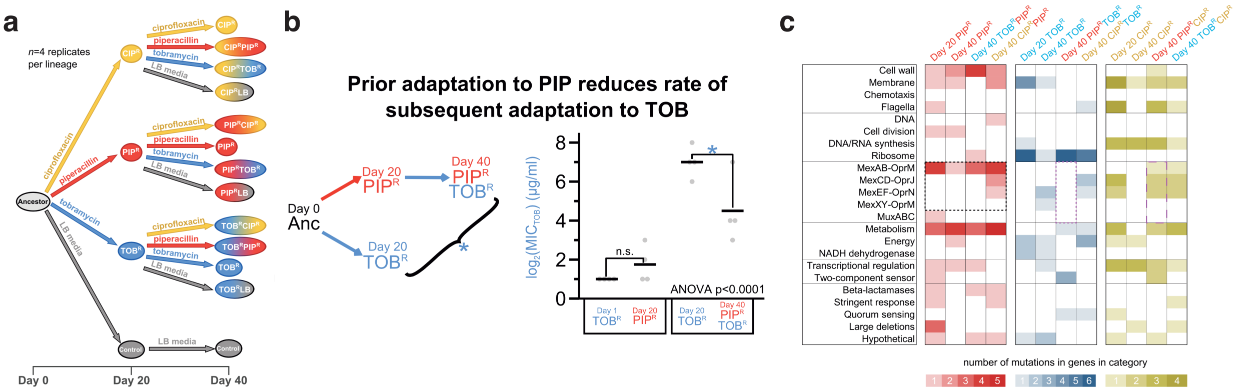

Fig. 2.9 **a**, A schematic picture describing the design of the laboratory evolution experiment. **b**, Stress resistance to TOB for strains that directly evolved under TOB for 20 days and strains that evolved under TOB after evolution to PIP for 20 days. **c**, The number of unique mutations in a gene of a specific functional class, shown by the color intensity. For the Day 40 strains, the acquired mutations during day 21 – 40 are shown and not those that were acquired until day 20. Figures reproduced from [95].

affects resistance acquisition to a different drug (Fig. 2.9a). For example, TOB resistance was more difficult to acquire for the strains that initially adapted to PIP (Day 40 PIP^R TOB^R)compared to the strains that directly evolved in TOB (Day 20 TOB^R) (Fig. 2.9b). The effects of the adaptation history was observed not only in the resistance profiles but also in the mutation profiles of the evolved strains. For instance, the Day 40 PIP^R TOB^R strain had kept their mutations related to the MexAB-OprM efflux pump, while the Day 20 TOB^R strain did not acquire any mutations related to this functional unit, which supposedly led to the observed difference in PIP resistance[9]. Interestingly, the authors suggested that the established collateral sensitivity relation concerning aminoglycosides [56, 57, 59] might not be observed for *P. aeruginosa*[10]. These results of Yen and Papin imply the effect of an initial adaptation to a certain drug on future evolutionary dynamics. In other words, the collateral sensitivity networks observed in laboratory evolution experiments might depend heavily on the starting point (i.e. the initial genetic/phenotypic background).

2.2 Inferring fitness landscapes from experimental data

Along with laboratory evolution experiments, a complementary approach to evaluate evolutionary dynamics stems from the idea of the fitness landscape. The fitness (or adaptive) landscape was first introduced by Sewall Wright where he coined the problem of evolution as “a mechanism by which the species may continually find its way from lower to higher peaks” in the landscape [97, 98] (Fig. 2.10). In the fitness landscape, each inner state (e.g. genotype) of an organism is associated with a fitness value (e.g. growth rate) in a fixed environment. By assuming that the organism continually optimizes its fitness, we would be able to predict the direction of evolution and also the predictability from the “ruggedness” of the landscape. Since the introduction of fitness landscapes, considerable theoretical efforts have been devoted to understand the relation between landscape properties and evolutionary outcomes [99–103]. In addition, starting with the pioneering work of Weinreich *et al*, the investigation of empirical landscapes based on mutation libraries and massive genotype sequencing have led to our understanding of accessible evolutionary trajectories and the underlying epistatic interactions [70, 71, 104–108]. In this section, we review the experimental studies which construct empirical fitness landscapes based on genotypes and discuss their implications.

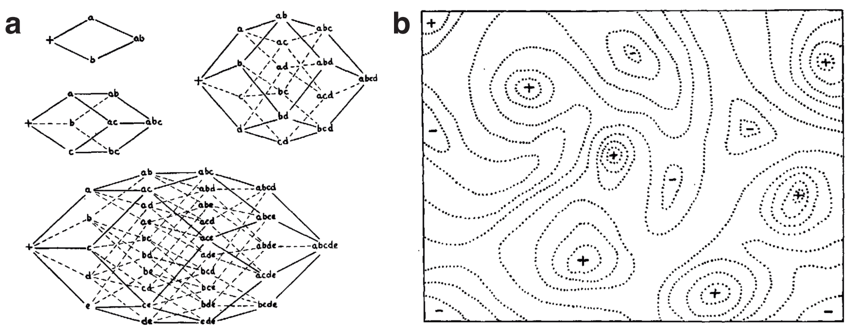

Fig. 2.10 **a**, A schematic image of the genotype space constituted by 2 – 5 genes. The dimensionality of the genotype space increases as the number of genes increase, resulting into a hypercube of genotypes. **b**, A representation of a high dimensional ($\sim \mathcal{O}(10^3)$) fitness landscape projected on a two dimensional plane. The dotted lines represent the contours with respect to fitness. Figure reproduced from [97].

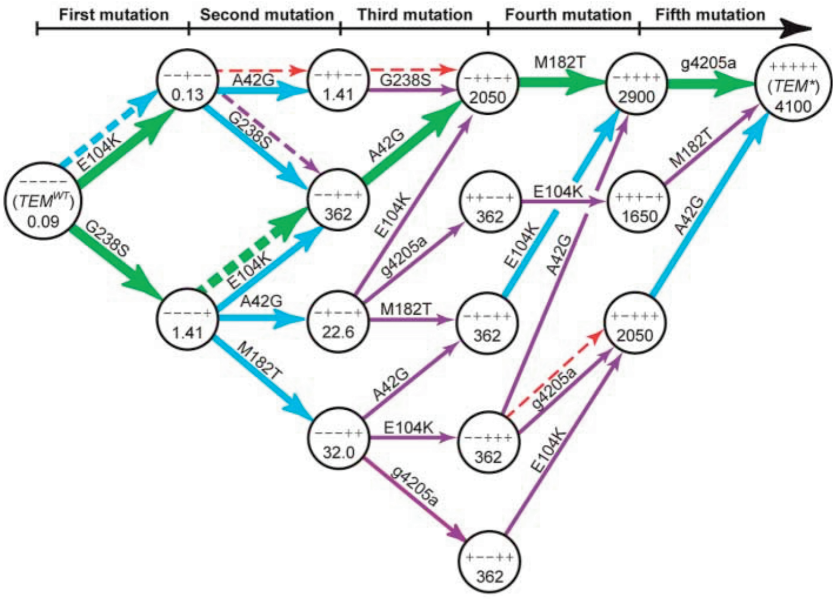

Fig. 2.11 The genotype-fitness landscape for β-lactamase. Genotypes that do not contribute to high probability mutational trajectories are omitted. The +/− for each node represents the presence/absence of the following mutations on β-lactamase: g4205a, A42G, E104K, M182T, and G238S. The numbers on each node indicate cefotaxime resistance (μg/mL) for each genotype. The thickness and color of the arrows show the relative transition probabilities. Figure reproduced from [70].

The β-lactamase landscape reveals constrained mutational pathways to fitter proteins

Weinreich and his colleagues constructed a fitness landscape for β-lactamase a protein that mediates resistance to β-lactam antibiotics, based on its genotype [70]. Five point mutations (g4205a, A42G, E104K, M182T, and G238S)[11] for the *TEM*-1 β-lactamase were selected and all $2^5 = 32$ possible *E. coli* mutants were constructed. The fitness of these mutants were measured through their resistance to cefotaxime. Assuming that the time for mutation fixation is far less than the time between mutations, the authors were able to estimate the probabilities of 120 mutational trajectories from TEM^{wt} to TEM^* (the genotype with all five mutations). Interestingly, this analysis suggested that only a few mutational trajectories are accessible for improving TEM^{wt} for cefotaxime resistance (Fig. 2.11). Although only

[9] The effects of adaptation history on resistance acquisition were also recently discussed in [96].

[10] At the same time, the authors note that they only used one antibiotic per drug class. For reference, in [60] where they also use *P. aeruginosa* (PAO1), TOB adapted strains rarely exhibited collateral sensitivity to other drugs. However, strains that evolved in ciprofloxacin, erythromycin, doxycycline and other stresses showed sensitivity to tobramycin and other aminoglycosides indicating an asymmetric tradeoff relation.

[11] These five mutations jointly increase resistance to cefotaxime by a factor of $\sim 100,000$.

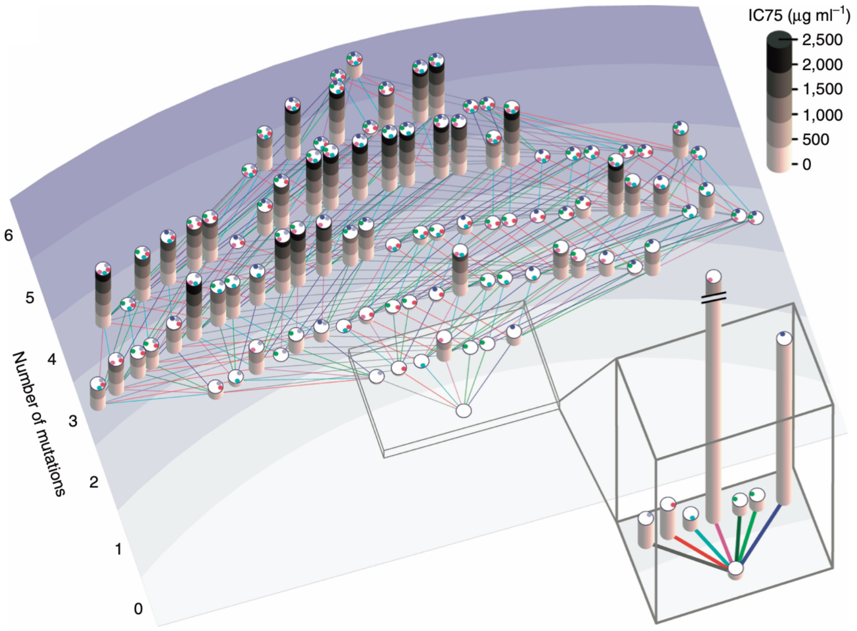

Fig. 2.12 The genotype-fitness landscape for the DHFR gene and its promoter. The landscape is based on the trimethoprim resistance of mutants with the combinations of five mutations for DHFR (P21L, A26T, L28R, I94L and W30G/R) plus one for its promoter (-35C>T). Figure reproduced from [71].

five mutations are considered, this work highlights how epistatic interactions could constrain accessible paths in protein evolution[12]. Note, Fig. 2.11 which shows the mutational trajectories from *TEM*$^{\text{wt}}$ to *TEM** could be considered as a fitness landscape provided by genotypes. Since every genotype (except *TEM*$^{\text{wt}}$) has at least one mutation that increases resistance, there are no local optima in this landscape which makes it single-peaked.

Epistatic interactions expand the number of indirect paths in the DHFR landscape

Fitness landscapes are not always single-peaked. Palmer and his colleagues constructed the fitness landscape of trimethoprim spanned by the genotypes of DHFR [71]. The genotypes (-35C>T (promoter), P21L, A26T, L28R, I94L and W30G/R (two different mutations)) were selected from a previous laboratory evolution for trimethoprim resistance [55], and were recombined to *E. coli* resulting in $2^5 \times 3^1 = 96$ DHFR mutants. Fitness was quantified through the resistance to trimethoprim. This comprehensive investigation of mutants and fitness produced a 'rugged' fitness landscape with 11 peaks where no gain/loss of a mutation is able to increase trimethoprim resistance (Fig. 2.12). Interestingly, the authors point out that this 'ruggedness' is generated by higher order interactions than pairwise interactions since every possible pair of mutations coexists in at least one adaptive peak. One of the main claims of this paper was that indirect paths, where the loss of an initially advantageous mutation leads to higher resistance, provide escape pathways from evolutionary 'dead-ends' and increase the number of genotypes that could access to a certain peak.

Cryptic genetic variation increases accessible mutational trajectories

Zheng *et al* studied the effect of cryptic genetic variation, standing genetic variation that does not normally contribute to phenotypic variation, on accessible pathways to fitter phenotypes. They first performed four rounds of directed evolution subject to stabilizing selection for *E. coli* with YFP, where the cells were subject to PCR mutagenesis and 20% of the cells with yellow flourescence within a narrow range around the median were selected for the next round. They next performed four rounds of directed evolution on the population with increased cryptic variation through stabilized selection and the population with zero cryptic variation towards green flourescence. As a result, they found that the populations that evolved from populations with cryptic variation (V^{C}_{1-4}) tended to acquire higher green

[12] Ever since this pioneering work by Weinreich *et al*, many subsequent studies have been performed investigating the details of the *TEM*-1 β-lactamase landscape [109–111].

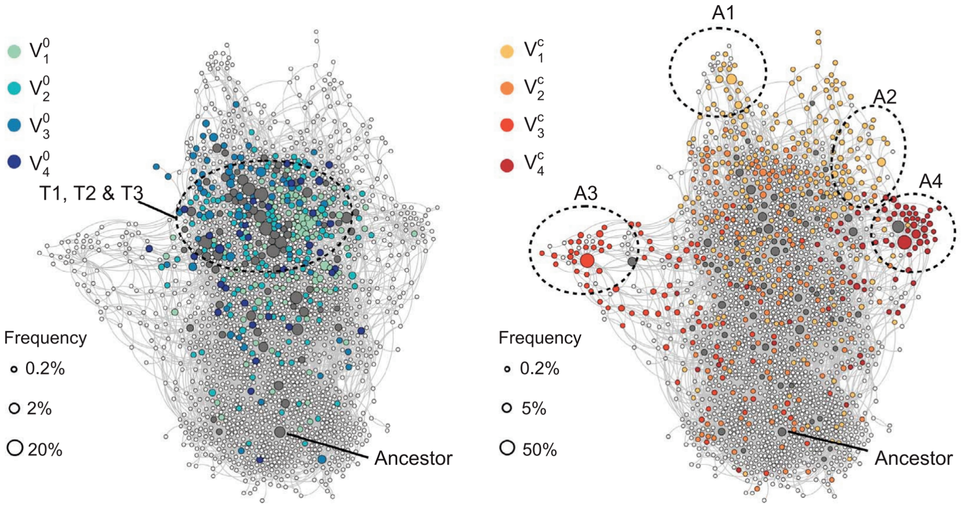

Fig. 2.13 The map of genotypes observed in the evolution experiment in [107]. Nodes represent genotypes and are connected if they differ in a single amino acid. The size of circles represent the genotype's frequency, and the color represents the four population replicates where V^C_{1-4} and V^0_{1-4} represent the evolved populations that evolved from populations with/without cryptic variation, respectively. T1, T2, T3 represent the dominating genotypes for the populations that evolved green flourescence directly from V^0. A1, A2, A3, A4 represent the dominating genotypes for the populations that evolved from V^C_4. Figure reproduced from [107].

flourescence than those that evolved from populations with no cryptic variation (V^0_{1-4}). To find out how cryptic genetic variation facilitated evolution, the authors used single-molecule real-time sequencing to genotype 500 to 1,000 evolved variants for each population (Fig. 2.13). The authors found that genotypes that led to high green flourescence were constituted with genotypes that were not readily accessible from the populations with zero cryptic variation (e.g. one of the strains acquired high flourescence by the combination of mutations that were only beneficial when combined). However, the the populations with cryptic variation already had intermediate genotypes that mediate paths to higher flourescence which led to accelerated evolution.

The map of genotypes in Fig. 2.13 is similar to a fitness landscape spanned by the relevant genotypes for green flourescence evolution. The colors correspond to the final generations of the populations which could be thought as surrogate values for fitness. However, because the map is projecting the high-dimensional genotype space to a 2D plane, the structure of the landscape (e.g. 'ruggedness') could not be handled readily.

2.3 Why can we expect evolution to be predictable?

The predictability (i.e. chance and necessity) of evolution is a classic topic in evolutionary biology [45, 48, 49]. Because evolutionary processes involve a certain level of stochasticity, many have been skeptical about the predictability in evolution. In addition, the lack/sparseness of experimental data kept us from directly comparing theories with experiments. This situation is now changing due to the emergence of massive evolutionary information from laboratory evolution experiments combined with high-throughput sequencing/phenotyping and the construction of empirical fitness landscapes. As we have seen in the previous sections, these experimental approaches give us several reasons to be optimistic of predicting evolution [49, 78].

First, the empirical fitness landscapes in section 2.2 have shown that epistatic interactions, where the fitness effect of a certain mutation depends on the presence/absence of other mutations, could constrain possible mutational trajectories to fitter phenotypes. Fitness effects brought by individual genes are not always additive and thus, epistatic interactions could sculpt fitness landscapes so that only a limited number of evolutionary trajectories are possible [70, 112]. However, it should also be noted that epistatic interactions could also bring "ruggedness" (i.e. multiple local optima) to the fitness landscape [71, 107], leading to multiple probable outcomes for evolution [108]. Although rugged landscapes lead to sequence level stochasticity, it is fair to say that epistatic interactions could constrain possible evolutionary pathways within the high-dimensional genotype space, leading to predictability in evolution. Indeed, several studies suggest that diminishing-returns epistasis could lead to predictable trajectories in fitness evolution despite the sequence level stochasticity [113, 114].

Second, the massively parallel laboratory evolution experiment by Tenaillon *et al* and other studies suggest that despite the diversity in the sequence level, evolution could lead to convergence in coarse grained units such as genes, operons and functional units [54, 55, 58, 59]. Biological networks such as metabolic/regulatory networks have redundancies and thus, different single nucleotide and amino acid changes could lead to similar functional effects. What is interesting is that evolution somehow leads to repeatable outcomes in the high level features such as phenotypes. These repeatable features are important building blocks for predicting evolution. Studies of antibiotic resistance evolution show that evolution to a certain drug leads to repeatable collateral sensitivity relations between different drugs. One such example is the tradeoff between aminoglycosides and drugs which are pumped out the cell by efflux pumps [56, 57, 59]. We have seen that this tradeoff could be explained by the balance of the PMF across the inner membrane [57]. These interactions between drugs could be supported by the changes in the gene expression space. In fact, Suzuki *et al* show that the changes in drug resistance (including collateral sensitivities) for strains that evolved in various stress environments could be accurately predicted using only ~ 8 genes. These results suggest the existence of evolutionary constraints which limit possible phenotypic outcomes through evolution. In addition, we might be able to elucidate such constraints by extensively studying the relation between gene expression profiles and resistance profiles of diverse stresses. This is what we aim to do in Chapter 4 where we study evolutionary constraints through the multi-omics data acquired from *E. coli* strains adapted to a diverse range of stresses.

However, recent studies also suggest that collateral sensitivities between drugs are not so simple. The laboratory evolution experiments of Yen and Papin suggest that collateral sensitivities between drugs depend on the mutational background of the strain [95]. The study by the Kishony group also supports this idea where they observe that bacteria could find a narrow mutational path to cross resistance between drugs that initially showed sensitivity to each other [94]. These studies suggest that in order to acquire a full understanding of evolutionary constraints for resistance evolution, we need to study the progression of phenotypes from different mutational backgrounds. This is the motivation for Chapter 5 where we study resistance evolution from multiple starting points and seek to infer the structure of the underlying fitness landscape. Overall, laboratory evolution experiments and empirical fitness landscapes have revealed the existence of constraints which lead to repeatable features in evolution. The identification of such evolutionary constraints would be essential for constructing a framework for predicting evolution.

Chapter 3

Machine learning methods for high-dimensional biological data

Here, we briefly review the recent advances in the field of machine learning and its applications for high-dimensional data acquired in biological studies. In Section 3.1, we explain the basic and popular dimension reduction methods for analyzing and visualizing high-dimensional data. Next in Section 3.2, we discuss effective approaches for cases where the number of features p is much larger than the number of samples N, which is a typical situation in biology. Finally in Section 3.3, we briefly review methodologies using artificial neural networks for analyzing biological data.

3.1 Dimension reduction techniques for high-dimensional biological data

3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) was first introduced by Karl Pearson in 1901 where he seeked to represent a high-dimensional system with a "best-fitting" straight line or plane [115]. The method was later independently developed and named as principal components by Harold Hotelling in 1936 [116]. In PCA, we first consider N observations in the $\mathbb{R}^p$ space as an input $N \times p$ matrix $\mathbf{X}$[1]. Here we assume $\bar{x}_i = 0$ for the mean of observations $x_i \in \mathbb{R}^p$, otherwise we replace all observations with their centered versions $\tilde{x}_i = x_i - \bar{x}_i$. The singular value decomposition of $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{3.1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $N \times p$ and $p \times p$ orthogonal matrices with the columns of $\mathbf{U}$ and $\mathbf{V}$ spanning the column and row space of $\mathbf{X}$, respectively. $\mathbf{D}$ is a diagonal matrix with entries of $d_1 \geq d_2 \geq \ldots d_p \geq 0$ which are called the singular values of $\mathbf{X}$. Following from eq. (3.1), we have

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T \tag{3.2}$$

which gives us the eigen decomposition of $\mathbf{X}\mathbf{X}^T$. Since the covariance matrix of the observations are given as $\mathbf{S} = \mathbf{X}^T\mathbf{X}/N$, eq. (3.2) also provides the eigen decompositions of S as well. Here, the eigenvectors v_j (columns of $\mathbf{X}$) are called the principal components directions of $\mathbf{X}$. Importantly, the first principal component defined as $z_1 = \mathbf{X}v_1$, has the largest variance amongst all normalized linear

[1] Here, we basically follow the math introduced in [61].

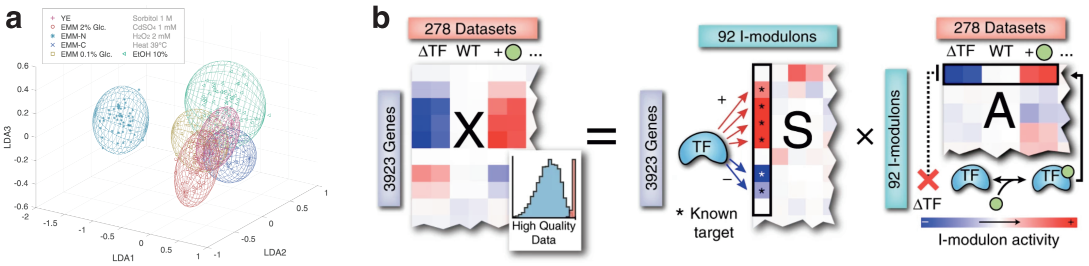

Fig. 3.1 **a**, LDA axes separate *E. coli* populations which have grown in different environments, in the Raman spectra space. Figure reproduced from [118]. **b**, The schematic representing the strategy to recover independent regulons (the set of genes regulated by a given regulator) from transcriptome datasets using ICA. Figure reproduced from [119].

combinations of $\mathbf{X}$ where $\mathrm{Var}(\mathbf{z}_1) = \mathrm{Var}((\mathbf{X}v_1)) = d_1^2/N$. The subsequent principal components $\mathbf{z}_j$ have ordered variance of d_j^2/N and are orthogonal to the earlier ones. It should be noted that linear regression with L2 regularization also projects the data on the principal components and shrinks the components with low-variance more than those with high-variance. Thus, the principal components could be considered as the "best-fitting" linear approximations for the observed data $\mathbf{X}$.

PCA is widely used for reducing the dimensions of high-dimensional biological data. Since it can preserve the majority of variance, it is also often used for preprocessing before applying other methods (for example see [117]). In the context of laboratory evolution, PCA has been used to probe the convergence in high-dimensional transcriptome profiles [93] and stress resistance profiles [60] (see Fig. 2.5). One caveat of PCA is that it does not necessarily provide the optimal projections when trying to predict a independent target vector $y \in \mathbb{R}^p$ from $\mathbf{X}$. We will discuss in details how to obtain a linear combination of features with both high variance and significant correlation with the target y in section 3.2.

3.1.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) provides a linear method for the classification of K classes. LDA models class densities as multivariate Gaussian distributions and it assumes that all classes have a common covariance matrix $\Sigma_k = \Sigma$ for all k. When $K = 2$, the decision boundary provided by LDA corresponds with that provided by linear regression optimized via least squares. Simply put, LDA provides the linear boundaries which separate the K classes the most under the assumption of Gaussian distributed classes with the same covariance. For example, LDA with PCA preprocessing has been used in [118] to find the most discriminatory bases which separate different populations in the Raman spectra space (Fig. 3.1a).

Note that vanilla LDA for $K > 2$ classes do not provide orthogonal axes. It should also be noted that since LDA depends on all of the data, it becomes sub-optimal when the data is distributed in a broad non-Gaussian manner. In such cases, the optimal separating hyperplane could be acquired from linear support vector machines or logistic regression.

3.1.3 Independent Component Analysis

Multivariate data is often viewed as mixed signals from individual latent sources which typically cannot be observed directly. Independent Component Analysis (ICA) aims to recover these latent signals by assuming that the latent sources are statistically independent[2]. To do so, ICA tries to find an orthogonal $\mathbf{A}$ such that the independent sources S could be recovered from the observations X by $S = \mathbf{A}^T X$. This is often done through the minimization of mutual information of $I(S) = I(\mathbf{A}^T X)$ which leads to the identification of orthogonal sources which are far from Gaussian and have the most independence[3]. For example, Sastry *et al* applied ICA to RNA-seq datasets of *E. coli* to find independently regulated modules

[2] When we assume that the latent sources are uncorrelated rather than statistically independent, the latent sources correspond to principal components.

[3] When applying ICA, it is important that the data is pre-whitened so that $\mathrm{Cov}(X) = \mathbf{I}$.

