HAL
open science

The genotype-phenotype relationship across different scales

Henry Kemble

To cite this version:

Henry Kemble. The genotype-phenotype relationship across different scales. Molecular biology. Université Sorbonne Paris Cité, 2018. English. NNT : 2018USPCC178 . tel-02438077

HAL Id: tel-02438077

https://theses.hal.science/tel-02438077v1

Submitted on 14 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole doctorale Frontières du Vivant (474)

Laboratoire Infection, Antimicrobials, Modelling, Evolution (IAME)

Equipe Quantitative Evolutionary Microbiology (QEM)

The genotype-phenotype relationship across different scales

Par Henry Kemble

Thèse de doctorat de la Biologie des Systèmes et l'Évolution

Dirigée par Olivier Tenaillon

Présentée et soutenue publiquement à Paris le 31 octobre

Président du jury : Courtier-Orgogozo, Virginie / Directrice de Recherche / Université Paris Diderot

Rapporteurs : Bank, Claudia / Directrice de Recherche / Gulbenkian Institute ; Bataillon, Thomas / Directeur de Recherche / Aarhus University

Examinateurs : Félix, Marie-Anne / Directrice de Recherche / l'Ecole Normale Supérieure ; Martin, Gullaume / Chargé de Recherche / Université Montpellier

Directeur de thèse : Tenaillon, Olivier / Directeur de Recherche / Université Paris Diderot

Co-directeur de thèse : Nghe, Philippe / Maître de Conférences / École Supérieure de Physique et de Chimie Industrielles

Except where otherwise noted, this work is licensed under
https://creativecommons.org/licenses/by-nc-nd/4.0/

Titre : La relation génotype-phénotype vue à différentes échelles

Résumé : Avec la révolution moléculaire en biologie, une compréhension des mécanismes de la relation génotype-phénotype est devenue possible. Récemment, les progrès réalisés dans la synthèse et le séquençage de l'ADN ont permis le développement d'expériences de *deep-mutational scanning* capable de quantifier divers phénotypes pour un ensemble de génotypes sur toute la longueur d'un gène. Ces ensembles de données sont non seulement intéressants en eux-mêmes, mais permettent également de tester de manière rigoureuse des modèles phénotypiques quantitatifs. Nous avons utilisé cette technologie pour caractériser les cartes séquence-fitness de 3 systèmes bactériens modèles: un régulateur global, la CRP, une enzyme de résistance aux antibiotiques, la β-lactamase, et une petite voie métabolique constituée des enzymes AraA et AraB. Ces systèmes ont été choisis pour éclairer les rôles de différentes caractéristiques dans la formation de la relation génotype-fitness (réseaux de régulations, stabilité des protéines et flux métabolique). Nous constatons que la tendance globale des effets sur le fitness semble prévaloir sur les tendances spécifiques. Ceci nous conduit à penser qu'une grande partie de la relation entre le génotype et le fitness pourrait être expliquée à partir de la forme des fonctions de phénotype-fitness. Par ailleurs, nous voyons que la caractérisation de la relation génotype-fitness dans différents systèmes peut être un moyen puissant d'obtenir des informations sur les phénotypes pertinents.

Mots clefs : évolution, génétique, biologie des systèmes, *deep-mutational scanning*, paysages adaptatifs, épistasie, expression génique, métabolisme, toxicité, régulateurs globaux

Title: The genotype-phenotype relationship across different scales

Abstract: With the molecular revolution in Biology, a mechanistic understanding of the genotype-phenotype relationship became possible. Recently, advances in DNA synthesis and sequencing have enabled the development of deep-mutational scanning experiments, capable of scoring comprehensive libraries of genotypes for a variety of phenotypes over the length of entire genes. Such datasets are not only interesting in themselves, but also allow rigorous testing of quantitative phenotypic models. We used this technology to characterise sequence-fitness maps for 3 model bacterial systems: a global regulator, CRP, an antibiotic-resistance enzyme, β-lactamase, and a small metabolic pathway, consisting of the enzymes AraA and AraB. These different systems were chosen to illuminate the roles of different mechanistic features in shaping the genotype-fitness relationship (regulatory wiring, protein stability and metabolic flux). We find that smooth patterns of fitness effects tend to prevail over idiosyncrasy, indicating that much of the genotype-fitness relationship could be understood from the global shape of smooth underlying phenotype-fitness functions. On the flip side, we see that characterising the genotype-fitness relationship in different systems can be a powerful way to glean phenotypic insights.

Keywords: evolutionary genetics, systems biology, deep-mutational scanning, fitness landscapes, epistasis, gene expression, metabolism, toxicity, global regulators

To Margot and Anne-Laure,

My favourite sets of phenotypes

Acknowledgements

First, I must unquestionably thank my wife, Lou. Your patience, your generosity, your sacrifice. Your ability to listen to monologues on bacteria for 4 uncertain years, without flinching. You have taught me the meaning of both chance and necessity.

And my family: mum and dad, for teaching me so early to pursue what I love, and the rest will follow; and grandma, for your unconditional support.

And my family-in-France: Josiane, Laoreato, Marie-Claire, Mamie and Papi, for making Paris my new home.

I thank the jury members for kindly agreeing to take this on, especially the reviewers, Claudia Bank and Thomas Bataillon, and for all their work which has been a guiding force in this thesis.

I thank the Thesis Advisory Committee, Luis-Miguel Chevin and Didier Mazel, for their patience, advice and encouragement.

I thank Olivier Tenaillon, the most generous supervisor I could have hoped for, who has taught me so much. I'll miss his enthusiasm, his creativity and his kindness. And Philippe Nghe, who has taught me all the rest, for always taking the time to lend his sharp eye. I'll also miss *his* enthusiasm, his creativity and his kindness. They made an excellent team.

I thank Erick Denamur, for welcoming me so warmly among the ranks, for all his generosity and his considered advice.

I thank all those teachers who have inspired me along the way: Wei-Jun Liang and the CRI team for showing me what Science really is, and for restoring the joy of learning. Pablo Ibañez Cruceyra, the most natural and generous teacher, for teaching me the tricks of the cloning trade that made all this work possible. And Maude Guillier and Marie-Eve Val for being so willing to impart their practical knowledge.

Lastly, I thank the past and present members of the lab for all the help, discussions and company along the way. Especially Alejandro Couce, for endless conversation and boundless enthusiasm, André Birgy, for allowing me to reap the fruits of his hard-won technical optimisation, Audrey Chapron, Catherine Eisenhauer and Mélanie Magnan for so much help with the experiments, Hervé Le Nagard for all the technical support and tinkering, and Antoine Bridier-Nahmias for diverse titbits of help and advice. And finally, I thank Matt Deyell and Andrzej Prokopat at the ESPCI for their help and their tireless technical development.

Table of contents

1 Introduction 8
1.1 Part I: Overview 9
1.1.1 Conceptual origins of the genotype/phenotype distinction 9
1.1.2 Zooming in: from the chromosome to the codon 12
1.1.3 The code is cracked! Long live the code! 14
1.1.4 Where are we now? Step-by-step from genotype to phenotype in the era of molecular biology 16
1.1.4.1 Expression 16
1.1.4.2 Protein structure, macromolecular assembly and molecular function 17
1.1.4.3 Functional molecular networks 21
1.1.4.4 Cell physiology 25
1.1.4.5 Development 27
1.1.4.6 Organismal anatomy and physiology 28
1.1.4.7 Summing up 29
1.2 Scope of the rest of this thesis 30
1.3 Part II : Experimental Insights 31
1.3.1 Properties emerging from the genotype-phenotype relationship 31
1.3.1.1 One gene 31
1.3.1.2 Two genes 58
1.3.1.3 The genome 62
1.3.1.4 The environment 71
1.4 Outline of the original research chapters included in this thesis 72
2 Single-Mutation Fitness Landscape of a Global Transcriptional Regulator across Environments 75
2.1 Introduction 76
2.1.1 Global transcriptional regulators 78
2.1.2 The cyclic AMP receptor protein (CRP) of *Escherichia coli* 79
2.1.3 Choice of experimental environments 81
2.2 Results 82
2.2.1 Optimisation of experimental conditions 82

2.2.2 Mutant library quality 85
2.2.3 Fitness estimation and experimental noise 88
2.2.4 Distribution of fitness effects (DFE) 92
2.2.5 Fitness correlations between environments 95
2.2.6 Sequence-fitness maps 98
2.3 Discussion 100
2.3.1 Technical considerations 100
2.3.2 Underlying phenotype-fitness landscape 102
2.3.3 Gain-of-function mutations 105
2.3.4 Sequence-fitness maps as a functional resource 106
2.4 Methods 107
2.4.1 Supplementary Tables 128
2.5 References 131
3 The thermodynamic roots of pairwise epistasis in the α-helix of β-lacatamase TEM-1: local *versus* global epistasis 138
3.1 Introduction 139
3.2 Results 141
3.3 Discussion 150
3.4 Methods 152
3.5 Supplementary figures 173
3.6 References 174
4 Flux, toxicity and protein expression costs shape genetic interaction in a metabolic pathway 176
4.1 Introduction 177
4.2 Results 180
4.3 Discussion 188
4.4 Methods 189
4.5 Supplementary figures 224
4.6 Supplementary tables 237
4.7 References and notes 242
5 Discussion 248
6 References 258

Table of figures

Figure 1.1 The concept of separation between genotype and phenotype

Figure 1.2 A remarkable (and the exclusive) figure from Johannsen's paper coining the term phenotype

Figure 1.3 The simplified "Central Dogma" of Molecular Biology by the end of the 1960s

Figure 1.4 The discovery of RNA splicing provided a direct visualisation of the fact that the correspondence between DNA base sequence and polypeptide amino acid sequence was not always as direct as first assumed

Figure 1.5 S. cerevisiae protein-protein interactome network representations constructed from three different data types

Figure 1.6 Conformational diversity of a ligand bound to different proteins

Figure 1.7 An early attempt at a comprehensive cell-scale kinetic model

Figure 1.8 Reconstruction of the E. coli metabolic network

Figure 1.9 Whole-cell model of a bacterium

Figure 1.10 The most common use of phage display

Figure 1.11 FACS-seq, a common particle-sorting method for high-throughput genotype-phenotype mapping

Figure 1.12 HiTS-FLIP, a specialised method to directly quantify DNA affinity landscapes

Figure 1.13 EMPIRIC, a general method for high-throughput genotype-fitness mapping

Figure 1.14 A sample of experimentally characterised DMEs in various proteins, all showing multi-modality

Figure 1.15 Illustration of the thermodynamic hypothesis for DMEs in proteins

Figure 1.16 A sample of experimentally characterised elasticity functions, all of a saturating concave form

Figure 1.17 Expression-fitness functions for a diverse set of protein-coding yeast genes

Figure 1.18 Types of pairwise epistasis possible for different types of mutation pairs

Figure 1.19 Trends of epistasis predicted by thermodynamic model of mutation effects

Figure 1.20 Two-dimensional activity-fitness functions predicted from Metabolic Control Analysis

Figure 1.21 bTRACE, a general high-throughput method for analysing the effect of known genome-wide mutation combinations

Figure 1.22 The canonical isotropic Fisher's Geometric Model of Adaptation (FGMA) in

two dimensions

Figure 1.23 A global network of gene-gene interaction profile similarities

Figure 2.1 Choice of experimental conditions

Figure 2.2 Sequencing coverage and quality of barcoded mutant library

Figure 2.3 Anomalous barcode detection

Figure 2.4 Barcode and mutant dynamics during competitive growth

Figure 2.5 Experimental noise characterised by independent barcode sets

Figure 2.6 Distribution of fitness effects (DFE) of single amino acid substitutions in CRP

Figure 2.7 Correlations of fitness effects between environments

Figure 2.8 Non-monotonous fitness effect correlation and optimum overshooting

Figure 2.9 Sequence-fitness maps for single amino acid substitutions across the entire length of CRP, in 4 environments

Figure 3.1 Single- and double- mutation fitness effects

Figure 3.2 Pairwise epistasis

Figure 3.3 Stability and context dependency

Figure 3.4 Deviations from the stability model

Figure 3.S1 Correlation between MIC of amoxicillin and fitness for single mutants

Figure 3.S2 Correlation between MIC of amoxicillin and fitness for double mutants

Figure 3.S3 Distribution of fitness effects of stop-codon mutations

Figure 4.1 Quantitative mapping of fitness interactions between expression variants of two metabolic genes in expression-modifying environments

Figure 4.2 Fitness effects of promoter mutations across backgrounds and environments

Figure 4.3 Strength, types and trends of epistasis across environments

Figure 4.4 Mechanistic basis of heterogeneous, environmentally dependent epistasis

Figure 4.S1 Construction and characterisation of barcoded promoter-mutant plasmid library

Figure 4.S2 Sequencing coverage and quality of barcoded mutant library

Figure 4.S3 Mutant dynamics during pooled competition assays under different inducer concentrations

Figure 4.S4 Measurement precision and reproducibility

Figure 4.S5 Fitness effects of single and double mutations across environments

Figure 4.S6 Epistasis across environments

Figure 4.S7 Correlations between individual fitness effects and epistasis

Figure 4.S8 Performance of flux-toxicity-expression burden model

Figure 4.S9 Goodness-of-fit comparison of different phenotype-fitness models

Figure 4.S10 Flux-fitness relationship predicted by model

Figure 4.S11 Fitness surface coloured by predicted epistasis category in Env2

List of Abbreviations

AraA: L-arabinose isomerase

AraB: L-ribulokinase

aTc: anyhdrotetracycline

ATP: adenosine triphosphate

BCM: Base Competition Medium

bTRACE: Barcoded Tracking of Combinatorial Engineered Libraries

cAMP: cyclic AMP

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

CRP: Cyclic AMP Receptor Protein

DFE: Distribution of Fitness Effects

DME: Distribution of Mutational Effects

DNA: deoxyribonucleic acid

EMPIRIC: Extremely Methodical and Parallel Investigation of Randomized Individual Codons

FACS: Fluorescence-Activated Cell Sorting

FBA: Flux Balance Analysis

FGMA: Fisher's Geometric Model of Adaptation

FRT: Flippase Recognition Target

GEO: Gene Expression Omnibus

HiTS-FLIP: High-Throughput Sequencing-Fluorescent Ligand Interaction Profiling

HPLC: High-Performance Liquid Chromatography

IPTG: isopropyl β-D-1-thiogalactopyranoside

LB: Lysogeny Broth; Luria-Bertani medium

LED: Light-Emitting Diode

MCA: Metabolic Control Analysis

MCMC: Markov Chain Monte Carlo

MH: Mueller-Hinton medium

MIC: Minimum Inhibitory Concentration

NaCl: sodium chloride

NAD: nicotinamide adenine dinucleotide

NGS: Next-Generation Sequencing

NNS: a triplet of nucleotides, with N = A,C,G or T, and S = G or C

OD: Optical Density

ORF: Open Reading Frame

PCR: Polymerase Chain Reaction

PPP: Pentose Phosphate Pathway

RNA: ribonucleic acid

UV: Ultra-Violet

WT: wildtype

1 Introduction

1.1 Part I: Overview

1.1.1 Conceptual origins of the genotype/phenotype distinction

The need to distinguish genotype from phenotype was first suggested (implicitly) by Gregor Mendel (Mendel, 1866). Based on his observations of the disappearance and reappearance of an apparently discrete trait in pea plants (white flowers) over two successive generations, Mendel proposed the existence of discrete internal "factors" (genotype) that are passed from one generation to the next. These factors interact through some logic within an individual plant to manifest its visible features (phenotype), but are not themselves altered by this process, being passed to the next generation re-assorted but individually untouched. Hence the fundamental distinction that the agent of heredity, genotype, causes phenotype, but phenotype does not affect genotype (Figure 1.1). Indeed, the "epistemic cut" separating genotype from phenotype may lie at the origin of biological evolution itself (Pattee, 2001).

A physical explanation for such a cut was first provided by Francis Galton (Galton, 1876), whose pool of hereditary elements ("stirp") was divided into "patent" ones that developed into different cell types (phenotype), and "latent" ones that were transmitted to the next generation (genotype). More famously, August Weismann (Weismann, 1892) proposed the distinction between the *Keimplasma* (germplasm) and

Somatoplasma (somatoplasm) tissues in multicellular organisms. According to Weismann, the source of gametes, the germplasm, was physically isolated from environmental influence, while the somatoplasm forming the body of an organism (phenotypes) developed throughout an individual's lifetime and was sensitive to the environment. It was left to Wilhelm Johannsen, however, to introduce the current terminology of *genotype* and *phenotype* ((Johannsen, 1911); Figure 1.2), in the process conceptualising the active field of developmental biology as the study of how genotypes lead to phenotypes (Figure 1.1).

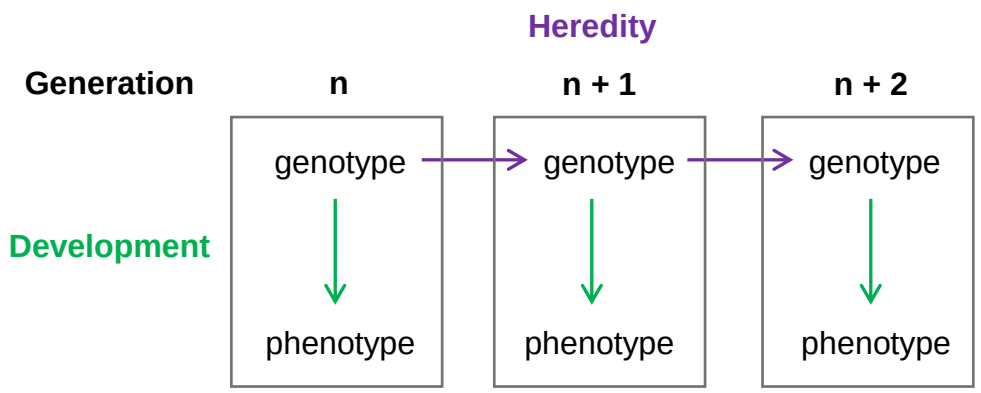
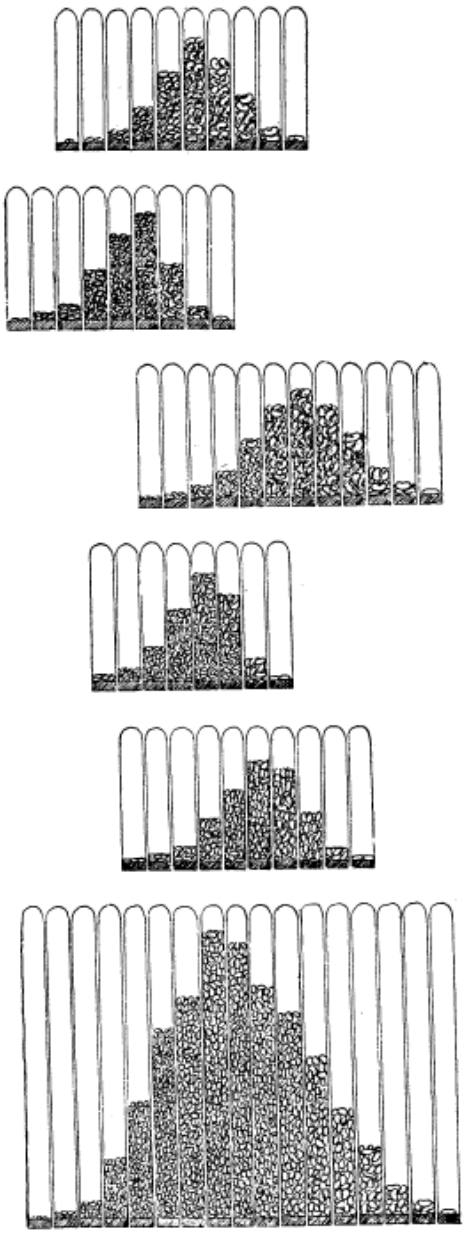


Figure 1.1: The concept of separation between genotype and phenotype. The genotype is passed directly from one generation to the next, and develops into the phenotype within each individual (boxes). The phenotype, however, cannot itself be transmitted through generations, nor can it have any reciprocal effect on the genotype.

136 THE AMERICAN NATURALIST [VOL. XLV



EXPLANATION OF DIAGRAMS

DIAGRAMS SHOWING FIVE DIFFERENT PURE LINES OF BEANS AND A "POPULATION" FORMED BY THEIR UNION. In each case the beans enclosed in glass-tubes are marshalled in equidistant classes of length; identical classes are superposed. The pure lines show transgressive fluctuation: it is mostly impossible to state by simple inspection of any individual bean the line to which it belongs.—The fluctuations about the average length (the phenotype) within the pure lines as well as in the mixed population show no characteristic difference.

Figure 1.2: A remarkable (and the exclusive) figure from Johannsen's paper coining the term *phenotype*. "All "types" of organisms, distinguishable by direct inspection or only by finer methods of measuring or description, may be characterized as "phenotypes"". [Excerpt and figure from (Johannsen, 1911)].

1.1.2 Zooming in: from the chromosome to the codon

The material basis of the genotype (carrier of Mendel's factors, Galton's latent stirp and Johannsen's genes) was unequivocally narrowed down to the chromosome by 1915 (Morgan et al., 1915), although some cytologists had been propounding the chromosome theory of inheritance since 1902 (Boveri, 1902; Sutton, 1902; Wilson, 1902; Stevens, 1905; Wilson, 1905; Carothers, 1913; Crow and Crow, 2002). While this crucial insight provided a mechanism for heredity (*ie.* transmission of the genotype through generations), it revealed nothing about the mechanism of the development of phenotype from genotype. Understanding this would require an even finer characterisation of the genetic material, starting with its chemical identity as DNA (Avery et al., 1944; Hershey and Chase, 1952), and quickly followed by its molecular structure as the double helix (Watson and Crick, 1953). Once again, the immediate beneficiary of the base-pairing double helix discovery was heredity – the molecular mechanism of template-based DNA replication for the transmission of genotype through generations, as proposed by Koltsov in 1927 with remarkable prescience (Soyfer, 2001), suggested itself directly from the structure (Watson and Crick, 1953), and was powerfully demonstrated in bacteria 5 years later (Meselson and Stahl, 1958). In the same year, Francis Crick finally made explicit a molecular basis for the genotype-phenotype relationship, based on the emerging picture from studies on protein synthesis and the genetic code (Crick, 1958). His "Central Dogma" (it turns

out that his understanding of the definition of "dogma" was mistaken (Judson, 1979)) held that information, in the form of monomer sequence, could be transferred from nucleic acid to protein (whose versatile functional roles in causing phenotypes were by then well-recognised [see especially (Beadle and Tatum, 1941; Pauling et al., 1949)]), but not from protein to protein or from protein to nucleic acid. Messenger RNA (mRNA) was then confirmed as the informational intermediate leading from DNA to protein (Brenner et al., 1961; Gros et al., 1961), and by the end of 1966 the complete genetic code had been cracked (Nirenberg et al., 1966). The molecular revolution of Biology was firmly established (Figure 1.3).

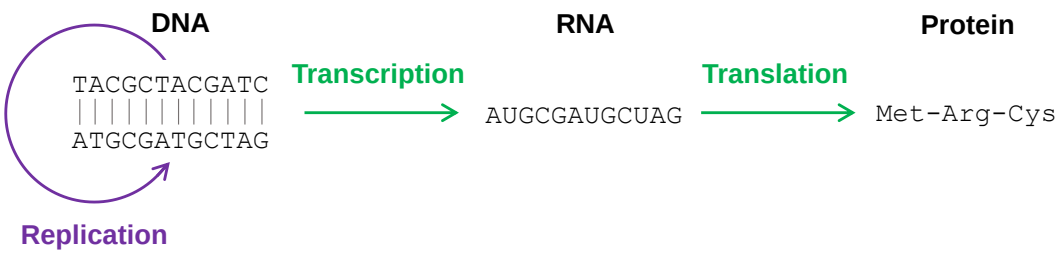


Figure 1.3: The simplified "Central Dogma" of Molecular Biology by the end of the 1960s. DNA, the genotype molecule, is replicated by a semi-conservative mechanism enabled by specific base-pairing between sister strands. The sequence of bases in the DNA "coding strand" is copied to messenger RNA (transcription), and messenger RNA (mRNA) serves as the direct template for protein synthesis (translation). The resulting sequence of amino acid residues is determined, essentially unambiguously, by the DNA base sequence, via the (redundant) genetic code. As proteins are well-known to be directly responsible for major cell functions, and are therefore agents of phenotype, this picture provides an obvious analogy with Figure 1.1, and so a molecular basis for the separation of genotype and phenotype.

1.1.3 The code is cracked! Long live the code!

The demonstration of a simple chemical code by which the genotype is translated to molecules of established biochemical function was surely beyond the wildest dreams of the pre-revolution pioneers hoping for an understanding of the genotype-phenotype relationship. Given a sequence of DNA bases, we could now predict the precise sequence of amino acid residues in the polypeptide chain it encoded. Although not quite: around a decade later it was found that, in eukaryotic cells, the base sequence of mRNAs could be processed ("spliced") before being translated into polypeptides, with a single initial transcript potentially giving rise to multiple differently rearranged "splice variants" (Berget et al., 1977; Chow et al., 1977; Goldberg et al., 1978; Nevins and Darnell, 1978), and so the code governing this transcript processing would also have to be cracked (Figure 1.4). And even then, the resulting theoretical polypeptide chains are still a far cry from the kind of organismal traits biologists are interested in. First, they are theoretical – they may or may not be synthesised at different times, places and amounts within an organism, and so we must understand the rules of their dynamical expression. Second, it is generally found that protein functionality does not result directly from amino acid sequence but from 3-dimensional structure (Wright and Dyson, 1999). This structure can depend on post-translational modifications and non-covalent interactions with other molecules, so we must understand the rules of protein folding, modification and intermolecular interaction, as well as the ultimate structure-

function relationship. Third, macroscopic phenotypes do not result from the action of a single protein, or even a single supramolecular assembly, but rather from a set of them, functionally interconnected through the regulated processes of metabolism, physiology and development (Schadt, 2009). We thus need a quantitative understanding of *their* rules, too.

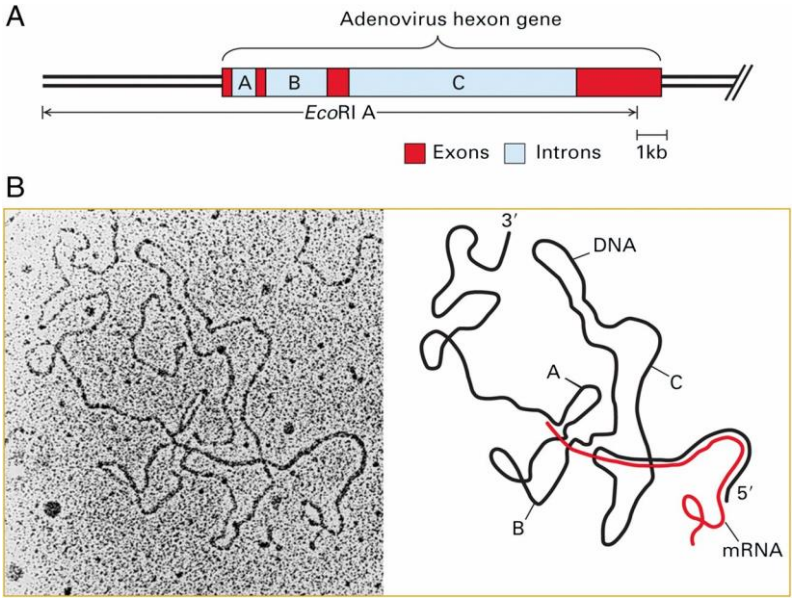


Figure 1.4: The discovery of RNA splicing provided a direct visualisation of the fact that the correspondence between DNA base sequence and polypeptide amino acid sequence was not always as direct as first assumed. The first experiments used electron microscopy to visualise DNA-RNA hybrid molecules prepared *in vitro*. The DNA is the transcribed strand of a eukaryotic viral genome fragment and the RNA is messenger RNA that has been transcribed from part of this fragment. The two strands can therefore hybridise *via* complementary base-pairing (which is the mechanism by which RNA is accurately transcribed from DNA). The complex structure of the hybrid (**B**) thus revealed that the mRNA is not simply a direct transcript of the encoding DNA, but that the mRNA must have been processed by the excision of several regions (*introns*) and the concatenation of the regions flanking these introns (*exons*) (**A**). To complicate things further, it was also found that a single DNA sequence could result in multiple different mRNA transcripts (*alternative splicing*). [Figure from (Berk, 2016)].

1.1.4 Where are we now? Step-by-step from genotype to phenotype in the era of molecular biology

1.1.4.1 Expression

The simplicity of the genetic code translating mRNA base sequence to amino acid sequence appears, however, to have been a "once-in-a-lifetime" gift to biologists (see (Jantzen and Danks, 2008) for a formal explanation of what may make translation such a special case). The base sequence determinants of gene *expression* are not nearly as clear-cut, although considerable progress has been made in unravelling them over the last decades. This progress began with the alignment of sequences from known gene regulatory regions of different species, leading to the discovery of reasonably conserved consensus sequences such as the Pribnow box for bacterial transcription (Pribnow, 1975) and the Shine-Dalgarno sequence for bacterial translation (Shine and Dalgarno, 1975). These sequences govern expression by binding the macromolecular assemblies directly responsible for transcription and translation, the rate of these processes being determined in part by the strength of such binding. The relationship between sequence and expression is thus a quantitative one, determined directly by physico-chemistry, in contrast to the genetic code which seems to have largely transcended this realm into that of symbolism (Jantzen and Danks, 2008).

Since these initial discoveries, molecular biologists have reported countless sequence-dependent mechanisms governing the dynamics of cellular RNA and protein levels, such as nucleosome organization (Kaplan et al., 2010; Hornung et al., 2012), epigenetic marks (Wachter et al., 2014), transcription factor binding (Slattery et al., 2014; Levo et al., 2015), messenger ribonucleoprotein particle formation (Gehring et al., 2017), ribonuclease processivity (Pertzev, 2006), mRNA 3' end processing (Shalem et al., 2015), mRNA folding (Kudla et al., 2009; Anderson-Lee et al., 2016), ribosome binding (Salis et al., 2009), translation elongation rate (Gingold and Pilpel, 2014) and protein modification (Basu and Plewczynski, 2010), not to mention those capable of altering the encoded *sequence via* mRNA splicing (Julien et al., 2016; Rosenberg et al., 2015) and editing (Tan et al., 2017). An impressive number of models are available to predict the effect of sequence on these processes, and the rapidly developing technologies of –omics and machine learning are enabling their continued refinement (Libbrecht and Noble, 2015).

1.1.4.2 Protein structure, macromolecular assembly and molecular function

Now that we have considered the journey from DNA sequence to the dynamical synthesis of a (modified) polypeptide chain, we can move to the next step: protein folding, a discipline all by itself. The prediction of 3-dimensional structure from amino acid sequence was until recently considered an insurmountable challenge (Dill et al.,

2007; R. Shenoy and Jayaram, 2010). But advances in statistical mechanics and homology modelling have now placed it firmly within our grasp, to the point where (simple) novel folds can be accurately designed at atomic resolution (Kuhlman et al., 2003). Prediction accuracy can still of course be much improved, however, especially in the case of large multi-domain proteins and membrane proteins (Dill et al., 2007; R. Shenoy and Jayaram, 2010), and in reality proteins exist not as static structures but as a dynamic ensemble of *conformers,* so protein *motion* must also be considered (Monzon et al., 2017). Proteins often function as multimers or supramolecular assemblies, and so the next step in the chain of causality from genotype to organismal phenotype, and one of the new frontiers of protein science, is physical protein-protein interaction. Computational approaches for the *qualitative* prediction of interactions from both amino acid sequence and 3-D structure have been developed that perform fairly well, and again these are likely to improve as more large-scale protein-interaction datasets are incorporated (Ventura, 2005; Lee et al., 2007; Tuncbag et al., 2008; Kamisetty et al., 2014; Celaj et al., 2017; Snider et al., 2015; Uetz et al., 2000; Ito et al., 2001; Butland et al., 2005; Stelzl et al., 2005; Krogan et al., 2006; Tarassov et al., 2008; Yu et al., 2008) (Figure 1.5). These interactions may be modulated, however, by post-translational modifications like ubiquitylation, and the “code” governing the resulting *signalling networks* is far from being understood (Lothrop et al., 2013; Yau and Rape, 2016).

Finally, the molecular function/activity arising from these folded proteins and protein assemblies must be deduced. Whereas 3-D structure prediction can successfully be performed from first principles, this is clearly untenable at present for function prediction, not least because function in this context is such a poorly defined concept (Smith et al., 2003). However, when chemical mechanisms are well-understood, it is in principle possible to determine quantitative "functional" parameters like association rate constants from structure, but the necessary molecular simulations are computationally expensive (and so not compatible with the –omics *zeitgeist*) and this field is in its infancy (Gabdoulline et al., 2003; Garcia-Viloca, 2004; Gabdoulline et al., 2007). More promisingly, progress has also been made in assigning qualitative molecular function based on homology to domains or proteins of experimentally deduced function (Lee et al., 2007; Sadowski and Jones, 2009; R. Shenoy and Jayaram, 2010). Ironically (given the textbook importance of structure in defining function and consequently the enormous efforts invested in solving structures), sequence homology has proven more fruitful in this respect than structural similarity, perhaps due to the extreme fluidity in the structure-function relationship (Stockwell and Thornton, 2006; Sadowski and Jones, 2009) (Figure 1.6), but maybe simply *because* there are so few solved structures for comparison (Lee et al., 2007).

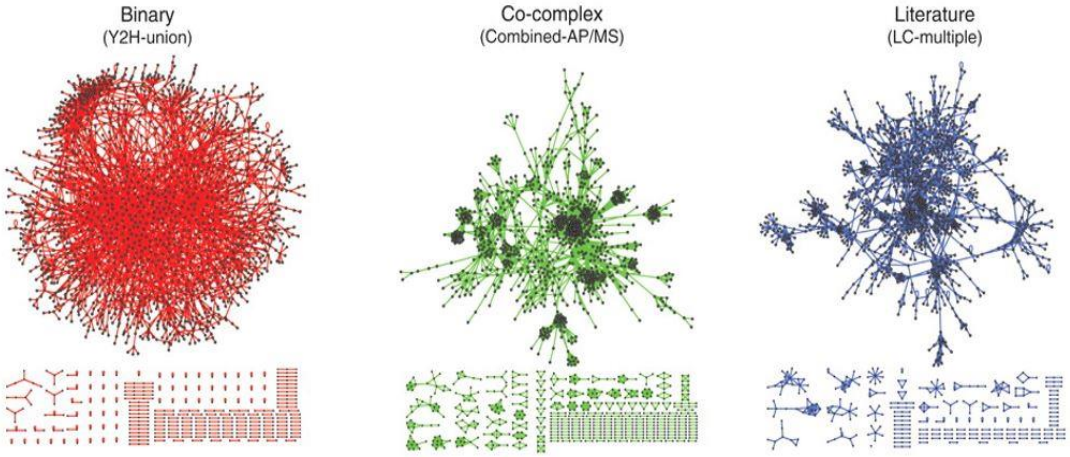


Figure 1.5: *S. cerevisiae* protein-protein interactome network representations constructed from three different data types. Nodes are proteins and edges represent physical interactions. Data types are (left-right) yeast-2-hybrid, coaffinity purification-mass spectrometry, and literature-curated interactions. [Figure from (Yu et al., 2008)].

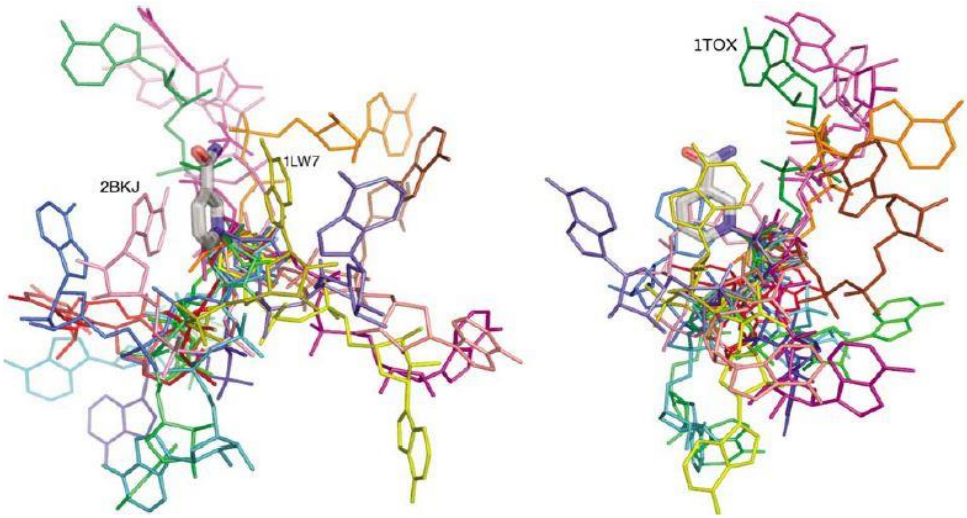


Figure 1.6: Conformational diversity of a ligand bound to different proteins. Nicotinamide adenine dinucleotide (NAD) is shown bound to 19 different representative proteins, with its different conformations (colours) superimposed by aligning them on the rigid nicotinamide ring (highlighted). Two viewing angles (left and right) are shown, and a few unusual conformations are labelled. Even a simple molecular function like ligand binding is seen to be structurally diverse. [Figure from (Stockwell and Thornton, 2006)].

1.1.4.3 Functional molecular networks

Now that we have an individual protein (assembly) performing a particular molecular function, we must question how many such functions combine through interconnected networks of signalling, gene regulation and metabolism, and ultimately physiology, to result in a measurable phenotype. It began to be realised in the 1930s, with the establishment of nonequilibrium thermodynamics, that the self-organization occurring in living systems (Schrodinger, 1944) must depend somehow on the *integration* of multiple molecular processes (Westerhoff and Palsson, 2004). By the 1960s, experiments then started to make clear that certain biological phenomena, such as feedback inhibition in metabolic pathways (Yates and Pardee, 1957), chemiosmotic ATP synthesis (Mitchell, 1961), oscillating metabolite levels (Chance et al., 1964) and the complex regulation of gene expression in response to the environment (Beckwith, 1967), could only be understood by considering biochemical reactions in their context of integrated systems, rather than by the prevailing reductionist approach of considering them individually, marking the dawn of Systems Biology (Westerhoff and Palsson, 2004). Early models of such processes, and later those at the scale of the entire cell (Joshi and Palsson, 1989; Novak and Tyson, 1995), were successfully built based directly on the kinetics of individual reactions (Figure 1.7). The laws of enzyme kinetics were by then well-understood, their origins tracing back to the beginning of the 20th century (Henri, 1902; Michaelis and Menten, 1913; Cornish-Bowden, 2013).

They are generally simple enough to have enabled the development of analytical frameworks for the study of multi-enzyme systems (including signalling pathways), such as Metabolic Control Analysis, capable of solving without the need for simulation the fluxes and steady-state concentrations given information on the individual reactions (Kacser and Burns, 1973; Heinrich and Rapoport, 1974; Westerhoff and Chen, 1984; Krauss and Brand, 2000). The great limitation of the kinetic approach to cell-scale modelling, however, is the requirement for accurate knowledge of kinetic parameters and rate constants, which is lacking for the vast majority of known biochemical reactions (see above) (Westerhoff and Palsson, 2004; Dandekar et al., 2014; Hartmann and Schreiber, 2015).

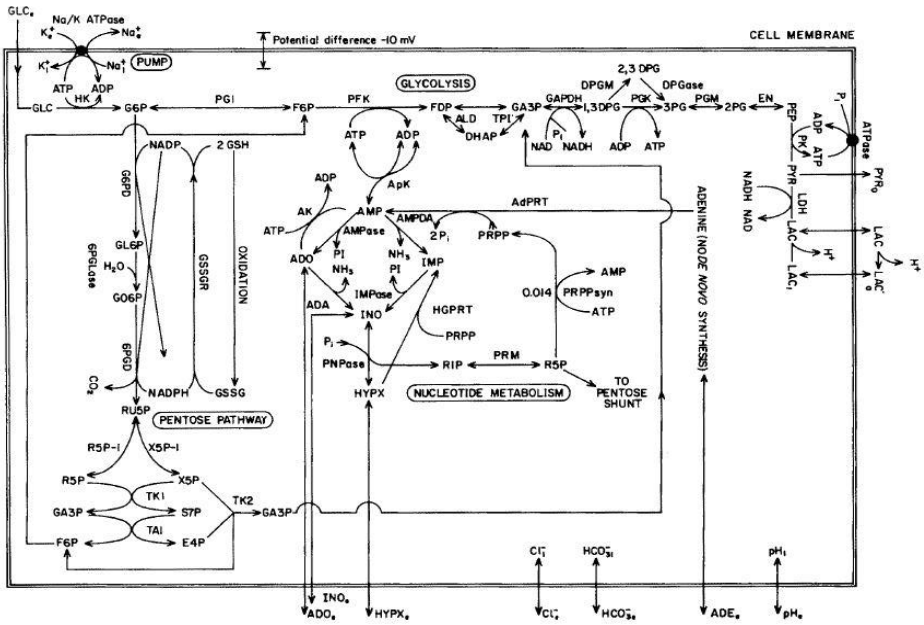


Figure 1.7: An early attempt at a comprehensive cell-scale kinetic model. The model describes human red blood cell metabolism using previously measured kinetic constants and the imposed physico-chemical constraints of osmotic balance and electroneutrality. [Figure from (Joshi and Palsson, 1989)].

To overcome this limitation, qualitative modelling approaches requiring less detailed data, in particular stoichiometric and Boolean approaches, have been developed (Ay and Arnosti, 2011; Chandrasekaran and Price, 2010; Hyduke and Palsson, 2010; O'Brien et al., 2015; Machado et al., 2016; Brunk et al., 2018). The primary data limitation for these approaches now becomes the network structure, the accuracy of which may be crucial, considering the network is an integrated system. The remarkable success of biochemists in unravelling metabolic pathways since the early 20th century means that this limitation is least problematic for modelling metabolism. Effectively complete metabolic network reconstructions now exist for several microbes and cell types from multicellular organisms, and new reconstructions can be built using increasingly available annotated genome data (O'Brien et al., 2015) (Figure 1.8). Constraint-based stoichiometric analysis of these reconstructions, such as Flux Balance Analysis (FBA), has proved especially promising, achieving some major successes beginning with the prediction of the optimal growth rate of *E. coli* on an unfavourable carbon source (Ibarra et al., 2002) and the growth effects of single-gene deletions in *S. cerevisiae* (Famili et al., 2003), and now including an impressive number of applications across medical and industrial biotechnology (Milne et al., 2009; O'Brien et al., 2015). Genome-scale regulatory and signalling networks have proven more challenging to reconstruct, in part due to their position as signal-flow networks rather than mass-flow networks like metabolism, making their wiring less constrained over evolutionary time (Hyduke and Palsson, 2010). Regulatory networks are the more

advanced of the two, with that of *E. coli* probably being the most complete. A recent *in silico* study revealed, however, that even for *E. coli*, an average of only ~27% of genes found to be differentially expressed across experimental environments could be directly explained by the network reconstruction (Fang et al., 2017). Improvements in chromatin immunoprecipitation (ChIP), transcriptomic and comparative genomics methods are all helping improve this situation, but progress is certainly slower than for metabolic networks (Novichkov et al., 2010; Fang et al., 2017). Large-scale signalling networks have proven least amenable to experimental elucidation, with their reconstruction relying primarily on manual curation (Hyduke and Palsson, 2010). Finally, several different approaches have been taken to model the integration of metabolic networks with regulatory and/or signalling networks, but this work is still in its infancy (Arias et al., 2015; Chandrasekaran and Price, 2010; Gonçalves et al., 2013; Ryll et al., 2014).

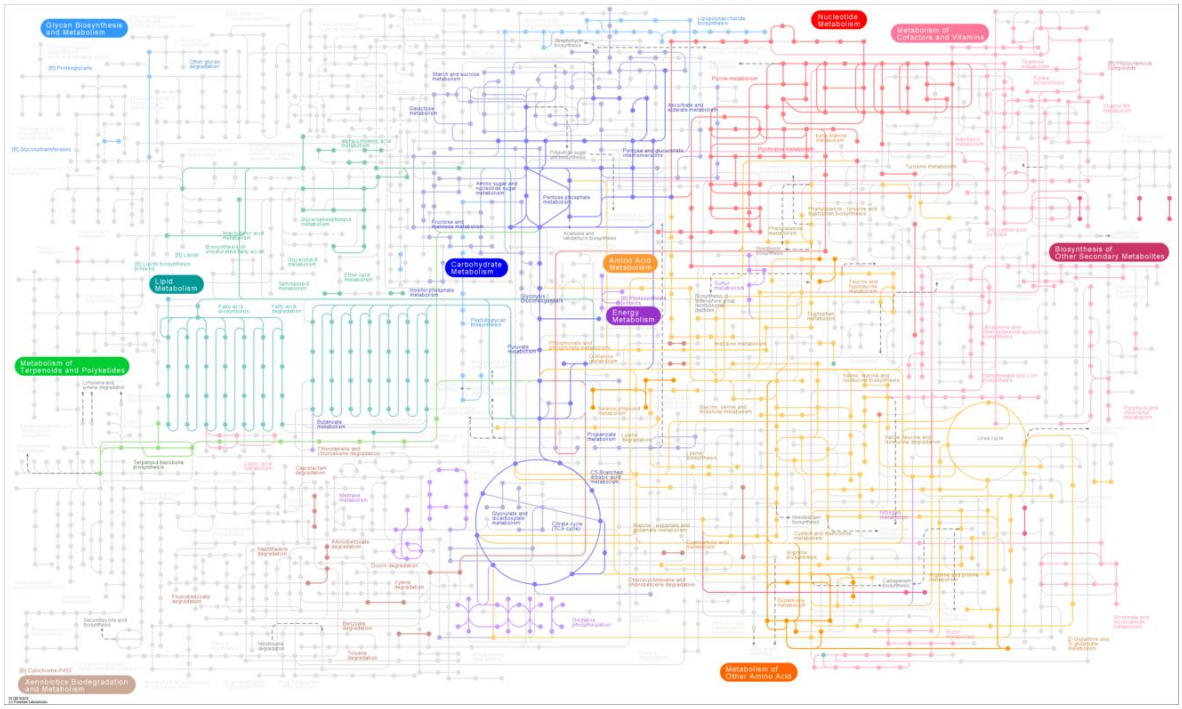


Figure 1.8: Reconstruction of the *E. coli* metabolic network. The *E. coli K-12 MG1655* metabolic network structure is shown by the coloured nodes and edges. Nodes are metabolites and edges represent reactions. The network was built from the KEGG PATHWAY database, which is compiled from existing literature (Kanehisa et al., 2014).

1.1.4.4 Cell physiology

Ultimately, these network models describe the transformation of a set of input molecules into a set of output molecules, *via* (regulated) metabolism. And so it would seem that we have still not even reached a cell-level phenotype. So how have such models managed to predict a trait like cellular growth rate? The answer is that they must include some "cell physiological" function mapping molecules to cell growth rate, like the oft-used biomass function of FBA. The biomass function of a particular cell type must account for at least the macromolecular content of a cell, and ideally also

the energetic requirements of cell maintenance and growth processes, from biosynthesis to error-checking in transcription, along with all necessary vitamins, cofactors and elements. This information is of course incomplete, but continues to grow with the help of the FBA framework itself, the refinement of biomass functions for certain unicellular model organisms being an active field of research (Feist and Palsson, 2010). A landmark in this area was the 2012 construction of a comprehensive whole-cell model of a simple bacterium, going the whole way from genome to individual molecules to detailed cellular physiology, based entirely on information from existing literature and databases (Karr et al., 2012) (Figure 1.9). The model was capable of predicting observed gene essentiality with 79% accuracy. Despite abstracting away many molecular kinetic details, such molecular-resolution models still of course require an ambitious amount of information, and so no further ones have been built to date. Rather, there has been an increasing trend towards highly coarse-grained models focussing on the relationship between macromolecular composition and microbial growth rate, which together have provided several novel insights into the laws of microbial growth (Bosdriesz et al., 2015; Giordano et al., 2016; de Jong et al., 2017; Scott et al., 2014; Weiße et al., 2015).

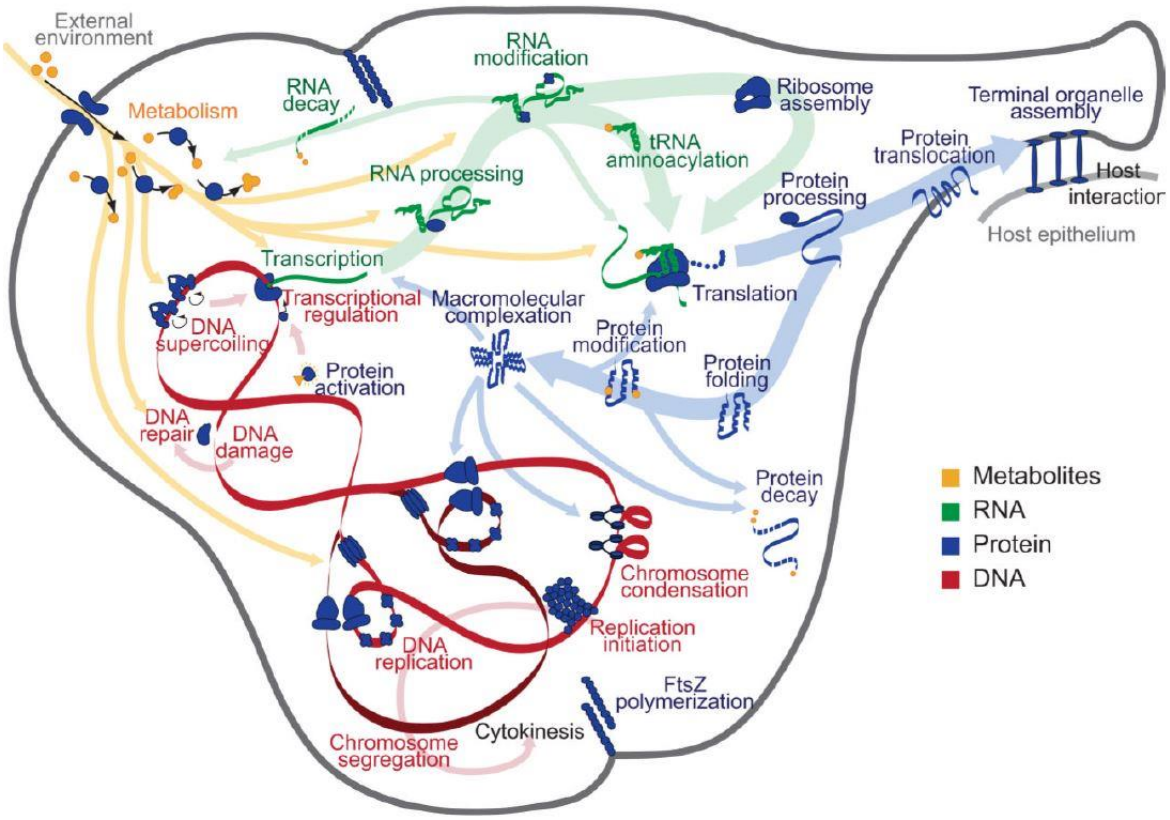


Figure 1.9: Whole-cell model of a bacterium. *Mycoplasma genitalium*, a simple bacterium containing 525 genes, was modelled *in silico* by dividing it into 28 functional processes involving DNA, RNA, proteins and metabolites, and building sub-models based on various formalisms for each from existing data (Kanehisa et al., 2014). The model predicts phenotypes from genotype by considering detailed molecular biology, metabolism and cellular physiology. [Figure from (Karr et al., 2012)].

1.1.4.5 Development

In the case of microbes, whole-cell models amount to whole-organism models, and so we have reached the end of the chain of causality between genotype and organismal phenotype. For multicellular organisms, however, integration of cell-level behaviour into the classical biological hierarchy of tissue-organ-organism through the process of

development must now be considered. The first quantitative mechanistic models of development can be traced back to the 1950s, with Alan Turing's diffusion-driven instability demonstration (Turing, 1952), which described how a spatially periodic pattern could arise from an initially homogenous system due to intrinsic noise. A molecular basis for Turing's abstract "morphogens" soon revealed itself with the discovery of diffusible growth factors, encouraging the construction of explicit tissue pattern formation models such as Lewis Wolpert's French Flag model (Wolpert, 1969). Since then, quantitative developmental biology has progressed somewhat irregularly, but many models now exist describing cell differentiation, tissue patterning and embryogenesis in terms of morphogens, cell growth and migration, cell-cell contacts and cell/tissue mechanics. Advances in imaging, single-cell and omics technologies are supporting more detailed studies of increasingly more model systems, but the field remains one that treats its experimental models on a case-by-case basis – the prospect of a set of general quantitative principles of development is at present difficult to envision (Davidson and Baum, 2012; Salazar-Ciudad and Jernvall, 2010; Vasieva et al., 2013; Yuan et al., 2017).

1.1.4.6 Organismal anatomy and physiology

Finally, we reach the top of the hierarchy, associated to the ancient disciplines of organismal anatomy and physiology (of multicellular organisms, of course). With the rise of molecular biology, funding for research in physiology, once at the centre of

biology, began to wane considerably. Efforts have therefore been made over the last 20 years to restore its standing, by establishing physiology within a quantitative and integrative framework capable of uniting the traditionally disparate sub-disciplines concerned with the various organ systems (Hunter, 2016). Some of these sub-disciplines, notably neurology (Hodgkin and Huxley, 1952), cardiology (Noble, 1960; Vik, 2011) and osteology/myology (Alexander, 2003), have a more solid history of mathematical modelling, but the others are now catching up, its importance being increasingly widely recognised (Gavaghan et al., 2006). Although there are enormous gaps in our quantitative understanding of biology at the organismal scale, the establishment of ambitious research programmes such as The Physiome Project (Hunter, 2016) and The OpenWorm Project (Gleeson et al., 2015), which aim to model whole organisms at the physiological level, inspires hope. But while the incorporation of behavioural rules into these models may be feasible for *C. elegans* in the not-too-distant future, the same can hardly be said for humans (Faisal and Stephens, 2009).

1.1.4.7 Summing up

To sum-up the state of our mechanistic understanding of the genotype-phenotype relationship, many promising inroads have been made, and models of different scales are being fruitfully combined. But there exist errors and patent large gaps in our knowledge at every step of the way, explaining why we cannot yet design *a de novo* genome containing the instructions for an organism with a set of pre-defined

phenotypes (*de novo* genome synthesis itself being no longer a limiting factor (Cello et al., 2002; Gibson et al., 2010; Annaluru et al., 2014)). And it should be noted here, first, that the genotype-phenotype relationship may be modulated by the environment at every step of the way. And second, that even if we had a complete specification of an organism's genotype and environmental history, we could not in general predict phenotype with certainty, due to internal stochastic effects caused by the low copy number of important biomolecules in each cell/sub-cellular compartment. Stochastic effects, resulting in phenotypic heterogeneity in a population of clonal cells in a homogenous environment, have been shown to contribute to key biological processes, such as antibiotic persistence in *E. coli* and cell differentiation in higher organisms (Balaban, 2004; Kærn et al., 2005; Bressloff, 2017). Such processes can therefore only be understood within a probabilistic framework.

1.2 Scope of the rest of this thesis

Given the many uncertainties and knowledge gaps sketched out above, the rest of this thesis will not aim to predict phenomes from genomes, but rather focus on some basic *properties* (distribution of mutational effects and epistasis) that emerge from this mechanistic view of the genotype-phenotype (including fitness) relationship in systems of various sizes and spans (with the highest phenotypic level being limited to unicellular organismal fitness). As well as their fundamental importance to biology as a whole, these properties are of great interest to the applied fields of biotechnology and

medicine (Lehner, 2013). The global structure of Part II (One gene, A few genes, Many genes) was inspired by the programme of the Evolutionary Systems Biology (ESB) 2018 conference, Cambridge, UK.

1.3 Part II : Experimental Insights

1.3.1 Properties emerging from the genotype-phenotype relationship

1.3.1.1 One gene

1.3.1.1.1 Experimental history of comprehensive genotype-phenotype mapping of single genes

Empirical snapshots of the local, or occasionally global, genotype-phenotype map for short sequences have recently become possible through creative combinations of high-throughput mutagenesis and phenotyping technologies. The technical challenge may be broken down as follows: the number of possible nucleic acid or protein sequence variants grows exponentially with sequence length; to draw statistical conclusions about the genotype-phenotype relationship for even very short sequences (*eg.* > 5-mer), a large number (> 10^3) of sequence variants must be characterised; currently, the only economical way of generating large numbers of sequence variants is in bulk; to characterise a phenotype of a large number of pooled sequence variants, not only do we need a high-throughput phenotyping technology, but also a high-throughput way to trace measured phenotype of individual variants back to their sequence.

The earliest strategy used to achieve this last challenge was *phage display*, originally developed for engineering purposes (Smith, 1985; Scott and Smith, 1990). Phage display is an *in vitro* system which allows biochemical phenotypes of peptides, such as binding affinity, to be easily traced back to DNA sequence. A peptide of interest is fused to a viral coat protein, making it accessible for biochemical analysis while remaining physically associated to the DNA from which it is expressed (Figure 1.10). Soon after the invention of phage display, alternative molecular display systems were developed (bacterial, yeast, ribosomal, mRNA), also for engineering goals (Levin and Weiss, 2006), but it was not until about 20 years after that a display strategy was used to attempt the first comprehensive local genotype-phenotype characterisation (Pál et al., 2006).

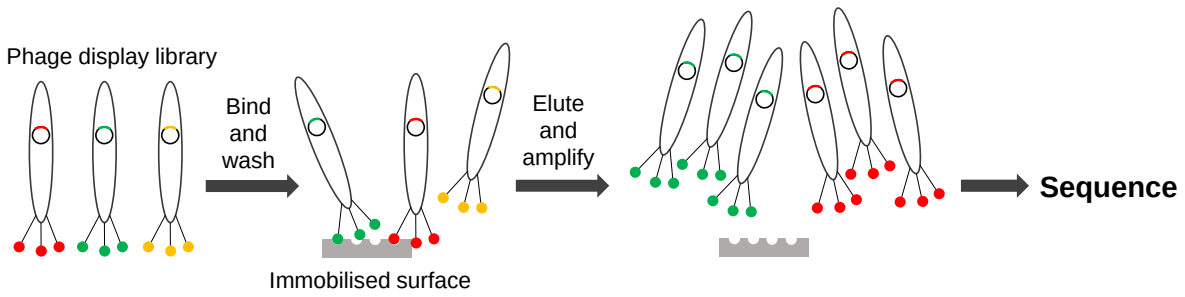


Figure 1.10: The most common use of phage display. A phage display library is constructed by high-throughput bulk mutagenesis, containing up to trillions of variants of a polypeptide of interest, which are expressed on the outside of phage particles (solid coloured circles). A biochemical selection is then applied to the polypeptide variants, most typically based on binding affinity. Selected phage particles can be amplified in bacteria and sequenced, linking genotype to biochemical phenotype. Multiple selection and analysis cycles may be performed if desired.

In this study, single point mutations were systematically introduced into the 35-aa receptor-binding site of human growth hormone using Kunkel mutagenesis (Kunkel et al., 1991), a phage display library was constructed, the library was screened for binding to either a structure-specific antibody or the human growth hormone receptor (to assess structural stability and receptor binding affinity, respectively), and a sample of screened clones was Sanger-sequenced to achieve quasi-quantitative measures for these two phenotypes. This first comprehensive *deep-mutational scanning* study produced several novel insights: the native protein fold was highly tolerant to mutations in the targeted solvent-exposed positions (with the established exceptions of cysteine and proline), with all positions showing a similar level of robustness; hydrophobic residues were generally more stabilising than hydrophilic ones, a very surprising result for a solvent-exposed region; at the majority of positions, mutations existed that resulted in both greater stability and stronger receptor binding than the wildtype; binding affinity was less robust to mutation than overall stability, with robustness in this case being position-specific; experimentally determined binding affinity robustness did not relate strongly to sequence conservation across species; physicochemical similarity of residue side-chains did not correlate strongly with phenotypic effects (Pál et al., 2006).

Although hugely informative, only a few studies employing this methodological strategy have been performed, as it still involves one low-throughput step – Sanger

sequencing – and so is rather labour-intensive. Microarray binding assays provide a higher-throughput approach for the specific case of assessing binding of a library of short nucleic acid sequences to a protein (Badis et al., 2009), but ultimately it was the arrival of massively parallel sequencing technologies that allowed large-scale genotype-phenotype mapping studies to flourish. Similarly to the aforementioned Sanger-sequencing study, the earliest such study used phage-display of human WW domain variants coupled with (weak) selection for binding to its peptide ligand, with Illumina sequencing of pre- and post-selection libraries to again give a quasi-quantitative measure of binding affinity (Fowler et al., 2010). The conclusions differed substantially, however: this time, 97% of the library variants bound the ligand less tightly than did the wildtype; mutational intolerance of the different positions correlated strongly with evolutionary conservation; the core ligand binding region was generally intolerant to mutation; each position appeared to possess a unique mutational preference spectrum; and global thermodynamic instability was concluded to be the primary determinant of binding affinity for the majority of variants. Some of the inconsistencies between these two studies may have a technical basis, but they also provide a first hint that the genotype-phenotype relationship, even for a given biochemical phenotype (here, ligand binding), may vary substantially between protein domains and as an extension whole proteins.

With the advent of high-throughput sequencing, one of the new limitations now becomes the types of phenotypes which allow easy coupling of phenotype measurement with sequencing. Creative approaches have been developed based on molecular display to assess a mutant protein library for binding to other proteins, DNA, RNA and small molecule ligands, as well as amenable enzymatic activities such as ubiquitination (Fowler and Fields, 2014; Starita et al., 2013). Further, particle sorting techniques like cell and microfluidic-droplet sorting can be used to place variants into phenotypic (*eg.* fluorescence) bins which are then subjected to deep-sequencing (Fowler and Fields, 2014; Kinney et al., 2010; McLaughlin Jr et al., 2012; Noderer et al., 2014; Sarkisyan et al., 2016; Whitehead et al., 2012) (Figure 1.11). These strategies are all only quasi-quantitative, however, as molecular display experiments use mutant frequency change over selection cycles as a proxy for protein stability, binding or activity, and sorting techniques can only place mutants into a limited number of discrete phenotypic bins.

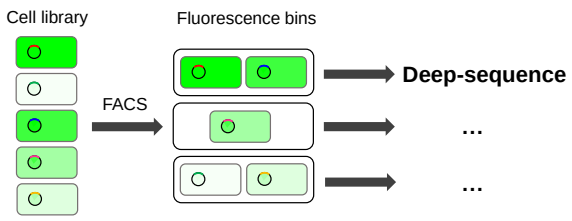


Figure 1.11: FACS-seq, a common particle-sorting method for high-throughput genotype-phenotype mapping. A cell library is constructed containing variants of a sequence of interest linked somehow to cell fluorescence (*eg.* a regulatory sequence controlling expression of a fluorescent reporter). Fluorescence-activated cell sorting (FACS) is then used to physically sort cells into bins based on chosen fluorescence ranges, and each bin is deep-sequenced, linking genotype to a simple cell-level phenotype.

To overcome this limitation, some *in vitro*, fluorescence-coupled microarray- and flow-cell-based techniques have been developed to *directly* measure the biochemical (binding affinity) phenotypes of a large number of mutants in parallel (Boyle et al., 2017; Buenrostro et al., 2014; Geertz et al., 2012; Maerkl and Quake, 2009; Nutiu et al., 2011; Shultzaberger et al., 2012) (Figure 1.12). Such methods tend to require highly specialised equipment/expertise, however, and so their output is limited to a small number of laboratories. In the specific case that a gene product functions by processing nucleic acids, direct catalytic measurements can also be made at high-throughput (Guenther et al., 2013).

