Evolution of transcriptional regulation in *Escherichia coli*

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Luise Wolf
aus Deutschland

Basel, 2014

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Erik van Nimwegen (Fakultätsverantwortlicher)

Prof. Dr. Dirk Bumann (Korreferent)

Basel, den 10.12.2013

Prof. Dr. Jörg Schibler
(Dekan)

Contents

Abstract **1**

1 Introduction **3**

1.1 The importance of gene regulation in bacteria 3
1.2 Mechanisms of gene regulation in bacteria 4
1.3 Gene expression is a stochastic process 5
1.4 Evolution of gene regulation 7
1.5 Outline of the thesis 8

2 Limited regulatory accuracy implies selection for noisy gene expression **9**

2.1 Abstract 10
2.2 Introduction 10
2.3 Main text 10
2.4 Discussion 16
2.5 Materials and Methods 17
2.6 Supplementary Text 29

3 σ^{70} binding is a prerequisite for expression but not predictive for transcript levels **63**

3.1 Introduction 63
3.2 Main part 65
  3.2.1 Characterization of the initial library 65
  3.2.2 Similar sequences exhibit similar expression 67
  3.2.3 Evolution of transcription factor binding sites 69
  3.2.4 Evolutionary dynamics of sequence features during promoter sequence evolution 72
3.3 Discussion 79
3.4 Methods 82
  3.4.1 Library preparation of promoter sequences for deep sequencing 82
  3.4.2 Pre-processing of Sequencing reads 83
  3.4.3 Genotyping individual sequences by Sanger Sequencing 83
  3.4.4 Phenotyping individual sequences by flow cytometry 83

3.4.5 Unification and clustering of sequences based on similarity 83
3.4.6 Prediction of binding sites for transcription factors in *E.coli* 84
3.4.7 Prediction of σ^{70} free binding energies 84
3.5 Acknowledgements 85
3.6 Supplementary information 85
3.6.1 Supplementary figures 85
3.6.2 Set of barcodes for multiplexed deep sequencing 86

4 The predictability of molecular evolution during functional innovation 89
4.1 Abstract 91
4.2 Significance 91
4.3 Introduction 91
4.4 Results and Discussion 93
4.5 Materials and Methods 103
4.5.1 Gene deletion strains 103
4.5.2 Experimental evolution 103
4.5.3 Sequence analysis 103
4.5.4 Assaying the effects of intergenic mutations on protein expression . . 103
4.5.5 Plasmid rescue of deletion genotypes 103
4.5.6 Analysis of network distances 104
4.5.7 Statistics 104
4.6 Acknowledgements 104
4.7 SI Appendix 104
4.7.1 SI Appendix, Materials and Methods 104
4.7.2 SI Discussion 110

5 Summary and future perspectives 133

References 135

Acknowledgments 145

Curriculum Vitae 147

Abstract

During gene expression, transcription initiation marks the first step towards synthesis of functional proteins. Expression levels of specific types of RNA molecules in the cell depend on the underlying genotype of the promoter sequence. Prediction of expression levels from the promoter sequence alone can have important implications for the design of artificial promoters. In this work, we explored promoter determinants that cause differences in expression levels and tracked how a certain level can be reached by a directed evolution experiment in *E.coli* . Promoter sequences were evolved from a million random sequences with selection on expression level and high mutation rate. Mapping of expression phenotypes to the underlying promoter genotypes revealed what sequence features determine the rate of transcription. If no differential expression is required, incorporation of σ^{70} binding sites allows expression. However, predicted affinity of σ^{70} to bind to a promoter sequence in different promoter contexts is not explanatory in terms of expression levels, suggesting that other sequence features determine the rate of transcription. Furthermore, separation of functional promoter sequences to non-regulatory sequences is promoted by high AT content as well as preference of generally longer promoter sequences. Recovery of an essential missing gene function can also be obtained by overexpression of other genes present in the genome by changing the strength of σ^{70} binding to the promoter sequence. Small changes in the expression level were shown to have a severe impact on the fitness of the organism. The amount of deviation away from the optimal expression level in clonal promoter populations has been shown to depend on the promoter's genotype. We are presenting an evolutionary model to explain under which regulatory settings selection favors high variance in expression levels between cells.