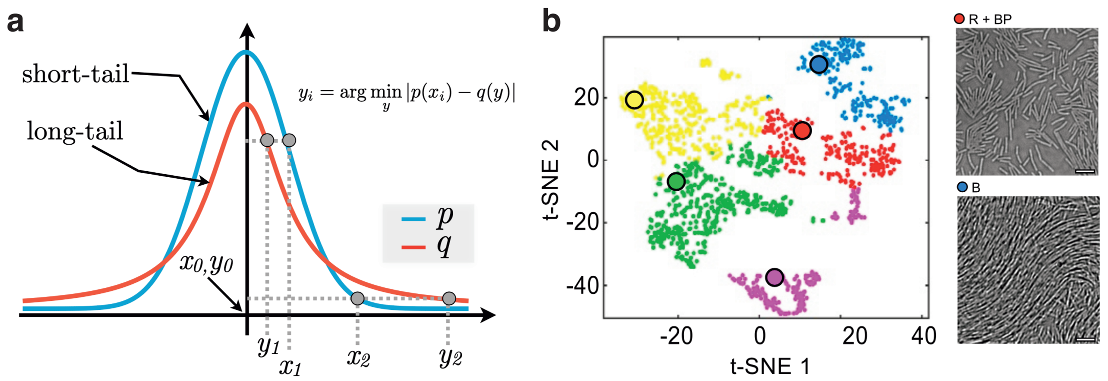

Fig. 3.2 **a**, A schematic illustration of t-SNE. x_i corresponds to the original data points in the high-dimensional space while y_i corresponds to the low-dimensional points embedded by t-SNE. Figure reproduced from [63]. **b**, t-SNE was used to classify five different macroscopic phases within a swarm of *Bacillus subtilis*. Observables such as cell density, aspect ratio, speed, etc. were used for applying t-SNE. Figure reproduced from [122].

in the transcriptional regulatory network (Fig. 3.1b)[4] [119].

3.1.4 Non-linear methods: t-SNE, UMAP, PHATE and others

t-distributed Stochastic Neighbor Embedding (t-SNE), proposed by Maaten and Hinton in 2008 [65], aims to construct a low-dimensional embedding that preserves local neighbor relations while repelling points that were far apart in the original space. In t-SNE, the two-point neighbor relations are associated with the following probability distribution,

$$p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}, \tag{3.3}$$

where $x_i, x_j.x_k \in \mathbb{R}^s$ are the points in the original space and σ_i is a free parameter (usually set by the ‘perplexity’ parameter). The N points in the original space are embedded in a low-dimensional space where the neighbor relations are associated as

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq j}\left(1 + \|y_i - y_k\|^2\right)^{-1}}, \tag{3.4}$$

where $y_i, y_j, y_k \in \mathbb{R}^t$ $(t < s)$ are the points embedded in the low-dimensional space. Importantly, q_{ij} is set to a probability distribution with a long tail which keeps close neighbors even closer while repelling the points that were far apart (Fig. 3.2a). t-SNE tries to find a q_{ij} that has similar neighbor relations to the original space by minimizing the following Kullback-Leibler divergence:

$$KL(p\|q) \equiv \sum_{ij} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right), \tag{3.5}$$

where $p_{ij} = (p_{i|j} + p_{j|i})/2N$ is the symmetrized distribution of $p_{i|j}$.

t-SNE is often used to visualize the high-dimensional structures of the data in a two- or three-dimensional space. For example, Jeckel *et al* applied t-SNE to distinguish five different macroscopic phases in *Bacillus subtilis* swarms (e.g. biofilm phase, high-density rafting phase) via 14 observables extracted from microscope images, in an unsupervised manner (Fig. 3.2b) [122]. t-SNE is also often used to visualize the variation within single-cell RNA-seq data [123, 124].

Although t-SNE is widely used for visualizing high-dimensional data in low-dimensional spaces, it should be noted that the nonlinear transformation in t-SNE makes it hard to interpret the metric and

[4] Recently, the Palsson group has also used Non-negative Matrix Factorization (NMF) for finding potential regulons from transcriptome data of *E. coli* [120]. NMF was also used to find differentially expressed gene regions within single-cell RNA-seq data in [121].

global relations in the low-dimensional space. Recently, alternative methods such as Uniform Manifold Approximation and Projection (UMAP) [125, 126] and Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) [68], are emerging in the field. These methods typically convert transcriptome measurements into a connected graph and then perform mathematical operations (e.g. solve diffusion equations) over this inferred graph [127, 128]. Interestingly, several studies suggest that methods such as UMAP and PHATE tend to preserve both global and local structures of the dataset. However, as long as these methods 'flatten' the original data into low-dimensional representations, there always exists a risk that the structure in the original high-dimensional space gets distorted. Low-dimensional visualizations are convenient, but ideally, they should serve only as aids for other (more powerful) analyses in the high-dimensional space.

3.2 Handling the $p \gg N$ problem

In biological data, it is often the case that the number of features p is much larger than the number of samples N ($p \gg N$)[5]. For such settings, high variance and overfitting becomes a severe problem. One might think that regularization methods such as LASSO (least absolute shrinkage and selection operator) and ridge regression [129, 130] would be effective. Although these typical regularization methods could capture sufficient correlations between the features when $p < N$, the lack of information to efficiently estimate the high-dimensional covariance matrix keeps us from finding the sufficient features in a $p \gg N$ setting (for example, see Fig. 18.1 in [61]). In this section, we will review 'supervised principal components' which is a simple approach to overcome this $p \gg N$ problem. This method will be the key for our analyses in Chapter 4.

3.2.1 Supervised principal components

Supervised principal components was introduced by Bair, Tibshirani and their colleagues [66,67]. The idea of supervised principal components is to find a latent variable that represents the effective degrees of freedom for the target in interest[6]. For example, Bair *et al* aimed to find a linear combination of gene expressions that efficiently describes the latent cell type responsible for lymphoma [67]. The question is, how can we find this latent variable?

Principal component analysis is an effective way to find latent variables that exhibit large variance in the dataset. However, the problem here is that the components with the largest variables are not necessarily the ones that have significant correlation with the target in interest. Supervised principal components seeks to find a linear combination of features that exhibits both large variance and significant correlation with the target. To do so, we use the following algorithm:

1. First, we restrict the model's attention to features that have a sizable correlation with the target. This can be done by calculating univariate regression coefficients for the target as a function of each feature. For example, Bair *et al* used the Cox's proportional hazards regression model to calculate the univariate regression coefficients[7] [66, 67]. This results in p regression coefficients $\theta_1 \dots \theta_p$ for the p features.
2. We next set a threshold θ and compute the first m principal components using only the features that satisfy $\theta_i > \theta$.
3. We finally optimize θ and m through cross validation by using the m principal components in a regression model (e.g. linear regression) to predict the target.

Fig. 3.3a shows the calculated supervised principal component for the survival time of lymphoma patients [67]. The calculated supervised principal component shows good correspondence with the survival time of the patients. Bair *et al* further compare this method with partial least squares (PLS) and

[5] For example, the transcriptome dataset we use in Chapter 4 is a dataset of $p = 4492 \gg N = 192$.

[6] If we use a naive fully supervised approach, we would be able to acquire the genes that have the strongest correlation with the outcome. However, these genes would partially but not perfectly correlate with the underlying effective degrees of freedom that are responsible for the target dynamics. The idea here is that a latent variable as a combination of multiple genes might do a better job in predicting the target dynamics than using individual genes.

[7] This is because Bair *et al* were solving a survival problem in [66,67]. The details of the regression model is not an essential ingredient for the construction of supervised principal components. Different regression models such as random forest regression could be used instead as we will see in Chapter 4.

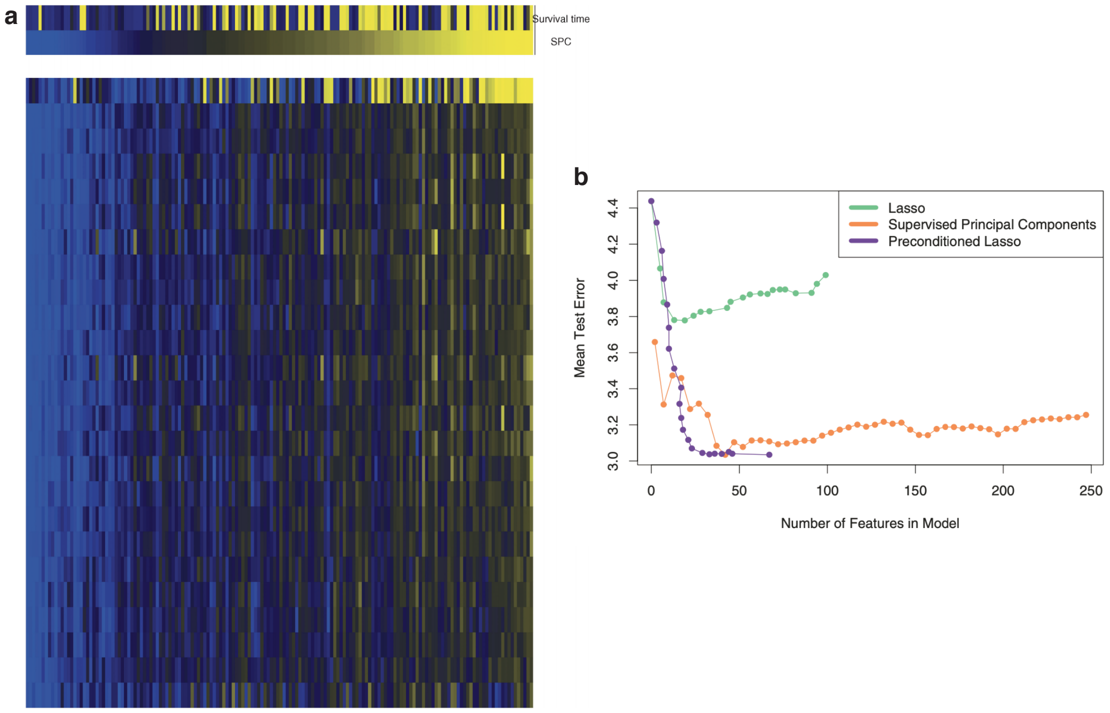

Fig. 3.3 **a**, The survival time of lymphoma patients (top row) and the values of the first supervised principal component (bottom row) for each patient (top panel). The principal component was calculated based on the 25 genes with the top Cox scores. The bottom panel shows the top 25 genes that have significant correlation (high Cox score) with the survival time. Figure reproduced from [67]. **b**, Test errors for the lasso, supervised principal components, pre-conditioned lasso for a simulated dataset with $N = 100$ samples and $p = 5000$ features. Each model is indexed by the number of nonzero features. Figure reproduced from [61].

ridge regression in [67], and show that supervised principal components could significantly reduce the test error while other methods suffer from the very high dimensions under a $p \gg N$ setting. In Fig. 3.3b, the comparison of lasso and supervised principal components is shown for a different simulated dataset [61]. We can see that lasso starts to overfit before it can reach the prediction level of supervised principal components under a $p \gg N$ setting[8].

3.2.2 Random Forest

The original work concerning supervised principal components uses Cox's proportional hazards regression model for calculating univariate regression coefficients for each feature. Since the Cox's regression model is for predicting the survival time for patients, we use the random forest regression model instead in Chapter 4. Here in this section, we give a brief introduction for random forest regression models.

To explain random forests, we must explain the concept of a 'decision tree'. Decision tree models partition the feature space into a set of 'rectangles' and fit a constant in each one of them in order to perform classification/regression. In Fig. 3.4, we show an example of regression using a single decision tree. The maximum depth (`max_depth`) of the tree is the most important hyperparameter. The more deep the tree is, the more vulnerable it becomes to overfitting (see Fig. 3.4d,g). A tree with a maximum depth of one, introduces a partition at $X = 0.403$ and fits -1.322 and 1.108 for the left and right points of the partition, respectively. The position of the partition is computed based on the decrease of the mean squared error (MSE)[9]. If the dataset has multiple features, the decision tree looks for the feature which

[8] Pre-conditioned lasso also shows remarkable results in Fig. 3.3b. In pre-conditioned lasso, we first predict the targets using supervised principal components. We then train a lasso-regularized linear model to predict the predictions given by the supervised principal components. This pre-conditioning using supervised principal components is considered to remove the unneeded noise that typical lasso models tend to overfit on. Pre-conditioned lasso is also useful for reducing the number of features in a $p \gg N$ setting, which is important for interpretational uses. For more details, consult chapter 18 in [61].

[9] For classification tasks, one can use the Gini impurity instead of the MSE.

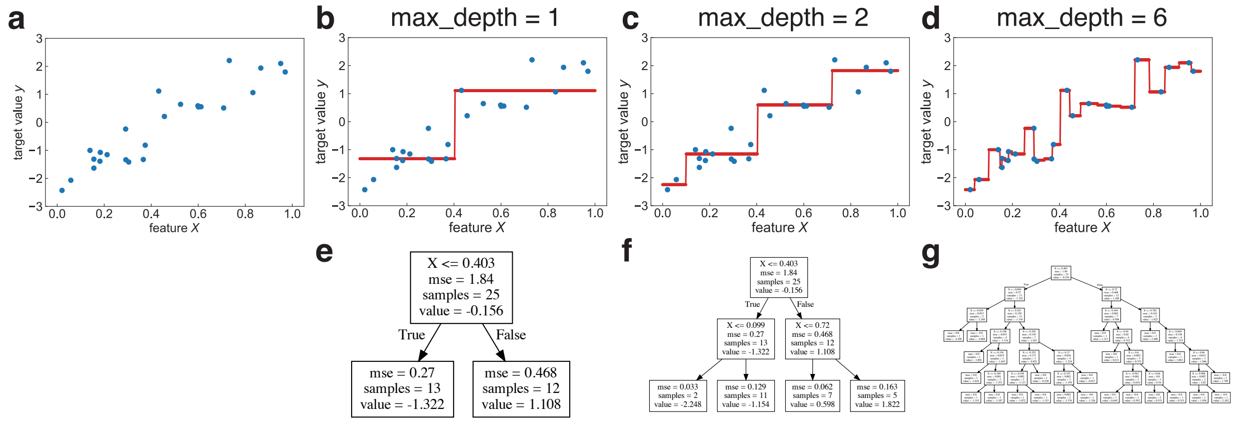

Fig. 3.4 An example of regression based on a decision tree with different depths. **a**, We use randomly distributed points around $y = 4(X - 0.5)$ as target values to predict. **b**,**c**,**d**, The results of regression (red) using a decision tree with maximum depths of 1, 2, 6, respectively. **e**,**f**,**g**, The branching structure for each decision tree model responsible for the results in b, c and d, respectively.

can lower the MSE the most. The decision tree continues to set partitions until (i) the depth of the tree reaches the `max_depth`, (ii) the MSE can no longer decrease[10]. Importantly, the importance of each feature can be calculated through its total contribution to the decrease of MSE. For example, the first branch in the decision tree in Fig. 3.4e reduces the MSE by $36.874 = 25 \times 1.84 - (13 \times 0.27 + 12 \times 0.468)$.

The Random forest model, which is basically an ensemble of decision trees, were introduced by Breiman in 2001 [64][11]. It is based on the idea of bagging or bootstrap aggregation[12] which aims to improve the model's performance by averaging many noisy, but approximately unbiased models. Ensembles work poorly if the noise in the individual models are correlated. Thus, there are basically two randomization processes through the construction of random forests. First, each decision tree is trained on different subsets of the data (bootstrapped data). Second, only a random subset of features could be used at each split of each decision tree. These two randomization processes reduce correlations between the decision trees and leads to an improvement in performance when averaged. This is especially important in the $p \gg N$ setting since there should exist multiple models in the hypothesis space $\mathcal{H}$ that all give the same performance on the training data, and by averaging these models, we would be able to reduce the risk of choosing the wrong hypothesis [132]. As we have seen for individual decision trees, random forest models can also compute the contribution of each feature for prediction (feature importance)[13]. This makes the random forest model a popular choice not only for prediction but also feature selection. In Chapter 3, we will see how the feature selection by random forest models could contribute to downstream analysis (in our case, supervised PCA).

3.3 Artificial neural networks and biological data

The recent advances in artificial neural network models (deep learning) has expanded the range of problems we can deal with biological data. While this big wave of artificial neural networks started from the breakthrough in image classification [133], its applicability now is not limited to image data. Artificial neural networks can now be used in a much broader range of tasks / data domains including and not limited to:

- cell segmentation tasks [134]
- aligning single-cell RNA-seq data with mass cytometry (proteome) data [135] through the utiliza-

[10] As we can see in Fig. 3.4d, decision trees easily overfit when the maximum depth is too large. The maximum depth of the decision tree is a hyperparameter and can be decided through cross validation.

[11] However, the term "random forest" and the essence of building multiple trees in randomly selected subspaces of the feature space, was first introduced in [131].

[12] The term *bootstrap* comes from the saying "pull oneself up with one's bootstrap". Although this saying is often referred to the story in "The Surprising Adventures of Baron Munchausen", there is actually no reference to bootstraps in the Munchausen tales where Baron pulls himself up using his pigtail (and not his bootstrap).

[13] The theoretical characterization of the feature importance in random forests has been performed in [132].

tion of generative adversarial networks (GAN) [136]
- identifying new cell types in single-cell RNA-seq data through meta-learning [137]
- constructing a latent space for shape variations in cultured cells [138] and 3D mandibles [139]
- efficiently sampling statistically independent states of folded proteins from Boltzmann distributions [140]
- identifying mutations that have large biological impact and are also biologically viable [141, 142] using an analogy with natural language processing

and other miscellaneous tasks.

Although neural networks are emerging in a wide range of topics, the preparation of sufficient data still remains as a hurdle to overcome, especially in biology. Recently, self-supervised learning has been suggested as an effective pre-training method for natural language processing [143] and image classification [144, 145]. The idea of self-supervised learning is to utilize unlabeled data to improve task performance when only a few labeled data is available. In image based tasks, for example, self-supervised methods formulate pre-training tasks such as predicting the representation of one part of an image from those of other parts of the image [144, 145]. It has also been shown that these methods can be used as auxillary tasks for regularizing segmentation models in the small labeled data regime [69]. Self-supervised learning might be the key for a broader and more effortless usage of artificial neural networks for biological data.

Chapter 4

Probing evolutionary constraints from high-dimensional multi-omics data

Biological data is often acquired as high dimensional data. However, there is accumulating evidence in the field that hints us the existence of a low dimensional manifold which underlies the dynamics of living things. One of the origins of this low dimensional manifold are understood as evolutionary constraints: biases and limitations of the genotypes or phenotypes that arise in biological systems. In the context of drug resistance evolution, evolutionary constraints appear as cross-resistance/sensitivity relations in the acquired drug resistance profiles for example. These evolutionary constraints are expected to contribute to enabling the prediction and control of microbial evolution. Although previous works have shed light on several constraints for drug resistance acquirement, we still lack a systematic understanding of evolutionary constraints. In this chapter, we will focus on the drug resistance evolution of *Escherichia coli* by analyzing a multi-omics dataset acquired from laboratory evolution in a variety of 48 conditions in total. We will see how the utilization of machine learning such as random forest regression and supervised PCA can contribute to probing the low dimensional manifold of *E. coli*'s phenotypes. We will also discuss how our analyses could contribute to deepen our understandings of evolutionary constraints.

Related publications by author:
Tomoya Maeda*, Junichiro Iwasawa*, Hazuki Kotani, Natsue Sakata, Masako Kawada, Takaaki Horinouchi, Aki Sakai, Kumi Tanabe, and Chikara Furusawa, "High-throughput laboratory evolution reveals evolutionary constraints in *Escherichia coli*". *Nature Communications* **11**, 5970 (2020). *bioRxiv* doi: 10.1101/2020.02.19.956177 (*co-first authors)

Contribution:
The author (J.I.) conducted all data analyses under the advice and supervision of C.F.. Biological interpretation of the analyses was done by J.I., T.M., and C.F.. All experiments and data acquisition were performed by T.M. with the support of H.K., N.S., M.K., T.H., A.S., and K.T., under the supervision of C.F..

4.1 Introduction

The emergence of antibiotic resistance and multidrug-resistant bacteria is a growing global health concern [51, 52, 79]. Since bacteria can easily acquire resistance even to novel drugs, the development of novel antibiotics is not necessarily an effective approach to combatting antibiotic resistance [53, 146]. Thus, it is important to understand the mechanism itself of antibiotic (or stress) resistance evolution [55–59, 147]. One promising approach to understand the mechanism to evolution is to identify the evolutionary constraints, "a bias or limitation in genotypic or phenotypic variation that a biological system produces" [148], which shape resistance evolution [46, 52]. Although genotypic/phenotypic variations have a high dimensional nature, evolutionary constraints could keep the dynamics on a low dimensional manifold within the high dimensional space, making the evolutionary dynamics predictable and maybe even controllable [49, 78, 81, 149]. Notably, previous studies have revealed that the acquisition of resistance to a certain drug is often associated with resistance/sensitivity to a different drug which is coined as cross resistance / collateral sensitivity, and that such phenomena could be widely observed among different drugs [55–60]. As one can easily imagine, this interconnectedness of cross-resistances and collateral sensitivities could be interpreted as the scars of evolutionary constraints on accessible phenotypes. Thus, it is crucial to reveal the network of cross-resistances and collateral sensitivities over a wide variety of drugs (or stresses) and elucidate the biological mechanisms which underlie such networks to understand evolutionary constraints[1]. In other words, the investigation of the relations between the phenotypic space describing the cross resistance / collateral sensitivity networks (i.e. the resistance space) and the space describing the internal biological processes (i.e. the gene expression space, genotypic space) are crucial for understanding drug resistance evolution (Fig. 4.1). If we could find a latent space in the gene expression space which corresponds to the low dimensional dynamics in the resistance space (i.e. a gene expression – resistance mapping), we would be able to investigate the origins of evolutionary constraints which would lead to the basis for making an effective framework for predicting evolution.

However, elucidating the relations between the resistance space and the gene expression space leads to a two-fold problem. One is that the variety of constraints underlying drug resistance evolution would be hindered if we only investigate the resistance acquisition network based on a limited number of stresses. Although an extensive exploration of the network of cross-resistances / collateral sensitivities is crucial for investigating evolutionary constraints, we still lack such exploration over a wide range of stresses mainly due to the experimental cost. Another problem is that when we acquire omics data (e.g. transcriptome data) to investigate the relation between resistance acquisition and the underlying biological processes, it is usually difficult to find corresponding signals in the omics data due to the $p \gg N$ problem where the dimensionality of the dataset p is much larger than the number of samples N [61]. Unless we figure out how to find the signals in the high dimensionality dataset, we will not be able to find an effective latent space for investigating evolutionary constraints.

In this chapter, we will provide an approach to the problems above by combining high-throughput

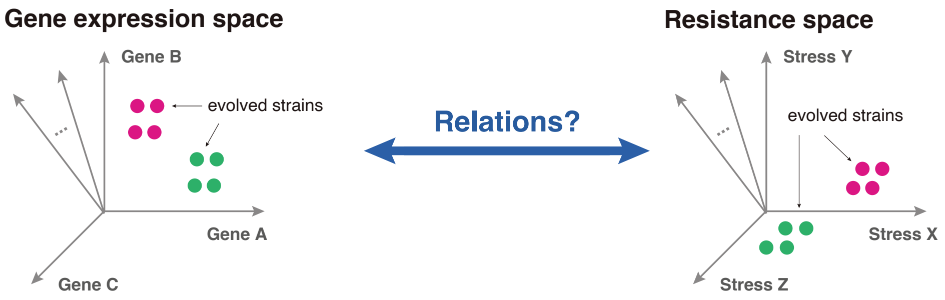

Fig. 4.1 One of the central questions in this chapter is the following: How can we relate the gene expression space to the stress resistance space?

[1] The understanding of the network of cross-resistances and collateral sensitivities are important not only for the understanding of evolutionary constraints, but also for clinical reasons. For example, the cyclic use of two drugs with collateral sensitivity were demonstrated to suppress resistance evolution [56]. Thus, the investigation of drug interactions might provide alternative strategies for such clinical approaches to antibiotic resistance.

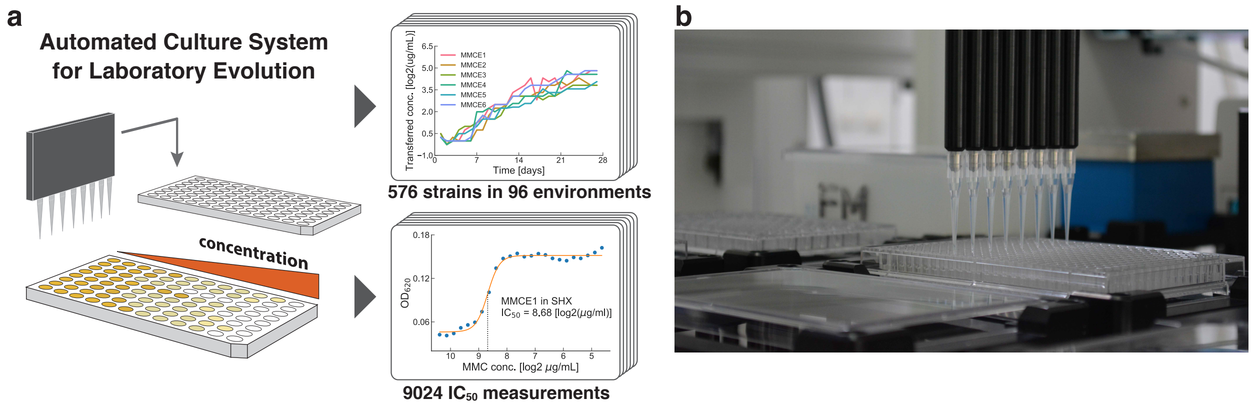

Fig. 4.2 **a**, Schematic image of the automated culture system for laboratory evolution [151]. Examples for the time series for resistance evolution and resistance (IC_{50}) measurements are shown together. **b**, A photograph of the automated culture system taken by Junichiro Iwasawa in RIKEN, Osaka. **a** was reproduced from [150].

laboratory evolution and machine learning based analyses. We first performed high-throughput laboratory evolution by using an automated culture system which allowed us to explore drug resistant phenotypes in a unprecedented scale. In details, we performed laboratory evolution of *Escherichia coli* under 47 stressors with a wide variety of action mechanisms (Fig. 4.2, Table 4.1)[2]. For each of the evolved strains, we collected their genome, transcriptome[3], and resistance profiles resulting to a multi-omics dataset. Next, to avoid the $p \gg N$ problem, we utilized a method called "Supervised PCA" [61, 66, 67] which allowed us to extract a latent space in the gene expression space that maximizes the dependency between gene expression and stress resistance. Through the analysis using supervised PCA, we were able to find low dimensional structures within the distribution of the evolved strains, suggesting the existence of constraints underlying resistance evolution dynamics. We further validate our analyses of the low dimensional structures by reconstructing commonly observed mutations in the parent strain. We also report an interesting phenomenon coined as "decelerated evolution", in which the resistance of the evolved strains under β-lactams are overtaken by strains evolved in different stressors. Our analyses show how that the combination of high-throughput laboratory evolution and machine learning can open new avenues to the investigation of evolutionary constraints.