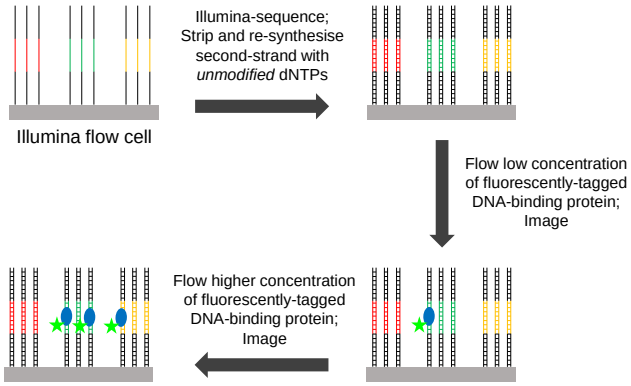


Figure 1.12: HiTS-FLIP, a specialised method to directly quantify DNA affinity landscapes. ~100 million unique DNA sequences are clustered and sequenced on an Illumina flowcell, and the chemically modified second-strand DNA is stripped away and rebuilt with unmodified dNTPs. Increasing concentrations of a DNA-binding protein of interest, tagged with a fluorescent reporter, are introduced, and binding to each cluster is visualised by fluorescence imaging. Each immobilised clonal DNA cluster is now associated to a sequence and a series of fluorescence measurements for different binding protein concentrations, allowing *in vitro* dissociation constants (K_d) to be precisely estimated for all sequences, thus linking genotype to a basic biochemical parameter at very high throughput. [Figure based on (Nutiu et al., 2011)].

One particular phenotype has proven generally amenable to easy and affordable experimental genotype-phenotype coupling, however. It allows precise, fully quantitative estimates, and is (fortuitously?) of great general interest to biologists, being a highly integrated trait of direct evolutionary significance: competitive cellular fitness (Hietpas et al., 2011) (Figure 1.13). In the original EMPIRIC (*Extremely Methodical and Parallel Investigation of Randomized Individual Codons*) experiment, a 9-a.a. region of *S. cerevisiae* Hsp90 (an essential and highly conserved eukaryotic chaperone protein), chosen to include positions covering a range of residue microenvironments and conservation levels, was comprehensively mutagenized to create a library of all possible single point-mutants. The fitness impact of all mutations was then directly measured in bulk, by growing the library in conditions where expression of the native Hsp90 copy was repressed (transferring to fresh medium at fixed time intervals to maintain exponential growth), and using deep-sequencing to sample mutant abundances at several time-points over a few days. Because the wildtype sequence was included in the library and the wildtype generation time was known, the competitive fitness of each mutant could be estimated relative to the wildtype as a selection coefficient, by taking the change in the ratio of mutant to wildtype reads as a function of wildtype generation time.

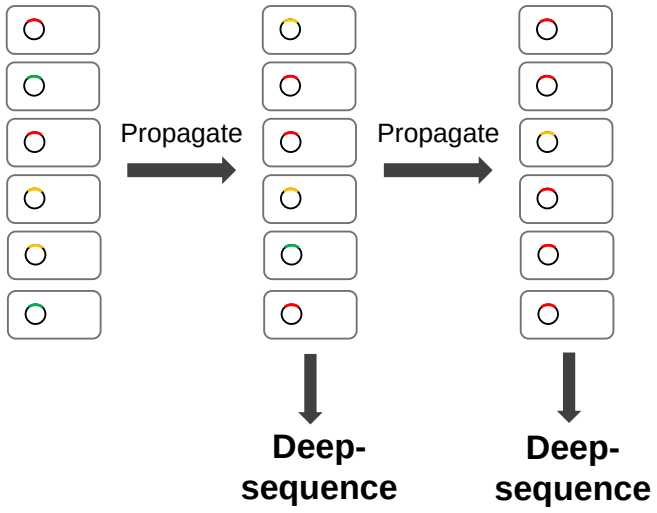


Figure 1.13: EMPIRIC, a general method for high-throughput genotype-fitness mapping. A cell library is constructed containing variants of a sequence of interest. The library is propagated under the chosen conditions, and deep-sequencing is applied at several time-points to track the change in variant frequencies over time, linking genotype to a highly integrated quantitative phenotype, cell fitness.

With this dataset, a truly quantitative and comprehensive analysis of individual mutation effects (on fitness) became possible for a small protein region, providing an important novel empirical insight: the distribution of fitness effects was bimodal, with a fairly equal proportion of mutations being either strongly deleterious or nearly neutral, consistent with a *nearly neutral model* of molecular evolution (Ohta, 1973). Two expected results were also confirmed: synonymous substitutions had far smaller effects on fitness than did non-synonymous ones (but see also (Agashe et al., 2016; Cuevas et al., 2012; Zwart et al., 2018)), and hydrophobic residues were more interchangeable with each other than were polar ones.

The interest in characterising genotype-fitness maps for other systems was immediately recognised, and experiments have now been performed on genes covering a diverse range of functions including ubiquitin, poly(A)-binding protein, antibiotic resistance enzymes, a small nucleolar RNA, tRNAs, metabolic enzymes and regulatory regions (Bernet and Elena, 2015; Chan et al., 2017; Dandage et al., 2018; Domingo et al., 2018; Firnberg et al., 2014; Jacquier et al., 2013; Klesmith et al., 2015; Li and Zhang, 2018; Li et al., 2016; Melamed et al., 2013; Melnikov et al., 2014; Puchta et al., 2016; Roscoe et al., 2013; Wrenbeck et al., 2017). Experiments vary in how precisely fitness is measured, and in how artificial the sequence-fitness relationship is: indeed, due to the ease of massively parallel genotype-fitness mapping with this approach, many artificial systems have been devised in which fitness is an indirect readout of some other phenotype, such as gene expression (Shultzaberger et al., 2010, 2012), protein-protein binding affinity (Diss and Lehner, 2018) or protein stability (Kim et al., 2013). Caution should therefore be taken in comparing studies, as non-linearities between different phenotypic levels likely have a major influence on the observed properties of the genotype-phenotype relationship, as will be discussed below.

This last decade has thus provided a rich pool of experimental data with which to explore statistically the genotype-phenotype relationship across different scales, and the following will summarise some principal conclusions.

1.3.1.1.2 Distribution of mutational effects (DME) in single genes

The distribution of effects of individual new mutations is of profound importance to medical and evolutionary genetics, and continues to be a hotly debated topic (Eyre-Walker and Keightley, 2007). The deep mutational scanning studies outlined above have revealed, perhaps unsurprisingly, that it differs between coding and non-coding regions, different types of gene and even different regions within genes.

1.3.1.1.2.1 The DME in single proteins

In proteins, whether the focal phenotype is biochemical functionality (Lagator et al., 2017a; McLaughlin Jr et al., 2012; Sarkisyan et al., 2016) or a highly integrated trait like fitness (Bank et al., 2014, 2015; Chan et al., 2017; Diss and Lehner, 2018; Firnberg et al., 2014; Hietpas et al., 2011; Jacquier et al., 2013; Jiang et al., 2013, 2016; Klesmith et al., 2015; Mavor et al., 2016; Melamed et al., 2013; Melnikov et al., 2014; Roscoe et al., 2013; Wrenbeck et al., 2017), the DME appears to be universally multi-modal, typically with near-neutral and highly deleterious modes and a vanishing fraction of beneficial effects (an almost trivial exception to this is the DME on expression or related phenotypes in repressor proteins (Lagator et al., 2017a); or, more generally, when decrease-of-function of one phenotype leads to an increase of a downstream phenotype, the DME is expected to be flipped) (Figure 1.14). The same multi-modality is found for whole viral genomes, which are dense with protein-coding sequences (Carrasco et al., 2007; Domingo-Calap et al., 2009; Peris et al., 2010; Sanjuan et al., 2004a). A popular and conceptually simple mechanistic hypothesis for

this is as follows: a globally determined property of proteins is stability; stability is therefore potentially affected by amino acid changes at many positions, while positions contributing to a limiting step in direct protein function are likely to be rare; protein folding is cooperative, resulting in a “thermodynamic cliff” in the stability-folding function; protein mutants therefore tend to lie either at the plateau at the top of this cliff (where the majority of the molecular ensemble is “correctly” folded), which is likely where the wildtype resides, or at the bottom of it (majority of ensemble is unfolded); the number of natively-folded molecules is likely a key determinant of protein activity, and the phenotype being measured depends either directly or indirectly on this activity (Wylie and Shakhnovich, 2011) (Figure 1.15). This hypothesis is partially supported by some empirical studies (Bank et al., 2015; Firnberg et al., 2014; Jacquier et al., 2013; Olson et al., 2014; Sarkisyan et al., 2016).

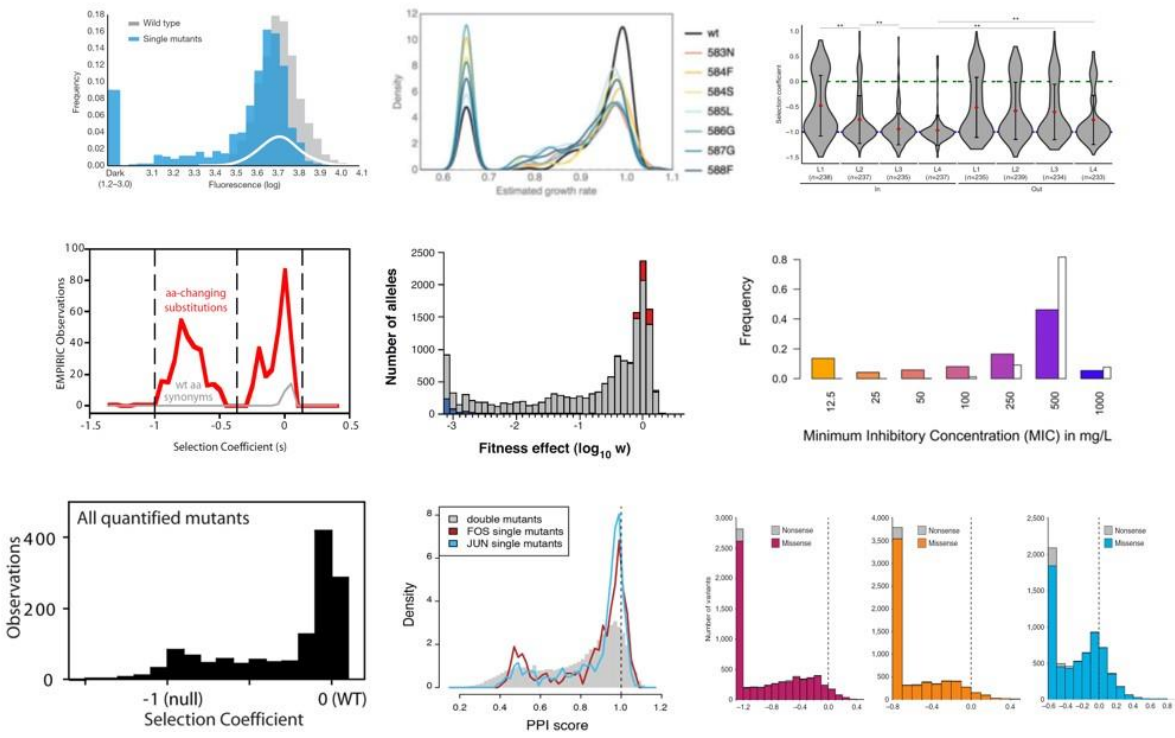


Figure 1.14: A sample of experimentally characterised DMEs in various proteins, all showing multi-modality. DMEs are shown for both full-length proteins and small protein regions. Mutated proteins include a fluorescent protein, a chaperone, metabolic enzymes, ubiquitin, transcription factor subunits and an antibiotic resistance enzyme. Measured phenotypes include fluorescence, minimum inhibitory antibiotic concentration (MIC) and cell fitness. [Figures from (Bank et al., 2015; Chan et al., 2017; Diss and Lehner, 2018; Firnberg et al., 2014; Hietpas et al., 2011; Jacquier et al., 2013; Roscoe et al., 2013; Sarkisyan et al., 2016; Wrenbeck et al., 2017)].

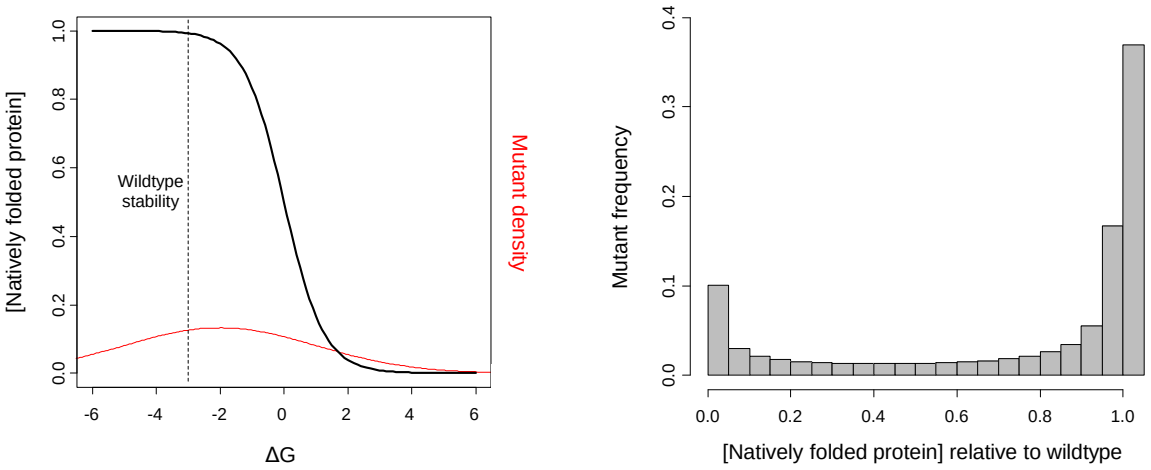


Figure 1.15: Illustration of the thermodynamic hypothesis for DMEs in proteins. Left panel - Black sigmoid curve shows the fraction of natively folded protein molecules as a function of the free energy of folding, ΔG, following: $P^{nat} = \frac{1}{1+e^{\Delta G/k_b T}}$, where k_b is the Boltzman constant and T is temperature ($k_b T$ is set here to 0.62, as in (Wylie and Shakhnovich, 2011)). Dashed line marks a hypothetical wildtype protein stability (-3 kcal/mole), located on the plateau of the sigmoid, for illustration. Red curve shows a hypothetical distribution of mutant ΔG values, resulting from a DME on ΔG that is Gaussian with a mean of +1, following (Wylie and Shakhnovich, 2011), but here with a larger standard deviation of 3. Right panel: The resulting DME on the *relative fraction of natively folded protein molecules*, which is bimodal under these parameter values. The standard deviation of the DME on ΔG had to be increased relative to (Wylie and Shakhnovich, 2011) to observe this bimodality, as the original model includes an extra (semi-step-) function in which protein functionality is set to 0 when $\Delta G > 0$ (rationalised by an aggravating toxic effect of misfolded protein molecules). Such steepening of the sigmoid causes bimodality to appear with a smaller range of ΔG effects.

Whatever the mechanism(s) responsible for changes in protein activity, however, the precise form of the distribution of mutational effects must depend on the quantitative relationship between activity and the measured phenotype, and on the activity of the wildtype. For example, three orthologous wildtype indole-3-glycerol phosphate

synthases were found to have their neutral mode shifted towards beneficial effects, showing they were sub-optimal for fitness under the chosen experimental conditions (Chan et al., 2017) (top right panel of Figure 1.14). An elegant demonstration of the impact of a non-linear activity-fitness relationship was provided by the EMPIRIC creators (Jiang et al., 2013). They characterized this relationship (a fitness *elasticity function* in the language of Metabolic Control Analysis (Kacser et al., 1995)) for their system, the Hsp90 chaperone protein, and found a strongly concave, saturating function similar to a binding curve. Interestingly, similar-shaped functions are found to generally describe activity-*flux* or activity-fitness relationships across enzymes and organisms, in line with expectations from Metabolic Control Analysis (Bershtein et al., 2013; Dykhuizen et al., 1987; Jiang et al., 2013; Kacser and Burns, 1981; Lunzer, 2005) (Figure 1.16). As is often found, wildtype Hsp90 activity lay safely on this plateau, far from the fitness shoulder, at endogenous expression levels. This meant that even mutations with an intermediate effect on *activity* could be nearly-neutral to fitness. By then measuring the distribution of fitness effects at increasingly reduced expression levels, such latent activity effects were shown to have greater and greater impacts on fitness. Indeed, the large number of mutations inferred to have intermediate, rather than nearly-neutral, effects on activity suggested that the dominant mechanism for activity changes in this particular mutated region was via direct molecular function (substrate binding), rather than global stability.

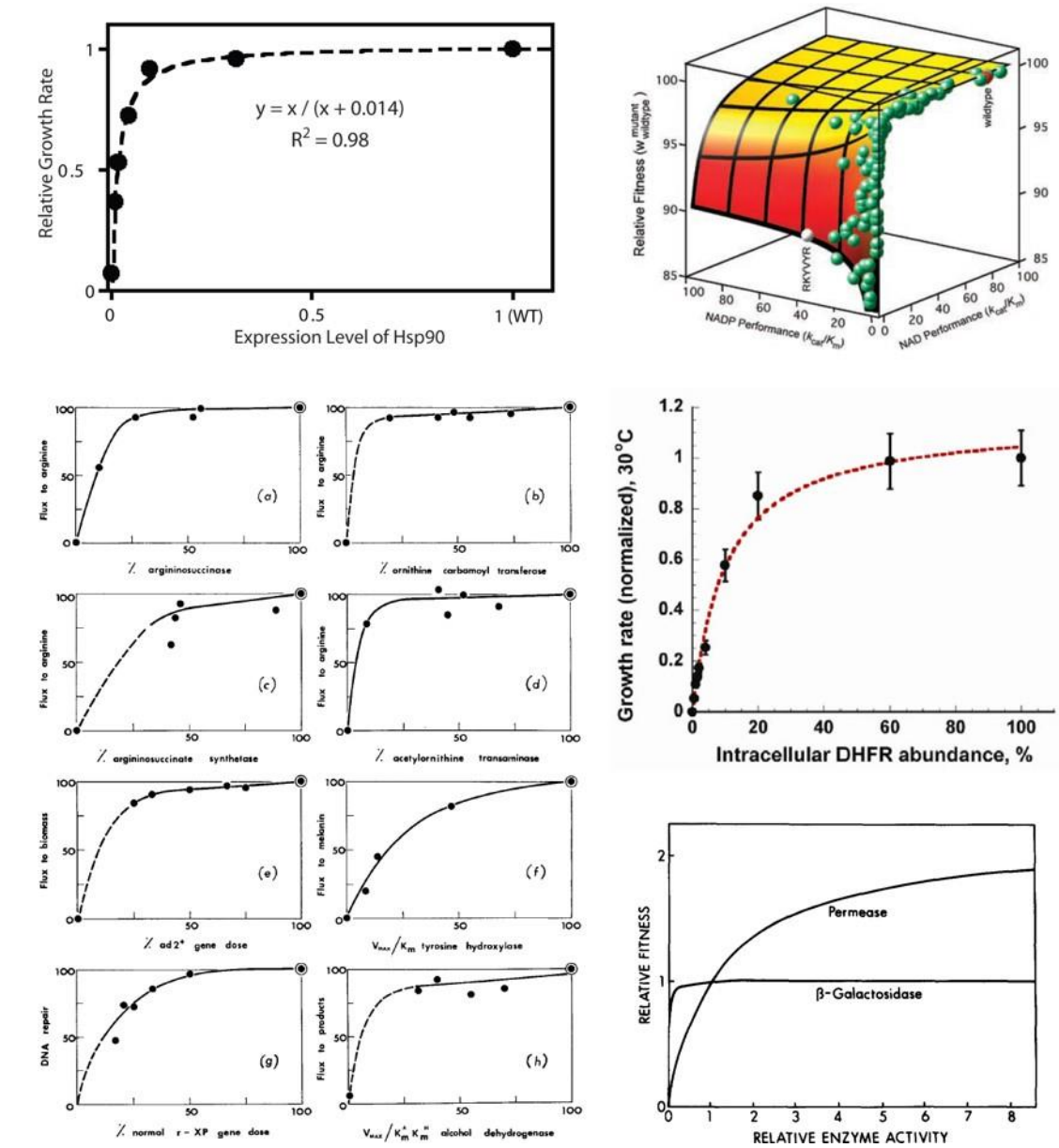


Figure 1.16: A sample of experimentally characterised elasticity functions, all of a saturating concave form. Proteins include a chaperone, a sugar transporter, a sugar hydrolase, an acid reductase, an acid dehydrogenase, four enzymes from an arginine biosynthesis pathway, alcohol dehydrogenase, an amino acid hydroxylase and an aminoimidazole carboxylase. Organisms include *Saccharomyces cerevisiae*, *Escherichia coli*, *Neurospora crassa*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*. Measured phenotypes include growth rate, metabolic flux and DNA repair rate. [Figures from (Bershtein et al., 2013; Dykhuizen et al., 1987; Jiang et al., 2013; Kacser and Burns, 1981; Lunzer, 2005)].

The DME in proteins, as will be found in other sequences, is thus shaped by forces at several scales, from the intramolecular level to the intermolecular interface level to the cell-system level and beyond. A source of bias to be wary of is therefore the choice of experimental system: it is often desirable to focus on systems suspected to show linear, or at least monotonic, relationships across these levels (*eg.* the activity-fitness function), but more complex relationships may well be common (Chou et al., 2014; Dekel and Alon, 2005; Drummond and Wilke, 2008; Perfeito et al., 2011; Rokyta et al., 2011; Serohijos and Shakhnovich, 2014; Serohijos et al., 2012; Shultzaberger et al., 2010; Towbin et al., 2017). A striking demonstration of this was provided recently for the case of *expression*-fitness relationships – these were characterized in parallel for 81 diverse genes in the yeast, *S. cerevisiae*, by inserting in front of each gene a library of synthetic promoters of known strength (Keren et al., 2016). In addition to the commonly found concave function, a variety of other forms were uncovered, including step-like ones, invariant ones, peaked ones (*eg.* for certain genes with regulatory function; **Chapter 1**) and even multi-peaked ones (Figure 1.17). Importantly, genes from the same pathway or complex tended to display similarly expression-fitness curves, suggesting that these are shaped primarily by the cell-level function of a gene, rather than its specific biochemistry. The consequences of these different elasticity functions for the DME across different genes should therefore be a fruitful avenue for future research.

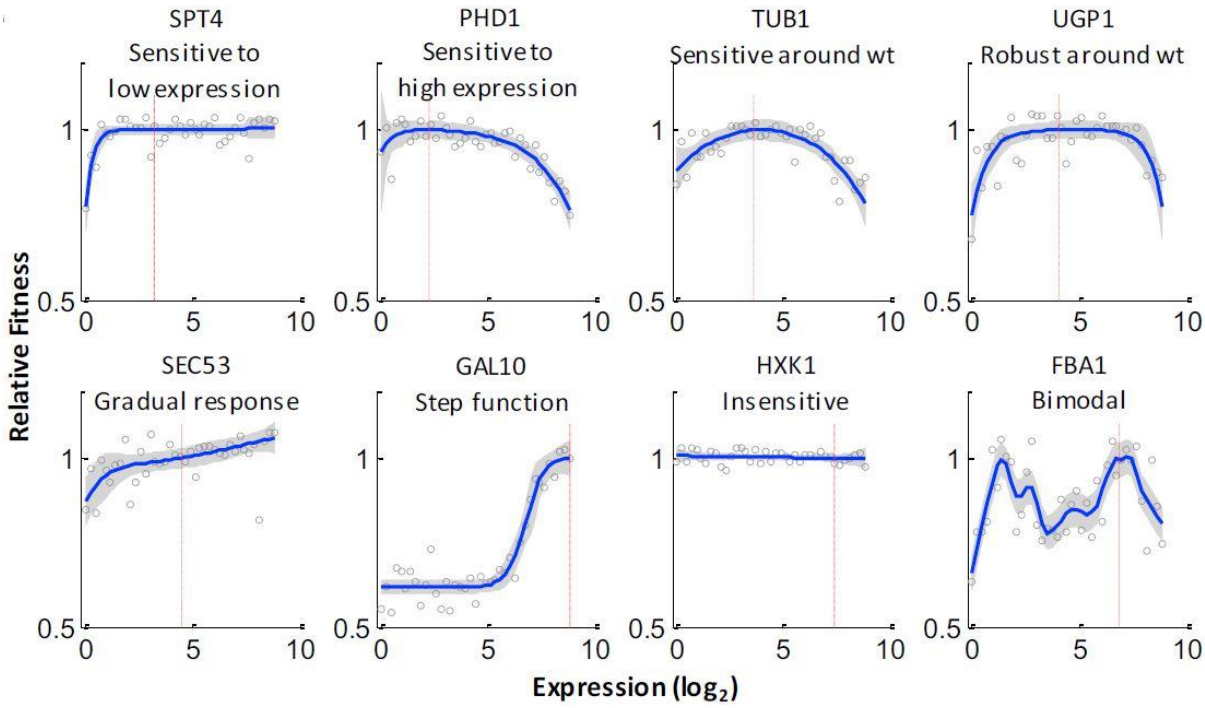


Figure 1.17: Expression-fitness functions for a diverse set of protein-coding yeast genes. Red lines mark wildtype expression levels. [Figure from (Keren et al., 2016)].

1.3.1.1.2.2 The DME in single functional non-coding RNAs

The DME for the few functional non-coding RNA sequences (*cis*-regulatory region, tRNA, snoRNA, twister ribozyme) so far examined appears to follow similar rules to proteins: deleterious and nearly-neutral modes and a miniscule proportion of beneficial mutations, again probably shaped in part by both a thermodynamic stability threshold and common saturating concave activity-fitness functions (Bendixsen et al., 2017; Bernet and Elena, 2015; Kobori and Yokobayashi, 2016; Li et al., 2016; Puchta et al., 2016). This is perhaps not surprising because, first, as for proteins, non-coding RNA function depends strongly on structure, and second, there is no obvious reason why the

dependence of cell fitness on RNA-regulated activities or direct RNA activities (especially those chosen for experimental investigation) would be fundamentally different to its dependence on protein activities.

1.3.1.1.2.3 The DME in single *cis*-regulatory DNA regions

In the case of *cis*-regulatory DNA sequences, measured DMEs are generally unimodal (Badis et al., 2009; Kinney et al., 2010; Lagator et al., 2016, 2017a; Warren et al., 2006), although they can be multimodal in more complex regulatory contexts (Lagator et al., 2017b, 2017a) or when a fraction of the sites function through specific base-pairing with other nucleic acids (and so are highly sensitive to mutation) (Boyle et al., 2017). The majority of mutations decrease DNA binding to regulators (again presumably because wildtype sequences are optimised for relatively strong binding), but of course the impact of this on downstream phenotypes depends if the regulation is activational, repressive or some complex mixture of both (Lagator et al., 2017b). The biophysical reason for a more uniform DME on direct biochemical phenotypes for *cis*-regulatory DNA sequences than for proteins and non-coding RNAs is not clear: the thermodynamics of DNA-protein binding can result in similar energy-phenotype functions to those of macromolecular folding described earlier (Lagator et al., 2017b; Mustonen et al., 2008; Vilar, 2010). It may be that in reality they are less steep (for example, because deleterious effects from misfolded molecules no longer contribute), or that *cis*-regulatory mutations tend to have smaller effects on binding energy than do

protein and non-coding RNA mutations on folding energy, and so sample a more local region of the energy space (see Figure 1.15). In any case, one result of this difference is that the DME in *cis*-regulatory DNA sequences might be even more sensitive to the precise forms of downstream phenotype-phenotype transformations than in proteins and non-coding RNA, as the direct phenotype space is more evenly explored.

1.3.1.1.3 Epistasis within single genes (intragenic epistasis)

Epistasis, in this thesis, describes interactions between mutations. More precisely, it is defined here as the deviation of an observed phenotype value from that expected if the constituent individual mutations combined additively on the log-scale (*ie.* multiplicatively on an absolute scale) (2000) (Figure 1.18). Epistasis is increasingly acknowledged to be critically important for medical and evolutionary genetics and bio-engineering: among other things, it confounds prediction of mutational effects, constrains adaptive paths, determines the benefit of sex and hinders efforts to increase yields and activities of industrially useful substances (Badano and Katsanis, 2002; Breen et al., 2012; Dipple and McCabe, 2000; Hansen, 2013; Kimura and Maruyama, 1966; Kondrashov, 1988; Kondrashov and Kondrashov, 2001; Manolio et al., 2009; Niederberger et al., 1992; Phillips, 2008; Scriver and Waters, 1999; Weinreich, 2006). Intragenic epistasis is logically shaped by the same forces as those shaping the DME, such as thermodynamics and elasticity functions (Lehner, 2011; de Visser et al., 2011),

and its basic statistical properties will now be summarised for the systems in which large-scale measurements have been made.

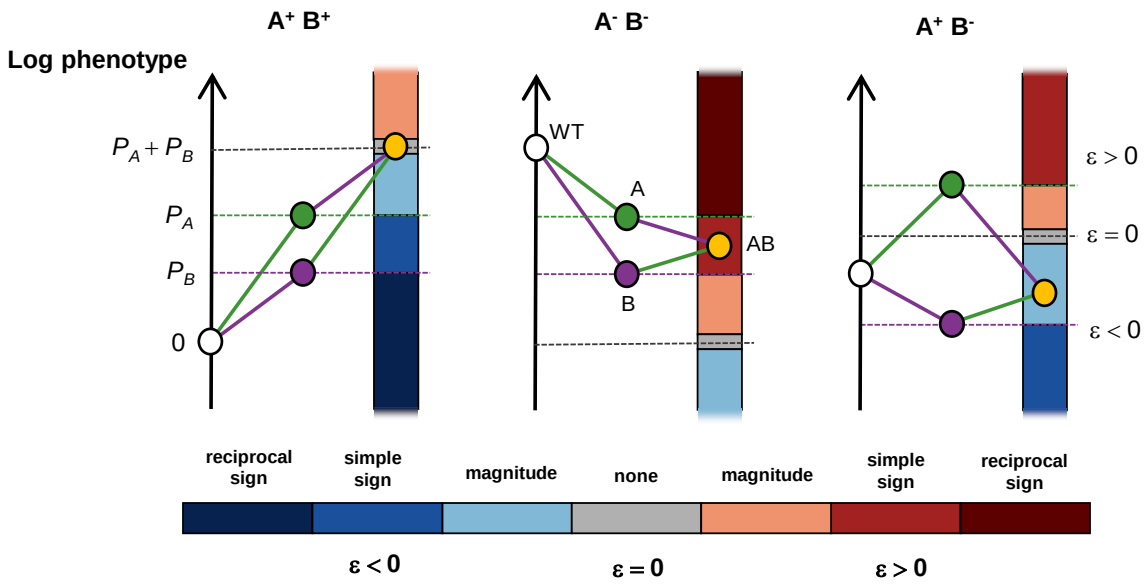


Figure 1.18: Types of pairwise epistasis possible for different types of mutation pairs. 'A' and 'B' are mutations, and superscript '+' and '-' denote that these individual mutations increase or decrease, respectively, the value of the measured phenotype, P. In all cases, the white point is wildtype and the orange point is the AB double mutant. The grey dashed line marks the sum of P_A and P_B, which is the *expected* value for the double mutant. Epistasis measures the deviation from this expectation, which may be either negative or positive, and it can be categorised as either magnitude (the direction of mutational effects do not depend on the other mutation) or sign type. Sign epistasis can be further categorised as simple (effect of one mutation changes sign in presence of the other) or reciprocal (effects of both mutations change sign in the presence of the other). The three examples shown are (left-right): no epistasis between a pair of positive-effect mutations, positive simple sign epistasis between a pair of negative-effect mutations, and negative magnitude epistasis between a positive-effect and negative-effect mutation.

1.3.1.1.3.1 Intragenic epistasis within single proteins

Within proteins, epistasis appears prevalent and predominantly negative, with a unimodal distribution. As the majority of mutations tend to reduce the value of the measured phenotype, this epistasis mainly represents *synergistic* interactions between deleterious mutations: the combined impact of multiple mutations is often *worse* than "the sum of the parts" (Bank et al., 2015, 2016; Melamed et al., 2013; Olson et al., 2014; Sarkisyan et al., 2016) (but (Araya et al., 2012; Diss and Lehner, 2018) are exceptions). Under the stability threshold model outlined above, such synergistic negative epistasis would indeed be expected to prevail between mutations having mildly destabilising effects, because, beginning from the stability plateau, the downwards slope initially becomes steeper as stability is decreased. As the threshold is crossed, however, the slope levels off again, which can result in positive epistasis being detected between highly destabilising mutations if the protein is non-essential or if an experimental measurement limit is approached (Bank et al., 2015; Bershtein et al., 2006; Diss and Lehner, 2018; Jacquier et al., 2013; Sarkisyan et al., 2016; Wylie and Shakhnovich, 2011) (Figure 1.19). Differences in the destabilising effects of mutations, wildtype stability and measurement precision/range could therefore explain discrepancies between studies as to the pervasiveness of epistasis as well as the relative fractions of positive and negative interactions. Negative epistasis between mutations of mildy deleterious biochemical effect could also result from concave saturating elasticity functions (described above) (Bank et al., 2015; Szathmary, 1993), for the same reason that it results from the stability function, and analogously to the classical molecular

hypothesis of genetic dominance and recessivity (Kacser and Burns, 1981). Of course, as for the DME, many specific cases will deviate from these general expectations (**Chapter 3**) and vary across proteins, such as specific local structural interactions which could generate sign epistasis, for example (sign epistasis is not predicted by monotonic functions like the common sigmoidal and concave ones discussed so far). Further, any single mutation could potentially effect multiple molecular phenotypes, including stability, activity (existing or new!) per folded molecule, folding kinetics, aggregation propensity, degradation rate and post-translational modification (DePristo et al., 2005; Echave and Wilke, 2017; Shah et al., 2015; Sikosek and Chan, 2014), so simplistic models should be treated as what they are.

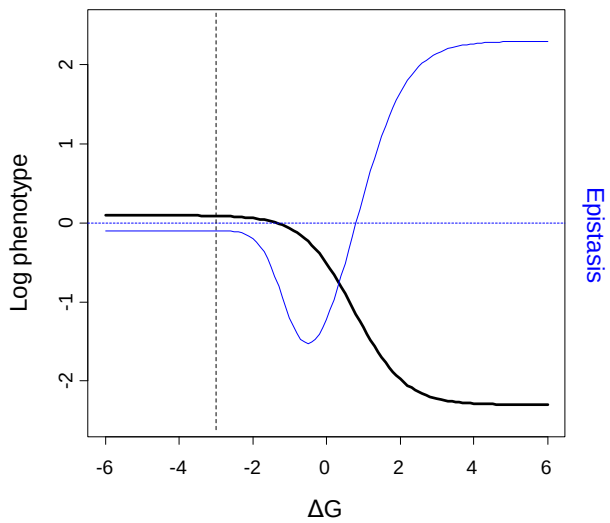


Figure 1.19: Trends of epistasis predicted by thermodynamic model of mutation effects. Black sigmoid curve shows the logarithm of a phenotype that increases proportionally with the fraction of natively folded protein molecules as a function of the free energy of folding, ΔG. A small background value (phenotype of 0.1 in the absence of any correctly folded molecules) has been applied to capture the situation for non-essential genes and/or the effect of measurement background/limits. If the phenotype can truly go to 0 in the limit of very high ΔG, the stability curve is no longer sigmoidal on this log scale, but has a concave shape, causing the epistasis curve (see below) to become increasingly negative, in a linear fashion, as ΔG increases. In reality, however, even if the phenotype really does approach zero, experimental limitations will likely result either in a background phenotype value in the absence of correctly folded protein, resulting in a log-sigmoid as shown here, or in a threshold being applied below which all mutants are considered null and therefore not considered for epistasis analysis. The formula and parameter values are the same as those in Figure 1.15, just with the addition of the 0.1 background phenotype. Dashed vertical line marks a hypothetical wildtype protein stability (-3 kcal/mole), located on the plateau of the sigmoid, for illustration. Blue curve shows the epistasis that would occur between pairs of mutations of identical ΔG effects, each of which individually displaces ΔG from the wildtype value to the value indicated by the x-axis. Dashed horizontal line marks the boundary between positive and negative epistasis (*ie.* zero epistasis). A transition from negative to positive epistasis is seen to occur as mutations become more strongly destabilising, due to the sigmoid shape of the stability curve. The shape of the epistasis curve could explain why both negative and positive epistasis is observed between mutations within proteins, as well as the existence of certain correlations between mutation effect size and epistasis (see below).

In addition to prevailing negative epistasis, a trend of "increasing losses" is sometimes detected in proteins (at least for mildly deleterious mutations), further supporting the idea that there exists some source of concavity in the phenotypic landscape (Bank et al., 2015; Diss and Lehner, 2018; Sarkisyan et al., 2016). This trend manifests as a positive correlation between the sum of the magnitude of individual mutation effects and the magnitude of epistasis they experience when combined. The flip-side of "increasing losses" between the predominantly deleterious mutations of the DME is "diminishing returns" between the rare mutations from the beneficial tail of the DME (MacLean et al., 2010; Nagel et al., 2012; Schenk et al., 2013): synergistic negative epistasis increases with the *increasing* downward slope moving away from the plateau, and antagonistic negative epistasis increases with the *decreasing* upward slope. The evolutionary outcome of these trends should be a more rapid purging of deleterious mutation combinations (negative selection), but a slower rate of adaptation resulting from beneficial combinations (positive selection), than would be expected in the absence of epistasis.

1.3.1.1.3.2 Intragenic epistasis within functional non-coding RNAs

Epistasis between mutations residing in the same non-coding RNA molecule, as for the DME and again probably for the same reasons, appears similar to the case of proteins: it is common and usually biased towards negativity (Bendixsen et al., 2017; Domingo et al., 2018; Kobori and Yokobayashi, 2016; Li and Zhang, 2018; Li et al., 2016;

Puchta et al., 2016). Notably, these experimental results are in direct contrast to the predominance of positive epistasis predicted by earlier computational RNA folding models (Wilke and Christoph, 2001; Wilke et al., 2003), but it is not clear whether this is due to the binary nature of these models or the fact that they do not use naturally evolved sequences as their starting point (Bendixsen et al., 2017) - when higher-order epistasis was examined experimentally, positive and negative epistasis were found to be equally prevalent, with many cases of sign-epistasis, leading the authors to conclude a rugged multi-dimensional fitness landscape on the *global* level ((Domingo et al., 2018); cf. (Bank et al., 2016) for a protein region).

1.3.1.1.3.3 Epistasis within single *cis*-regulatory DNA regions

Epistasis within *cis*-regulatory DNA has also been studied experimentally in a limited number of contexts. Within a mammalian *Rhodopsin* promoter region bound by at least two transcription factors, epistasis for expression was found to be biased towards negativity (Kwasnieski et al., 2012). More strikingly, in a recent study of the direct effect of target-site mutations on dCas9-DNA binding, negative epistasis for initial association rate between the universally deleterious single mutations was found to be ubiquitous – binding was essentially always worse than would be predicted from the simple addition of individual mutation effects (Boyle et al., 2017). dCas9-DNA binding may be rather unrepresentative, however, as its function is clearly not regulatory (it is immune), and it is mediated by DNA:RNA base-pairing. Another team recently

examined the expression epistasis between mutations within both the same and different operators in two different promoters: the *E. coli araBAD* promoter (Lagator et al., 2016) and the lambda bacteriophage P_R promoter (Lagator et al., 2017b). Although these two promoters possess substantially different architectures, both exhibited a predominance of negative epistasis, whether mutations resided in the same or different operators and whether active repressor proteins were present in high concentration or not. The authors showed that, at least for the simple case where expression is determined by DNA binding to a single regulator, this is to be expected from thermodynamic considerations: the free binding energy-expression curve resulting from their generic model is, once again, of a concave, saturating form (Lagator et al., 2017b). Notably, though, in the lambda bacteriophage P_R promoter, the fraction of positive interactions increased when active repressor concentration was increased, as did the frequency of the most extreme form of interaction, reciprocal sign epistasis (from 8% to 66%). This is explained by the fact that repressor binding sites (operators) and RNA polymerase binding sites overlap in this promoter, and so promoter mutations will tend to effect binding to both of these – one of which decreases transcription and the other of which increases transcription. The proposed thermodynamic model suggests that under these constraints, the nature of epistasis now depends on both the concentrations of repressor and RNA polymerase and on the sign and relative magnitude of the individual mutation effects on binding to both proteins (Lagator et al., 2017b).

Cis-regulatory DNA sequences have thus proved to be an excellent model system for studying how epistasis can arise through the inherent *molecular* pleiotropy of mutations: they directly affect multiple molecular phenotypes (here, binding energy for different regulatory proteins), often differentially, and the *measured* phenotype (here, expression) is then some simple (**Chapter 4**) or complex function of these multiple input phenotypes. In reality, as for proteins and non-coding RNAs, many more molecular phenotypes than those typically considered are potentially impacted by an individual mutation, such as DNA structure (Rajkumar et al., 2013), binding site accessibility (Levo and Segal, 2014), regulator protein cooperativity (Todeschini et al., 2014) and long-range DNA looping (Levine et al., 2014), likely accounting for the significant fraction of observed interactions that are not explained by simple binding energy models (Lagator et al., 2017b). Current empirical studies thus present a minimal mechanistic picture of the potential sources of epistatic interactions, rightly so, and the goal now should be to find additional molecular phenotypes that may account for a substantial amount of the unexplained epistasis.

Finally, *cis*-regulatory sequences appear to play a major role in evolutionary processes (Wittkopp and Kalay, 2012; Wray, 2007), yet none of the studies above have assessed epistasis at the level of fitness. Just as for proteins and RNA, in addition to the mechanisms just discussed, this is bound to be also shaped by the expression-fitness elasticity functions of the regulated genes. Fortunately, these are far more accessible to

high-throughput genome-wide characterization than are pure activity-fitness functions (Keren et al., 2016).

1.3.1.2 Two genes

Adding just one more gene, let alone several, to the kinds of high-throughput studies that have provided such rich insights for single genes generally requires more than simply twice the effort (let's call it experimental effort epistasis). For many systems, the problem is that the length of DNA fragments required to carry two different genes exceeds those amenable to the highest-throughput sequencing technologies (max. ~1kb with Illumina). Presumably largely for this reason, very few deep-mutational scanning studies have assayed more than one gene or even regulatory sequence simultaneously, but unique DNA barcodes that allow a long sequence to be broken up for sequencing and then reassembled *in silico* provide one workaround (Sarkisyan et al., 2016).

1.3.1.2.1 Epistasis between two genes (intergenic epistasis)

1.3.1.2.1.1 Intergenic epistasis between two physically interacting partners

A recent exception was the measurement of intermolecular epistasis between a library of point-mutants of the leucine zipper domains of two proto-oncogenes, FOS and JUN (Diss and Lehner, 2018). The products of these two genes physically interact through these domains to form a transcription factor complex, AP-1. In this case, the technical challenge of long sequencing fragments was overcome by isolating the two small leucine zipper domains, which are presumed to function in a modular manner, and cloning

them adjacent to each other in the absence of the rest of the native proteins. An artificial complementation assay, in which intermolecular FOS-JUN binding drives the assembly of a drug-resistance enzyme, was used to link binding strength to an easily measurable phenotype (cell growth in the presence of the drug). Further, the relationship between abundance of the complementation complex and growth rate is well-characterised, and expected to be approximately linear, removing the common added complication of non-linear elasticity functions discussed above. Intermolecular epistasis was found to be common and just slightly biased towards negativity. Importantly though, a characteristic relationship between the effect of single mutations and the interaction they experienced when combined suggested that the epistasis could be partly explained by a sigmoidal thermodynamic fitness landscape similar to that proposed to explain the DME and epistasis within single genes (but this time based on intermolecular binding rather than folding) (Figure 1.19). Indeed, applying this model removed the correlation between individual effects and epistasis, and increased the percentage of explained variance in double-mutant phenotype scores from ~86% (under a simple multiplicative model) to ~89%. This is a promising conclusion, as it implies that general rules for single genes may also be applicable to some degree for multiple genes whose products interact physically. It may also, however, be a particularity of the system, as the folding of leucine zippers like FOS and JUN is known to be coupled to their binding (Patel et al., 1990; Thompson et al., 1993) - studies on other pairs of

binding partners are therefore necessary to uncover to what extent rules can be generalised from single proteins to multi-protein complexes.

An impressive effort was also recently made to characterise the epistasis between a large number of mutants of a particular repressor protein and also of its *cis*-regulatory DNA target at low-throughput, using a fluorescent reporter to measure expression (Lagator et al., 2017a). Epistasis was detected for about half of the 150 pair-wise interactions tested, the majority of which were positive. As before, much, but not all, of this epistasis could be rationalized in terms of the promoter architecture, which contained overlapping binding sites for the mutated repressor protein and RNA polymerase, raising hope that a set of rules may exist that adequately predicts epistasis both within promoters, and between promoters and regulators, when promoter architecture is known.

1.3.1.2.1.2 Intergenic epistasis between two functionally interacting partners

Although epistasis between protein-protein and protein-DNA binding partners is of great importance, the majority of genes in a given genome are expected to interact *indirectly*, through metabolic, regulatory and signalling networks. As explained in the first part of the introduction, metabolic networks are perhaps the most tractable of these, being based on simple mass-flow. Metabolic Control Analysis (MCA) provides a rigorous framework to explore how pathway phenotypes such as flux and metabolite concentrations depend on the activity of several enzymes simultaneously, and thus

enables predictions of inter-enzyme epistasis (Szathmary, 1993). Several small-scale studies provide support for the validity of MCA in general (Chou et al., 2014; Dykhuizen et al., 1987), but its rich predictions regarding the epistasis between genes connected by metabolic pathways have until now not been tested (**Chapter 4**). Importantly, the nature of fitness epistasis is predicted to vary considerably depending on which pathway phenotypes are under selection (flux, steady-state metabolite levels), the pathway position of any selected metabolites relative to the two enzymes considered, and the type of selection operating (directional, stabilising) (Szathmary, 1993) (Figure 1.20). This points to the necessity of uncovering which phenotypes are typically under selection if we are to use such systems-models to predict sequence-fitness relationships. Epistasis between two or more genes in signal-flow networks has also never been tested on a systematic scale, but a recent small-scale study on a synthetic gene regulatory cascade found a surprisingly high frequency of sign-epistasis simply at the level of expression (briefly, explained by the fact that changes in the activity of one regulator shift the optimal activity of other regulators) (Nghe et al., 2018). Sign epistasis has strong consequences for both evolution and the predictability of mutation effects, and so regulatory and signalling networks are a key area of future study.

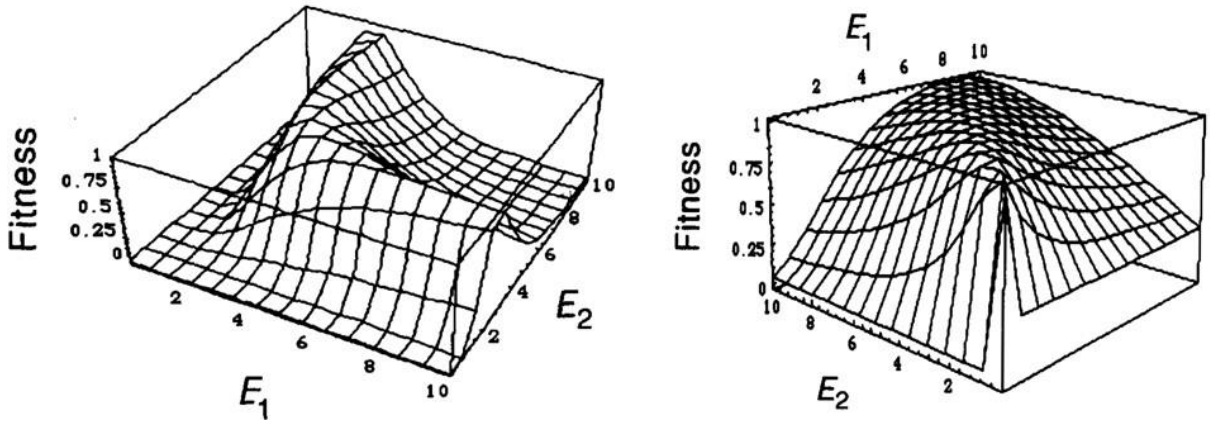


Figure 1.20: Two-dimensional activity-fitness functions predicted from Metabolic Control Analysis. E_1 and E_2 are the activities of two adjacent enzymes in a linear metabolic pathway. In both landscapes, fitness is assumed to depend solely on the steady-state concentration of a pathway intermediate, in a Gaussian manner (*ie.* stabilising selection is assumed to operate on the intermediate). The only difference is that, in one case, the intermediate lies downstream of the two enzymes (left), and in the other, it lies between them (right). The two landscapes have strikingly different forms, resulting in different expectations of inter-enzyme epistasis. Further, in both cases, trends of inter-enzyme epistasis will depend on the position of the wildtype and the distribution of mutation effects on enzyme activity. [Figures from (Szathmary, 1993)].

1.3.1.3 The genome

1.3.1.3.1 Experimental approaches for genome-wide genotype-phenotype mapping

Scaling up deep-mutational scanning experiments to the scale of the genome is at present out of reach: bottlenecks include the precise, genome-wide introduction of individual mutations (mutagenesis efficiency and accuracy), sequencing costs and linking mutations at distant loci. The first is improving with advances in genome engineering, particularly from CRISPR-Cas9-based methods (Barbieri et al., 2017;

Haimovich et al., 2015; Roy et al., 2018). The second is continuing to improve, following a long-term trend of decreasing costs (but see (Muir et al., 2016) for the alternative challenge of managing increasing amounts of data); and the third is becoming more feasible with emulsion-based generalised DNA assembly technologies that encapsulate single cells and enable distal DNA sites to be linked by sequencing (either by directly ligating mutated sites adjacent to each other (Haliburton et al., 2017; Zeitoun et al., 2015) or, more scalably, by ligating them to a cell-specific DNA barcode (Zeitoun et al., 2017)) (Figure 1.21).

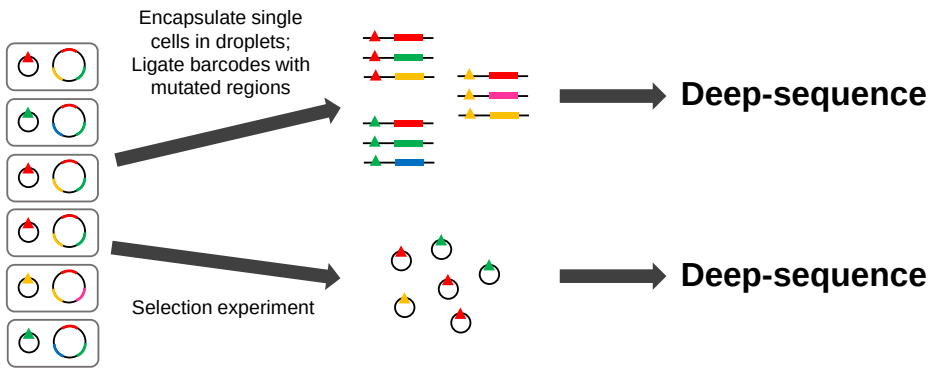


Figure 1.21: bTRACE, a general high-throughput method for analysing the effect of known genome-wide mutation combinations. A multiplex genome engineering method is used to construct a library in which each cell can contain multiple mutations throughout the genome, and this library is itself transformed with a library of plasmids carrying highly diverse DNA barcodes (triangles), such that each cell now contains a unique barcode. Single cells are then encapsulated in emulsion droplets, where they are lysed, and a targeted binary PCR assembly reaction is performed to ligate barcodes adjacent to chosen mutated genomic regions. The emulsion is then broken, and deep-sequencing of the assembled product pool allows reconstruction of the complete genotype associated to each barcode. In parallel, the library can be phenotyped by one of the deep-sequencing techniques discussed previously, with only the small DNA barcodes now needing to be sequenced, allowing genome-wide mutation combinations to be linked to an amenable trait at high throughput [Figure based on (Zeitoun et al., 2017)].

In the meantime, a plethora of "functional genomics" systematic genome-wide studies have been performed, especially in yeast, that measure the effects (typically fitness) of different types of large perturbations in single or multiple (maximum of 3) genes (deletion, overexpression, knockdown, transposon insertion; typically one or two perturbations per gene) (Baba et al., 2006; Babu et al., 2014; Boutros et al., 2004; Breslow et al., 2008; Collins et al., 2007; Costanzo et al., 2010, 2016; Davierwala et al., 2005; Decourty et al., 2008; Douglas et al., 2012; Fuchs et al., 2010; Gagarinova et al., 2016; Giaever et al., 2002; Jaffe et al., 2017; Kamath et al., 2003; Kuzmin et al., 2018; Nichols et al., 2011; Onge et al., 2007; van Opijnen et al., 2009; Roguev et al., 2008; Sameith et al., 2015; Schuldiner et al., 2005; Szappanos et al., 2011; Tischler et al., 2006; Tong, 2004; Tsherniak et al., 2017; Ursell et al., 2017). These provide rich functional datasets, but it is not clear that any genotype-phenotype inferences would generalise to the less extreme perturbations (*eg.* point mutations) often found in nature. Further, the majority are biased towards altering functions of known genes. An approach to study genome-wide genotype-fitness relationships at the other extreme is mutation accumulation experiments and the analysis of natural DNA sequence data. These benefit from sampling *naturally-occurring* mutations, which may be very different to those introduced experimentally (even for point mutations), and having the potential to capture mutations of effects too weak to be quantified directly, but they rely purely on inferences (Eyre-Walker and Keightley, 2007). In between these two extremes (in terms of artificiality) are experiments that directly measure the

effects of randomly induced or collected mutations (Bonhoeffer, 2004; Carrasco et al., 2007; Domingo-Calap et al., 2009; Peris et al., 2010; Sanjuan et al., 2004a; Szafraniec et al., 2003; Wloch et al., 2001), or of mutations that are detected during experimental evolution (Caudle et al., 2014; Chou et al., 2011, 2014; Flynn et al., 2013; Khan et al., 2011; Kryazhimskiy et al., 2014; Kvitek and Sherlock, 2011; Rokyta et al., 2011; Venkataram et al., 2016), all of which provide datasets orders of magnitude smaller than do systematic perturbation studies.

1.3.1.3.2 The genome-wide DME

An early finding of the genome-wide perturbation studies was that, in most organisms and in permissive conditions, the majority of genes are inessential (Baba et al., 2006; Gerdes et al., 2003; Giaever et al., 2002; Kim et al., 2010; Sassetti et al., 2003; Viswanatha et al., 2018; Yamamoto et al., 2009; Zhang and Lin, 2009). A notable exception is the “minimal bacterium”, *Mycoplasma genitalium* (Glass et al., 2006; Hutchison III et al., 1999), which was in fact the first organism in which gene essentiality was examined directly (Hutchison III et al., 1999), demonstrating a frequent irony in experimental biology: the first choice of experimental system is usually based on convenience, which in some cases results in it being an utterly unrepresentative outlier. “Chemical genomics” approaches which phenotype genome-wide perturbation libraries in many different defined environments, perhaps unsurprisingly, reduce the fraction of inessential genes by revealing *conditionally*

essential genes that are necessary for growth in at least one of the environments tested (Nichols et al., 2011). The interpretation of these classifications (essential, conditionally essential, inessential) is not clear, however, as they are clearly fully-dependent on the environments tested (see **Environment**).

Interestingly, it seems from studies reporting quantitative fitness effects of genome-wide perturbations that the DFE of single-gene deletions/disruptions (Baryshnikova et al., 2010; van Opijnen et al., 2014; Wang et al., 2018) might be qualitatively similar to the DFE of randomly sampled/induced genome-wide mutations (Eyre-Walker and Keightley, 2007), itself similar to the DFE of single point-mutations in proteins and non-coding RNAs: a nearly-neutral mode with a heavy negative tail, a very deleterious/lethal mode (sometimes overlooked) and a small proportion of beneficial effects (see (Bataillon and Bailey, 2014; Eyre-Walker and Keightley, 2007) for more fine-scale properties), and uncovering the precise reasons for this universality should prove to be a highly worthwhile endeavour.

In the meantime, top-down heuristic phenotype-fitness models have provided useful unifying frameworks with which to capture such common trends found across these different scales and species, as they do not rely on system-specific mechanistic details. In particular, Fisher's Geometric Model of Adaptation (FGMA), originally proposed simply as a convenient metaphor for phenotypic adaptation (Fisher, 1930), can correctly predict the oft-observed shifted reflected Γ-shape of the nearly-neutral mode

of the DFE (although not generally the strongly deleterious/lethal mode) (Bank et al., 2014; Bataillon and Bailey, 2014; Chevin et al., 2010; Jacquier et al., 2013; Martin, 2014; Martin and Lenormand, 2006; Tenaillon, 2014; Trindade et al., 2012). FGMA assumes that fitness depends on a certain number of independent traits and that, in a given environment, there is a single optimum value for each of these traits, with fitness decreasing smoothly with increasing distance from the optimum (generally modelled as a Gaussian function). Mutations are assumed to be partially or fully pleiotropic (in which case their directional effect in phenotype space is completely unconstrained), and their phenotypic effects are typically drawn from a multivariate normal distribution with a mean of zero (Figure 1.22). When the number of idealized traits under selection is not too large (which appears to generally be the case) and the wildtype resides close to the optimum (*ie.* it is well-adapted), the DFE takes the aforementioned shifted reflected Γ-shape, with a small number of weakly beneficial mutations, a near-neutral mode and a heavy deleterious tail. In addition to the DFE, FGMA predicts the common patterns of epistasis observed between beneficial mutations in systems of all scales (see also below): a general predominance of negative (antagonistic) epistasis and, more specifically, a trend of diminishing returns (Blanquart et al., 2014; Martin et al., 2007; Rokyta et al., 2011). These agreements are not so surprising when we consider the similarities of a Gaussian FGMA to the sigmoidal and simple concave phenotype-fitness functions previously demonstrated to explain these patterns: in all cases, as the fitness maximum is approached, the upwards

slope becomes increasingly flat (*ie.* they are all locally concave at high fitness). One difference is that, as opposed to the sigmoidal and simple concave functions, the non-monotonicity of FGMA also predicts sign epistasis (for example, a mutation that is beneficial in a maladapted background may become deleterious in a well-adapted background due to optimum overshooting (Blanquart et al., 2014)). Overall, the success of FGMA suggests that at least some of the repeatedly observed properties of the genotype-fitness relationship may be remarkably predictable without the need for any mechanistic knowledge.

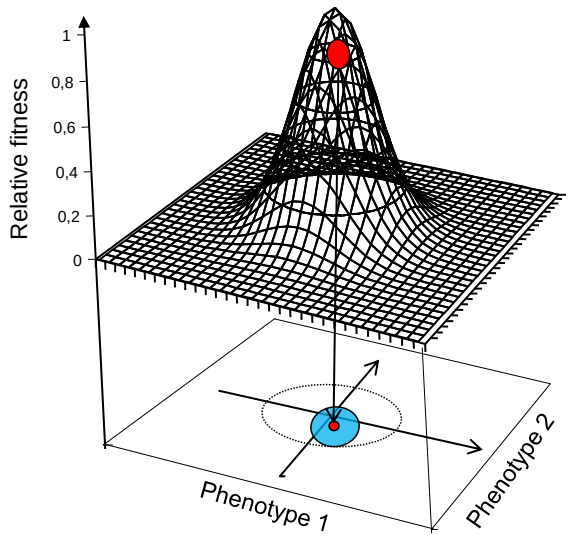


Figure 1.22: The canonical isotropic Fisher's Geometric Model of Adaptation (FGMA) in two dimensions. The fitness surface here is described by an isotropic multivariate Gaussian, centred at the origin of phenotype space. The red point (shown both on the fitness surface and projected onto phenotype space) represents a wildtype genotype, and the blue circle represents an isotropic cloud of (mild) one-step mutations. Under such a model, when the wildtype is near the optimum, the DFE will contain a small fraction of weakly beneficial mutations (those that move closer to the optimum) and a heavy tail of deleterious mutations. [Figure adapted from (Gros et al., 2009)].

1.3.1.3.3 Genome-wide epistasis (intergenic epistasis)

The nature of epistasis appears to be less general across systems. The genome-wide deletion analyses (like the single gene and regulatory sequence studies) tend to find a predominance of negative interactions, though still alongside a substantial proportion of positive ones (Babu et al., 2014; Costanzo et al., 2010, 2016; Onge et al., 2007; Roguev et al., 2008; Szappanos et al., 2011), but it should be noted that these have for now mainly been performed in yeast. In line with Flux Balance Analysis (FBA) pair-wise gene perturbation predictions (He et al., 2010; Segrè et al., 2005; Szappanos et al., 2011), comparing interaction profiles for different genes has provided information on the topology of their functional connections, providing "wiring diagrams" of cell function (Costanzo et al., 2016) (Figure 1.23). For example, genes from redundant pathways tend to undergo negative synergistic interactions with each other, and genes from the same pathway tend towards positive antagonistic interactions (perhaps contributing to the differential proportions of negative and positive epistasis) (Avery and Wasserman, 1992; Battle et al., 2010; Beltrao et al., 2010; Breslow et al., 2008; Fu et al., 2017; Lehner, 2011; Onge et al., 2007; Ye et al., 2005). As stated above, though, it is not clear that these rules should generalise to mutations other than the complete loss-of-function ones making up the vast majority of these datasets (hypomorphic alleles of essential genes are the exception) (**Chapter 4;** (Szathmary, 1993; Xu et al., 2012)).

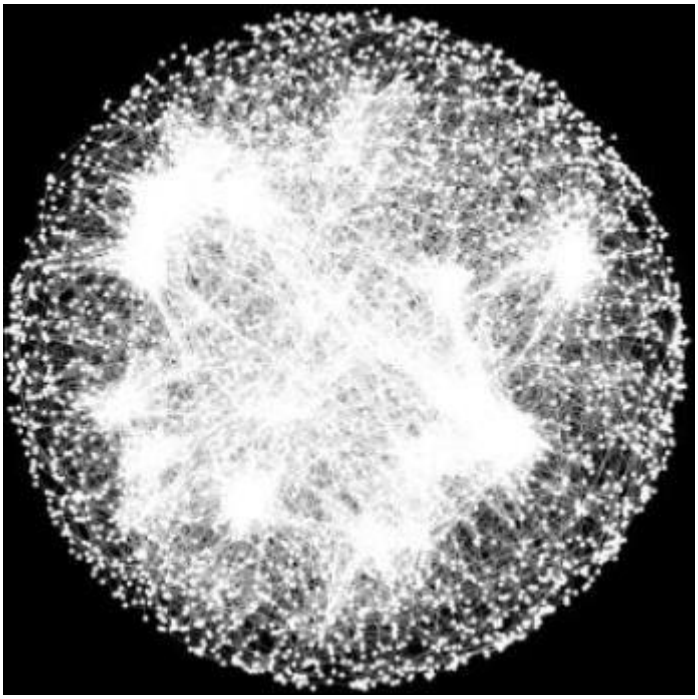


Figure 1.23: A global network of gene-gene interaction profile similarities. Nodes are yeast genes and edges connect genes with similar genome-wide fitness interaction profiles, revealing functional modules. [Figure from (Costanzo et al., 2016)].