Chapter 1

Introduction

1.1 The importance of gene regulation in bacteria

Gene regulation is a fundamental and essential process that is present in all organisms and allows cells to alter their endogenous RNA and protein concentrations. Variations in the concentrations can be achieved by changes in the transcription and translation rates. Individual promoter activities span a wide range and gene regulation allows promoters to be active only in specific conditions, e.g. such that a metabolic enzyme will only be expressed in the presence of a specific carbon source like lactose (Jacob & Monod, 1961). The ability to control the expression of RNA and protein molecules that are only needed in certain conditions saves energy.

Gene expression is as a dynamic process consisting of multiple steps, with various layers for control. One of the keystones in molecular biology was the postulation of its central dogma by Crick in 1958 (Crick, 1958, Crick, 1970), describing the flow of genetic information in a cell. Before a protein gets expressed, it passes several control check points and modification steps. The amount of protein in the cell can be quantitatively described by a thermodynamic model, incorporating transcription and translation rates as well as RNA and protein degradation rates (Swain et al., 2002). The main determinant of translational efficiency and thus translation rate is the ribosomal binding site (RBS) upstream of the translational start site (Lee et al., 2013). ‘AGGAGG’ is the consensus sequence found in *E.coli* (Vimberg et al., 2007) and is known to influence ribosomal binding positively. Also, the spacing between the RBS and translational start play a role in the translation initiation rate (Vellanoweth & Rabinowitz, 1992). Secondary structure and folding of the RNA, especially around the RBS, can lower translation rate (Salis et al., 2009), and this leads to avoidance of nucleotides around the translation start site complementary to the RBS (Molina & van Nimwegen, 2008). Codon usage can influence translation elongation, thereby affecting expression level (Roymondal et al., 2009, Welch et al., 2009) and the favorite start codon ‘ATG’ used in *E.coli* creates a positive effect on translation efficiency (Vellanoweth & Rabinowitz, 1992). Rare codons are preferred at the N-terminus for highly expressed genes, most probably because of reduced RNA secondary structures in this area (Goodman et al., 2013).

Besides regulation of transcription and translation rate, RNA and protein amounts are also determined by RNA and protein half lives and protein degradation (Maurizi, 1992). Synthesized proteins are generally stable (Goldberg & John, 1976) and have much longer half lives than the generation time in *E.coli* . With generation times in favorable growth conditions of only around 20 minutes in *E.coli* (Wang et al., 2010), most proteins will be diluted by the growth rate and not degradation of the proteins themselves. In contrast to proteins, RNAs in *E.coli* have very short half-lives with a median of only 3.7 minutes (Bernstein et al., 2004).

The ability to control gene expression offers the possibility to respond dynamically to changing environments. Bacteria are often faced with changing environments over time, and for genes that are only needed in specific conditions, differential gene expression is favorable. The ability to change the (expression) phenotype over time in response to the environment is called ‘phenotypic plasticity’ (Price et al., 2003). In bacteria, this is helped by the organization of genes into operons (Jacob & Monod, 1961) allowing co-regulation of genes. In order for bacteria to change their expression profile under external stimuli, signals have to be integrated by, for instance, a sensor kinase (Krell et al., 2010, Stock et al., 2000) and transmitted by a response regulator to change the transcriptional program. Single cell organisms like bacteria are especially susceptible to changes in their surrounding environment (Boor, 2006) and are able to react upon small changes by changing their transcriptional profile.