4.2 Basic statistics of the dataset acquired from high-throughput laboratory evolution

High-throughput laboratory evolution was performed using an automated culture system (Fig. 4.2, 4.3) [151] for 47 stressors covering a wide range of action mechanisms (Table 4.1, Fig. 4.4a) to systematically investigate drug-resistant phenotypes. Originally, six independent culture lines were propagated in parallel for each stressor to evaluate the reproducibility of the stress resistance evolution dynamics. In total, 288 independent culture series were maintained (47 stressors plus a control without any stressor × six replicates) for 27 daily passages corresponding to approximately 250–280 generations[4]. Figures 4.2a

[2] In the original work [150], laboratory evolution was performed under 95 stressors. However, due to the limitation in experimental capacity, we have picked up 47 stressors for further analysis. In this dissertation, we focus on the 47 stressors from the very beginning to avoid confusion.

[3] It should be noted that although there exist several previous works investigating drug resistance using laboratory evolution, most just collect the drug resistance profiles with the evolved strains' genotypes, and not gene expression profiles. However, as other studies of microbial populations suggest, phenotypic changes such as changes in regulatory or metabolic pathways are much more repeatable than changes in the genotype space [49]. The investigation of these repeated phenotypic changes is what we really need when we desire to build a framework for predicting evolution, and thus, the collection of the transcriptome is an essential feature of our study.

[4] Here, the initial OD is set to $\mathrm{OD} = 0.00015$ which corresponds to $\sim 7,500$ cells per 50 μL. In the current system, the minimum threshold for transferring cells is set to $\mathrm{OD} = 0.09$ which corresponds to a 600-fold growth of cells in 24h. On the other hand, the OD of full growth is $\mathrm{OD} = 0.2$ which corresponds to a 1333-fold growth of cells in 24h. Therefore, the generations of cells N_{gen} could be bound as, $27 \times \log_2 600 < N_{\mathrm{gen}} < 27 \times \log_2 1333 \Rightarrow 249 < N_{\mathrm{gen}} < 280$.

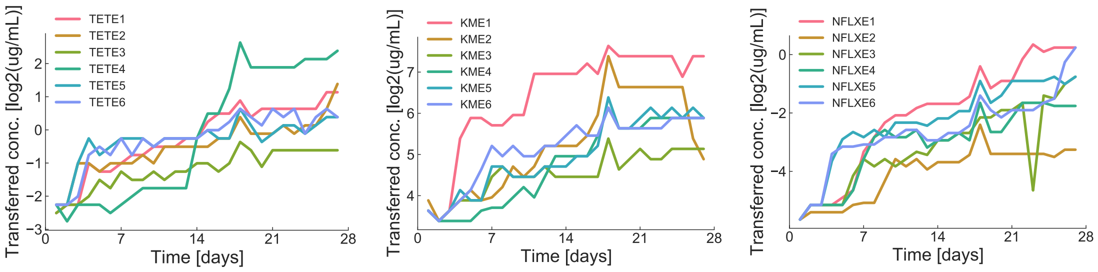

Fig. 4.3 Time series for the resistance over the 27 days of the laboratory evolution experiment for tetracyclin (TET), kanamycin (KM) and norfloxacin (NFLX). The time series for all 48 environments are shown in Fig. 4.23 and 4.24. Figures were modified and reproduced from [150].

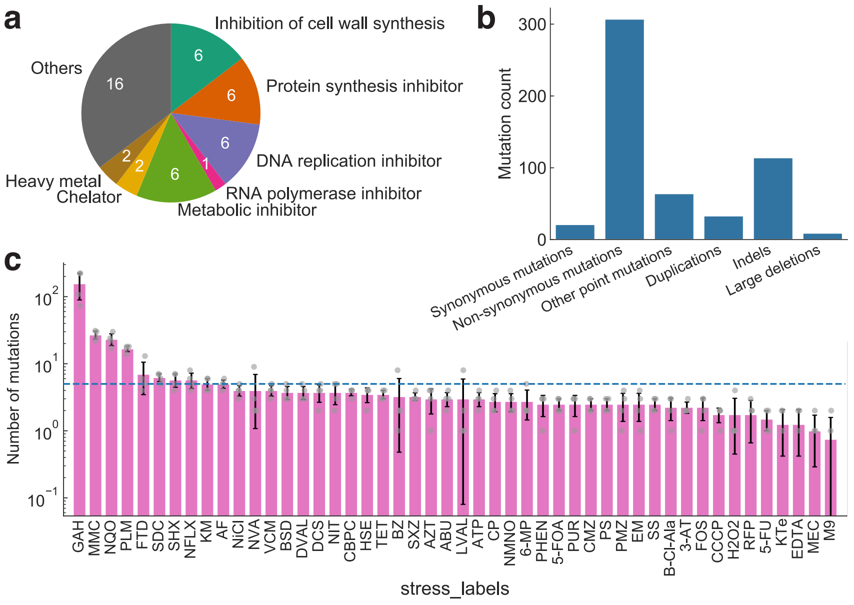

Fig. 4.4 **a**, Categories of the mechanism of action for the 47 stresses selected for resistance (IC$_{50}$) measurements. The 47 stresses are given in Table 4.1. **b**, Distribution of mutation events for the evolved strains according to its mutation type, except for strains evolved in GAH, NQO, and MMC. Other point mutations include mutations in intergenic/noncoding regions. **c**, The mean number of identified mutations for the four evolved strains in each environment are shown. The error bars represent the standard deviation. The blue dashed line is a guide to the eye showing the level of five mutations. Figures reproduced from [150].

and 4.3 show examples of the time course of transferred concentrations (which could be considered close to the minimal inhibitory concentration, MIC) during laboratory evolution. Here, the strains are named by combining the abbreviations of the stressor (e.g. MMC) and its replication number (E1 – E6). The time courses of transferred concentrations for all stressors are shown in Fig. 4.23, 4.24. Among the 47 stressors, a significant increase in the transferred concentrations was observed for all 47 stressors (Mann-Whitney U-test, false discovery rate (FDR) < 5%). For further phenotypic and genotypic analyses, the top four evolved strains showing higher IC$_{50}$values among the six, isolated from each stressor (and the control), resulting to $(47 + 1) \times 4 = 192$ evolved strains, were selected.

4.2.1 Genotypic changes

To investigate genetic alterations underlying the observed resistance, genome resequencing analyses of the 192 evolved strains were performed (Fig. 4.4). To show the variation in the number of mutations among strains evolved in the same environment, the mean number and standard deviation of mutations for the four strains are shown in Fig. 4.4c, where we can see that most of the strains had less than five mutations. In fact, 147/192 (76.6%) of the evolved strains had less than five mutations. Among the 47 stressors, the highest number of mutations were observed in stresses such as glutamic acid γ-hydrazide

Abbreviation	Stress name	Biological target
5-FOA	5-Fluoroorotic Acid Monohydrate	DNA
5-FU	5-Fluorouracil	DNA
6-MP	6-Mercaptopurine monohydrate	DNA
ABU	DL-2-Aminobutyric acid	Unknown in bacteria
AF	Acriflavine	DNA
ATP	Amitriptyline Hydrochloride	Unknown in bacteria
AZT	Aztreonam	Peptidoglycan
BSD	Blasticidine S hydrochloride	Protein translation
BZ	Benserazide Hydrochloride	Aromatic-L-amino-acid
CBPC	Carbenicillin Sodium Salt	Peptidoglycan
CCCP	Carbonyl cyanide 3-chlorophenylhydrazone	Oxidative phosphorylation
CMZ	Cefmetazole sodium salt	Peptidoglycan
CP	Chloramphenicol	50S ribosome
DCS	D-Cycloserine	Peptidoglycan
DVAL	D-Valine	Extracellular polysaccharide
EM	Erythromycin	50S ribosome
FOS	Fosfomycin disodium salt	Peptidoglycan
GAH	L-Glutamic acid gamma-hydrazide	Glutamate decarboxylase
H2O2	Hydrogen peroxide	Oxidative stress
HSE	L-Homoserine	Glutamate dehydrogenase
KM	Kanamycin Sulfate	30S ribosome
K2TeO3	Potassium Tellurite (IV)	Oxidative stress
MEC	Mecillinam	Peptidoglycan
MMC	Mitomycin C	DNA
NFLX	Norfloxacin	DNA gyrase
NiCl2	Nickel(II) Chloride	Oxidative stress
NMNO	N-methyl-N-octylamine	Unknown in bacteria
NQO	4-Nitroquinoline 1-oxide	DNA
NVA	DL-3-hydroxynorvaline	Aspartate and homoserine kinase
PLM	Phleomycin	DNA
PMZ	Promethazine Hydrochloride	Histamine H1 receptor
PUR	Puromycin Dihydrochloride	Protein translation
RFP	Rifampicin	RNA polymerase
SDC	Sodium Dichromate Dihydrate	Oxidative DNA damage
SHX	DL-Serine hydroxamate	Serine-tRNA ligase
SS	Sodium salicylate	Metal chelater
SXZ	Sulfisoxazole	Folic acid biosynthesis
TET	Tetracycline	30S ribosome
VCM	Vancomycin Hydrochloride	Peptidoglycan

Table 4.1 Full names of the stresses used in this study. This table was reproduced from the supplementary information of [150].

(GAH) with 157 ± 67 mutations, 4-nitroquinoline-1-oxide (NQO) with 23 ± 5 mutations, and mitomycin C (MMC) with 27 ± 4 mutations. Although GAH has not been previously recognized as a mutagen, the results show that GAH has a higher mutagenic activity than known mutagens such as NQO and MMC. To estimate the ratio of beneficial mutations among the evolved strains, we calculated the ratio of nonsynonymous to synonymous mutations per site ($\mathrm{d}N/\mathrm{d}S$) for the evolved strains. Excluding the strains which evolved in GAH, NQO, and MMC, 21 and 307 mutations were identified as synonymous and nonsynonymous mutations, respectively (Fig. 4.4b)[5]. For strains evolved in stresses other than GAH, NQO and MMC, the ratio of nonsynonymous to synonymous mutations per site was 5.26, implying that

[5] Here, we excluded the strains which evolved in GAH, NQO, and MMC, since the inclusion of mutagens leads to a lower-biased estimate of $\mathrm{d}N/\mathrm{d}S$.

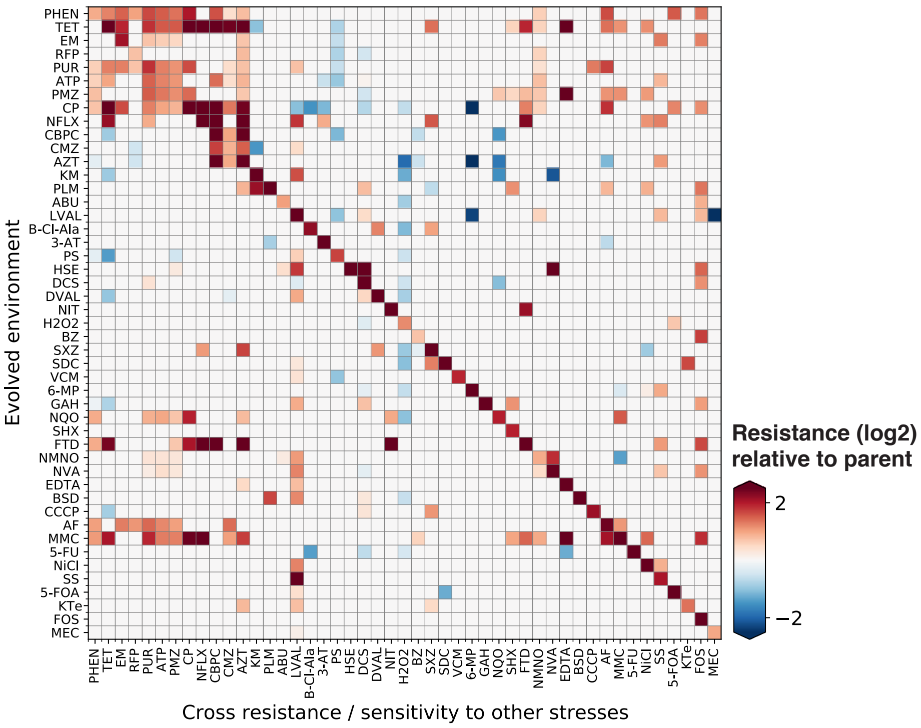

Fig. 4.5 Identified combinations of stresses that exhibited either cross-resistance or collateral sensitivity for each of the four strains which evolved in the same environment. Figure reproduced from [150].

approximately 80% of the nonsynonymous mutations were beneficial[6].

4.2.2 Stress-pairs exhibiting cross-resistance / collateral sensitivity

The exploration of stress pairs exhibiting cross-resistance / collateral sensitivity, a phenomenon where an evolved strain in a certain stress gains resistance / sensitivity to another stress, is important for clinical purposes such as drug cycling [56]. It is also important for the control of stress resistance evolution since cross resistances and collateral sensitivities can be considered as scars of evolutionary constraints. The changes in stress resistance profiles were quantified by measuring the half-maximal inhibitory concentration (IC_{50})[7] of all 47 chemicals for each evolved strain (9024 measurements in total), relative to the parent strain (details of this quantification are given in **??**). These IC_{50}measurements allowed us to study how common cross-resistance / collateral sensitivity occur. By comparing the four strains evolved in the same stress and 13 independent IC_{50}measurements of the parent MDS42 strain, we found that 336 and 157 pairs of stressors exhibited cross-resistance and collateral sensitivity, respectively, within the possible 2162 combinations (Mann-Whitney U-test, double-sided, FDR < 5%, Fig. 4.5). These information provide a basis for the prediction and control of stress resistance evolution. However, it should be noted that the measurements here are based on the strains evolved only from a single genetic background (the parent MDS42 strain). What is really needed is an extensive investigation of stress pairs exhibiting cross-resistance / collateral sensitivity based on strains evolved from different genetic backgrounds[8]. We will discuss this in more details in the next chapter.

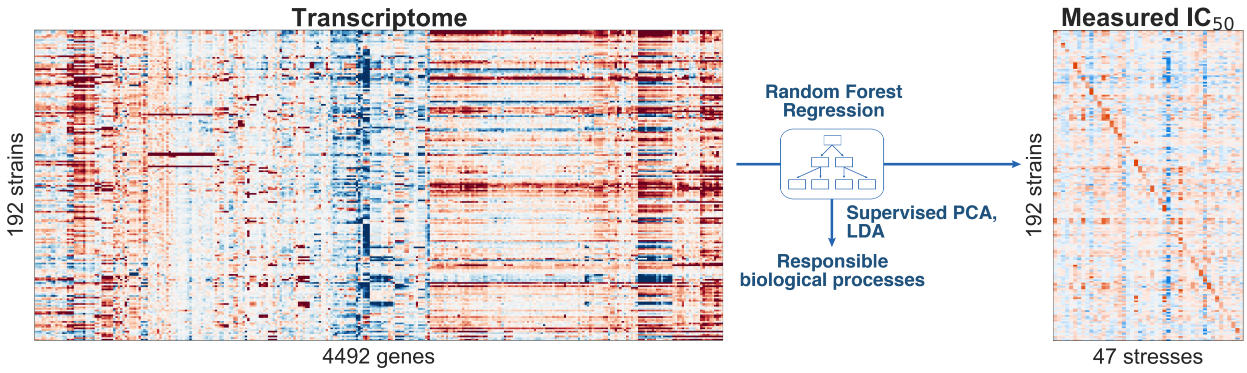

Fig. 4.6 Schematic image of the procedure of the analyses in this study. The IC_{50}changes were predicted through the random forest regression model using the gene expression profiles. The genes with high importance for the regression model were used for Supervised PCA and LDA for extracting the underlying biological processes for stress resistance acquisition.

4.3 Supervised PCA bridges gene expression and stress resistance

To investigate the changes in gene expression which underlie stress resistance acquisition, we performed transcriptome analysis on the 192 evolved strains (for the whole dataset, see Supplementary Data 4 in [150]). Note, all evolved strains were cultured without addition of stressors to standardize the culture condition for the transcriptome analysis[9]. To explore the latent space in the gene expression space which corresponds to resistance evolution, we will perform dimension reduction on the gene expression data using Supervised Principal Component Analysis (PCA) [66, 67]. As we have seen in Chapter 3, Supervised PCA allows us to extract a subspace in which the dependency between gene expression and stress resistance is maximized, which is not assured by vanilla PCA.

4.3.1 Preprocessing for Supervised PCA using random forest regression

The essence of supervised PCA is to construct a latent space that exhibits both large variance and significant correlations with the target in interest (e.g. stress resistance). In order to do so, we need to estimate the correlation between each feature and stress resistance and restrict the attention of PCA to features that have a sizable correlation. We thus constructed a random forest regression model[10] that predicts the relative IC_{50}s for each of the 47 stresses from the 4,492 $\log_{10}$-transformed gene expression levels for all 192 evolved strains (Fig. 4.6, 4.7) [11]. Because the changes in IC_{50}varied within the 47 stresses, we normalized the IC_{50}changes by multiplying 1, 0.5, or 0.25 depending on the maximum fold

[6] The ratio of beneficial mutations of the evolved strains could be estimated through the ratio of nonsynonymous to synonymous mutations per site ($\mathrm{d}N/\mathrm{d}S$) which was calculated as 5.26 for the current experiment [54]. Under the assumption that $\mathrm{d}N/\mathrm{d}S$ should be 1.0 under strict neutrality, the ratio of beneficial mutations y can be calculated as $y = (5.26-1.0)/5.26 = 0.810$.

[7] The genotypic and phenotypic analyses here are performed against a isolated single clone. It has been confirmed that the IC_{50}s of the single clones were nearly identical to that of the corresponding endpoint cultures. In fact the mean of IC_{50}(isolated clones) - IC_{50}(endpoint cultures) was 0.18 ± 0.22 (95% confidence interval).

[8] A pioneering work in this direction was recently conducted in [95].

[9] Of course, the expression of some genes would only be induced or suppressed in the presence of a stressor. In this study, we neglected the environment-dependent expression changes and collected the gene expression profiles in the no-drug condition, to compare expression profiles of the evolved strains without environmental-dependent biases. However, it would be interesting to collect the gene expression profiles under various environmental conditions, e.g., 192 strains × 47 stress environments = 9024 conditions, to unveil both the environment-specific regulatory responses and their evolution

[10] Ensembling methods including the random forest model are now one of the most powerful tools in machine learning. The reasons underlying this success is explained by Louppe in his PhD dissertation as follows [132]:

> "The first reason is statistical. When the learning set is too small, a learning algorithm can typically find several models in the hypothesis space $\mathcal{H}$ that all give the same performance on the training data. Provided their predictions are uncorrelated, averaging several models reduces the risk of choosing the wrong hypothesis. The second reason is computational. Many learning algorithms rely on some greedy assumption or local search that may get stuck in local optima. As such, an ensemble made of individual models built from many different starting points may provide a better approximation of the true unknown function than any of the single models. Finally, the third reason is representational. In most cases, for a learning set of finite size, the true function cannot be represented by any of the candidate models in $\mathcal{H}$. By combining several models in an ensemble, it may be possible to expand the space of representable functions and to better model the true function."

[11] Here, we used the scikit-learn implementation `sklearn.ensemble.RandomForestRegressor`.

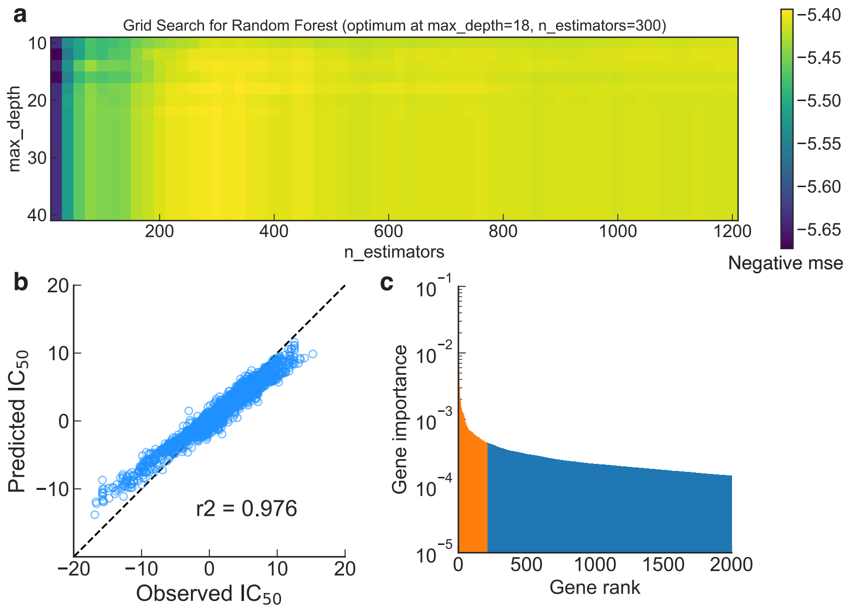

Fig. 4.7 **a**, Negative mean squared error from the 4-fold cross validation for the random forest model. Two hyperparameters (`n_estimators`: number of trees, `max_depth`: max depth of each tree in the forest) were optimized through a grid search. The following values for hyperparameters were obtained: `n_estimators` $= 300$, `max_depth` $= 18$. **b**, The predicted IC_{50} values predicted from the whole gene expression dataset using the random forest with `n_estimators` $= 300$, `max_depth` $= 18$. **c**, The sorted gene importance levels computed through the random forest model. The top 213 genes (orange) were used for supervised PCA. **c** was reproduced from [150].

changes of the IC_{50}s[12]. To avoid overfitting, a grid search over the number of trees (16 values between 10 and 40) and the max depth of each tree (60 values between 20 and 1,200) was performed using a 4-fold cross validation method (Fig. 4.7a). The set of hyperparameters, (number of trees, max depth) = (300, 18), which provided the lowest mean squared prediction error averaged over the 4-fold validation sets, was selected for further analysis. Using these hyperparameters, we trained the random forest regression model using the 4,492 gene expression levels as features and the resistance changes for all 47 stresses as targets. Through this procedure, we were able to extract the feature importance for each gene (Fig. 4.25)[13]. The number of genes to use in the latter PCA process were coarsely chosen by looking at the inflection point of the decay curve for feature importance (Fig. 4.7b). To fine-tune the number of genes N to use, we calculated the class dissimilarity $W_1 5$ (explained below) in the 47 dimensional resistance space based on the expression profiles of the top N genes with high feature importance (Fig. 4.12c).

4.3.2 Performing Supervised PCA

Supervised PCA and hierarchical clustering

To explore the latent space for stress resistance, we performed Supervised PCA [66, 67] based on the 213 genes extracted by the random forest regression model[14]. Supervised PCA based on the expressions of these genes revealed the existence of clusters of evolved strains in the dimension reduced gene expression space (Fig. 4.8). Especially, when observing the supervised PCA space through tSNE [65] and PHATE [68], the existence of several distinct clusters can be clearly observed (Fig. 4.9). To clarify these clusters, we performed hierarchical clustering in the supervised PCA space. The nontrivial procedure of hierarchical clustering is how to define the resulting number of clusters. One heuristic method is called the elbow method where you compute the class dissimilarity W_n for each number of

[12] This normalization corresponds to the chemical gradient step size used for the determination for IC_{50}changes.

[13] Here, the feature importance (gene importance) is evaluated by the decrease of the mean squared prediction error at each branch of each decision tree through the feature_importance attribute of the `RandomForestRegressor` function. The 213 genes and their feature importances computed by the random forest model is given in Fig. 4.25.

[14] Note that in the original work, feature selection prior to PCA was performed through the Cox score which measures the correlation between the gene's expression and patient survival.

classes and choose the number of clusters n where $W_n - W_{n-1}$ takes its maximum. Here, W_n is defined as

$$W_n = \sum_k \sum_{i \in c_k} |r_i - \mu_{c_k}|^2 , \tag{4.1}$$

where k, c_k, μ_{c_k} is the class's index, the set of elements for each class, and class k's centroid, respectively, and r_i, i represents the location of each strain, and its index, respectively. For each number of classes n, we calculated W_n, and its derivative ($W_n - W_{n-1}$) in the 47 dimensional resistance space and searched for the number of classes where $W_n - W_{n-1}$ sharply decreased (Fig. 4.12a,b)[15]. As a result, the optimal number of classes was determined to be 15. The results of hierarchical clustering including all 15 classes are given in Fig. 4.10 and a simplified version omitting one class and three singletons are shown in Fig. 4.11.