On the other hand, epistasis between naturally-occurring mutations in viral and cellular genomes appears to be biased towards positive interactions, with positive epistasis more common between the relatively frequent deleterious mutations and negative epistasis more common between rarely-occurring beneficial mutations, *ie.* an overall trend of antagonism (Bonhoeffer, 2004; Burch, 2004; Caudle et al., 2014; Chou et al., 2011, 2014; Flynn et al., 2013; Khan et al., 2011; Kryazhimskiy et al., 2014; Lalić and Elena, 2012; Maisnier-Patin et al., 2005; Rokyta et al., 2011; Sanjuan et al., 2004b; Schoustra et al., 2016; de Visser et al., 2011). This antagonism represents a kind of genomic buffering process: combinations of deleterious mutations are “less bad”

than expected from simple additivity, and beneficial combinations are “less good” (and beneficial/deleterious combinations have rarely been studied explicitly), but a true mechanistic explanation is lacking for now. Further, the trend of diminishing returns between beneficial mutations found in single genes is often also found at the genome scale, suggesting analogously the saturation, and sometimes even an optimum overshoot (potentially causing sign epistasis) (Rokyta et al., 2011), by successive mutations of some phenotype that is contributing to fitness, or of the phenotype-fitness function itself (reviewed in (Berger and Postma, 2014)). These general fitness-level trends found for real mutations are rather encouraging for the predictability of adaptive dynamics, despite the underlying genetic and even phenotypic complexity ((Kryazhimskiy et al., 2014); see FGMA discussion above).

1.3.1.4 The environment

It is clear that the environment affects genotype-phenotype relationships, and so for a complete understanding of them we must consider them across different environments. Indeed, the few studies that explore large mutant sets across large numbers of environments find fundamental changes, such as the proportion of essential genes (Nichols et al., 2011). Even such ambitious large-scale studies explore only a vanishing fraction of potentially relevant fixed environments, though, not to mention dynamic ones.

Studies examining environmental effects therefore tend to have the ambition of proof-of-principle, rather than exhaustive sampling, and these have produced myriad examples of the environmental-dependence of both mutation effects and epistasis (although it should be noted that this is an area potentially ripe for publication bias). It remains extremely difficult to form any general conclusions, however, as "environment" is such a broad term. It may refer, for example, to the concentration of small molecules (gene expression inducers, enzyme substrates, cofactors, antibiotics) which have some expected specific role the system of study (Dean, 1995; Lagator et al., 2016, 2017b; Melnikov et al., 2014; Nghe et al., 2018; Shultzaberger et al., 2010; de Vos et al., 2013, 2015; Wrenbeck et al., 2017), precise physico-chemical parameters known to matter in *in vitro* studies (Hayden and Wagner, 2012; Hayden et al., 2011), or more general "pleiotropic" factors such as temperature, chemical stresses, complex nutrients or even host (Bank et al., 2014; Caudle et al., 2014; Dandage et al., 2018; Flynn et al., 2013; Fragata et al., 2018; Hietpas et al., 2013; Jagdishchandra Joshi and Prasad, 2014; Lalić et al., 2011; Li and Zhang, 2018; Mavor et al., 2016). It will be important going forward to develop a more systematic approach to the environment, focussing on relevant, informative experimental conditions.

1.4 Outline of the original research chapters included in this thesis

The three research chapters presented here aim to help fill some of the gaps in our understanding of the genotype-phenotype relationship highlighted in the Introduction.

Chapter 1 describes the results of a deep-mutational scanning experiment on a global transcriptional regulator in *E. coli*. Fitness was measured for a comprehensive set of single codon-substituted CRP mutants in environments containing different concentrations of CRP's specific allosteric activator, cAMP, and sodium chloride, a stressor molecule known to induce a CRP-mediated stress response. This study addresses the lack of empirical DFEs for highly pleiotropic regulators, which are frequent targets for adaptation to new environments and whose behaviour is likely to be highly influenced by the nature of the complex networks in which they are embedded.

Chapter 3 describes the results of a deep-mutational scanning experiment on the model β-lactamase antibiotic-resistance protein, TEM-1. Fitness effects were measured at high precision for a comprehensive set of single and double codon substitutions in a small region encoding an α-helix, which is expected to generally have no direct role in TEM-1 function. Any fitness changes caused by these mutations are thus expected to be exerted primarily through structural effects. This study adds to existing fitness and epistasis data for single model proteins, contributing a new level of precision and a well-controlled system designed to precisely unravel the contribution of different phenotypic dimensions to the genotype-fitness relationship.

Chapter 4 describes the results of a deep-mutational scanning experiment on two promoters driving the expression of two genes encoding enzymes that participate in the same metabolic pathway. Fitness effects were measured at ultra-high precision for a comprehensive set of single-nucleotide substitutions in the core region of one or both promoters. Artificial promoters were used so that their activity could also be independently controlled by chemical inducers, allowing fitness and epistasis to be measured from three different regions of expression space. This study addresses the lack of intergenic epistasis data for genes interacting indirectly through their common participation in molecular pathways, and reveals the existence of remarkably diverse types and trends of fitness epistasis for such a simple system, which can be explained by the inherent molecular pleiotropy of mutation effects on a few key phenotypes (here, flux, metabolite toxicity and expression burden).

2 Single-Mutation Fitness Landscape of a Global Transcriptional Regulator across Environments

2.1 Introduction

The Distribution of Fitness Effects (DFE) of new mutations is one of the key evolutionary parameters, influencing for example the rate of and potential for adaptation (Chevin et al., 2010; Hoffmann and Sgrò, 2011), the maintenance of quantitative and molecular genetic variation (Charlesworth et al., 1995; Hill, 2010), the rate of genomic decay from Muller's Ratchet (Loewe, 2006), and the benefit of sex (Otto and Lenormand, 2002; Peck et al., 1997) (reviewed in (Bataillon and Bailey, 2014; Eyre-Walker and Keightley, 2007; Keightley and Eyre-Walker, 2010)). Moreover, it is central to our understanding of complex disease (Eyre-Walker and Keightley, 2007).

The most direct and efficient strategy for characterising the DFE is to introduce a large number of random point mutations into the organism of interest and quantify each of their effects with a fitness assay (with the only limitations being the sensitivity of the assay to weak-effect mutations and the fact that random mutations may not reflect the actual spectrum of new mutations) (Eyre-Walker and Keightley, 2007). At present, such an approach is only feasible for the smallest of viral genomes (Carrasco et al., 2007; Domingo-Calap et al., 2009; Peris et al., 2010; Sanjuan et al., 2004). On a finer scale, the DFE for individual genes may also prove informative. Indeed, the recent development of high-throughput, Next Generation Sequencing-enabled bulk competition assays (Hietpas et al., 2011) has made it possible to rapidly and

comprehensively characterise all possible single nucleotide or codon substitutions in a given gene (Klesmith et al., 2015; Kowalsky et al., 2015; Li et al., 2016; Melnikov et al., 2014; Roscoe et al., 2013; Wrenbeck et al., 2017). Such comprehensive single-mutation fitness landscapes have now been mapped for regions or the entire length of a variety of genes, encoding products including a chaperone protein (Bank et al., 2014; Hietpas et al., 2011, 2013; Jiang et al., 2013), ubiquitin (Mavor et al., 2016; Roscoe et al., 2013), poly(A)-binding protein (Melamed et al., 2013), antibiotic-resistance enzymes (Dandage et al., 2018; Firnberg et al., 2014; Jacquier et al., 2013; Melnikov et al., 2014), metabolic enzymes (Chan et al., 2017; Jiang et al., 2016; Klesmith et al., 2015; Wrenbeck et al., 2017), a tRNA (Li and Zhang, 2018; Li et al., 2016) and a small nucleolar RNA (Puchta et al., 2016).

The emerging picture from these studies is that the DFE for single genes tends to be bimodal, with weakly deleterious and highly deleterious/lethal modes and a vanishing fraction of beneficial mutations (the proportion of mutations in each of these categories varies widely across studies, however). Encouragingly, a similar trend is found at the scale of the whole genome (Eyre-Walker and Keightley, 2007), giving hope that the DFE in single genes may indeed be informative for the genome-wide DFE. In addition, the large-scale sequence-fitness maps obtained from these deep-mutational scanning experiments have proved to be rich resources for inferring structural and mechanistic details (Chan et al., 2017; Hietpas et al., 2011; Jiang et al., 2013, 2016; Mavor et al.,

2016; Melamed et al., 2013; Melnikov et al., 2014; Roscoe et al., 2013). Single-gene sequence-fitness maps are therefore well established as valuable sources of insight in both evolutionary and molecular biology.

2.1.1 Global transcriptional regulators

An important class of genes for which such maps are currently lacking is those encoding global transcriptional regulators. These provide particularly intriguing subjects because they are frequently found to be among the first targets of adaptation to new environments, both in the laboratory and in nature (Damkiaer et al., 2013; Hindré et al., 2012; Saxer et al., 2014; Tenaillon et al., 2012). They are also interesting at the molecular level, due to their typically relaxed requirements for DNA sequence recognition (Badis et al., 2009; Shultzaberger et al., 2012; Slattery et al., 2014) and their ability to function through multiple mechanisms and modes of action (*ie.* a single regulator can act as both activator and repressor) (Browning and Busby, 2016). Further, some have been found to be remarkably capable of developing novel molecular functionality *via* altered DNA-binding specificity, resulting in the recognition of new sets of targets and thus potentially regulatory network rewiring (Shultzaberger et al., 2012). Such rewiring should be of profound importance to evolution, as it entails a transition in genome-wide genetic constraints, altering the future set of potentially adaptive paths (Hindré et al., 2012). Finally, the DFE of a gene depends fundamentally on its product's activity-fitness function (Jiang et al., 2013), and this is

likely to be rather different for global regulators than for the majority of genes characterised so far (for whom it is typically expected to be monotonic). This distinction is reasoned on an evolutionary level, as the known environmentally-responsive role of regulators suggests that there is some regulator activity level that is optimal for fitness in a given environment (Towbin et al., 2017), and also on a molecular level, as global regulators are embedded within complex regulatory networks, including feedback loops (Seshasayee et al., 2006), suggesting the existence of complex activity-fitness functions.

2.1.2 The cyclic AMP receptor protein (CRP) of *Escherichia coli*

We therefore chose to characterise the single-mutation fitness landscape of a canonical global transcriptional regulator, the cyclic AMP (cAMP) receptor protein (CRP) of *Escherichia coli* (Kolb et al., 1993). CRP is a conditionally essential homodimeric protein implicated in the regulation of an array of physiological processes: most famously, carbon metabolism (Görke and Stülke, 2008; Kolb et al., 1993), but also nitrogen assimilation (Mao et al., 2007), iron uptake (Zhang et al., 2005), osmoregulation (Balsalobre et al., 2006; Landis et al., 1999), biofilm formation (Jackson et al., 2002) and multidrug resistance (Nishino et al., 2008), to name a few. CRP responds to the cell state *via* the small signalling molecule, cAMP, which allosterically activates CRP's C- terminal helix-turn-helix DNA-binding domain upon

binding its N-terminal region (Fic et al., 2009; Kolb et al., 1993). It directly regulates > 500 target genes through binding their promoter regions (Gama-Castro et al., 2016), but has also been found to interact with thousands of weaker sites throughout the chromosome (Grainger et al., 2005). CRP regulates transcription initiation most often by activation, but also by repression, through diverse mechanisms, acting both alone and in combination with other CRP dimers and/or other transcription factors (TFs) (Kolb et al., 1993). Phylogenetically, CRP belongs to a well-characterized bacterial transcription factor family which shows conservation of certain key structural characteristics (cAMP-binding and DNA-binding domains) (Körner et al., 2003; Matsui et al., 2013; Soberón-Chávez et al., 2017). Although this conservation leads members of the family to bind similar DNA sequences, promoter divergence causes their target gene repertoires to differ substantially between species, and thus also their physiological roles (Soberón-Chávez et al., 2017).

CRP provides an excellent model for our purposes because: a) it is well-characterised at the genetic, structural and physiological levels (Kolb et al., 1993); b) its activity can be easily experimentally modulated, by applying cAMP to the growth medium (Towbin et al., 2017); c) it is short (209 amino acids), allowing it to be characterised in its entirety; d) CRP activity-fitness functions have recently been characterised under a variety of conditions (Towbin et al., 2017); e) CRP mutants have been

documented to play a role in adaptation to new environments (Basak and Jiang, 2012; Gayán et al., 2017; Sievert et al., 2017).

2.1.3 Choice of experimental environments

In selecting conditions in which to which to perform fitness mapping, we reasoned that the external environment could affect the fitness effects of CRP mutations by two principle means. The first is to change the relationship between transcriptomic/proteomic state and fitness (for example, the fitness effect of expressing a particular CRP-regulated metabolic enzyme will clearly depend on the presence or absence of its substrate in the environment). The second is by changing the relationship between CRP activity changes and transcriptomic/proteomic state (either directly, by cAMP-mediated control of CRP activity, or indirectly, by control of the activity of other regulators with which CRP interacts). By definition, environments in which CRP is physiologically adaptive potentially affect both these relationships. Varying external cAMP concentration, however, should predominantly affect the latter relationship, by changing wildtype CRP activity (its only known effect in *E. coli*), allowing us to capture the effects of changing the latter relationship without changing the former. As we can only increase, but not decrease, native cAMP concentration experimentally, we required a growth source resulting in low endogenous cAMP levels. We thus chose glucose, which meets this requirement and is also the preferred sugar of *E. coli* and the majority of studied microorganisms (Bettenbrock et al., 2007; Görke

and Stülke, 2008). To examine the impact of an environment inducing a physiologically adaptive cAMP/CRP-mediated response, and thus likely affecting both relationships, we chose sodium chloride (NaCl), an experimentally robust molecule which causes hyperosmotic stress (Balsalobre et al., 2006). We thus chose to study the single-mutation fitness landscape of CRP in the four following controlled environments: glucose, glucose + cAMP, glucose + NaCl, and glucose + cAMP + NaCl.

2.2 Results

2.2.1 Optimisation of experimental conditions

Due to technical limitations, our mutant library was designed to be expressed from a multicopy plasmid rather than from its native position on the chromosome, and so we first constructed a set of three *E. coli* K12 MG1655 control strains with which to assess the impact of this and to optimise experimental conditions: one with no copies of *crp* (*crp*$^-$ [pBC-*crp*$^-$]), one with the native chromosomal copy of *crp* (*crp*$^+$ [pBC-*crp*$^-$]) and one with *crp* translocated from the chromosome to a multicopy plasmid (*crp*$^-$ [pBC-*crp*$^+$]). We then measured the growth of monocultures of these three strains in different test conditions. We found that 400mM NaCl, a typical concentration used in previous studies of the CRP-mediated osmotic stress response, did not allow detectable growth of the *crp*$^-$ [pBC-*crp*$^-$] strain under our conditions. This is in contrast to one of the previous studies (Balsalobre et al., 2006), perhaps because of differences in the

mineral media compositions. As we were interested in using a stress condition that still permitted growth of *crp*-null mutants, we reduced NaCl concentration to 100mM, finding this to reduce the growth rate of the *crp*$^-$ [pBC-*crp*$^-$] strain by ~25%, while causing no measurable growth defect in the crp$^+$ strains (Figure 2.1A). This condition therefore allows the impact of a stressful environment on the DFE to be assessed with the wildtype growth rate being kept relatively constant, and so was selected for the high-throughput competition assays. We also note that the *crp*-null strain grows slower than the *crp*-overexpressing strain both in the presence and absence of NaCl, demonstrating that overexpressed *crp*, although somewhat artificial, was still adaptive in these conditions.

For cAMP, we found that external cAMP concentrations within the range of those used in previous physiological studies (10 mM, (Bren et al., 2016)) abolished growth of the overexpressing *crp*$^-$ [pBC-*crp*$^+$] strain (but not the single-copy *crp*$^+$ [pBC-*crp*$^-$] strain). This demonstrates a growth hypersensitivity to cAMP caused by *crp* overexpression, perhaps aggravated by the native *crp* feedback loops (Kremling et al., 2007) which we chose to leave intact, highlighting that our experimental system can behave rather differently to the natural case. A low-concentration cAMP titration experiment was therefore performed on the *crp*$^-$ [pBC-*crp*$^+$] strain (Figure 2.1B), revealing a monotonic decrease of growth rate with increasing external cAMP concentration, and so again suggesting the presence of an optimum in the CRP

activity-fitness function (Towbin et al., 2017). We opted to use for the mutant competition assays the cAMP concentrations causing *crp*$^-$ [pBC-*crp*$^+$] growth to be most similar to the *crp*-null strain (which is insensitive to cAMP), to calibrate the system such that loss-of-function mutations would be expected to be beneficial, and gain-of-function mutations, deleterious: 0.7 mM in the absence of NaCl, and 1.2 mM in the presence of NaCl. Growth rates of *crp*$^-$ [pBC-*crp*$^+$], representing the experimental "wildtype" strain, and *crp*$^-$ [pBC-*crp*$^-$] are provided for convenience in Figure 2.1C for the 4 environments chosen for deep-mutational fitness scanning.

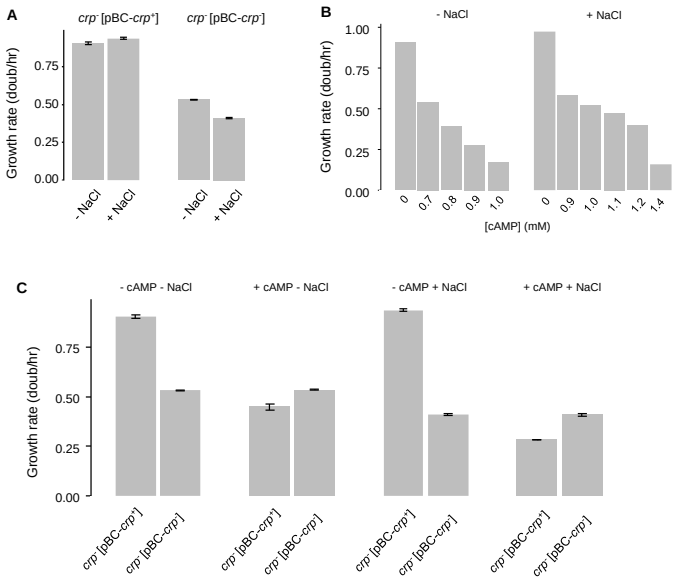


Figure 2.1: Choice of experimental conditions. A. The effect of 100 mM NaCl on growth rate of the strain harbouring *crp* on a plasmid (*crp*⁻ [pBC-*crp*⁺]) and the *crp*-null strain (*crp*⁻ [pBC-*crp*⁻]). Mean exponential growth rate is shown with the SEM of 2 independent replicates from different days. **B.** The effect of cAMP concentration on *crp*⁻ [pBC-*crp*⁺] exponential growth rate, with or without 100 mM NaCl. **C.** Growth rates of *crp*⁻ [pBC-*crp*⁺] and *crp*⁻ [pBC-*crp*⁻] in the 4 conditions chosen for deep-mutational fitness scanning. NaCl was added at 100 mM, and cAMP at 0.7 mM in the absence of NaCl and 1.2 mM in the presence of NaCl. Mean exponential growth rate is shown with the SEM of 2 independent replicates from different days.

2.2.2 Mutant library quality

To achieve a comprehensive full-length mutant *crp* library, a gene-tiling approach (Firnberg and Ostermeier, 2012; Kowalsky et al., 2015) was used in which 3 plasmid sub-libraries were created, each targeting a different third of the *crp* open reading frame (ORF) (see Methods). The sub-libraries were designed to together contain every possible single NNS codon substitution between the start and stop codons, with NNS accommodating all of the 20 amino acids as well as the amber stop codon. Carryover

of the wildtype template during mutagenesis ensures the wildtype *crp* DNA sequence is also present at high frequency (Firnberg and Ostermeier, 2012; Hietpas et al., 2011). Each plasmid sub-library was also intergenically tagged with unique DNA barcodes containing 20 randomised positions (Mavor et al., 2016; Sarkisyan et al., 2016). Once a first sequencing step was performed to associate these barcodes with their accompanying *crp* genotype, *crp* genotypes could be tracked through competition by simply sequencing the short barcode region. This strategy also means that internal replicates are present for every genotype.

The majority of expected *crp* DNA genotypes were present at a satisfactory sequencing coverage at t_0 of the competition assays (Figure 2.2A; median expected genotype counts = 734). At the barcode level, ~93% of all expected genotypes were associated to > 9 high-confidence barcodes (defined as barcodes present at > 5 read counts), with a median of 80 high-confidence barcodes per expected genotype, providing a large number of internal independent replicates (Figure 2.2B).

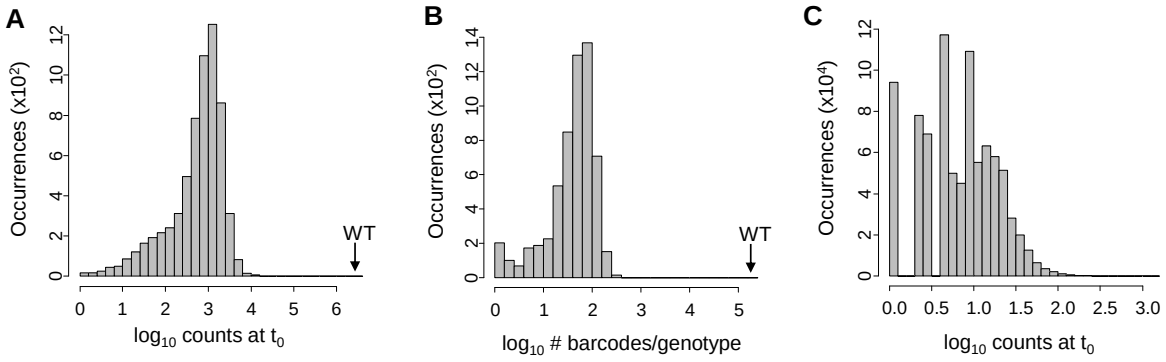


Figure 2.2. Sequencing coverage and quality of barcoded mutant library. **A.** The total read coverage of all expected *crp* DNA genotypes at t_0 of competition assays, computed by summing the counts of all barcodes associated to each genotype. WT is the wildtype DNA sequence. **B.** The number of unique barcodes (present at >5 reads at t_0 of competition assays) associated to each *crp* DNA genotype. **C.** The total read coverage of all unique barcodes at t_0 of competition assays.

The cost of this barcode richness, however, was a relatively low level of *barcode* sequencing coverage (Figure 2.2C; median of 12 reads/high-confidence barcode at t_0). As a consequence, barcodes were not used individually to compute independent fitness estimates for each genotype, but they did allow anomalous lineages associated to a particular genotype to be filtered out before aggregating the remaining barcode counts and computing a single genotype fitness estimate. This outlier removal step was found to be critical for accurate fitness estimation, as even very rare undetected off-target beneficial mutations (likely introduced during mutagenesis) can have enormous impacts on the apparent frequency change of unfit genotypes. It was particularly important for the wildtype genotype, not only because it serves as a reference, but also because it was so abundant (Figures 2A-B), causing undetected mutations in the

wildtype background to be numerous enough to potentially skew fitness estimates substantially (Figure 2.3). This problem was aggravated by our choice of certain environments in which undetected loss-of-function mutations, expected to be relatively common, were beneficial compared to the wildtype (Figure 2.3, right panel).

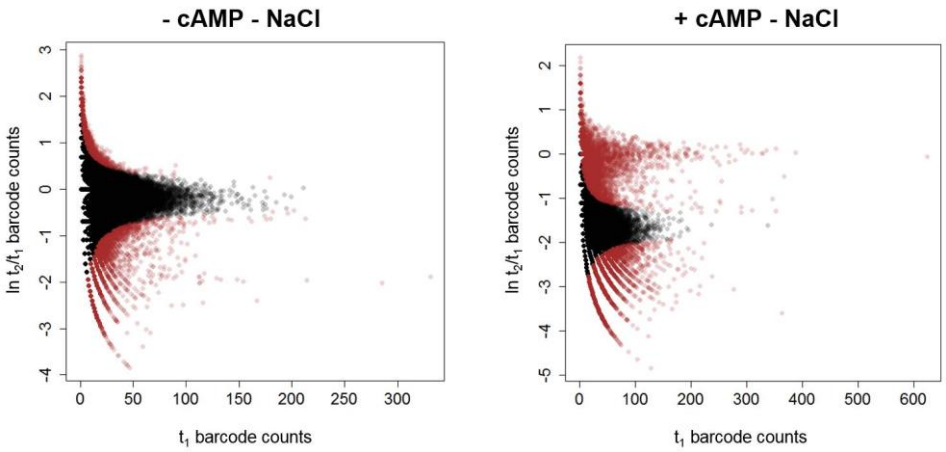


Figure 2.3. Anomalous barcode detection. The log ratio of t_2 to t_1 counts, against t_1 counts, is shown for all barcodes associated to the wildtype *crp* DNA genotype in 2 environments (n = 257,952 and 198,584, left-right). Outlier barcodes (brown) were detected with a 2-tailed Poisson test ($\log_{10}$ p-value < -10) (see Methods).

2.2.3 Fitness estimation and experimental noise

A key assumption in estimating fitness from competition data is that selection remained constant throughout the experiment. We found that this was not the case for the 2 cAMP-containing environments (Figure 2.4), and so to reduce any confounding effects to a minimum we chose to quantify fitness based on just 2 adjacent time-points: t_1 and t_2, with t_0-t_1 being left out to allow time for physiological acclimatisation to the competition media (see Methods). The experimental noise in these fitness estimates

was assessed by re-computing them using 2 fully independent sets of barcodes (Figure 2.5), and we found that it was far lower for the 2 environments containing cAMP than for the 2 without it (Pearson's r = 0.96 *vs.* 0.71 and 0.78), due presumably to the very different DFEs: cAMP results in a large number of mutations being highly beneficial, reducing counting noise (the wildtype has a below-average fitness, causing it to reduce in frequency over time (Figure 2.4A), which could counteract this noise-reduction, but this does not occur as it is present in the initial library at extremely high frequency (Figure 2.2A)). A further consequence is a much larger range of fitness effects, also improving the correlation.

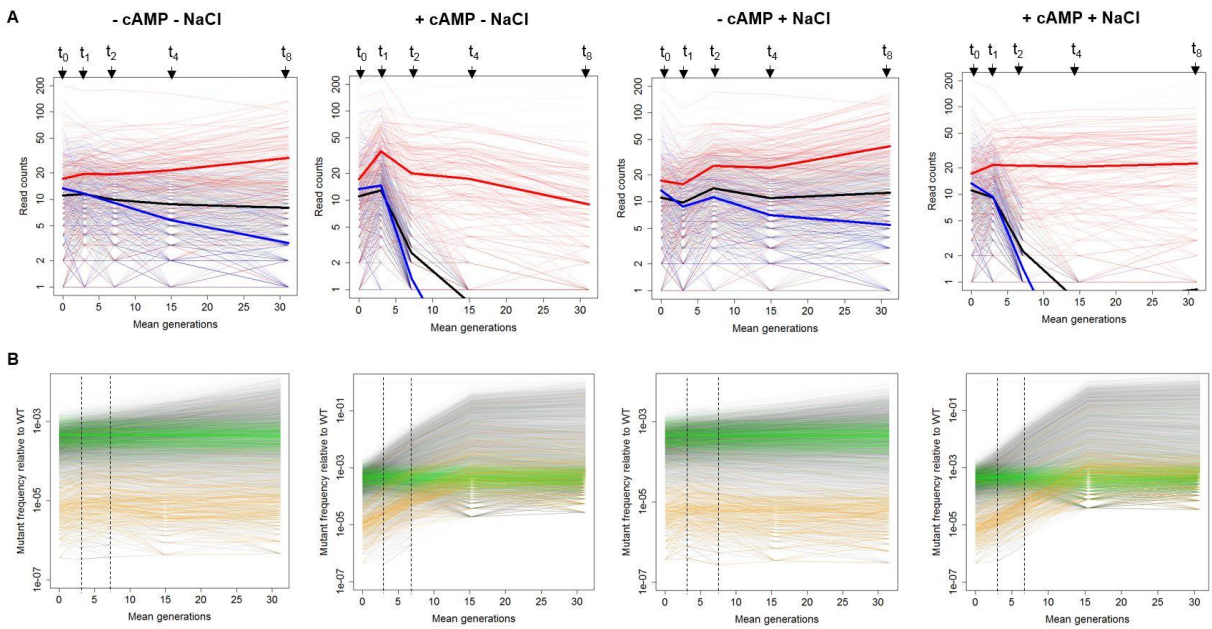


Figure 2.4. Barcode and mutant dynamics during competitive growth. A. Example trajectories are shown for a random sample of 1,000 barcodes associated to the wildtype (black), and all barcodes associated to 2 *crp* codon mutants (red and blue). Thick lines are the mean number of barcode counts. **B.** Barcode-grouped trajectories are shown for all mutant *crp* DNA genotypes relative to the wildtype. Synonymous genotypes are coloured green, and genotypes containing a stop codon within the first 150 codons are coloured orange. After barcode outlier filtering, read counts for all barcodes belonging to a particular mutant were summed and normalized to total WT *crp* DNA read counts. Dashed lines indicate time-window chosen for fitness estimation.

We also found that stop codons present in the first ~3/4 of *crp*, representing expected *crp* null-mutants, were associated with an especially high degree of measurement uncertainty (Figure 2.5), even when they increased in frequency (Figure 2.4B, panels 2 and 4), clearly due to their low starting abundance (Figure 2.4B). This problem arises because, due to *crp*'s highly pleiotropic nature, there exist no truly permissive media for library cloning and outgrowth, and so selection had acted on the mutant library

prior to the competition assays. In the rich medium (Lysogeny Broth) we used, null mutations were deleterious and so, although growth prior to competition was kept to a minimum, they were depleted from the library. This pre-assay selection problem is aggravated by the fact that, as we have shown, barcode-mutation assignment carries some uncertainty. The result is that a fraction of barcodes assigned to any one mutant may be misassigned, with some of these erroneous lineages being fitter than the correctly assigned ones. During the unwanted pre-assay selection, these misassigned lineages would rise in frequency relative to the correctly assigned ones, without detection, and by the time of assay they could represent the majority of barcodes if their fitness relative to the correct barcodes was sufficiently high. This explains how some early stop codon genotypes can be estimated as beneficial with rather high confidence in the conditions (- cAMP) where they are expected to be deleterious, for example. Overall, these confounding factors strongly limit our ability to quantify the effects of mutations that were highly deleterious in the pre-competition medium (comprising, at least, null-like mutants), but we can see from Figure 2.4B that they represent a minority of the library.

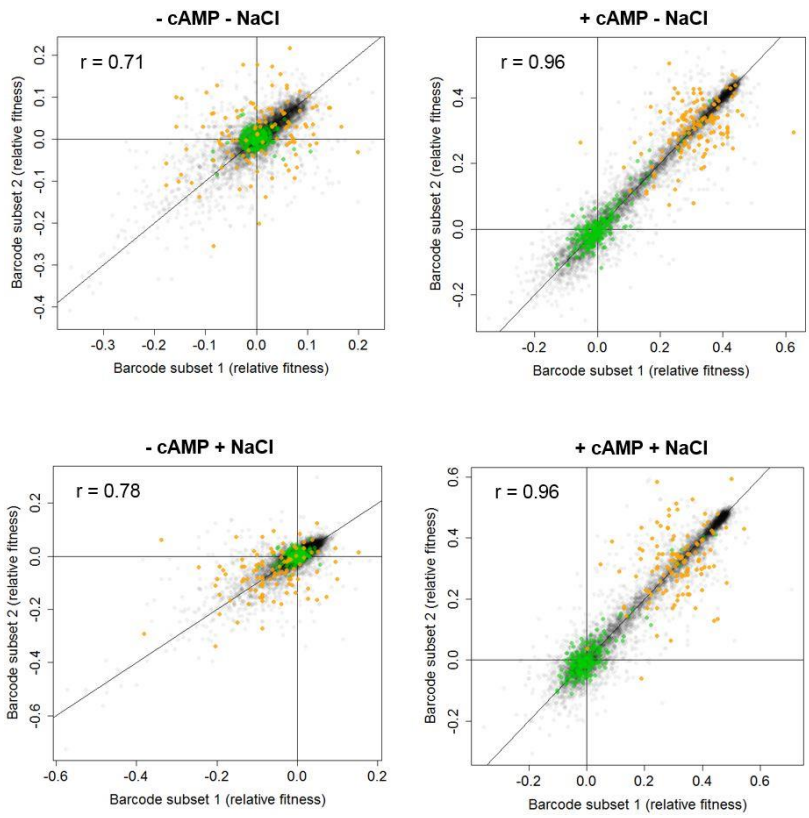


Figure 2.5. Experimental noise characterised by independent barcode sets. After barcode outlier filtering, barcodes associated to each *crp* mutant DNA genotype and the wildtype DNA genotype were randomly split into 2 equal-sized subsets. For each subset, log fitness relative to the wildtype was computed using the sum of mutant and wildtype barcode subset counts (see Methods). The Pearson correlation coefficient for the 2 resulting sets of independent fitness estimates is indicated. Synonymous genotypes are coloured green, and genotypes containing a stop codon within the first 150 codons are coloured orange.

2.2.4 Distribution of fitness effects (DFE)

With an awareness of the distribution of experimental noise in our fitness estimates, we next assessed the DFE across environments at the level of protein sequence, by pooling all barcodes associated to a particular amino acid change. Fitness estimates were obtained for 89-92% of all 4,180 possible single amino acid substitutions

(including to the amber stop codon), with 2 positions being essentially uncovered (0 or 1 of the 20 possible substitutions quantified, likely due to missing degenerate primers in the mutagenesis step). In line with the monoculture growth measurements, the DFE was substantially transformed by the presence of cAMP, and far less so by NaCl (Figure 2.6). As seen already, the DFE of early stop codons is far broader than would be expected (as they should all be similarly null-like), due to increased counting noise and the potential takeover of misassigned barcodes, but they can still provide an indication of where null-like mutants are expected to lie in each distribution (except in the noisiest case of the –cAMP –NaCl environment). The DFE of synonymous mutations provides another biological control of fitness estimates, as the majority are expected to be very nearly neutral, and we find them to indeed be centred close to zero and far narrower than the corresponding DFEs of amino acid substitutions.

The extensive overlap between the DFEs of synonymous mutations and early stop codons in the no-cAMP environments suggests, however, that the level of experimental noise for low-fitness mutants is too high to permit reliable interpretation of the overall DFE in these conditions. The change in selection pressure caused by cAMP alleviates this problem, in part by making null-like mutants beneficial (see above), and results in a clear separation between the synonymous and early stop codon DFEs. Interestingly, it also results in the appearance of bimodality in the DFE of single amino acid substitutions, a very common observation for experimental DFEs in general (Bank et

al., 2014, 2015; Bernet and Elena, 2015; Chan et al., 2017; Diss and Lehner, 2018; Eyre-Walker and Keightley, 2007; Firnberg et al., 2014; Hietpas et al., 2011; Jacquier et al., 2013; Jiang et al., 2013, 2016; Klesmith et al., 2015; Li et al., 2016; Mavor et al., 2016; Melamed et al., 2013; Melnikov et al., 2014; Puchta et al., 2016; Roscoe et al., 2013; Wrenbeck et al., 2017) . Typically, the modes are centred close to the wildtype and close to null-mutants, with null-mutations being lethal or very deleterious. In this system, a nearly-neutral mode is indeed apparent, and the other mode lies fairly close to that of the (beneficial) null-mutants, but is clearly shifted to the right. The activity of this large set of mutants most likely lies somewhere between that of the wildtype and the null-mutants, both of which are less fit than them, and their location at the extreme-right of the DFE (Bataillon and Bailey, 2014) in turn suggests that this intermediate activity represents a rather broad fitness optimum, perhaps resulting from the negative feedback in the CRP-cAMP signalling network (You et al., 2013). In more natural conditions, when the wildtype lies on this plateau, this would provide a strong source of genetic robustness, buffering fitness against activity-changing mutations (Denby et al., 2012; Marciano et al., 2014, 2016).

mutations (Denby et al., 2012; Marciano et al., 2014, 2016).

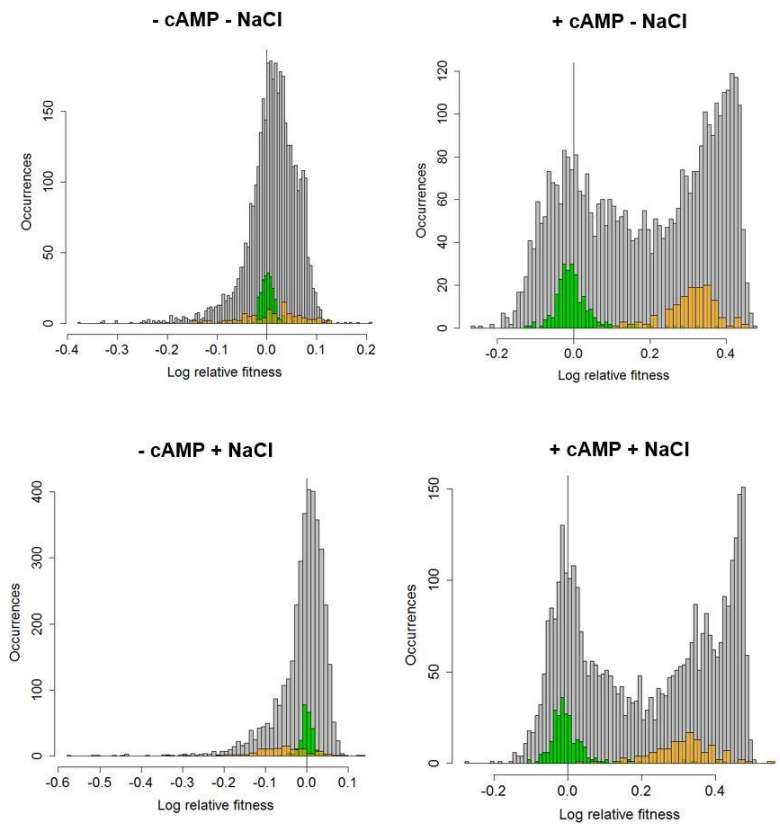


Figure 2.6. Distribution of fitness effects (DFE) of single amino acid substitutions in CRP. The DFE of synonymous mutations is shown in green, and the DFE of amber stop codons (in the first 150 positions) in orange. In all cases, log fitness is estimated relative to the wildtype *crp* DNA sequence.

2.2.5 Fitness correlations between environments

The correlation between fitness effects across environments has important implications for evolution and the predictability of mutational effects in general, and it may also provide hints about mechanism and underlying phenotypes. We found that there was a significantly positive correlation across all environments, and, as suggested by the DFEs, the correlation was far stronger (and close to the identity line) between conditions differing by just NaCl presence rather than cAMP presence (Figure 2.7, top

panels *vs.* bottom panels). In all cases, however, a strong, smooth trend was apparent, suggesting the correlations are shaped predominantly by some relatively simple global phenotypic mechanism rather than by idiosyncratic mutation-environment interactions. In the environments differing by cAMP presence, this trend, although noisy, was clearly non-monotonic, remarkably reminiscent of those in the *ara* system of Chapter 4, which were indicative of overshoots of a phenotypic optimum in some states or environments but not in others (Figure 2.7, bottom panels; Figure 2.8A). In this case, the pattern can be explained in terms of the discrimination of mutants who, in one environment, have similar fitness but lie on 2 sides of a phenotypic (*ie.* CRP activity) optimum (here, this seems to apply to the case of early stop codon mutants and a set of more weakly affected mutants) (Figure 2.8C). In a 2nd environment, however, their respective positions relative to the optimum are shifted, and large differences in their fitness become apparent (Figure 2.8B). Another result of such environmental shifting of activity with respect to an optimum is that mutations can switch between being beneficial and deleterious in different environments (Figure 2.7, upper-left quartile of bottom panels; Figure 2.8B-C). The presence of a phenotypic optimum is in line with known features of the CRP activity-fitness relationship (Towbin et al., 2017) and the conclusions made from the DFE shapes. Finally, the extensive breadth of the inferred CRP activity optimum is again hinted at by the density of mutants lying at the top-right "corner" of the non-monotonic correlation trends: although cAMP induces large changes in CRP activity, a substantial

proportion of mutants find their way close to the apparent maximal fitness when it is both present and absent.

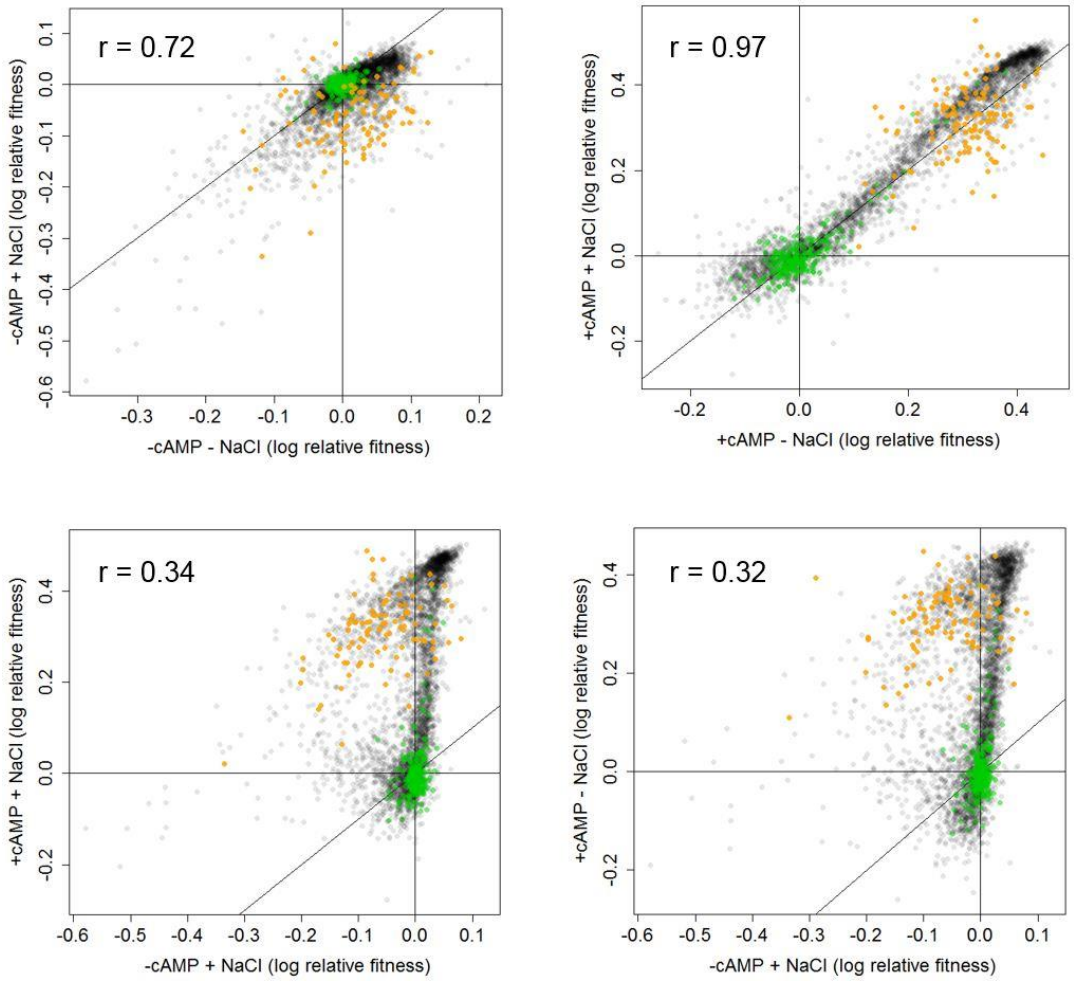


Figure 2.7. Correlations of fitness effects between environments. Synonymous mutations are shown in green, and mutants containing an early stop codon (first 150 positions) in orange. Pearson's correlation coefficients are provided.

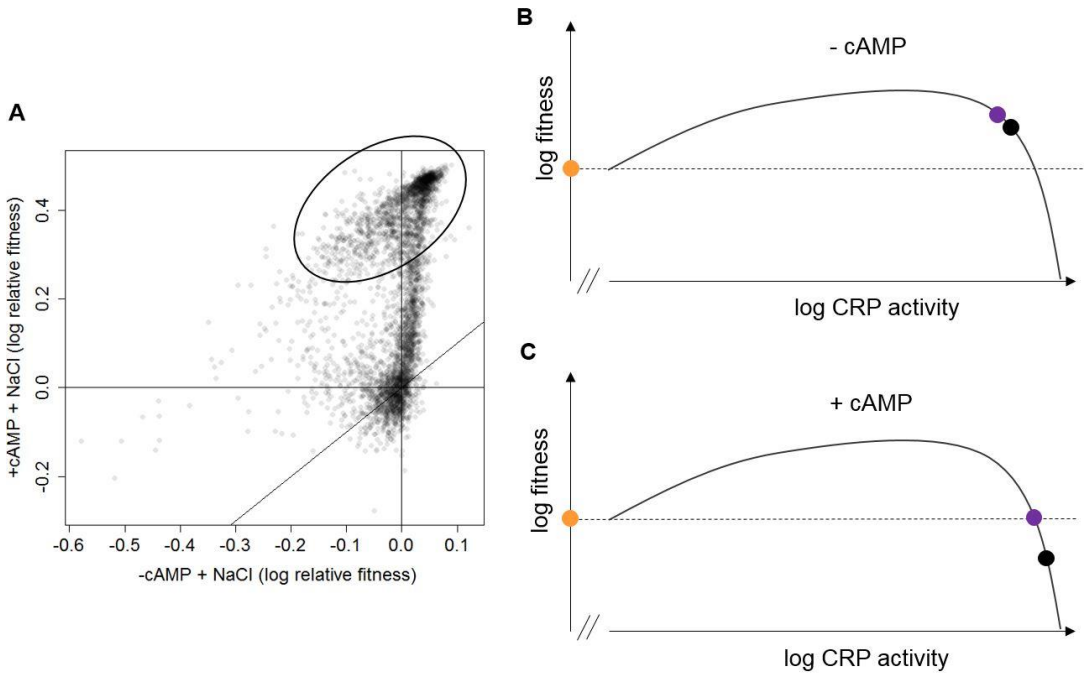


Figure 2.8. Non-monotonous fitness effect correlation and optimum overshooting. A. The same plot as that in Figure 2.7, bottom left panel, but with colouring removed for clarity. The region of apparent non-monotony is circled. **B.** A hypothetical CRP activity-fitness function is shown, along with the position of 3 hypothetical genotypes, in the absence of cAMP: wildtype (black), a null-mutant (orange) and a weakly-deactivated mutant (purple). Dashed line shows log fitness in the absence of CRP activity. Note that the wildtype has a slightly above-optimal CRP activity (in line with its being carried on a multicopy plasmid in our experimental system). The weakly deactivating mutation is then slightly beneficial, and the null-mutation is deleterious. **C**. As for **B**, but in the presence of cAMP, which increases the activity of the non-null-mutants (null-mutants are completely infunctional, and so do not respond to cAMP). The result is that both mutations now become similarly beneficial with respect to the wildtype, which causes non-monotonicity and sign-changes in the correlation between their fitness effects in the 2 environments.

2.2.6 Sequence-fitness maps

Comprehensive sequence-fitness maps provide an unprecedented resource both for understanding protein evolution and for detailed structural and mechanistic insights. It

is immediately apparent from ours and others', for example, that positions can vary greatly in their overall sensitivity to mutation, and some have more specific constraints on residue physico-chemistry than others. They may also point to currently unexplored mutations, positions or whole regions warranting functional characterisation. Unfortunately, we have not yet had the opportunity to analyse these maps in any detail, but they are provided here as a community resource (Figure 2.9). We also note that, when viewed as a classical functional screen (with function here being the highly integrated trait of fitness), the use of artificial environments designed to transform the direction and magnitude of fitness effects can be extremely useful: the presence of activity-increasing cAMP clearly amplifies the signal in many cases.

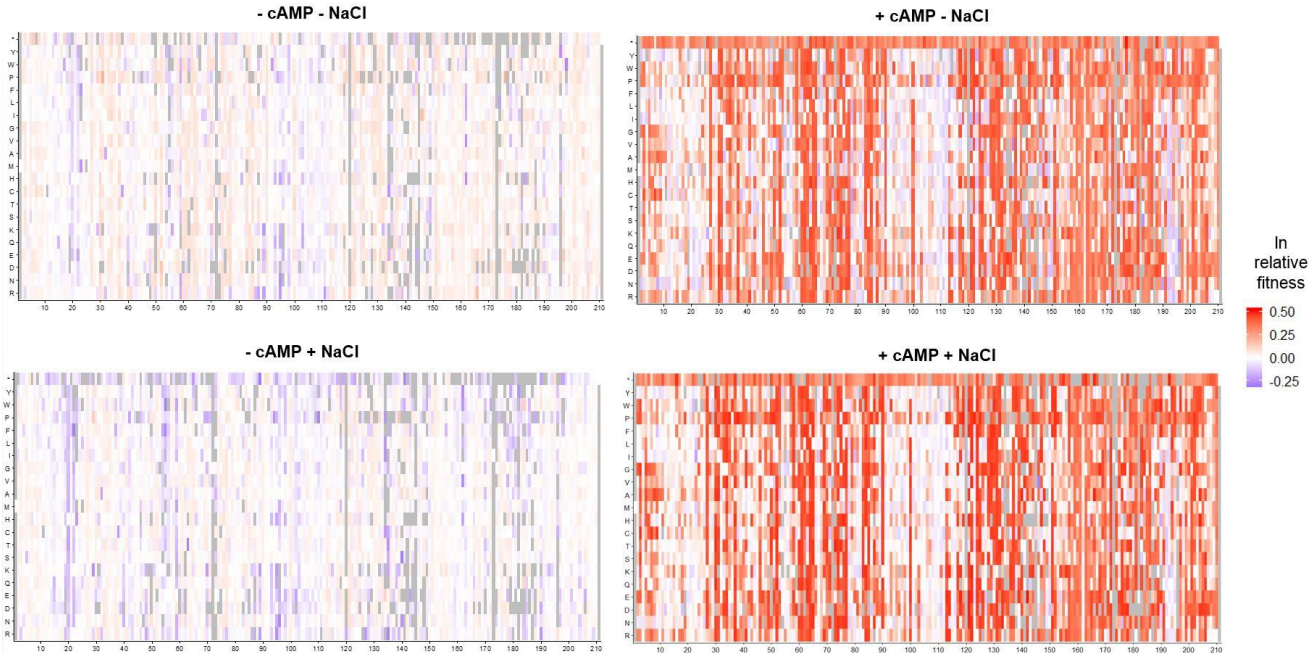


Figure 2.9. Sequence-fitness maps for single amino acid substitutions across the entire length of CRP, in 4 environments. Residues on the y-axes are ordered according to chemical similarity, and the amber stop codon is at the top. Red indicates beneficial mutations, white, neutral, and blue, deleterious. Proline (P) is known to be particularly disruptive at many sites and so can serve as a biological control of data quality. We find that it stands out clearly, as expected (right panels, 4th row from top).

2.3 Discussion

2.3.1 Technical considerations

This (unfinished) research project aimed to comprehensively characterise the single amino acid substitution fitness landscape of a bacterial global transcriptional regulator, CRP. Two experimental issues warranting discussion were encountered along the way. The first was experimental realism: monoculture growth measurements of control strains revealed that, in certain environments, our multicopy genetic system could

behave very differently to the natural single-copy one, with the multicopy wildtype even being inviable under cAMP concentrations permitting growth of the single-copy wildtype. One way to alleviate this could be to use a weaker non-native promoter that matches multicopy expression to the native single-copy expression (Hietpas et al., 2011). The complex native *crp* promoter region is an integral part of its function, however, embedding it within the context of a large regulatory network, and so such a system would likely lack key features of the natural one. More precise genetic manipulations of RNA polymerase- or ribosome-binding sites may permit comparable expression levels between the multicopy and single-copy system while leaving the regulatory network intact, but this would be a substantial endeavour, with no guarantees of true equivalence between the experimental and native systems. The ideal solution would therefore be mutagenesis of *crp* at its native chromosomal locus (Li et al., 2016), a difficult task, but one that is becoming more feasible with advances in genome engineering (Haimovich et al., 2015). The discrepancy between our experimental system and the natural one means that the reported mutational fitness effects can of course not be interpreted as a direct estimation of fitness effects in the natural system, but they can still allow highly relevant evolutionary and functional inferences to be made.

The second experimental issue was the high level of uncertainty in fitness estimates for null-like mutants. This is a common problem in large-scale sequencing-based

competition assays, as their rapid depletion from the population during competition results in low read counts, and is often partially dealt with by assigning a lower bound representing the expected fitness of null-mutants. In our case, the problem was aggravated by low read counts *before* competition, as well as the possibility that fitter, misassigned barcode lineages had begun to takeover, as the highly pleiotropic nature of CRP meant that no non-selective media could be found for library cloning and outgrowth. This is ultimately a less extreme version of the technical challenge associated with characterising essential genes, for which, by definition, permissive growth conditions do not exist. A workaround for these is to maintain the mutated copy of the gene alongside a functional copy, which can then be "shutoff" or removed immediately prior to competition (Hietpas et al., 2011). The engineering of such a system is, again, technically difficult, and the possibility of unwanted residual wildtype activity remaining during competition is a serious concern, but such a strategy is likely the most economical one. In terms of the results presented here, the noise is far less problematic for the cAMP-containing environments (in which null-mutations are beneficial), and, in general, conclusions can still be reached with an awareness of it.

2.3.2 Underlying phenotype-fitness landscape

Both the DFEs and the correlations of fitness effects between environments allow inferences to be made regarding an underlying phenotypic dimension. First, the 2 more reliable amino acid substitution DFEs (those in the cAMP-containing environments)

show a mode shifted to the right of the DFEs of both synonymous mutations and early stop codons. The order, by increasing fitness is: synonymous, early stop codons, amino acid substitutions, implying that complete loss-of-function increases fitness, but another class of mutations increases fitness even more. As loss-of-function mutations form a zero-activity bound, this implies non-monotonicity in the CRP activity-fitness function. One possibility is that the wildtype activity lies in a fitness valley, and these mutations increase activity and thus fitness. But from our knowledge of CRP's biology and the apparently universal predominance of loss-of-function mutations over gain-of-function ones, a far more reasonable hypothesis is that these mutations decrease activity to an intermediate level between the wildtype and null-mutants, and this level lies at an optimum in the activity-fitness function (see Figure 2.8B-C). The large amplitude of this high-fitness mode and its truncation-like right tail then suggest that this apparently maximal fitness can be reached by a substantial proportion of mutants, implying a very broad fitness plateau (*ie.* a wide range of CRP activities result in near-maximal fitness). Second, the correlations between fitness effects in environments differing by cAMP presence show a smooth non-monotonous trend, with a group of mutations switching between being beneficial and deleterious in the 2 environments. This is a strong sign of the fact that changing environments shifts the position of genotypes in the activity-fitness landscape relative to an optimum (Figure 2.8B-C; see Chapter 4), as would be entirely expected from the known role of cAMP: it is a highly specific activator of CRP, and so should change its activity while having minimal

impact on the activity-fitness function itself. Further, the high density of points in the top-right corner of these correlations again points to the breadth of the inferred fitness plateau: a large number of mutants lie on it in both the presence and absence of cAMP, even though it is expected to shift their activity considerably.

The next step of this project will be to test these hypotheses using quantitative models. Existing knowledge of the CRP activity-fitness landscape comes from experiments disrupting one of its native feedback loops and controlling activity *via* external cAMP (Towbin et al., 2017). These experiments indeed reveal a function with a single optimum, although not a particularly broad one. The broad plateau inferred from our results then likely relies on CRP being embedded in its native regulatory context, and on the fact that mutations can affect CRP activity/cAMP, rather than simply total CRP activity (You et al., 2013). This raises the tantalising possibility that global transcriptional regulators, due to their being embedded within networks typically possessing feedback, may in general be endowed with broad plateaus in their activity-fitness landscapes (potentially selected for to buffer against small physiological fluctuations in total activity) (Denby et al., 2012; Marciano et al., 2014, 2016). Such plateaus could then make them particularly robust to activity-changing mutations, in turn increasing cryptic genetic variation and the potential for innovation (Wagner, 2012).

More generally, the strong, smooth trends in the correlations between environments reiterate the fact that simple global phenotypic dimensions may often explain many of the observed interactions between mutations and environment (Hietpas et al., 2013; Li and Zhang, 2018), as well as epistatic interactions (Bank et al., 2015; Diss and Lehner, 2018; Sarkisyan et al., 2016), providing a very promising route to the prediction of mutational fitness effects.

2.3.3 Gain-of-function mutations

As explained in the introduction, the cAMP-containing environments were designed to result in gain-of-function mutations, but not loss-of-function ones, having a deleterious fitness effect (Figure 2.8C). Our results indicate that such mutations may constitute a significant fraction of the complete set, but we have not yet had the opportunity to quantify this. To accomplish this, bootstrapping of barcodes will be performed to obtain confidence intervals on the fitness estimates, allowing the rigorous classification of significantly deleterious mutations. Such a classification would provide a lower bound on the number of mutations that have an effect other than simply reduction of wildtype CRP activity (a lower bound because these mutations could also confer a net fitness benefit, although this would presumably be rarer). Mechanistically, these mutations could act by increasing the existing activity of CRP, as for the well-known mutations making it constitutively active in the absence of cAMP (Youn et al., 2006). Alternatively, they could create a novel (deleterious) function. Novel functionality

affecting fitness may not be as rare as one would think, due to CRP's multiple intermolecular interfaces (Kolb et al., 1993): for example, mutations affecting RNA polymerase-contacting regions but not the DNA-binding domain could result in CRP blocking transcription where it used to activate it (similarly to the novel function of dCas9 (Dominguez et al., 2016)), or sequestering RNA polymerase. Mutations could also alter CRP's DNA binding repertoire (Shultzaberger et al., 2012). Further, although not "gain-of-function" mutations, those causing more general stresses such as toxic misfolding would also be included in this category (Bednarska et al., 2013). Ultimately, such an analysis would provide an assessment of the proportion of mutations that either increase wildtype activity or are not well described by a simple, one-dimensional phenotype-fitness function (*ie.* CRP activity), both of which bear important implications for evolution and the predictability of mutation effects.

2.3.4 Sequence-fitness maps as a functional resource

The full-length sequence-fitness maps provided here should provide valuable information for protein biochemists. It is complementary to maps based on *in vitro* assays (Fowler and Fields, 2014), pointing to variants and regions relevant for fitness. Comparisons between such different types of maps could also be extremely insightful, revealing how the enormously complex *in vivo* context of a global regulator and its relationship to fitness transforms conclusions based on *in vitro* functional biochemistry.

2.4 Methods

General microbiology and molecular biology

Lysogeny Broth (LB) powder, agar, salts, sugars, growth supplements and antibiotics were all purchased from Sigma-Aldrich. Bacteria were cultured in LB, unless otherwise stated. M9 base medium consisted of 1X M9 salts supplemented with 1mM MgSO4 and 100 µM CaCl2. Glucose was used at a concentration of 0.4% w/v throughout. Ampicillin (amp) was used at 100 µg/ml, chloramphenicol (cm) at 10 µg/ml and streptomycin (str) at 50 µg/ml. Bacterial cultures were grown at 37°C (with shaking at 200 rpm for liquid cultures; Multitron, Infors HT), unless otherwise stated, and culture stocks were stored at -80°C in LB with 40% glycerol. For electroporation, DNA was added to 50 µl homemade electro-competent cells (unless otherwise stated), transferred to a 1mm-gap electroporation cuvette (VWR) and submitted to a pulse of 1,800 V (Electroporator 2510, Eppendorf). Cells were immediately transferred to fresh LB for recovery at 37°C (unless otherwise stated) with shaking for 30-90 minutes, before being plated on the appropriate selective media and left to grow overnight.

All enzymes and molecular biology reagents were purchased from NEB, unless otherwise stated. Primers were purchased from IDT or Eurofins, and designed with the help of Primer3 (Rozen and Skaletsky, 2000). For sensitive applications like barcoding and NGS library preparation, primers were ordered HPLC-purified, otherwise they

were ordered desalted. UltraPure agarose was supplied by Invitrogen, and all agarose gels were stained with SYBR Safe (Thermo Scientific) and visualised with a GelDoc XR+ imager (Bio-Rad). The GeneRuler 1kb Plus ladder (Thermo Scientific) was used for DNA fragment size estimation.

All plasmids used in this study, excluding the mutant library, are detailed in Table 2.S1. DNA fragments used in cloning are detailed in Table 2.S2. Primers, excluding those used for library mutagenesis, are provided in Table 2.S3.

Plasmid construction

Two plasmids, pBC -*crp*$^+$ and pBC -*crp*$^-$, were constructed for the initial optimisation of experimental conditions, with an intermediate plasmid, p-*crp*$^+$, also serving as the template for library mutagenesis. Plasmid pBC -*crp*$^+$ is derived from pSkunk3-BLA (Firnberg and Ostermeier, 2012), a low/medium-copy phagemid containing a bacterial *p15A ori* and a phage *f1 ori* (necessary for the library mutagenesis step), and was constructed by replacing the *bla* gene cassette used in the original study with a *crp* cassette. The *crp* cassette was based on that from (Zhang et al., 2012), containing the complete native *crp* promoter region (including two CRP-cAMP binding sites, four Fis binding sites and three annotated *crp* promoters), the ORF (with 20bp of its upstream ribosome binding site-containing region and 16bp of its downstream region) and an ectopic *rrnB* T1 terminator for strong transcriptional termination, with these 3 components being separated by restriction sites. The original *aadA1* Str/Sp-resistance

cassette from pSkunk3-BLA was then swapped with a *bla* β-lactamase cassette attached to a short random DNA barcode, to be comparable with the mutant library design (see *Library Creation*). Plasmid pBC-*crp*$^-$ was derived from pBC-*crp*$^+$ by excision of the entire *crp* cassette followed by self-ligation of the plasmid backbone. Plasmids, DNA fragments and PCR primers used in the construction of these plasmids are detailed in Tables S1-3, respectively, and the detailed cloning methods follow.

The DNA fragments used to construct plasmids p-*crp*$^+$, pBC-*crp*$^+$ and pBC-*crp*$^-$ come from either PCR amplification or from direct restriction digestion of purified plasmid DNA, and were joined by either standard restriction-ligation or by Gibson Assembly (Gibson et al., 2009) (in which case, overlaps of ~40 nucleotides were used). PCR amplifications were all performed with Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific) in its High-Fidelity buffer, following the manufacturer's recommendations. Restriction enzymes were used according to the manufacturer's instructions. After PCR amplification and/or digestion, DNA fragments were either verified by electrophoresis and column-purified (QIAquick PCR Purification Kit, QIAGEN) or, when necessary, gel purified (QIAquick Gel Extraction Kit, Qiagen). Gel-purification was always followed by a 2nd clean-up (QIAquick PCR Purification Kit, QIAGEN) to improve DNA quality for ligation. When DNA termini blunting was necessary, mung bean nuclease (NEB) was applied before gel extraction, following the manufacturer's recommendations. For gel extractions, agarose gels were

stained with SYBR Safe (Thermo Scientific), and DNA was visualised with blue light to avoid UV-induced DNA damage (Blue Transilluminator, Pearl Biotech). A NanoDrop ND-1000 spectrophotometer (Thermo Scientific) was used to determine DNA concentration for all fragments prior to ligation/Gibson Assembly. Standard ligation and Gibson Assembly were performed using T4 ligase and Gibson Assembly Master Mix (NEB), respectively, according to the manufacturer's recommendations (T4 ligase was then inactivated by heating at 65°C for 10 mins). In both cases, DNA was subsequently microdialysed against water for > 30 mins (MF-Millipore, Merck), and 1-5 µl were electroporated into 50 µl electrocompetent cells. *DH5α ΔaraBA* was used as the cloning strain in all cases. After electroporation, cells were recovered in 1 ml LB for 30-90 mins at 37°C with shaking at 200 rpm, plated on LB-agar in the presence of the antibiotic indicated in Table 2.S1 and incubated overnight at 37°C. Plasmid DNA was purified from several colonies (Plasmid Mini Kit, QIAGEN) and verified by both restriction analysis and Sanger sequencing of the ligated region.