1.2 Mechanisms of gene regulation in bacteria

The first step in protein expression from DNA is the initiation of transcription upstream of the translational start site. Being the first link in the expression chain, gene expression levels in bacteria are mainly determined by the rate of transcription initiation (Lloyd et al., 2001). Given that the number of molecules involved in the transcriptional regulation of genes is small, molecules involved in transcriptional regulation have to be shared and correctly distributed. Promoter sequences particularly compete for binding of the RNA polymerase (RNAP) and do so by attracting it with binding sites for the sigma subunits of the RNAP holoenzyme (Maeda et al., 2000). Attractiveness of the promoter sequence for RNAP binding changes across conditions as specific transcription factors promoting or preventing RNAP binding alter their activity profile over environments (Rolfe et al., 2012).

Transcription factors change into the active state by modifications like phosphorylation (Re et al., 2002) or oligomerization (Myers et al., 2013). Most factors bind their target genes with their effector molecules bound (Balderas-Martínez et al., 2013), which enables fast switching to their active mode instead of producing transcription factors upon stimulation. Regulatory proteins can facilitate initiation of transcription (activators) or lower the rate of transcription (repressors). Many factors affect in both modes of control as dual regulators, depending on the target gene they are acting on (Balderas-Martínez et al., 2013, Salgado et al., 2004).

Epigenetic modifications of DNA can also alter the interaction between DNA and the binding proteins, which is mainly achieved by methylation in bacteria (Casadesús & Low, 2006).

1.3 Gene expression is a stochastic process

The expression phenotype observed is determined by the promoter genotype, the environment the cell is faced with, and the internal state the cell finds itself in. Additionally, there is a noise component introducing variation in the phenotypes observed (Raj & van Oudenaarden, 2008). Gene expression is a stochastic process (Elowitz et al., 2002, Raser & O'Shea, 2005, Cai et al., 2006, McAdams & Arkin, 1997, Kaern et al., 2005) due to the randomness associated with individual reactions during gene expression, limiting its precision. This implies that a population of genetically identical cells show inter-cell variation in the number of gene products observed in a particular environment (Elowitz et al., 2002). The reaction kinetics of the processes of transcription and translation should be described using a stochastic rather than a deterministic model (Munsky et al., 2012) to account for the uncertainty involved. As molecules involved in the gene expression process are small in number, stochastic effects can play a crucial role. For instance, each promoter sequence has a certain probability to be transcribed in a given condition that depends, for example, on the concentration of RNAP molecules in the cell. RNAP molecules diffuse in the three-dimensional cell space and every now and then bind to promoter regions and initiate transcription. Under the assumption that the variation in average protein number $\langle p \rangle$ follows a Poisson distribution, variance σ_p^2 equals $\langle p \rangle$ (Arriaga, 2009, Thattai & van Oudenaarden, 2001). Following Poisson behavior, the squared coefficient of variation $\frac{\sigma_p^2}{\langle p \rangle^2}$ (CV^2) scales with $\frac{1}{\langle p \rangle}$. Deviations from this behaviour (Bar-Even et al., 2006, Swain et al., 2002) reveal that the stochasticity of biochemical reactions during expression are not the only sources for variation observed, specifically for high expression (Taniguchi et al., 2010). The total noise measured $\mathrm{CV}^2_{\mathrm{tot}}$ is composed of the sum of an intrinsic component $\mathrm{CV}^2_{\mathrm{int}}$ and an extrinsic component $\mathrm{CV}^2_{\mathrm{ext}}$ (Swain et al., 2002). Extrinsic noise arises from the heterogeneity between cells in the number and activity of cellular components involved in gene expression (Swain et al., 2002, Raser & O'Shea, 2004, Elowitz et al., 2002, Raser & O'Shea, 2005, Kaern et al., 2005).