Finding genes underlying the modular classes via LDA

Hierarchical clustering in the supervised PCA space resulted in the elucidation of modular classes of expression profiles (Fig. 4.10, 4.11,)[16]. Strikingly, strains in the same class were those that did not necessarily evolve in the same stress, nor stress category. Similar phenotypic convergence of drug-resistant strains has previously been observed in clinically isolated strains of *Pseudomonas aeruginosa* [60]. To elucidate characteristic gene expression for each class, we applied linear discriminant analysis (LDA), which allowed us to extract the most discriminative set of genes for each class, through the observation of each decision boundary (Fig. 4.10b, 4.11b). Here, LDA was performed by using the `LinearDiscriminantAnalysis` function from the scikit-learn package. The strains were given binary labels for LDA: one for the strains which belonged to the class of interest, and zero for the other strains. To extract the important genes which characterized each class, we looked for the top weighted genes in each LDA axis, which corresponded to the genes which contributed to the decision boundary for the binary labeled strains. Due to the nature of LDA, the top weighted genes sometimes showed the characteristic genes not for the class in interest, but for the other classes. Thus, we further selected the genes which had more than a two-fold change in gene expression compared with the parent strain, within the top weighted genes in each LDA axis.

Evaluating the correspondence between the Supervised PCA space and Stress resistance space

To investigate how the classes of gene expression profiles correspond to stress resistance, we observed the relative $\mathrm{IC}_{50}$for each of the 47 stresses of each evolved strain sorted based on the hierarchical clustering in the supervised PCA space (Fig. 4.10c, 4.11c). As shown in the figures, the classes in the supervised PCA space correspond well with the stress resistance patterns. To quantitatively evaluate the correspondence between the resistance space and the supervised PCA space[17], we used the class dissimilarity (W_n) measure. In details, we computed W_{15} in the resistance (IC_{50}) space based on the clustering results in the resistance space[18], supervised PCA space, genotypic (mutation) space, and the whole expression space, respectively (Fig. 4.13a)[19]. W_n is a measure of the sum of "compactness" of each cluster in the space of interest. Because W_{15} in Fig. 4.13a is measured in the IC_{50}space, it is natural that W_{15} based on the IC_{50}space takes a minimum among the others. Interestingly, W_{15} based on the supervised PCA space takes the next smallest value, which is smaller than that based on the genotypic space nor the whole expression space. This suggests that the supervised PCA space is offering good representations in the expression space that corresponds well with the resistance space. In other words, the topological relationships between the strains in the resistance space could be accurately represented

[15] Because we wanted to explore the underlying biological mechanisms for resistance acquisition in detail, we limited the lower bound of n to 10. When we computed W_n for $n > 0$, $W_n - W_{n-1}$ took its maximum at $n = 8$.

[16] Here, hierarchical clustering was applied to the 36 dimensional supervised PCA space which corresponds to 90% of the total variance. The Ward's method was used for clustering.

[17] Here, we considered the neighboring (or topological) relations in each space as a measure of correspondence.

[18] For the resistance space, hierarchical clustering was applied based on the 47 relative IC_{50}s, to cluster the 192 strains to 15 classes.

[19] For the resistance space, hierarchical clustering was applied based on the 47 relative IC_{50}s, to cluster the 192 strains to 15 classes. For the mutation space, hierarchical clustering was applied based on the one-hot encoding which reflects the information of the presence of a mutation. For the expression space, hierarchical clustering was applied to the whole 4,492-dimension gene expression space. To construct a baseline, the class dissimilarity was calculated for randomly clustered classes in the resistance space as well.

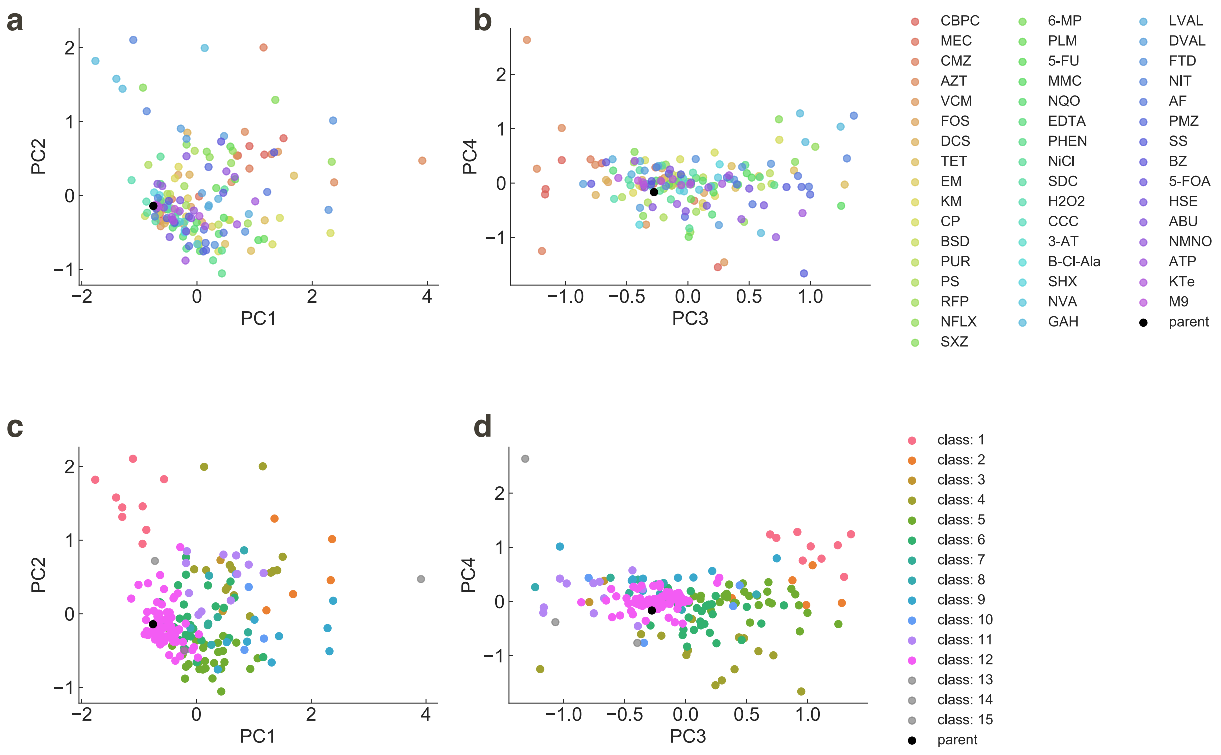

Fig. 4.8 **a,b**, Distribution of the 192 evolved strains and the parent strain in the supervised PCA space. The four principal components are shown and the colors denote the evolved environment for each strain. **c,d**, The same data as in **a,b** with colors denoting the classes defined by hierarchical clustering. Figures reproduced from [150].

in a subspace of the gene expression space[20].

One might wonder what happens if we performed our analyses based on the results of hierarchical clustering in the resistance space. Because we are looking for a representation that bridges the gene expression space and the resistance space, clustering directly in the resistance space might also provide us nice representations. However, direct clustering in the resistance actually breaks apart some of the important gene expression clusters (e.g. the cluster with high *acrR* expression, the cluster with high *prlF* expression) (Fig. 4.26). We can also see that direct clustering in the resistance space is not a good strategy through the calculation of W_{15} measured in the whole gene expression space (Fig. 4.13b). Here, we computed W_{15} in the whole gene expression space based on the clustering results in the whole expression space, supervised PCA space, resistance space, and the genotypic (mutation) space, respectively. We can see that although clustering in the supervised PCA space leads to low W_{15} close to that of the whole expression space, W_{15} based on clustering in the resistance space leads to a high value. These results might suggest that the effective degrees of freedom in the resistance space is not sufficient to recover necessary information in the gene expression space, while the supervised PCA space obtains good representations conserving information for both the expression and resistance space[21]. Overall, our results suggest that the supervised PCA space, preserving sufficient information, bridges the gene expression space and the stress resistance space.

[20] Of course, the results of Fig. 4.13a might differ slightly when we have more experimental samples. It is important to be aware that the awful results of W_{15} based on the genotypic space and the whole expression space partially stems from the $p << N$ problem, where the high dimensionality of the data hinders important signals of the data. However, as in [59, 93], there is growing evidence in the field that microbial evolution can be well represented in a low dimensional space, and we can consider this result as a consequence of the manifold hypothesis.

[21] Indeed, when we performed PCA for the IC_{50}space, supervised PCA space, and the whole expression space, the number of dimensions corresponding to 90% variance was 18, 36, and 70, respectively.

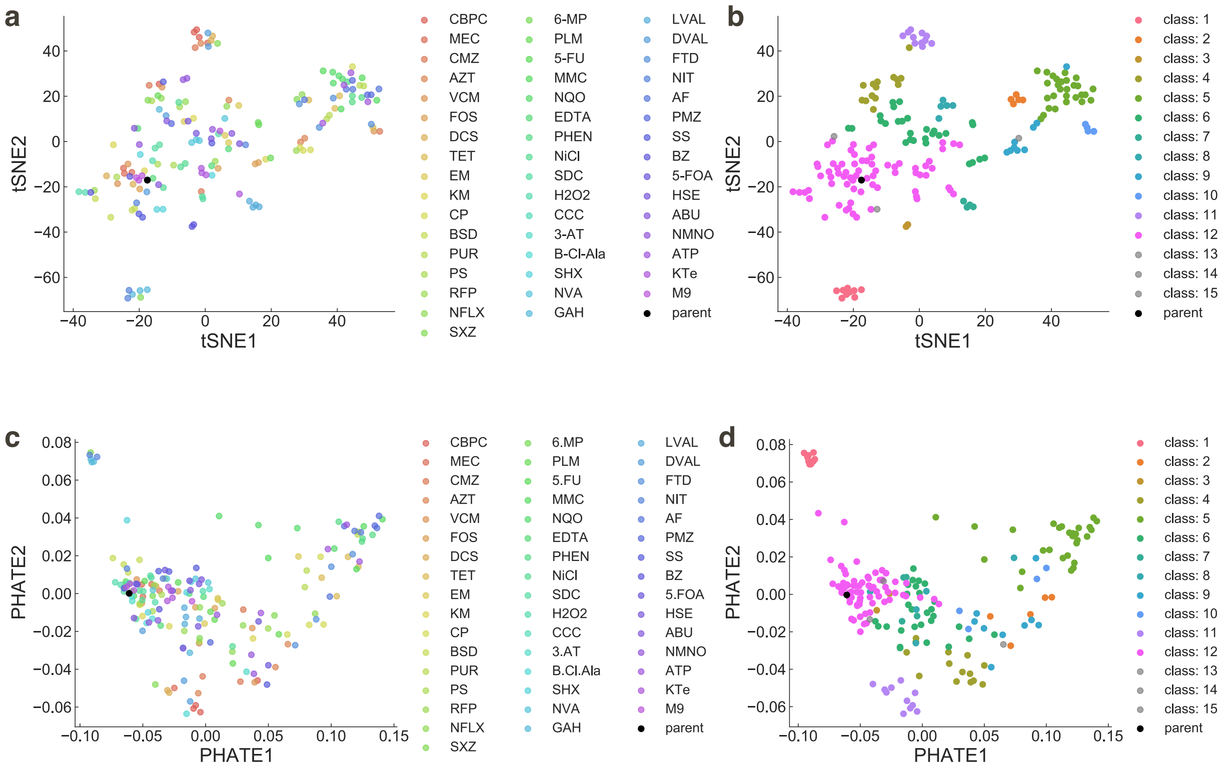

Fig. 4.9 **a**, The distribution of the evolved strains in the 36 dimension supervised PCA space embedded to a two-dimensional space using tSNE. Colors denote the evolved environment. **b**, The same data as in **a** with colors denoting the classes defined by hierarchical clustering. **c**, The distribution of the evolved strains in the 36 dimension supervised PCA space embedded to a two-dimensional space using PHATE. Colors denote the evolved environment. **d**, The same data as in **c** with colors denoting the classes defined by hierarchical clustering. **a**, **b** were reproduced from [150].

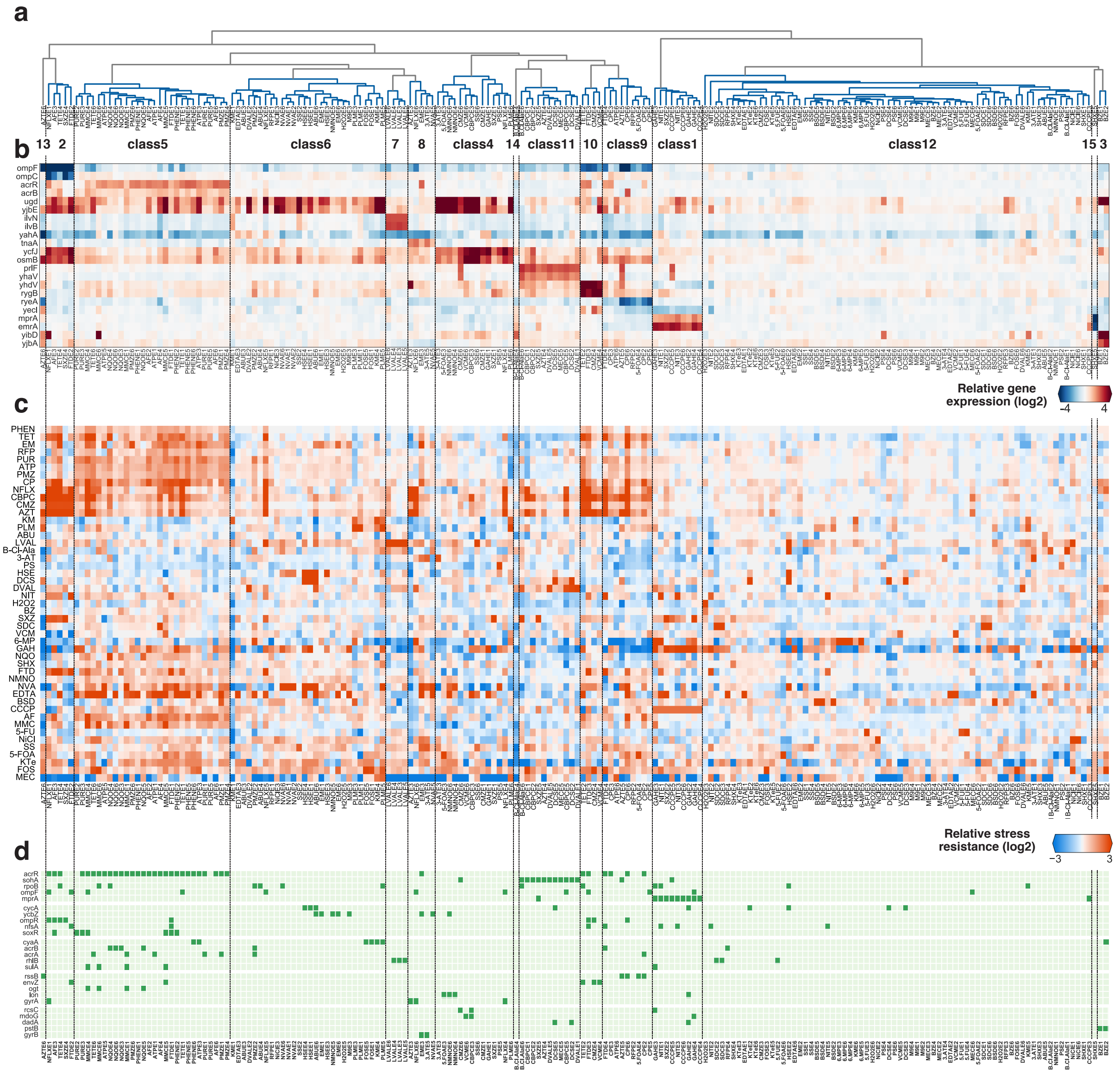

Fig. 4.10 Overview of the phenotypic clusters obtained from hierarchical clustering in the supervised PCA space. **a**, Dendrogram of the hierarchical clustering performed in the supervised PCA space. **b**, Gene expression profiles for the evolved strains. The genes were selected by selecting genes that had the highest weight for separating each specific class from the others via LDA and had more than a two-fold difference in average compared to the parent strain. **c**, The stress resistance (IC_{50}) profiles for the evolved strains. The colors for the ticks correspond to the action mechanisms of the stresses which are shown in the bottom of the figure. **d**, Mutation profiles for characteristic genes for each cluster. Genes that were enriched in each cluster were selected through the Fisher exact test (double-sided, $p < 0.01$). Genes that were mutated in more than seven strains are also shown.

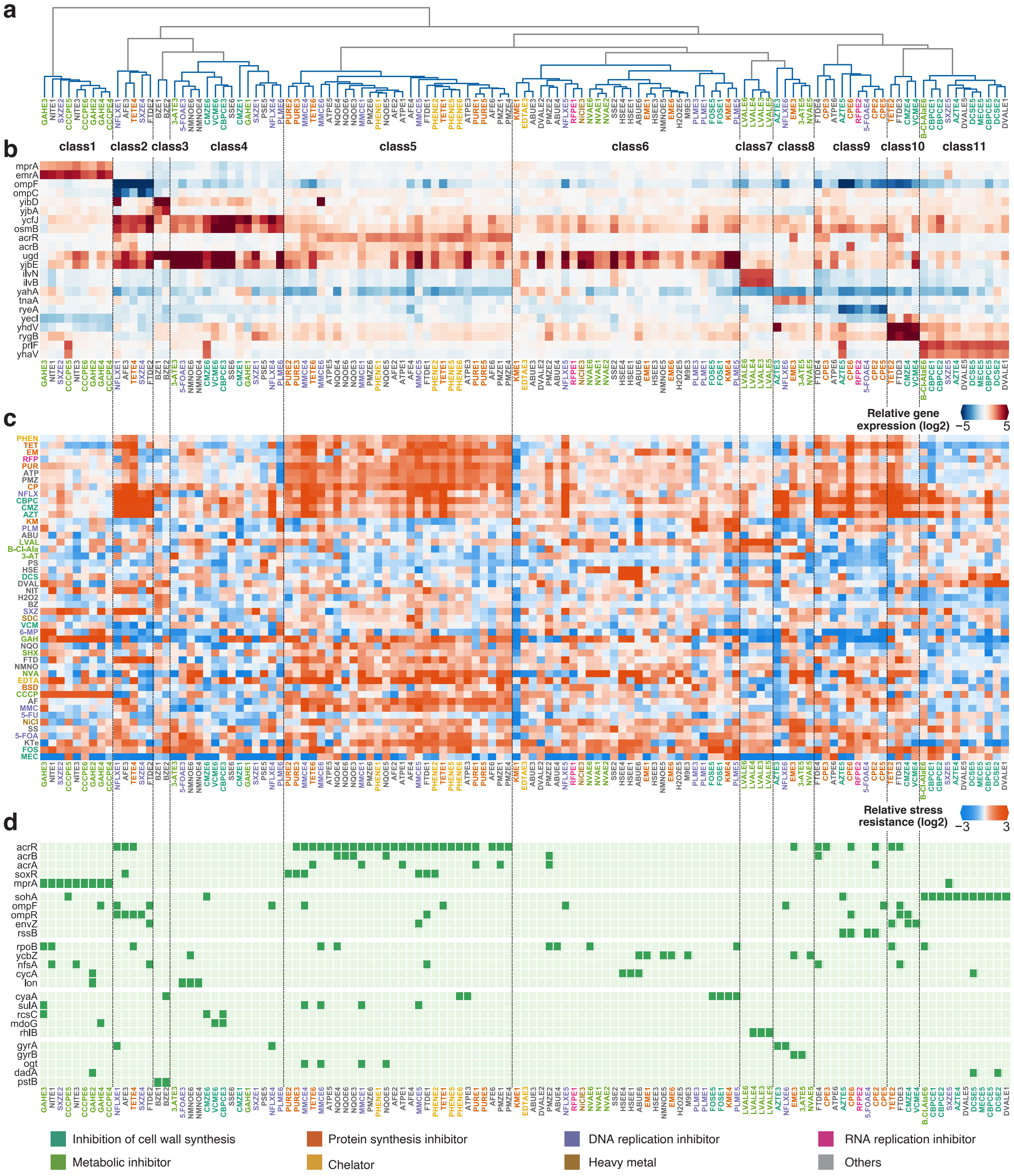

Fig. 4.11 The overview of the phenotypic clusters obtained from hierarchical clustering in the supervised PCA space omitting one cluster and three singletons for visibility. **a**, Dendrogram of the hierarchical clustering performed in the supervised PCA space. **b**, Gene expression profiles for the evolved strains. The genes were selected by selecting genes that had the highest weight for separating each specific class from the others via LDA and had more than a two-fold difference in average compared to the parent strain. **c**, The stress resistance (IC_{50}) profiles for the evolved strains. The colors for the ticks correspond to the action mechanisms of the stresses which are shown in the bottom of the figure. **d**, Mutation profiles for characteristic genes for each cluster. Genes that were enriched in each cluster were selected through the Fisher exact test (double-sided, $p < 0.01$). Genes that were mutated in more than seven strains are also shown. Figures reproduced from [150].

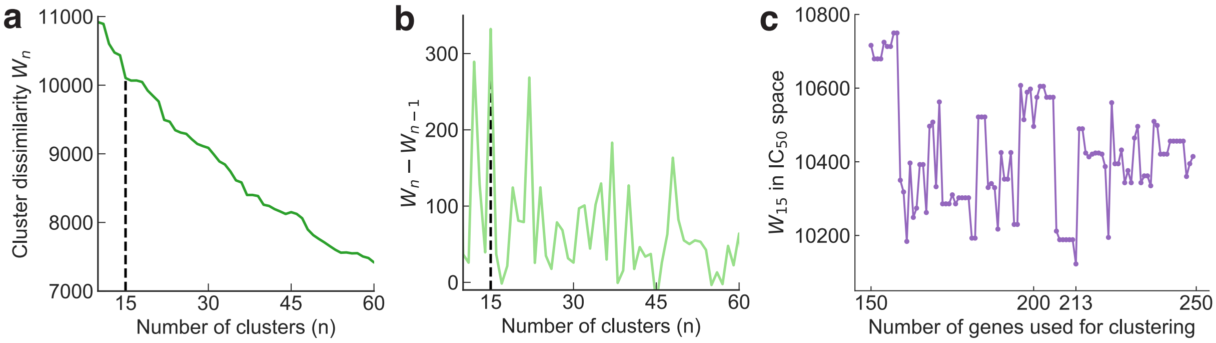

Fig. 4.12 **a**, Class dissimilarity W_n for different number of classes: n when clustering in the supervised PCA space constructed from 213 genes. **b**, The difference in class dissimilarity $W_n - W_{n-1}$ when clustering in the supervised PCA space constructed from 213 genes. **c**, Class dissimilarity W_{15} for the results of hierarchical clustering in the supervised PCA spaces constructed by different number of genes. The genes are sorted in the order of gene importance given by the random forest model. **c** was reproduced from [150].

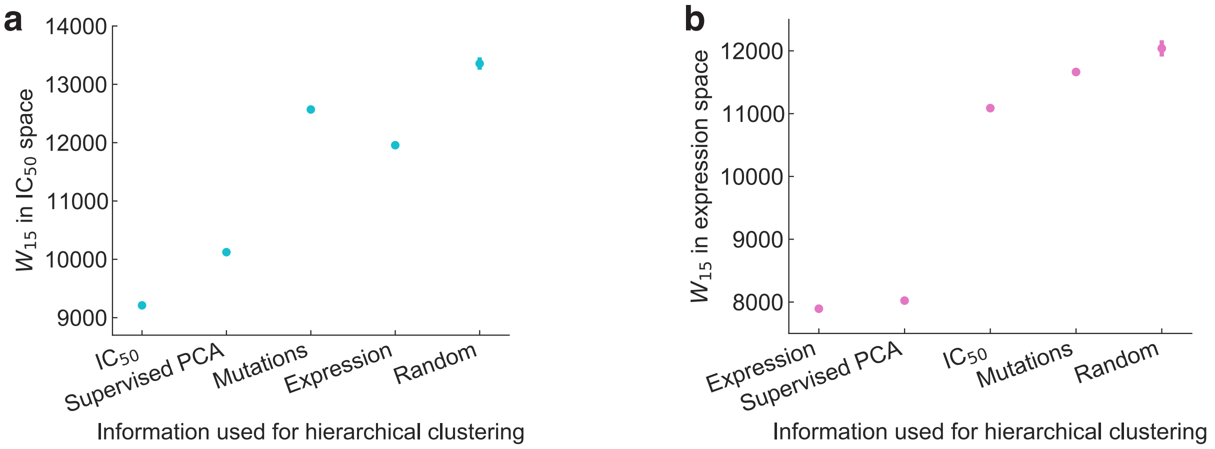

Fig. 4.13 **a**, Class dissimilarity W_n in the IC$_{50}$ space for the 15 classes which were defined by hierarchical clustering based on the supervised PCA expression space, IC$_{50}$ space, mutations, and full gene expression space, respectively, are shown. The mean and standard deviation for the class dissimilarity for ten runs of randomly clustered results in the IC$_{50}$ space are also shown. **b**, Class dissimilarity W in the 4492 dim. gene expression space for the results of hierarchical clustering in other spaces are shown. The mean and standard deviation for the class dissimilarity for ten runs of randomly clustered results in the gene expression space are also shown. Figures reproduced from [150].

4.3.3 Relation between the supervised PCA space and the genotypic space

Up to now, we have focused on the relations between the supervised PCA space and the stress resistance space, and have seen that the modular classes in the supervised PCA space corresponds well with the resistance space. We will next look into the relations between the classes in the supervised PCA space and the genotypic (or mutation) space. To our surprise, relatively clear relationships between the genotypic space and the supervised PCA space was observed (Fig. 4.11d) As shown, the patterns of fixed mutations coincided well with the modular classes in the supervised PCA space (Fig. 4.11a,b). This is a nontrivial correspondence since no genotypic information was used for the hierarchical clustering, suggesting that the identified mutations play a meaningful role in the modular gene expression classes. For example, all evolved strains in class 1 had mutations in *mprA* which encodes a repressor for multidrug resistance pump EmrAB. On the other hand, all class 11 strains had a mutation in *prlF* which encodes the antitoxin for the PrlF (SohA)-YhaV toxin-antitoxin (TA) system.