Strain engineering

The plasmid host strain for both monoculture growth measurements and mutant library competition assays was *E. coli K12 MG1655 Δcrp*. The *crp* gene deletion was performed in *E. coli K12 MG1655* (A. Couce; Coli Genetic Stock Centre #6300) using the standard method of Datsenko and Wanner (Datsenko and Wanner, 2000). *E. coli K12 MG1655* was made electrocompetent, electroporated with 10 ng plasmid pKD46

DNA, and transformants were selected on LB-agar with 100 µg/ml ampicillin at 30°C. Several colonies were then re-isolated under the same conditions. The *cat* chloramphenicol-resistance cassette was PCR-amplified from pKD3 (Datsenko and Wanner, 2000) using primer pair KO-*crp*-fwd and KO-*crp*-rev and a 2:1 mix of GoTaq/Pfu DNA polymerases (Promega). PCR product was verified by 1% agarose gel electrophoresis, column-purified (QIAquick PCR Purification Kit, QIAGEN) and spectrophotometrically quantified (NanoDrop ND-1000). A pre-culture of a single pKD46-transformed colony was grown overnight (LB-amp) at 30°C and then diluted 100x into LB-amp with 0.2% L-arabinose and grown at 30°C to an OD600nm of ~0.7 (BioMate 3S, Thermo Scientific; 3-5 hours). The culture was made electrocompetent, electroporated with ~200 ng of the purified PCR product, and recombinants were selected on LB-agar with 10 µg/ml chloramphenicol at 37°C, for curing of pKD46. Several colonies were then re-isolated under the same conditions, and tested in parallel for pKD46 curing by plating on LB-amp and checking for colonies after an overnight growth at 30°C. Several of the re-isolated colonies were verified by colony-PCR, using 3 primer pairs as in (Datsenko and Wanner, 2000). The gene-specific primers were verif-*crp*-fwd and verif-*crp*-rev, and the common *cat* primers were *c1* and *c2* from (Datsenko and Wanner, 2000). The 3 primer pairs were thus: verif-*crp*-fwd/verif-*crp*-rev, verif-*crp*-fwd/c1 and verif-*crp*-rev/c2. GoTaq DNA polymerase (Promega) was used for amplification, following the manufacturer's recommendations, and PCR products were analysed by agarose gel electrophoresis (1.5%). The *cat* cassette was

then removed as described in (Datsenko and Wanner, 2000). For this, a pre-culture of a single recombineered colony was grown overnight (LB-cm, 37°C) and then diluted 100x into LB-cm and grown at 37°C to an OD600nm of ~0.7 (BioMate 3S, Thermo Scientific; 2-4 hours). The culture was made electrocompetent, electroporated with 10 ng plasmid pCP20 DNA, and transformants were selected on LB-agar with 100 µg/ml ampicillin at 30°C. Several colonies were then re-isolated under the same conditions, and then again in the absence of ampicillin at 42°C, to cure pCP20. Finally, several colonies were streaked in parallel on LB (37°C, purification), LB-cm (37°C, verify *cat* loss) and LB-amp (30°C, verify pCP20 loss). The loss of the *cat* cassette through FRT recombination was verified molecularly for several clones by colony-PCR, using the same primer pairs and conditions described above for *cat* insertion verification. The PCR product resulting from amplification with primer pair verif-*crp*-fwd/verif-*crp*-rev was also Sanger-sequenced (GATC; using the amplification primers) as a final verification.

Monoculture Growth Measurements

Optical density was monitored using a home-made turbidometer which allows the quasi-continuous, parallel turbidimetric measurement of cultures growing in standard glass culture tubes (5ml cultures in 15ml tubes) in a standard shaking incubator. Each tube has its own light source (~600 nm LED) and phototransistor, and measurements are recorded every 10 seconds for a range of emitted light intensities. Different light

intensities provide optimal results for different cell density ranges, and so appropriate intensities can be selected after data collection depending on the growth phase of interest (lower intensity for lower cell density, and visa versa). The apparatus offers several advantages over turbidometric microplate readers, the current method of choice for high-throughput growth measurements of microorganisms in liquid culture. First, the two major problems of growth in microplates, sample evaporation and low aeration (Hermann et al., 2003), are avoided due to the use of standard laboratory culture volumes. Aeration is further improved as the setup allows OD to be measured during continuous agitation, whereas microplate readers require shaking to be stopped for each reading. The standard laboratory conditions used in our system thus allow for better cell growth than do microplates, resulting in a more stable exponential phase and so enabling classical exponential growth rate estimates, rather than the alternative, inherently less stable metric of maximum growth rate. Second, the ability of our apparatus to measure OD during continuous agitation allows measurements to be taken at extremely high frequency ("quasi-continuously"), allowing easy noise-filtering when necessary and increasing the confidence of growth parameter estimates. Finally, due to the longer path-length in our set-up and the use of a range of light intensities, it has a lower detection threshold and so allows growth to be accurately observed at lower cell densities. This permits growth rates to be estimated from earlier in the exponential phase, helping to avoid the many pitfalls of using turbidity to

estimate cell number at high cell density/after long growth times (Stevenson et al., 2016).

For these growth assays, 5 ml M9 + 0.4% glucose + 100 µg/ml ampicillin was inoculated with the appropriate strain in a 50 ml Falcon tube and grown overnight at 37°C, with shaking at 200 rpm. Overnight cultures were diluted 1,000x into 5 ml of the appropriate medium (same as the pre-culture medium, with or without the indicated concentrations of NaCl and cAMP) in 15 ml glass culture tubes, and growth was monitored with the instrument described above at 37°C and 200 rpm shaking. Tube positions were randomised for each trial in case of any subtle positional effects. Growth curves were processed using a home-made R (version 3.4.3) script, and exponential growth rates were estimated using the lm() function, taking a universal (low) OD window ($0.02 < OD_{LED} < 0.14$).

Library Creation

The initial *crp* mutant library was constructed by the PFunkel method (Firnberg and Ostermeier, 2012) using a gene-tiling approach (Firnberg and Ostermeier, 2012; Kowalsky et al., 2015), resulting in 3 pooled sub-libraries each consisting of mutations targeted to a different third of the ORF. First, uracil-containing single-stranded DNA was produced from plasmid p-*crp*$^+$ following (Firnberg and Ostermeier, 2012), except for the final centrifugation step which was performed at 26,200 xg for 1h at 4°C to

increase yields. DNA concentration was quantified using the Qubit ssDNA Assay Kit (ThermoFisher Scientific).

For each of the 3 sub-libraries, PFunkel mutagenesis was then carried out as described in (Kowalsky et al., 2015), but with an amplification step of 20 mins, using 1 µg of ssDNA as template in a total volume of 100 µl. Mutagenic primers, each containing a single NNS triplet and together covering every codon between the *crp* start and stop codon, were designed using the online QuikChange Primer Design module (Agilent, Santa Clara, CA), as recommended in (Kowalsky et al., 2015). Mutated DNA was purified with the innuPREP PCRpure Kit (Analytik Jena) and eluted in 15 µl of DNAse/RNAse free water. 2 µl were then electroporated into 20 µl of electrocompetent DH5α cells, which were incubated in 500 µl of LB for 1 hour at 37°C with shaking at 200 rpm. These cells were then plated on LB-agar with 50 µg/ml streptomycin and incubated overnight at 37°C. Enough transformations were performed to obtain ~40,000 – 60,000 colonies for each sub-library, to avoid loss of complexity. For each sub-library, plasmid DNA was purified from 4 colonies (Plasmid Mini Kit, QIAGEN) and the *crp* region was Sanger-sequenced (GATC) as a preliminary test of library quality, which was satisfactory. In sub-libraries 1 (in which the first third of *crp* is mutated) and 3, 3 of the clones showed a single unique NNS codon substitution, and 1 was wildtype. In sub-library 2, 3 of the clones showed a single unique NNS codon substitution, and 1 showed 2 NNS codon substitutions. Two clones in total showed

signs of a SNP, which could result either from accidental isolation of multiple clones, or from a single clone containing two different sequences. Neither case is problematic, as plasmid DNA from all clones was then pooled together for the barcoding step. Finally, for each sub-library, all colonies were scraped off the agar into 40% LB-glycerol (3 mL/plate), mixed thoroughly, and mixed plasmid DNA was directly purified from 800 µl of this suspension (Plasmid Mini Kit, QIAGEN).

Library Barcoding

Association of mutations with short unique DNA barcodes has proved a powerful method for the deep-mutational scanning of long DNA sequences (Mavor et al., 2016; Sarkisyan et al., 2016). In addition, their high diversity helps overcome the problem of PCR and sequencing errors in the mutated gene, and they provide internal replicates for each genotype. We therefore tagged our plasmid sub-libraries with barcodes containing 20 random nucleotides, split into 4 blocks of 5 (Levy et al., 2015) to aid their alignment to a reference: $N_5ATN_5ATN_5ATN_5$. Barcodes were inserted immediately downstream of *crp*'s T1 transcriptional terminator, and so are expected to be effectively neutral for fitness.

In detail, primers oKH160309a and oKH160104b (Table 2.S3) were used at a concentration of 0.5 µM each to PCR-amplify *bla* from plasmid pKD3 (Datsenko and Wanner, 2000), using Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific) in its High-Fidelity buffer, following the manufacturer's recommendations.

Cycling conditions were: 98°C for 30 secs, followed by 30 cycles of 98°C for 10 secs, 60°C for 30 secs and 72°C for 25 secs, with a final extension step of 72°C for 3 mins. PCR product quality was checked by agarose gel electrophoresis, after which the product was column-purified (QIAquick PCR Purification Kit, QIAGEN) and quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). The purified product was then digested for 1 hour with SpeI-HF restriction enzyme (NEB CutSmart buffer), while each of the purified plasmid sub-libraries obtained above was digested for 1 hour with BstZ17I and SpeI-HF restriction enzymes (NEB CutSmart buffer). Digested DNA was again column-purified (QIAquick PCR Purification Kit, QIAGEN) and quantified with a NanoDrop ND-1000 spectrophotometer. 60 ng of each digested sub-library was then ligated in a 1:4 molar ratio with the *bla*/barcode-containing insert in a total volume of 20 µl. The ligation was carried out at 16°C overnight using T4 DNA ligase (NEB T4 DNA ligase reaction buffer), which was then deactivated by heating at 65°C for 10 mins. The ligate was microdialysed against water for 30 mins (MF-Millipore, Merck), after which several transformations were performed as follows: 1-2 µl were electroporated into 15µl commercially-prepared ElectroMAX DH5α-E electrocompetent cells (Invitrogen); cells were recovered in 500 µl LB for 30 mins (to minimise cell replication) at 37°C with shaking at 200rpm, plated on LB-agar with 100 µg/ml ampicillin and incubated overnight at 37°C. For each sub-library, plasmid DNA was purified from 2 colonies (QIAquick PCR Purification Kit, QIAGEN) for Sanger sequencing (GATC) of the *crp* and barcode

regions as a preliminary test of barcoding efficiency. All 6 colonies contained a unique, correctly inserted barcode. Of the *crp* sequences, 1 of the 6 was wildtype, 1 contained 2 NNS codon substitutions and the other 4 contained single, unique NNS codon substitutions. Some clones again showed signs of SNPs, but this was not problematic, as plasmid DNA from all clones would again be purified and pooled in the next step. For each sub-library, an estimated 200,000 colonies (each expected to carry a unique barcode) were scraped off the agar into 40% LB-glycerol (3 ml/plate), and plasmid DNA was purified directly from 800 µl of each resulting cell suspension (QIAprep Spin Miniprep Kit, Qiagen) after thorough mixing.

Barcode-mutation association: sample preparation

To reveal the *crp* ORF sequence linked to each barcode, each plasmid sub-library was PCR-amplified with a different forward primer binding just upstream of the mutated *crp* region and a common reverse primer binding just downstream of the barcode. A standard 2-step PCR protocol was used to add the technical sequences necessary for paired-end Illumina MiSeq sequencing of barcode-*crp* amplicons. Both PCR steps (< 15 cycles each) were performed in emulsion (Micellula DNA Emulsion & Purification Kit, Roboklon) to avoid recombination that would create false barcode-mutation associations, under conditions optimised according to the manufacturer's recommendations, and using KAPA HiFi HotStart ReadyMixPCR Kit (Kapa Biosystems).

For the first PCR, primer pairs were oCRPlink-1-fwd/oCRPlink-rev for sub-library 1, oCRPlink-2-fwd/oCRPlink-rev for sub-library 2, and oCRPlink-3-fwd/oCRPlink-rev for sub-library 3 (Table 2.S3). These primers contain adaptors for a 2nd PCR at their 5' extremities, followed by fully randomised hexamers added to increase amplicon diversity to facilitate MiSeq flow-cell clustering. Purified products from the 1st PCR were each quantified fluorometrically (dsDNA HS Assay Kit with a QuBit 2.0, Thermo Scientific), and used as templates in the 2nd emulsion PCR step, which employed a different pair of P5/P7 Nextera Index Kit primers (Illumina) for each sub-library (these add Illumina adaptors and sub-library multiplexing indexes). Each resulting amplicon sub-library was gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 1.5% agarose gel and a 20,000X dilution was quantified by qPCR using KAPA Library Quantification Kit for Illumina (Kapa Biosystems) on a LightCycler 480 (Roche), following the manufacturer's recommendations.

The amplicon sub-libraries are composed of DNA fragments of the structure: P5 - i5 - N_6 PCR tag – *crp* region - N_{20} plasmid barcode - N_6 PCR tag - i7 - P7, with the *crp* region containing the whole ORF in sub-library 1 (total amplicon length ~ 1.2 kb), the C-terminal two thirds in sub-library 2 (total amplicon length ~ 1 kb), and the C-terminal third in sub-library 3 (total amplicon length ~ 0.7 kb). 300nt paired-end MiSeq sequencing allowed us to sequence the entire mutated region of each *crp* sub-library (and slightly further) on Read 1 and the plasmid barcode on Read 2 (note that

Reads 1 and 2 do not overlap). For this, a 600-cycle MiSeq Reagent Kit v3 (Illumina) was used, and DNA was loaded at a concentration of 10 pM with a 15 % PhiX DNA spike-in (PhiX Control v3, Illumina). Several runs were performed, and preliminary quality filtering and demultiplexing by the standard MiSeq software package (Illumina) resulted in an output of ~10 M read pairs for each sub-library, giving an expected coverage of ~50X for each plasmid barcode.

Barcode-mutation association: analysis

MiSeq reads were processed using the Mothur (Schloss et al., 2009) (version 1.37.6) software package via the following steps: reads were quality-filtered by size (>250 bases), number of uncalled bases (<3 Ns) and length of the longest homopolymer stretch, another indicator of overall read quality (<13 bases). Sequences of the mutated region of *crp* were extracted from Read 1, and plasmid barcode sequences from Read 2, by Needleman alignment to reference sequences (default alignment parameters). Reads for which either the *crp* or barcode region contained insertions or did not generate a full alignment with the reference were discarded. The Mothur Precluster algorithm was then used to cluster barcode sequences differing by a Hamming distance of 1, with the aim of correcting for PCR and sequencing errors (the potential barcode diversity is so high ($> 1x10^{12}$) that the presence of immediately neighbouring sequences is very likely due to these errors). The algorithm uses sequence abundance to decide the "true" (majority) barcode sequence for each cluster, and to

decide where a sequence clusters if it has >1 immediate neighbour. After de-gapping and re-grouping barcode sequences to account for any alignment ambiguities resulting from small deletions, barcode clusters were used to build a dictionary assigning each "true" barcode sequence to a consensus sequence for the mutated *crp* region. Barcode clusters were only considered if they contained >4 reads, and an associated consensus *crp* region sequence was only assigned if the most common base at every considered *crp* position occurred at a frequency of ≥ 0.75.

Transformation of host strain with plasmid library

To move the barcoded plasmid sub-libraries into the final host strain, while avoiding the creation of transformants harbouring multiple unique plasmids (Goldsmith et al., 2007), several transformations were performed as follows, with plasmid concentration kept fairly low: 10 ng of each purified plasmid sub-library obtained above were electroporated into 50 µl electrocompetent *MG1655 Δcrp* cells; cells were recovered in 500 µl LB for 30 mins at 37°C with shaking at 200rpm, dilutions were plated on LB-agar with 100 µg/ml ampicillin and incubated overnight at 37°C. For each sub-library, an estimated 1-3 million colonies were scraped off the agar into LB-glycerol (40%), and this cell suspension was aliquoted and stored at -80°C after thorough mixing.

Bulk competition assays

The final mutant sub-libraries (*MG1655 Δcrp* transformed with barcoded plasmid sub-libraries) were pooled and competed over ~30 mean generations (~3 days) in the 4 different competition media. Cell density was kept low during competition (OD_{600} < 0.15) by serial transfer into fresh medium, in order to maintain the culture in exponential phase and to avoid large changes in medium composition. Large volumes of media (100 ml) were used to avoid severe population bottlenecks during serial transfer (>10^8 cells each transfer). Plasmid DNA was purified from the culture at several time-points for HiSeq sequencing of plasmid barcodes. Plasmid barcode abundance serves as a proxy for the abundance of cells carrying that particular barcode. The change in frequency over time of a barcode thus provides an estimate of competitive fitness for the lineage carrying that barcode (Hietpas et al., 2011). Since we know the *crp* sequence associated to each barcode, this in turn provides us with a distribution of fitness estimates for every mutant.

The base competition medium (BCM) consisted of M9 + 0.4% glucose, with 100 µg/ml ampicillin to select against plasmid loss. The 4 competition assays were performed in: BCM, BCM + 0.7 mM cAMP, BCM + 100mM NaCl, and BCM + 100mM NaCl + 1.2mM cAMP. In detail, 1 ml of frozen cell stock of each of the sub-libraries was washed by pelleting, resuspending in 50 ml M9 + 0.4% glucose, pelleting again, and resuspending again in M9 + 0.4% glucose. Cell density of each washed sub-library was quantified (BioMate 3S, Thermo Scientific), and they were then co-diluted in equal cell

quantity into 200 ml BCM (in a 500 ml container) to result in a total blank-subtracted OD_{600} ~0.05 (200 µl read by Varioskan microplate reader, Thermo Scientific). This common starting-culture was recovered for ~5 hours at 37°C with shaking at 200 rpm, reaching an OD_{600} of 0.11 (t_0), before being concentrated 50X by pelleting 100 ml and resuspending in 2 ml BCM. 270 µl of this cell concentrate was then diluted into 100 ml of each competition medium (in 250 ml containers), aiming for an OD_{600} of ~0.015. In all cases, media were pre-warmed at 37°C prior to transfer to keep temperature constant and detect any contamination. The 4 cultures were left to grow (37°C, 200 rpm) to an OD_{600} of ~0.12 (~3 mean generations; t_1), and 6.25 ml of each culture was then transferred to 93.75 ml fresh competition media (16X dilution). The 4 cultures were again left to grow to an OD_{600} of ~0.12 (~4 mean generations; t_2), and the transfer procedure was repeated until t_8, for a total of ~31 mean generations of competition. The precise number of mean generations between each sampling was calculated from measured OD_{600} values and used for estimating fitness. At every transfer, plasmid DNA was also purified from a 50 ml sample of culture (QIAprep Spin Miniprep Kit, Qiagen) and quantified fluorometrically (dsDNA HS Assay Kit with a QuBit 2.0, Thermo Scientific) for eventual HiSeq sequencing of plasmid barcodes. The rest remaining after this and transfer was pelleted, resuspended in LB-40% glycerol and stored at -80°C as an archive.

Barcode-sequencing of competed mutants: sample preparation

To track plasmid barcode frequencies throughout the competition experiments, barcodes were PCR-amplified from plasmid DNA in 2 steps, as for *Barcode-mutation association*, to add technical sequences necessary for 100nt overlapping paired-end Illumina HiSeq sequencing. This was performed for the sample from time-points t_0, t_1, t_2, t_4, and t_8 (approximately 0, 3, 7, 15 and 31 mean generations).

In detail, at each selected time-point, 20 ng of purified plasmid DNA was PCR-amplified in a 40 µl reaction using 0.6 µM each of primers oCRP-BCseq-fwd and oCRP-BCseq-rev (Table 2.S3). These primers contain adaptors for a 2nd PCR at their 5' extremities, followed by fully randomised hexamers to increase amplicon diversity, as in *Barcode-mutation association*. In this case, the randomized hexamers were also used to detect PCR duplicates arising from the 2nd PCR (Levy et al., 2015). KAPA HiFi HotStart ReadyMixPCR Kit (Kapa Biosystems) was used for amplification, under the following cycling conditions (cycle number was kept low to reduce PCR errors and artefacts): 95°C for 3 mins, followed by 13 cycles of 98°C for 20 secs, 58°C for 30 secs and 68°C for 30 secs, with a final extension step of 68°C for 2 mins. Amplicons (~200 bp) were gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 2% agarose gel and quantified fluorometrically (dsDNA HS Assay Kit with a QuBit 2.0, Thermo Scientific). A 2nd 40 µl PCR was then performed using ~8 ng of each amplicon as template and 0.6 µM each of a P5 and P7 Nextera Index Kit primer (Illumina) to add Illumina adaptors and multiplexing indexes. KAPA HiFi HotStart

ReadyMixPCR Kit (Kapa Biosystems) was again used for amplification, under the following cycling conditions: 95°C for 3 mins, followed by 13 cycles of 98°C for 20 secs, 55°C for 30 secs and 68°C for 30 secs, with a final extension step of 68°C for 5 mins. These ~300 bp amplicons, of the structure, P5 - i5 - N_6 PCR tag - N_{20} plasmid barcode - N_6 PCR tag - i7 - P7, were gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 2% agarose gel and sent to IntegraGen (Evry, France) for qPCR-based quantification, equimolar pooling and 100nt paired-end HiSeq-4000 sequencing (Illumina). Preliminary quality filtering and demultiplexing (Integragen, Evry, France) resulted in ~19-43 million read pairs per time-point per competition experiment, giving, for each point, an expected barcode coverage of ~32-72X.

Barcode-sequencing of competed mutants: analysis

HiSeq sequencing reads were processed using the Mothur (Schloss et al., 2009) (version 1.37.6) software package by the following steps: Forward and reverse reads were joined into contigs using Mothur's make.contigs command with the default parameters. Contigs were then quality-filtered by size (<131bp, as longer contigs imply forward and reverse reads could not be properly overlapped), number of uncalled bases (no Ns) and length of longest homopolymer stretch, an indicator of overall read quality (<9 bases). To remove the majority of PCR duplicates arising from the 2nd PCR (made possible by randomised hexamers introduced on each side of the barcode during the 1st PCR (Levy et al., 2015)), if a particular complete contig was present more than once,

only one copy was kept. Barcode sequences were then extracted after aligning contigs to the reference sequence (Needleman global alignment). Reads containing insertions or not generating a full alignment with the reference were discarded. After de-gapping and re-clustering barcode sequences to account for any alignment ambiguities resulting from small deletions, the number of occurrences of each barcode was tabulated across all time-points for each competition experiment. Finally, a custom R (v.3.4.3) script was used to merge these barcode counts tables with the barcode- mutation dictionary generated in *Barcode-mutation association*. Only barcodes associated to wildtype *crp* DNA sequences or those containing a single mutated codon, and containing no mutations in the sequenced region outside the ORF, were considered for further analysis.

Estimation of competitive fitness from Illumina sequencing data

We found that competitive fitness could be rather inconstant over the course of competition in the cAMP-containing environments (Figure 2.4). Moreover, by t_4, a substantial number of lower-fitness mutants begin to escape detection completely, and so to avoid any bias in fitness estimates we consider only the frequency changes between t_1 and t_2 for all environments. Mutations were analysed at the level of either DNA (for analysis of synonymous effects) or amino acid sequence. Under either definition, we began by removing outlier barcodes associated to the wildtype and all mutants using a 2-tailed Poisson test for $P(counts_{t_1} | \, counts_{t_2}, \lambda)$ and

$P(counts_{t_2} | counts_{t_1}, \lambda)$, with λ computed from the ratio of the sum of all relevant barcode counts at t_1 and at t_2. If the lowest of the 2 log p-values was less than -10, the barcode was declared an outlier and removed. For each mutant, the remaining barcode counts were summed and normalised to total remaining wildtype barcode counts, and mutant log relative fitness, F^{rel}, was computed as the log ratio of this frequency at t_2 to that at t_1, normalized by the number of mean generations that had elapsed between them. Mutants associated to < 5 unique barcodes, or whose total t_1 abundance was $<$ 10 read counts, were considered unreliable and discarded. All steps of fitness analysis were performed with custom R (v.3.4.3) scripts.

2.4.1 Supplementary Tables

Plasmid name	Description	DNA fragments used for construction (this study)	Construction method / Supplier	Antibiotic used for selection
pKD3 (Datsenko and Wanner, 2000)	PCR template plasmid for Datsenko-Wanner gene deletion, containing a *cat* Cm-resistance cassette flanked by *FRT* sites and an R6Kγ *pir*-dependent *ori*. Also used as PCR template for *bla* amplification in library barcoding step	-	Lab stocks	Cm
pKD46 (Datsenko and Wanner, 2000)	Plasmid with L-arabinose-inducible λ Red expression cassette for Datsenko-Wanner recombineering; temperature-sensitive *ori* (repA101ts) for easy curing	-	Lab stocks	Amp
pCP20 (Datsenko and Wanner, 2000)	Plasmid with yeast *FLP* recombinase expression cassette for Datsenko-Wanner resistance-gene excision; temperature-sensitive *ori* (repA101ts) for easy curing	-	Lab stocks	Amp
pSkunk3-BLA (Firnberg and Ostermeier, 2012)	Phagemid containing *p15A* and *f1 ori*s, *bla* β-lactamase gene and *aadA1* Str/Sp-resistance gene. Used for backbone of plasmid library and all plasmids constructed in this study	-	E. Firnberg and M. Ostermeier	Str
p-*crp*$^+$	pSkunk3-BLA backbone, with *bla* replaced by a *crp* cassette; used as template for *crp* mutagenesis	aKH150603a, aKH150603b, aKH150603c, pSKUNK-bkb	Gibson Assembly (Gibson et al., 2009)	Str
pBC-*crp*$^+$	p-*crp*$^+$, with *aadA1* replaced by a *bla* cassette linked to a random DNA barcode; used in initial optimisation of experimental conditions	aKH160316a, p-*crp*$^+$-bkb	Restriction-ligation	Amp
pBC-*crp*$^-$	pBC-*crp*$^+$, with entire *crp* cassette excised; used in initial optimisation of experimental conditions	pBC-*crp*$^+$-bkb	Restriction-ligation	Amp

Table 2.S1. Plasmids used in this study. Amp: ampicillin (100 µg/ml); Cm: chloramphenicol (10 µg/ml); Str: streptomycin (50 µg/ml).

DNA fragment name	Description/Creation	PCR template or digested plasmid	Primers used for PCR (blank if fragment comes directly from plasmid digestion)	Restriction enzymes used (either post-PCR or directly on plasmid)
aKH150603a	Entire native *crp* promoter region, with an upstream extension overlapping the EcoRV extremity of pSKUNK-bkb and a downstream SacI site introduced. PCR-amplification (upstream and downstream extensions introduced on primers)	*E. coli* K12 MG1655 genomic DNA	oKH150603a, oKH150603b	-
aKH150603b	*crp* coding sequence, along with native upstream and downstream regions, with an upstream extension overlapping aKH150603a and a downstream extension overlapping aKH150603c. PCR-amplification (upstream and downstream extensions introduced on primers)	*E. coli* K12 MG1655 genomic DNA	oKH150603c, oKH150615d	-
aKH150603c	*rrnB* T1 transcriptional terminator, with an upstream KpnI site introduced and a downstream extension overlapping the SpeI extremity of pSKUNK-bkb (upstream and downstream extensions introduced on primers)	*E. coli* K12 MG1655 genomic DNA	oKH150603e, oKH150603f	-
aKH160316a	*bla* cassette, with a randomised barcode region inserted downstream followed by a SpeI site, and an upstream extension containing a short region missing from p-crp^+-bkb	pKD3	oKH160309a, oKH160104b	SpeI
p-crp^+-bkb	p-crp^+ backbone, with *aadA1* excised. Double-digest of p-crp^+ followed by column purification	p-crp^+	-	BstZ17I, SpeI
pBC-crp^+-bkb	pBC-crp^+ backbone, with entire *crp* cassette excised. Double-digest of pBC-crp^+ followed by mung bean nuclease blunting and gel extraction	pBC-crp^+	-	EcoRV, SpeI
pSKUNK-bkb	pSkunk3-BLA backbone, containing *ori*s and *aadA1* Str/Sp-resistance gene. Double-digest of pSkunk3-BLA followed by gel extraction	pSkunk3-BLA	-	EcoRV, SpeI

Table 2.S2. DNA fragments used for cloning in this study.

Primer name	Sequence (5' -> 3')
c1 (Datsenko and Wanner, 2000)	TTATACGCAAGGCGACAAGG
c2 (Datsenko and Wanner, 2000)	GATCTTCCGTCACAGGTAGG
KO-crp-fwd	GGCGTTATCTGGCTCTGGAGAAAGCTTATAACAGAGGATAACCGCGCATGGTGTAGGCTGGAGCTGCTTC
KO-crp-rev	CTACCAGGTAACGCGCCACTCCGACGGGATTAACGAGTGCCGTAAACGACCATATGAATATCCTCCTTAG
oCRP-BCseq-fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNCTACAAACTCTTCCTGTCGTC
oCRP-BCseq-rev	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNCAAGATCCGGCCACGATGC
oCRPlink-1-fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNCATAACAGAGGATAACCGCG
oCRPlink-2-fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNTCTCCTATCTGAATCAGGGTG
oCRPlink-3-fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNCAACCTGGCGTTCCTCGAC
oCRPlink-rev	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNGTCCGGCGTAGAGGATCTG
oKH150603a	AGCCAGAAAACCGAATTTTGCTGGGTGGGCTAACGATATCAGAGTACGCGTACTAACCAAATCGCGCAAC
oKH150603b	TTATGAGCTCTCTCCAGAGCCAGATAACGCCGCTGTCT
oKH150603c	AGAGACAGCGGCGTTATCTGGCTCTGGAGAGAGCTCATAACAGAGGATAACCGCGCATG
oKH150603e	GTTTGGTACCCAGGCATCAAATAAAACGAAAGGCTCAG
oKH150603f	AGCGCGTCGGCCGGTCGAATGCATAAGCTTACTAACTAGTTGTAGATATGACGACAGGAAGAGTTTGT
oKH150615d	GCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTGATGCCTGGGTACCCGCCACTCCGACGGGATTA
oKH160104b	TACACTCCGCTAGCGCTGATGTCCGGCGGTGCCAGGTGGCACTTTTCGGG
oKH160309a	TTTTTACTAGTGGTACCTTNNNNNATNNNNNATNNNNNATNNNNNATCTTCAGATCCTCTACGCCGG
verif-*crp*-fwd	TTTCCTGACAGAGTACGCGT
verif-*crp*-rev	GCGTTAATCCGGTCAGCAAA

Table 2.S3. PCR primers used in this study, excluding those used for mutagenesis

2.5 References

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. Science *324*, 1720–1723.

Balsalobre, C., Johansson, J., and Uhlin, B.E. (2006). Cyclic AMP-Dependent Osmoregulation of crp Gene Expression in Escherichia coli. Journal of Bacteriology *188*, 5935–5944.

Bank, C., Hietpas, R.T., Wong, A., Bolon, D.N., and Jensen, J.D. (2014). A Bayesian MCMC Approach to Assess the Complete Distribution of Fitness Effects of New Mutations: Uncovering the Potential for Adaptive Walks in Challenging Environments. Genetics *196*, 841–852.

Bank, C., Hietpas, R.T., Jensen, J.D., and Bolon, D.N.A. (2015). A Systematic Survey of an Intragenic Epistatic Landscape. Molecular Biology and Evolution *32*, 229–238.

Basak, S., and Jiang, R. (2012). Enhancing E. coli Tolerance towards Oxidative Stress via Engineering Its Global Regulator cAMP Receptor Protein (CRP). PLoS ONE *7*, e51179.

Bataillon, T., and Bailey, S.F. (2014). Effects of new mutations on fitness: insights from models and data: Effects of new mutations on fitness. Annals of the New York Academy of Sciences *1320*, 76–92.

Bednarska, N.G., Schymkowitz, J., Rousseau, F., and Van Eldere, J. (2013). Protein aggregation in bacteria: the thin boundary between functionality and toxicity. Microbiology *159*, 1795–1806.

Bernet, G.P., and Elena, S.F. (2015). Distribution of mutational fitness effects and of epistasis in the 5' untranslated region of a plant RNA virus. BMC Evolutionary Biology *15*.

Bettenbrock, K., Sauter, T., Jahreis, K., Kremling, A., Lengeler, J.W., and Gilles, E.-D. (2007). Correlation between Growth Rates, EIIACrr Phosphorylation, and Intracellular Cyclic AMP Levels in Escherichia coli K-12. Journal of Bacteriology *189*, 6891–6900.

Bren, A., Park, J.O., Towbin, B.D., Dekel, E., Rabinowitz, J.D., and Alon, U. (2016). Glucose becomes one of the worst carbon sources for E.coli on poor nitrogen sources due to suboptimal levels of cAMP. Scientific Reports *6*.

Browning, D.F., and Busby, S.J.W. (2016). Local and global regulation of transcription initiation in bacteria. Nature Reviews Microbiology *14*, 638–650.

Carrasco, P., de la Iglesia, F., and Elena, S.F. (2007). Distribution of Fitness and Virulence Effects Caused by Single-Nucleotide Substitutions in Tobacco Etch Virus. Journal of Virology *81*, 12979–12984.

Chan, Y.H., Venev, S.V., Zeldovich, K.B., and Matthews, C.R. (2017). Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. Nature Communications *8*, 14614.

Charlesworth, D., Charlesworth, B., and Morgan', M.T. (1995). The Pattern of Neutral Molecular Variation Under the Background Selection Model. Genetics *141*, 1619–1632.

Chevin, L.-M., Lande, R., and Mace, G.M. (2010). Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. PLoS Biology *8*, e1000357.

Damkiaer, S., Yang, L., Molin, S., and Jelsbak, L. (2013). Evolutionary remodeling of global regulatory networks during long-term bacterial adaptation to human hosts. Proceedings of the National Academy of Sciences *110*, 7766–7771.

Dandage, R., Pandey, R., Jayaraj, G., Rai, M., Berger, D., and Chakraborty, K. (2018). Differential strengths of molecular determinants guide environment specific mutational fates. PLOS Genetics *14*, e1007419.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proceedings of the National Academy of Sciences *97*, 6640–6645.

Denby, C.M., Im, J.H., Yu, R.C., Pesce, C.G., and Brem, R.B. (2012). Negative feedback confers mutational robustness in yeast transcription factor regulation. Proceedings of the National Academy of Sciences *109*, 3874–3878.

Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. ELife *7*, e32472.

Domingo-Calap, P., Cuevas, J.M., and Sanjuán, R. (2009). The Fitness Effects of Random Mutations in Single-Stranded DNA and RNA Bacteriophages. PLoS Genetics *5*, e1000742.

Dominguez, A.A., Lim, W.A., and Qi, L.S. (2016). Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. Nature Reviews Molecular Cell Biology *17*, 5–15.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nature Reviews Genetics *8*, 610–618.

Fic, E., Bonarek, P., Gorecki, A., Kedracka-Krok, S., Mikolajczak, J., Polit, A., Tworzydlo, M., Dziedzicka-Wasylewska, M., and Wasylewski, Z. (2009). cAMP Receptor Protein from *Escherichia coli* as a Model of Signal Transduction in Proteins – A Review. Journal of Molecular Microbiology and Biotechnology *17*, 1–11.

Firnberg, E., and Ostermeier, M. (2012). PFunkel: Efficient, Expansive, User-Defined Mutagenesis. PLoS ONE *7*, e52031.

Firnberg, E., Labonte, J.W., Gray, J.J., and Ostermeier, M. (2014). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. Molecular Biology and Evolution *31*, 1581–1592.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods *11*, 801–807.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J.S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J.A., et al. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Research *44*, D133–D143.

Gayán, E., Cambré, A., Michiels, C.W., and Aertsen, A. (2017). RpoS-independent evolution reveals the importance of attenuated cAMP/CRP regulation in high hydrostatic pressure resistance acquisition in E. coli. Scientific Reports *7*.

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

Goldsmith, M., Kiss, C., Bradbury, A.R.M., and Tawfik, D.S. (2007). Avoiding and controlling double transformation artifacts. Protein Engineering Design and Selection *20*, 315–318.

Görke, B., and Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. Nature Reviews Microbiology *6*, 613–624.

Grainger, D.C., Hurd, D., Harrison, M., Holdstock, J., and Busby, S.J.W. (2005). Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. Proceedings of the National Academy of Sciences *102*, 17693–17698.

Haimovich, A.D., Muir, P., and Isaacs, F.J. (2015). Genomes by design. Nature Reviews Genetics *16*, 501–516.

Hermann, R., Lehmann, M., and Büchs, J. (2003). Characterization of gas-liquid mass transfer phenomena in microtiter plates: Gas-Liquid Mass Transfer in Microtiter Plates. Biotechnology and Bioengineering *81*, 178–186.

Hietpas, R.T., Jensen, J.D., and Bolon, D.N.A. (2011). Experimental illumination of a fitness landscape. Proceedings of the National Academy of Sciences *108*, 7896–7901.

Hietpas, R.T., Bank, C., Jensen, J.D., and Bolon, D.N.A. (2013). Shifting fitness landscapes in response to altered environments. Evolution *67*, 3512–3522.

Hill, W.G. (2010). Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B: Biological Sciences *365*, 73–85.

Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2012). New insights into bacterial adaptation through in vivo and in silico experimental evolution. Nature Reviews Microbiology *10*, 352–365.

Hoffmann, A.A., and Sgrò, C.M. (2011). Climate change and evolutionary adaptation. Nature *470*, 479.

Jackson, D.W., Simecka, J.W., and Romeo, T. (2002). Catabolite Repression of Escherichia coli Biofilm Formation. Journal of Bacteriology *184*, 3406–3410.

Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., et al. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. Proceedings of the National Academy of Sciences *110*, 13067–13072.

Jiang, L., Mishra, P., Hietpas, R.T., Zeldovich, K.B., and Bolon, D.N.A. (2013). Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. PLoS Genetics *9*, e1003600.

Jiang, L., Liu, P., Bank, C., Renzette, N., Prachanronarong, K., Yilmaz, L.S., Caffrey, D.R., Zeldovich, K.B., Schiffer, C.A., Kowalik, T.F., et al. (2016). A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance. Journal of Molecular Biology *428*, 538–553.

Keightley, P.D., and Eyre-Walker, A. (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philosophical Transactions of the Royal Society B: Biological Sciences *365*, 1187–1193.

Klesmith, J.R., Bacik, J.-P., Michalczyk, R., and Whitehead, T.A. (2015). Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. ACS Synth. Biol. *4*, 1235–1243.

Kolb, A., Busby, S., Buc, I.I., Garges, S., and Adhya, S. (1993). Transcriptional Regulation by cAMP and its Receptor Protein. Annual Review of Biochemistry *62*, 749–795.

Körner, H., Sofia, H.J., and Zumft, W.G. (2003). Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. FEMS Microbiology Reviews *27*, 559–592.

Kowalsky, C.A., Klesmith, J.R., Stapleton, J.A., Kelly, V., Reichkitzer, N., and Whitehead, T.A. (2015). High-Resolution Sequence-Function Mapping of Full-Length Proteins. PLOS ONE *10*, e0118193.

Kremling, A., Bettenbrock, K., and Gilles, E. (2007). Analysis of global control of Escherichia coli carbohydrate uptake. BMC Systems Biology *1*, 42.

Landis, L., Xu, J., and Johnson, R.C. (1999). The cAMP receptor protein CRP can function as an osmoregulator of transcription in Escherichia coli. Genes & Development *13*, 3081–3091.

Levy, S.F., Blundell, J.R., Venkataram, S., Petrov, D.A., Fisher, D.S., and Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature *519*, 181–186.

Li, C., and Zhang, J. (2018). Multi-environment fitness landscapes of a tRNA gene. Nature Ecology & Evolution *2*, 1025–1032.

Li, C., Qian, W., Maclean, C.J., and Zhang, J. (2016). The fitness landscape of a tRNA gene. Science *352*, 837–840.

Loewe, L. (2006). Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. Genetical Research *87*, 133–159.

Mao, X.-J., Huo, Y.-X., Buck, M., Kolb, A., and Wang, Y.-P. (2007). Interplay between CRP-cAMP and PII-Ntr systems forms novel regulatory network between carbon metabolism and nitrogen assimilation in Escherichia coli. Nucleic Acids Research *35*, 1432–1440.

Marciano, D.C., Lua, R.C., Katsonis, P., Amin, S.R., Herman, C., and Lichtarge, O. (2014). Negative Feedback in Genetic Circuits Confers Evolutionary Resilience and Capacitance. Cell Reports *7*, 1789–1795.

Marciano, D.C., Lua, R.C., Herman, C., and Lichtarge, O. (2016). Cooperativity of Negative Autoregulation Confers Increased Mutational Robustness. Physical Review Letters *116*.

Matsui, M., Tomita, M., and Kanai, A. (2013). Comprehensive Computational Analysis of Bacterial CRP/FNR Superfamily and Its Target Motifs Reveals Stepwise Evolution of Transcriptional Networks. Genome Biology and Evolution *5*, 267–282.

Mavor, D., Barlow, K., Thompson, S., Barad, B.A., Bonny, A.R., Cario, C.L., Gaskins, G., Liu, Z., Deming, L., Axen, S.D., et al. (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. ELife *5*, e15802.

Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA *19*, 1537–1551.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T.S. (2014). Comprehensive mutational scanning of a kinase *in vivo* reveals substrate-dependent fitness landscapes. Nucleic Acids Research *42*, e112–e112.

Nishino, K., Senda, Y., and Yamaguchi, A. (2008). CRP Regulator Modulates Multidrug Resistance of Escherichia coli by Repressing the mdtEF Multidrug Efflux Genes. The Journal of Antibiotics *61*, 120–127.

Otto, S.P., and Lenormand, T. (2002). Resolving the paradox of sex and recombination. Nature Reviews Genetics *3*, 252.

Peck, J.R., Barreaut, G., and Heath, S.C. (1997). Imperfect Genes, Fisherian Mutation and the Evolution of Sex. Genetics *145*, 1171–1199.

Peris, J.B., Davis, P., Cuevas, J.M., Nebot, M.R., and Sanjuan, R. (2010). Distribution of Fitness Effects Caused by Single-Nucleotide Substitutions in Bacteriophage f1. Genetics *185*, 603–609.

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., and Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. Science *352*, 840–844.

Roscoe, B.P., Thayer, K.M., Zeldovich, K.B., Fushman, D., and Bolon, D.N.A. (2013). Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. Journal of Molecular Biology *425*, 1363–1377.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for General Users and for Biologist Programmers. Methods in Molecular Biology *132*, 365–386.

Sanjuan, R., Moya, A., and Elena, S.F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences *101*, 8396–8401.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. Nature *533*, 397–401.

Saxer, G., Krepps, M.D., Merkley, E.D., Ansong, C., Deatherage Kaiser, B.L., Valovska, M.-T., Ristic, N., Yeh, P.T., Prakash, V.P., Leiser, O.P., et al. (2014). Mutations in Global Regulators Lead to Metabolic Selection during Adaptation to Complex Environments. PLoS Genetics *10*, e1004872.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology *75*, 7537–7541.

Seshasayee, A.S., Bertone, P., Fraser, G.M., and Luscombe, N.M. (2006). Transcriptional regulatory networks in bacteria: from input signals to output responses. Current Opinion in Microbiology *9*, 511–519.

Shultzaberger, R.K., Maerkl, S.J., Kirsch, J.F., and Eisen, M.B. (2012). Probing the Informational and Regulatory Plasticity of a Transcription Factor DNA–Binding Domain. PLoS Genetics *8*, e1002614.

Sievert, C., Nieves, L.M., Panyon, L.A., Loeffler, T., Morris, C., Cartwright, R.A., and Wang, X. (2017). Experimental evolution reveals an effective avenue to release catabolite repression via mutations in XylR. Proceedings of the National Academy of Sciences *114*, 7349–7354.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences *39*, 381–399.

Soberón-Chávez, G., Alcaraz, L.D., Morales, E., Ponce-Soto, G.Y., and Servín-González, L. (2017). The Transcriptional Regulators of the CRP Family Regulate Different Essential Bacterial Functions and Can Be Inherited Vertically and Horizontally. Frontiers in Microbiology *8*.

Stevenson, K., McVey, A.F., Clark, I.B.N., Swain, P.S., and Pilizota, T. (2016). General calibration of microbial growth in microplate readers. Scientific Reports *6*.

Tenaillon, O., Rodriguez-Verdugo, A., Gaut, R.L., McDonald, P., Bennett, A.F., Long, A.D., and Gaut, B.S. (2012). The Molecular Diversity of Adaptive Convergence. Science *335*, 457–461.

Towbin, B.D., Korem, Y., Bren, A., Doron, S., Sorek, R., and Alon, U. (2017). Optimality and sub-optimality in a bacterial growth law. Nature Communications *8*, 14123.

Wagner, A. (2012). The role of robustness in phenotypic adaptation and innovation. Proceedings of the Royal Society B: Biological Sciences *279*, 1249–1258.

Wrenbeck, E.E., Azouz, L.R., and Whitehead, T.A. (2017). Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nature Communications *8*, 15695.

You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C.W., Wang, Y.-P., Lenz, P., Yan, D., and Hwa, T. (2013). Coordination of bacterial proteome with metabolism by cyclic AMP signalling. Nature *500*, 301.

Youn, H., Kerby, R.L., Conrad, M., and Roberts, G.P. (2006). Study of Highly Constitutively Active Mutants Suggests How cAMP Activates cAMP Receptor Protein. Journal of Biological Chemistry *281*, 1119–1127.

Zhang, H., Chong, H., Ching, C.B., and Jiang, R. (2012). Random mutagenesis of global transcription factor cAMP receptor protein for improved osmotolerance. Biotechnology and Bioengineering *109*, 1165–1172.

Zhang, Z., Gosset, G., Barabote, R., Gonzalez, C.S., Cuevas, W.A., and Saier, M.H. (2005). Functional Interactions between the Carbon and Iron Utilization Regulators, Crp and Fur, in Escherichia coli. Journal of Bacteriology *187*, 980–990.

3 The thermodynamic roots of pairwise epistasis in the α-helix of β-lacatamase TEM-1: local *versus* global epistasis

Authors: André Birgy*, Harry Kemble*, Jimmy Mullaert, Karine Panigoni, Audrey Chapron, Jérémie Chatel, Melanie Magnan, Hervé Jacquier and Olivier Tenaillon

Affiliations: Infection, Antimicrobials, Modelling, Evolution, INSERM, Unité Mixte de Recherche 1137, Université Paris Diderot, Université Paris Nord, 75018 Paris, France.

Abstract: The interactions between mutations on fitness, or epistasis, affects genome evolution together with our ability to predict individual mutation effects. The mechanistic bases of epistasis remain however largely unknown. To quantify the extent and molecular bases of epistasis, we focused on a structural component of a protein and made a comprehensive library of more than 15,000 double mutants in the 11 amino-acid alpha-helix of beta-lactamase TEM-1. The pervasive epistasis observed was largely explained by a thermodynamic model of protein stability that sorted mutations as inactivating, destabilizing, neutral or stabilizing. Yet, deviations from that prediction were consistently found as the distance to the active site decreased and when the interacting residues were in contact. Our results suggest that even in a small structural component of a protein, both macroscopic and microscopic interactions shape the epistasis landscape.

3.1 Introduction

Sequences of the first proteins triggered the emergence of molecular evolution and bioinformatics in the 1960s (Hagen 2000). Yet, more than 50 years later, despite a massive number of available protein sequences and a pressing demand from human genetic disease and synthetic biology, the prediction of non-synonymous mutation effects remains a challenging task. Recently, protein deep mutational scans, in which the impacts of all possible single amino acid changes in a protein are investigated, offered new perspectives to the study of nonsynonymous mutations (Fowler & Fields 2014). However, one of the first lessons from these approaches was that mutation impact could vary with genetic background (Bank et al. 2015, 2016; Jacquier et al. 2013). These variations limit the power of descriptive mutation scans, calling for an integrated understanding of mutation effects and especially of their interactions.

Epistasis refers to the context dependency of mutation effects. In genetics, epistasis refers to interactions between mutations in general; in population genetics, pairwise epistasis refers more precisely to mutation interactions that translate to non-additivity of log-fitness effects. Epistasis between mutation A and B can be quantitatively estimated as the deviation between the observed log-fitness of the double mutants, AB, and the sum of the log fitness of both individual mutations (A and B). Under this strict definition, epistasis has been predicted to have a large impact on many facets of evolution, from the evolution of mutation rate and recombination (de Visser & Elena

2007), to the diversity of adaptive paths and the repeatability of adaptation (De Visser & Krug 2014). An integrated vision of epistasis may be obtained top-down with phenomenological models that capture its global properties (Gros et al. 2009; Martin et al. 2007), but a bottom-up mechanistic approach is needed when it comes to predicting the effects of individual mutations.

As proteins generally operate in a folded state, mutations' impacts on proteins have mostly been investigated through their impacts on that fold or its affinity with a substrate. For epistatic interactions, two mutually non-exclusive mechanistic visions have emerged. With compensatory mutations, characterized by two independently deleterious mutations that when combined outcompete at least one of the single mutants, the idea of key-lock local interactions suggested itself. Alternatively, the existence of mutations with a global impact on protein stability (Bloom et al. 2005) hinted that the cooperative nature of protein folding could also result in epistatic effects, this time at a more global level (Wylie & Shakhnovich 2011). The extent of both types of interactions and the overall prevalence of epistasis remain unclear, however.

To investigate the molecular determinants of epistatic interactions, we generated a comprehensive library of more than 15,000 single and double mutants within an alpha-helix of beta-lactamase TEM-1. TEM-1 is an extremely successful antibiotic-resistance gene that is now present in about 35% of E. coli natural isolates (EARS-Net, France).

We focused on an 11 amino acid alpha-helix, from residue 119 to 128 (Figure 3.1A), as alpha-helices are the most characterized and frequent secondary structure in protein folds. For the sake of generality this alpha-helix is not involved in the active site, it is just a structural component of the enzyme. The mutants, who cover more than 76% of all possible double mutants, were analyzed for their impact on protein activity, measured through the minimum inhibitory concentration (MIC), and more importantly through their impact on fitness, which allows a proper estimation of epistasis.

3.2 Results

Mutants were produced in bulk and associated to genetic barcodes. Changes in frequency of the barcodes estimated through sequencing were used to compute (i) fitness through 30 generations of evolution in 8 mg/l of amoxicillin, an antibiotic degraded by TEM-1 and (ii) Minimum Inhibitory concentrations through challenges with 1, 2, 4, 8 and 16 mg/l of that antibiotic. The consistency of the signature of the multiple barcodes covering a given genotype as well as the very high correlation between MIC and fitness for both single (r=0.984) and double mutants (r=0.963) supported the robustness of the data produced (Supplementary Figure 3.S1 and 3.S2). Mutants with stop codons could be used to define a lower threshold for log-fitness: below a value of 1, mutants were considered selectively lethal (Supplementary Figure 3.S3).

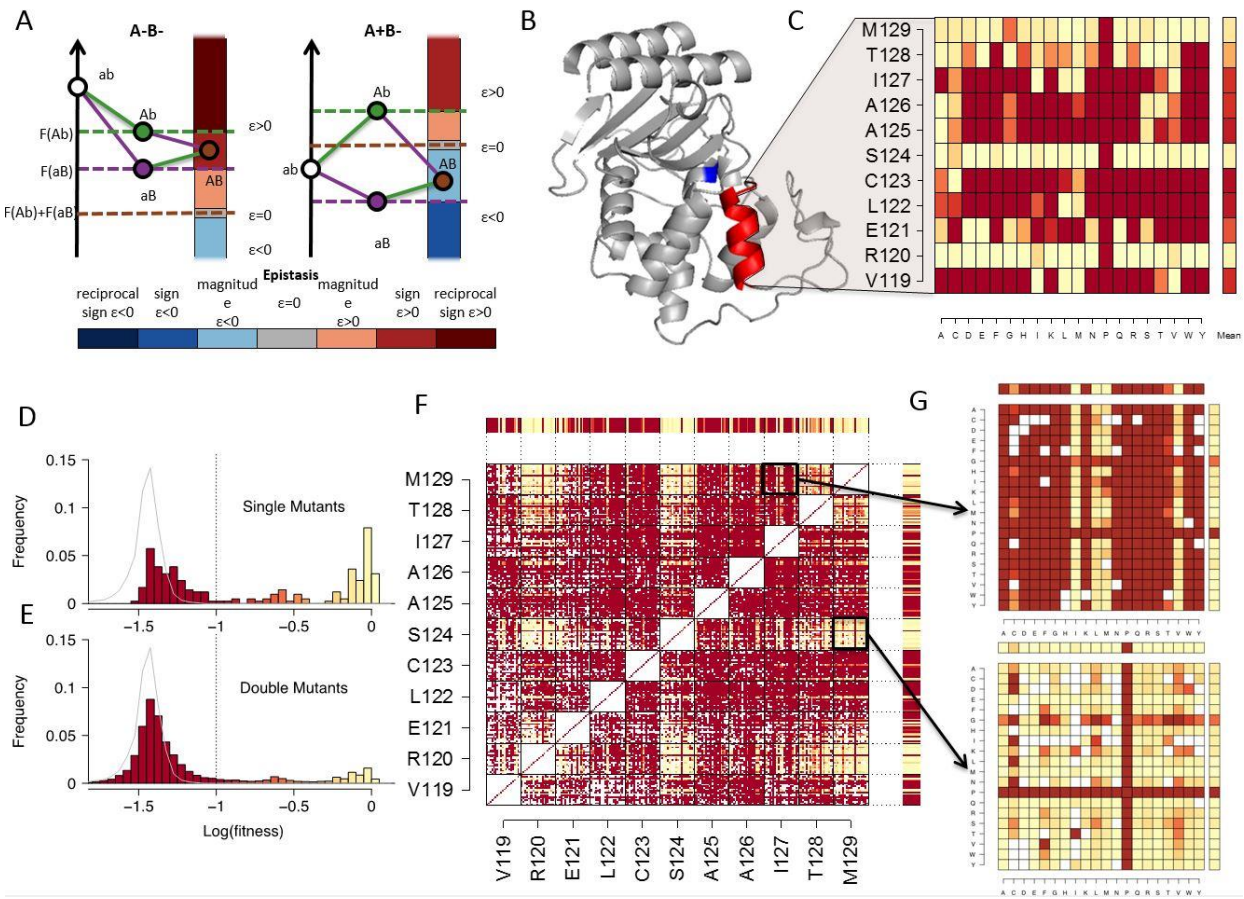


Figure 3.1. Single- and double- mutation fitness effects. **A.** Pairwise epistasis is a measure of the deviation of the observed fitness of a double mutant from the sum of the fitnesses of its constituent single mutants, on a log scale. It can also be qualitatively categorized as magnitude, sign and reciprocal sign, as well as positive or negative. The figures illustrate how this categorization functions in the case of a pair of deleterious mutations (left) and a pair including a deleterious (b to B) and a beneficial mutation (a to A) (right). **B.** 3D structure of beta-lactamase TEM-1. In red, the alpha-helix of interest and in blue, the serine residue of the active site. **C.** The effects on fitness of all single mutations per residue. **D**, **E.** Distribution of fitness effects of single (**D**) and double (**E**) mutants with the distribution of fitness effects of stop codons in grey. Below the dotted line, mutants are considered non-functional. **F**. Fitness of the double mutants with missing data in white. **G.** Zoom on the double mutant fitness map involving residues I127 and M129 (top) and S124 and M129 (bottom).

The distribution of fitness effects of single mutations had a trimodal structure with close to 50% lethal mutants (Figure 3.1D). This suggested an overall important role of the alpha-helix. The different residues had very different patterns, with 3 sites permissive to mutations, while the others were much more sensitive (Figure 3.1C). As expected, proline, which is known to be incompatible with alpha-helix structure, was lethal or close to lethal at all sites (fitness<-0.95) (Figure 3.1C). The distribution of double mutation effects also appeared to be tri-modal, with an even larger fraction of lethal genotypes (77%) (Figure 3.1E). A dominance effect emerged: mutant combinations including a lethal mutation were lethal (Figure 3.1FG). Out of the 11,477 double mutants comprising at least one lethal mutation, only 59 (0.5%) had a log fitness higher than -0.75 (Figure 3.1F). Only 7 (0.06% of total) resulted from the combination of two deleterious mutations, an instance of sign epistasis in which one of the mutations is deleterious in one background and beneficial in another. This general dominance effect clarifies the partial success of methods based on residue conservation (Adzhubei et al. 2013; Ng & Henikoff 2003) to predict mutational effects: large effects such as inserting a proline within an alpha-helix are effectively context-independent. This also suggests that the extent of pairwise key-lock epistatic compensation is very limited among mutations of large effects.

We then focused on the quantification of epistasis and noticed that double mutant fitness deviated substantially from that expected (Figure 3.2A). Epistasis could not be

computed for lethal double mutants that were predicted to be lethal based on their constituent single mutants. Excluding these cases, and restricting the dataset to mutants whose fitness is estimated with high accuracy, we could compute a distribution of epistasis that was both broadly distributed around zero and biased towards negative values (Figure 3.2B), as observed in other experiments based on proxies of protein function rather than on true fitness. Yet some large positive epistasis was also found, especially among pairs including a beneficial mutation and a deleterious one (Figure 3.2C). We then looked at the fitness effects of individual mutations across all different genetic backgrounds (Figure 3.2D). Mutations exhibited highly contrasting patterns. First, 67 lethal mutations were lethal across all backgrounds. Second, 115 deleterious mutations, including some lethals, had their fitness positively correlated with background fitness (Figure 3.2DE). Third, 17 mutations showed an overall context-independence in their effects (Figure 3.2DE). These mutations had small effects on fitness, with 10 having less than a 1% effect, 6 less than 5% and 1 a 12% effect. Finally, 10 mutations, with marginally positive fitness effects in the ancestral background (9 of them with log(fitness) <0.01, and log(fitness)>-0.01 for the other) had some marked fitness benefits in deleterious backgrounds (Figure 3.2DE). Strikingly, excluding mutations that were lethal in all backgrounds, 88% of mutations exhibited some strong form of context dependency that was structured by background fitness. This consistency suggests a macroscopic force at play, such as protein stability.

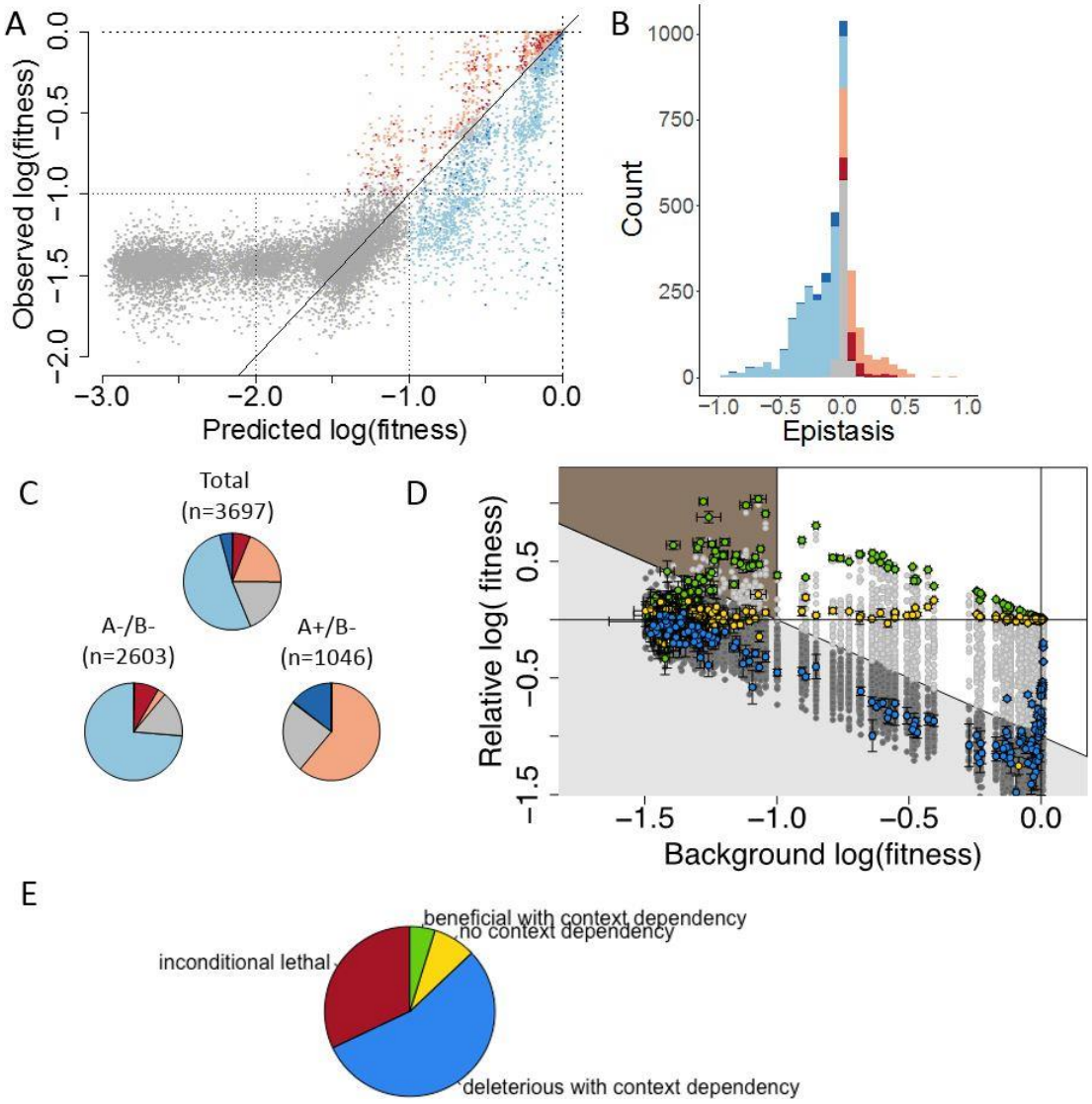


Figure 3.2. Pairwise epistasis. **A**. Log-fitness effects of double mutants, against the sum of the single mutant log-fitnesses. Grey mutants of observed fitness and predicted fitness lower than -1 cannot be used to compute epistasis. The colours of the other points represent the form of epistasis detected using the colour code of figure 1A. **B**. Distribution of epistasis using the same colour code, excluding mutants with non-measurable epistasis. **C**. Categorization of epistasis for all mutations, pairs of deleterious mutations (A-/B-), or pairs involving one deleterious and one beneficial mutation (A+/B-). **D**. Relative log-fitness effect of all mutations against the log-fitness of the different backgrounds in which they were found. The values for three focal mutations, L122A, R120K and S124E, are highlighted in blue, yellow and green respectively. **E**. The fraction of mutations falling into unconditionally inactivating, deleterious with context dependency, no context dependency and beneficial with context dependency categories is presented.

Protein stability controls the amount of protein in a functional fold. It can be directly connected to fitness in the case of an antibiotic resistance gene (Jacquier et al. 2013). Upon change of stability, the amount of functioning protein changes according to the free energy of the reference sequence (△G0) and the impact of the mutation (△△G): $Pnat = \frac{1}{1+e^{\Delta G_0+\Delta\Delta G}}$. The resulting relative fitness of a mutant can be computed as $log(\frac{W}{W0}) = log(1 + e^{\Delta G_0}) - log(1 + e^{\Delta G_0+\Delta\Delta G})$. Depending on the mutant △△G, this model produces patterns of fitness effects according to background fitness similar to those observed (Figure 3.3A).