Expression of each gene exhibits a certain level of total noise in a given environment (Silander et al., 2012) and is reproducibly measurable, showing that the noise associated with each gene is a property of the sequence underlying its regulation (Newman et al., 2006, Bar-Even et al., 2006, Silander et al., 2012, Raser & O'Shea, 2004, Blake et al., 2003, Golding et al., 2005, Carey et al., 2013). A strong relationship between mean protein levels $\langle p \rangle$ and their variations are observed across different taxa, including deviations from the general trend as well (Newman et al., 2006, Bar-Even et al., 2006, Silander et al., 2012, Taniguchi et al., 2010, Carey et al., 2013).

Genes expressing at a similar level but showing substantially different levels of noise raise the question as to why they differ.

One common explanation for the noise differences observed is that genes important for growth have experienced selection for lowering their noise levels. Growth rate can depend on expression level of genes, as overexpression of genes can be a costly enterprise (Wagner, 2005, Shachrai et al., 2010). On the other hand, underexpression of genes important for growth can also lead to a reduction in growth rate. Growth rate and expression level are thus directly intertwined (Babu & Aravind, 2006, Dekel & Alon, 2005, Fong et al., 2005, Rowley et al., 1992). This makes expression level a direct target of selection: gene expression level is a phenotypic trait that is selected upon (Rifkin et al., 2003, Lemos et al., 2005, Gilad et al., 2006). That is why different genes exhibit a wide range of expression levels when measured using promoter-fluorescence reporter gene fusions (Silander et al., 2012). If gene expression level is subject to natural selection, any variation away from the optimum level decreases the fitness of an individual cell. On the population level, fitness decreases with the amount of variance observed away from the optimum. Selection has indeed been shown to minimize noise levels (Lehner, 2008, Fraser et al., 2004, Wang & Zhang, 2011). Genes essential for the growth of an organism in defined conditions have been shown to exhibit less variation in their expression levels (Silander et al., 2012, Wang & Zhang, 2011, Dong et al., 2011, Newman et al., 2006, Li et al., 2010). Also, noise was considerably lower for genes that were highly conserved across taxa, as well as for genes belonging to certain functional categories like building block biosynthesis (Silander et al., 2012) such as synthesis of amino acids or nucleotides and genes known to be dosage-sensitive (Lehner, 2008). However, fitness of individuals is not affected by variation in the expression of all genes. The promoter architectures of these genes may evolve without taking into consideration their noise levels.

At the same time, there is also evidence that selection acted to increase the noise levels of some genes. Functional categories associated with these genes are, for instance, stress response and energy metabolism (Silander et al., 2012). In environments fluctuating over time, noise might be considered an adaptive trait (Kaern et al., 2005, Zhang et al., 2009, Kussell & Leibler, 2005) as sensing mechanisms are more cost intensive to maintain and have a longer response time. The stochastic switching of the phenotype is a phenomenon known as 'bet-hedging' (Beaumont et al., 2009, Haccou & Iwasa, 1995, Thattai & van Oudenaarden, 2004). Populations that have an selective advantage by division of labor can also show elevated levels of phenotypic noise. During infection of *Salmonella Typhimurium*, self-destructive cooperation of a subpopulation laid the foundation for a successful infection (Ackermann et al., 2008).

The third possible explanation is that most genes try to lower their noise levels but due to other traits they are selected upon, selection on their noise levels becomes less important. Genes that show a high variability in their expression levels across conditions (with high phenotypic plasticity) tend to have higher levels of noise associated with their expression levels (Lehner, 2010, Bajić & Poyatos, 2012). Incorporation of transcription factor binding sites into the promoter region generate dependencies between the regulating factor and its target gene. This might constitute another unavoidable source of variation introduced in the gene expression level (Woo & Li, 2011, Sanchez et al., 2011).

If variability in expression levels is subject to natural selection or only a side-effect of other evolutionary forces has not been fully evaluated.