However, interestingly, the correspondence between the genotypic space and supervised PCA space was not perfect. This could be seen where strains in the same class gene expression class did not necessarily share the same mutations. For example, the evolved strains in class 5 exhibited an increased expression of *acrB* which encodes a component of the AcrAB/TolC multidrug efflux pump, and most of the class 5 strains (26/28) had a mutation in *acrR*, a repressor for *acrAB* (Fig. 4.11b,d). However, the other two strains also showed an increase in *acrB* expression without an *acrR* mutation. Another example could be seen in class 8 where the five strains consistently had increased expression of *tnaA* which encodes tryptophanase, whereas four out of the five strains had mutations in genes encoding DNA gyrase subunit A or B (*gyrA* or *gyrB*, Fig. 4.11b,d)[22]. Here, NVAE5, which did not have a mutation in *gyrA* nor *gyrB*, also showed an increased expression of tnaA. These two results seen in class 5 and 8 suggest the existence of multiple paths in the genotypic space for *E. coli* to reach desired expression and resistance levels. In other words, the convergence through evolution in the genotypic space is not as good as the phenotypic space. This leads to the motivation of Chapter 5.

In class 2, 9, and 10, where strains exhibited resistance to cell wall inhibitors and other stresses, a decreased expression of *ompF* was commonly observed. *ompF* encodes the outer membrane porin and decrease in *ompF* expression can be caused by either inactivation of the OmpR/EnvZ two-component system or RssB, which is a regulator of the alternative sigma factor RpoS [153–155]. Indeed, all strains in class 2 and class 10 had mutations in either *ompR* or *envZ*, and four out of nine strains in class 9 had mutations in *rssB*. Although these strains commonly had a decreased *ompF* expression, their expression levels for genes such as *ompC* and *rygB* differed (Fig. 4.11b)[23]. These differences in the gene expression profile could be guessed as one of the reasons for the diversification of the three *ompF* classes. Interestingly, although all three classes showed resistance to β-lactams such as carbenicillin (CBPC) and cefmetazole (CMZ)), resistance levels to stresses such as sulfisoxazole (SXZ) and DL-3-hydroxynorvaline (NVA) differed between classes (e.g. strains in class 2 and 10 exhibited resistance to SXZ, while strains in class 9 did not, Fig. 4.10c, 4.11c). These results suggest that different classes in the supervised PCA space correspond to different internal cell states, leading to different stress resistance mechanisms.

A certain level of stochasticity was also observed where the four strains which evolved under the same stress were not always categorized in the same class. For example, none of the four SXZ evolved strains shared the same class (SXZE2 in class 1, SXZE4 in class 2, SXZE1 in class 4, and SXZE5 in class 11). Moreover, each SXZ evolved strain showed different gene expression and resistance patterns, indicating a rugged fitness landscape with multiple local peaks. Not only SXZ, but also none of the four norfloxacin (NFLX) evolved strains shared the same class. Intriguingly, these local peaks were accessible not only by SXZ and NFLX evolved strains, but also by strains which evolved in other stresses (for example, see class 2 where strains evolved in NFLX, SXZ, AF, TET, and FTD share the same class[24]). These results suggest that evolution under the same selection pressure does not necessarily lead to the same phenotype [94, 108], and that these local peaks in the fitness landscape are reachable through the evolution under different stresses. Overall, the phenotypic classes in the supervised PCA

[22] The relation between the *gyrB* mutation and increased TnaA production was previously discussed in [152].

[23] *ompC* encodes a porin and *rygB* encodes a small RNA involved in the regulation of the outer membrane composition.

[24] It should be noted that these drugs do not share the same mechanism of action. This suggests that constraints leading to the modular gene expression classes do not necessarily depend on the mechanism of action of the stresses.

space, loosely corresponded with the genotypic space, while we could observe the existence of multiple genotypic pathways to reach an optimum in the fitness landscape. In addition, these optima were shared by strains evolved in diverse stresses suggesting that evolutionary constraints do not necessarily rely on the drug action mechanisms.

4.4 Commonly mutated genes underlie the evolutionary constraints for stress resistance

As we have seen in the previous section, several genes were commonly mutated within the 192 evolved strains, and some showed good correspondence with the classes in the supervised PCA space. However, up to now, we have only been focusing on correlations in the dataset and not causality. Thus, to verify the effects of the commonly mutated genes, 64 of the representative mutations within the 192 evolved strains were introduced to the parent strain using multiplex automated genome engineering (MAGE) [156] (Table 4.2, 4.3)[25]. To compare the phenotypes of the mutant strains with the 192 evolved strains, we quantified the changes in the 47 dimensional IC_{50}space for the 64 reconstructed mutant strains.

We first asked whether the cross-resistance and collateral sensitivities observed within the evolved strains could be reproduced by the 64 reconstructed mutant strains. Accordingly, we calculated the Pearson's correlation coefficient R between the IC_{50}s of all 47 stresses within the 192 evolved strains. We recognized that some stress pairs showed high positive correlation suggesting cross resistance. For example, evolved strains resistant to CBPC tended to exhibit resistance to aztreonam (AZT) as well ($R = 0.95$, Fig. 4.14a), both of which constitute β-lactam stresses. On the other hand, there existed pairs that showed negative correlation suggesting collateral sensitivity (e.g. TET and B-Cl-Ala, Fig. 4.14c). We then calculated correlation coefficients for the reconstructed mutant strains. Interestingly, many of the correlations observed in the 192 evolved strains were also observable within the 64 mutant strains, even though each mutant has only been introduced a single mutation. For example see the correlations between CBPC & AZT and TET & B-Cl-Ala (Fig. 4.14b,d). To quantify the generality of this correspondence, we compared the pair correlation coefficients for the mutant strains R_{mutant} and the evolved strains R_{evolved} and measured the correlation between those coefficients ($R = 0.66$, Fig. 4.15a,b). Since transporters and porins are major stress resistance mechanisms and a significant portion of the representative mutations included genes related to such mechanisms (e.g. *acrR, ompF*), one might think that the high correlation between the evolved strains and mutant strains are also caused by them. To answer this question, we calculated the pair correlation coefficients within the mutant strains, excluding 18 strains that are related to transporters and porins (*dctA, uraA, sstT, livM, potA, oppA, cycA, yhjE, glpT, ompF, glnP, metN, ptsP, frlA, gabP, potH, mprA*, and *acrR*). Interestingly, we could still observe a high correlation between the mutant strains and the evolved strains with a correlation of $R = 0.57$ (Fig. 4.15c). These results indicate that the observed cross resistance / collateral sensitivity patterns (i.e. evolutionary constraints) are rooted in the representative mutations. In the following subsections, we will investigate the mechanisms of resistance through a portion of the representative mutations and the corresponding reconstructed mutants (Fig. 4.16). We show a schematic figure with the resistance mechanisms described below, in Fig. 4.17.

4.4.1 Biological mechanisms responsible for the phenotypic classes

Class 1 and the EmrAB/TolC efflux pump

All strains in class 1 had a mutation in *mprA* and had high expression levels for *mprA* and *emrA* (Fig. 4.11b,d). The activation of EmrAB/TolC which is an efflux pump, is regulated by *mprA* and results in resistance to a previously identified substrate CCCP (uncoupling agent) [157]. The resistance profiles of both the evolved strains and the reconstructed mutant suggest that the *mprA* mutation leads not only to CCCP resistance, but also to substrates such as chloramphenicol (CP, protein synthesis inhibitor) and phleomycin (PLM, DNA intercalator).

[25] Here, we focused on the genes that were commonly mutated in more than two strains. Exceptionally, genes that shared the same operon (e.g. *cyoA, cyoB, cyoE*), were included even if they were confirmed in a single strain. For mutations that seemed to have disabled the genes, a NheI site containing a TAG stop codon was introduced immediately downstream of its start codon and one base was inserted to introduce a frameshift mutation. See Supplementary Data 3 in [150] for details of the introduced mutations.

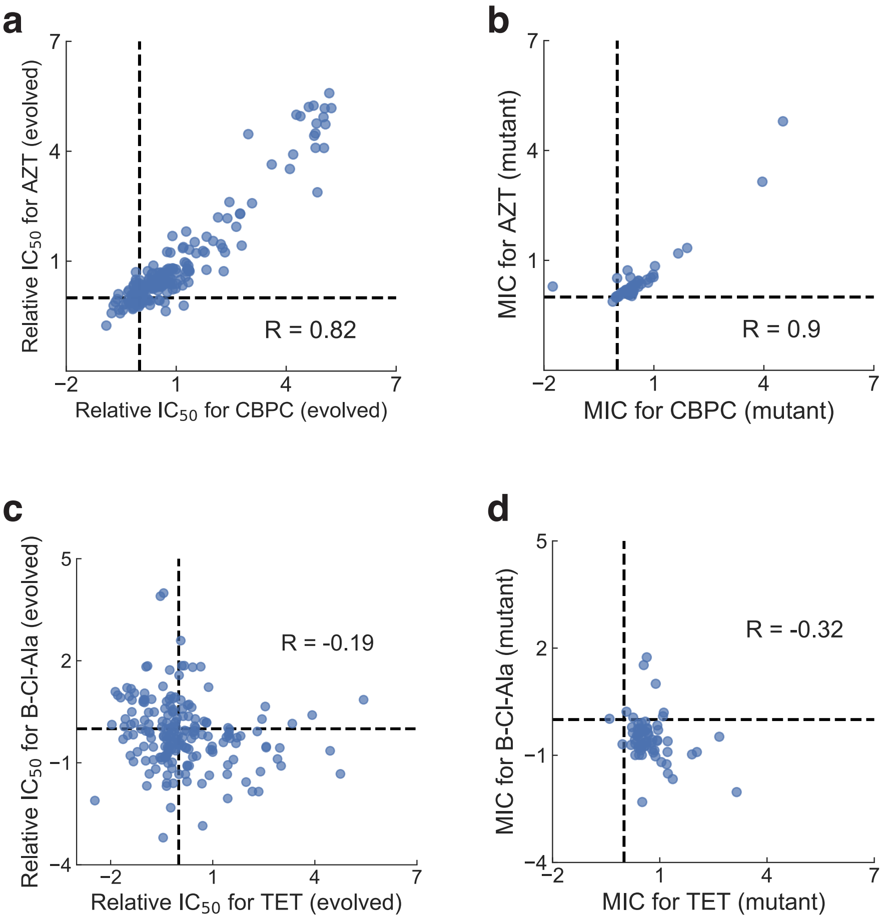

Fig. 4.14 Examples of pair relations of stress resistance acquisition. **a**, **b**, Relationship between the IC$_{50}$ values of CBPC/AZT for the 192 evolved strains and 64 site directed mutant strains, respectively. R denotes Pearson's correlation coefficient. **c**, **d**, Relationship between the IC$_{50}$ values of TET/B-Cl-Ala for the 192 evolved strains and 64 site directed mutant strains, respectively. **a** and **d** were modified and reproduced from [150].

Class 2,10 and the OmpF porin

All strains in class 2 and 10 had a mutation in either *ompR* or *envZ* which are both regulators for the OmpF porin, and these strains all had a decreased expression of *ompF*. We identified that the inactivation of the OmpF porin results in resistance not only to previously described substrates such as CBPC (cell wall synthesis inhibitor) [158, 159], but also novel substrates such as 1,10-phenanthroline (PHEN, chelator), puromycin (PUR, protein synthesis inhibitor), and other chemicals.

Class 5 and the AcrAB/TolC efflux pump

Al strains except PURE2 amd PURE5 in class 5 had a mutation in *acrR* which is a regulator for the AcrAB/TolC efflux pump. Indeed, all strains in class 5 exhibited an increased expression of *acrB* (Fig. 4.11b,d). We identified that the inactivation of the repressor AcrR results in resistance not only to previously described substrates such as tetracycline (TET) and erythromycin (EM) [160–162], which are both known as protein synthesis inhibitors, but also to substrates that have not been reported yet such as NVA (threonine analog) and NQO.

Class 9 and *rssB*

The strains in class 9 showed a tendency of sensitivity to metabolic inhibitors such as L-valine (LVAL), β-chloro-L-alanine (B-Cl-Ala), 6-mercaptopurine monohydrate (6-MP), GAH, and 3-amino-1,2,4-triazole (3-AT) (Fig. 4.11c, 4.16a). Since 4/9 strains in class 9 had a mutation in *rssB* (Fig. 4.11d), we speculated that this *rssB* mutation could be one of the reasons for the observed collateral sensitivities. Indeed, we have been able to observe a two to five-fold change in sensitivity to the stresses noted above in the reconstructed *rssB* mutant strain (Fig. 4.16c). Since class 9 strains and the reconstructed *rssB* strain both show resistance to cell wall inhibitors and other stresses (AZT, CBPC, TET), our results indicate a

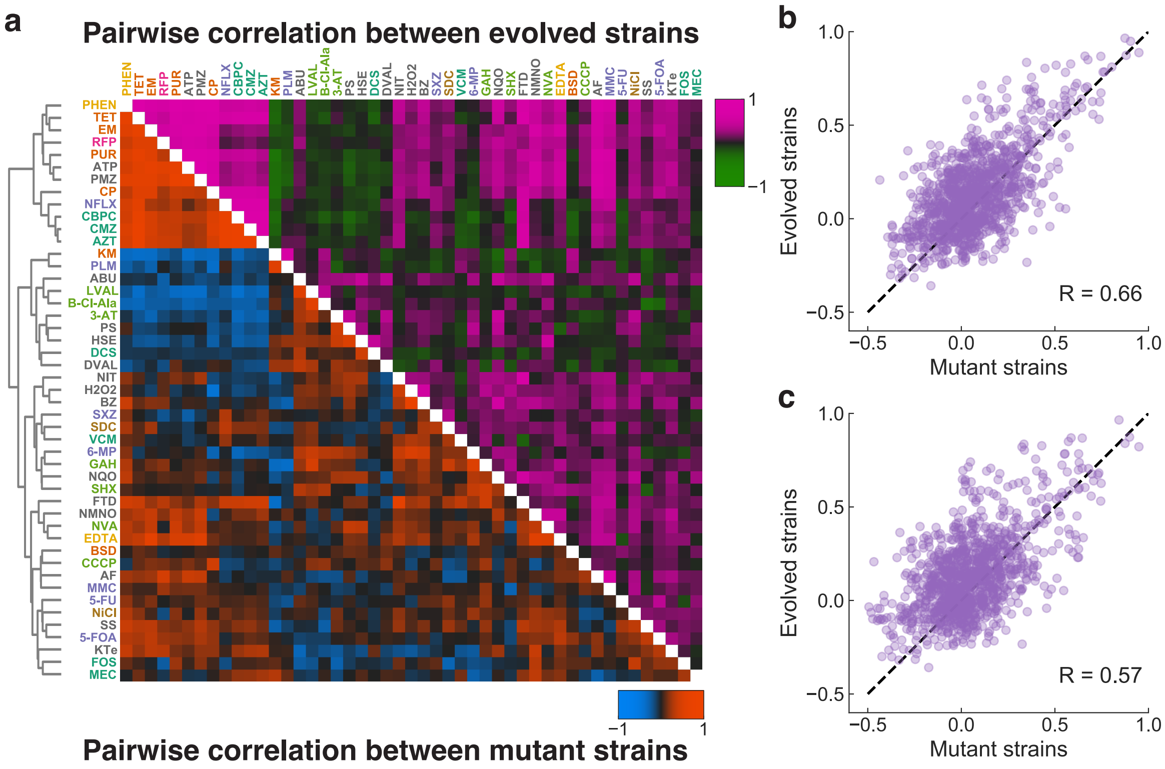

Fig. 4.15 **a**, Pearson's correlation coefficient for all pairwise combinations for the stress resistance acquisition over the 192 evolved strains (upper right) and 64 mutant strains (lower left), respectively. The order of stresses were determined through hierarchical clustering on the mutant's stress resistance profiles. **b**, The relation between the correlation coefficients for stress resistance acquisition of the 64 mutant strains and the 192 evolved strains. R denotes Pearson's correlation coefficient. **c**, The relation between the correlation coefficients for stress resistance acquisition of the 46 mutant strains and the 192 evolved strains. Here, 18 mutants with transporter related mutations (*dctA, uraA, sstT, livM, potA, oppA, cycA, yhjE, glpT, ompF, glnP, metN, ptsP, frlA, gabP, potH, mprA*, and *acrR*) were excluded from the calculation for correlation. Figures reproduced from [150].

trade-off between these stresses and metabolic inhibitors. It has been reported that while *E. coli* strains with higher *rpoS* levels show increased resistance to several external stresses [163], they also exhibit decreased carbon source availabilities and poor competitiveness for low concentrations of nutrients due to the competition between RpoS and the house-keeping sigma factor RpoD (sigma 70) [164]. Since RssB facilitates the degradation of RpoS, the collateral sensitivities to several metabolic inhibitors in class 9 evolved strains could also be caused by the sigma factor competition.

Class 11 and the PrlF/YhaV toxin-antitoxin system

All evolved strains in class 11 carried the same mutation in *prlF* (*sohA*), a duplication of TTCAACA sequences located 272 bp downstream of the start codon. Although the contribution of PrlF-YhaV to stress resistance has not yet been reported, We found that these evolved strains, and the reconstructed *prlF* mutant strain, commonly exhibited resistance to CBPC, AZT, and DVAL (Fig. 4.11c, 4.16a). All 11 strains in class 11 showed a decreased expression of *ompF* (Fig. 4.11b), which was also confirmed in the reconstructed *prlF* mutant strain through qRT-PCR analysis. These results suggest that cross-resistance to CBPC, AZT, and DVAL by the *prlF* mutation is at least partially caused by decreased expression of *ompF*. Since YhaV is a translation-dependent RNase [165], this decrease in ompF expression might be caused through the alteration of global gene expression. Class 11 strains and the *prlF* mutant strain also showed sensitivity to hydrogen peroxide (H_2O_2), benserazide (BZ), and NQO (Fig. 4.11c, 4.16b). These results suggest that DVAL, CBPC resistance, acquired through the *prlF* mutation, leads to a trade-off for H_2O_2, BZ, and NQO. Previous studies reported that *E. coli* mutant strains lacking superoxide dismutase showed increased susceptibility to H_2O_2 mediated killing [166]. Indeed, all strains in class 11 and the

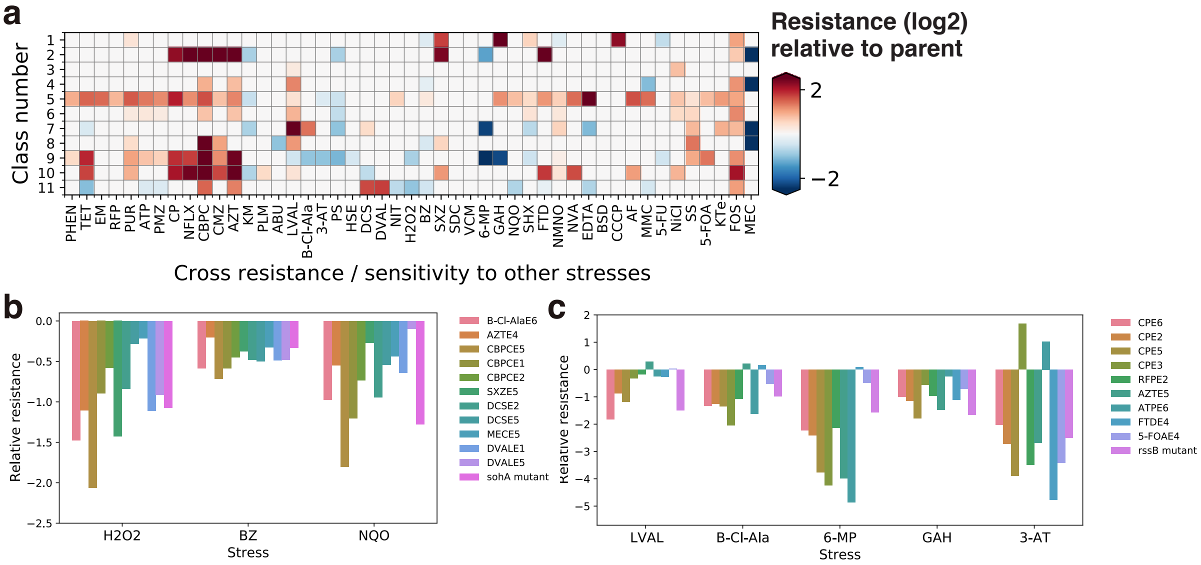

Fig. 4.16 **a**, Combinations of stresses which exhibited either cross-resistance or collateral sensitivity for the strains in each class in the supervised principal component analysis (PCA) space. The combinations were detected by the Mann-Whitney U-test (false discovery rate, FDR < 0.05), and the colors indicate the resistance to the stress relative to the parent strain. **b,c**, Stress resistance relative to the parent strain for strains in class 11 and class 9, respectively. Resistance levels for the rssB and prlF mutant are also shown for comparison. Figures reproduced from [150].

reconstructed *prlF* mutant strain consistently exhibited a 0.45 ± 0.11-fold decrease in *sodB* expression[26]. This suggests that the observed H_2O_2 sensitivity is caused by the degradation of sodB through the YhaV toxin.

4.4.2 Are single mutations sufficient to explain drug resistance?

Although the correlation coefficients of the stress pairs in Fig. 4.15 between the mutant and evolved strains showed a good correspondence, there did exist some cases where the resistance profiles of the single mutant strains differed from that of the evolved strains with the corresponding mutations. For example, all evolved strains in class 1 (Fig. 2) had mutations in *mprA* and no other common mutations, strongly suggesting the contribution of this mutation to the common phenotypic changes in class 1. However, the reconstructed mutant strain of *mprA* exhibited a similar, yet significantly different resistance profile for certain stresses, such as SXZ (Fig. 4.18). Interestingly, the majority of the reconstructed mutant strains showed smaller phenotypic changes compared to the evolved strains. For example, this could be seen in Fig. 4.19a, where we show the distance from the parent strain for all mutant and evolved strains. These results suggest that the single mutant strains are not sufficient to express the distances in the resistance space although they can roughly explain the direction of the phenotypic changes as we see in Fig. 4.15 and Fig. 4.19c. These differences between the evolved strains and reconstructed mutant strains might suggest the contribution of epistatic interactions between the multiple mutations to the resistance changes since most of the evolved strains had more than two mutations [167]. There may also be a non-genetic contribution, which will be difficult to explain simply by the phenotype-genotype mapping presented in this study [168]. It would be interesting to identify the effects of multiple mutations and non-genetic adaptations on stress resistance, which is a promising direction of future work.

4.5 Decelerated evolution against β-lactam antibiotics

In the adaptive evolution to β-lactams (i.e., CMZ and CBPC), we found that certain strains, evolved under specific stresses, acquired higher resistances to β-lactams than strains that were directly selected by β-lactams. For example in Fig. 4.20a, we can observe that strains evolved in TET acquired higher

[26] *sodB* encodes (Fe) superoxide dismutase.

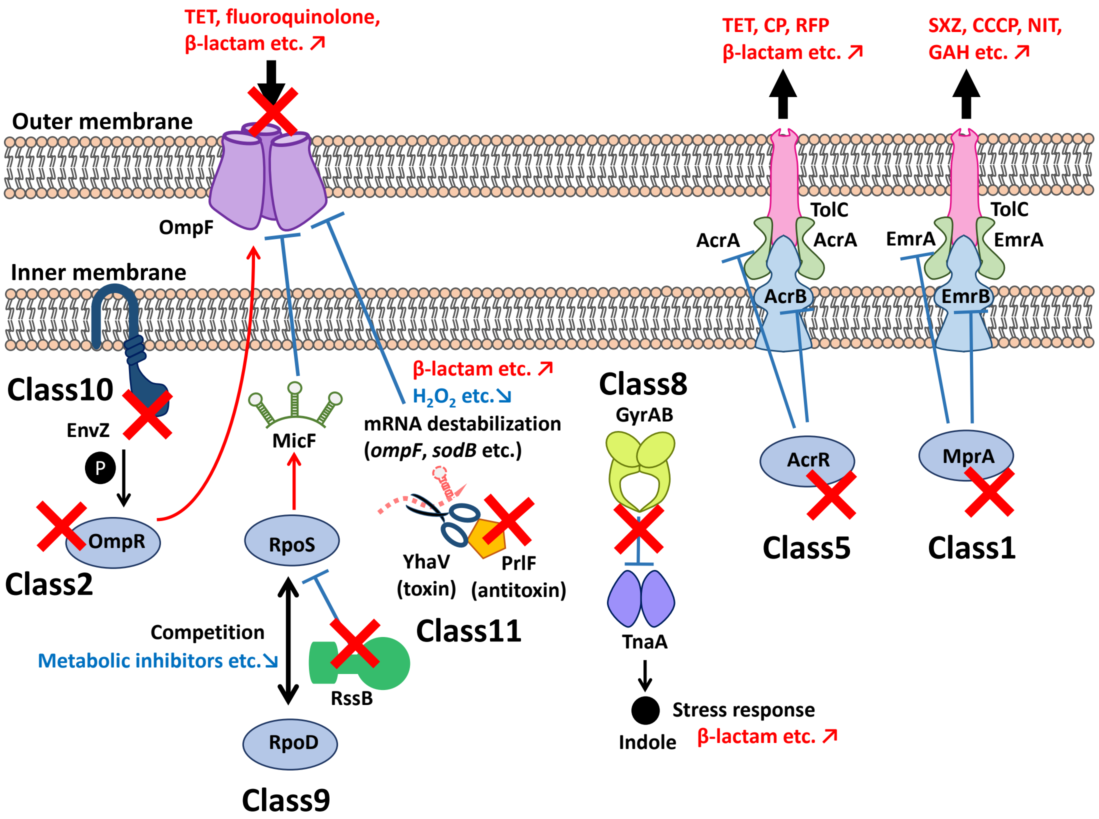

Fig. 4.17 A schematic image of the stress resistance mechanisms responsible for the elucidated clusters in the supervised PCA space. Typical stresses which the strains showed resistance (red) and sensitivity (blue) are shown. The schematics are used by courtesy of Tomoya Maeda. Figures reproduced from [150].