We used a goodness-of-fit approach (Methods) to estimate a △△G0 of -2.9 kcal/mol, and estimated a △△G value for each of the single mutants. We found a correlation of 0.948 between the observed and predicted fitness under the stability model, compared with a 0.890 correlation under the no-epistasis, additive model. Hence the model provides an improvement, validating quantitatively the likely role of protein stability. Most importantly, the stability model captures the overall background dependency of the mutations' fitness effects (Figure 3.3B) and reproduces the shape and breadth of the distribution of epistasis (Figure 3.3C), with a correlation of 0.75 between observed and predicted epistasis. It suggests therefore that a large fraction of epistasis between non-synonymous mutations arises not through local interactions but mostly through a global interaction at the level of protein stability.

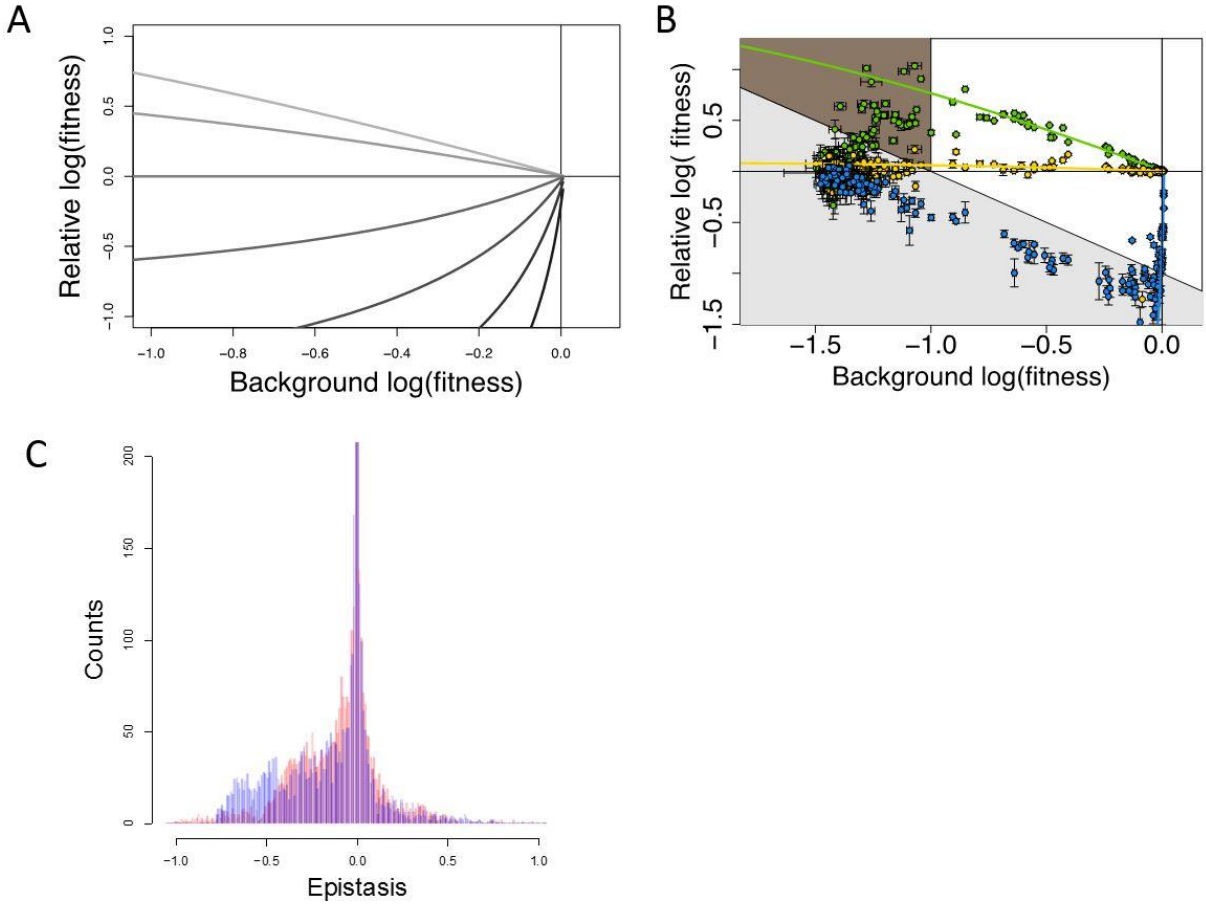


Figure 3.3. Stability and context dependency. **A.** The relationship between background log-fitness and new mutations' relative log-fitness predicted by the model of stability is presented. The modelled protein has a free energy of -2.9 kcal mol^{-1} and the impact of mutations, $\Delta\Delta G$, is -1, 0.5, 0, 0.5, 1, 1.5 and 2 from light grey to black. **B.** The lines represent the fit of the model for the 3 mutants from figure 2D. **C.** In red is the distribution of epistasis as presented in Figure 3.2B, and overlaid on it in blue is the distribution of epistasis obtained with the fitted stability model.

A deeper look at the data suggests, however, that the global stability model is not sufficient. First, if we focused on fitting a model based only on residues separated by less than 6Å, the correlation decreased to 0.925, while when only distant pairs (>6 Å) were considered, the correlation improved to 0.964. This implies that interactions between physically close were less well explained by a global model than distant ones.

Accordingly, a maximum likelihood model was used (Methods) to quantify error to the stability model. A model with two different errors was best supported, and found an error for sites at less than 6 Å of 1.84 times greater than the one found for sites further away. For some of the local interactions, forces other than stability seemed to be at play. For instance, mutation R120D and M11W showed signs of both stabilizing and destabilizing effects, the positive effects being restricted to residues in direct contact (Figure 3.4ABC). The R120D mutation leads to a change in charge that is deleterious for distant interactions, but became beneficial when associated to departure from the E121 charged amino acid at the neighbouring amino residue (Figure 3.4B). These interactions, unexplained by the stability model, represent what we refer to as idiosyncratic epistasis. They may result from non-additivity of the $\Delta\Delta G$, but also from the existence of some local forces at play that are not directly linked to stability.

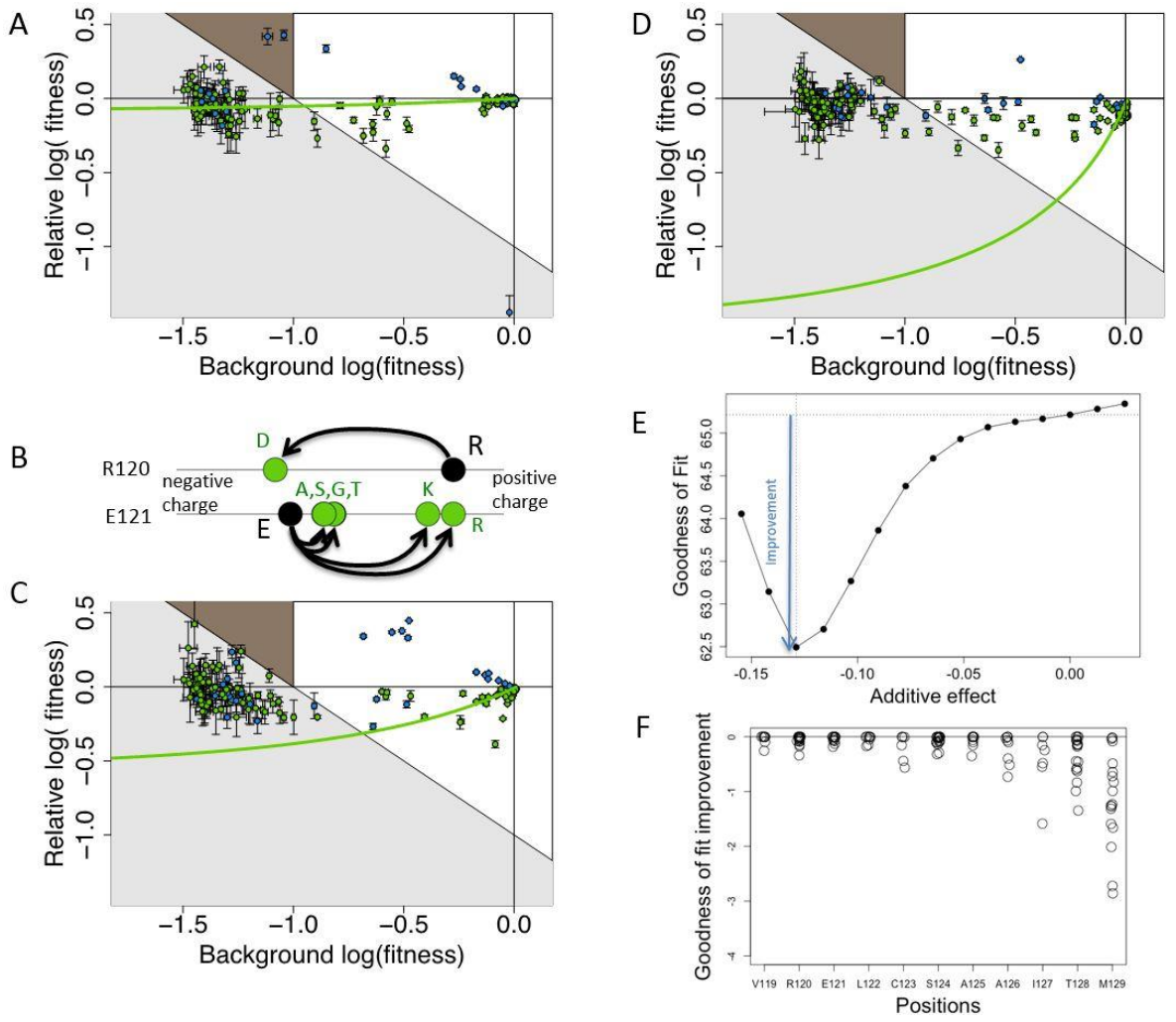


Figure 3.4. Deviations from the stability model. **A**, **C** and **D**. Relative log-fitness according to background fitness for three mutants: R120D (**A**), M129W (**C**) and M129H (**D**). Double mutants involving residues at more than 6Å of distance result in green points, and the ones involving residues at less than 6Å result in blue points. The green line is the fit of the stability model. **B**. At residue 120, the decrease of charge associated with the R to D mutation compensates mutations at residue 121 that increase the charge. **E**. Adding an additive component to the fitness of mutation M129H improves the fit to the data. The maximum improvement obtained for that mutation occurs when the additive fraction is equal to the fitness of the M11H mutation (vertical dotted line). The arrow represents the intensity of the goodness of fit improvement. **F**. Goodness of fit improvement associated with the incorporation of an additive component to fitness according to mutated residue. Residue M129 is closer to the active site.

With the overall quality of the stability model, it is tempting to use it as a new neutral model and to compute deviation from it as a new form of epistasis. However, a closer look at the data revealed that some mutations seemed to behave almost additively with the other mutations (Figure 3.4D). We tested therefore if decomposing fitness into a stability component and an additive component would result in a better overall fit (Methods) (Figure 3.4E). We found a substantial improvement for a fraction of the mutations. When looking at the 25 showing the largest improvement, 14 occurred at site M129, 6 at T128, 2 at I127, 2 at A126 and 1 at C123, revealing clearly a gradient along the alpha helix (Figure 3.4F). Knowing that residue M129 is the closest to the active site, we suspect that another global phenotype linked to protein activity rather than stability may be under selection in the protein, and that both phenotypes interact independently. The additive combination of multiple global phenotypes prevents the development of a simple alternative to epistasis, but suggests rather that both epistasis and deviation from the stability model should be considered next to uncover the molecular determinants of epistasis.

3.3 Discussion

Even though we examined interactions among residues in a local 3D structure of the protein, we showed that most mutations exhibit a macroscopic pattern of epistasis. These patterns can be captured by a simple biophysical model of protein stability that

predicts the emergence of epistasis based on the additive effects of mutations on the overall stability of the protein. Hence, additivity at the phenotypic level, here, summing the independent ΔΔG values, results in a macroscopic epistasis due to the nonlinearity of the mapping from phenotype to fitness (Otwinowski et al. 2018). The importance of this form of macroscopic epistasis at the protein level is reminiscent of the negative epistasis found genome-wide in experimental evolution (Chou et al. 2011; Khan et al. 2011; Kryazhimskiy et al. 2014; Wiser et al. 2013). Interestingly, this macroscopic epistasis also occurs between mutations affecting residues in contact. Hence, while it is tempting to interpret these interactions as the result of some key-lock mechanisms, epistasis may be resulting simply from the impact of the mutations on a global property of the protein. Many alternative mutations with similar effects on stability would have had similar effects.

Our precise estimates of fitness also allowed us to identify certain deviations from the stability model. Both forms of deviation suggest the existence of phenotypes other than stability that affect fitness in an independent manner. Some of these phenotypes may also generate a global pattern of epistasis as we could imagine for the affinity to the substrate, but others, like the conservation of charge at residues 120-121, may be very local traits. This apparent complexity challenges our ability to predict mutation effects in two ways: first, local interactions seem hard to predict, and second, stability effects are predictable but rely on precise knowledge of the overall stability of the protein.

However, two factors give us hope. First, deep mutational scans can, like the one performed here, provide some quite robust estimates of the macroscopic parameters of mutations. As the context-dependency is based on many mutations, this provides some power to precisely quantify the impact of mutations on a simple model of stability, or even in some more complex models including multiple phenotypes as we have shown here and has been done in alternative models based on reporter proteins (Otwinowski & Wilke). Second, local idiosyncratic interactions could result in the long term in some specific coevolution patterns between pairs of sites (Weigt et al. 2009). These patterns are specifically the ones that can be inferred from multiple sequence alignment, and have been shown to allow the prediction of mutational effects (Figliuzzi et al. 2016).

3.4 Methods

Strains and plasmids

E. coli strains used in this study were **XL1-Blue** (Agilent, Santa Clara, CA) of genotype *recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F' proAB lacI1qZΔM15 Tn10 (Tetr)]*; **CJ236** (NEB) of genotype *FΔ(HindIII)::cat (Tra+ Pil+ CamR)/ ung-1 relA1 dut-1 thi-1 spoT1 mcrA*; **DH5α** (Invitrogen) of *genotype F– Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17 (rK–, mK+) phoA supE44* **λ**– *thi-1 gyrA96 relA1*; **DH10b Electromax** (ThermoFisher Scientific) of genotype *F- mcrA*

Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139Δ(ara, leu)7697 galU galK λ-rpsL nupG.

Phagemid **pSkunk3-TEM-1** was obtained graciously from Elad Firnberg and Marc Ostermeier. The plasmid **pSkunk-TEM-helix** was created by inserting an *NcoI* restriction site 2 bases before the beginning of the alpha-helix to mutagenize it using single-step PFunkel mutagenesis. Using the same protocol, we also inserted *XhoI* and *NotI* restrictions sites surrounding the streptomycin/spectinomycin (Str/Spec) resistance gene. Plasmid **pKD3** was used to amplify the *cat* gene encoding chloramphenicol (Cm) resistance (Datsenko & Wanner 2000). The plasmid **pSkunk-TEM-helix-Cm** was created by swapping the Str/Spec resistance gene with *cat* and adding a DNA barcode of 20 degenerate nucleotides.

Targeted mutagenesis

Targeted mutagenesis was performed using the PFunkel mutagenesis strategy (Firnberg & Ostermeier 2012). Uracil-containing single-strand DNA of pSkunk3-TEM-helix was produced as published by (Firnberg & Ostermeier 2012; Kowalsky et al. 2015) (except the final centrifugation step, which was performed at 26200 xg for 1h at 4°C). DNA was quantified using the Qubit® ssDNA Assay Kit (ThermoFisher Scientific). Mutagenesis was then performed as previously described with 1 µg of ssDNA used as template in a total volume of 100 µl. The only difference was the elongation step, which was 15 min. Thus, the reaction cycling conditions were 95°C for

3 min, followed by 55°C for 90 sec, and 68°C for 15 min and 45°C for 15 min. We used the innuPREP PCRpure Kit (Analytik Jena) to purify DNA, and eluted it in 15 µl of distilled DNAse/RNAse-free water. 2 µl were then electroporated in 20 µl of DH5α electrocompetent cells and incubated with 500 µl of LB medium for 1 hour at 37°C with shaking at 250 rpm. The transformation was plated on LB-agar with 50 µg/ml streptomycin and incubated overnight at 37°C. PCR verification and Sanger sequencing were performed on isolated colonies using primers TEM-pSKUNK-DIM-F and TEM-pSKUNK-DIM-R to check for mutagenesis efficiency. Mutants were stocked in LB-glycerol 40% after an overnight culture at 37°C in LB media containing 50 µg/ml streptomycin.

Comprehensive PFunkel mutagenesis

Primers containing all combinations of two NNS degenerate codons (N is either A, T, G or C; S is either G or C) in the 11 codons of the targeted alpha-helix were designed with 20 fixed base-pairs surrounding the alpha helix.

Protocols were performed as published by (Firnberg & Ostermeier 2012; Kowalsky et al. 2015) for single-site mutagenesis, with an amplification step of 20 mins. Purification was carried out using the innuPREP PCRpure Kit (Analytik Jena), and DNA was eluted in 15µl of distilled water. 2 µl were electroporated in 20 µl of Dh5a electrocompetent cells and incubated with 500 µl of LB media for 1 hour at 37°C with shaking at 250 rpm. The transformation was plated on LB agar with 50 µg/ml

streptomycin and incubated overnight at 37°C. A pool of 150,000 colonies was scraped from the LB agar plates (245 mm X 245 mm, Greiner Bio-one) in LB-glycerol 40% and stored at -80°C. After pooling all colonies together, plasmids were extracted from a sample using a plasmid Miniprep Kit (Qiagen, Valencia, CA), forming the library of mutants.

Mutant barcoding

10 µg of purified plasmid from the library of mutants was digested with NotI, XhoI and NcoI (buffer 3.1 (NEB)), in 500ul total reaction volume, gel extracted (band of 3,350 bp) using Qiagen Gel Extraction kit and then also cleaned a 2nd time with Qiagen PCR Purification kit. The final concentration was 25 ng/ul. Plasmid pKD3 was used as template for PCR amplification of cat using specific primers that also contained overlapping regions of pSkunk-TEM-helix (for subsequent Gibson Assembly). The forward primer also contained a non-overlapping region with a DNA barcode consisting of 20 degenerate nucleotides. Phusion High-Fidelity DNA Polymerase (New England Biolabs) was used with reaction cycling conditions: 98°C for 30 sec, followed by 35 cycles of 98°C for 10 sec, and 62°C for 30 sec, 72°C for 15 sec and a final extension at 72°C for 2 min.

The plasmid pSkunk-TEM-helix-Cm was created by switching the spectinomycin/streptomycin resistance gene with the *cat* cassette amplified previously, using Gibson Assembly (NEB), allowing integration of the DNA barcode. The Gibson

reaction was carried out with 3µl of 25 ng/ul of plasmid fragment and 1.3 µl of 88 ng/µl of the barcode-CmR-amplicon (1:5 molar ratio backbone:insert), in a total of 20 µl reaction mix, and incubated at 50°C for 1 hour. The total volume of Gibson reaction was dialysed with water for 30 mins and 4 µl were electroporated into 20 µl of DH10b Electromax competent cells. These were then incubated in 500 µl of LB media for 1 hour at 37°C with shaking at 250 rpm. The transformants were plated on LB agar with 25 µg/ml chloramphenicol and incubated overnight at 37°C. A pool of ~2 x 106 colonies was scraped from the LB-agar plates (245 mm X 245 mm, Greiner Bio-one) into LB-glycerol 40% and frozen at -80°C.

Selection experiments

Selection with 8g/l of amoxicillin: 1ml of the frozen barcoded library cell stock was cultured in MH broth at 37°C with shaking at 250 rpm from OD_{600} 0.2 to OD_{600} 0.4 without antibiotics (named T0). Then, 3.2 ml of this first culture were used to re-inoculate 96.8 ml MH broth supplemented with 8g/l amoxicillin (corresponding to 2 dilutions below the MIC of TEM-1) until OD_{600} 0.2 (this point is called T1). This re-inoculation of 3.2 ml of culture of OD_{600} 0.2 into 96.8 ml fresh MH broth supplemented with 8 g/l amoxicillin was repeated until approximately 40 total population-averaged generations (T8). Half of these cultures were pelleted for plasmid extraction and half were pelleted and re-suspended in LB-glycerol 40% for storing at -80°C.

Selection with different amoxicillin concentrations: From the first culture, named T0, we used 3.2 ml to re-inoculate a total of 100 ml of MH broth supplemented with various amoxicillin concentrations ranging from 16 g/l to 1 g/l, and left the cultures to grow until OD_{600} 0.2 at 37°C with shaking at 250 rpm. Half of these cultures was pelleted for freezing at -80°C and half was washed 2 consecutive times with sterile physiological serum and grown overnight in fresh MH medium without antibiotics. Finally, half of this overnight culture was plasmid-extracted and half was pelleted and re-suspended in LB-glycerol 40% for storing at -80°C.

Library Preparation and Deep Sequencing

The protocol is carried out in 2 steps.

Combining barcodes and alpha-helix sequences: the first step is to reveal which barcodes are linked to which mutations in the T0 library. For that, a two-step PCR method was used to amplify the corresponding part of the gene, including the alpha helix sequence on 5' part and barcode sequence on 3' extremity, and to add the Illumina sequencing adaptor and multiplex barcode sequences. In detail, plasmid DNA concentration was determined using Qubit fluorometric quantification (ThermoFisher scientific) and normalized to 2.5 ng/µl. 12.5 ng of DNA was used for the 1st PCR using specific primers and allowed the attachment of an adaptor that is necessary for the 2nd PCR. Between specific primers and adaptors, 6 degenerate nucleotides were inserted in order to increase the diversity of DNA to facilitate MiSeq clustering.

Kapa Hifi Hotstart Ready Mix PCR Kit polymerase (Kapa Biosystems) was used for amplification. The reaction cycling conditions were 95°C for 30 sec, followed by 12 cycles of 95°C for 10 sec, 55°C for 30 sec, 68°C for 30 sec and a final extension at 68°C for 5 min.

After gel purification using Qiagen gel extraction kit (Valencia, CA), DNA was quantified using Qubit fluorometric quantification and DNA concentration was normalized. The 2nd PCR was performed using 5 ng of DNA using primers commercialized by Illumina in the Nextera Index Kit allowing dual indexing. The reaction cycling conditions were the same as previously but only 11 cycles were performed using Kapa Hifi Hotstart Ready Mix PCR Kit polymerase (Kapa Biosystems). After gel-purification with Qiagen gel extraction kit (Valencia, CA), quantification using qPCR kapa Hifi Hotstart (Kapa Biosystems) on a Light cycler 480 Roche was performed with reaction cycling conditions of 95°C for 5 min, followed by 35 cycles of 95°C for 30 sec and 60°C for 45 sec as specified by Kapa Biosystems.

This library, corresponding to the first time-point of evolution (T0), was diluted to 12 pM and loaded on the MiSeq with a mix of 10% PhiX DNA (PhiX Control v3, illumina) as sequencing control and to increase diversity. Three MiSeq V3 2x75 bp paired-end runs (Illumina technology) were performed for this part, resulting in a total of > 40M reads, for an expected ~20x coverage of barcode diversity. The paired-end reads are non-overlapping, with the alpha helix sequence on Read 1 and the barcode

sequence on Read 2. The amplification protocol was changed for the third MiSeq run: in order to decrease recombination that arises during PCR, we used an emulsion-PCR protocol (Micellula, following the manufacturer's guidelines).

Barcoding sequencing: The second step consists of sequencing the barcodes alone at the different time-points (T0, T1, T2, T4, T6 and T8), and different amoxicillin concentrations (16, 8, 4, 2, 1 g/L). For this, a similar protocol was carried out using oligonucleotides that surround the barcode region, employing the same 2-step PCR based method and similar conditions. In this case, the 6 degenerate nucleotides inserted on either side of the barcode region during the 1st PCR also allowed us to remove PCR duplicates arising from the 2nd PCR. All libraries corresponding to the different evolution time-points and amoxicillin concentrations were quantified using a qPCR-based method (Integragen) and pooled in equal molar quantity. They were then sequenced on a HiSeq4000 with a 2x100 bp paired-end kit (Illumina technology) by Integragen society, to give overlapping reads of the barcode region. The run resulted in ~300M raw paired-end reads, and so ~27M for each of the 11 time-points/conditions. This gives a barcode coverage of ~14x for each time-point/condition.

Sequence analysis

Barcode-mutant association: The following steps were performed using the Mothur software package (Schloss et al. 2009): raw reads from all sequencing runs were pooled together and quality-filtered by size (>69 bases), number of uncalled bases (<3 Ns)

and length of longest homopolymer stretch, an indicator of overall read quality (<13 bases). Alpha helix and barcode sequences were extracted from Read 1 and Read 2, respectively, after alignment to the reference sequences (Needleman global alignment). Reads for which either the alpha helix or barcode region contained insertions or did not generate a full alignment with the reference were discarded. The Mothur precluster algorithm was then used to cluster barcode sequences differing by a Hamming distance of 1, with the aim of correcting for PCR and sequencing errors (the potential barcode diversity is so high that the presence of immediately neighbouring sequences is very likely due to these errors). The algorithm uses sequence abundance to decide the "true" (majority) sequence for each cluster, and to decide where a sequence clusters if it has >1 immediate neighbor. After de-gapping and re-clustering barcode sequences to account for any alignment ambiguities resulting from small deletions, barcode clusters were used to build a dictionary assigning each "true" barcode sequence to an alpha helix sequence. Due to the high rate of PCR-derived recombination observed (caused by the long homologous region between the barcode region and alpha helix sequence, and resulting in molecules with swapped barcodes), a haplotype-based strategy was used for this step rather than one in which each nucleotide is considered independently. This is because the small number of mutations present in each mutant means that, at any particular position, the majority of molecules will possess the WT base, and so a high recombination rate can result in consensus alpha helix sequences in which mutant bases are assigned as WT. The efficiency of this strategy was ensured by

the short length of the mutagenized region and high quality of the reads, meaning that most reads did not contain a single error in the regions of interest and so were not wasted. Briefly, a custom Python script was used to perform the following: for each barcode cluster (consisting of reads whose barcode sequences are identical to or the immediate neighbor of the inferred "true" barcode sequence), the paired alpha helix sequences were fetched; the number of occurrences of each resulting alpha helix sequence was tabulated; if the cluster contains more than 2 reads in total, the most abundant alpha helix sequence is >= 5x more abundant than the second-most abundant alpha helix sequence, and the most abundant alpha helix sequence contains no Ns, then the most abundant alpha helix sequence is assigned to the "true" barcode sequence for that cluster (else the cluster is discarded).

Barcode counting: The following steps were performed using the Mothur software package (Schloss et al. 2009): demultiplexed forward and reverse reads were joined into contigs using Mothur's make.contigs command with the default parameters, which takes into account the Phred score to assign (or not) a base when there is disagreement between forward and reverse reads. Contigs were then quality-filtered by size (<151bp, as longer contigs imply forward and reverse reads could not be properly overlapped), number of uncalled bases (no Ns) and length of longest homopolymer stretch, an indicator of overall read quality (<13 bases). To remove the majority of PCR duplicates arising from the 2nd PCR (made possible by the 6 degenerate

nucleotides introduced on each side of the barcode during the 1st PCR), if a particular contig was present more than once, only one copy was kept. Barcode sequences were then extracted after aligning full contigs to the reference sequence (Needleman global alignment). Reads containing insertions or not generating a full alignment with the reference were discarded. Next, the Mothur precluster algorithm was used to cluster barcode sequences differing by a Hamming distance of 1, with the aim of correcting for PCR and sequencing errors, as described above for the barcode-mutant association. After de-gapping and re-clustering barcode sequences to account for any alignment ambiguities resulting from small deletions, the number of occurrences of each "true" barcode was tabulated across all time-points/conditions. Finally, a custom R script was used to merge the barcode-mutant dictionary generated above with the barcode counts table.

Based on previous work in which we found no clear effect of synonymous mutations, we combined all synonymous mutations into a single allele.

Quality control of barcodes

Multiple barcodes were associated to the each genotypes. Several processes may lead to variability in the signal provided by the different barcodes. First, though we used some correction and some emulsion PCR to try to correct that bias, some recombination may occur during the PCR between the part of the protein and the barcode and escape our detection procedures. Hence, a barcode may appear to be associated to the

focal genotype, but may indeed correspond to an alternative genotype. Even if a barcode is associated properly to its alpha-helix genotype, we have not sequenced the whole protein. Consequently, an undetected mutation may affect the protein elsewhere and result in a modified behaviour of that barcode. To limit the effect of these outliers, that are often barcodes associated with loss of function or maximal fitness, we first did a screen to filter outlier barcodes.

For that purpose, we computed the change in the focal genotype to wild-type genotype frequency over the first cycle of evolution (T0 to T1), using the sum of all barcodes linked the focal genotype. $K_j = \left(\frac{\sum_i BC_{ij}^1}{Wt^1}\frac{Wt^0}{\sum_i BC_{ij}^0}\right)$, in which BC_{ij}^1 is the number of reads matching the jth barcode associated to genotype i at time 1, and Wt^1 the number of reads matching barcodes associated to wild type sequence. The value of K corresponds to an estimate of fitness over one cycle. Then, for each individual barcode we can compute based on BC_{ij}^0 the estimated number of reads expected at T1. If the barcode is following the overall trend we expect

$$K_{ij} = \left(\frac{BC_{ij}^1}{Wt^1}\frac{Wt^0}{BC_{ij}^0}\right) = K_j \,.$$

We expect therefore BC_{ij}^1 to be distributed with a Poisson law of parameter, $\frac{\sum_i BC_{ij}^1}{\sum_i BC_{ij}^0}\, BC_{ij}^0$.

All barcodes, with a p-value lower than 10^{-5} were assumed to reject that model and to be the result of some of the artefacts previously mentioned. They were later on discarded, and reads matching all the remaining barcodes were combined to estimate fitness. Furthermore, barcodes with less than 10 counts for the combined time T0 and T1 were excluded as well as mutants with less than 4 barcodes.

Computing fitness

The frequency of a genotype is supposed to evolve through time according to its relative fitness with the following law: $\frac{G_i^t}{Wt^t} = \frac{G_i^0}{Wt^0}(f_i)^t$, such that $log\left(\frac{G_i^t}{Wt^t}\right) = t\ F_i + log\left(\frac{G_i^0}{Wt^0}\right)$, with $F_i = \log(f_i)$. Fitness is therefore the slope of the change in the ratio of genotype to wild-type frequency. However, because we are dealing with counts, there is a lower bound to the genotype frequency, and computing the slope may be affected by low values, especially for deleterious mutants who tend to disappear rapidly. To compute the slope, we first computed the number of points that could be reliably used for that purpose. We used a moment matching approach in which we matched the pattern of our distribution of relative counts through time with a distribution of 0s and 1s that has similar variance and mean. For each genotype, we computed the variance V_i and mean M_i of $\frac{G_i^t}{Wt^t}$ across the 5 time points (T0, T1, T2, T4 and T6).

We then computed the number points to be used for the slope as $nbp_i = ceiling\left(\frac{5}{1+\frac{V_i}{(M_i)^2}}\right)$, in which ceiling corresponds to the function rounding to the next integer.

The underlying idea is to compute a distribution composed of 0 and 1 that has a standardized mean and variance similar to the one observed for $\frac{G_i^t}{Wt^t}$ across time points. Let us consider two extremes cases. If a mutant is very deleterious, the ratios $\frac{G_i^t}{Wt^t}$ at

the different time points are X, Y, 0, 0, 0, with X>>Y. The mean is $M \approx X/5$, the variance $V \approx \frac{X^2}{5} - \frac{X^2}{25} = \frac{4X^2}{25}$. With these approximations, we find $nbp_i = \frac{5}{1+\frac{V_i}{(M_i)^2}} = 1$, which means that the distribution is well matched with a distribution 1,0,0,0,0. In practice, if Y>0 we have $\frac{5}{1+\frac{V_i}{(M_i)^2}} > 1$ and $nbp_i = 2$. This means for our purposes that the best signal is to be extracted from the first two points. If conversely the mutant is neutral, its ratio to wildtype will remain constant through time: X, X, X, X, X. With such a distribution we have $M \approx X$ and $V = 0$, so $nbp_i = 5$. The slope then has to be estimated using all points.

Computing MIC

For MIC determination, we first used wildtype counts at different concentrations to estimate the change in frequency of the various genotypes with antibiotic concentration. We identified a subset of clones which increased in frequency over the wildtype at the highest concentration. That set of clones was used as reference. Similarly to the determination of the number of points used to estimate fitness, we used a moment matching approach to identify the concentrations at which the mutant is eradicated by the antibiotic, mimicking the retention of the mutant with a step down function. In detail, the normalised change in ratio of counts towards the reference set was computed through time, leading for each mutant to a set of values of the form x_{i0}, x_{i1}, x_{i2}, x_{i4}, x_{i8}, x_{i16}, with x_{i0}=1 the initial normalized frequency and the

other values reflecting the relative maintenance or loss of the mutant with increasing concentration. The variance V_i and mean M_i of x_i were used in the following formula to compute a quantitative MIC: $MIC_i = \frac{6}{1+\frac{V_i}{(M_i)^2}}$.

Fitness of stop codons

Many genotypes are corresponding to non-functioning protein due to non-sense mutations or frameshifts. We used the distribution of fitness effects of non-sense mutations (n=2,045) to define a threshold value corresponding to gene inactivation. Estimates of fitness are much more noisy for largely deleterious mutants as counts may come to close to zero at the first time point. The noise prevents precise measures of epistasis in that range. A stringent cut-off of log fitness of -1 was used, meaning that the log-fitness of mutants with value of -1 were assigned that minimal value. This value has a Z score of 4.8 and corresponds to less than 5% chance that one of the 15,000 mutants is assigned as an inactivating mutation while it is not.

Estimating error on fitness measurement

We quantified the error rate in three ways:

a) Biological semi-replicate

The distribution of mutation effects is very sensitive to the antibiotic concentration, which are difficult to reproduce with high precision on a daily bases. We therefore estimate that internal controls such as the variability of fitness estimates based on

barcodes are more appropriate than biological replicates that will represent a slightly different environment. Nevertheless, the day of the experiment, we did a partial replicate using the same exact media and antibiotic dilution. Rather than evolving the population for 30 generations, we evolved them only for 4 in the presence of the same antibiotic concentration, but then contrary to the presented experiment, we allowed a 24 hours recovery growth in the absence of antibiotic. A correlation of 0.982 was found between estimates of fitness in the two conditions.

b) Using Independent barcodes to estimate noise

To use internal controls to estimate noise in the fitness, we exploited the multiple barcodes found for each genotype. After the aberrant barcodes have been filtered out, we then used two approaches. In the first one we estimated fitness from two fully independent sets of barcodes picked randomly. The correlations between the two independent set was 0.942 when we included the non-functional ones (log fitness<-1), and higher than 0.994 when we assigned them a threshold fitness of -1 and 0.976 excluding non-functional mutants.

c) Using bootstrapping of barcodes to estimate noise

To estimate noise more properly, we used a bootstrapping approach. For each genotype having J barcodes, a set of J barcodes sampled with resampling from that set of barcodes was performed 100 times. The mean and variance of fitness among the

bootstrap replicates was then measured. 50% of mutants with log-fitness higher than -1 had a standard deviations of less than 0.01, 70% of less than 0.02, 92% of less than 5%.

However, using either bootstrapping or the two sets of barcodes, some mutants with low barcode counts and a low number of barcodes still exhibited quite a high level of noise. As these may confound the estimate of epistasis, we excluded them from any analysis involving epistasis. We used three criteria to do so. For non-functioning mutants, we excluded mutants with a coefficient of variation (standard deviation divided by mean) of more than 20%. For mutants with fitness higher than -1, we kept the mutants with a coefficient of variation of lower than 12% or of an absolute value lower than 0.03 (the coefficient of variation being infinite for clones of mean close to 0). These empirical criteria were used to balance the quality of the data used to infer epistasis and the intensity of the filtering. These stringent filters reduced the number of mutant from 18,050 to 15,526 (86% retention), but affected mostly non-functioning mutants as the number of functioning mutants decreased from 3,252 to 3,066 (94% retention).

After these filterings, that are fully independent of any measure of epistasis, but just linked to experimental noise, the correlation between the fitness estimates based on the two independent sets of barcodes was 0.954, 0.998 when non-functional were attributed a fitness of -1 and 0.993 excluding the non-functional mutants.

Estimation of the thermodynamic model parameters

To fit the parameters of the thermodynamics model of stability, we had to assign each single mutant a free entropy value, $\Delta\Delta G$, reflecting the mutations' impact on the overall stability of the protein ΔG_0. Though measures of ΔG_0 have been performed *in vitro*, the cellular environment in which the mutants are evaluated could substantially affect this value. We therefore also estimated ΔG_0.

Ideally, for the single mutants, there is a direct connection from estimated fitness to $\Delta\Delta G$. Using such a transformation does improve the signal, but is limited for two reasons. First, for mutants that have fitness lower than -1, this method does not give any $\Delta\Delta G$ value, even if some of them are compensated by stabilizing double mutants. Second, it appears that our protein is quite stable, which results in stabilizing mutants having a very minor effect on fitness as they are on the plateau side of the stability to fitness. For these important mutants, noise in the fitness estimation can result in very important change in the $\Delta\Delta G$ estimation, and infinite values could even be computed if the observed fitness is higher than the maximal one predicted by the model. We therefore decided to use all double mutant fitness effects to estimate the $\Delta\Delta G$ of the single mutants.

We first noticed that when plotting the predictions of the stability model on graphs like Figure 3.2D, the curves connecting the effects of a focal mutation according to the fitness of the diverse genetic background to which it was associated in the double

mutants depended dominantly on ΔΔG of the focal mutant and on the measured fitness of the single mutants. Accordingly, ΔΔG of each mutant could be estimated independently using these graphs. Two further complications limited our ability to perform a simple optimization. First, for deleterious mutants, the curve is almost vertical and any deviation from that curve due to noise result in a very bad fit. Second, as discussed in the main text, we expected and found various ways in which the double mutants may deviate from the stability model. It appeared that these outliers affected significantly the fit if taken into account. To circumvent these two issues, we decided to estimate ΔΔG of each mutant using a goodness-of-fit based on the distance to the theoretical curve and to restrict the fit to the 75% of points closest to the curve. In other words we allowed 25% of the points to deviate from the stability model.

With this strategy we could compute an overall goodness of fit for all mutants and vary in a gradual way the value of ΔG_0 to find the one with the best fit. A value of -2.9 kcal was found to be optimal. Using that value we could compute the ΔΔG of all mutants.

Estimation of the additive part of the fitness component

The stability model predicts that fitness of a single mutant and double mutants are

$$F_i = Log\left(\frac{1 + e^{\frac{\Delta G_0}{RT}}}{1 + e^{\frac{\Delta G_0 + \Delta\Delta G_i}{RT}}}\right) = g(\Delta\Delta G_i)$$

$$F_{ij} = Log\left(\frac{1 + e^{\frac{\Delta G_0}{RT}}}{1 + e^{\frac{\Delta G_0 + \Delta\Delta G_i + \Delta\Delta G_j}{RT}}}\right) = g\left(\Delta\Delta G_i + \Delta\Delta G_j\right).$$

To include an additive component, we redefined log fitness as

$$F_i = h_i + Log\left(\frac{1 + e^{\frac{\Delta G_0}{RT}}}{1 + e^{\frac{\Delta G_0 + \Delta\Delta G_i}{RT}}}\right) = h_i + g(\Delta\Delta G_i)$$

$$F_{ij} = h_i + h_j + Log\left(\frac{1 + e^{\frac{\Delta G_0}{RT}}}{1 + e^{\frac{\Delta G_0 + \Delta\Delta G_i + \Delta\Delta G_j}{RT}}}\right) = h_i + h_j + g\left(\Delta\Delta G_i + \Delta\Delta G_j\right).$$

For one mutation at a time, we then tried various fractions of additive effects, ranging from 120% of the observed fitness to -10% of it. We computed the ΔΔG of all mutants taking into account the additive part in the fit. In other words, for the focal mutation that has the additive part, the graph of figure 2D used to fit $\Delta\Delta G_i$ has to be modified. On the x-axis, we have F_j and on the ordinates $F_{ij} - F_j - h_i$ rather than $F_{ij} - F_j$, as we want on that graph only the stability contribution to fitness. For the other mutants, only the points corresponding to the double mutant including the focal mutation with the additive part has to be modified with abscise $F_i - h_i$ and ordinate$F_{ij} - F_j - h_i$.

For each mutant, we could compute how large goodness-of-fit gain was due to adding an additive part. In all cases showing a large benefit, the additive fraction to be added represented at least 60% of the observed fitness.

Estimation of the error part of the stability model prediction

Using the whole data set, we could estimate an error to the model using a maximum likelihood framework. The ΔΔG values were fixed. For each mutant, we either used the standard deviation σ_{ij} of the fitness estimate or a fitted version of this deviation using a loess regression of the standard error according to the log fitness value. With the fitted error we fitted with single and double mutants independently. Then, we estimated that the deviation of the observed fitness to the one predicted with the stability model resulting from a random deviation due to the experimental measure of the mutant and an overall random deviation from the model. This deviation from the model could be either the same for all pairs of mutations or could be different for residues in contact or not, *ie.* two model parameters. These parameters of noise were optimized using a Monte Carlo Markov Chain with Metropolis Hasting sampling. The two-error model was always much better than the single-error one, and always suggested a higher deviation from the model for residues in contact compared to distant residues.

For the single error model:

$$Log(Lk(\sigma_m)) \sim -\sum_{i,j,i\neq j} Log\left({\sigma_{ij}}^2 + {\sigma_m}^2\right) - \sum_{i,j,i\neq j} \frac{\left(F_{ij} - g(\Delta\Delta G_i + \Delta\Delta G_j)\right)^2}{{\sigma_{ij}}^2 + {\sigma_m}^2}.$$

For the double error model:

$$Log\left(Lk(\sigma_{mp,}\sigma_{md})\right) \sim -\sum_{i,j,i\neq j} Log\left({\sigma_{ij}}^2 + {\sigma_{mp}}^2\delta_{ij} + {\sigma_{md}}^2(1-\delta_{ij})\right) - \sum_{i,j,i\neq j} \frac{\left(F_{ij} - g(\Delta\Delta G_i + \Delta\Delta G_j)\right)^2}{{\sigma_{ij}}^2 + {\sigma_{mp}}^2\delta_{ij} + {\sigma_{md}}^2(1-\delta_{ij})},$$

with $\delta_{ij} = 1$ if the side-chains of residues carrying mutation i and j are less than 6Å away and 0 otherwise.

3.5 Supplementary figures

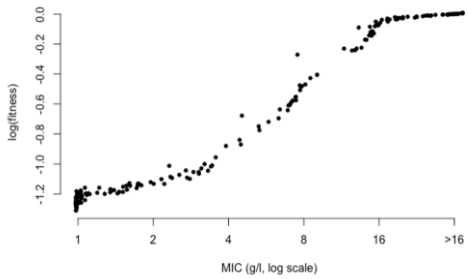


Figure 3.S1. Correlation between MIC of amoxicillin and fitness for single mutants. r = 0.984.

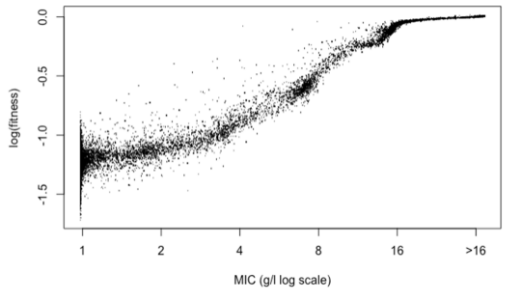


Figure 3.S2. Correlation between MIC of amoxicillin and fitness for double mutants. r = 0.963.

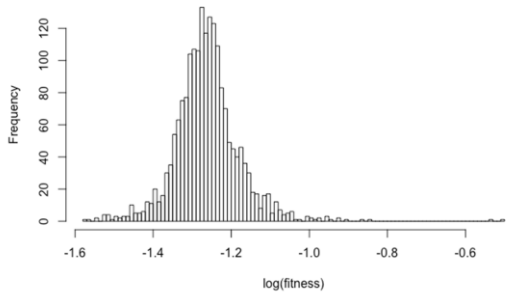


Figure 3.S3. Distribution of fitness effects of stop-codon mutations. They have been used to define a lower threshold for log-fitness: below a value of -1, mutants were considered selectively lethal.

3.6 References

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr. Protoc. Hum. Genet. 76(1):7.20.1-7.20.41

Bank C, Hietpas RT, Jensen JD, Bolon DNA. 2015. A systematic survey of an intragenic epistatic landscape. Mol. Biol. Evol. 32(1):229–38

Bank C, Matuszewski S, Hietpas RT, Jensen JD. 2016. On the (un)predictability of a large intragenic fitness landscape. Proc. Natl. Acad. Sci. U. S. A. 113(49):14085–90

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U A. 102(3):606–11

Chou H-H, Chiu H-C, Delaney NF, Segrè D, Marx CJ. 2011. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. Science. 332(6034):1190–92

Datsenko KA, Wanner BL. 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. 97(12):6640–45

de Visser JA, Elena SF. 2007. The evolution of sex: empirical insights into the roles of epistasis and drift. Nat Rev Genet. 8(2):139–49

De Visser JAG, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. Nat. Rev. Genet. 15(7):480–90

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. Mol. Biol. Evol. 33(1):268–80

Firnberg E, Ostermeier M. 2012. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. PLoS ONE. 7(12):e52031

Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. Nat. Methods. 11(8):801–7

Gros PA, Le Nagard H, Tenaillon O. 2009. The evolution of epistasis and its links with genetic robustness, complexity and drift in a phenotypic model of adaptation. Genetics. 182(1):277–93

Hagen JB. 2000. The origins of bioinformatics. Nat. Rev. Genet. 1(3):231–36

Jacquier H, Birgy A, Nagard HL, Mechulam Y, Schmitt E, et al. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc. Natl. Acad. Sci. 110(32):13067–72

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 332(6034):1193–96

Kowalsky CA, Klesmith JR, Stapleton JA, Kelly V, Reichkitzer N, Whitehead TA. 2015. High-Resolution Sequence-Function Mapping of Full-Length Proteins. PLOS ONE. 10(3):e0118193

Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 344(6191):1519–22

Martin G, Elena SF, Lenormand T. 2007. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat Genet. 39(4):555–60

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31(13):3812–14

Otwinowski J, McCandlish DM, Plotkin JB. 2018. Inferring the shape of global epistasis. Proc. Natl. Acad. Sci. 201804015

Otwinowski J, Wilke C. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. Mol. Biol. Evol.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. 75(23):7537–41

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl. Acad. Sci. 106(1):67–72

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-Term Dynamics of Adaptation in Asexual Populations. Science. 342(6164):1364–67

Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U A. 108(24):9916–21

4 Flux, toxicity and protein expression costs shape genetic interaction in a metabolic pathway

Authors: Harry Kemble[1,2], Catherine Eisenhauer[1], Alejandro Couce[1], Audrey Chapron[1], Mélanie Magnan[1], Gregory Gautier[3,4], Hervé Le Nagard[1], Philippe Nghe[2], Olivier Tenaillon[1]

Affiliations: [1]Infection, Antimicrobials, Modelling, Evolution, INSERM, Unité Mixte de Recherche 1137, Université Paris Diderot, Université Paris Nord, 75018 Paris, France.

[2]École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI Paris), PSL Research University, UMR CNRS-ESPCI CBI 8231, 10 Rue Vauquelin, 75231 Paris Cedex 05, France.

[3]Centre de Recherche sur l'Inflammation, INSERM, UMRS 1149, 75018 Paris, France.

[4]Laboratoire d'Excellence INFLAMEX, Université Paris Diderot, Sorbonne Paris Cité, 75018 Paris, France.

Abstract: Our ability to predict the impact of mutations on traits relevant for disease and evolution remains severely limited by the dependence of their effects on the genetic background and environment. Even when molecular interactions between genes are known, it is unclear how these translate to organism-level interactions between alleles. We therefore characterized the interplay of genetic and environmental dependencies in determining fitness by quantifying ~4,000 fitness interactions between expression variants of two metabolic genes, in different environments. We detect a remarkable variety of environment-dependent interactions, and demonstrate they can be quantitatively explained by a mechanistic model accounting for catabolic flux, metabolite toxicity and expression costs. Complex fitness interactions between mutations can therefore be predicted simply from their simultaneous impact on a few connected molecular phenotypes.

4.1 Introduction

Despite its centrality to medical and evolutionary genetics, our ability to predict the impact of mutations on even the apparently simplest of organismal traits (1–8), let alone complex ones (9), remains minimal. Three of the main factors proposed to account for this "missing heritability" (9) are: the large number of possible alleles at any locus, each having a potentially different impact on a gene's function; interaction between alleles at different loci (intergenic epistasis), such that their combined effect is not simply the sum of their individual effects; and interaction between genotype and environment, such that different genotypes respond to the environment in different ways (1–9). A promising inroad is the increasingly refined characterization of molecular interaction networks enabled by –omics approaches (10). Metabolic networks are the best-characterized of these, and are of great practical interest for medicine and engineering, but even for metabolic genes it remains unclear how functional interactions at the molecular level translate to allelic interactions at the level of integrated traits relevant for disease, industry and adaptation (11).

We therefore developed an experimental system with which to systematically quantify the fitness interactions occurring between many alleles of two metabolic genes from the same pathway. Further, the design enabled us to probe the dependence of these interactions on environmentally modulated gene expression, a common non-genetic mechanism for the modification of physiological traits (5, 12).

Our system was composed of the genes (*araA* and *araB*) encoding the enzymes responsible for the first two steps of the well-studied Escherichia coli L-arabinose-utilization pathway (13): L-arabinose isomerase (AraA) and L-ribulokinase (AraB), who together transform the sugar, L-arabinose, into the intermediate, L-ribulose-5-phosphate (Fig. 4.1A). L-ribulose-5-phosphate enters the pentose phosphate pathway (PPP) of central metabolism via further enzymatic reactions, ultimately supporting cell growth, but like many intermediates (14, 15), its accumulation is toxic, retarding growth (16). Environmental modulation of gene expression was achieved by placing each of the two genes under an independent, trans-regulated chemically-inducible promoter.

For each promoter, 36 single-base variants were constructed, along with the initial "wildtype" sequence, and combined with all variants of the other promoter (Fig. 4.1B). The organismal phenotype, competitive fitness, was then measured for the entire set of 1,369 genotypes under three different inducer concentration combinations (Figs. 4.1C-D). Fitness was measured by tagging the mutant library with unique DNA barcodes (tens to thousands per genotype) (Figs. 4.S1-2), culturing the pooled library for ~30 mean generations, and tracking barcode frequencies over time with Next-Generation Sequencing (Fig. 4.S3). The barcodes act as internal replicates for every genotype, enabling precise fitness estimates at high-throughput (log relative fitness, F^{rel}, median

standard deviation of 0.0011 for single mutants and 0.0047 for double mutants; Fig. 4.S4).

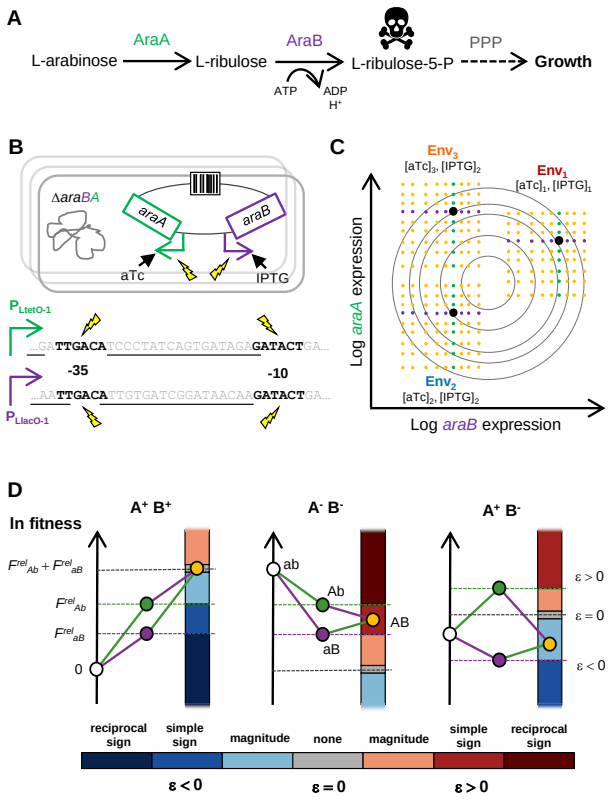


Fig. 4.1. Quantitative mapping of fitness interactions between expression variants of two metabolic genes in expression-modifying environments. (**A**) L-arabinose pathway of *E. coli*. (**B**) *araA* and *araB* were placed under the control of inducible promoters, making their expression sensitive to the concentration of their respective inducers, anhydrotetracycline (aTc) and isopropyl β-D-1-thiogalactopyranoside (IPTG). A barcoded library of mutant promoter combinations was constructed, with mutations targeting the -35 and -10 RNA-polymerase binding hexamers (black letters). Underlined bases are annotated repressor binding sites. (**C**) Competitive fitness was measured under different inducer concentrations defining three environments. $P_{LtetO-1}$ single mutants – green; $P_{LlacO-1}$ single mutants – purple; double mutants - orange. Contours are hypothetical fitness isoclines. (**D**) Epistasis was quantified for all mutant promoter pairs across environments. Epistasis can be categorized as either magnitude or sign type. Sign epistasis is further categorized as simple (effect of one mutation changes sign in presence of the other) or reciprocal (effects of both mutations change sign in the presence of the other). Capitalized letters represent mutant alleles of $_{PLtetO-1}$-*araA* and $P_{LlacO-1}$-*araB*. Superscript plus and minus denote that individual alleles are beneficial or deleterious, respectively.

4.2 Results

The overall distribution of fitness effects depended critically on the inducer environment, *ie.* the *trans*-regulatory input (Fig. 4.2A; Fig. 4.S5A; Data S1). The proportion of beneficial effects varied from 88% in Env_1 (median $F^{rel} = 0.12$) to 51% in Env_3 (median $F^{rel} = -0.03$) and 12% in Env_2 (median $F^{rel} = -0.12$). Further, the correlation of fitness effects between environments ranged from strongly positive (Env_1-Env_3, Pearson's $r = 0.74$, $p < 2.2x10^{-16}$) to weakly negative (Env_1-Env_2, Pearson's $r = -0.11$, $p = 1x10^{-4}$) (Fig. 4.S5B), demonstrating that fitness in one environment can be an extremely poor predictor of fitness in other environments due simply to expression differences. At the level of individual alleles, all but one had changing patterns of effects across environments (Fig. 4.2B). In some environments, they were universally beneficial or deleterious across genetic backgrounds, and in others they switched between being beneficial and deleterious depending on the allele at the second promoter. This pervasive and inconsistent variability poses a clear challenge for the prediction of mutation effects.

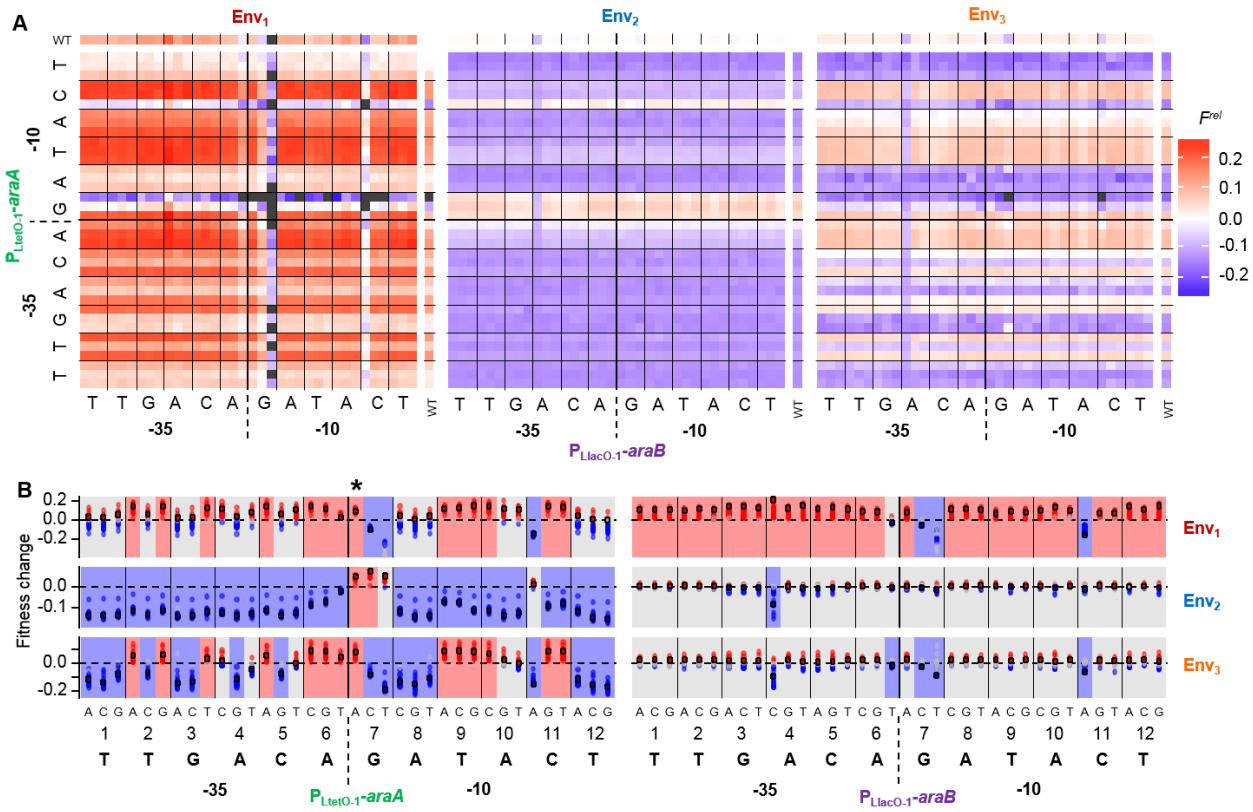


Fig. 4.2. Fitness effects of promoter mutations across backgrounds and environments. (**A**) Genotypes are colored according to the natural logarithm of their fitness relative to the wildtype (F^{rel}). Grey denotes unquantifiable fitness effects. Letters show wildtype bases, and the 3 mutations at each position are ordered alphabetically, as in **B**. Single promoter mutants make up the right-most column (*araA*) and top row (*araB*). Inducer concentrations were: 20 ng/ml aTc and 30 µM IPTG (Env_1); 5 ng/ml aTc and no IPTG (Env_2); 200 ng/ml aTc and no IPTG (Env_3). (**B**) Fitness changes when an allele of one promoter is added to alleles of the second promoter. Large points indicate the "background" promoter is wildtype. Red, blue and grey points indicate positive, negative and non-significant fitness changes, respectively. Red, blue and grey rectangles indicate, in that environment, an allele can be beneficial but never deleterious, deleterious but never beneficial, or both beneficial and deleterious. G7A of $P_{LtetO-1}$-*araA* (*) is the only allele conferring a qualitatively consistent fitness effect (beneficial) across all backgrounds and environments.

To further characterize how the effects of mutations in one gene depended on the allele present at the other gene, we computed epistasis (*17*) for all mutation pairs in each environment. Epistasis evaluates quantitatively and qualitatively how the log fitness of

a double mutant deviates from the sum of that of the constituent single mutants (Fig. 4.3A, Fig. 4.S6A). Epistasis was found to be pervasive (89%, 39% and 81% of pairs in Env_{1-3}, respectively), heterogeneous and environment-dependent. A trend of antagonism reported for several other systems (*18*) was recovered between pairs of individually beneficial (negative epistasis in 89%, 72% and 100% in Env_{1-3}, respectively) and individually deleterious (positive epistasis 100% (1/1), 97% and 98%, respectively) mutations, while interactions between a beneficial and a deleterious mutation could be mostly positive or mostly negative, depending on the environment and on which gene carried the beneficial/deleterious mutation. This epistatic diversity extended to individual mutation pairs, with more than 20% interacting both positively and negatively across environments (Figs. 4.S6B-C). Notably, sign epistasis, an extreme interaction which occurs when the sign of a mutation effect changes in the presence of a second mutation (Fig. 4.1D), represented 31% of significant interactions in Env_1, 17% in Env_2 and 34% in Env_3.

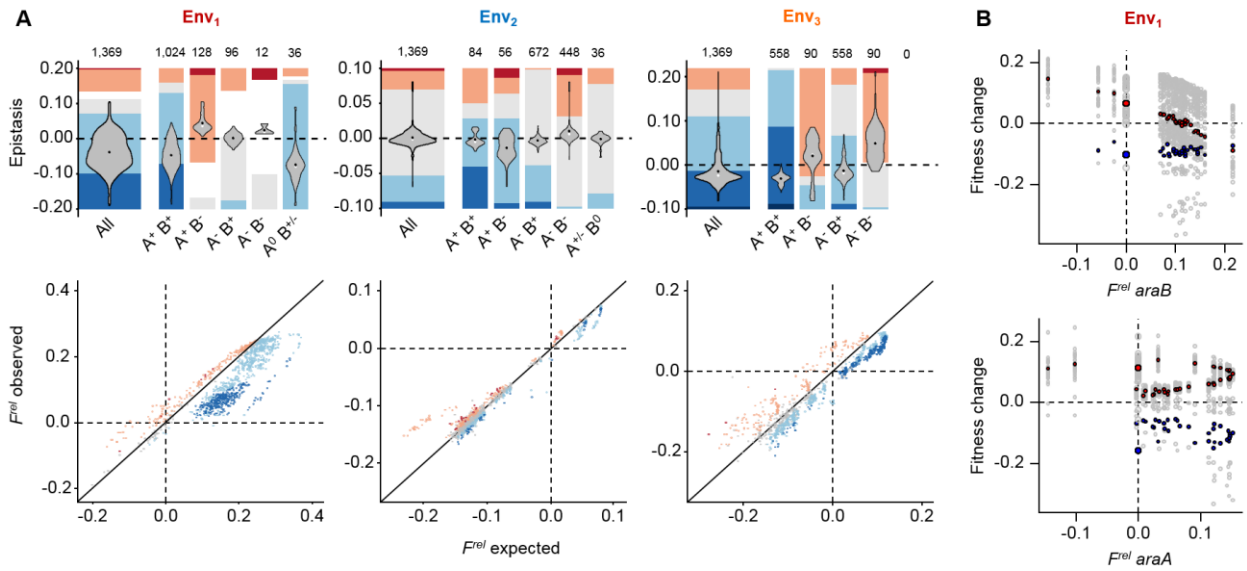


Fig. 4.3. Strength, types and trends of epistasis across environments. (**A**) Violins show epistasis for different kinds of mutation pairs (white point - median; black point - mean). Mutation pairs may contain mutations that are individually both beneficial (A^+ B^+), both deleterious (A^- B^-) or mixed (A^+ B^- and A^- B^+), or one of which confers an undetectable effect (A^0 $B^{+/-}$ and $A^{+/-}$ B^0). The number of each such pair is provided. Stacked bars show fractions of different epistasis types (colors as Fig. 4.1D, with white where epistasis could not be computed). Scatterplots show fitness of double mutants against that expected if mutation effects combined additively. Points colored as in Fig. 4.1D. (**B**) Relationship between background fitness and the fitness change induced by mutations in the second promoter, in Env_1. Top: *araA* promoter mutations added to existing *araB* promoter mutations; bottom: inverse case. Colored points highlight particular alleles. Top: $P_{LtetO-1}$-*araA* alleles T2C (red) and G7C (blue). Bottom: $P_{LlacO-1}$-*araB* alleles T1A (red) and C11A (blue). Large points show effects in the wildtype background.