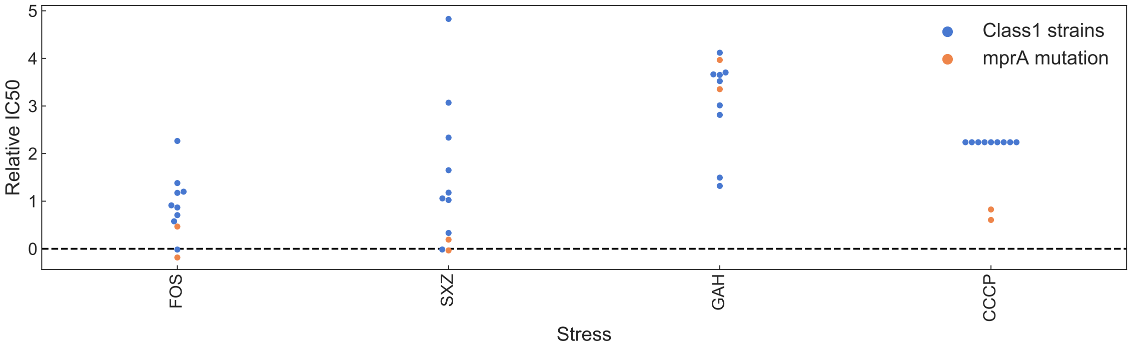

Fig. 4.18 The IC_{50} values relative to the parent MDS42 strain are shown for the class 1 evolved strains and the *mprA* reconstructed mutant. It could be observed that the *mprA* mutant shows a lower resistance level than the class 1 evolved strains for stresses such as SXZ and CCCP.

resistance levels to CBPC than the strains that directly evolved in CBPC. The same tendency could be observed between NFLX and CMZ evolved strains (Fig. 4.20b). We coined this phenomenon as "decelerated evolution" to β-lactams, and this phenomenon seemed to be reflected in the difference in the mutation profile among the evolved strains. Especially, we found that the evolved strains that exhibited the highest resistances to CBPC and CMZ (the *overtakers*) tended to have mutations in genes related to the membrane porin protein OmpF, i.e., *ompF*, *ompR* and *envZ* (Fig. 4.20c). In contrast, the strains evolved under CBPC or CMZ had fewer mutations in OmpF related genes (one out of eight evolved strains) in comparison with the overtaking strains with high β-lactam resistance ($p = 0.04$, Fisher's exact test, Fig. 4.20c). This result might suggest that in the current laboratory evolution setup, the fixation of mutations related to OmpF is suppressed under the addition of β-lactams, even though they can increase their resistance to the drug.

Several explanations could be thought for this "decelerated evolution".

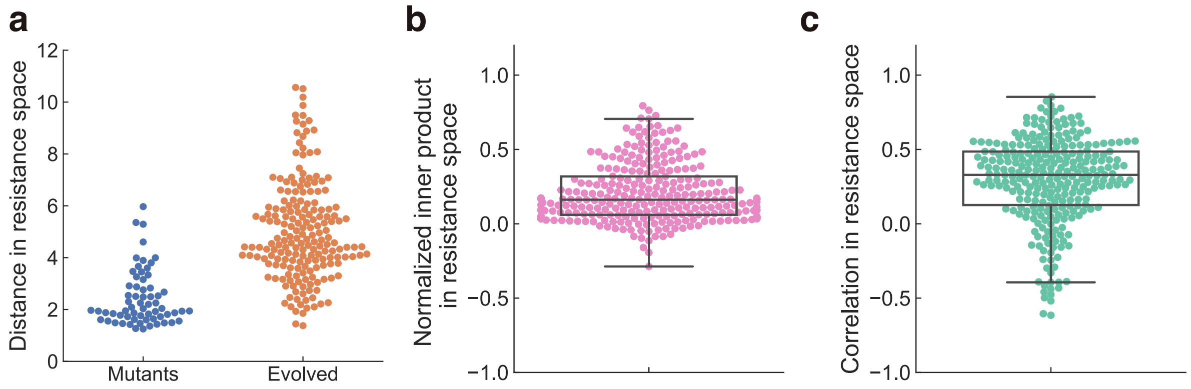

Fig. 4.19 **a**, The distances in the 47 dim. resistance space for the 64 mutants and 192 evolved strains. **b**, The normalized inner product in resistance space: $v_{\mathrm{evolved}} \cdot v_{\mathrm{mutant}}/|v_{\mathrm{evolved}}|$, where v_{evolved}, v_{mutant} denotes the difference from the parent strain as a 47 dim. vector in the resistance space for the evolved and mutant strains, respectively. The inner products were calculated between the single mutant strains and the evolved strains which had a mutation in the same gene. **c**, Pearson's correlation coefficients in the 47 dim. resistance space calculated between single mutant strains and the evolved strains which had a mutation in the same gene.

(i) The *ompF* mutation might lead to a fitness cost under β-lactams.

(ii) Negative epistasis might exist between a *ompF* related mutation and a common mutation among the β-lactam evolved strains.

(iii) The mutation rate under β-lactams might be suppressed, making it hard for the strains to find *ompF* related mutations although they are beneficial.

To test hypothesis (i), we measured the growth rate of the parent MDS42 strain and the *ompF* mutant[27] under various concentrations of CMZ and CBPC (Fig. 4.20a,b). However, the growth rate for the *ompF* mutant strain was the comparable with (or even higher than) the parent MDS42 strain at all CMZ and CBPC concentrations, suggesting that there is no fitness cost associated with the *ompF* mutation. To test hypothesis (ii), we compared the IC_{50}values of the *ompF* mutant strain with that of the *ompF+prlF* mutant strain. The *prlF* mutation was observed in 4/8 strains of the CMZ and CBPC strains which was the most common mutated gene within the eight strains. Thus, if somehow *prlF* was mutated in the early stage of evolution and there existed a negative epistatic interaction between *prlF* and *ompF* related mutations, the frequency of *ompF* mutations could be suppressed. However, we found that the IC_{50}values of the *ompF+prlF* mutant strain has the same level of $\mathrm{IC}_{50}$to CMZ and CBPC compared to the *ompF* mutant strain (Fig. 4.21c). For the last hypothesis (iii), we measured the mutation frequency for the parent MDS42 strain under β-lactam stresses: CBPC, CMZ and stresses that frequently had overtakers: CP, NFLX, TET[28]. Since the mutation frequency did not differ between β-lactams and other stresses, we concluded that the mutation frequency could not explain the observed decelerated evolution (Fig. 4.21d). Since the three possible explanations were all denied, we suggest another possible hypothesis which suggests that the contributions of the *ompF* mutation might not be observable when the cell wall state is not stable, especially under cell wall inhibitors such as β-lactams. Since the addition of β-lactams reportedly induces bulge formation leading to cell lysis [169], a decrease in *ompF* expression and the disruption of OmpF function may not contribute to β-lactam resistance. However, we have not been able to test this hypothesis sufficiently, and the explanation of this interesting phenomenon "decelerated evolution" is left for future work.

[27] Here, the *ompF* mutant was made by inserting a NheI sequence to disable OmpF production.

[28] Here, the mutation rate was measured as the number of colonies growing on rifampicin after a overnight culture under the addition of each stress.

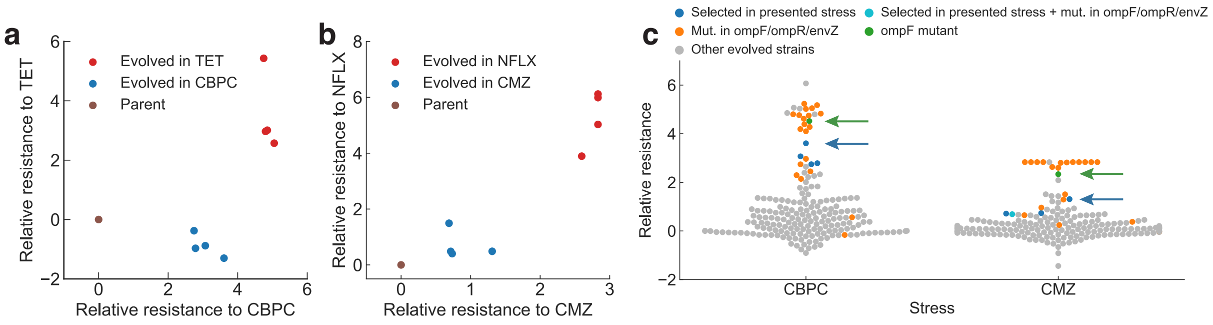

Fig. 4.20 **a,b**, Decelerated evolution observed within the evolved strains. Relative log2 (IC_{50}) for evolved strains in CBPC and TET (a), and relative log2 (IC_{50}) for evolved strains in norfloxacin (NFLX) and cefmetazole (CMZ) (b). **c**, Relative IC_{50} values for CBPC and CMZ for all 192 evolved strains. Many of the strains which exhibit resistance higher than the CBPC and CMZ evolved strains had a mutation in ompF or its regulators ompR and envZ (orange). The CBPC and CMZ resistance of the ompF introduced strain also exhibited higher resistance (green, green arrow) than the CBPC and CMZ evolved strains (blue, cyan, denoted by a blue arrow). Figures reproduced from [150].

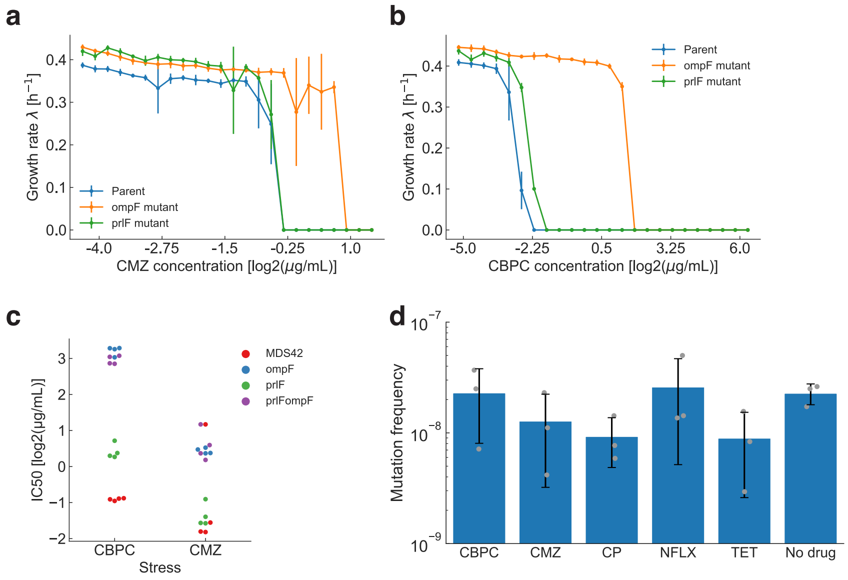

Fig. 4.21 **a,b**, Growth rates for the parent strain, *ompF* mutated strain, and the *prlF* mutant. Growth rates were measured in 24 concentration levels of cefmetazole (CMZ) and carbenicillin (CBPC), respectively. **c**, IC_{50} levels measured for different stresses for the parent strain, *ompF* mutated strain, *prlF* mutated strain, and the *prlF*/*ompF* doubled mutated strain, respectively. *d*, Mutation frequencies for MDS42 strains under addition of IC_{50} concentrations of β-lactam stresses (i.e. CBPC, CMZ) and other antibiotics in which the evolved strains acquired high resistance to β-lactam stresses (i.e. CP, NFLX, TET). Error bars represent the standard deviation of three independent experiments. Figures reproduced from [150].

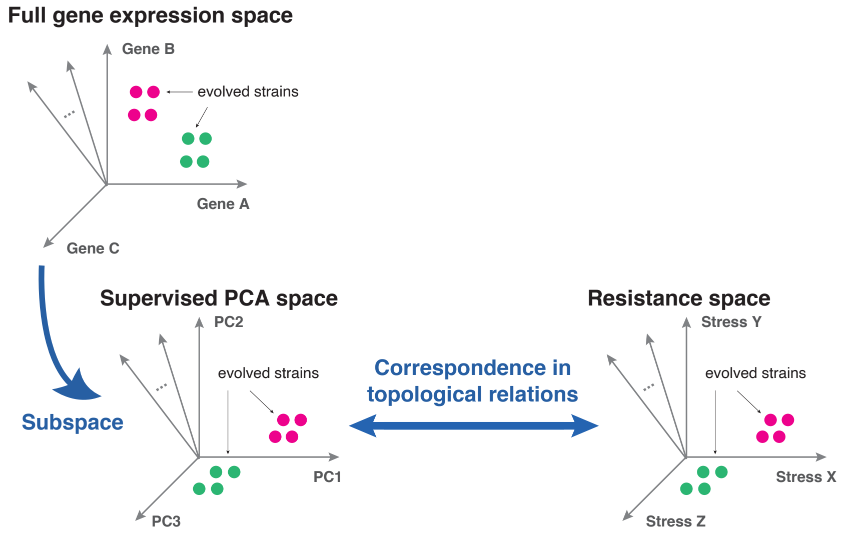

Fig. 4.22 Our study indicates that topological (neighboring) relations in the stress resistance space can be accurately represented by the supervised PCA space which is a subspace of the whole gene expression space.

4.6 Discussion

In this study, we performed laboratory evolution of *E. coli* under various heterogenous stress conditions which allowed us to elucidate the molecular mechanisms associated with resistance acquisition to antibiotics and non-antibiotics stressors. Our analysis using supervised PCA revealed the existence of modular phenotypic classes both in the gene expression space and the stress resistance space, suggesting close interactions between changes in gene expression and stress resistance (Fig. 4.11). Using the class dissimilarity measure W_n, we have shown that the supervised PCA space represents the topological relationships in the stress resistance space better than the whole 4492 dimensional gene expression space (Fig. 4.13a). This suggests that the supervised PCA space, which is a subspace of the whole gene expression space, is constituted by representations that accurately express the stress resistance phenotypes (Fig. 4.22). In other words, the supervised PCA space could be interpreted as the latent space for stress resistance.

The distribution of the evolved strains in the supervised PCA space revealed that the drug resistant phenotypes could be classified into a few number of classes. This indicates that *E. coli* only arm a few number of strategies for stress resistance, suggesting that resistance acquiring dynamics could be predictable. Interestingly, strains in the same phenotypic class did not necessarily evolve in the same stress nor stresses with similar mechanisms of action suggesting that evolutionary constraints could be constituted by different mechanisms (e.g. the chemical properties of drugs). It has also been observed that strains that evolved in the same stress did not necessarily belong to the same phenotypic class, suggesting the stochasticity in evolution. We believe that our analyses, together with the data provided in this study, provide the basis for understanding evolutionary constraints and the stochasticity which underlie stress resistance evolution.

Note, our analyses based on supervised PCA in the gene expression space has several caveats. First, the identification of distinct classes in the supervised PCA space (Fig. 4.10, 4.11) was based on gene expression changes, and thus, our analysis could not detect resistance acquiring mechanisms with little if any gene expression changes. For example, three out of four evolved strains in 6-MP had mutations in the *hpt* gene, encoding hypoxanthine phosphoribosyltransferase, which suggests a contribution of this *hpt* mutation to the 6-MP resistance phenotype. However, these 6-MP resistant strains exhibited little expression changes compared to the parent MDS42 strain which kept us from identifying the phenotypic class the 6-MP evolved strains belong to. Such evolved strains with minor expression changes were assigned to class 12 (Fig. 4.10)[29].

[29] As we can see from Fig. 4.10c, the strains in class 12 exhibit local changes in their IC_{50}profiles. While there is a possibility that these localized changes might be an artifact caused by the bias in the selected 47 stresses, this observed locality suggest that the class 12 strains acquire resistance by changing their phenotype in a limited manner.

Another limitation of our study is caused by the limitation of the conditions used for the transcriptome analysis. It is true that some genes will only be induced or suppressed in the presence of a stressor, and these genes could also contribute to stress resistance [170]. However, due to the expensive experimental cost, we neglected environment-dependent gene expression changes and measured the transcriptome only in the no-stress environment. It would clearly be interesting to collect the transcriptome profiles under the presence of each drug for the parent strain and all evolved strains. In this case, we would be able to measure how long-term evolution would bias the direction of short-term adaptation which would definitely contribute to constructing a theoretical framework for evolution.

4.7 Supplementary Figures and Tables

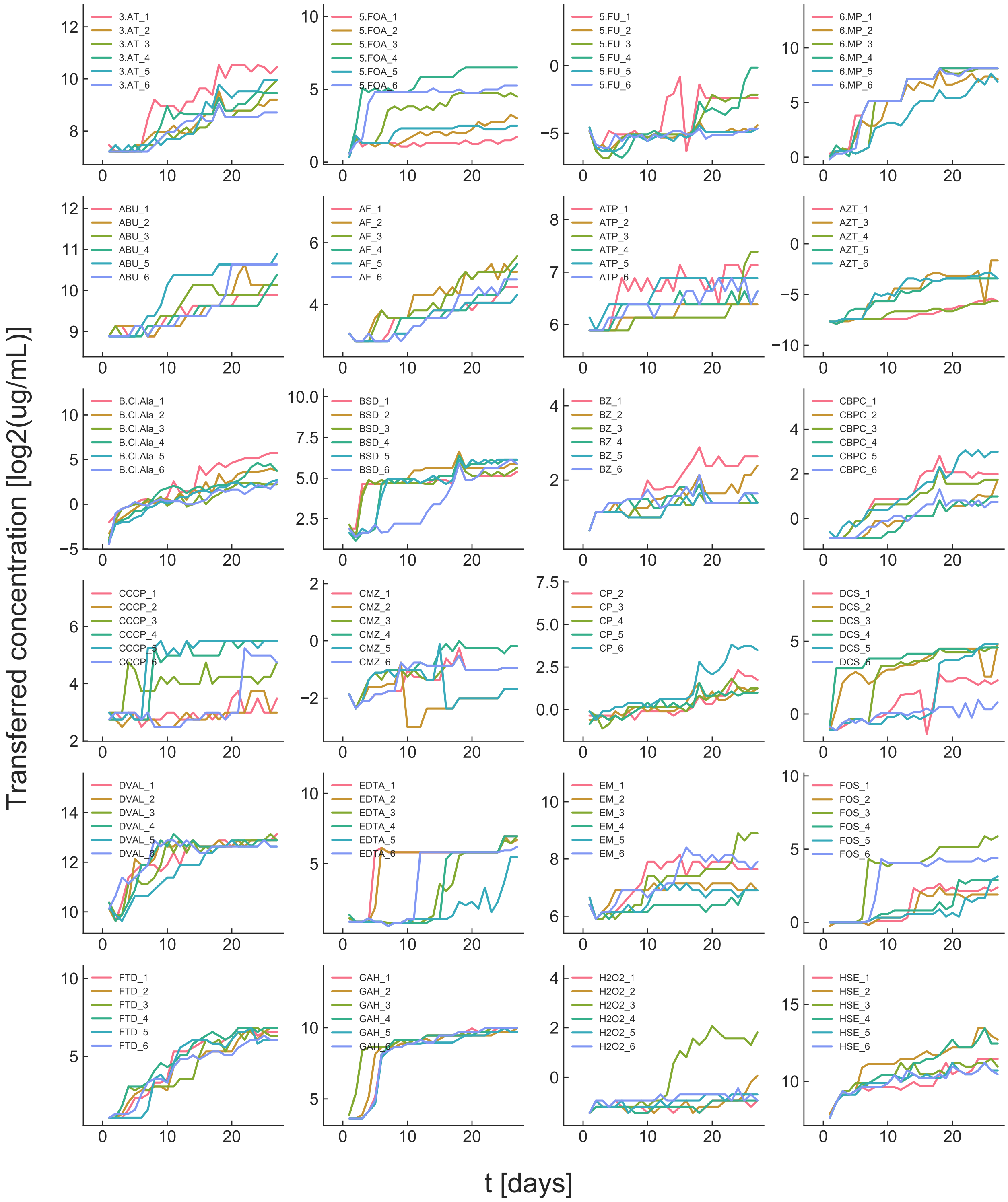

Fig. 4.23 Time series for the resistance over the 27 days of the laboratory evolution experiment for all 48 environments. The resistance is measured as the highest concentration of the drug which the strain could grow in each day. The strain which grew in this highest concentration was selected and transferred to a fresh medium each day. Figures modified and reproduced from [150].

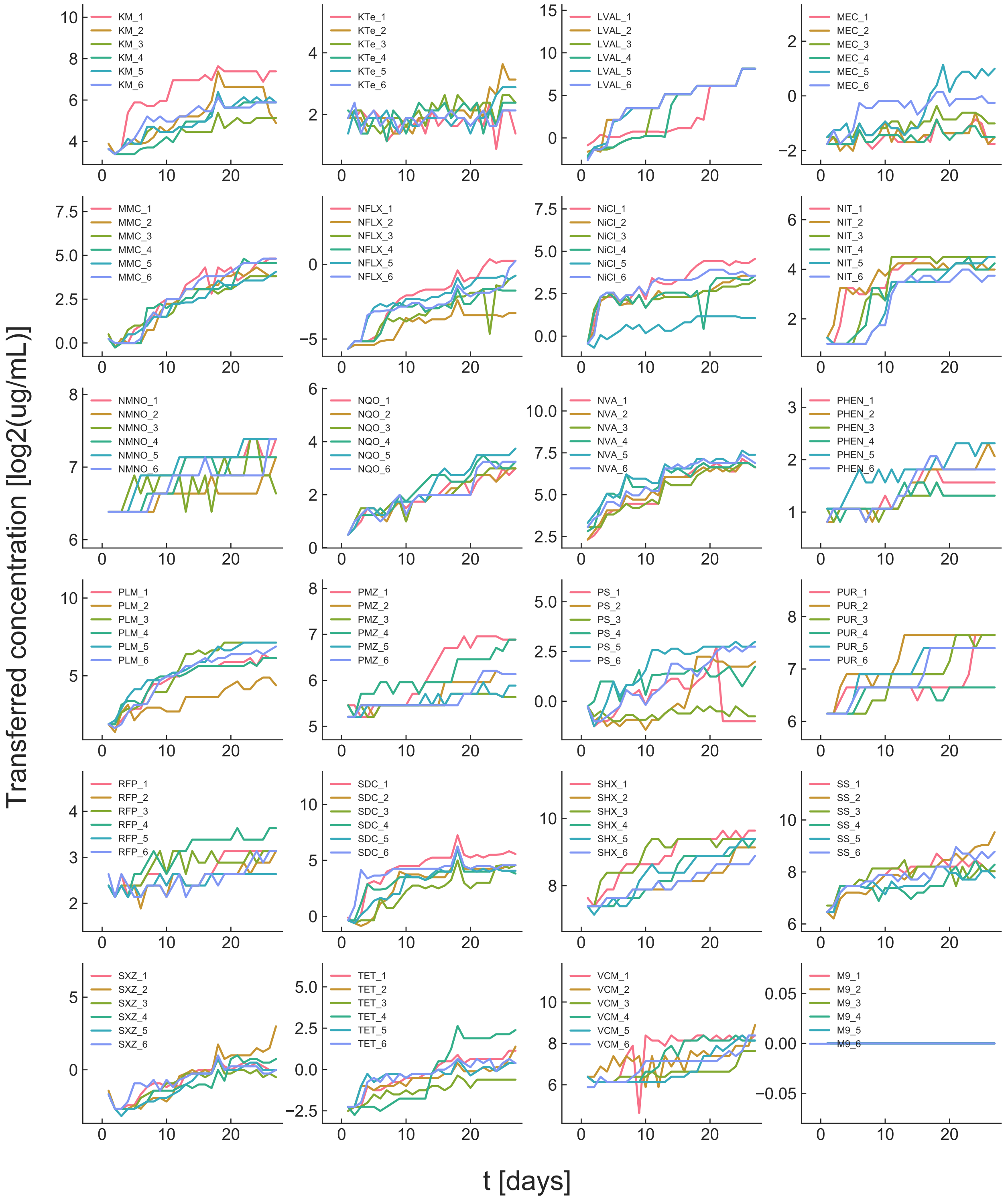

Fig. 4.24 Continued from Fig. 4.23. Figures modified and reproduced from [150].

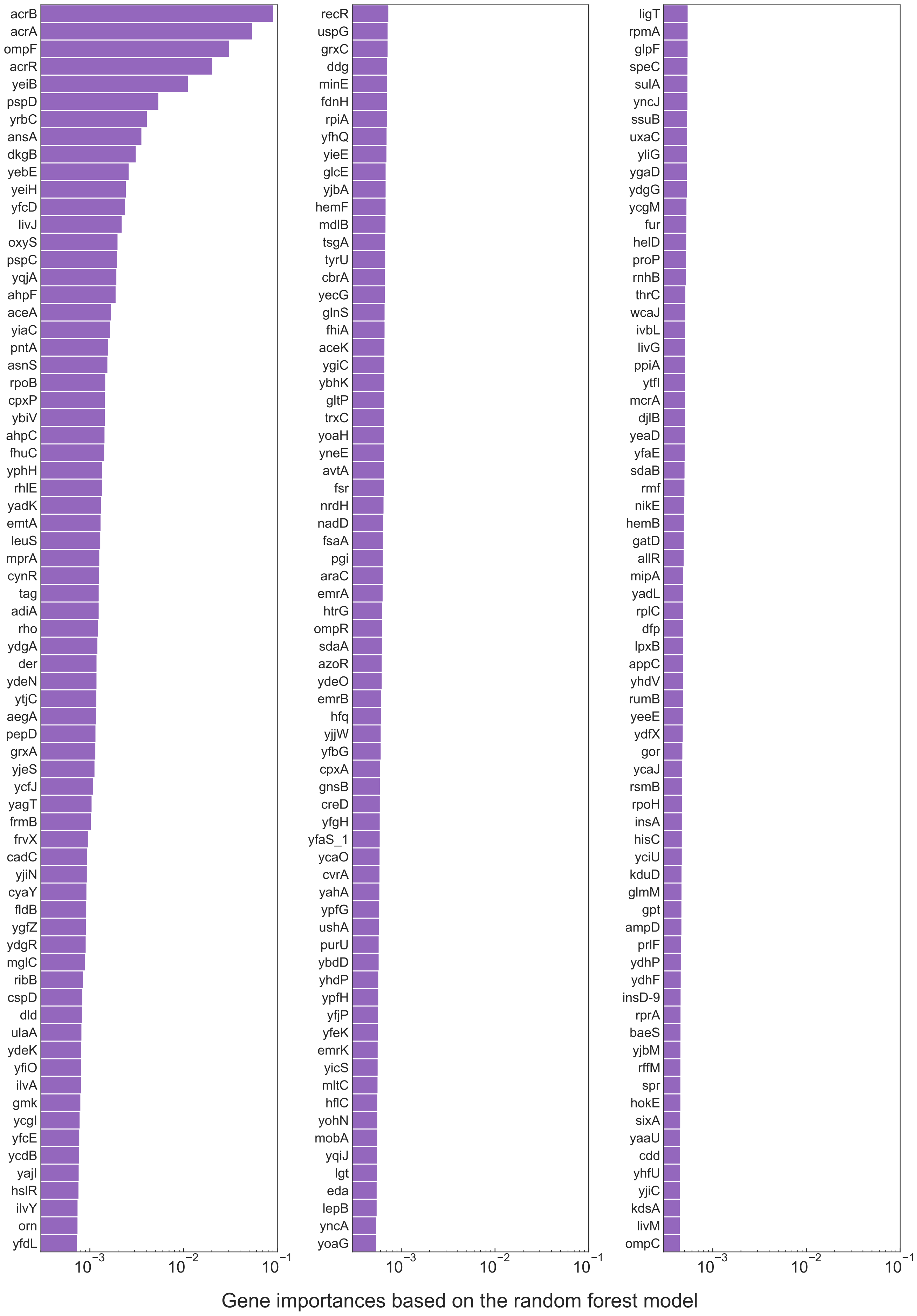

Fig. 4.25 Gene importances calculated by the random forest regression model. The 213 genes that were used for supervised PCA are shown.

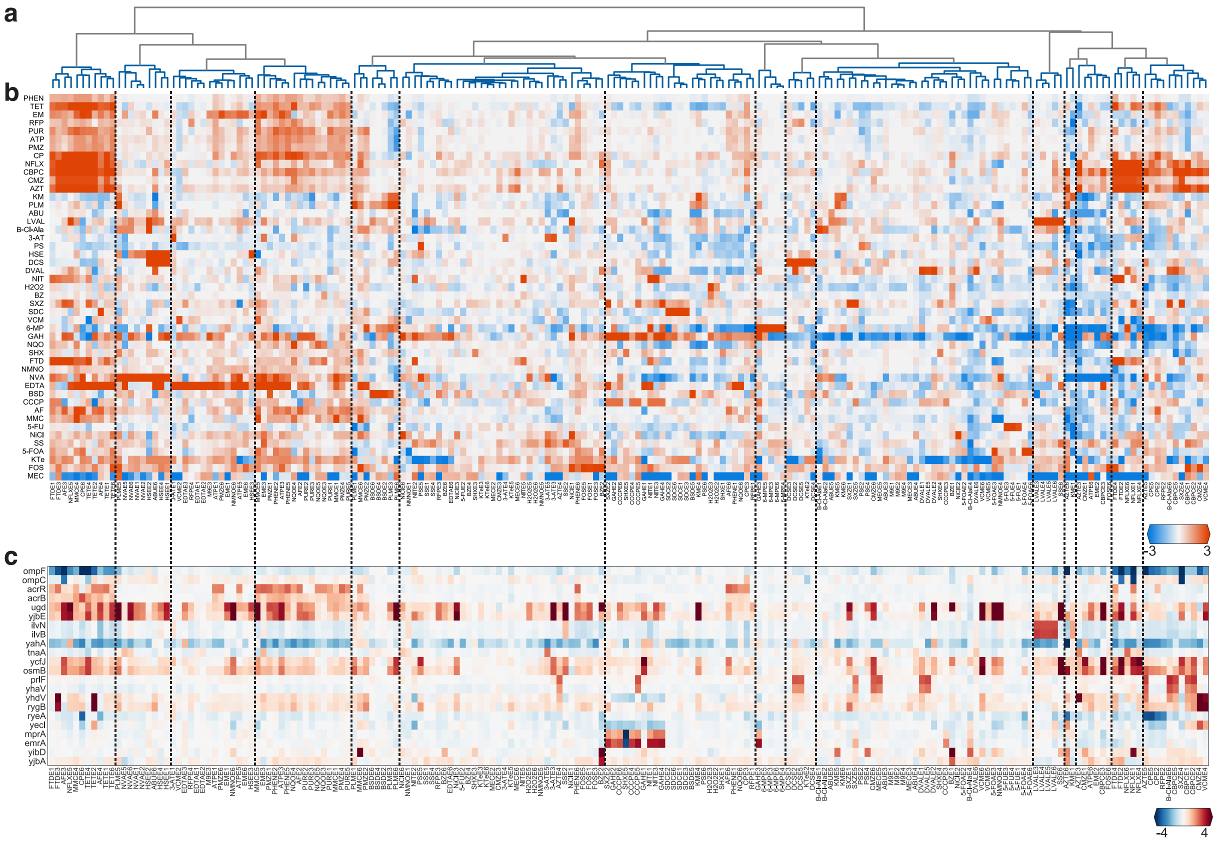

Fig. 4.26 **a**, Dendrogram of the hierarchical clustering performed in the resistance space. Ward's method was applied for clustering using the log2 resistance levels for all 47 stresses. **b**, Stress resistance profiles sorted based on hierarchical clustering in the resistance space. **c**, Gene expression profiles sorted based on hierarchical clustering in the resistance space. The same genes in Fig. 4.10b are shown.

Genes	Strains with a mutation in the corresponding gene
acrR	CPE6, CPE2, CPE3, AFE3, AFE2, AFE4, AFE6, NFLXE1, MMCE6, MMCE4, MMCE1, MMCE5, ATPE3, ATPE1, ATPE5, PHENE2, PHENE5, PHENE6, PHENE1, TETE4, TETE1, TETE2, TETE6, NQOE5, NQOE6, NQOE4, NQOE3, PMZE1, PMZE4, PMZE6, FTDE4, FTDE1, FTDE3, EME3, PURE3, PURE1
apt	3.ATE1, 3.ATE4
baeS	AZTE3, CBPCE1
corA	NiClE1, NiClE6, NiClE2, BZE2
cyaA	KME4, FOSE5, FOSE1, BZE2, ATPE3, PLME5, PHENE6
cycA	KTeE2, DCSE4, DCSE3, DCSE5, GAHE2, HSEE4, HSEE2, HSEE1, ABUE6
cyoA	KME5
cyoB	KME6, PLME3, PLME5, PLME6, PLME1
cyoE	KME1
dacA	MMCE5, VCME2, VCME5, GAHE1
dadA	DCSE2, DCSE5, GAHE2
dctA	5.FOAE4, 5.FOAE6, 5.FOAE3
folM	SDCE6, SDCE2, SDCE3, SXZE2, SXZE4, SXZE1
frlA	SDCE2, SDCE3, GAHE3
gabP	SHXE3, SHXE1, SHXE5, SHXE4
gadB	GAHE1, GAHE4
glnP	GAHE1, GAHE4
glpT	FOSE3, FOSE6
glyT	NVAE1, NVAE2
gshA	DCSE2, GAHE3, GAHE1, DVALE6, DVALE2
gyrA	AZTE3, NFLXE1, NFLXE6, NFLXE4
hisR	3.ATE4
hisS	3.ATE3, GAHE2
hrpA	MMCE4, MMCE1, NQOE5
ilvL	AFE6, SSE1, GAHE4
iscR	PLME6, PLME1
livM	B.Cl.AlaE1, B.Cl.AlaE6
lon	GAHE2, 5.FOAE3, NMNOE6, NMNOE4
metN	GAHE3, GAHE2, GAHE4
mipA	AFE6, NFLXE1, NFLXE4
mprA	SXZE2, SXZE5, NITE3, NITE1, GAHE3, GAHE2, GAHE4, CCCPE4, CCCPE5, CCCPE6, CCCPE3

Table 4.2 The full list of the 64 representative mutated genes which were introduced to the MAGE mutant strains.

Genes	Strains with a mutation in the corresponding gene
msbA	GAHE2, NMNOE5, NMNOE1
nfsA	NITE3, NITE2, NITE1, NITE5, FTDE4, FTDE1, FTDE2, FTDE3
nuoG	KME4
ompF	CPE5, CBPCE5, NFLXE1, NFLXE6, NFLXE5, NFLXE4, MMCE4, MECE6, TETE1, GAHE4, FTDE2, FTDE3
oppA	KME5, KME6, PLME3, BSDE2, BSDE6
oxyR	H2O2E3, H2O2E2, PURE2
potA	PLME3, PLME6, PLME1, BSDE5, BSDE2, BSDE6
potH	MMCE1
prlF	CMZE6, B.Cl.AlaE6, AZTE5, AZTE4, CBPCE5, CBPCE1, CBPCE2, SXZE5, DCSE2, DCSE5, MECE5, CCCPE5, DVALE1, DVALE5
ptsP	KME5, PLME5, PLME6, PLME1
purR	6.MPE6, 6.MPE4, GAHE3
rfe	KME4, NiClE3, VCME4, PLME5, PLME6
rhlB	5.FUE2, LVALE3, LVALE4, LVALE5, SDCE2, SDCE3
rne	LVALE6, KTeE5
rob	AFE4, TETE4, TETE6, PMZE4
rplM	H2O2E5, EME6, EME2, NMNOE5
rpoB	B.Cl.AlaE6, KME5, MMCE6, PLME5, NVAE6, TETE4, TETE2, NQOE4, PMZE2, NITE1, GAHE3, HSEE2, ABUE4
rpoC	B.Cl.AlaE1, B.Cl.AlaE2, AZTE5, FOSE5, MMCE5, PLME1
rssB	CPE6, CPE2, AZTE6, AZTE5, 5.FOAE4
sdaA	GAHE3, GAHE2, GAHE1, GAHE4
sdhA	PHENE6, NITE5, SSE6, SSE1, SSE4
serA	MMCE5, SHXE1, SHXE4
sodB	NQOE5, NQOE4
soxR	AFE3, MMCE4, MMCE5, PHENE2, FTDE1, PURE2, PURE3
sstT	B.Cl.AlaE1, B.Cl.AlaE2, NVAE1, NVAE2, GAHE2
sulA	MMCE6, MMCE4, MMCE1, MMCE5, GAHE3
uraA	5.FUE4, 5.FUE3, 5.FUE1
ybjK	AFE3, AFE2, AFE4, AFE6
ycbZ	NVAE5, M9E3, EME3, EME6, EME1, ABUE6, NMNOE5, NMNOE6
ycfQ	KTeE2, FTDE4
yhjE	DVALE6, DVALE5
yjcO	MMCE6, MMCE4
zupT	NQOE6, NQOE3, NITE5, GAHE4

Table 4.3 Continued from Table 4.2.

Chapter 5

Dynamical multi-stress data reveal the phenotypic fitness landscape of *Escherichia coli*

Methods for predicting and controlling trajectories of evolution are crucial not only for tackling drug resistance bacteria, but also for extending our horizons of evolutionary biology. Among the various concepts around evolution, the concept of the fitness landscape has been frequently invoked since it offers information on the predictability (e.g. directions, convergence) of evolution. Thus, constructing empirical landscapes from experimental data is an effective approach to develop methods for predicting evolution. Although previous studies have constructed fitness landscapes based on a comprehensive study of mutations on specific genes, the high dimensionality of the genotypic space keeps us from building a fitness landscape that is capable for predicting evolution for the whole cell which is constituted by a complex network of $\mathcal{O}(10^3)$ genes in the case of *Escherichia coli*. Here, we tackle this problem by inferring the fitness landscape for stress resistance evolution based on the phenotypic space which has much fewer effective dimensions than the genotypic space. By using stress resistance along the trajectories of evolution as a probe for both phenotypes and fitness, we infer the fitness landscape which underlies the resistance evolution dynamics. We show how the structures of the inferred landscapes correspond with biological mechanisms and tradeoffs for resistance evolution. We further discuss how the inferred phenotype-fitness landscapes could contribute to the prediction and control of evolution.

Related publications by author:

Junichiro Iwasawa, Tomoya Maeda, Hazuki Kotani, Masako Kawada and Chikara Furusawa, "Dynamical multi-stress data from laboratory evolution reveals the phenotypic landscape for *Escherichia coli*". *Manuscript in preparation.*

Contribution:

The author (J.I.) conducted all laboratory evolution experiments and the acquisition of stress resistance profiles under the advice of T.M.. The genotype data for the evolved strains were acquired by T.M., H.K., and M.K. All data analyses were performed by J.I. under the supervision of C.F..

The contents are censored since they are scheduled to be published in a scientific journal.
本章については、5年以内に雑誌等で刊行予定のため、非公開。

Chapter 6

Conclusion and Outlook

I will not follow where the path may lead,
but I will go where there is no path, and I will leave a trail.

from Muriel Strode, “Wind-Wafted Wild Flowers” (1903)

Throughout this dissertation, we have investigated the dynamics of stress resistance evolution for *Escherichia coli* through laboratory evolution and machine learning. As we have seen in Chapter 2, recent studies concerning massively parallel laboratory evolution experiments with high-throughput sequencing/phenotyping have provided an unprecedented amount of data on evolutionary dynamics. The main theme running through these studies was that despite the diversity in the sequence level, evolution could lead to convergence in coarse grained features such as phenotypes, suggesting the existence of evolutionary constraints. However, it is not always easy to decipher the evolutionary constraints from the acquired high-dimensional data. This is why we reviewed the recent advances in machine learning, especially the methods applicable to biological data where the $p \gg N$ problem is prevalent in Chapter 3. Equipped with the right tools, we would be able to identify the essential signals which is in our case the responsible evolutionary constraints, from the noisy high-dimensional data.

In Chapter 4, utilizing supervised PCA, we have seen that the *E. coli* which evolved under 48 various conditions converge into a few number of phenotypic classes. We have elucidated the underlying biological processes for each class through LDA, and further confirmed them by observing the resistance profiles of the reconstructed single mutant strains. In addition, consistent with previous studies, the elucidated phenotypic classes suggested that the phenotypes show better convergence than the genotypes, although the genotypes also showed a certain level of correspondence between the phenotypic classes. Importantly, we have shown that the supervised PCA space, a subspace of the whole gene expression space, corresponds well with the stress resistance space. This suggests that stress resistance dynamics of *E. coli* could be represented by a low-dimensional manifold within the high-dimensional gene expression space, and that supervised PCA could be an effective method to elucidate this manifold. Overall, our supervised PCA based analysis revealed the existence of evolutionary constraints in stress resistance evolution and the underlying biological mechanisms, providing the bases for the prediction and control of evolution.

Interestingly, the constraints underlying the stress resistance profiles did not necessarily depend on the action mechanisms of the stresses. Thus it is natural to ask, what are the origins of the evolutionary constraints for stress resistance evolution? We believe that chemical properties of the stresses could be one candidate. Several studies suggest that the occurrence of cross resistance between two stresses correlates with the similarity of their chemical fingerprints [58, 171]. Consistent with these reports, we found that strains with decreased expression of *ompF* tended to become resistant to chemicals with large molecular weight and/or high hydrophobicity. These results suggest that chemical properties of the stresses could be a factor that constrains resistance evolution. However, it should be noted that there exist cases that chemical properties solely cannot explain the observed constraints. Yen and Papin report that *Pseudomonas aeruginosa* with different adaptation histories acquired different resistance profiles through evolution [95], suggesting that not only chemical properties, but also the genetic background influences the patterns in resistance acquisition (see also [96]). Further investigation on how genetic backgrounds and chemical properties influence stress resistance acquisition could be a key to unravel the origins of evolutionary constraints for resistance evolution.

In Chapter 5, we proposed the phenotype-fitness landscape based on stress resistance profiles as a method for predicting and controlling evolution. Although fitness landscapes visualize the evolutionary constraints and provide information on the predictability of evolution, the high-dimensionality of the genotype space kept us from drawing a comprehensive genotype-fitness landscape capable of predicting evolution. We thus focused on the resistance profiles instead of genotypes since (i) evolution leads to better convergence on the phenotypic level, (ii) stress resistance profiles correspond with the subspace of the gene expression space making it an appropriate candidate for probing evolutionary dynamics. The directions of evolution predicted by the inferred fitness landscapes showed agreement with the observed experimental trajectories, suggesting that the landscapes are capable for predicting evolution. We further demonstrated that trajectories of evolution could be controlled in a fine-grained manner by utilizing the information from the inferred fitness landscapes. Indeed, the data we have used for inferring the fitness landscapes is based on a limited number of stresses and initial states (genetic backgrounds). Thus, there is a possibility that the structure of the landscapes changes when using different stresses or/and the genetic background of the strain changes. It is important to detect such changes since they indicate the existence of novel evolutionary constraints, and it should be noted that we will be able to actively investigate such novel constraints by using the inferred landscapes as hypotheses. We believe that the inferred fitness landscapes could accelerate the investigation of evolutionary constraints and lead to the true control of evolution.

Our results in Chapter 4 and 5 revealed the evolutionary constraints underlying evolution, providing an upbeat picture for predicting evolution. Nevertheless, our understanding of evolutionary constraints including its dependence on different genetic backgrounds is not complete and further studies are still needed for a comprehensive theory of evolution. Naively, this would lead to an exhausting investigation through an immense number of experimental combinations and thus, this might be where we could make use of theory. Theories could inform us of what we should be looking for and the possible structures of the data we collect. This is especially important when living in an era where we have access to exponentially more data than we had before. In the case of evolution, I personally believe theoretical studies of the genotype-phenotype map and their consequences for the phenotype-fitness landscape should pave the way for an efficient investigation of evolutionary constraints. Intense collaborations between theory, experiment and data analysis should lead to a systematic understanding of constraints and a predictive theory of biology.

Physics is a framework of explaining a phenomenon by "identifying details that matter" [172]. We are often cast on the curse to believe that we simply do not have enough data for a general theory. In the context of stress resistance evolution, for instance, there indeed are many unknown molecular mechanisms underlying the evolutionary dynamics which seem to keep us from building a unifying theory. While it may be true that we lack sufficient data, it still would be worth to take a look back at the 1800s when thermodynamics was constructed. Thermodynamics is one of the most successful examples where physicists were able to coarse grain and acquire a macroscopic theory of a many body system. The interesting part about thermodynamics is that it was formulated without knowing the microscopic details of the system. If we were not able to construct thermodynamics in the 1800s, we might have continued to dig into the microscopic details and might have been standing hopeless in front of the complex microscopic world [173]. Of course gases and living things are different, but still, we should sometimes recall the importance of taking a step back and try to construct a theory by identifying the coarse grained, macroscopic variables that matter for the system. We hope that our study could be one of the stepping stones for constructing such a macroscopic theory of biological systems.

References

[1] E. Schrödinger, "What is life," 1944.

[2] W. Bialek, *Biophysics: searching for principles*. Princeton University Press, 2012.

[3] W. Bialek, "Perspectives on theory at the interface of physics and biology," *Reports on Progress in Physics*, vol. 81, no. 1, p. 012601, 2017.

[4] L. Galvani, "De viribus electricitatis in motu musculari. commentarius," *De Bonoiensi Scientiarum et Artium Intituo atque Academie Commentarii*, vol. 7, pp. 363–418, 1791.

[5] J. D. Watson and F. H. Crick, "Genetical implications of the structure of deoxyribonucleic acid," *Nature*, vol. 171, no. 4361, pp. 964–967, 1953.

[6] F. H. Crick, "On protein synthesis," in *Symposia of the Society for Experimental Biology*, vol. 12, p. 8, 1958.

[7] H. C. Berg, "Bacterial behaviour," *Nature*, vol. 254, no. 5499, pp. 389–392, 1975.

[8] H. C. Berg and E. M. Purcell, "Physics of chemoreception," *Biophysical Journal*, vol. 20, no. 2, pp. 193–219, 1977.

[9] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek, "Probing the limits to positional information," *Cell*, vol. 130, no. 1, pp. 153–164, 2007.

[10] G. Tkačik, C. G. Callan, and W. Bialek, "Information flow and optimization in transcriptional regulation," *Proceedings of the National Academy of Sciences*, vol. 105, no. 34, pp. 12265–12270, 2008.

[11] D. Fuller, W. Chen, M. Adler, A. Groisman, H. Levine, W.-J. Rappel, and W. F. Loomis, "External and internal constraints on eukaryotic chemotaxis," *Proceedings of the National Academy of Sciences*, vol. 107, no. 21, pp. 9656–9659, 2010.

[12] S. Uda, T. H. Saito, T. Kudo, T. Kokaji, T. Tsuchiya, H. Kubota, Y. Komori, Y. Ozaki, and S. Kuroda, "Robustness and compensation of information transmission of signaling pathways," *Science*, vol. 341, no. 6145, pp. 558–561, 2013.

[13] S. Ito and T. Sagawa, "Maxwell's demon in biochemical signal transduction with feedback loop," *Nature Communications*, vol. 6, no. 1, pp. 1–6, 2015.

[14] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer, "Multidimensional optimality of microbial metabolism," *Science*, vol. 336, no. 6081, pp. 601–604, 2012.

[15] T. Großkopf, J. Consuegra, J. Gaffé, J. C. Willison, R. E. Lenski, O. S. Soyer, and D. Schneider, "Metabolic modelling in a dynamic evolutionary framework predicts adaptive diversification of bacteria in a long-term evolution experiment," *BMC Evolutionary Biology*, vol. 16, no. 1, pp. 1–15, 2016.

[16] B. Niebel, S. Leupold, and M. Heinemann, "An upper limit on Gibbs energy dissipation governs cellular metabolism," *Nature Metabolism*, vol. 1, no. 1, pp. 125–132, 2019.

[17] J. F. Yamagishi, N. Saito, and K. Kaneko, "Advantage of Leakage of Essential Metabolites for Cells," *Physical Review Letters*, vol. 124, p. 048101, 2020.

[18] C. M. Waters and B. L. Bassler, "Quorum sensing: cell-to-cell communication in bacteria," *Annual Review of Cell and Developmental Biology*, vol. 21, pp. 319–346, 2005.

[19] E. Kussell and S. Leibler, "Phenotypic diversity, population growth, and information in fluctuating

environments," *Science*, vol. 309, no. 5743, pp. 2075–2078, 2005.

[20] T. Danino, O. Mondragón-Palomino, L. Tsimring, and J. Hasty, "A synchronized quorum of genetic clocks," *Nature*, vol. 463, no. 7279, pp. 326–330, 2010.

[21] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, "Statistical mechanics for natural flocks of birds," *Proceedings of the National Academy of Sciences*, vol. 109, no. 13, pp. 4786–4791, 2012.

[22] J. Gautrais, F. Ginelli, R. Fournier, S. Blanco, M. Soria, H. Chaté, and G. Theraulaz, "Deciphering interactions in moving animal groups," 2012.

[23] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, "Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[24] P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, M. Y. Byron, and A. P. Batista, "Neural constraints on learning," *Nature*, vol. 512, no. 7515, pp. 423–426, 2014.

[25] K. Kaneko and C. Furusawa, "Macroscopic theory for evolving biological systems akin to thermodynamics," *Annual Review of Biophysics*, vol. 47, pp. 273–290, 2018.

[26] K. R. Moon, J. S. Stanley III, D. Burkhardt, D. van Dijk, G. Wolf, and S. Krishnaswamy, "Manifold learning-based methods for analyzing single-cell RNA-sequencing data," *Current Opinion in Systems Biology*, vol. 7, pp. 36–46, 2018.

[27] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, "Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model," *Physical Review X*, vol. 10, p. 041044, 2020.

[28] G. Grégoire and H. Chaté, "Onset of Collective and Cohesive Motion," *Physical Review Letters*, vol. 92, p. 025702, 2004.

[29] M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, M. Rao, and R. A. Simha, "Hydrodynamics of soft active matter," *Review of Modern Physics*, vol. 85, pp. 1143–1189, 2013.

[30] S. Ramaswamy, "Active matter," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, no. 5, p. 054002, 2017.

[31] K. Kawaguchi, R. Kageyama, and M. Sano, "Topological defects control collective dynamics in neural progenitor cell cultures," *Nature*, vol. 545, no. 7654, pp. 327–331, 2017.

[32] T. Mano, J.-B. Delfau, J. Iwasawa, and M. Sano, "Optimal run-and-tumble–based transportation of a Janus particle with active steering," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. E2580–E2589, 2017.

[33] D. Nishiguchi, J. Iwasawa, H.-R. Jiang, and M. Sano, "Flagellar dynamics of chains of active Janus particles fueled by an AC electric field," *New Journal of Physics*, vol. 20, no. 1, p. 015002, 2018.

[34] G. Duclos, R. Adkins, D. Banerjee, M. S. Peterson, M. Varghese, I. Kolvin, A. Baskaran, R. A. Pelcovits, T. R. Powers, A. Baskaran, *et al.*, "Topological structure and dynamics of three-dimensional active nematics," *Science*, vol. 367, no. 6482, pp. 1120–1124, 2020.

[35] Y. Maroudas-Sacks, L. Garion, L. Shani-Zerbib, A. Livshits, E. Braun, and K. Keren, "Topological defects in the nematic order of actin fibres as organization centres of Hydra morphogenesis," *Nature Physics*, 2020.

[36] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel Type of Phase Transition in a System of Self-Driven Particles," *Physical Review Letters*, vol. 75, pp. 1226–1229, 1995.

[37] J. Toner and Y. Tu, "Long-Range Order in a Two-Dimensional Dynamical XY Model: How Birds Fly Together," *Physical Review Letters*, vol. 75, pp. 4326–4329, 1995.

[38] J. Toner and Y. Tu, "Flocks, herds, and schools: A quantitative theory of flocking," *Physical Review E*, vol. 58, pp. 4828–4858, 1998.

[39] J. Toner, "Reanalysis of the hydrodynamic theory of fluid, polar-ordered flocks," *Physical Review E*, vol. 86, p. 031918, 2012.