Confronted with such a variety of interactions, we asked whether they might be understood simply in terms of the quantitative fitness effects of the interacting mutations, as has been found for some other mutation sets (*19*). We found that the effects of individual mutations were weakly predictive of the type and value of epistasis they exhibited with mutations at the second promoter (Fig. 4.3A scatterplots). In all environments, there was a significantly negative correlation between the sum of

individual fitness effects and the value of epistasis (Pearson's r = -0.36, -0.37, -0.51 in $Env_{1\text{-}3}$, respectively; $p < 2.2x10^{-16}$ for all), a trend of diminishing returns that appears common across experimental systems (*19–22*) (Fig. 4.S7A). However, when the two genes were considered separately, the relationship between individual fitness effects and epistasis was found to be markedly different between *araA* and *araB*: the negative correlation was stronger for $P_{LtetO\text{-}1}$-*araA* mutations being added to existing $P_{LlacO\text{-}1}$-*araB* mutations than for the inverse case (Figs. 4.S7B-C; Pearson's r = -0.67, -0.73, -0.63 in $Env_{1\text{-}3}$, $p < 2.2x10^{-16}$ for all, *vs.* 0.12, -0.20 and -0.34, $p < 1.6x10^{-5}$ for all), in which the correlation can even be positive, an extremely rare trend in existing studies (*19*). Moreover, we found that the average trend was in some cases strikingly non-monotonic (Figs. 4.S7B-C), revealing that different alleles of a particular promoter can cause similar fitness changes on their own but interact very differently with alleles at the second promoter.

The relationship between individual mutation effects and epistasis was further complicated by the fact that it could be different for different alleles of the same promoter. For example, in Env_1, the numerous beneficial $P_{LtetO\text{-}1}$-*araA* mutations caused the average negative trend with $P_{LlacO\text{-}1}$-*araB* background fitness, while the rare deleterious $P_{LtetO\text{-}1}$-*araA* mutations showed no such trend (Fig. 4.3B, top panel). For individual $P_{LlacO\text{-}1}$-*araB* mutations in $P_{LtetO\text{-}1}$-*araA* backgrounds, the relationship was consistently non-monotonic, but had a different average direction for individually

beneficial or deleterious alleles (Fig. 4.3B, bottom panel). Moreover, the trend for a given allele could vary greatly with the environment (Figs. 4.S7B-C). These results demonstrate that genes interacting simply through their common participation in a linear pathway can exhibit complex, allele- and environment-dependent trends of epistasis. The smooth patterns exemplified by Fig. 4.3B, however, suggest that they may in principle be understood from an underlying phenotypic mechanism.

To this end, we constructed a quantitative model of the metabolic pathway, where fitness results from a balance between the benefit of flux (*23*) and the costs of intermediate toxicity (*14*, *24*, *25*) and AraA and AraB protein expression (*26–28*). Log fitness was computed as $F = \left(\omega + u\varphi - \frac{v}{1/\eta - \varphi}\right)\left(1 - \theta_A A - \theta_B B\right)$, with ω a basal growth rate, u and v terms describing the catabolic benefit and toxicity cost of pathway flux (φ), A and B the cellular activity of the two enzymes, and θ the cost of enzyme expression. Flux depended on AraA and AraB activities as $\varphi = \frac{1}{1/A + 1/B + \eta}$ (*25*, *29*).

Each promoter mutation was then characterized as a change in the activity (*via* expression) of AraA or AraB. Because most mutations lay outside of the repressor binding sites governing promoter inducibility (Fig. 4.1B), the fold-change in activity caused by each mutation was kept constant across inducer environments. Parameters describing the fitness function, wildtype activities in the 3 environments and

expression effects of individual mutations were then optimized to fit the observed data (Data S2; Fig. 4.S8A).

The fitted model is in excellent agreement with our data, yielding r^2 values of 0.98 between experimental and simulated fitness effects and 0.82 between experimental and simulated epistasis coefficients (Fig. 4.4A-B; Fig. 4.S8B-C; see Fig. 4.S9 for more minimal models). Notably, the model is capable of recapitulating the diverse and complex trends of epistasis seen in the data. In particular, we find that the non-monotonic relationships between single-mutant fitness and the fitness impact of alleles at the second promoter are well explained by the single mutants lying at two sides of a phenotypic optimum (Fig. 4.4B). Such overshooting, which is also the cause of sign epistasis (Fig. 4.4C) (*30*), is relatively common in our dataset, mostly because L-ribulose-5-phosphate toxicity results in an optimum in the flux-fitness relationship (*24*, *25*) (Fig. 4.S10). Two alleles of the same gene may thus result in similar fitness changes individually but cause substantially different expression levels and fluxes, resulting in different responses to mutations at the second gene. This is principally due to enzymes possessing different degrees of flux control on each side of the optimum, with lower levels of one resulting in the second having less control.

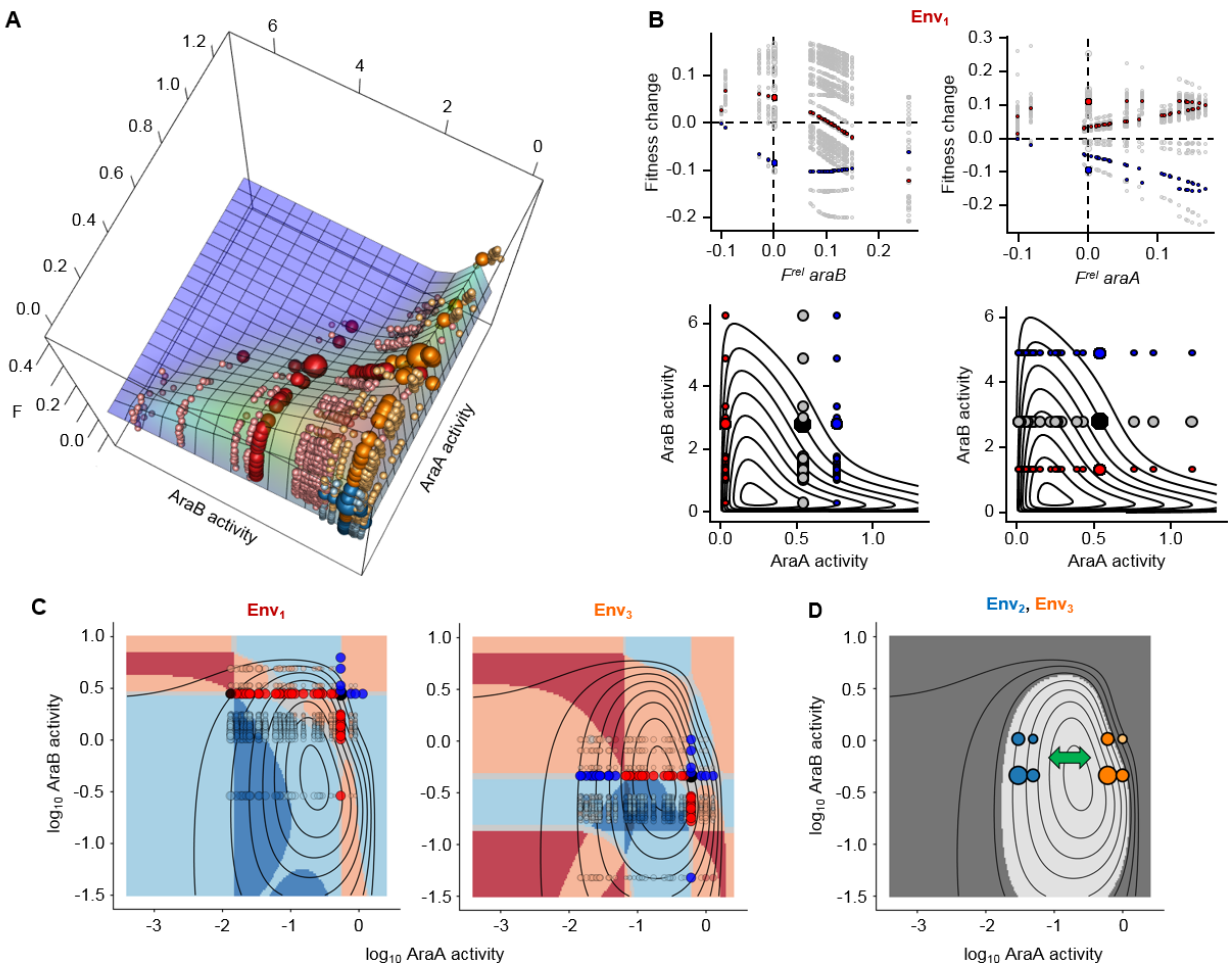


Fig. 4.4. Mechanistic basis of heterogeneous, environmentally dependent epistasis. (**A**) Fitted activity-fitness model. Spheres are positioned according to predicted activity levels and observed F^{rel} ($Env_{1\text{-}3}$ – red, blue, orange). Three largest spheres are wildtype, intermediate-sized spheres are single mutants, small pale spheres are double mutants. (**B**) Upper plots recapitulate Fig. 4.3B. Lower plots show highlighted genotypes within fitness landscape (black point is wildtype; other large points are single mutants, grey for the gene considered as carrying the "background" alleles). (**C**) Fitness surface on log activity scale, colored by predicted intergenic epistasis type (colors as Fig. 4.1D; determined as non-significant (grey) if magnitude < 0.005). Large black point is wildtype. Smaller, opaque blue, red and black points are single mutants, colored by observed F^{rel} (deleterious, beneficial and neutral, respectively). Transparent points are double mutants, colored by observed epistasis type and sized by epistasis strength. (**D**) Dark grey marks area below a hypothetical disease threshold (40% of maximum fitness). Points are four genotypes in Env_2 (blue) and Env_3 (orange): wildtype (largest), C11A of $P_{LtetO\text{-}1}$-*araA* and G7T of $P_{LlacO\text{-}1}$-*araB* (intermediate size), and the resulting double-mutant (smallest). Green arrow represents a change in activity levels caused by non-genetic factors like ageing or environment. A disease state results here from one combination of alleles and environment (pale orange).

4.3 Discussion

The model reveals how the biology underlying a linear pathway can result in heterogeneous, environmentally dependent intergenic interactions. When fitness depends solely on flux (*23*, *25*), the nature of epistasis should be guaranteed by pathway topology alone (*25*). Under the slightly more complex selection pressure resulting from metabolite toxicity (*24*, *25*) and gene expression costs, however, interactions can be both positive and negative. We find that epistatic categories form several localized zones over the fitness landscape, their size and position dependent on the wildtype phenotype, controlled here by the environment (Fig. 4.4C; Fig. 4.S11). Encouragingly, these zones are generally large and orderly enough to be predictable, but only through knowledge of the underlying landscape and the position of the relevant genotypes within it.

The importance of this knowledge becomes immediately apparent when considering the existence of a disease threshold (Fig. 4.4D). The two alleles shown can lead to disease, but only when they co-occur, and only in one particular environment. The model thus provides a mechanism by which potential physiological defects can be manifested, aggravated or alleviated by particular combinations of alleles and environments (*1*–*7*, *9*). Insight into intergenic fitness landscapes for other biological systems, and for genes connected by more complex topologies, will be indispensable for progress across medicine, bio-engineering and evolution.

4.4 Methods

General microbiology and molecular biology

Lysogeny Broth (LB) powder, agar, salts, sugars, growth supplements, antibiotics and inducers were all purchased from Sigma-Aldrich. Bacteria were cultured in LB, unless otherwise stated. Liquid LB was the standard Lennox formulation, except for when blasticidin-S was included, in which case the Luria low-salt formulation (0.5 g/L NaCl) was used. LB-agar always contained the Luria low-salt formulation. M9 base medium consisted of 1X M9 salts supplemented with 1mM $MgSO_4$ and 100 µM $CaCl_2$. Unless otherwise stated, L-arabinose was used at a concentration of 0.03% w/v. Ampicillin (amp) was used at 100 µg/ml, chloramphenicol (cm) at 10 µg/ml, streptomycin (str) at 50 µg/ml, blasticidin-S (bsd) at 100 µg/ml and erythromycin (erm) at 20 µg/ml. Bacterial cultures were grown at 37°C (with shaking at 200 rpm for liquid cultures; Multitron, Infors HT), unless otherwise stated, and culture stocks were stored at -80°C in LB with 40% glycerol. For electroporation, DNA was added to 50 µl homemade electro-competent cells (unless otherwise stated), transferred to a 1mm-gap electroporation cuvette (VWR) and submitted to a pulse of 1,800 V (Electroporator 2510, Eppendorf). Cells were immediately transferred to fresh LB for recovery at 37°C (unless otherwise stated) with shaking for 30-90 minutes, before being plated on the appropriate selective media and left to grow overnight.

All enzymes and molecular biology reagents were purchased from NEB, unless otherwise stated. Primers were purchased from IDT or Eurofins, and designed with the help of Primer3 (*31*). For sensitive applications like barcoding and NGS library preparation, primers were ordered HPLC-purified, otherwise they were ordered desalted. UltraPure agarose was supplied by Invitrogen, and all agarose gels were stained with SYBR Safe (Thermo Scientific) and visualised with a GelDoc XR+ imager (Bio-Rad). The GeneRuler 1kb Plus ladder (Thermo Scientific) was used for DNA fragment size estimation.

All plasmids used in this study, excluding the mutant library, are detailed in Table 4.S1. DNA fragments used in cloning are detailed in Table 4.S2. Primers, excluding those used for promoter mutagenesis, are provided in Table 4.S3. All strains are detailed in Table 4.S4. Primers used in promoter mutagenesis are provided in Table 4.S5.

Plasmid construction

Our library creation strategy depended on two plasmids, pKH1511c and pKH1511d, which were created in this study. pKH1511c serves as the library "backbone", carrying all the necessary elements of the final plasmid library except for the $P_{LtetO-1}$ and $P_{LlacO-1}$ promoters destined to drive *araA* and *araB* expression, respectively: *p15A* origin-of-replication, *lacI-tetR* repressor cassette (for inducibility of $P_{LtetO-1}$ and $P_{LlacO-1}$), *araA* and *araB*. *araA* and *araB* ORFs (with their upstream ribosome binding site-containing

regions) are divergently oriented (with each followed by an artificial transcriptional terminator), and are separated by 2 restriction sites to allow easy insertion of divergently oriented $P_{LtetO-1}$ and $P_{LlacO-1}$ promoters. pKH1511d serves as a template for amplification of a *bsd* blasticidin S-resistance cassette with primers containing the $P_{LtetO-1}$ and $P_{LlacO-1}$ variant sequences, allowing their eventual insertion into pKH1511c (Figure 4.S1). pKH1511d replication is pir-dependent, abolishing the occurrence of false-positive colonies caused by PCR template carryover during library cloning. Plasmids, DNA fragments, PCR primers and bacterial strains used in the construction of these two plasmids are detailed in Tables 4.S1-4, respectively, and the detailed cloning methods follow.

The DNA fragments used to construct pKH1503a, pKH1511c and pKH1511d come from either PCR amplification or from direct restriction digestion of purified plasmid DNA, and were joined by either standard restriction-ligation or by Gibson Assembly (*32*) (in which case, overlaps of ~40 nucleotides were used). PCR amplifications were all performed with Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific) in its High-Fidelity buffer, following the manufacturer's recommendations. Restriction enzymes were used according to the manufacturer's instructions. When found necessary to reduce the occurrence of false-positive colonies, DNA was treated with calf intestinal alkaline phosphatase (to reduce vector self-ligation) and/or DpnI (to digest PCR template). After PCR amplification and/or digestion, DNA fragments

were either verified by electrophoresis and column-purified (QIAquick PCR Purification Kit, QIAGEN) or, when necessary, gel purified (QIAquick Gel Extraction Kit, Qiagen). Gel-purification was always followed by a 2nd clean-up (QIAquick PCR Purification Kit, QIAGEN) to improve DNA quality for ligation. For gel extractions, agarose gels were stained with SYBR Safe (Thermo Scientific), and DNA was visualised with blue light to avoid UV-induced DNA damage (Blue Transilluminator, Pearl Biotech). A NanoDrop ND-1000 spectrophotometer (Thermo Scientific) was used to determine DNA concentration for all fragments prior to ligation/Gibson Assembly. Standard ligation and Gibson Assembly were performed using T4 ligase and Gibson Assembly Master Mix (NEB), respectively, according to the manufacturer's recommendations (T4 ligase was then inactivated by heating at 65°C for 10 mins). In both cases, DNA was subsequently microdialysed against water for > 30 mins (MF-Millipore, Merck), and 1-5 µl were electroporated into 50 µl electrocompetent cells. DH5α ΔaraBA was used as the cloning strain except when the plasmid was pir-dependent, in which case PIR1 was used. After electroporation, cells were recovered in 1 ml LB for 30-90 mins at 37°C with shaking at 200 rpm, plated on LB-agar in the presence of the antibiotic indicated in Table 4.S1 and incubated overnight at 37°C. Plasmid DNA was purified from several colonies (QIAquick PCR Purification Kit, QIAGEN) and verified by both restriction analysis and Sanger sequencing of the insert region.

Strain engineering and adaptation

The final library host strain, *E. coli* MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo} (Table 4.S4), was originally designed to possess a rewired D-arabinose metabolism (*33–35*), in which *araB* (but not *araA*) participates. D-arabinose was not used in this study, however, so this feature (D-ara$^{+/evo}$ Δ*fucK*) is not relevant here. In addition, *araA* and *araB* ORFs were removed from the chromosome, to allow them to be expressed exclusively from plasmids (the 3rd gene of the *araBAD* operon, *araD*, was kept on the chromosome under the control of its native L-arabinose-responsive promoter, as were the transcriptional regulator gene, *araC*, and the transporter genes, *araE*, *araFGH* and *araJ*; given the all-or-nothing response of the positive feedback loop governing L-arabinose uptake, all these genes are expected to be maximally induced by internal L-arabinose by the time of fitness measurement (*36*)). Further, *lacIZYA* was replaced by a *cat* chloramphenicol-resistance cassette. This allows the use of IPTG to control the artificial promoter, $P_{LlacO-1}$, in the absence of any effects resulting from induction of the native *lac* operon, and the absence of *lacY* also causes this control to be titratable rather than all-or-nothing (*37*). Finally, this strain was transformed with plasmid pKH1503a (which carries an *araBA* cassette under the control of $P_{LlacO-1;}$ Table 4.S1) and briefly adapted to M9 with alternating D-/L-arabinose (see above) in the presence of a low concentration of IPTG. This adaptation step was included to allow

fixation of any mutations conferring a very high fitness advantage to our engineered strain in our approximate experimental conditions, to avoid them interfering with mutant library competition experiments. Detailed strain engineering methods follow.

Details of the final library host strain, and all intermediates used in its creation, are provided in Table 4.S4. Gene knockouts were performed using the method of Datsenko and Wanner (*38*). The relevant strain was made electrocompetent, electroporated with 10 ng plasmid pKD46 DNA, and transformants were selected on LB-agar with 100 µg/ml ampicillin at 30°C. Several colonies were then re-isolated under the same conditions. The cat chloramphenicol-resistance cassette was PCR-amplified from pKD3 (*38*) using primer pairs KO-araBA-fwd/KO-araBA-rev for araBA, KO-lacIZYA-fwd/KO-lacIZYA-rev for lacIZYA and KO-fucK-fwd/KO-fucK-rev for fucK, and a 2:1 mix of GoTaq/Pfu DNA polymerases (Promega). PCR products were verified by 1% agarose gel electrophoresis, column-purified (QIAquick PCR Purification Kit, QIAGEN) and spectrophotometrically quantified (NanoDrop ND-1000). A pre-culture of a single pKD46-transformed colony was grown overnight (LB-amp) at 30°C and then diluted 100x into LB-amp with 0.2% L-arabinose and grown at 30°C to an OD_{600nm} of ~0.7 (BioMate 3S, Thermo Scientific; 3-5 hours). The culture was made electrocompetent, electroporated with ~200 ng of the purified PCR product, and recombinants were selected on LB-agar with 10 µg/ml chloramphenicol at 37°C, for curing of pKD46. Several colonies were then re-isolated under the same conditions, and

tested in parallel for pKD46 curing by plating on LB-amp and checking for colonies after an overnight growth at 30°C. Several of the re-isolated colonies were verified by colony-PCR, using 3 primer pairs for each knockout (*38*). The gene-specific primers are verif-araBA-fwd/verif-araBA-rev for araBA, verif-lacIZYA-fwd/verif-lacIZYA-rev for lacIZYA and verif-fucK-fwd/verif-fucK-rev for fucK, and the common cat primers are c1 and c2 from reference (*38*). For each knockout, the 3 primer pairs were: gene-specific fwd/gene-specific rev, gene-specific fwd/c1 and gene-specific rev/c2. GoTaq DNA polymerase (Promega) was used for amplification, following the manufacturer's recommendations, and PCR products were analysed by agarose gel electrophoresis (1.5%). In the case of *araBA* and *fucK*, we wished to retain function of the remaining genes in their respective operons, and so the cat cassette was removed as described in reference (*38*). For this, a pre-culture of a single recombineered colony was grown overnight (LB-cm, 37°C) and then diluted 100x into LB-cm and grown at 37°C to an OD_{600nm} of ~0.7 (BioMate 3S, Thermo Scientific; 2-4 hours). The culture was made electrocompetent, electroporated with 10 ng plasmid pCP20 DNA, and transformants were selected on LB-agar with 100 µg/ml ampicillin at 30°C. Several colonies were then re-isolated under the same conditions, and then again in the absence of ampicillin at 42°C, to cure pCP20 (*38*). Finally, several colonies were streaked in parallel on LB (37°C, purification), LB-cm (37°C, verify *cat* loss) and LB-amp (30°C, verify pCP20 loss). The loss of the *cat* cassette through FRT recombination was verified molecularly for several clones by colony-PCR, using the same primer pairs and conditions described

above for *cat* insertion verification. The PCR products resulting from amplification with the gene-specific primer pairs were also Sanger-sequenced (GATC; using the amplification primers) as a final verification.

Adaptations were performed as described in Table 4.S4. For the initial adaptation step, pre-cultures were grown overnight in LB, washed twice in an equal volume of M9, and 1 ml washed cells were diluted in 100 ml of the appropriate adaptation media. Once growth became apparent, cultures were serially transferred in a volume of 20 ml, being left to grow for ~24 hours between each transfer, at which point they were diluted ~100x into fresh media. After adaptation, colonies were isolated on agar plates containing the same media used for adaptation.

To cure the plasmid from MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo}, a pre-culture was grown overnight in LB-cm, and dilutions were plated on LB-cm with 2% ribitol and 200 µM IPTG. IPTG induces *araBA* from the plasmid, and AraB converts ribitol to the toxic compound ribitol phosphate (*39*), rendering plasmid-harbouring cells unable to grow. Several colonies were tested and confirmed for plasmid loss by streaking on LB-str and by colony-PCR (primers oKH150401c/oKH150202d, GoTaq (Promega)), with comparison to control colonies grown in the absence of ribitol. The final plasmid-less host strain was also tested once more for its marker-less *ΔaraBA* and *ΔfucK* deletions using colony-PCR (primer pairs verif-araBA-fwd/verif-araBA-rev and verif-fucK-fwd/verif-fucK-rev, as above).

Library creation strategy

On the evolutionary scale, direct changes in the total cellular activity of a particular enzyme can occur through either regulatory mutations, which alter the concentration of active enzyme, or structural mutations, which can effect both active enzyme concentration and kinetic parameters. A common target of regulatory mutations is the promoter (*40*), which controls a protein's expression level by determining transcription rate, and we decided to focus on promoter mutations in this study. We first placed *araA* and *araB* under the control of the well-known artificial, chemically-inducible promoters, $P_{LtetO-1}$ and $P_{LlacO-1}$, developed by Lutz and Bujard (*41*). They are each regulated by a single transcription factor (*tetR* repressor for $P_{LtetO-1}$ and *lacI* repressor for $P_{LlacO-1}$), and can be specifically induced to different levels by addition of a small, non-metabolisable compound (aTc for $P_{LtetO-1}$ and IPTG for $P_{LlacO-1}$). We focussed mutagenesis on the RNA polymerase-binding sites (-35 and -10 hexamers) of the two promoters, as these sites are known to be the most significant determinants of expression level in the core promoter (*42*, *43*). Conveniently, these sites are identical between $P_{LtetO-1}$ and $P_{LlacO-1}$, coming from phage lambda P_L in both cases (*41*). For each promoter, we constructed all possible single-bp substitutions over this 12bp region (36 mutants for each promoter), along with the wildtype sequence. All 37 sequence variants of the two promoters were combined together, resulting in a plasmid library

containing: all 1,296 double-promoter mutants, all 36 single-promoter mutants for each promoter (one promoter is mutated, the other is wildtype) and the full wildtype (both promoters are wildtype). The majority of mutations in the RNA polymerase-binding sites are expected to have little or no effect on repressor binding, and their relative effect on expression should be similar across different inducer concentrations (*44*, *45*). However, one of the -10 bases on $P_{LtetO-1}$ overlaps with a *tet* operator, and three of the -35 bases on $P_{LlacO-1}$ are expected to overlap with a *lac* operator (*41*) (Fig. 4.S1), meaning that the effect on expression of mutations at these positions could depend strongly on inducer concentration (*46*).

The overall structure of the plasmid on which the library is based is shown in Fig. 4.S1. *araA* and *araB* are divergently expressed from $P_{LtetO-1}$ and $P_{LlacO-1}$ promoters, respectively. These two promoters are separated from each other by a short *bsd* blasticidin S resistance cassette (*47*), in order to reduce any physical interactions between them. The presence of a resistance cassette between the promoters also considerably increased cloning efficiency, as explained below, and *bsd* in particular was chosen for its small size (396 bp ORF), making it possible to sequence both promoters on a single amplicon using paired-end Illumina technology (Fig. 4.S1). The promoters' repressors, *tetR* and *lacI*, were included on the plasmid.

Plasmid molecules were also intergenically tagged with unique DNA barcodes, similarly to reference (*48*) (Fig. 4.S1). These were used to help overcome the problem of PCR

and sequencing errors and to increase the precision of mutant fitness estimates by providing many independent frequency trajectories for each mutant (Figs. 4.S2-4). The barcodes thus also allowed us to account for anomalous lineages containing off-target mutations (present in the initial library) and *de novo* mutations (arising during competition assays). They consist of 20 random nucleotides, split into 4 blocks of 5 (*49*) to avoid the creation of restriction sites used in a later sequencing step: $N_5ATN_5ATN_5ATN_5$. Barcodes were inserted downstream of the *lacI-tetR* cassette, far from the $P_{LtetO-1}$ and $P_{LlacO-1}$ promoters, to avoid any effects on *araA* and *araB* expression, and so are expected to be effectively neutral for fitness (Fig. 4.S1). Care was taken throughout to avoid loss of library complexity (Fig. 4.S2), and quality controls were employed at each step of library construction.

The pooled plasmid library was constructed using standard restriction-ligation cloning (Fig. 4.S1). Due to their short length, promoter sequences could be introduced facing outwards on the 5' ends of PCR primers that were used to amplify a *bsd*(*47*) blasticidin S-resistance cassette from plasmid pKH1511d ($P_{LtetO-1}$ on forward primers and $P_{LlacO-1}$ on reverse primers). This was done using primer pools with randomised nucleotides at each of the 12 target positions for each promoter. The primers also contained restriction sites on their 5' extremities, allowing the resulting amplicon pool to be ligated into the library backbone, pKH11511c, in the desired orientation. The resulting plasmid library was transformed into DH5α Δ*araBA* and colonies were

selected on blasticidin S. This strategy ensured that the occurrence of false-positive colonies from undigested or self-ligated vector was negligible, as a functional *ori* could only come from pKH11511c (the pKH1511d *ori* is *pir*-dependent), while *bsd* was only present in pKH1511d. Due to the use of fully-randomised nucleotides at each target position and the combinatorial way in which variants of the two promoters were cloned together, the expected genotype frequencies in this initial library are: 1/16 for WT, 1/192 for each of the 72 single-promoter mutants and 1/2304 for each of the 1,296 double-promoter mutants. With this in mind, an estimated 40,000 colonies were harvested in this step to avoid loss of library complexity. Barcodes were added in a 2nd round of restriction-ligation cloning, introduced *via* a randomised PCR primer. The primer, containing fully-randomised nucleotides at 20 positions, was used to amplify the *bla* β-lactamase gene from plasmid pKD3 (*38*), and the resulting amplicon pool was swapped with the *aadA1* streptomycin/spectinomycin resistance gene in the plasmid library backbone. The primer contains restriction sites on its 5' extremity, one of which is used for this ligation, and another of which allows the barcodes to be moved closer to the mutated promoter region in a later step (see *Barcode-promoter association*). The barcoded plasmid library was again transformed into DH5α Δ*araBA* and colonies were this time selected on ampicillin. False-positive colonies were avoided for the same reason as above, as pKD3 also has a pir-dependent *ori*. An estimated 100,000 colonies were harvested during this step, with the vast majority expected to contain a unique barcode. Expected barcode richness was thus: 6,250 for WT, 521 for

each single-promoter mutant and 43 for each double-promoter mutant. In a final step, the engineered host strain, MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo}, was transformed with this barcoded plasmid library, and an estimated 600,000 colonies were harvested after selection on ampicillin. Detailed library construction methods follow.

Library creation methods

To create the initial library, two promoter-containing primer sets, oPtetLib-fwd and oPlacLib-rev, were each pooled in equimolar quantity (Table 4.S5). These two primer pools were then used together at a concentration of 0.5 µM each pool to PCR-amplify *bsd* from plasmid pKH1511d, using Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific) in its High-Fidelity buffer, following the manufacturer's recommendations. Cycling conditions were: 98°C for 30 secs, followed by 35 cycles of 98°C for 10 secs, 60°C for 30 secs and 72°C for 15 secs, with a final extension step of 72°C for 2 mins. PCR product quality was checked by agarose gel electrophoresis, after which the product was column-purified (QIAquick PCR Purification Kit, QIAGEN) and quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). The purified product and plasmid pKH1511c were then both digested for 90 mins with XhoI and SacI-HF restriction enzymes (NEB CutSmart buffer), and digested DNA was again column-purified (QIAquick PCR Purification Kit, QIAGEN) and quantified with a NanoDrop ND-1000

spectrophotometer. 70ng of the pKH1511c vector fragment was ligated in a 1:3 molar ratio with the *bsd/*promoter-containing insert in a total volume of 20 µl. The ligation was carried out at 16°C overnight using T4 DNA ligase (NEB T4 DNA ligase reaction buffer), which was then deactivated by heating at 65°C for 10 mins. The ligate was microdialysed against water for 30 mins (MF-Millipore, Merck), after which several transformations were performed as follows: 3 µl were electroporated into 50 µl electrocompetent DH5α Δ*araBA* cells; cells were recovered in 500 µl low-salt (Miller) LB for 1 hour at 37°C with shaking at 200 rpm, plated on LB-agar with 100 µg/ml blasticidin-S and incubated overnight at 37°C. Colony-PCR and Sanger sequencing (GATC) of the mutated promoter region was performed on 4 of the resulting colonies as a preliminary test of library quality, and all 4 clones had a unique promoter genotype with a single base substitution in the target region of either one or both promoters, as expected. An estimated 40,000 colonies were scraped off the agar into LB-glycerol (40%), and plasmid DNA was purified from a sample of this cell suspension (QIAprep Spin Miniprep Kit, Qiagen) after thorough mixing.

To barcode the plasmid library, primers oBarcodeBla-fwd and oBarcodeBla-rev (Table 4.S3) were used at a concentration 0.5 µM each to PCR-amplify *bla* from plasmid pKD3 (*38*), using Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific) in its High-Fidelity buffer, following the manufacturer's recommendations. Cycling conditions were: 98°C for 30 secs, followed by 30 cycles of 98°C for 10 secs,

60°C for 30 secs and 72°C for 25 secs, with a final extension step of 72°C for 3 mins. PCR product quality was checked by agarose gel electrophoresis, after which the product was column-purified (QIAquick PCR Purification Kit, QIAGEN) and quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). The purified product was then digested for 1 hour with SpeI-HF restriction enzyme (NEB CutSmart buffer), while the purified plasmid library obtained above was digested for 1 hour with BstZ17I and SpeI-HF restriction enzymes (NEB CutSmart buffer). Digested DNA was again column-purified (QIAquick PCR Purification Kit, QIAGEN) and quantified with a NanoDrop ND-1000 spectrophotometer. 60 ng of the digested library was ligated in a 1:4 molar ratio with the *bla*/barcode-containing insert in a total volume of 20 µl. The ligation was carried out at 16°C overnight using T4 DNA ligase (NEB T4 DNA ligase reaction buffer), which was then deactivated by heating at 65°C for 10 mins. The ligate was microdialysed against water for 30 mins (MF-Millipore, Merck), after which several transformations were performed as follows: 1 µl was electroporated into 15µl commercially-prepared ElectroMAX DH5α-E electrocompetent cells (Invitrogen); cells were recovered in 500 µl LB for 30 mins (to minimise cell replication) at 37°C with shaking at 200rpm, plated on LB-agar with 100 µg/ml ampicillin and incubated overnight at 37°C. The use of commercially prepared electrocompetent cells was necessary due to reduced cloning efficiency at this step, possibly due to the ligation reaction involving blunt ends. Plasmid DNA was purified from 3 colonies (QIAquick PCR Purification Kit, QIAGEN) for Sanger sequencing

(GATC) of the mutated promoter and barcode regions as a preliminary test of barcoding efficiency. All 3 colonies were found to possess a unique promoter genotype, as before, along with a unique, correctly-inserted barcode. An estimated 100,000 colonies were scraped off the agar into LB-glycerol (40%), and plasmid DNA was purified from a sample of this cell suspension (QIAprep Spin Miniprep Kit, Qiagen) after thorough mixing.

To move the barcoded plasmid library into the final host strain, while avoiding the creation of transformants harbouring multiple unique plasmids (*50*), several transformations were performed as follows, with plasmid concentration kept fairly low: 5 ng of the purified barcoded plasmid library obtained above were electroporated into 50 µl electrocompetent MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo} cells; cells were recovered in 500 µl LB for 30 mins at 37°C with shaking at 200rpm, plated on LB-agar with 100 µg/ml ampicillin and incubated overnight at 37°C. An estimated 600,000 colonies were scraped off the agar into LB-glycerol (40%), and this cell suspension was aliquoted and stored at -80°C after thorough mixing.

Barcode-promoter association

To reveal the $P_{LtetO-1}$ and $P_{LlacO-1}$ promoter sequences linked to each barcode sequence, barcodes were first brought closer to the promoters by excision of the intervening region from the plasmid followed by re-circularisation (*48*). PCR-amplification was

then used to add the technical sequences necessary for paired-end Illumina MiSeq sequencing of barcode-promoter amplicons (Fig. 4.S1B).

To first move barcodes closer to the promoter region, the purified barcoded plasmid library was digested for 90 mins with XhoI, SalI-HF and SphI restriction enzymes (NEB CutSmart buffer). The largest fragment (~5.5 kb), which contains the mutated promoters and the barcode, was gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 1% agarose gel and quantified with a NanoDrop ND-1000 spectrophotometer before being self-ligated. XhoI and SalI are isocaudamers, so they create complementary cohesive ends, but the sequence resulting from ligation between these ends is no longer recognised by either enzyme (SphI cuts within the region being discarded, and was simply included to ease gel extraction of the desired fragment). Because of this, they can be included in the reaction mix during self-ligation of the purified fragment to help reduce intermolecular ligation (undesired intermolecular ligation events which recreate XhoI and SalI sites can be reversed, releasing the original monomers and so increasing the efficiency of the desired intramolecular ligation reaction (*48*, *51*)). Due to the inclusion of these restriction enzymes, the self-ligation reaction was carried out in a restriction enzyme buffer, with ATP added for ligase activity. Additionally, the concentration of DNA and ligase was substantially reduced compared to standard ligation reactions to further reduce the occurrence of intermolecular ligation. The self-ligation reaction mix thus consisted of: 1X NEB

restriction buffer 2 supplemented with 100 µg/ml BSA and 1 mM ribo-ATP (NEB), 30 ng DNA, 1 U each of XhoI and SalI-HF and 800 U of T4 DNA ligase, in a total volume of 200 µl. Inspired by the strategy of reference (*51*), the reaction was cycled 50 times between 37°C (restriction enzyme and ligase activity optimum) for 5 mins and 16°C (promote annealing of DNA termini) for 15 mins. A final 37°C incubation was carried out for 15 mins to promote digestion of any remaining XhoI and SalI sites, followed by one of 65°C for 20 minutes to inactivate all enzymes. The ligate was concentrated to ~20 µl using a SpeedVac concentrator (Savant DNA 120, Thermo Scientific) and then microdialysed against water for 90 mins (MF-Millipore, Merck). As a preliminary test of the success of this ligation step, a portion of the ligate was used in a transformation to allow isolation and sequencing of several re-circularised plasmids: 2 µl were electroporated into 50 µl electrocompetent DH5α Δ*araBA* cells; cells were recovered in 500 µl LB for 30 mins at 37°C with shaking at 200 rpm, plated on LB-agar with 100 µg/ml ampicillin and incubated overnight at 37°C; plasmid DNA was purified from 6 colonies (QIAquick PCR Purification Kit, QIAGEN) for Sanger sequencing (GATC) of the ligated region containing the mutated promoters and barcode. All 6 clones were found to possess the expected linking sequence between promoters and barcode, and all plasmids were inferred to be monomeric due to the high Phred scores of the chromatograms (suggesting the presence of a single unique barcode on each re-circularised plasmid).

With the re-circularised DNA placing barcodes in proximity to their respective mutated promoters, this region was then PCR-amplified in a 40 µl reaction using 25 ng of the ligated DNA as template and 0.6 µM each of primers oLinkBarcode-fwd and oLinkBarcode-rev (Table 4.S3). These primers contain adaptors for a 2nd PCR at their 5' extremities, followed by fully randomised hexamers added to increase amplicon diversity to facilitate MiSeq flow-cell clustering. KAPA HiFi HotStart ReadyMixPCR Kit (Kapa Biosystems) was used for amplification, under the following cycling conditions (cycle number was kept low to reduce PCR errors and artefacts): 95°C for 3 mins, followed by 15 cycles of 98°C for 20 secs, 60°C for 30 secs and 68°C for 30 secs, with a final extension step of 68°C for 2 mins. The amplicon (~0.9 kb) was gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 1.5% agarose gel and quantified fluorometrically (dsDNA HS Assay Kit with a QuBit 2.0, Thermo Scientific). A 2nd 40 µl PCR was then performed using 5 ng of this amplicon as template and 0.6 µM each of a P5 and P7 Nextera Index Kit primer (Illumina) to add Illumina adaptors and multiplexing indexes. KAPA HiFi HotStart ReadyMixPCR Kit (Kapa Biosystems) was again used for amplification, under the following cycling conditions (cycle number was again kept low): 95°C for 30 secs, followed by 12 cycles of 95°C for 10 secs, 55°C for 30 secs and 68°C for 30 secs, with a final extension step of 68°C for 5 mins. The amplicon library (~1 kb) was gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 1.5% agarose gel and a 20,000X dilution was quantified by qPCR using KAPA Library

Quantification Kit for Illumina (Kapa Biosystems) on a LightCycler 480 (Roche), following the manufacturer's recommendations.

The resulting amplicon library is composed of DNA fragments of the structure: P5 - i5 - N_6 PCR tag - $P_{LtetO-1}$ (rev) - *bsd* (rev) - $P_{LlacO-1}$ - N_{20} plasmid barcode - N_6 PCR tag - i7 - P7, which are ~1 kb in size (close to the size-limit for reliable MiSeq sequencing). 300nt paired-end MiSeq sequencing allowed us to sequence the entire $P_{LtetO-1}$ promoter on Read 1 and the plasmid barcode and entire $P_{LlacO-1}$ promoter on Read 2 (note that Reads 1 and 2 do not overlap). For this, a 600-cycle MiSeq Reagent Kit v3 (Illumina) was used, and DNA was loaded at a concentration of 12pM, with a 20% PhiX DNA spike-in (PhiX Control v3, Illumina). Preliminary quality filtering and demultiplexing by the standard MiSeq software package (Illumina) resulted in an output of > 22M read pairs, giving an expected coverage of > 220X for each plasmid barcode.

MiSeq reads were processed using the Mothur (*52*) (version 1.37.6) software package *via* the following steps: reads were quality-filtered by size (>199 bases), number of uncalled bases (<3 Ns) and length of the longest homopolymer stretch, another indicator of overall read quality (<9 bases). Entire $P_{LtetO-1}$ sequences were extracted from Read 1, and barcode sequences and entire $P_{LlacO-1}$ from Read 2, by Needleman alignment to reference sequences (default alignment parameters). Reads for which either the $P_{LtetO-1}$, $P_{LlacO-1}$ or barcode region contained insertions or did not generate a full alignment with the reference were discarded. The Mothur Precluster algorithm was

then used to cluster barcode sequences differing by a Hamming distance of 1, with the aim of correcting for PCR and sequencing errors (the potential barcode diversity is so high ($> 1x10^{12}$) that the presence of immediately neighbouring sequences is very likely due to these errors (Fig. 4.S2C)). The algorithm uses sequence abundance to decide the "true" (majority) sequence for each cluster, and to decide where a sequence clusters if it has >1 immediate neighbour. After de-gapping and re-grouping barcode sequences to account for any alignment ambiguities resulting from small deletions, barcode clusters were used to build a dictionary assigning each "true" barcode sequence to a $P_{LtetO-1}$ and $P_{LlacO-1}$ sequence. Due to a high rate of PCR-derived recombination (*53*) being observed (caused by the extensive homology between all fragments, and resulting in some molecules displaying incorrect barcode-promoter associations), a haplotype-based strategy was used for this step rather than one in which each nucleotide is considered independently as in reference (*48*). This is because the small number of mutations expected to be present in each mutant (0-2) means that, at any particular position, the majority of molecules will possess the WT base. If the consensus $P_{LtetO-1}$ and $P_{LlacO-1}$ sequences attached to a particular barcode are computed by considering each nucleotide independently, a high recombination rate can thus result in mutant bases being assigned as the WT base. The haplotype-based strategy, executed in Python (v3.5), consists of the following steps: for each barcode cluster (consisting of reads whose barcode sequences are identical to or the immediate neighbour of the inferred "true" barcode sequence), the associated complete $P_{LtetO-1}$-

$P_{LlacO-1}$ concatenate sequences were grouped; the number of occurrences of each of these 108-nt $P_{LtetO-1}$-$P_{LlacO-1}$ sequences was tabulated; if the cluster contained more than 2 read pairs in total, the most abundant concatenate $P_{LtetO-1}$-$P_{LlacO-1}$ sequence is ≥ 5x more abundant than the second-most abundant one, and the most abundant concatenate $P_{LtetO-1}$-$P_{LlacO-1}$ sequence contains no Ns (uncalled bases), then this $P_{LtetO-1}$-$P_{LlacO-1}$ sequence is assigned to the "true" barcode sequence for that cluster (else the cluster is discarded). This stringent requirement is aimed at reducing barcode-promoter mis-assignments caused by PCR and sequencing errors, PCR-derived recombination or intermolecular ligation during the first step of barcode-promoter association, as well as to avoid any barcodes that may be linked to multiple promoter genotypes. Only barcodes associated to promoter genotypes for which the entire promoter regions contain no unexpected mutations were considered for further analysis.

Mutant library competition assays

The final mutant library (host strain transformed with barcoded plasmid library) was competed over ~30 mean generations (~3 days) in the presence of L-arabinose and different concentrations of the inducers, aTc and IPTG. Cell density was kept low during competition ($OD_{600} < 0.2$) by serial transfer into fresh medium, in order to maintain the culture in exponential phase and to avoid large changes in medium composition. Large volumes of media (100 ml) were used to avoid severe population bottlenecks during serial transfer (> 1 x 10^8 cells each transfer). Plasmid DNA was

purified from the culture at several time-points for HiSeq sequencing of plasmid barcodes. Plasmid barcode abundance serves as a proxy for the abundance of cells carrying that particular barcode. The change in frequency over time of a barcode thus provides an estimate of competitive fitness for the lineage carrying that barcode (*54*). Since we know the $P_{LtetO-1}$-$P_{LlacO-1}$ sequence associated to each barcode (see *Barcode-promoter association*), this in turn provides us with a distribution of fitness estimates for every mutant.

The base competition medium consisted of M9 + 0.1% casamino acids (for basal growth) + 0.03% L-arabinose, with 100 µg/ml ampicillin to select against plasmid loss. A preliminary competition experiment was performed under inducer concentrations of 20 ng/ml aTc and 30 µM IPTG, expected to endow the wildtype with near-maximal fitness (although this was found to be inaccurate). A second round of competition experiments was carried out at a later date and was comprised of three different inducer concentration combinations. One duplicated those of the initial experiment to check reproducibility (Figs. 4.S3-4), and the other two were: 5 ng/ml aTc and no IPTG, and 200 ng/ml aTc and no IPTG. No IPTG was chosen to reduce *araB* expression as much as possible, as the preliminary experiments suggested that the wildtype over-expressed *araB* even in the absence of inducer (*28*), due to promoter leakiness. The range of aTc was chosen to explore the full range of achievable *araA* expression.

In detail, a sample of the frozen library cell stock was thawed and diluted in 200 ml of M9 + 0.5% casamino acids (with 100 µg/ml ampicillin), in a 500 ml flask, for a final blank-subtracted OD_{600} of 0.12 (200 µl read by Varioskan microplate reader, Thermo Scientific). This common starting-culture was recovered for ~3.5 hours at 37°C with shaking at 200 rpm, reaching an OD_{600} of 0.3, before being washed with 200 ml of M9 + 0.1% casamino acids. Washed cell pellets (each coming from 50 ml of the original culture) were resuspended directly in 100 ml of the different competition media, for an effective 2X dilution of the original culture (OD_{600} of ~0.15; flasks of competition media were always pre-warmed at 37°C to keep temperature constant and detect any contamination, with aTc, IPTG and ampicillin being added at the time of transfer to avoid degradation). These cultures were then acclimatised to their respective competition media for ~2.25 hours (37°C, 200 rpm), reaching an OD_{600} of 0.23-0.28, to allow time for stable induction by aTc, IPTG and L-arabinose. These acclimatised cultures were taken as t_0, and so plasmid DNA was purified from a 50 ml sample of each culture (QIAprep Spin Miniprep Kit, Qiagen) and quantified fluorometrically (dsDNA HS Assay Kit with a QuBit 2.0, Thermo Scientific) for eventual HiSeq sequencing of plasmid barcodes (the rest remaining after this and transfer was pelleted, resuspended in LB-40% glycerol and stored at -80°C as an archive). 3.2 ml of each culture was transferred to 100 ml fresh competition media (~32X dilution) and left to grow (37°C, 200 rpm) to an OD_{600} of ~0.12 (3-4 mean generations). DNA was purified from a 50 ml sample of each culture (t_1), as before, and 3.2 ml of each culture was

again transferred to 100 ml of fresh competition media and left to grow to an OD_{600} of ~0.12 (~5 mean generations). This procedure was repeated until t_6 (or t_8 in an initial experiment), for a total of ~29 mean generations of competition (or ~39), over which time the impact of *de novo* mutation appears low (Fig. 4.S3). The precise number of mean generations between each sampling was calculated from OD_{600} values and used for estimating fitness.

Barcode-sequencing of competed mutant library

To track plasmid barcode frequencies throughout the competition experiments, barcodes were PCR-amplified from plasmid DNA in 2 steps, as for *Barcode-promoter association*, to add technical sequences necessary for 100nt overlapping paired-end Illumina HiSeq sequencing. This was performed for time-points t_0, t_1, t_2, t_4, t_6 and t_8 (approximately 0, 4, 9, 19, 29 and 39 mean generations) for the preliminary experiment, and t_1, t_2, t_4 and t_6 for the later experiments. These time-points were chosen with the aim of obtaining precise fitness estimates for both large-effect and small-effect mutations.

In detail, at each selected time-point, 20 ng of purified plasmid DNA was PCR-amplified in a 40 µl reaction using 0.6 µM each of primers oBarcodeSeq-fwd and oBarcodeSeq-rev (Table 4.S3). These primers contain adaptors for a 2nd PCR at their 5' extremities, followed by fully randomised hexamers to increase amplicon diversity, as in *Barcode-promoter Association.* In this case, the randomized hexamers were also

used to detect PCR duplicates arising from the 2nd PCR[45]. KAPA HiFi HotStart ReadyMixPCR Kit (Kapa Biosystems) was used for amplification, under the following cycling conditions (cycle number was kept low to reduce PCR errors and artefacts): 95°C for 3 mins, followed by 12 cycles of 98°C for 20 secs, 60°C for 30 secs and 68°C for 30 secs, with a final extension step of 68°C for 2 mins. Amplicons (~200 bp) were gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 2% agarose gel and quantified fluorometrically (dsDNA HS Assay Kit with a QuBit 2.0, Thermo Scientific). A 2nd 40 µl PCR was then performed using 5-8 ng of each amplicon as template and 0.6 µM each of a P5 and P7 Nextera Index Kit primer (Illumina) to add Illumina adaptors and multiplexing indexes. KAPA HiFi HotStart ReadyMixPCR Kit (Kapa Biosystems) was again used for amplification, under the following cycling conditions: 95°C for 3 mins, followed by 13 cycles of 98°C for 20 secs, 55°C for 30 secs and 68°C for 30 secs, with a final extension step of 68°C for 5 mins. These ~300 bp amplicons, of the structure, P5 - i5 - N_6 PCR tag - N_{20} plasmid barcode - N_6 PCR tag - i7 - P7, were gel-purified (QIAquick Gel Extraction Kit, Qiagen) using a 2% agarose gel and sent to IntegraGen (Evry, France) for qPCR-based quantification, equimolar pooling and 100nt paired-end HiSeq-4000 sequencing (Illumina). Preliminary quality filtering and demultiplexing (Integragen, Evry, France) resulted in ~18 M read pairs per time-point per competition experiment, giving, for each point, an expected barcode coverage of ~200X and an expected mutant coverage of >14,000X.

HiSeq sequencing reads were processed using the Mothur (*52*) (version 1.37.6) software package by the following steps: Forward and reverse reads were joined into contigs using Mothur's make.contigs command with the default parameters. Contigs were then quality-filtered by size (<131bp, as longer contigs imply forward and reverse reads could not be properly overlapped), number of uncalled bases (no Ns) and length of longest homopolymer stretch, an indicator of overall read quality (<9 bases). To remove the majority of PCR duplicates arising from the 2nd PCR (made possible by randomised hexamers introduced on each side of the barcode during the 1st PCR (*49*)), if a particular full contig was present more than once, only one copy was kept. Barcode sequences were then extracted after aligning contigs to the reference sequence (Needleman global alignment). Reads containing insertions or not generating a full alignment with the reference were discarded. Next, the Mothur precluster algorithm was used to cluster barcode sequences differing by a Hamming distance of 1, with the aim of correcting for PCR and sequencing errors, as described in *Barcode-promoter association.* After de-gapping and re-clustering barcode sequences to account for any alignment ambiguities resulting from small deletions, the number of occurrences of each "true" barcode was tabulated across all time-points for each competition experiment. Finally, a custom R (v.3.4.3) script was used to merge these barcode counts tables with the barcode-promoter mutant dictionary generated in *Barcode-promoter association.*

Estimation of competitive fitness and epistasis

We found that competitive fitness was not constant over the course of competition, with, for example, a possible period of physiological adaptation between t_0 and t_2 for certain inducer environments (Fig. 4.S3). By t_6, a substantial number of lower-fitness mutants begin to escape detection completely, and so to avoid any bias in fitness estimates we consider only the frequency changes between t_2 and t_4 (two time-points). We begin by removing outlier barcodes associated to the wildtype genotype, to avoid any systematic biases coming from inaccurate wildtype estimates. This was done by computing the log ratio of t_4 to t_2 counts for all wildtype barcodes and removing those giving values > 1.5x the inter-quartile range above (below) the upper (lower) quartile. We also removed all barcodes giving < 8 reads at t_2 from our dataset. For every remaining mutant barcode, i, we then estimate its log fitness relative to the wildtype as:

$$F_i^{rel} = \frac{\ln\left(\frac{f_i^{t_4}}{\sum f_{wt}^{t_4}}\right) - \ln\left(\frac{f_i^{t_2}}{\sum f_{wt}^{t_2}}\right)}{t_4 - t_2},$$

where f_i is the frequency of a mutant barcode, $\sum f_{wt}$ is the total frequency of all wildtype barcodes and $t_4 - t_2$ is the number of mean generations between the two time-points considered (~9). We now estimate log relative fitness of a mutant g, F_g^{rel}, as the median of that of its associated barcodes, $F_{g_i}^{rel}$. We use the median barcode fitness as a fitness estimate for each promoter genotype as a convenient way to filter out the many

sources of error in a competition experiment (especially undetected mutations introduced during library construction, *de novo* mutations arising during competition and barcode-promoter misassignments due to PCR and sequencing errors) (Fig. 4.S3A). The number of eligible barcodes for each promoter genotype ranges from a few to thousands (exact numbers are provided in Data S1). Some barcodes disappear from our sequencing sample by t_4, and so are given an F_i^{rel} of *–Inf.* In Env_1 and Env_3, for a very few genotypes this is the case for the majority of their barcodes, and we identify these mutants as being less fit than the wildtype but cannot estimate total/marginal fitness effects or epistasis for them.

To estimate the precision of mutant fitness estimates, we used standard bootstrapping of the eligible mutant and wildtype barcodes (n=1,000), each time computing the mutants' fitness, F_g^{rel}, as the median fitness of their associated barcodes, $F_{g_i}^{rel}$. The same 1,000 sets of randomly sampled wildtype barcodes were used as the references for all mutants. The bootstrap distributions were then used to determine significance (empirical 95% confidence) for non-neutrality of total (F_g^{rel}) and marginal (F_g^{marg}) fitness effects, non-zero epistasis, simple sign epistasis and reciprocal sign epistasis, pairing bootstrap F_g^{rel} estimates by sampled wildtype barcode set when necessary.

The marginal fitness change induced by adding mutation A to the genetic background B is defined as $F_{A|B}^{marg} = F_{AB}^{rel} - F_B^{rel}$, and epistasis between mutations A and B is defined as $\varepsilon_{AB} = F_{AB}^{rel} - (F_A^{rel} + F_B^{rel})$ (*17*).

Phenotype-fitness model

We consider a linear metabolic reaction path,

$$S \underset{A}{\rightarrow} X \underset{B}{\rightarrow} Y \underset{C}{\rightarrow} Z \rightarrow \cdots$$

where S is the substrate (L-arabinose) concentration. As shown in references (*25*, *29*), for S and Z fixed, the steady-state flux for non-saturated enzymes and the intermediate concentration are respectively given by:

$$\varphi = \frac{1}{1/A + 1/B + \eta} \quad (1)$$

$$Y = D - \frac{\varphi}{1/A + 1/B} \quad (2)$$

where A and B are proportional to the maximum reaction rates provided by each enzyme, η is the inverse of the maximal flux, φ_{max}, as imposed by the fixed pathway steps, and D is a certain function of S and equilibrium constants (see reference (*29*) for detailed expressions). We note that the flux, φ, is an increasing function of A and B and saturates at φ_{max} for very efficient enzymes A and B, or very high concentrations of them. However, at high fluxes, the hypothesis of unsaturated downstream enzymes breaks down, and a reaction step becomes limiting, such that the concentrations of metabolic intermediates may build up to toxic levels.

To account for such saturation, we extend the model above by considering the full Reversible Michealis-Menten (RMM) form for the step C instead of its first order approximation (similar reasoning applies for longer paths). At steady-state, all reaction speeds must be equal, giving for the third step:

$$\varphi = \frac{\alpha Y - \beta Z}{1 + \gamma Y + \delta Z} \quad (3)$$

where $\alpha, \beta, \gamma, \delta$ are the RMM parameters for C. Equivalently, expressing Y as a function of φ:

$$Y = \frac{\beta Z + (1 + \delta Z)\varphi}{\alpha - \gamma\varphi} \quad (4).$$

We could eliminate Y by combining (2) and (4), Z being fixed, and obtain an exact expression for φ. Note that expression (1) is recovered for $\delta = \gamma = 0$, as this corresponds to the unsaturated case. In the general case, φ would still be an increasing function of A and B and saturate at a certain value, but its expression becomes more complicated.

Instead of using the full expression of φ, we report here an approximation with less parameters, which consistently recovers the monotonicity with A and B, and the limit regimes for unsaturated and saturated downstream steps. For this, we simply keep expression (1) for the flux, and set its saturation by the saturation of the reaction catalysed by C, as obtained in the limit of very high Y in (4):

$$\varphi_{max} = 1/\eta = \alpha/\gamma\,.$$

With this, expression (4) becomes:

$$Y = \frac{P + Q\varphi}{\varphi_{max} - \varphi} \quad (5)$$

where P and Q are functions of the fixed downstream enzyme properties and concentrations. We note in particular that Y diverges when the flux becomes maximal, meaning that the downstream reaction is saturated, leading to an accumulation of Y.

We now assume fitness to be a function of flux and the toxic intermediate (L-ribulose-5-phosphate) concentration, Y, and that there exist constants e and f such that, from (1) and (5):

$$F = e\varphi - fY = e\varphi - f\frac{P + Q\varphi}{\varphi_{max} - \varphi} \quad (6).$$

This expression can be further simplified by considering the low and high flux regimes:

For $\varphi << \varphi_{max}$, (6) behaves as $F = -fP/\varphi_{max} + u\varphi$, with $u = e - f(Q + P/\varphi_{max})/\varphi_{max}$, the offset $-fP/\varphi_{max}$ being determined solely by properties of the fixed downstream enzyme, C. Thus, any fitness change due to mutations in A and B is of the form $u\varphi$.

For $\varphi \sim \varphi_{max}$, the first term of (6) remains finite while the second with numerator $v = f(P + Q\varphi_{max})$ diverges. Thus, replacing e by u as defined in the regime $\varphi << \varphi_{max}$ has a negligible contribution.

Introducing a basal growth rate, ω, supplied by alternative nutrients in the medium (casamino acids), fitness is then well approximated by:

$$F = \omega + u\varphi - \frac{v}{\varphi_{max} - \varphi} \quad (7).$$

In addition to flux and toxic metabolite concentration, gene expression burden can also contribute to fitness changes (*26*, *28*, *55*, *56*). Following the observation that protein expression burden depends on metabolic state (*27*, *57*), we include an expression cost factor in which ϑ_A and ϑ_B describe the cost of increasing cellular enzyme activity, including potential contributions from both the amount of expression and the specific enzyme activity constants:

$$F = \left(\omega + u\varphi - \frac{v}{1/\eta - \varphi}\right)(1 - \theta_A A - \theta_B B) \quad (8).$$

This expression is considered valid only when both factors are positive. Expressions (1) and (8) together define a fitness surface in the two-dimensional space of AraA and AraB activities, described by the 6 independent parameters, ω, u, v, ϑ_A, ϑ_B and η.

The entire model consists of 83 parameters: the 6 detailed immediately above; 5 defining the "wildtype" activity levels (AraA and AraB activities for the 3 inducer environments, with Env_2 and Env_3 having the same wildtype AraB activity, as both contained the same IPTG concentration); and 72 defining the relative impact of the single mutations (36 for each gene) on enzyme expression/activity. For a given parameter set, the fitness, F^{rel}, of the 72 single mutants and 1,296 double mutants was

computed in each of the 3 environments, relative to the respective "wildtype" fitness. The 83-parameter model was fitted on 4,079 data points, corresponding to the computable set of relative fitnesses (Fig. 4.2A) of the 1,368 mutants measured in 3 different environments.

The model was fit using multiple Monte Carlo Markov Chains (MCMC) (*58*). Parameters were generated randomly from uniform distributions, both initially and at each step of the chain for a randomly chosen parameter (bounds are provided in Fig. 4.S8A and Data S2; bounds for expression effects of inducer concentrations and a few mutations were guided by experimental expression measurements (data not shown)). 800 chains, each of 300,000 steps, were simulated, and for each chain the parameter set giving the best fit with measured fitness values was stored (residuals were weighted to give equal consideration overall to single and double mutants, and were also normalised to the mean fitness effect in the environment from which they came). The distribution of goodness-of-fit values from the 800 chains was multi-modal (*ie.* convergence was not guaranteed), with ~5-10% of the chains achieving a best fit residing in the lowest peak. We take the best of all these parameter sets as the most likely fit, but the distributions of parameter values from the best 2.5% of chains are also provided in Fig. 4.S8A and Data S2.

Several fitness function variations containing less parameters than the one presented in the main text were fit in the same way, and we conclude that flux, toxicity and gene

expression burden must all be accounted for to explain the observed fitness and epistasis values (Fig. 4.S9).

4.5 Supplementary figures

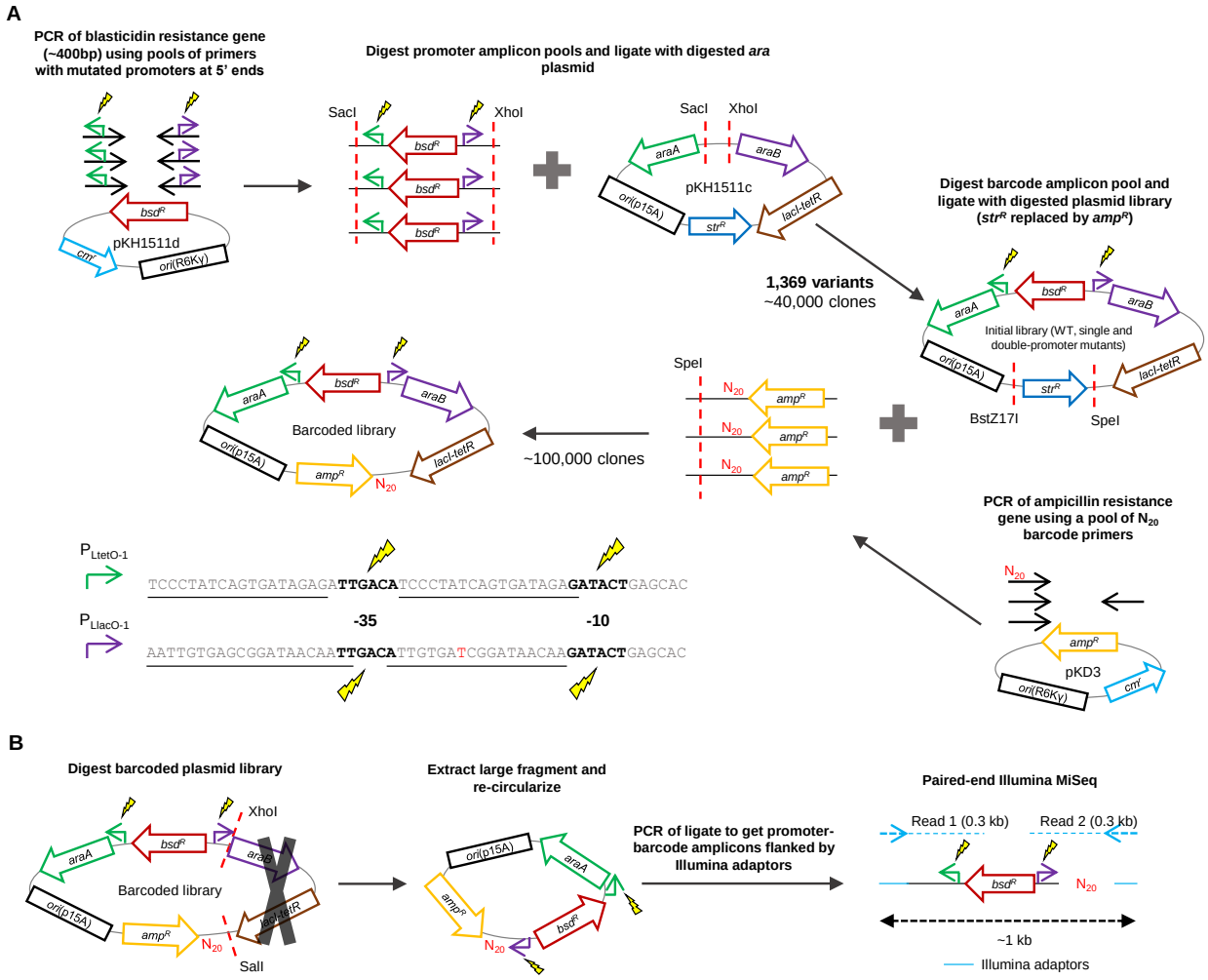


Fig. 4.S1.