[40] D. Nishiguchi, K. H. Nagai, H. Chaté, and M. Sano, "Long-range nematic order and anomalous fluctuations in suspensions of swimming filamentous bacteria," *Physical Review E*, vol. 95, p. 020601, 2017.

[41] D. Geyer, A. Morin, and D. Bartolo, "Sounds and hydrodynamics of polar active fluids," *Nature Materials*, vol. 17, no. 9, pp. 789–793, 2018.

[42] S. Tanida, K. Furuta, K. Nishikawa, T. Hiraiwa, H. Kojima, K. Oiwa, and M. Sano, "Gliding

filament system giving both global orientational order and clusters in collective motion," *Physical Review E*, vol. 101, p. 032607, 2020.

[43] H. Soni, N. Kumar, J. Nambisan, R. K. Gupta, A. Sood, and S. Ramaswamy, "Phases and excitations of active rod–bead mixtures: simulations and experiments," *Soft Matter*, 2020.

[44] J. Iwasawa, D. Nishiguchi, and M. Sano, "Algebraic correlations and anomalous fluctuations in ordered flocks of Janus particles fueled by an AC electric field," *arXiv e-prints*, p. arXiv:2011.14548, 2020.

[45] S. J. Gould, *Wonderful Life: the Burgess Shale and the Nature of History*. WW Norton & Company, 1990.

[46] T. M. Conrad, N. E. Lewis, and B. Ø. Palsson, "Microbial laboratory evolution in the era of genome-scale science," *Molecular Systems Biology*, vol. 7, no. 1, p. 509, 2011.

[47] A. E. Lobkovsky and E. V. Koonin, "Replaying the tape of life: quantification of the predictability of evolution," *Frontiers in Genetics*, vol. 3, p. 246, 2012.

[48] J. B. Losos, *Improbable destinies: Fate, chance, and the future of evolution*. Penguin, 2017.

[49] M. Lässig, V. Mustonen, and A. M. Walczak, "Predicting evolution," *Nature Ecology & Evolution*, vol. 1, no. 3, pp. 1–9, 2017.

[50] Z. D. Blount, R. E. Lenski, and J. B. Losos, "Contingency and determinism in evolution: Replaying life's tape," *Science*, vol. 362, no. 6415, 2018.

[51] K. Bush, P. Courvalin, G. Dantas, J. Davies, B. Eisenstein, P. Huovinen, G. A. Jacoby, R. Kishony, B. N. Kreiswirth, E. Kutter, *et al.*, "Tackling antibiotic resistance," *Nature Reviews Microbiology*, vol. 9, no. 12, pp. 894–896, 2011.

[52] A. C. Palmer and R. Kishony, "Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance," *Nature Reviews Genetics*, vol. 14, no. 4, pp. 243–248, 2013.

[53] M. Baym, L. K. Stone, and R. Kishony, "Multidrug evolutionary strategies to reverse antibiotic resistance," *Science*, vol. 351, no. 6268, 2016.

[54] O. Tenaillon, A. Rodríguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, and B. S. Gaut, "The molecular diversity of adaptive convergence," *Science*, vol. 335, no. 6067, pp. 457–461, 2012.

[55] E. Toprak, A. Veres, J.-B. Michel, R. Chait, D. L. Hartl, and R. Kishony, "Evolutionary paths to antibiotic resistance under dynamically sustained drug selection," *Nature Genetics*, vol. 44, no. 1, pp. 101–105, 2012.

[56] L. Imamovic and M. O. Sommer, "Use of collateral sensitivity networks to design drug cycling protocols that avoid resistance development," *Science Translational Medicine*, vol. 5, no. 204, pp. 204ra132–204ra132, 2013.

[57] V. Lázár, G. Pal Singh, R. Spohn, I. Nagy, B. Horváth, M. Hrtyan, R. Busa-Fekete, B. Bogos, O. Méhi, B. Csörgő, *et al.*, "Bacterial evolution of antibiotic hypersensitivity," *Molecular Systems Biology*, vol. 9, no. 1, p. 700, 2013.

[58] V. Lázár, I. Nagy, R. Spohn, B. Csörgő, Á. Györkei, Á. Nyerges, B. Horváth, A. Vörös, R. Busa-Fekete, M. Hrtyan, *et al.*, "Genome-wide analysis captures the determinants of the antibiotic cross-resistance interaction network," *Nature Communications*, vol. 5, no. 1, pp. 1–12, 2014.

[59] S. Suzuki, T. Horinouchi, and C. Furusawa, "Prediction of antibiotic resistance by gene expression profiles," *Nature Communications*, vol. 5, no. 1, pp. 1–12, 2014.

[60] L. Imamovic, M. M. H. Ellabaan, A. M. D. Machado, L. Citterio, T. Wulff, S. Molin, H. K. Johansen, and M. O. A. Sommer, "Drug-driven phenotypic convergence supports rational treatment strategies of chronic infections," *Cell*, vol. 172, no. 1-2, pp. 121–134, 2018.

[61] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[62] B. Efron and T. Hastie, *Computer age statistical inference*, vol. 5. Cambridge University Press, 2016.

[63] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, "A high-bias, low-variance introduction to machine learning for physicists," *Physics Reports*, vol. 810, pp. 1–124, 2019.

[64] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[65] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[66] E. Bair and R. Tibshirani, “Semi-supervised methods to predict patient survival from gene expression data,” *PLoS Biology*, vol. 2, no. 4, p. e108, 2004.

[67] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, “Prediction by supervised principal components,” *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 119–137, 2006.

[68] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy, “Visualizing structure and transitions in high-dimensional biological data,” *Nature Biotechnology*, vol. 37, no. 12, pp. 1482–1492, 2019.

[69] J. Iwasawa, Y. Hirano, and Y. Sugawara, “Label-Efficient Multi-Task Segmentation using Contrastive Learning,” *arXiv e-prints*, p. arXiv:2009.11160, 2020.

[70] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, “Darwinian evolution can follow only very few mutational paths to fitter proteins,” *Science*, vol. 312, no. 5770, pp. 111–114, 2006.

[71] A. C. Palmer, E. Toprak, M. Baym, S. Kim, A. Veres, S. Bershtein, and R. Kishony, “Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes,” *Nature Communications*, vol. 6, no. 1, pp. 1–8, 2015.

[72] K. Kaneko, C. Furusawa, and T. Yomo, “Universal Relationship in Gene-Expression Changes for Cells in Steady-Growth State,” *Physical Review X*, vol. 5, p. 011014, 2015.

[73] C. Furusawa and K. Kaneko, “Formation of dominant mode by evolution in biological systems,” *Physical Review E*, vol. 97, p. 042410, 2018.

[74] T. U. Sato and K. Kaneko, “Evolutionary dimension reduction in phenotypic space,” *Physical Review Research*, vol. 2, p. 013197, 2020.

[75] M. Tikhonov, S. Kachru, and D. S. Fisher, “A model for the interplay between plastic tradeoffs and evolution in changing environments,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 16, pp. 8934–8940, 2020.

[76] G. Kinsler, K. Geiler-Samerotte, and D. A. Petrov, “Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation,” *eLife*, vol. 9, p. e61271, 2020.

[77] F. Pinheiro, O. Warsi, D. I. Andersson, and M. Lässig, “Predicting trajectories and mechanisms of antibiotic resistance evolution,” *arXiv e-prints*, p. arXiv:2007.01245, 2020.

[78] C. Furusawa, T. Horinouchi, and T. Maeda, “Toward prediction and control of antibiotic-resistance evolution,” *Current opinion in Biotechnology*, vol. 54, pp. 45–49, 2018.

[79] P. Center, A. Alliance, J. Network, A. E. Hub, and A. Insurance, “Antibiotic resistance,” *Public Health*, 2019.

[80] V. M. D’Costa, C. E. King, L. Kalan, M. Morar, W. W. Sung, C. Schwarz, D. Froese, G. Zazula, F. Calmels, R. Debruyne, *et al.*, “Antibiotic resistance is ancient,” *Nature*, vol. 477, no. 7365, pp. 457–461, 2011.

[81] M. O. Sommer, C. Munck, R. V. Toft-Kehler, and D. I. Andersson, “Prediction of antibiotic resistance: time for a new preclinical paradigm?,” *Nature Reviews Microbiology*, vol. 15, no. 11, pp. 689–696, 2017.

[82] R. E. Lenski, “Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations,” *The ISME Journal*, vol. 11, no. 10, pp. 2181–2194, 2017.

[83] J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim, “Genome evolution and adaptation in a long-term experiment with *Escherichia coli*,” *Nature*, vol. 461, no. 7268, pp. 1243–1247, 2009.

[84] B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski, and M. M. Desai, “The dynamics of molecular evolution over 60,000 generations,” *Nature*, vol. 551, no. 7678, pp. 45–50, 2017.

[85] Z. D. Blount, C. Z. Borland, and R. E. Lenski, “Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 23, pp. 7899–7906, 2008.

[86] E. Toprak, A. Veres, S. Yildiz, J. M. Pedraza, R. Chait, J. Paulsson, and R. Kishony, “Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition,” *Nature Protocols*, vol. 8, no. 3, pp. 555–567, 2013.

[87] M. Lukačišinová, B. Fernando, and T. Bollenbach, “Highly parallel lab evolution reveals that epistasis can curb the evolution of antibiotic resistance,” *Nature Communications*, vol. 11, no. 1, pp. 1–14, 2020.

[88] M. Zampieri, T. Enke, V. Chubukov, V. Ricci, L. Piddock, and U. Sauer, “Metabolic constraints on the evolution of antibiotic resistance,” *Molecular Systems Biology*, vol. 13, no. 3, p. 917, 2017.

[89] R. M. Lennen, K. Jensen, E. T. Mohammed, S. Malla, R. A. Börner, K. Chekina, E. Özdemir, I. Bonde, A. Koza, J. Maury, *et al.*, “Adaptive laboratory evolution reveals general and specific chemical tolerance mechanisms and enhances biochemical production,” *bioRxiv*, p. 634105, 2019.

[90] A. Porse, L. J. Jahn, M. M. Ellabaan, and M. O. Sommer, “Dominant resistance and negative epistasis can limit the co-selection of de novo resistance mutations and antibiotic resistance genes,” *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020.

[91] T. Horinouchi, S. Suzuki, H. Kotani, K. Tanabe, N. Sakata, H. Shimizu, and C. Furusawa, “Prediction of cross-resistance and collateral sensitivity by gene expression profiles and genomic mutations,” *Scientific Reports*, vol. 7, no. 1, pp. 1–11, 2017.

[92] S. S. Fong, A. R. Joyce, and B. Ø. Palsson, “Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states,” *Genome Research*, vol. 15, no. 10, pp. 1365–1372, 2005.

[93] T. Horinouchi, S. Suzuki, T. Hirasawa, N. Ono, T. Yomo, H. Shimizu, and C. Furusawa, “Phenotypic convergence in bacterial adaptive evolution to ethanol stress,” *BMC Evolutionary Biology*, vol. 15, no. 1, p. 180, 2015.

[94] Y. J. Jiao, M. Baym, A. Veres, and R. Kishony, “Population diversity jeopardizes the efficacy of antibiotic cycling,” *bioRxiv*, p. 082107, 2016.

[95] P. Yen and J. A. Papin, “History of antibiotic adaptation influences microbial evolutionary dynamics during subsequent treatment,” *PLoS Biology*, vol. 15, no. 8, p. e2001586, 2017.

[96] M. Hoeksema, M. J. Jonker, S. Brul, and B. H. Ter Kuile, “Effects of a previously selected antibiotic resistance on mutations acquired during development of a second resistance in *Escherichia coli*,” *BMC Genomics*, vol. 20, no. 1, p. 284, 2019.

[97] S. Wright, “The roles of mutation, inbreeding, crossbreeding, and selection in evolution,” *Proceedings of The Sixth International Congress of Genetics*, vol. 1, pp. 356–366, 1932.

[98] S. Wright, “Surfaces of selective value revisited,” *The American Naturalist*, vol. 131, no. 1, pp. 115–123, 1988.

[99] S. Kauffman and S. Levin, “Towards a general theory of adaptive walks on rugged landscapes,” *Journal of Theoretical Biology*, vol. 128, no. 1, pp. 11–45, 1987.

[100] V. Mustonen and M. Lässig, “From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation,” *Trends in Genetics*, vol. 25, no. 3, pp. 111–119, 2009.

[101] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, “Predictability of evolutionary trajectories in fitness landscapes,” *PLoS Computational Biology*, vol. 7, no. 12, p. e1002302, 2011.

[102] S. Wang and L. Dai, “Evolving generalists in switching rugged landscapes,” *PLoS Computational Biology*, vol. 15, no. 10, p. e1007320, 2019.

[103] V. Sachdeva, K. Husain, J. Sheng, S. Wang, and A. Murugan, “Tuning environmental timescales to evolve and maintain generalists,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12693–12699, 2020.

[104] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans, “Empirical fitness landscapes reveal accessible evolutionary paths,” *Nature*, vol. 445, no. 7126, pp. 383–386, 2007.

[105] J. A. G. De Visser and J. Krug, “Empirical fitness landscapes and the predictability of evolution,” *Nature Reviews Genetics*, vol. 15, no. 7, pp. 480–490, 2014.

[106] J. A. G. de Visser, S. F. Elena, I. Fragata, and S. Matuszewski, “The utility of fitness landscapes and big data for predicting evolution,” 2018.

[107] J. Zheng, J. L. Payne, and A. Wagner, “Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks,” *Science*, vol. 365, no. 6451, pp. 347–353, 2019.

[108] D. Nichol, J. Rutter, C. Bryant, A. M. Hujer, S. Lek, M. D. Adams, P. Jeavons, A. R. Anderson, R. A. Bonomo, and J. G. Scott, “Antibiotic collateral sensitivity is contingent on the repeatability of evolution,” *Nature Communications*, vol. 10, no. 1, pp. 1–10, 2019.

[109] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. de Visser, “Quantifying the adaptive potential of an antibiotic resistance enzyme,” *PLoS Genetics*, vol. 8, no. 6, p. e1002783, 2012.

[110] H. Jacquier, A. Birgy, H. Le Nagard, Y. Mechulam, E. Schmitt, J. Glodt, B. Bercot, E. Petit, J. Poulain, G. Barnaud, *et al.*, “Capturing the mutational landscape of the beta-lactamase TEM-1,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 32, pp. 13067–13072, 2013.

[111] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, “A comprehensive, high-resolution map of a gene’s fitness landscape,” *Molecular Biology and Evolution*, vol. 31, no. 6, pp. 1581–1592, 2014.

[112] M. F. Schenk, I. G. Szendro, M. L. Salverda, J. Krug, and J. A. G. De Visser, “Patterns of epistasis between beneficial mutations in an antibiotic resistance gene,” *Molecular Biology and Evolution*, vol. 30, no. 8, pp. 1779–1787, 2013.

[113] H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segrè, and C. J. Marx, “Diminishing returns epistasis among beneficial mutations decelerates adaptation,” *Science*, vol. 332, no. 6034, pp. 1190–1192, 2011.

[114] S. Kryazhimskiy, D. P. Rice, E. R. Jerison, and M. M. Desai, “Global epistasis makes adaptation predictable despite sequence-level stochasticity,” *Science*, vol. 344, no. 6191, pp. 1519–1522, 2014.

[115] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[116] H. Harold, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[117] D. Kobak and P. Berens, “The art of using t-SNE for single-cell transcriptomics,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[118] K. J. Kobayashi-Kirschvink, H. Nakaoka, A. Oda, F. K. Ken-ichiro, K. Nosho, H. Fukushima, Y. Kanesaki, S. Yajima, H. Masaki, K. Ohta, *et al.*, “Linear Regression Links Transcriptomic Data and Cellular Raman Spectra,” *Cell Systems*, vol. 7, no. 1, pp. 104–117, 2018.

[119] A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, and B. O. Palsson, “The *Escherichia coli* transcriptome mostly consists of independently regulated modules,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[120] X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, J. T. Yurkovich, C. J. Lloyd, Y. Gao, L. Yang, and B. O. Palsson, “Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. 10286–10291, 2017.

[121] H. Matsumoto, T. Hayashi, H. Ozaki, K. Tsuyuzaki, M. Umeda, T. Iida, M. Nakamura, H. Okano, and I. Nikaido, “An NMF-based approach to discover overlooked differentially expressed gene regions from single-cell RNA-seq data,” *NAR Genomics and Bioinformatics*, vol. 2, no. 1, 2019. lqz020.

[122] H. Jeckel, E. Jelli, R. Hartmann, P. K. Singh, R. Mok, J. F. Totz, L. Vidakovic, B. Eckhardt, J. Dunkel, and K. Drescher, “Learning the space-time phase diagram of bacterial swarm expansion,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 5, pp. 1489–1494, 2019.

[123] J. A. Briggs, C. Weinreb, D. E. Wagner, S. Megason, L. Peshkin, M. W. Kirschner, and A. M. Klein, “The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution,” *Science*, vol. 360, no. 6392, 2018.

[124] D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, and A. M. Klein, “Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo,” *Science*, vol. 360, no. 6392, pp. 981–987, 2018.

[125] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv e-prints*, p. arXiv:1802.03426, 2018.

[126] T. Sainburg, L. McInnes, and T. Q. Gentner, “Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning,” *arXiv e-prints*, p. arXiv:2009.12981, 2020.

[127] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein, “Fundamental limits on dynamic inference from single-cell snapshots,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 10, pp. E2467–E2476, 2018.

[128] D. E. Wagner and A. M. Klein, “Lineage tracing meets single-cell omics: opportunities and challenges,” *Nature Reviews Genetics*, pp. 1–18, 2020.

[129] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[130] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,”

Technometrics, vol. 12, no. 1, pp. 55–67, 1970.

[131] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[132] G. Louppe, "Understanding Random Forests: From Theory to Practice," *arXiv e-prints*, p. arXiv:1407.7502, 2014.

[133] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing systems*, 2012.

[134] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nature Methods*, pp. 1–14, 2019.

[135] M. Amodio and S. Krishnaswamy, "MAGAN: Aligning biological manifolds," *arXiv preprint arXiv:1803.00385*, 2018.

[136] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing systems*, pp. 2672–2680, 2014.

[137] M. Brbić, M. Zitnik, S. Wang, A. O. Pisco, R. B. Altman, S. Darmanis, and J. Leskovec, "MARS: discovering novel cell types across heterogeneous single-cell experiments," *Nature Methods*, vol. 17, no. 12, pp. 1200–1206, 2020.

[138] D. Imoto, N. Saito, A. Nakajima, G. Honda, M. Ishida, T. Sugita, S. Ishihara, K. Katagiri, C. Okimura, Y. Iwadate, *et al.*, "Comparative mapping of crawling-cell morphodynamics in deep learning-based feature space," *bioRxiv*, 2020.

[139] M. Tsutsumi, D. Koyabu, N. Saito, and C. Furusawa, "Characterization of biological morphology by using machine learning." *To be published*.

[140] F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," *Science*, vol. 365, no. 6457, p. eaaw1147, 2019.

[141] B. Hie, E. Zhong, B. Bryson, and B. Berger, "Learning Mutational Semantics," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[142] B. Hie, E. D. Zhong, B. Berger, and B. Bryson, "Learning the language of viral evolution and escape," *Science*, vol. 371, no. 6526, pp. 284–288, 2021.

[143] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[144] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International Conference on Machine Learning*, pp. 4182–4192, PMLR, 2020.

[145] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[146] B. Plackett, "Why big pharma has abandoned antibiotics," *Nature*, vol. 586, no. 7830, pp. S50–S52, 2020.

[147] M. Baym, T. D. Lieberman, E. D. Kelsic, R. Chait, R. Gross, I. Yelin, and R. Kishony, "Spatiotemporal microbial evolution on antibiotic landscapes," *Science*, vol. 353, no. 6304, pp. 1147–1151, 2016.

[148] A. Wagner, *The origins of evolutionary innovations: a theory of transformative change in living systems*. OUP Oxford, 2011.

[149] B. Papp, R. A. Notebaart, and C. Pál, "Systems-biology approaches for predicting genomic evolution," *Nature Reviews Genetics*, vol. 12, no. 9, pp. 591–602, 2011.

[150] T. Maeda, J. Iwasawa, H. Kotani, N. Sakata, M. Kawada, T. Horinouchi, A. Sakai, K. Tanabe, and C. Furusawa, "High-throughput laboratory evolution reveals evolutionary constraints in *Escherichia coli*," *Nature Communications*, vol. 11, no. 1, p. 5970, 2020.

[151] T. Horinouchi, T. Minamoto, S. Suzuki, H. Shimizu, and C. Furusawa, "Development of an automated culture system for laboratory evolution," *Journal of Laboratory Automation*, vol. 19, no. 5, pp. 478–482, 2014.

[152] H. H. Lee, M. N. Molla, C. R. Cantor, and J. J. Collins, "Bacterial charity work leads to population-wide resistance," *Nature*, vol. 467, no. 7311, pp. 82–85, 2010.

[153] S. Norioka, G. Ramakrishnan, K. Ikenaka, and M. Inouye, "Interaction of a transcriptional activator, OmpR, with reciprocally osmoregulated genes, ompF and ompC, of *Escherichia coli*.," *Journal of Biological Chemistry*, vol. 261, no. 36, pp. 17113–17119, 1986.

[154] K. E. Gibson and T. J. Silhavy, "The LysR homolog LrhA promotes RpoS degradation by modulat-

ing activity of the response regulator sprE," *Journal of Bacteriology*, vol. 181, no. 2, pp. 563–571, 1999.

[155] A. H. Delcour, "Outer membrane permeability and antibiotic resistance," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1794, no. 5, pp. 808–816, 2009.

[156] 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.

[157] O. Lomovskaya, K. Lewis, and A. Matin, "EmrR is a negative regulator of the *Escherichia coli* multidrug resistance pump EmrAB.," *Journal of Bacteriology*, vol. 177, no. 9, pp. 2328–2334, 1995.

[158] K. J. Harder, H. Nikaido, and M. Matsuhashi, "Mutants of *Escherichia coli* that are resistant to certain beta-lactam compounds lack the ompF porin," *Antimicrobial Agents and Chemotherapy*, vol. 20, no. 4, pp. 549–552, 1981.

[159] C. Balague and E. G. Vescovi, "Activation of multiple antibiotic resistance in uropathogenic *Escherichia coli* strains by aryloxoalcanoic acid compounds," *Antimicrobial Agents and Chemotherapy*, vol. 45, no. 6, pp. 1815–1822, 2001.

[160] H. Okusu, D. Ma, and H. Nikaido, "AcrAB efflux pump plays a major role in the antibiotic resistance phenotype of *Escherichia coli* multiple-antibiotic-resistance (Mar) mutants.," *Journal of Bacteriology*, vol. 178, no. 1, pp. 306–308, 1996.

[161] D. Ma, D. Cook, M. Alberti, N. Pon, H. Nikaido, and J. Hearst, "Molecular cloning and characterization of acrA and acrE genes of *Escherichia coli*.," *Journal of Bacteriology*, vol. 175, no. 19, pp. 6299–6313, 1993.

[162] A. Mazzariol, G. Cornaglia, and H. Nikaido, "Contributions of the AmpC β-lactamase and the AcrAB multidrug efflux system in intrinsic resistance of *Escherichia coli* K-12 to β-lactams," *Antimicrobial Agents and Chemotherapy*, vol. 44, no. 5, pp. 1387–1390, 2000.

[163] J. L. Radzikowski, S. Vedelaar, D. Siegel, Á. D. Ortega, A. Schmidt, and M. Heinemann, "Bacterial persistence is an active σS stress response to metabolic flux limitation," *Molecular Systems Biology*, vol. 12, no. 9, p. 882, 2016.

[164] T. Ferenci, "Maintaining a healthy SPANC balance through regulatory and mutational adaptation," *Molecular Microbiology*, vol. 57, no. 1, pp. 1–8, 2005.

[165] O. Schmidt, V. J. Schuenemann, N. J. Hand, T. J. Silhavy, J. Martin, A. N. Lupas, and S. Djuranovic, "prlF and yhaV encode a new toxin–antitoxin system in *Escherichia coli*," *Journal of Molecular Biology*, vol. 372, no. 4, pp. 894–905, 2007.

[166] J. A. Imlay and S. Linn, "Mutagenesis and stress responses induced in *Escherichia coli* by hydrogen peroxide.," *Journal of Bacteriology*, vol. 169, no. 7, pp. 2967–2976, 1987.

[167] A. Wong, "Epistasis and the evolution of antimicrobial resistance," *Frontiers in Microbiology*, vol. 8, p. 246, 2017.

[168] D. Ghosh, B. Veeraraghavan, R. Elangovan, and P. Vivekanandan, "Antibiotic resistance and epigenetics: more to it than meets the eye," *Antimicrobial Agents and Chemotherapy*, vol. 64, no. 2, 2020.

[169] Z. Yao, D. Kahne, and R. Kishony, "Distinct single-cell morphological dynamics under beta-lactam antibiotics," *Molecular Cell*, vol. 48, no. 5, pp. 705–712, 2012.

[170] Z. Zhu, D. Surujon, J. C. Ortiz-Marquez, W. Huo, R. R. Isberg, J. Bento, and T. van Opijnen, "Entropy of a bacterial stress response is a generalizable predictor for fitness and antibiotic sensitivity," *Nature Communications*, vol. 11, no. 1, pp. 1–15, 2020.

[171] C. Barbosa, V. Trebosc, C. Kemmer, P. Rosenstiel, R. Beardmore, H. Schulenburg, and G. Jansen, "Alternative evolutionary paths to bacterial antibiotic resistance cause distinct collateral effects," *Molecular Biology and Evolution*, vol. 34, no. 9, pp. 2229–2244, 2017.

[172] M. Tikhonov, "Identifying details that matter: fruit fly development, genetic regulation, and microbial ecology," *Ph.D. Dissertation*, 2014.

[173] This was pointed out by Shin-ichi Sasa during a lecture by Kyogo Kawaguchi (Condensed Biophysics) in Dec. 2020.