Construction and characterisation of barcoded promoter-mutant plasmid library. (**A**) A blasticidin-resistance cassette (bsd^R) was amplified from pkH1511d using pools of primers carrying variants of the entire $P_{LtetO-1}$ (green arrow) and $P_{LlacO-1}$ (purple arrow) promoters at their 5' ends, flanked by *SacI* and *XhoI* restriction sites. The resulting amplicon pool (containing an expected 1,369 promoter variant combinations – see below) was digested with SacI and XhoI and ligated with a SacI-XhoI digest of plasmid pKH1511c. ~40,000 colonies were harvested after transformation with this ligate, from which plasmid DNA was then purified, giving an initial plasmid library. An ampicillin-resistance cassette (amp^R) was amplified from pKD3 using for forward priming a pool of primers containing a region of 20 fully randomised nucleotides (the barcode, N_{20}) at their 5' end, flanked by *a* SpeI restriction site. The resulting amplicon pool was digested with SpeI and ligated with a BstZ17I-

SpeI digest of the initial plasmid library (BstZ17I creates blunt ends). ~100,000 colonies were harvested after transformation with this ligate, each expected to harbour a plasmid with a unique barcode. Underlined regions of the $P_{LtetO-1}$ and $P_{LlacO-1}$ sequences are the repressor binding sites reported in reference (*41*). The repressor of $P_{LtetO-1}$ is TetR, and the repressor of $P_{LlacO-1}$ is LacI, both encoded on the constant region of the library plasmid (*lacI-tetR*). The red T in $P_{LlacO-1}$ differs from the original sequence reported in reference (*41*), and was used due to its appearance during an initial adaptation step (this modified sequence still allows titratable control of expression from $P_{LlacO-1}$ using IPTG, as verified by growth and expression measurements – see Table 4.S1). Black letters denote the -35 and -10 RNA-polymerase binding hexamers (note that 1 of the -10 nucleotides in $P_{LtetO-1}$, and 3 of the -35 nucleotides in $P_{LlacO-1}$, overlap with repressor binding sites). These hexamers were targeted for mutation: over these 12 sites, for each promoter, all 36 possible single-nucleotide substitutions were made, along with the wildtype, and the two sets of promoter variants were comprehensively combined. (**B**) To uncover which barcodes were linked to which promoter genotypes, the barcoded plasmid library was first digested with XhoI and SalI to remove the region between the $P_{LtetO-1}$ and $P_{LlacO-1}$ promoters and the barcode. The remaining section of the plasmids was re-circularised by ligation under conditions promoting intramolecular ligation. This ligate was used as template for PCR to amplify the newly created promoter-barcode region while adding Illumina adaptors to the amplicon termini. Finally, non-overlapping paired-end Illumina MiSeq sequencing was used to associate barcode sequences with promoter genotypes.

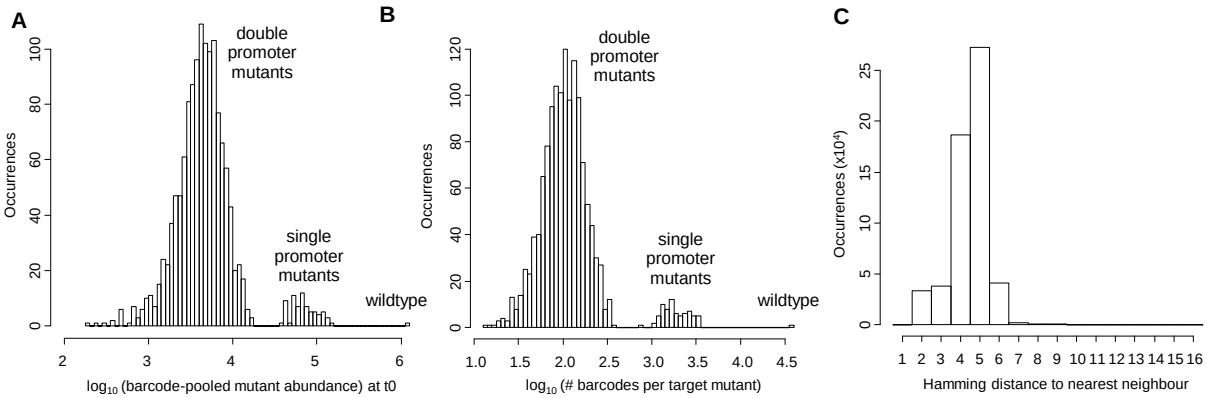


Fig. 4.S2.

Sequencing coverage and quality of barcoded mutant library. Data from t_0 of the preliminary competition experiment. (**A**) The total coverage (after pooling barcode counts) of each genotype is on the order of 10^3 for double mutants, 10^5 for single mutants and 10^6 for the "wildtype". These different ranges result directly from the library creation strategy. (**B**) The number of unique barcodes associated to each genotype is on the order of 10^2 for double mutants, 10^3 for single mutants and 10^4 for the wildtype. These different ranges also result directly from the library creation strategy. (**C**) Over all barcode sequences observed, the mean Hamming distance to a barcode's nearest neighbor is 4.5. The complete absence of immediately neighbouring sequences is due to the preclustering analysis, in which immediately neighbouring sequences were assumed to be the result of PCR and sequencing errors.

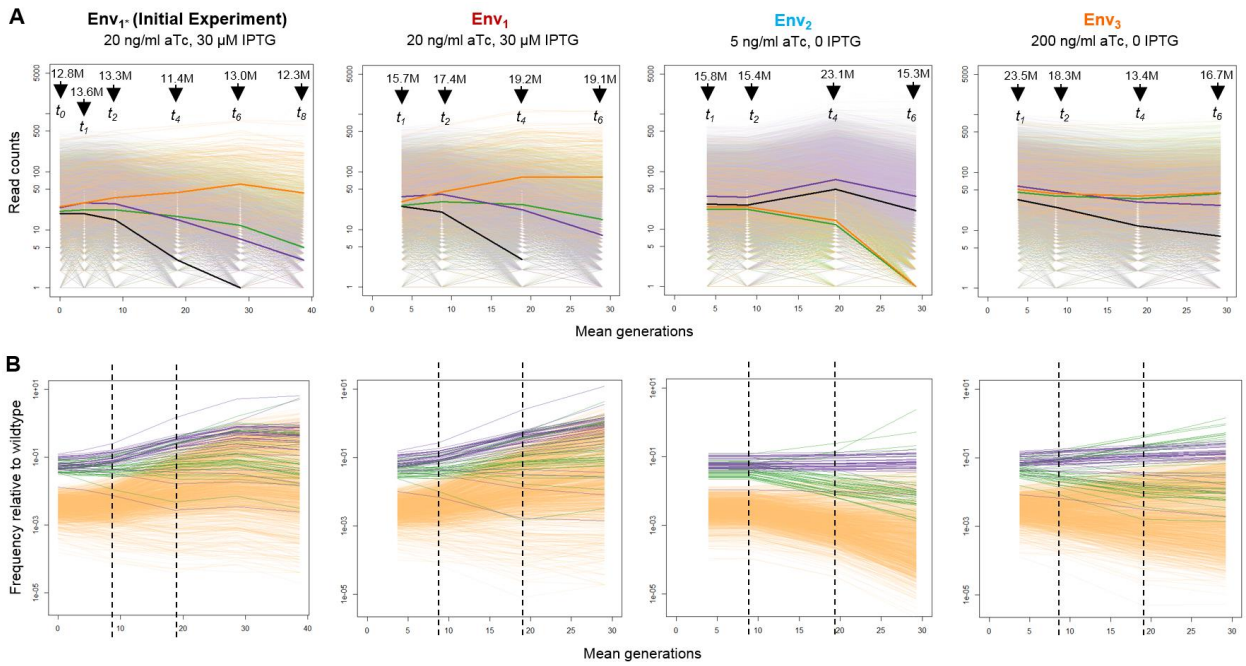


Fig. 4.S3.

Mutant dynamics during pooled competition assays under different inducer concentrations. (**A**) Example trajectories are shown for all barcodes associated to the wildtype (black), a single $P_{LtetO-1}$-*araA* mutant (green), a single $P_{LlacO-1}$-*araB* mutant (purple) and the resulting double mutant (orange). Thick lines show median read counts. Numbers are the total number of HiSeq reads obtained at each sampled time-point. (**B**) Barcode-grouped trajectories are shown for all 1,368 mutants relative to the wildtype. Colours as in **A.** At every time-point, read counts for all barcodes belonging to a particular mutant have been summed and normalized to WT read counts. Dashed lines indicate time-window chosen for fitness estimation.

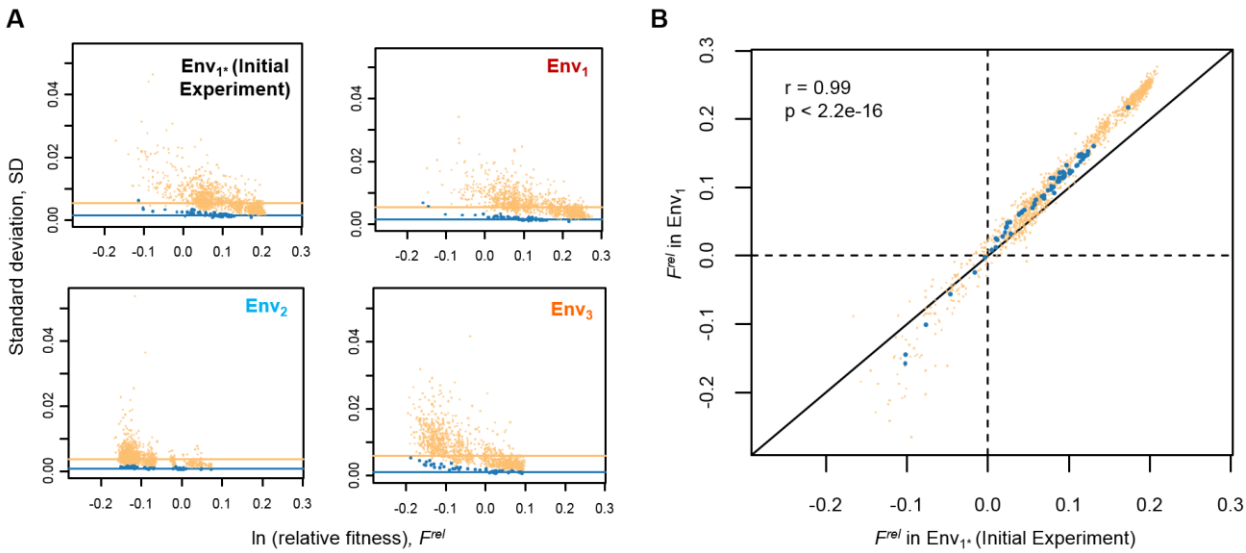


Fig. 4.S4.

Measurement precision and reproducibility. (**A**) Fitness estimates are plotted against their corresponding bootstrap standard deviations (SD) for the different competition assays. Single mutants (blue) yield more precise estimates as they are associated to more barcodes than double mutants (orange). Precision is lower for less-fit genotypes due to their more rapidly decreasing abundances and so higher counting noise. Lines show median SDs. (**B**) F^{rel} estimates are compared between two replicate experiments (Env_1 conditions; same mutant library stock). Colors as in **A**. Reproducibility is high (Pearson's r = 0.99, n = 1,344 mutants), but systematic differences are apparent, likely due to small differences in media composition.

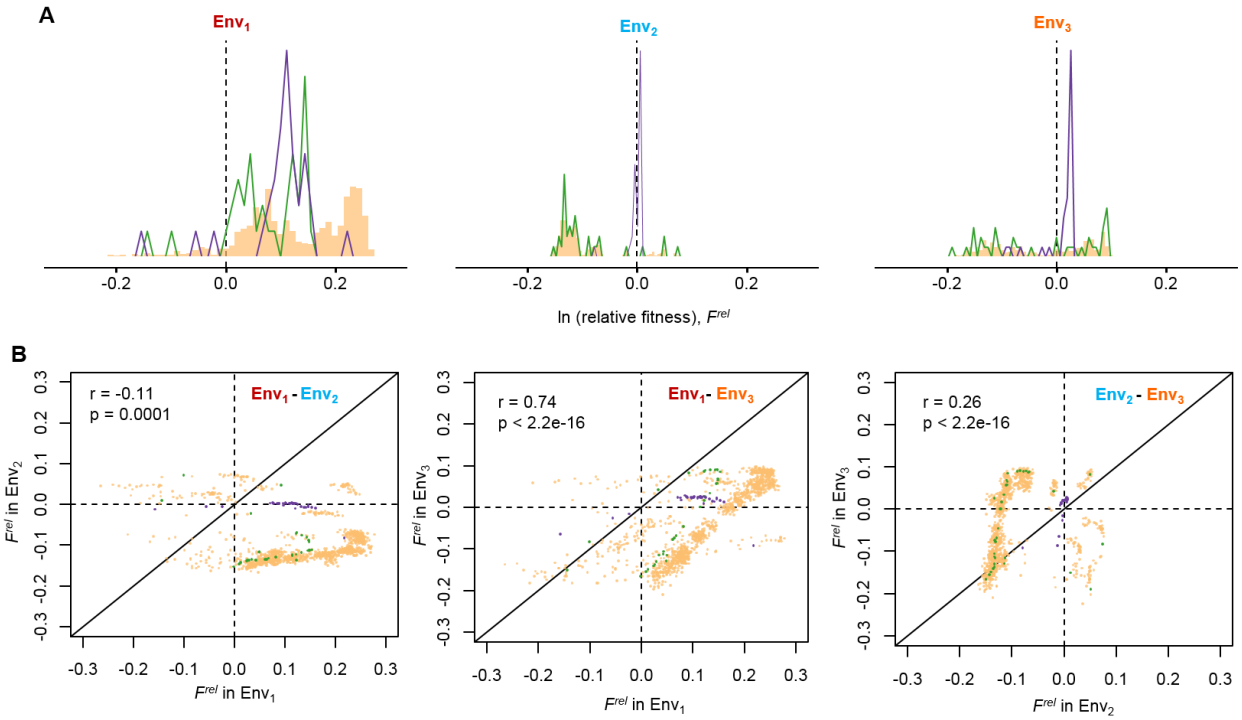


Fig. 4.S5.

Fitness effects of single and double mutations across environments. (**A**) Density distributions of fitness effects (F^{rel}) of single $P_{LtetO\text{-}1}$-*araA* mutants (green), single $P_{LlacO\text{-}1}$-*araB* mutants (purple) and double mutants (orange). (**B**) Correlations between mutant F^{rel} in different environments range from strongly positive to weakly positive and weakly negative, and can show strong signs of non-monotonicity. Pearson's r is shown, with n = 1,345, 1,345 and 1,366 mutants, left-right. Colours as in **A**.

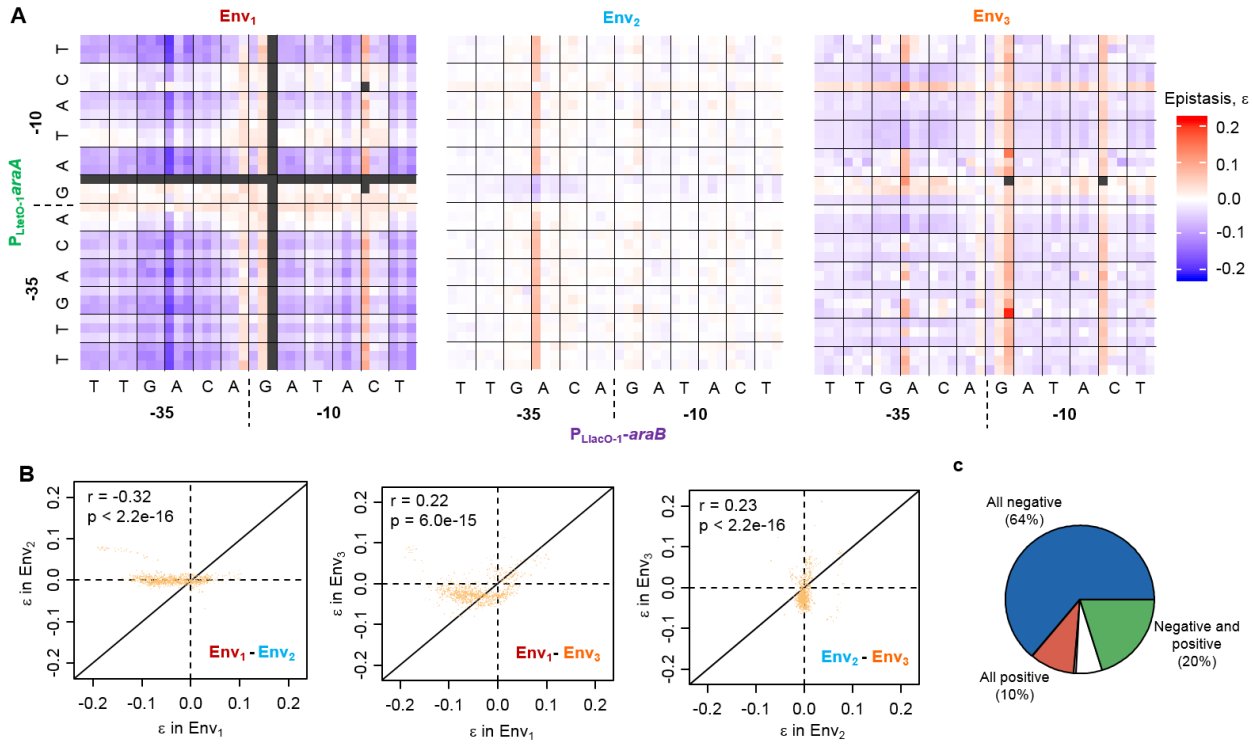


Fig. 4.S6.

Epistasis across environments. (**A**) Genotype-epistasis maps. "-35" and "-10" denote the RNA polymerase-binding hexamers. Letters show the wildtype base at each position. The three mutants at each position are ordered alphabetically, as in Fig. 2A. Grey denotes incomputable epistasis coefficients. (**B**) Correlations between epistasis coefficients in different environments, with Pearson's r (n = 1,223, 1,223 and 1,294 mutation pairs, left-right). (**C**) The fraction of mutation pairs (n=1,296) for which, across environments, epistasis can be positive but never negative (red), negative but never positive (blue), or both positive and negative (green). Pairs exhibiting no detectable epistasis in any environment are shown in grey, and those for which epistasis could not be computed in all environments are white.

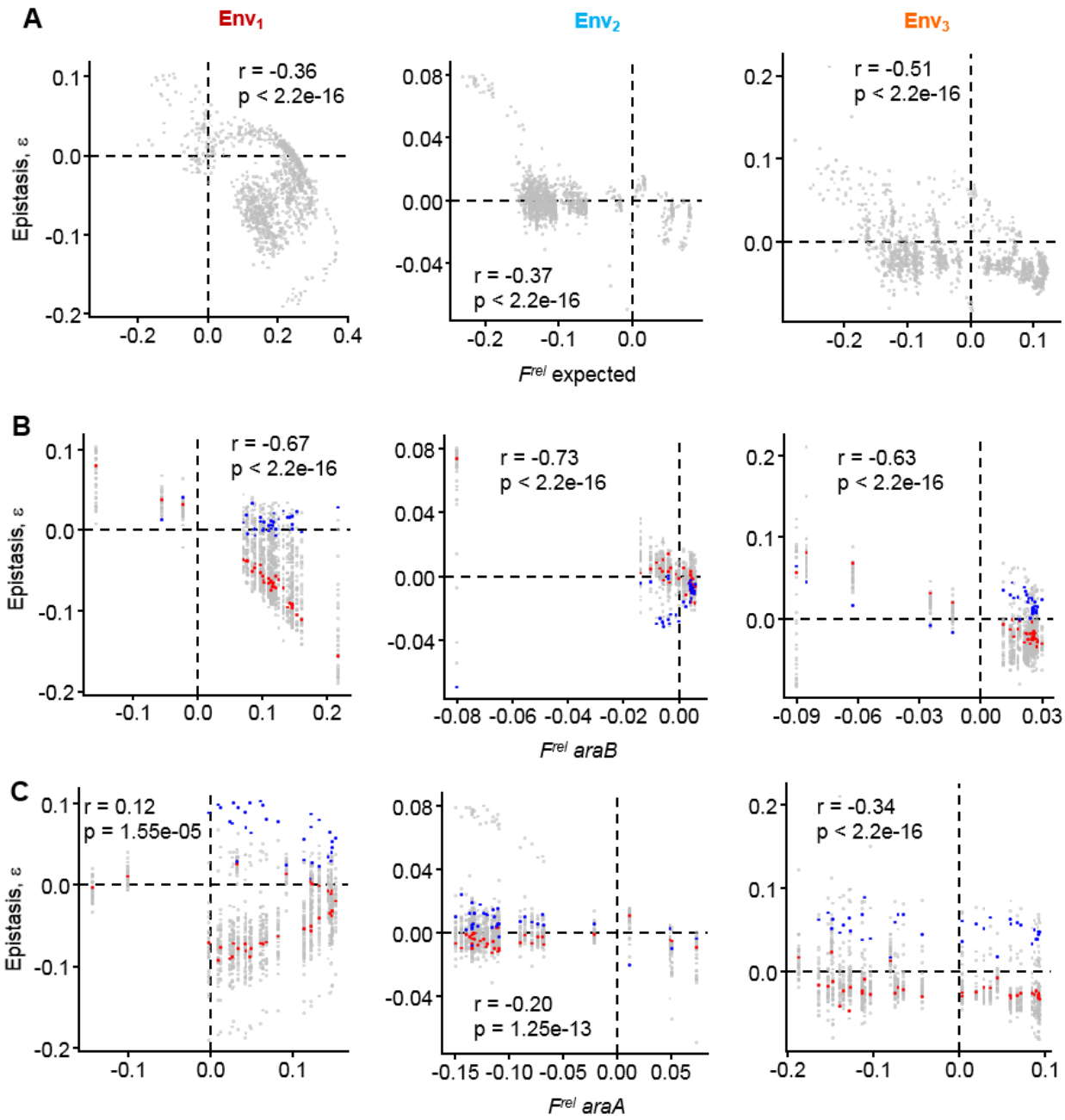


Fig. 4.S7.

Correlations between individual fitness effects and epistasis. (**A**) In all environments, the sum of the fitness effects of two individual mutations (F^{rel} expected) correlates negatively with the epistasis they experience when combined, a trend of diminishing returns and losses (Pearson's r, n = 1,223, 1,296 and 1,294 mutation pairs, $Env_{1\text{-}3}$). The relationship appears complex, however. (**B**) When $P_{LlacO\text{-}1}$-*araB* is considered alone, the negative correlation between fitness effects and epistasis is stronger, but in Env_2 and Env_3 there is evidence of non-monotonicity (Pearson's r, number of mutation pairs as for **A**). Different $P_{LtetO\text{-}1}$-*araA* alleles can cause different trends within an environment, and the same $P_{LtetO\text{-}1}$-*araA* allele can cause different

trends across environments (coloured alleles as for Fig. 3B, top panel). (**C**) When $P_{LtetO\text{-}1}$-*araA* is considered alone, the negative correlation between fitness effects and epistasis is weaker, and in Env_1 it even becomes positive, albeit strongly non-monotonous (Pearson's r, number of mutation pairs as for **A**). Different $P_{LlacO\text{-}1}$-*araB* alleles can cause different trends within an environment, and the same $P_{LlacO\text{-}1}$-*araB* allele can cause different trends across environments (coloured alleles as for Fig. 3B, bottom panel).

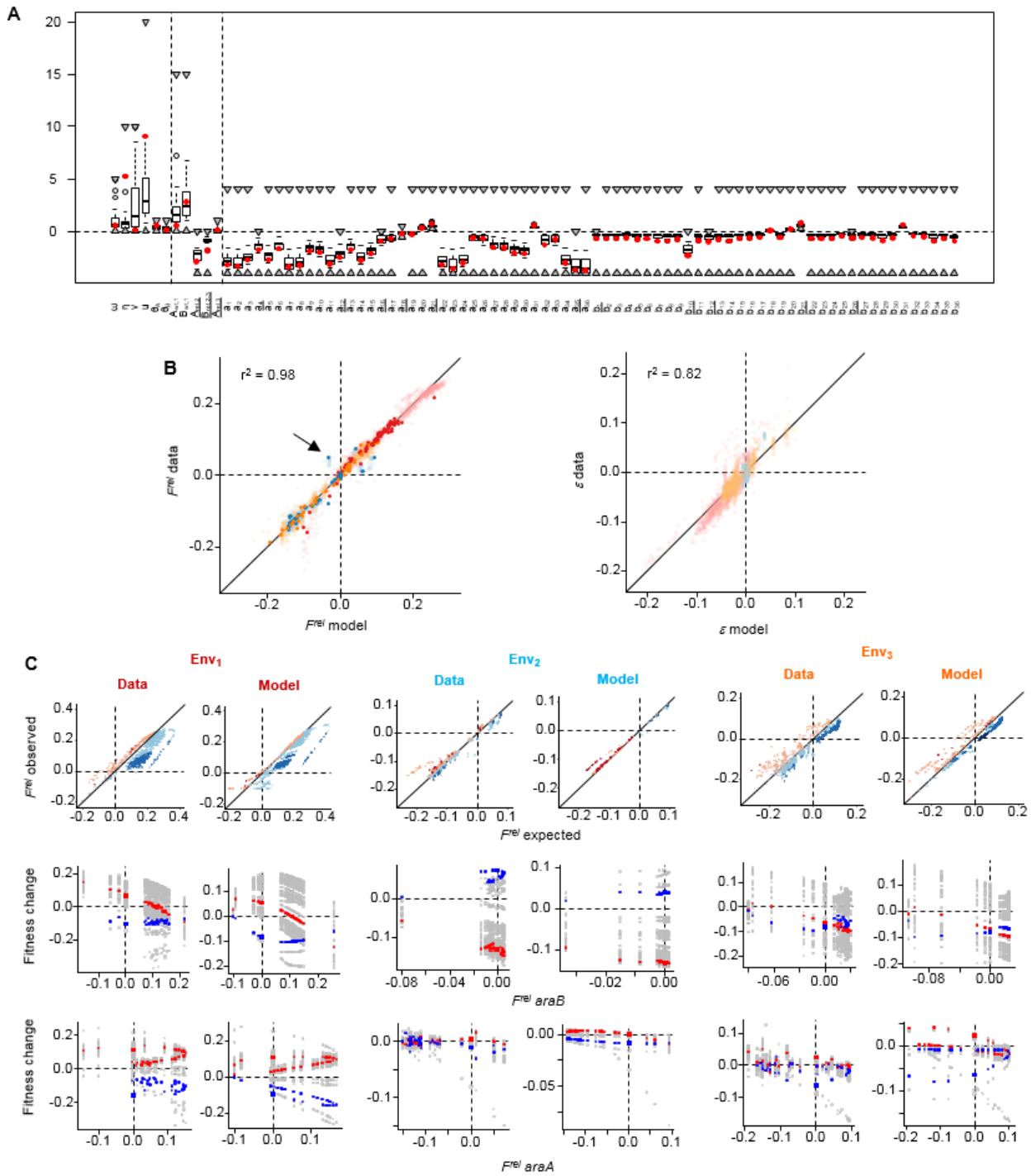


Fig. 4.S8.

Performance of flux-toxicity-expression burden model. (**A**) Parameter estimates. Boxplots show distributions from the best 2.5% of Markov chains (n = 800 chains). Red points show parameter estimates from the best chain. Triangles show bounds of the uniform prior distributions. Parameter descriptions are given in Data S2. Vertical dashed lines separate the fitness function parameters, parameters describing wildtype expression levels across environments, and the expression effect (natural logarithm) of mutations (ordered as in Fig. 2B), from left to right. Prior bounds of

underlined expression effect parameters were guided by expression measurements. The majority of mutations in both promoters are predicted to decrease expression (expression effect < 0), which is not surprising as the (identical) "wildtype" RNA polymerase-binding sequences are a Hamming distance of only 2 away from the bacterial consensus sequence, indicating near-maximal binding strength. (**B**) Correlations between observed values and those predicted by the model. Left – fitness (n = 4,079 mutant measurements); right – epistasis (n = 3,813 mutation pair measurements); $p < 2.2e\text{-}16$ for both. Opaque points are single-mutants. Points are coloured by environment, as in Fig. 4A. Arrow points to genotypes containing a qualitative outlier mutation, $P_{LtetO\text{-}1}$-*araA* G7A, which is also the only mutation to be beneficial in all environments (Fig. 2B), presumably because its effect on *expression* depends on the environment (supported by the fact that it lies in a repressor binding site (Fig. S1A)). (**C**) Comparison of epistatic trends from experimental data and model, across environments. Top row – as for Fig. 3A; lower two rows – as for Fig. 3B (same 4 alleles coloured in all environments). Looping is explained by single-mutants lying on two sides of a phenotypic optimum.

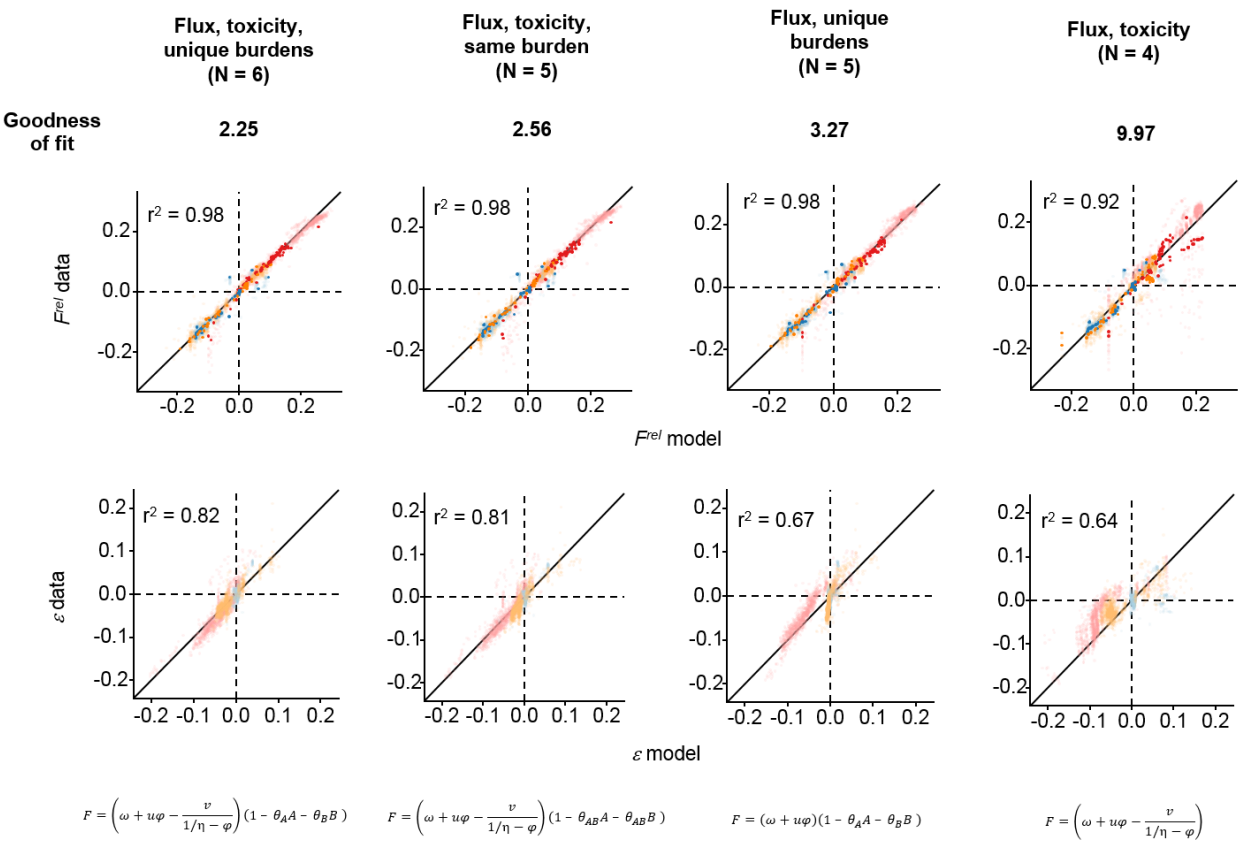


Fig. 4.S9.

Goodness-of-fit comparison of different phenotype-fitness models. Correlations between observed values and those predicted by different model variations. Top row – fitness (n = 4,079 mutant measurements); bottom row – epistasis (n = 3,813 mutation pair measurements); $p < 2.2e\text{-}16$ for all. Opaque points are single-mutants. Points are coloured by environment, as in Fig. 4A. Goodness of fit is calculated as the sum of the squared differences between all observed fitness effects and epistasis coefficients and those predicted by the models (n = 7,892). N is the number of parameters defining the fitness function for each model. From left to right: complete model used in main text; as complete model, except that expression burden per activity unit is the same for both proteins; as complete model, but no toxicity; as complete model, but no expression burden.

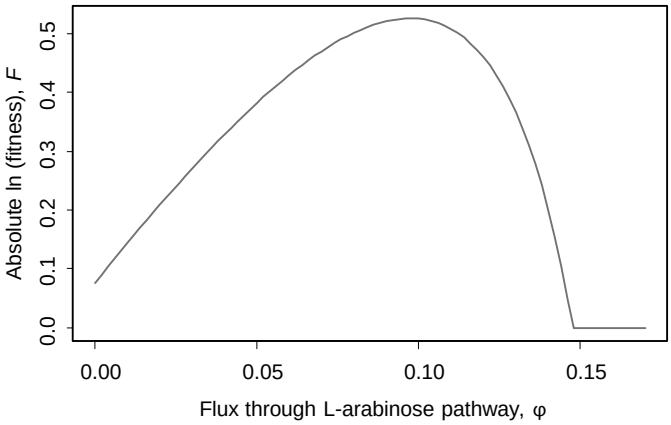


Fig. 4.S10.

Flux-fitness relationship predicted by model. The fitted model results in the existence of a particular flux that is optimal for fitness (*24*, *25*). As the flux exceeds this optimum, the rapid accumulation of the toxic intermediate, L-ribulose-5-phosphate, causes a steep fitness decline. The flux-fitness function diverges at very high fluxes (above the predicted range of our dataset), presumably as one or more of the simplifying assumptions underlying the enzyme activity-flux function starts to break down.

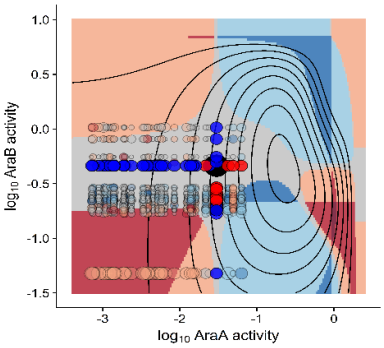


Fig. 4.S11.

Fitness surface coloured by predicted epistasis category in Env$_2$. As for Fig. 4C. The vast majority of interactions in this environment are predicted, and observed, to be weak (see blue points in Fig. S8C, right panel).

4.6 Supplementary tables

Plasmid name	Description	DNA fragments used for construction (this study)	Construction method / Supplier	Antibiotic used for selection	Accidental mutations / Sequence conflicts
pKD3 (*38*)	PCR template plasmid for Datsenko-Wanner (*38*) gene deletion, containing a *cat* Cm-resistance cassette flanked by *FRT* sites and an R6Kγ *pir*-dependent *ori*. Also used as PCR template for *bla* amplification in library barcoding step	-	Lab stocks	Cm	-
pKD46 (*38*)	Plasmid with L-arabinose-inducible λ Red expression cassette for Datsenko-Wanner (*38*) recombineering; temperature-sensitive *ori* (repA101ts) for easy curing	-	Lab stocks	Amp	-
pCP20 (*38*)	Plasmid with yeast *FLP* recombinase expression cassette for Datsenko-Wanner (*38*) resistance-gene excision; temperature-sensitive *ori* (repA101ts) for easy curing	-	Lab stocks	Amp	-
pSkunk3-BLA (*59*)	Phagemid containing *p15A* and *f1 ori*s, *bla* β-lactamase gene and *aadA1* Str/Sp-resistance gene. Used for backbone (*f1* phage *ori* not exploited in this study)	-	A. Birgy	Str	-
pZS4Int-1 (*41*)	*pSC101 ori*, *lacI* and *tetR* repressor genes under constitutive promoters, *attP* phage λ attachment site and *aadA1* Str/Sp -resistance gene. Used for *lacI* and *tetR*	-	A. Decrulle and I. Matic	Sp	G -> C at +246 of *tetR* ORF, causing Lys82 -> Asn82 (reported in other constructs, including reference (*60*)); 2 small insertions between *tetR* stop codon and its T1 terminator
pKH1503a	pSkunk3-BLA backbone, with *bla* replaced by: *araBA* under $P_{LlacO-1}$ inducible promoter (*41*) and *lacI* and *tetR* repressor genes under constitutive promoters (*41*)	pSkunk-bkb, aKH150312a, aKH150312b	Gibson Assembly	Str	-
pKH1503a^{evo}	Plasmid purified from a single colony (MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo} *[pKH1503a^{evo}]*) isolated after adaptation to alternating D- and L-arabinose. Sanger sequencing of *araBA*, *tetR* and *lacI*, along with their regulatory regions, revealed a single G –> C substitution in the 2nd *lac*O1 operator (-23 from TSS, in notation of reference (*41*)). This was found in 3/3 colonies tested from the evolved population, and was deliberately included in all future $P_{LlacO-1}$ -containing plasmids of this study (it was found through growth and expression measurements to still allow titratable expression control by IPTG)	-	Purified from a single colony isolated after MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* *[pKH1503a]* adaptation to alternating D- and L-arabinose	Str	-
pKH1511c	pKH1503a^{evo} backbone (rather than pKH1503a backbone, to exploit any unseen adaptive mutations arising during adaptation), with $P_{LlacO-1}$-*araBA* replaced by *araA* and *araB* in divergent orientation and promoter-less, separated by SacI and XhoI restriction sites to allow easy insertion of divergent promoters	aKH151120a, aKH151120b, aKH151120c	Restriction-ligation	Str	C -> A substitution (synonymous) at +1638 of *araB* ORF
pSW23T::*attP* (*61*)	*oriV*$_{R6Kγ}$ (*pir*-dependent replication), *attP* phage λ attachment site, *cat* Cm-resistance gene. Used for *pir*-dependent backbone to avoid template plasmid carryover during cloning	-	A. Soler and D. Mazel	Cm	-
pBSK-BSD1	pBluescript SK phagemid containing *pUC* and *f1 ori*s, *bsd* Bsd-resistance cassette and *bla* β-lactamase gene. Used for *bsd*	-	A. Couce (gene synthesis by Epoch Life Science, Inc, TX, USA)	Amp	-
pKH1511d	pSW23T::*attP* with *bsd* Bsd-resistance cassette inserted into multiple cloning site. Used to avoid plasmid carryover during future *bsd* cloning	pSW23T::*attP*-bkb, aKH151126a	Gibson Assembly	Cm	-

Table 4.S1.

Plasmids used in this study. Amp: ampicillin (100 µg/ml); Bsd: blasticidin; Cm: chloramphenicol (10 µg/ml); Spec: spectinomycin (50 µg/ml); Str: streptomycin (50 µg/ml).

DNA fragment name	Description/Creation	PCR template or digested plasmid	Primers used for PCR (blank if fragment comes directly from plasmid digestion)	Restriction enzymes used (either post-PCR or directly on plasmid)
pSkunk-bkb	pSkunk3 backbone, containing *ori*s and *aadA1* Str/Sp-resistance gene. Double-digest of pSkunk3-BLA to excise *bla*, followed by gel-extraction of backbone fragment	pSkunk3-BLA (*59*)	-	EcoRV, SpeI
aKH150312a	*lacI-tetR* constitutive expression cassette (*inc*. T1 terminator), with a downstream extension overlapping the SpeI extremity of pSkunk-bkb. PCR-amplification; overlap introduced on reverse primer	pZS4Int-1 (*41*)	oKH150312a, oKH150312b	-
aKH150312b	$P_{LlacO-1}$-*araBA* bicistronic cassette (*inc*. BBa_B1002 artificial terminator (BioBrick Foundation)), with an upstream extension overlapping the EcoRV extremity of pSkunk-bkb and a downstream extension overlapping the upstream extremity of aKH150312a. PCR-amplification; overlaps, $P_{LlacO-1}$ and BBa_B1002 all introduced on primers	*E. coli* K12 MG1655 genomic DNA	oKH150312c, oKH150312e	-
aKH151120a	pKH1503a^{evo} backbone, containing *oris*, *aadA1* Str/Sp-resistance gene and *lacI-tetR* ($P_{LlacO-1}$-*araBA* removed), with a downstream extension containing an NcoI site. PCR-amplification; extension introduced on reverse primer	pKH1503a^{evo}	oKH150312a, oKH151120a	SphI, NcoI
aKH151120b	*araB* coding region followed by BBa_B1004 artificial terminator (BioBrick Foundation), with an upstream extension containing SacI and XhoI restriction sites and a downstream extension containing an SphI restriction site. PCR-amplification; extensions and BBa_B1004 introduced on primers	pKH1503a^{evo}	oKH151120b, oKH151120c	SacI, SphI
aKH151120c	*araA* coding region followed by BBa_B1002 artificial terminator (BioBrick Foundation), with an upstream extension containing a SacI restriction site and a downstream extension containing an NcoI restriction site. PCR-amplification; extensions introduced on primers	pKH1503a^{evo}	oKH151120d, oKH151120e	SacI, NcoI
pSW23T::*attP*-bkb	Linearised pSW23T::*attP*. Double-digest of pSW23T::*attP* at Multiple Cloning Site	pSW23T::*attP* (*61*)	-	SpeI, SacII
aKH151126a	*bsd* Bsd-resistance cassette (*inc*. T1 terminator), with an upstream extension overlapping the SacII extremity of pSW23T::*attP*-bkb and a downstream extension overlapping the SpeI extremity of pSW23T::*attP*-bkb. PCR-amplification; overlaps introduced on primers	pBSK-BSD1	oKH151126a, oKH151203a	-

Table 4.S2.

DNA fragments used for cloning in this study.

Primer name	Sequence (5' -> 3')
oKH150202d	ATGGCAGAAATTCGAAAGC
oKH150312a	GCGGCATGCATTTACGTTGA
oKH150312b	AGCGCGTCGGCCGGTCGAATGCATAAGCTTACTAACTAGTGAGAGCGTTCACCGACAAAC
oKH150312c	AGCCAGAAAACCGAATTTTGCTGGGTGGGCTAACGATATCAATTGTGAGCGGATAACAATTGACATTGTGAGCGGATAACAAGATACTG AGCACACCCGTTTTTTTGGATGGAGTG
oKH150312e	TTTTGCACCATTCGATGGTGTCAACGTAAATGCATGCCGCGCGAAAAAACCCCGCCGAAGCGGGGTTTTTTGCGTTAGCGACGAAACC CGTAATAC
oKH150401c	ATTCATTAATGCAGCTGGC
oKH151120a	TTTTTCCATGGGATATCGTTAGCCCACCCAG
oKH151120b	TTTTTGAGCTCCACAGCTAACCTCGAGACCCGTTTTTTTGGATGGAGTG
oKH151120c	TTTTTGCATGCCGCGCGGCAAAACCCCGCCGAAGCGGGGTTTTCGGCGTTATAGAGTCGCAACGGCCT
oKH151120d	TTTTTGAGCTCTGCGACTCTATAAGGACACG
oKH151120e	TTTTTCCATGGGCGAAAAAACCCCGCCGA
oKH151126a	GATAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCCACTAGTGCGGCCGCGTGAGCCAGTGTGACTCTAGT
oKH151203a	CGTTTTATTTGATGCCTCTAGCACGCGTACCATGGAGCTCCACCGCGGATAGGAACTTCACGCTAGGG
KO-araBA-fwd	ACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGATGGAGTGAAACGATGGTGTAGGCTGGAGCTGCTTC
KO-araBA-rev	ATCAGGCGTTACATACCGGATGCGGCTACTTAGCGACGAAACCCGTAATACATATGAATATCCTCCTTAG
verif-araBA-fwd	TTGCATCAGACATTGCCGTC
verif-araBA-rev	GTTGGCTTCTAATACCTGGCG
KO-lacIZYA-fwd	GTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATGTGGTGTAGGCTGGAGCTGCTTC
KO-lacIZYA-rev	AGCGCAGCGTATCAGGCAATTTTTATAATTTAAACTGACGATTCAACTTTCATATGAATATCCTCCTTAG
verif-lacIZYA-fwd	GTGATGACTATCAACTGGCAC
verif-lacIZYA-rev	CTATTGCTGGCAAGCTGGTG
KO-fucK-fwd	TCCGGCTACCGGGCCTGAACAAGCAAGAGTGGTTAGCCGGATAAGCAATGGTGTAGGCTGGAGCTGCTTC
KO-fucK-rev	AAATTAACGGCGAAATTGTTTTCAGCATTTCACACTTCCTCTATAAATTCCATATGAATATCCTCCTTAG
verif-fucK-fwd	AACGCACCAACTCAACCTGG
verif-fucK-rev	TTGATGCGGATGATGTCAGG
oBarcodeBla-fwd	TTTTTACTAGTGGCGCGCCGTCGACTTNNNNNATNNNNNATNNNNNATNNNNNATNNNNNATCTTCAGATCCTCTACGCCGG
oBarcodeBla-rev	TACACTCCGCTAGCGCTGATGTCCGGCGGTGCCAGGTGGCACTTTTCGGG
oLinkBarcode-fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNCGTGTCCTTATAGAGTCGCAG
oLinkBarcode-rev	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNGTCCGGCGTAGAGGATCTG
oBarcodeSeq-fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNGTGAACGCTCTCACTAGTGG
oBarcodeSeq-rev	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNCAAGATCCGGCCACGATGC
c1 (*38*)	TTATACGCAAGGCGACAAGG
c2 (*38*)	GATCTTCCGTCACAGGTAGG

Table 4.S3.

PCR Primers used in this study, excluding those used directly for promoter mutagenesis.

Strain name	Description/Usage	Genotype	Engineering method / Supplier	Antibiotic / supplements used for selection / adaptation
K12 MG1655	"Wildtype" laboratory strain	F^- λ^- *ilvG*$^-$ *rfb*-50 *rph*-1	A. Couce; Coli Genetic Stock Centre #6300	-
PIR1	*pir*-expressing strain for cloning and maintenance of *pir*-dependent plasmids (thymidine auxotroph)	F^- Δ*lac*169 *rpo*S(am) *rob*A1 *cre*C510 *hsd*R514 *end*A *rec*A1 *uidA*(ΔMlu*I*)*::pir-116*	A. Soler and D. Mazel	Erm + dT
DH5α	Standard strain for plasmid cloning and maintenance	F^- λ^- Φ80/*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rK^-, mK^+) *pho*A *sup*E44 *thi*-1 *gyr*A96 *rel*A1	Lab stock	-
DH5α Δ*araBA*::*cat*	Intermediate for construction of DH5α Δ*araBA*	DH5α Δ*araBA*::*cat*	Datsenko-Wanner (pKD46) (*38*)	Cm
DH5α Δ*araBA*	Preliminary tests; used as alternative to DH5α in this study	DH5α Δ*araBA*::*FRT*	Datsenko-Wanner (pCP20) (*38*)	-
MG1655 Δ*araBA*::*cat*	Intermediate for construction of MG1655 Δ*araBA*	MG1655 Δ*araBA*::*cat*	Datsenko-Wanner (pKD46) (*38*)	Cm
MG1655 Δ*araBA*	Preliminary tests; intermediate for construction of MG1655 Δ*araBA* Δ*lacIZYA*::*cat* and MG1655 Δ*araBA* D-ara$^{+/evo}$	MG1655 Δ*araBA*::*FRT*	Datsenko-Wanner (pCP20) (*38*)	-
MG1655 Δ*araBA* Δ*lacIZYA*::*cat*	Preliminary tests	MG1655 Δ*araBA*::*FRT* Δ*lacIZYA*::*cat*	Datsenko-Wanner (pKD46) (*38*)	Cm
MG1655 Δ*araBA* D-ara$^{+/evo}$	MG1655 Δ*araBA* derivative able to metabolise D-arabinose using genes of the *fuc* operon, due to a *fucR* mutation rendering the operon D-arabinose-inducible. Further adapted to D-arabinose for ~ 60 generations, and a single colony isolated. Intermediate for construction of MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK*::*cat*	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo}	Incubated in M9 + D-arabinose until visible growth (6 days). Then, serially transferred in M9 + D-arabinose for ~ 60 generations before isolation of a single colony (see refs. (*33–35*))	D-arabinose
MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK*::*cat*	Intermediate for construction of MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK*	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo} Δ*fucK*::*cat*	Datsenko-Wanner (pKD46) (*38*)	Cm
MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK*	Intermediate for construction of MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat*	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo} Δ*fucK*::*FRT*	Datsenko-Wanner (pCP20) (*38*)	-
MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat*	Intermediate for construction of MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat [pKH1503a]*	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo} Δ*fucK*::*FRT* Δ*lacIZYA*::*cat*	Datsenko-Wanner (pKD46) (*38*)	Cm
MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat [pKH1503a]*	Intermediate for construction of MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo} *[pKH1503a]*	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo} Δ*fucK*::*FRT* Δ*lacIZYA*::*cat [pKH1503a]*	Plasmid transformation (electroporation)	Str
MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo} *[pKH1503a^{evo}]*	MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat [pKH1503a]* derivative adapted to alternating D- and L-arabinose in presence of 10µM IPTG for ~45 generations, and a single large colony isolated. Evolved plasmid (pKH1503a^{evo}) used as template for further plasmid constructs; intermediate for construction of MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo}	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo} Δ*fucK*::*FRT* Δ*lacIZYA*::*cat* D/L-ara^{evo} *[pKH1503a^{evo}]*	Incubated in M9 + 10µM IPTG + D-arabinose until visible growth (2 weeks). Then, serially transferred in M9 + 10µM IPTG + alternating D- and L-arabinose for ~45 generations before isolation of a single large colony	Alternating D- and L-arabinose (+ IPTG + Str)
MG1655 Δ*araBA* D-ara$^{+/evo}$ Δ*fucK* Δ*lacIZYA*::*cat* D/L-ara^{evo}	Final engineered/adapted plasmidless host strain for barcoded promoter-mutant plasmid library; able to utilize L-arabinose in presence of plasmid-expressed AraA and AraB, and D-arabinose in presence of plasmid-expressed AraB	MG1655 Δ*araBA*::*FRT* *fucR*$^{D\text{-}ara}$ D-ara^{evo} Δ*fucK*::*FRT* Δ*lacIZYA*::*cat* D/L-ara^{evo}	Plasmid curing	Ribitol (*39*) (+ IPTG + Cm)

Table 4.S4.

***E. coli* strains used in this study.** Cm: chloramphenicol (10 µg/ml); dT: thymidine (30 µg/ml); Erm: erythromycin (20 µg/ml); Str: streptomycin (50 µg/ml); IPTG: isopropyl β-D-1-thiogalactopyranoside. For adaptation, D- and L-arabinose were present at 0.3% and 0.2% w/v, respectively.

Primer name	Sequence (5' -> 3')
oPtetLib-fwd-1	TTTTTGAGCTCGTGCTC**AGTATC**TCTATCACTGATAGGGA**TGTCA*N***TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-2	TTTTTGAGCTCGTGCTC**AGTATC**TCTATCACTGATAGGGA**TGTC*N*A**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-3	TTTTTGAGCTCGTGCTC**AGTATC**TCTATCACTGATAGGGA**TGT*N*AA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-4	TTTTTGAGCTCGTGCTC**AGTATC**TCTATCACTGATAGGGA**TG*N*CAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-5	TTTTTGAGCTCGTGCTC**AGTATC**TCTATCACTGATAGGGA**T*N*TCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-6	TTTTTGAGCTCGTGCTC**AGTATC**TCTATCACTGATAGGGA***N*GTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-7	TTTTTGAGCTCGTGCTC**AGTAT*N***TCTATCACTGATAGGGA**TGTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-8	TTTTTGAGCTCGTGCTC**AGTA*N*C**TCTATCACTGATAGGGA**TGTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-9	TTTTTGAGCTCGTGCTC**AGT*N*TC**TCTATCACTGATAGGGA**TGTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-10	TTTTTGAGCTCGTGCTC**AG*N*ATC**TCTATCACTGATAGGGA**TGTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-11	TTTTTGAGCTCGTGCTC**A*N*TATC**TCTATCACTGATAGGGA**TGTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPtetLib-fwd-12	TTTTTGAGCTCGTGCTC***N*GTATC**TCTATCACTGATAGGGA**TGTCAA**TCTCTATCACTGATAGGGAGGCGCGCCGTGAGCCAGTGT GACTCTAGTAG
oPlacLib-rev-1	TTTTTCTCGAGGTGCTC**AGTATC**TTGTTATCCGATCACAA**TGTCA*N***TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-2	TTTTTCTCGAGGTGCTC**AGTATC**TTGTTATCCGATCACAA**TGTC*N*A**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-3	TTTTTCTCGAGGTGCTC**AGTATC**TTGTTATCCGATCACAA**TGT*N*AA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-4	TTTTTCTCGAGGTGCTC**AGTATC**TTGTTATCCGATCACAA**TG*N*CAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-5	TTTTTCTCGAGGTGCTC**AGTATC**TTGTTATCCGATCACAA**T*N*TCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-6	TTTTTCTCGAGGTGCTC**AGTATC**TTGTTATCCGATCACAA***N*GTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-7	TTTTTCTCGAGGTGCTC**AGTAT*N***TTGTTATCCGATCACAA**TGTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-8	TTTTTCTCGAGGTGCTC**AGTA*N*C**TTGTTATCCGATCACAA**TGTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-9	TTTTTCTCGAGGTGCTC**AGT*N*TC**TTGTTATCCGATCACAA**TGTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-10	TTTTTCTCGAGGTGCTC**AG*N*ATC**TTGTTATCCGATCACAA**TGTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-11	TTTTTCTCGAGGTGCTC**A*N*TATC**TTGTTATCCGATCACAA**TGTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG
oPlacLib-rev-12	TTTTTCTCGAGGTGCTC***N*GTATC**TTGTTATCCGATCACAA**TGTCAA**TTGTTATCCGCTCACAATTATAGGAACTTCACGCTAGGG

Table 4.S5.

Forward and reverse primer sets for promoter mutagenesis. -35 and -10 RNA polymerase-binding hexamers are in bold. N (italicised) denotes a mix of all 4 bases.

4.7 References and notes

1. C. R. Scriver, P. J. Waters, Monogenic traits are not simple: lessons from phenylketonuria. *Trends in Genetics*. **15**, 267–272 (1999).

2. J. L. Badano, N. Katsanis, Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews Genetics*. **3**, 779–789 (2002).

3. K. M. Dipple, E. R. B. McCabe, Modifier Genes Convert "Simple" Mendelian Disorders to Complex Traits. *Molecular Genetics and Metabolism*. **71**, 43–50 (2000).

4. B. Lanpher, N. Brunetti-Pierri, B. Lee, Inborn errors of metabolism: the flux from Mendelian to complex diseases. *Nature Reviews Genetics*. **7**, 449–459 (2006).

5. D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith, H. Kehrer-Sawatzki, Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics*. **132**, 1077–1130 (2013).

6. J. Vockley, P. Rinaldo, M. J. Bennett, D. Matern, G. D. Vladutiu, Synergistic Heterozygosity: Disease Resulting from Multiple Partial Defects in One or More Metabolic Pathways. *Molecular Genetics and Metabolism*. **71**, 10–18 (2000).

7. C. A. Argmann, S. M. Houten, J. Zhu, E. E. Schadt, A next generation multiscale view of inborn errors of metabolism. *Cell Metabolism*. **23**, 13–26 (2016).

8. J. Hou *et al.*, The hidden complexity of Mendelian traits across yeast natural populations (2016), doi:10.1101/039693.

9. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature*. **461**, 747–753 (2009).

10. E. J. O'Brien, J. M. Monk, B. O. Palsson, Using Genome-scale Models to Predict Biological Capabilities. *Cell*. **161**, 971–987 (2015).

11. L. Xu, B. Barker, Z. Gu, Dynamic epistasis for different alleles of the same gene. *Proceedings of the National Academy of Sciences*. **109**, 10420–10425 (2012).

12. M.-J. Favé *et al.*, Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nature Communications*. **9** (2018), doi:10.1038/s41467-018-03202-2.

13. R. Schleif, Regulation of the L-arabinose operon of Escherichia coli. *Trends in Genetics*. **16**, 559–565 (2000).

14. J. Ewald, M. Bartl, T. Dandekar, C. Kaleta, Optimality principles reveal a complex interplay of intermediate toxicity and kinetic efficiency in the regulation of prokaryotic metabolism. *PLOS Computational Biology*. **13**, e1005371 (2017).

15. P. J. O'Brien, R. Bruce, *Endogenous Toxins: Targets for Disease Treatment and Prevention* (Wiley-VCH, Hoboken, NJ, USA, 2010), vol. 1.

16. E. Englesberg *et al.*, L-arabinose-sensitive, L-ribulose 5-phosphate 4-epimerase-deficient mutants of Escherichia coli. *Journal of Bacteriology*. **84**, 137–146 (1962).

17. J. B. Wolf, E. D. Brodie III, M. J. Wade, Eds., in *Epistasis and the Evolutionary Process* (Oxford University Press, NY, USA, 2000), p. 10.

18. J. A. G. M. de Visser, T. F. Cooper, S. F. Elena, The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences*. **278**, 3617–3624 (2011).

19. D. Berger, E. Postma, Biased Estimates of Diminishing-Returns Epistasis? Empirical Evidence Revisited. *Genetics*. **198**, 1417–1420 (2014).

20. A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, T. F. Cooper, Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science*. **332**, 1193–1196 (2011).

21. H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segre, C. J. Marx, Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science*. **332**, 1190–1192 (2011).

22. C. Li, W. Qian, C. J. Maclean, J. Zhang, The fitness landscape of a tRNA gene. *Science*. **352**, 837–840 (2016).

23. D. E. Dykhuizen, A. M. Dean, D. L. Hart, Metabolic flux and fitness. *Genetics*. **115**, 25–31 (1987).

24. A. G. Clark, Mutation-selection balance and Metabolic Control Theory. *Genetics*. **129**, 909–923 (1991).

25. E. Szathmary, Do deleterious mutations act synergistically? Metabolic Control Theory provides a partial answer. *Genetics*. **133**, 127–132 (1993).

26. E. Dekel, U. Alon, Optimality and evolutionary tuning of the expression level of a protein. *Nature*. **436**, 588–592 (2005).

27. M. Kafri, E. Metzl-Raz, G. Jona, N. Barkai, The Cost of Protein Production. *Cell Reports*. **14**, 22–31 (2016).

28. H.-H. Chou, N. F. Delaney, J. A. Draghi, C. J. Marx, Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. *PLoS Genetics*. **10**, e1004149 (2014).

29. H. Kacser, J. A. Burns, The Molecular Basis Of Dominance. *Genetics*. **97**, 639–666 (1981).

30. F. Blanquart, G. Achaz, T. Bataillon, O. Tenaillon, Properties of selected mutations and genotypic landscapes under Fisher's Geometric Model. *Evolution*. **68**, 3537–3554 (2014).

31. S. Rozen, H. Skaletsky, Primer3 on the WWW for General Users and for Biologist Programmers. *Methods in Molecular Biology*. **132**, 365–386 (2000).

32. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*. **6**, 343–345 (2009).

33. D. J. Leblanc, R. P. Mortlock, The metabolism of d-arabinose: Alternate kinases for the phosphorylation of d-ribulose in Escherichia coli and Aerobacter aerogenes. *Archives of Biochemistry and Biophysics*. **150**, 774–781 (1972).

34. D. J. LeBlanc, R. P. Mortlock, Metabolism of D-Arabinose: a New Pathway in Escherichia coli. *Journal of Bacteriology*. **106**, 90–96 (1971).

35. D. J. LeBlanc, R. P. Mortlock, Metabolism of D-Arabinose: Origin of a D-Ribulokinase Activity in Escherichia coli. *Journal of Bacteriology*. **106**, 82–89 (1971).

36. G. Fritz *et al.*, Single Cell Kinetics of Phenotypic Switching in the Arabinose Utilization System of E. coli. *PLoS ONE*. **9**, e89532 (2014).

37. A. Khlebnikov, J. D. Keasling, Effect of lacY Expression on Homogeneity of Induction from the Ptac and Ptrc Promoters by Natural and Synthetic Inducers. *Biotechnology Progress*. **18**, 672–674 (2002).

38. K. A. Datsenko, B. L. Wanner, One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences*. **97**, 6640–6645 (2000).

39. L. Katz, Selection of AraB and AraC Mutants of Escherichia coli B/r by Resistance to Ribitol. *Journal of Bacteriology*. **102**, 593–595 (1970).

40. G. A. Wray, The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*. **8**, 206–216 (2007).

41. R. Lutz, Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*. **25**, 1203–1210 (1997).

42. R. K. Shultzaberger, D. S. Malashock, J. F. Kirsch, M. B. Eisen, The Fitness Landscapes of cis-Acting Binding Sites in Different Promoter and Environmental Contexts. *PLoS Genetics*. **6**, e1001042 (2010).

43. J. B. Kinney, A. Murugan, C. G. Callan, E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*. **107**, 9158–9163 (2010).

44. R. C. Brewster, D. L. Jones, R. Phillips, Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. *PLoS Computational Biology*. **8**, e1002811 (2012).

45. L. Bintu *et al.*, Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*. **15**, 116–124 (2005).

46. M. Lagator, T. Paixão, N. H. Barton, J. P. Bollback, C. C. Guet, On the mechanistic nature of epistasis in a canonical cis-regulatory element. *eLife*. **6** (2017), doi:10.7554/eLife.25192.

47. M. Kimura, A. Takatsuki, I. Yamaguchi, Blasticidin S deaminase gene from Aspergillus terreus (BSD): a new drug resistance gene for transfection of mammalian cells. *Biochimica et Biophysica Acta*. **1219**, 653–659 (1994).

48. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature*. **533**, 397–401 (2016).

49. S. F. Levy *et al.*, Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*. **519**, 181–186 (2015).

50. M. Goldsmith, C. Kiss, A. R. M. Bradbury, D. S. Tawfik, Avoiding and controlling double transformation artifacts. *Protein Engineering Design and Selection*. **20**, 315–318 (2007).

51. C. Pusch, H. Schmitt, N. Blin, Increased cloning efficiency by cycle restriction—ligation (CRL). *Technical Tips Online*. **2**, 35–37 (1997).

52. P. D. Schloss *et al.*, Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*. **75**, 7537–7541 (2009).

53. A. Meyerhans, J.-P. Vartanian, S. Wain-Hobson, DNA recombination during PCR. *Nucleic Acids Research.* **18**, 1687–1691 (1990).

54. R. T. Hietpas, J. D. Jensen, D. N. A. Bolon, Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences.* **108**, 7896–7901 (2011).

55. T. N. M. Nguyen, Q. G. Phan, L. P. Duong, K. P. Bertrand, Effects of carriage and expression of the Tn10 tetracycline-resistance operon on the fitness of Escherichia coli K12. *Molecular Biology and Evolution.* **6**, 213–225 (1989).

56. A. L. Koch, The protein burden of lac operon products. *Journal of Molecular Evolution.* **19**, 455–462 (1983).

57. M. S. Bienick *et al.*, The Interrelationship between Promoter Strength, Gene Expression, and Growth Rate. *PLoS ONE.* **9**, e109105 (2014).

58. S. Brroks, A. Gelman, G. L. Jones, X.-L. Meng, *Handbook of Markov Chain Monte Carlo* (Chapman and Hall/CRC, Boca Raton, Fl, USA, 2011), *Handbooks of Modern Statistical Methods.*

59. E. Firnberg, M. Ostermeier, PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS ONE.* **7**, e52031 (2012).

60. M. Gossen, H. Bujard, Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences.* **89**, 5547–5551 (1992).

61. A. Pant *et al.*, Effect of LexA on Chromosomal Integration of CTXφ in Vibrio cholerae. *Journal of Bacteriology.* **198**, 268–275 (2016).

Acknowledgments: We thank A. Birgy, A. Decrulle, I. Matic, M. Deyell, A. Soler and D. Mazel for providing genetic material and technical advice, A. Baron, J. Chatel, A. Bridier-Nahmias and the CRI cytometry facility for technical assistance, and L.-M. Chevin, B. Gaut, E. Denamur and C. Landry for critical reading of the manuscript. MiSeq sequencing was performed using equipment provided by the Genetics Department of Bichat-Claude Bernard Hospital. **Funding:** This work was supported by the European Research Council under the European Union's Seventh Framework Programme (ERC grant 310944 to O.T.). H.K. was supported by the Ecole Doctorale Frontières du Vivant (FdV) – Programme Bettencourt. **Author contributions:** H.E.K., P.N. and O.T. conceived the idea for the experiment; H.E.K. and O.T.

designed the experiments; H.E.K., C.E., A.C., A.E.C., M.A.M., G.G. and H.L.N. performed the experiments; H.E.K., P.N. and O.T. performed the analyses; H.E.K., P.N. and O.T. wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All genotype fitness estimates, along with their bootstrap 95% CIs and the number of replicates used to compute them, are provided in Supplementary Table 6. Raw and processed sequencing data has been submitted to GEO (accession number GSE115725). Custom code used in this study is available from the authors upon request.

5 Discussion

This thesis has been concerned with the properties emerging from the mapping between genotype and phenotype, a fundamental relationship in Biology whose conceptual origins date back to over 150 years ago (Mendel, 1866). Although entirely abstract initially, I described in the Introduction how Biology's molecular revolution endowed this mapping with a detailed material basis, encapsulated by Francis Crick's famed Central Dogma (Crick, 1958). This states, in its simplest form, that the genotype is nucleic acid and the phenotype results from the activity of proteins; protein residue sequence is encoded for by nucleic acid, *via* the genetic code, but such information cannot flow in reverse, nor between proteins.

Such an understanding raised the prospect that the purely statistical inferences made by pre-revolution geneticists might now be explained mechanistically: careful structural and biochemical characterisation of protein variants might reveal which precise genotypes would be expected to result in which precise phenotypes. The mass of data on molecular mechanisms collected over the last decades, along with the recent explosion of genome-sequence data, has revealed however that, except maybe for the very simplest of cases (*eg.* an early stop codon in an essential gene), predicting phenotype from genotype from first principles remains a formidable task.

An appealing solution is to focus on small, well-understood model systems. Recent methodological advances in DNA synthesis and sequencing have now made it possible to score comprehensive genotype libraries for a variety of phenotypes ("deep-

mutational scanning") (Fowler and Fields, 2014; Hietpas et al., 2011). Such datasets provide a complete, or locally complete, picture of the genotype-phenotype map for these systems, allowing the rigorous testing of mechanistic models. The research conducted for this thesis therefore leveraged this technology to perform deep-mutational fitness scanning experiments on 3 different model bacterial systems: a global transcriptional regulator, CRP, an antibiotic-resistance enzyme, β-lactamase, and a small metabolic pathway, consisting of the enzymes, AraA and AraB. These different systems were chosen to illuminate the roles of different mechanistic features in shaping the genotype-fitness relationship (protein stability, regulatory wiring and metabolic flux).

Chapter 1 describes the genotype-fitness map for all single amino acid substitutions of CRP, in 4 environments. Although the system was somewhat artificial, being located on a multicopy plasmid due to technical limitations, it appears to have provided information on the underlying CRP activity-fitness function, as hoped. First, both the shape of the DFEs and the correlations between fitness effects in different environments point to the existence of an optimum, intermediate activity of CRP resulting in maximum fitness in a given environment. This makes some sense intuitively, as CRP is known to respond adaptively to changes in the environment by varying its activity, and is also in line with the results of a recent study (Towbin et al., 2017). An apparently undocumented property of the activity-fitness function,

however, is that this activity optimum appears to in fact represent a very broad fitness plateau, allowing genotypes to "find their way there" easily, even in highly maladaptive artificial environments. This robustness could well result from the known negative feedback in the CRP regulatory network (You et al., 2013), and encouragingly, genetic robustness has been associated with the negative feedback motif in some recent studies (Denby et al., 2012; Marciano et al., 2014, 2016). These results therefore suggest regulatory network architecture to have a profound impact on the genotype-fitness relationship, with negative feedback potentially expanding the space of equally fit genotypes. This in turn highlights the need to consider higher-level organization when considering the genotype-phenotype relationship for even a single gene. Unfortunately, time ran out before hypotheses could be tested rigorously, but this will be done in the near future.

Another promising direction for the CRP data is the classification of mutations that either increase existing CRP activity or behave in a way that is inconsistent with the existence of a single phenotypic dimension (*ie.* CRP activity), an analysis that is enabled by the particular experimental conditions used. These would represent mutations interfering with other cellular processes in unforeseen ways, and as such would be critically important for the predictability of mutation effects.

Finally, the original idea of the CRP project was to perform a deep-mutational scan for expression levels of CRP-regulated genes using a new microfluidic technology, but

this remains under development. Coupling such measurements with the fitness data in the future could allow fitness changes to be understood from a finer-grained phenotypic level: the expression of the set of CRP's target genes rather than just CRP activity itself.

Chapter 3 describes the results of a project closer to completion, which characterised fitness effects and epistasis in a library of >15,000 single- and double-mutations in an 11-residue α-helix of the model antibiotic-resistance enzyme, β-lactamase TEM-1. Epistasis was found to occur frequently between mutations, much of which could be explained by a simple thermodynamic model of protein stability that enabled classification of mutations as inactivating, destabilising, neutral or stabilising. The power of such global phenotype-fitness models to explain so much of the variance observed in fitness data across model systems has been thoroughly discussed in this thesis, but it was particularly surprising to find it hold true for a small structural region of a protein: this might have been expected to show more idiosyncratic trends of epistasis between mutations, driven by the physical proximity of affected residues and so specific local physicochemical interactions. These kinds of interactions, although rarer, were however also detected as deviations from the thermodynamic model. As expected, they were found to be more frequent when one of the mutations was close to the active site, and among mutations in directly contacting residues. Assessing the influence of these different flavours of epistasis, global and local, on evolutionary

processes and the predictability of mutation effects should prove to be an interesting future direction.

Chpater 3 describes the most developed project, which characterised the epistasis arising between expression variants of the metabolic genes, *araA* and *araB*. The experiments were designed to address the extent to which mutations in one gene can depend on the state of other functionally connected genes, using a well-defined metabolic pathway as a model. We found epistasis to be pervasive and surprisingly diverse, with a significant proportion of both positive and negative interactions, synergistic and antagonistic interactions, and magnitude and sign interactions. Further, a diversity of *trends* of epistasis were also detected (assessed by the correlations between fitness effects and epistasis).

Fortunately, there exists a rigorous mathematical framework with which to explore the behaviour of such molecular pathways: Metabolic Control Analysis. We thus tested our data against several simple phenotypic models based on Metabolic Control Analysis (Heinrich and Rapoport, 1974; Kacser and Burns, 1973), each considering different sets of molecular phenotypes as contributing to fitness. When pathway flux alone is considered as a phenotype, the resulting 2-enzyme activity-fitness function is concave monotonic, leading to a large fitness plateau (Dykhuizen et al., 1987). Such a model could not explain our data, in part because it makes sign interactions impossible, while we found them to be abundant. We therefore considered phenotypes that could result

in a cost to enzyme expression, which could form the non-monotonicity in the activity-fitness landscape required for sign epistasis. One costly phenotype that is well-known in the *E. coli* arabinose pathway is the toxicity of one of the intermediates, L-ribulose-5-phosphate (Englesberg et al., 1962). Another more general, and now well established, one is protein expression burden (Dekel and Alon, 2005; Koch, 1983; Stoebel et al., 2008). The introduction of each of these phenotypes into our metabolic flux model improved its fit with the data considerably, as expected, but we found that all 3 molecular phenotypes (flux, toxicity, expression burden) needed to be considered to explain the full extent of observed fitness and epistasis trends. The final model demonstrates how surprisingly complex patterns of intergenic fitness interactions can emerge from a relatively simple, smooth underlying phenotype-fitness surface. Viewed from another perspective, it reveals how epistasis can emerge through molecular pleiotropy: as seen in the Introduction, a single mutation is likely to affect several partially dependent molecular phenotypes, each of which bears a potentially independent contribution to fitness. This can then cause complex patterns at the level of fitness even if each individual phenotype-fitness function is simple and monotonic.

Together, these 3 research projects suggest that it may indeed be feasible to understand properties of the genotype-fitness relationship from the bottom-up, at least for model systems. In these datasets, smooth trends tend to prevail over idiosyncrasy, indicating that much of the genotype-fitness relationship could be understood from the

global shape of smooth underlying phenotype-fitness functions. On the flip side, we have seen that characterising the genotype-fitness relationship in different systems can be a powerful way to glean phenotypic insights: the inferred broad plateau in the activity-fitness function of a global regulator, the stability model describing the activity of antibiotic-resistance enzyme variants, and the joint contribution of flux, toxicity and expression burden to fitness for a metabolic pathway.

Future directions for the study of model genetic systems include the development of technical tools to ease the characterisation of fitness effects in the most natural settings possible. For example, improved precision genome engineering methods to avoid the use of multicopy plasmids, and robust complementation strategies to avoid the loss of null-mutants prior to competition when non-selective growth conditions cannot be found. As we have seen, for certain genes, such as global regulators, key properties of the genotype-fitness relationship may depend enormously on their position within a network of interacting components, and so studying them in their natural context could provide insight as to how the effect of mutations in individual genes is shaped by network structure.

Another exciting possibility is the large-scale characterisation of mutation effects at several phenotypic scales simultaneously (eg. protein stability, protein activity, flux, expression, cell morphology, fitness), which could enable a direct and complete mechanistic description of the translation of genotype into high-level traits. Indeed,

certain phenotypes, such as metabolic flux (Sauer, 2004) and the set of –omes, have received very little direct attention, mostly due to the enormous technical challeges and cost involved in their high-throughput measurement/coupling to genotype libraries. One promising candidate, however, is the transcriptome, due to the fact that it can be sequenced by RNA-seq: using the same emulsion-based technology that can enable distal genomic sites to be linked together for many single cells (Figure 1.21), the transcriptome can in principle be quantified for single cells whilst linking this information to the cells' genotypes with the use of unique cellular DNA-barcodes (Adamson et al., 2016; Dixit et al., 2016). The transcriptomic impact of random mutations or a large set of transcriptional regulator mutations, for example, could thus be rapidly assessed and even linked to high-throughput fitness measurements.

Finally, the systematic analysis of the effects of genome-wide combinations of point-mutations still appears far out of reach, but a feasible next step might be introducing synthetic promoter libraries like those used in (Keren et al., 2016) in front of *pairs* of genes across the genome and measuring the fitness effects. Although clearly artificial, and in some cases breaking regulatory links that are ensured by native promoters, such an experiment could provide quantitative two-dimensional expression-fitness landscapes for many pairs of genes, which should be an extremely important component of the genotype-fitness relationship, and which for now we are almost

completely blind to (**Chapter 4;** see (Martin, 2016) for higher level 2-D trait-fitness landscapes in a multicellular organism).

With constantly improving sequencing, genetic engineering and –omics technologies, and the application of experimental creativity, our mechanistic understanding of the genotype-phenotype relationship across different scales can only continue to grow.

6 References

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867-1882.e21.

Agashe, D., Sane, M., Phalnikar, K., Diwan, G.D., Habibullah, A., Martinez-Gomez, N.C., Sahasrabuddhe, V., Polachek, W., Wang, J., Chubiz, L.M., et al. (2016). Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. Molecular Biology and Evolution *33*, 1542–1553.

Alexander, R.M. (2003). Modelling approaches in biomechanics. Philosophical Transactions of the Royal Society B: Biological Sciences *358*, 1429–1435.

Anderson-Lee, J., Fisker, E., Kosaraju, V., Wu, M., Kong, J., Lee, J., Lee, M., Zada, M., Treuille, A., and Das, R. (2016). Principles for Predicting RNA Secondary Structure Design Difficulty. Journal of Molecular Biology *428*, 748–757.

Annaluru, N., Muller, H., Mitchell, L.A., and Ramalingam, S. (2014). Total Synthesis of a Functional Designer Eukaryotic Chromosome. Science *344*, 55–58.

Araya, C.L., Fowler, D.M., Chen, W., Muniez, I., Kelly, J.W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proceedings of the National Academy of Sciences *109*, 16858–16863.

Arias, C.F., Catalán, P., Manrubia, S., and Cuesta, J.A. (2015). toyLIFE: a computational framework to study the multi-level organisation of the genotype-phenotype map. Scientific Reports *4*.

Avery, L., and Wasserman, S. (1992). Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends in Genetics *8*, 312–316.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. Journal of Experimental Medicine *79*, 137–158.

Ay, A., and Arnosti, D.N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. Critical Reviews in Biochemistry and Molecular Biology *46*, 137–151.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology *2*.

Babu, M., Arnold, R., Bundalovic-Torma, C., Gagarinova, A., Wong, K.S., Kumar, A., Stewart, G., Samanfar, B., Aoki, H., Wagih, O., et al. (2014). Quantitative Genome-Wide Genetic Interaction Screens Reveal Global Epistatic Relationships of Protein Complexes in Escherichia coli. PLoS Genetics *10*, e1004120.

Badano, J.L., and Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. Nature Reviews Genetics *3*, 779–789.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. Science *324*, 1720–1723.

Balaban, N.Q. (2004). Bacterial Persistence as a Phenotypic Switch. Science *305*, 1622–1625.

Bank, C., Hietpas, R.T., Wong, A., Bolon, D.N., and Jensen, J.D. (2014). A Bayesian MCMC Approach to Assess the Complete Distribution of Fitness Effects of New Mutations: Uncovering the Potential for Adaptive Walks in Challenging Environments. Genetics *196*, 841–852.

Bank, C., Hietpas, R.T., Jensen, J.D., and Bolon, D.N.A. (2015). A Systematic Survey of an Intragenic Epistatic Landscape. Molecular Biology and Evolution *32*, 229–238.

Bank, C., Matuszewski, S., Hietpas, R.T., and Jensen, J.D. (2016). On the (un)predictability of a large intragenic fitness landscape. Proceedings of the National Academy of Sciences *113*, 14085-14090.

Barbieri, E.M., Muir, P., Akhuetie-Oni, B.O., Yellman, C.M., and Isaacs, F.J. (2017). Precise Editing at DNA Replication Forks Enables Multiplex Genome Engineering in Eukaryotes. Cell *171*, 1453-1467.e13.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.-Y., Ou, J., San Luis, B.-J., Bandyopadhyay, S., et al. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nature Methods *7*, 1017–1024.

Basu, S., and Plewczynski, D. (2010). AMS 3.0: prediction of post-translational modifications. BMC Bioinformatics *11*, 1–15.

Bataillon, T., and Bailey, S.F. (2014). Effects of new mutations on fitness: insights from models and data: Effects of new mutations on fitness. Annals of the New York Academy of Sciences *1320*, 76–92.

Battle, A., Jonikas, M.C., Walter, P., Weissman, J.S., and Koller, D. (2010). Automated identification of pathways from quantitative genetic interaction data. Molecular Systems Biology *6*.

Beadle, G.W., and Tatum, E.L. (1941). Genetic control of biochemical reactions in Neurospora. Proceedings of the National Academy of Sciences *27*, 499–506.

Beckwith, J.R. (1967). Regulation of the Lac Operon. Recent studies on the regulation of lactose metabolism in Escherichia coli support the operon model. Science *156*, 597–604.

Beltrao, P., Cagney, G., and Krogan, N.J. (2010). Quantitative genetic interactions reveal biological modularity. Cell *141*, 739–745.

Bendixsen, D.P., Østman, B., and Hayden, E.J. (2017). Negative Epistasis in Experimental RNA Fitness Landscapes. Journal of Molecular Evolution *85*, 159–168.

Berger, D., and Postma, E. (2014). Biased Estimates of Diminishing-Returns Epistasis? Empirical Evidence Revisited. Genetics *198*, 1417–1420.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proceedings of the National Academy of Sciences *74*, 3171–3175.

Bernet, G.P., and Elena, S.F. (2015). Distribution of mutational fitness effects and of epistasis in the 5' untranslated region of a plant RNA virus. BMC Evolutionary Biology *15*.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D.S. (2006). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature *444*, 929–932.

Bershtein, S., Mu, W., Serohijos, A.W.R., Zhou, J., and Shakhnovich, E.I. (2013). Protein Quality Control Acts on Folding Intermediates to Shape the Effects of Mutations on Organismal Fitness. Molecular Cell *49*, 133–144.

Blanquart, F., Achaz, G., Bataillon, T., and Tenaillon, O. (2014). Properties of selected mutations and genotypic landscapes under Fisher's Geometric Model. Evolution *68*, 3537–3554.

Bonhoeffer, S. (2004). Evidence for Positive Epistasis in HIV-1. Science *306*, 1547–1550.

Bosdriesz, E., Molenaar, D., Teusink, B., and Bruggeman, F.J. (2015). How fast-growing bacteria robustly tune their ribosome concentration to approximate growth-rate maximization. FEBS Journal *282*, 2029–2044.

Boutros, M., Kiger, A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S., Heidelberg Fly Array Consortium, Paro, R., and Perrimon, N. (2004). Genome-Wide RNAi Analysis of Growth and Viability in Drosophila Cells. Science *303*, 832–835.

Boveri, T. (1902). Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. Verh. Der Phys.-Med. Ges. Zu Würzburg *35*, 67–90.

Boyle, E.A., Andreasson, J.O.L., Chircus, L.M., Sternberg, S.H., Wu, M.J., Guegler, C.K., Doudna, J.A., and Greenleaf, W.J. (2017). High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. Proceedings of the National Academy of Sciences *114*, 5461–5466.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535–538.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature *190*, 576–581.

Breslow, D.K., Cameron, D.M., Collins, S.R., Schuldiner, M., Stewart-Ornstein, J., Newman, H.W., Braun, S., Madhani, H.D., Krogan, N.J., and Weissman, J.S. (2008). A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nature Methods *5*, 711–718.

Bressloff, P.C. (2017). Stochastic switching in biology: from genotype to phenotype. Journal of Physics A: Mathematical and Theoretical *50*, 133001.

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature Biotechnology *36*, 272–281.

Buenrostro, J.D., Araya, C.L., Chircus, L.M., Layton, C.J., Chang, H.Y., Snyder, M.P., and Greenleaf, W.J. (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nature Biotechnology *32*, 562–568.

Burch, C.L. (2004). Epistasis and Its Relationship to Canalization in the RNA Virus 6. Genetics *167*, 559–567.

Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature *433*, 531–537.

Carothers, E.E. (1913). The mendelian ratio in relation to certain orthopteran chromosomes. Journal of Morphology *24*, 487–511.

Carrasco, P., de la Iglesia, F., and Elena, S.F. (2007). Distribution of Fitness and Virulence Effects Caused by Single-Nucleotide Substitutions in Tobacco Etch Virus. Journal of Virology *81*, 12979–12984.

Caudle, S.B., Miller, C.R., and Rokyta, D.R. (2014). Environment Determines Epistatic Patterns for a ssDNA Virus. Genetics *196*, 267–279.

Celaj, A., Schlecht, U., Smith, J.D., Xu, W., Suresh, S., Miranda, M., Aparicio, A.M., Proctor, M., Davis, R.W., Roth, F.P., et al. (2017). Quantitative analysis of protein interaction network dynamics in yeast. Molecular Systems Biology *13*, 934.

Cello, J., Paul, A.V., and Wimmer, E. (2002). Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template. *297*, 4.

Chan, Y.H., Venev, S.V., Zeldovich, K.B., and Matthews, C.R. (2017). Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. Nature Communications *8*, 14614.

Chance, B., Estabrook, R.W., and Ghosh, A. (1964). Damped sinusoidal oscillations of cytoplasmic reduced pyridine nucleotide in yeast cells. Proceedings of the National Academy of Sciences *51*, 1244–1251.

Chandrasekaran, S., and Price, N.D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences *107*, 17845–17850.

Chevin, L.-M., Martin, G., and Lenormand, T. (2010). Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution. Evolution *64*, 3213–3231.

Chou, H.-H., Chiu, H.-C., Delaney, N.F., Segre, D., and Marx, C.J. (2011). Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. Science *332*, 1190–1192.

Chou, H.-H., Delaney, N.F., Draghi, J.A., and Marx, C.J. (2014). Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. PLoS Genetics *10*, e1004149.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of Adenovirus 2 messenger RNA. Cell *12*, 1–8.

Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M., et al. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature *446*, 806–810.

Cornish-Bowden, A. (2013). The origins of enzyme kinetics. FEBS Letters *587*, 2725–2730.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The Genetic Landscape of a Cell. Science *327*, 425–431.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420–aaf1420.

Crick, F.H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology *12*, 138–163.

Crow, E.W., and Crow, J.F. (2002). 100 Years Ago: Walter Sutton and the Chromosome Theory of Heredity. Genetics *160*, 1–4.

Cuevas, J.M., Domingo-Calap, P., and Sanjuán, R. (2012). The Fitness Effects of Synonymous Mutations in DNA and RNA Viruses. Molecular Biology and Evolution *29*, 17–20.

Dandage, R., Pandey, R., Jayaraj, G., Rai, M., Berger, D., and Chakraborty, K. (2018). Differential strengths of molecular determinants guide environment specific mutational fates. PLOS Genetics *14*, e1007419.

Dandekar, T., Fieselmann, A., Majeed, S., and Ahmed, Z. (2014). Software applications toward quantitative metabolic flux analysis and modeling. Briefings in Bioinformatics *15*, 91–107.

Davidson, L.A., and Baum, B. (2012). Making waves: the rise and fall and rise of quantitative developmental biology. Development *139*, 3065–3069.

Davierwala, A.P., Haynes, J., Li, Z., Brost, R.L., Robinson, M.D., Yu, L., Mnaimneh, S., Ding, H., Zhu, H., Chen, Y., et al. (2005). The synthetic genetic interaction spectrum of essential genes. Nature Genetics *37*, 1147–1152.

Dean, A.M. (1995). A Molecular Investigation of Genotype by Environment Interactions. Genetics *139*, 19–33.

Decourty, L., Saveanu, C., Zemam, K., Hantraye, F., Frachon, E., Rousselle, J.-C., Fromont-Racine, M., and Jacquier, A. (2008). Linking functionally related genes by sensitive and quantitative characterization of genetic interaction profiles. Proceedings of the National Academy of Sciences *105*, 5821–5826.

Dekel, E., and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. Nature *436*, 588–592.

Denby, C.M., Im, J.H., Yu, R.C., Pesce, C.G., and Brem, R.B. (2012). Negative feedback confers mutational robustness in yeast transcription factor regulation. Proceedings of the National Academy of Sciences *109*, 3874–3878.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nature Reviews Genetics *6*, 678–687.

Dill, K.A., Ozkan, S.B., Weikl, T.R., Chodera, J.D., and Voelz, V.A. (2007). The protein folding problem: when will it be solved? Current Opinion in Structural Biology *17*, 342–346.

Dipple, K.M., and McCabe, E.R.B. (2000). Modifier Genes Convert “Simple” Mendelian Disorders to Complex Traits. Molecular Genetics and Metabolism *71*, 43–50.

Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. ELife *7*, e32472.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853-1866.e17.

Domingo, J., Diss, G., and Lehner, B. (2018). Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature *558*, 117–121.

Domingo-Calap, P., Cuevas, J.M., and Sanjuán, R. (2009). The Fitness Effects of Random Mutations in Single-Stranded DNA and RNA Bacteriophages. PLoS Genetics *5*, e1000742.

Douglas, A.C., Smith, A.M., Sharifpoor, S., Yan, Z., Durbic, T., Heisler, L.E., Lee, A.Y., Ryan, O., Göttert, H., Surendra, A., et al. (2012). Functional Analysis With a Barcoder Yeast Gene Overexpression System. Genes, Genomes, Genetics *2*, 1279–1289.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell *134*, 341–352.

Dykhuizen, D.E., Dean, A.M., and Hart, D.L. (1987). Metabolic flux and fitness. Genetics *115*, 25–31.

Echave, J., and Wilke, C.O. (2017). Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. Annual Review of Biophysics *46*, 85–103.

Englesberg, E., Anderson, R.L., Weinberg, R., Lee, N., Hoffee, P., Huttenhauer, G., and Boyer, H. (1962). L-arabinose-sensitive, L-ribulose 5-phosphate 4-epimerase-deficient mutants of Escherichia coli. Journal of Bacteriology *84*, 137–146.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nature Reviews Genetics *8*, 610–618.

Faisal, A., and Stephens, G. (2009). Quantitative Models of Natural Behaviour - Workshop 11 at CNS 2009 (Berlin, Germany).

Famili, I., Forster, J., Nielsen, J., and Palsson, B.O. (2003). Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proceedings of the National Academy of Sciences *100*, 13134–13139.

Fang, X., Sastry, A., Mih, N., Kim, D., Tan, J., Yurkovich, J.T., Lloyd, C.J., Gao, Y., Yang, L., and Palsson, B.O. (2017). Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. Proceedings of the National Academy of Sciences *114*, 10286–10291.

Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Current Opinion in Microbiology *13*, 344–349.

Firnberg, E., Labonte, J.W., Gray, J.J., and Ostermeier, M. (2014). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. Molecular Biology and Evolution *31*, 1581–1592.

Fisher, R.A. (1930). The genetical theory of natural selection (Oxford: Oxford University Press).

Flynn, K.M., Cooper, T.F., Moore, F.B.-G., and Cooper, V.S. (2013). The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. PLoS Genetics *9*, e1003426.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods *11*, 801–807.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nature Methods *7*, 741–746.

Fragata, I., Matuszewski, S., Schmitz, M.A., Bataillon, T., Jensen, J.D., and Bank, C. (2018). The fitness landscape of the codon space across environments. Heredity.

Fu, C., Deng, S., Jin, G., Wang, X., and Yu, Z.-G. (2017). Bayesian network model for identification of pathways by integrating protein interaction with genetic interaction data. BMC Systems Biology *11*.

Fuchs, F., Pau, G., Kranz, D., Sklyar, O., Budjan, C., Steinbrink, S., Horn, T., Pedal, A., Huber, W., and Boutros, M. (2010). Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Molecular Systems Biology *6*.

Gabdoulline, R.R., Kummer, U., Olsen, L.F., and Wade, R.C. (2003). Concerted Simulations Reveal How Peroxidase Compound III Formation Results in Cellular Oscillations. Biophysical Journal *85*, 1421–1428.

Gabdoulline, R.R., Stein, M., and Wade, R.C. (2007). qPIPSA: Relating enzymatic kinetic parameters and interaction fields. BMC Bioinformatics *8*, 373.

Gagarinova, A., Stewart, G., Samanfar, B., Phanse, S., White, C.A., Aoki, H., Deineko, V., Beloglazova, N., Yakunin, A.F., Golshani, A., et al. (2016). Systematic Genetic Screens Reveal the Dynamic Global Functional Organization of the Bacterial Translation Machinery. Cell Reports *17*, 904–916.

Galton, F. (1876). A theory of heredity. The Journal of the Anthropological Institute of Great Britain and Ireland *5*, 329–348.

Garcia-Viloca, M. (2004). How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. Science *303*, 186–195.

Gavaghan, D., Garny, A., Maini, P.K., and Kohl, P. (2006). Mathematical models in physiology. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences *364*, 1099–1106.

Geertz, M., Shore, D., and Maerkl, S.J. (2012). Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. Proceedings of the National Academy of Sciences *109*, 16540–16545.

Gehring, N.H., Wahle, E., and Fischer, U. (2017). Deciphering the mRNP Code: RNA-Bound Determinants of Post-Transcriptional Gene Regulation. Trends in Biochemical Sciences *42*, 369–382.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balázsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., et al. (2003). Experimental Determination and System Level Analysis of Essential Genes in *Escherichia coli* MG1655. Journal of Bacteriology *185*, 5673–5684.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., et al. (2010). Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. Science *329*, 52–56.

Gingold, H., and Pilpel, Y. (2014). Determinants of translation efficiency and accuracy. Molecular Systems Biology *7*, 481–481.

Giordano, N., Mairet, F., Gouzé, J.-L., Geiselmann, J., and de Jong, H. (2016). Dynamical Allocation of Cellular Resources as an Optimal Control Problem: Novel Insights into Microbial Growth Strategies. PLOS Computational Biology *12*, e1004802.

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O., and Venter, J.C. (2006). Essential genes of a minimal bacterium. Proceedings of the National Academy of Sciences *103*, 425–430.

Gleeson, P., Cantarelli, M., Currie, M., Hokanson, J., Idili, G., Khayrulin, S., Palyanov, A., Szigeti, B., and Larson, S. (2015). The OpenWorm Project: currently available resources and future plans. BMC Neuroscience *16*, P141.

Goldberg, S., Nevins, J., and Darnell, J.E. (1978). Evidence from UV Transcription Mapping that Late Adenovirus Type 2 mRNA Is Derived from a Large Precursor Molecule. J. VIROL. *25*, 5.

Gonçalves, E., Bucher, J., Ryll, A., Niklas, J., Mauch, K., Klamt, S., Rocha, M., and Saez-Rodriguez, J. (2013). Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. Molecular BioSystems *9*, 1576.

Gros, F., Hiatt, H., Gilbert, W., Kurland, C.G., Risebrough, R.W., and Watson, J.D. (1961). Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. Nature *190*, 581–585.

Guenther, U.-P., Yandek, L.E., Niland, C.N., Campbell, F.E., Anderson, D., Anderson, V.E., Harris, M.E., and Jankowsky, E. (2013). Hidden specificity in an apparently nonspecific RNA-binding protein. Nature *502*, 385–388.

Haimovich, A.D., Muir, P., and Isaacs, F.J. (2015). Genomes by design. Nature Reviews Genetics *16*, 501–516.

Haliburton, J.R., Shao, W., Deutschbauer, A., Arkin, A., and Abate, A.R. (2017). Genetic interaction mapping with microfluidic-based single cell sequencing. PLOS ONE *12*, e0171302.

Hansen, T.F. (2013). Why epistasis is important for selection and adaptation: Perspective. Evolution *67*, 3501–3511.

Hartmann, A., and Schreiber, F. (2015). Integrative Analysis of Metabolic Models - from Structure to Dynamics. Frontiers in Bioengineering and Biotechnology *2*.

Hayden, E.J., and Wagner, A. (2012). Environmental change exposes beneficial epistatic interactions in a catalytic RNA. Proceedings of the Royal Society B: Biological Sciences *279*, 3418–3425.

Hayden, E.J., Ferrada, E., and Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature *474*, 92–95.

He, X., Qian, W., Wang, Z., Li, Y., and Zhang, J. (2010). Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nature Genetics *42*, 272–276.

Heinrich, R., and Rapoport, T.A. (1974). A Linear Steady-State Treatment of Enzymatic Chains. General Properties, Control and Effector Strength. European Journal of Biochemistry *42*, 89–95.

Henri, V. (1902). Théorie générale de l'action de quelques diastases. Comptes Rendus de l'Académie Des Sciences *135*, 916–919.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. The Journal of General Physiology *36*, 39–56.

Hietpas, R.T., Jensen, J.D., and Bolon, D.N.A. (2011). Experimental illumination of a fitness landscape. Proceedings of the National Academy of Sciences *108*, 7896–7901.

Hietpas, R.T., Bank, C., Jensen, J.D., and Bolon, D.N.A. (2013). Shifting fitness landscapes in response to altered environments. Evolution *67*, 3512–3522.

Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology *117*, 500–544.

Hornung, G., Oren, M., and Barkai, N. (2012). Nucleosome Organization Affects the Sensitivity of Gene Expression to Promoter Mutations. Molecular Cell *46*, 362–368.

Hunter, P. (2016). The Virtual Physiological Human: The Physiome Project Aims to Develop Reproducible, Multiscale Models for Clinical Practice. IEEE Pulse *7*, 36–42.

Hutchison III, C.A., Peterson, S., Gill, S., Cline, R., White, O., Fraser, C., Smith, H., and J. Craig, V. (1999). Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. Science *286*, 2165–2169.

Hyduke, D.R., and Palsson, B.Ø. (2010). Towards genome-scale signalling-network reconstructions. Nature Reviews Genetics *11*, 297–307.

Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature *420*, 186–189.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences *98*, 4569–4574.

Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., et al. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. Proceedings of the National Academy of Sciences *110*, 13067–13072.

Jaffe, M., Sherlock, G., and Levy, S.F. (2017). iSeq: A New Double-Barcode Method for Detecting Dynamic Genetic Interactions in Yeast. G3: Genes|Genomes|Genetics *7*, 143–153.

Jagdishchandra Joshi, C., and Prasad, A. (2014). Epistatic interactions among metabolic genes depend upon environmental conditions. Mol. BioSyst. *10*, 2578–2589.

Jantzen, B., and Danks, D. (2008). Biological Codes and Topological Causation. Philosophy of Science *75*, 259–277.

Jiang, L., Mishra, P., Hietpas, R.T., Zeldovich, K.B., and Bolon, D.N.A. (2013). Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. PLoS Genetics *9*, e1003600.

Jiang, L., Liu, P., Bank, C., Renzette, N., Prachanronarong, K., Yilmaz, L.S., Caffrey, D.R., Zeldovich, K.B., Schiffer, C.A., Kowalik, T.F., et al. (2016). A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance. Journal of Molecular Biology *428*, 538–553.

Johannsen, W. (1911). The Genotype Conception of Heredity. The American Naturalist *45*, 129–159.

de Jong, H., Casagranda, S., Giordano, N., Cinquemani, E., Ropers, D., Geiselmann, J., and Gouzé, J.-L. (2017). Mathematical modelling of microbes: metabolism, gene expression and growth. Journal of The Royal Society Interface *14*, 20170502.

Joshi, A., and Palsson, B.O. (1989). Metabolic dynamics in the human red blood cell. Part I - A comprehensive kinetic model. Journal of Theoretical Biology *141*, 515–528.

Judson, H.F. (1979). The eighth day of creation: Makers of the revolution in biology (New York, NY, USA: Simon & Schuster, Inc.).

Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype–phenotype landscape for the alternative splicing of a human exon. Nature Communications *7*, 11558.

Kacser, H., and Burns, J.A. (1973). The control of flux. Symposia of the Society for Experimental Biology *27*, 65–104.

Kacser, H., and Burns, J.A. (1981). The Molecular Basis Of Dominance. Genetics *97*, 639–666.

Kacser, H., Burns, J.A., Kacser, H., and Fell, D.A. (1995). The control of flux: 21 Years on. Biochemical Society Transactions *23*, 341–366.

Kærn, M., Elston, T.C., Blake, W.J., and Collins, J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics *6*, 451–464.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kamisetty, H., Ghosh, B., Langmead, C.J., and Bailey-Kellogg, C. (2014). Learning Sequence Determinants of Protein: Protein Interaction Specificity with Sparse Graphical Models. In Research in Computational Molecular Biology, R. Sharan, ed. (Cham: Springer International Publishing), pp. 129–143.

Kaplan, N., Moore, I., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2010). Nucleosome sequence preferences influence in vivo nucleosome organization. Nature Structural & Molecular Biology *17*, 918–920.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell *150*, 389–401.

Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., Weinberger, A., Alon, U., Milo, R., and Segal, E. (2016). Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. Cell *166*, 1282-1294.e18.

Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., and Cooper, T.F. (2011). Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. Science *332*, 1193–1196.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nature Biotechnology *28*, 617–623.

Kim, I., Miller, C.R., Young, D.L., and Fields, S. (2013). High-throughput Analysis of *in vivo* Protein Stability. Molecular & Cellular Proteomics *12*, 3370–3378.

Kimura, M., and Maruyama, T. (1966). The mutational load with epistatic gene interactions in fitness. Genetics *54*, 1337–1351.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences *107*, 9158–9163.

Klesmith, J.R., Bacik, J.-P., Michalczyk, R., and Whitehead, T.A. (2015). Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. ACS Synth. Biol. *4*, 1235–1243.

Kobori, S., and Yokobayashi, Y. (2016). High-Throughput Mutational Analysis of a Twister Ribozyme. Angewandte Chemie International Edition *55*, 10354–10357.

Koch, A.L. (1983). The protein burden of lac operon products. Journal of Molecular Evolution *19*, 455–462.

Kondrashov, A.S. (1988). Deleterious mutations and the evolution of sexual reproduction. Nature *336*, 435.

Kondrashov, F.A., and Kondrashov, A.S. (2001). Multidimensional epistasis and the disadvantage of sex. Proceedings of the National Academy of Sciences *98*, 12089–12092.

Krauss, S., and Brand, M.D. (2000). Quantitation of signal transduction. The FASEB Journal *14*, 2581–2588.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Kryazhimskiy, S., Rice, D.P., Jerison, E.R., and Desai, M.M. (2014). Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science *344*, 1519–1522.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science *324*, 255–258.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science *302*, 1364–1368.

Kunkel, T.A., Bebenek, K., and McClary, J. (1991). Efficient site-directed mutagenesis using uracil-containing DNA. Methods in Enzymology *204*, 125–139.

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Mattiazzi Usaj, M., van Leeuwen, J., et al. (2018). Systematic analysis of complex genetic interactions. Science *360*, eaao1729.

Kvitek, D.J., and Sherlock, G. (2011). Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. PLoS Genetics *7*, e1002056.

Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proceedings of the National Academy of Sciences *109*, 19498–19503.

Lagator, M., Igler, C., Moreno, A.B., Guet, C.C., and Bollback, J.P. (2016). Epistatic Interactions in the Arabinose *Cis* -Regulatory Element. Molecular Biology and Evolution *33*, 761–769.

Lagator, M., Sarikas, S., Acar, H., Bollback, J.P., and Guet, C.C. (2017a). Regulatory network structure determines patterns of intermolecular epistasis. ELife *6*.

Lagator, M., Paixão, T., Barton, N.H., Bollback, J.P., and Guet, C.C. (2017b). On the mechanistic nature of epistasis in a canonical cis-regulatory element. ELife *6*.

Lalić, J., and Elena, S.F. (2012). Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. Heredity *109*, 71–77.

Lalić, J., Cuevas, J.M., and Elena, S.F. (2011). Effect of Host Species on the Distribution of Mutational Fitness Effects for an RNA Virus. PLoS Genetics *7*, e1002378.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology *8*, 995–1005.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. Trends in Genetics *27*, 323–331.

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. Nature Reviews Genetics *14*, 168–178.

Levin, A.M., and Weiss, G.A. (2006). Optimizing the affinity and specificity of proteins with molecular display. Mol. BioSyst. *2*, 49–57.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping Back to Leap Forward: Transcription Enters a New Era. Cell *157*, 13–25.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nature Reviews Genetics *15*, 453–468.

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A.C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. Genome Research *25*, 1018–1029.

Li, C., and Zhang, J. (2018). Multi-environment fitness landscapes of a tRNA gene. Nature Ecology & Evolution *2*, 1025–1032.

Li, C., Qian, W., Maclean, C.J., and Zhang, J. (2016). The fitness landscape of a tRNA gene. Science *352*, 837–840.

Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics *16*, 321–332.

Lothrop, A.P., Torres, M.P., and Fuchs, S.M. (2013). Deciphering post-translational modification codes. FEBS Letters *587*, 1247–1257.

Lunzer, M. (2005). The Biochemical Architecture of an Ancient Adaptive Landscape. Science *310*, 499–501.

Machado, D., Herrgård, M.J., and Rocha, I. (2016). Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction. PLOS Computational Biology *12*, e1005140.

MacLean, R.C., Perron, G.G., and Gardner, A. (2010). Diminishing Returns From Beneficial Mutations and Pervasive Epistasis Shape the Fitness Landscape for Rifampicin Resistance in Pseudomonas aeruginosa. Genetics *186*, 1345–1354.

Maerkl, S.J., and Quake, S.R. (2009). Experimental determination of the evolvability of a transcription factor. Proceedings of the National Academy of Sciences *106*, 18650–18655.

Maisnier-Patin, S., Roth, J.R., Fredriksson, Å., Nyström, T., Berg, O.G., and Andersson, D.I. (2005). Genomic buffering mitigates the effects of deleterious mutations in bacteria. Nature Genetics *37*, 1376–1379.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marciano, D.C., Lua, R.C., Katsonis, P., Amin, S.R., Herman, C., and Lichtarge, O. (2014). Negative Feedback in Genetic Circuits Confers Evolutionary Resilience and Capacitance. Cell Reports *7*, 1789–1795.

Marciano, D.C., Lua, R.C., Herman, C., and Lichtarge, O. (2016). Cooperativity of Negative Autoregulation Confers Increased Mutational Robustness. Physical Review Letters *116*.

Martin, C.H. (2016). Context dependence in complex adaptive landscapes: frequency and trait-dependent selection surfaces within an adaptive radiation of Caribbean pupfishes. Evolution *70*, 1265–1282.

Martin, G. (2014). Fisher's Geometrical Model Emerges as a Property of Complex Integrated Phenotypic Networks. Genetics *197*, 237–255.

Martin, G., and Lenormand, T. (2006). A General Multivariate Extension of Fisher's Geometrical Model and the Distribution of Mutation Fitness Effects across Species. Evolution *60*, 893–907.

Martin, G., Elena, S.F., and Lenormand, T. (2007). Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nature Genetics *39*, 555–560.

Mavor, D., Barlow, K., Thompson, S., Barad, B.A., Bonny, A.R., Cario, C.L., Gaskins, G., Liu, Z., Deming, L., Axen, S.D., et al. (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. ELife *5*, e15802.

McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA *19*, 1537–1551.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T.S. (2014). Comprehensive mutational scanning of a kinase *in vivo* reveals substrate-dependent fitness landscapes. Nucleic Acids Research *42*, e112–e112.

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. Verhandlungen Des Naturforschenden Vereines, Abhandlungern, Brünn *4*, 3–47.

Meselson, M., and Stahl, F.W. (1958). The replication of DNA in Escherichia coli. Proceedings of the National Academy of Sciences *44*, 671–682.

Michaelis, L., and Menten, M.L. (1913). Kinetik der Invertinwirkung. Biochemische Zeitschrift *49*, 333–369.

Milne, C.B., Kim, P.-J., Eddy, J.A., and Price, N.D. (2009). Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. Biotechnology Journal *4*, 1653–1670.

Mitchell, P. (1961). Coupling of Phosphorylation to Electron and Hydrogen Transfer by a Chemi-Osmotic type of Mechanism. Nature *191*, 144–148.

Monzon, A.M., Zea, D.J., Fornasari, M.S., Saldaño, T.E., Fernandez-Alberti, S., Tosatto, S.C.E., and Parisi, G. (2017). Conformational diversity analysis reveals three functional mechanisms in proteins. PLOS Computational Biology *13*, e1005398.

Morgan, T.H., Sturtevant, A.H., and Muller, H.J. (1915). The Mechanism of Mendelian Heredity (New York: Henry Holt and Co.).

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biology *17*.

Mustonen, V., Kinney, J., Callan, C.G., and Lassig, M. (2008). Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences *105*, 12376–12381.

Nagel, A.C., Joyce, P., Wichman, H.A., and Miller, C.R. (2012). Stickbreaking: A Novel Fitness Landscape Model That Harbors Epistasis and Is Consistent with Commonly Observed Patterns of Adaptive Evolution. Genetics *190*, 655–667.

Nevins, J.R., and Darnell, J.E. (1978). Groups of Adenovirus Type 2 mRNA's Derived from a Large Primary Transcript: Probable Nuclear Origin and Possible Common 3' Ends. J. VIROL. *25*, 13.

Nghe, P., Kogenaru, M., and Tans, S.J. (2018). Sign epistasis caused by hierarchy within signalling cascades. Nature Communications *9*.

Nichols, R.J., Sen, S., Choo, Y.J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K.M., Lee, K.J., Wong, A., et al. (2011). Phenotypic Landscape of a Bacterial Cell. Cell *144*, 143–156.

Niederberger, P., Prasad, R., Miozzari, G., and Kacser, H. (1992). A strategy for increasing an *in vivo* flux by genetic manipulations. The tryptophan system of yeast. Biochemical Journal *287*, 473–479.

Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., et al. (1966). The RNA Code and Protein Synthesis. Cold Spring Harbor Symposia on Quantitative Biology *31*, 11–24.

Noble, D. (1960). Cardiac Action and Pacemaker Potentials based on the Hodgkin-Huxley Equations. Nature *188*, 495–497.

Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., and Wang, C.L. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Molecular Systems Biology *10*, 748–748.

Novak, B., and Tyson, J.J. (1995). Quantitative analysis of a molecular model of mitotic control in fission yeast. Journal of Theoretical Biology *173*, 283–305.

Novichkov, P.S., Rodionov, D.A., Stavrovskaya, E.D., Novichkova, E.S., Kazakov, A.E., Gelfand, M.S., Arkin, A.P., Mironov, A.A., and Dubchak, I. (2010). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. Nucleic Acids Research *38*, W299–W307.

Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nature Biotechnology *29*, 659–664.

O'Brien, E.J., Monk, J.M., and Palsson, B.O. (2015). Using Genome-scale Models to Predict Biological Capabilities. Cell *161*, 971–987.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature *246*, 96–98.

Olson, C.A., Wu, N.C., and Sun, R. (2014). A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. Current Biology *24*, 2643–2651.

Onge, R.P.S., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R.W., Nislow, C., Roth, F.P., and Giaever, G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nature Genetics *39*, 199–206.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nature Methods *6*, 767–772.

van Opijnen, T., Lazinski, D.W., and Camilli, A. (2014). Genome-Wide Fitness and Genetic Interactions Determined by Tn-seq, a High-Throughput Massively Parallel Sequencing Method for Microorganisms: Tn-seq: High-Throughput Sequencing for Microorganisms. In Current Protocols in Molecular Biology, F.M. Ausubel, R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, and K. Struhl, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 7.16.1-7.16.24.

Pál, G., Kouadio, J.-L.K., Artis, D.R., Kossiakoff, A.A., and Sidhu, S.S. (2006). Comprehensive and Quantitative Mapping of Energy Landscapes for Protein-Protein Interactions by Rapid Combinatorial Scanning. Journal of Biological Chemistry *281*, 22378–22385.

Patel, L., Abate, C., and Curran, T. (1990). Altered protein conformation on DNA binding by Fos and Jun. Nature *347*, 572–575.

Pattee, H.H. (2001). The physics of symbols: bridging the epistemic cut. Biosystems *60*, 5–21.

Pauling, L., Itano, H.A., Singer, S.J., and Wells, I.C. (1949). Sickle Cell Anemia, a Molecular Disease. Science *110*, 543–548.

Perfeito, L., Ghozzi, S., Berg, J., Schnetz, K., and Lässig, M. (2011). Nonlinear Fitness Landscape of a Molecular Pathway. PLoS Genetics *7*, e1002160.

Peris, J.B., Davis, P., Cuevas, J.M., Nebot, M.R., and Sanjuan, R. (2010). Distribution of Fitness Effects Caused by Single-Nucleotide Substitutions in Bacteriophage f1. Genetics *185*, 603–609.

Pertzev, A.V. (2006). Characterization of RNA sequence determinants and antideterminants of processing reactivity for a minimal substrate of Escherichia coli ribonuclease III. Nucleic Acids Research *34*, 3708–3721.

Phillips, P.C. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics *9*, 855–867.

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proceedings of the National Academy of Sciences *72*, 784–788.

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., and Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. Science *352*, 840–844.

R. Shenoy, S., and Jayaram, B. (2010). Proteins: Sequence to Structure and Function – Current Status. Current Protein & Peptide Science *11*, 498–514.

Rajkumar, A.S., Dénervaud, N., and Maerkl, S.J. (2013). Mapping the fine structure of a eukaryotic promoter input-output function. Nature Genetics *45*, 1207–1215.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.-O., Hayles, J., et al. (2008). Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map in Fission Yeast. Science *322*, 405–410.

Rokyta, D.R., Joyce, P., Caudle, S.B., Miller, C., Beisel, C.J., and Wichman, H.A. (2011). Epistasis between Beneficial Mutations and the Phenotype-to-Fitness Map for a ssDNA Virus. PLoS Genetics *7*, e1002075.

Roscoe, B.P., Thayer, K.M., Zeldovich, K.B., Fushman, D., and Bolon, D.N.A. (2013). Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. Journal of Molecular Biology *425*, 1363–1377.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. Cell *163*, 698–711.

Roy, K.R., Smith, J.D., Vonesch, S.C., Lin, G., Tu, C.S., Lederer, A.R., Chu, A., Suresh, S., Nguyen, M., Horecka, J., et al. (2018). Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nature Biotechnology *36*, 512–520.

Ryll, A., Bucher, J., Bonin, A., Bongard, S., Gonçalves, E., Saez-Rodriguez, J., Niklas, J., and Klamt, S. (2014). A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models. Biosystems *124*, 26–38.

Sadowski, M.I., and Jones, D.T. (2009). The sequence–structure relationship and protein function prediction. Current Opinion in Structural Biology *19*, 357–362.

Salazar-Ciudad, I., and Jernvall, J. (2010). A computational model of teeth and the developmental origins of morphological variation. Nature *464*, 583.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nature Biotechnology *27*, 946–950.

Sameith, K., Amini, S., Groot Koerkamp, M.J.A., van Leenen, D., Brok, M., Brabers, N., Lijnzaad, P., van Hooff, S.R., Benschop, J.J., Lenstra, T.L., et al. (2015). A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions. BMC Biology *13*.

Sanjuan, R., Moya, A., and Elena, S.F. (2004a). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences *101*, 8396–8401.

Sanjuan, R., Moya, A., and Elena, S.F. (2004b). The contribution of epistasis to the architecture of fitness in an RNA virus. Proceedings of the National Academy of Sciences *101*, 15376–15379.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. Nature *533*, 397–401.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis: Genes required for mycobacterial growth. Molecular Microbiology *48*, 77–84.

Sauer, U. (2004). High-throughput phenomics: experimental methods for mapping fluxomes. Current Opinion in Biotechnology *15*, 58–63.

Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218–223.

Schenk, M.F., Szendro, I.G., Salverda, M.L.M., Krug, J., and de Visser, J.A.G.M. (2013). Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene. Molecular Biology and Evolution *30*, 1779–1787.

Schoustra, S., Hwang, S., Krug, J., and de Visser, J.A.G.M. (2016). Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. Proceedings of the Royal Society B: Biological Sciences *283*, 20161376.

Schrodinger, E. (1944). What is life? The Physical aspect of the living cell (Cambridge, UK: Cambridge University Press).

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., et al. (2005). Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile. Cell *123*, 507–519.

Scott, J., and Smith, G. (1990). Searching for peptide ligands with an epitope library. Science *249*, 386–390.

Scott, M., Klumpp, S., Mateescu, E.M., and Hwa, T. (2014). Emergence of robust growth laws from optimal regulation of ribosome synthesis. Molecular Systems Biology *10*, 747–747.

Scriver, C.R., and Waters, P.J. (1999). Monogenic traits are not simple: lessons from phenylketonuria. Trends in Genetics *15*, 267–272.

Segrè, D., DeLuna, A., Church, G.M., and Kishony, R. (2005). Modular epistasis in yeast metabolism. Nature Genetics *37*, 77–83.

Serohijos, A.W.R., and Shakhnovich, E.I. (2014). Contribution of Selection for Protein Folding Stability in Shaping the Patterns of Polymorphisms in Coding Regions. Molecular Biology and Evolution *31*, 165–176.

Serohijos, A.W.R., Rimas, Z., and Shakhnovich, E.I. (2012). Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. Cell Reports *2*, 249–256.

Shah, P., McCandlish, D.M., and Plotkin, J.B. (2015). Contingency and entrenchment in protein evolution under purifying selection. Proceedings of the National Academy of Sciences *112*, E3226–E3235.

Shalem, O., Sharon, E., Lubliner, S., Regev, I., Lotan-Pompan, M., Yakhini, Z., and Segal, E. (2015). Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control. PLOS Genetics *11*, e1005147.

Shine, J., and Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. Nature *254*, 34–38.

Shultzaberger, R.K., Malashock, D.S., Kirsch, J.F., and Eisen, M.B. (2010). The Fitness Landscapes of cis-Acting Binding Sites in Different Promoter and Environmental Contexts. PLoS Genetics *6*, e1001042.

Shultzaberger, R.K., Maerkl, S.J., Kirsch, J.F., and Eisen, M.B. (2012). Probing the Informational and Regulatory Plasticity of a Transcription Factor DNA–Binding Domain. PLoS Genetics *8*, e1002614.

Sikosek, T., and Chan, H.S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. Journal of The Royal Society Interface *11*, 20140419–20140419.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences *39*, 381–399.

Smith, G. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315–1317.

Smith, B., Williams, J., and Schulze, S. (2003). The Ontology of the Gene Ontology. AMIA Annual Symposium Proceedings 609–613.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. Molecular Systems Biology *11*, 848–848.

Soyfer, V.N. (2001). The consequences of political dictatorship for Russian science. Nature Reviews Genetics *2*, 723–729.

Starita, L.M., Pruneda, J.N., Lo, R.S., Fowler, D.M., Kim, H.J., Hiatt, J.B., Shendure, J., Brzovic, P.S., Fields, S., and Klevit, R.E. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proceedings of the National Academy of Sciences *110*, E1263–E1272.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell *122*, 957–968.

Stevens, N.M. (1905). A study of the germ cells of Aphis rosæ and Aphis œnotheræ. Journal of Experimental Zoology *2*, 313–333.

Stockwell, G.R., and Thornton, J.M. (2006). Conformational Diversity of Ligands Bound to Proteins. Journal of Molecular Biology *356*, 928–944.

Stoebel, D.M., Dean, A.M., and Dykhuizen, D.E. (2008). The Cost of Expression of Escherichia coli lac Operon Proteins Is in the Process, Not in the Products. Genetics *178*, 1653–1660.

Sutton, W.S. (1902). On the morphology of the chromosome group in Brachystola magna. Biological Bulletin *4*, 24–39.

Szafraniec, K., Wloch, D.M., Sliwa, P., Borts, R.H., and Korona, R. (2003). Small fitness effects and weak genetic interactions between deleterious mutations in heterozygous loci of the yeast Saccharomyces cerevisiae. Genetical Research *82*, 19–31.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. Nature Genetics *43*, 656–662.

Szathmary, E. (1993). Do deleterious mutations act synergistically? Metabolic Control Theory provides a partial answer. Genetics *133*, 127–132.

Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249–254.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Molina, M.M.S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in Vivo Map of the Yeast Protein Interactome. Science *320*, 1465–1470.

Tenaillon, O. (2014). The Utility of Fisher's Geometric Model in Evolutionary Genetics. Annual Review of Ecology, Evolution, and Systematics *45*, 179–201.

Thompson, K.S., Vinson, C.R., and Freire, E. (1993). Thermodynamic characterization of the structural stability of the coiled-coil region of the bZIP transcription factor GCN4. Biochemistry *32*, 5491–5496.

Tischler, J., Lehner, B., Chen, N., and Fraser, A.G. (2006). Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. Genome Biology 13.

Todeschini, A.-L., Georges, A., and Veitia, R.A. (2014). Transcription factors: specific DNA binding and specific gene regulation. Trends in Genetics *30*, 211–219.

Tong, A.H.Y. (2004). Global Mapping of the Yeast Genetic Interaction Network. Science *303*, 808–813.

Towbin, B.D., Korem, Y., Bren, A., Doron, S., Sorek, R., and Alon, U. (2017). Optimality and sub-optimality in a bacterial growth law. Nature Communications *8*, 14123.

Trindade, S., Sousa, A., and Gordo, I. (2012). Antibiotic resistance and stress in the light of Fisher's model. Evolution *66*, 3815–3824.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576.e16.

Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2008). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Briefings in Bioinformatics *10*, 217–232.

Turing, A. (1952). The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society B *237*, 37–72.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein–protein interactions inSaccharomyces cerevisiae. Nature *403*, 623–627.

Ursell, T., Lee, T.K., Shiomi, D., Shi, H., Tropini, C., Monds, R.D., Colavin, A., Billings, G., Bhaya-Grossman, I., Broxton, M., et al. (2017). Rapid, precise quantification of bacterial cellular dimensions across a genomic-scale knockout library. BMC Biology *15*.

Vasieva, O., Rasolonjanahary, M., and Vasiev, B. (2013). Mathematical modelling in developmental biology. Reproduction *145*, R175–R184.

Venkataram, S., Dunn, B., Li, Y., Agarwala, A., Chang, J., Ebel, E.R., Geiler-Samerotte, K., Hérissant, L., Blundell, J.R., Levy, S.F., et al. (2016). Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. Cell *166*, 1585-1596.e22.

Ventura, S. (2005). Sequence determinants of protein aggregation: tools to increase protein solubility. Microbial Cell Factories 8.

Vik (2011). Genotype-phenotype map characteristics of an in silico heart cell. Frontiers in Physiology.

Vilar, J.M.G. (2010). Accurate Prediction of Gene Expression by Integration of DNA Sequence Statistics with Detailed Modeling of Transcription Regulation. Biophysical Journal *99*, 2408–2413.

de Visser, J.A.G.M., Cooper, T.F., and Elena, S.F. (2011). The causes of epistasis. Proceedings of the Royal Society B: Biological Sciences *278*, 3617–3624.

Viswanatha, R., Li, Z., Hu, Y., and Perrimon, N. (2018). Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in Drosophila cells.

de Vos, M.G.J., Poelwijk, F.J., Battich, N., Ndika, J.D.T., and Tans, S.J. (2013). Environmental Dependence of Genetic Constraint. PLoS Genetics *9*, e1003580.

de Vos, M.G.J., Dawid, A., Sunderlikova, V., and Tans, S.J. (2015). Breaking evolutionary constraint with a tradeoff ratchet. Proceedings of the National Academy of Sciences *112*, 14906–14911.

Wachter, E., Quante, T., Merusi, C., Arczewska, A., Stewart, F., Webb, S., and Bird, A. (2014). Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. ELife *3*.

Wang, T., Guan, C., Guo, J., Liu, B., Wu, Y., Xie, Z., Zhang, C., and Xing, X.-H. (2018). Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. Nature Communications *9*.

Warren, C.L., Kratochvil, N.C.S., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., and Ansari, A.Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. Proceedings of the National Academy of Sciences *103*, 867–872.

Watson, J.D., and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weinreich, D.M. (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science *312*, 111–114.

Weismann, A. (1892). Das Keimplasma. Eine Theorie der Vererbung. (Jena: Fischer).

Weiße, A.Y., Oyarzún, D.A., Danos, V., and Swain, P.S. (2015). Mechanistic links between cellular trade-offs, gene expression, and growth. Proceedings of the National Academy of Sciences *112*, E1038–E1047.

Westerhoff, H.V., and Chen, Y.-D. (1984). How do enzyme activities control metabolite concentrations?. An additional theorem in the theory of metabolic control. European Journal of Biochemistry *142*, 425–430.

Westerhoff, H.V., and Palsson, B.O. (2004). The evolution of molecular biology into systems biology. Nature Biotechnology *22*, 1249–1252.

Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A., et al. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nature Biotechnology *30*, 543–548.

Wilke, C.O., and Christoph, A. (2001). Interaction between directional epistasis and average mutational effects. Proceedings of the Royal Society B: Biological Sciences *268*, 1469–1474.

Wilke, C.O., Lenski, R.E., and Adami, C. (2003). Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. BMC Evolutionary Biology 14.

Wilson, E.B. (1902). Mendel's principles of heredity and the maturation of the germ-cells. Science *16*, 991–993.

Wilson, E.B. (1905). Studies on chromosomes. I. The behavior of the idiochromosomes in hemiptera. Journal of Experimental Zoology *2*, 371–405.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics *13*, 59–69.

Wloch, D.M., Szafraniec, K., Borts, R.H., and Korona, R. (2001). Direct Estimate of the Mutation Rate and the Distribution of Fitness Effects in the Yeast Saccharomyces cerevisiae. Genetics *159*, 441–452.

Wolpert, L. (1969). Positional Information and the Spatial Pattern of Cellular Differentiation. Journal of Theoretical Biology *25*, 1–47.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics *8*, 206–216.

Wrenbeck, E.E., Azouz, L.R., and Whitehead, T.A. (2017). Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nature Communications *8*, 15695.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of Molecular Biology *293*, 321–331.

Wylie, C.S., and Shakhnovich, E.I. (2011). A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proceedings of the National Academy of Sciences *108*, 9916–9921.

Xu, L., Barker, B., and Gu, Z. (2012). Dynamic epistasis for different alleles of the same gene. Proceedings of the National Academy of Sciences *109*, 10420–10425.

Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., et al. (2009). Update on the Keio collection of Escherichia coli single-gene deletion mutants. Molecular Systems Biology *5*.

Yates, R.A., and Pardee, A.B. (1957). Control by uracil of formation of enzymes required for orotate synthesis. Journal of Biological Chemistry *227*, 677–692.

Yau, R., and Rape, M. (2016). The increasing complexity of the ubiquitin code. Nature Cell Biology *18*, 579–586.

Ye, P., Peyser, B.D., Pan, X., Boeke, J.D., Spencer, F.A., and Bader, J.S. (2005). Gene function prediction from congruent synthetic lethal interactions in yeast. Molecular Systems Biology *1*, E1–E12.

You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C.W., Wang, Y.-P., Lenz, P., Yan, D., and Hwa, T. (2013). Coordination of bacterial proteome with metabolism by cyclic AMP signalling. Nature *500*, 301.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science *322*, 104–110.

Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., et al. (2017). Challenges and emerging directions in single-cell analysis. Genome Biology *18*.

Zeitoun, R.I., Garst, A.D., Degen, G.D., Pines, G., Mansell, T.J., Glebes, T.Y., Boyle, N.R., and Gill, R.T. (2015). Multiplexed tracking of combinatorial genomic mutations in engineered cell populations. Nature Biotechnology *33*, 631–637.

Zeitoun, R.I., Pines, G., Grau, W.C., and Gill, R.T. (2017). Quantitative Tracking of Combinatorially Engineered Populations with Multiplexed Binary Assemblies. ACS Synthetic Biology *6*, 619–627.

Zhang, R., and Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Research *37*, D455–D458.

Zwart, M.P., Schenk, M.F., Hwang, S., Koopmanschap, B., de Lange, N., van de Pol, L., Nga, T.T.T., Szendro, I.G., Krug, J., and de Visser, J.A.G.M. (2018). Unraveling the causes of adaptive benefits of synonymous mutations in TEM-1 β-lactamase. Heredity.

(2000). Beyond the Average: The Evolutionary Importance of Gene Interactions and Variability of Epistatic Effects. In Epistasis and the Evolutionary Process, J.B. Wolf, E.D. Brodie III, and M.J. Wade, eds. (NY, USA: Oxford University Press), pp. 20-38.