1.4 Evolution of gene regulation

The availability of thousands of sequenced bacterial genomes has sped up our understanding of forces in the evolution of gene regulation in bacteria (McAdams et al., 2004). Evolution of regulatory regions besides coding regions was observed early on (King & Wilson, 1975) and has been proven to be adaptive (Blank et al., 2013, Wray et al., 2003, Wray, 2007, Gilad et al., 2006). Genetic changes in the regulatory region of a gene can evoke changes in the regulation that can be selected upon. These changes may allow a gene to react appropriately on changes in organismal development (Wray, 2007, King & Wilson, 1975) or organismal ability to respond to changes in the environment. Innovation of novel regulatory function is important, as innovation of novel gene functionality itself and and evolves from a given DNA sequence.

Gene and regulatory functionality may evolve from random, non-functional sequences (Carvunis et al., 2012, Tautz & Domazet-Losǒ, 2011, Kaessmann, 2010, Cai et al., 2008, Tsai et al., 2012). As new genes in bacteria are acquired via mechanisms like horizontal gene transfer or gene duplications (Ochman et al., 2000, Serres et al., 2009), new regulation may have to evolve *de novo* as well. Moreover, bacteria facing new environments may have to evolve novel functional regulation. If the regulatory sequence of a gene evolves new binding sites for transcription factors present in the genome, then the expression level of the gene becomes a function of the concentration or activity of the transcription factor regulating its expression.

Providing binding sites for transcription factors allows differential expression of the target gene over time. In the best case scenario, transcription levels of the gene that are most beneficial can be tracked over all environments the bacterium finds itself in. However, there are limitations to the precision of tracking the ideal gene expression level. Each promoter can only evolve binding sites for transcription factors encoded in the genome, or evolve a a new factor that is able to track the environmental needs. It has generally been observed, that bacteria living in more complex environments tend to have more sigma factors (Kill et al., 2005). The complexity of the environment is thus shaping the complexity of regulatory interactions observed in an organism (McAdams et al., 2004). This is also reflected in the number of transcription factors encoded in bacterial genomes: with increasing complexity of the genome in terms of gene numbers, the need for transcription factors grows to the power of two and not linearly (van Nimwegen, 2003). Besides evolution of novel binding sites in promoter sequences, novel transcription factors can evolve or get lost from the genome. As only presence of both a factor and its regulated site cause changes in the phenotypic state of the cell, these entities usually co-evolve (Hershberg & Margalit, 2006).

Evolutionary turnover of transcription factors across species gives us an idea about the speed of evolution at the regulatory versus the coding level. Although traces of conservation of transcription factors across genomes are present (Rajewsky et al., 2002), transcription factors generally evolve faster than their target genes (Lozada-Chávez et al., 2006), resulting in less conservation (Babu & Aravind, 2006, Babu et al., 2007).

1.5 Outline of the thesis

The work presented highlights several aspects of the molecular evolution of transcriptional regulation.

Limitations in the precision of transcription rate regulation are being discussed in Chapter 2 (Limited regulatory accuracy implies selection for noisy gene expression). Variation in expression levels between cells connected with a single regulatory region are under selection. The Chapter illustrates that the two concepts of noise-minimization (Lehner, 2008) and noise-favoritism (Kussell & Leibler, 2005) are not mutually exclusive but connected in a continuous space depending on the actual regulatory abilities of the organism. Selection for noise-incorporation in the promoter sequence is presented as a strategy to overcome regulatory incapacities that may be achieved by coupling to noisy transcription factors.

Chapter 3 (σ^{70} binding is a prerequisite for expression but not predictive for transcript levels) presents minimal requirements for a DNA sequence to be regulatory functional. Genotypic traits are being discussed that give rise to the expression phenotypes observed. Prediction of expression levels from a diverse set of promoter sequences alone is a difficult task, that requires more information than only the binding strength of sigma factors.

In Chapter 4 (The predictability of molecular evolution during functional innovation) it becomes evident that mutations in regulatory regions are more frequent than expected from their occurrence in the genome in the recovery of lost functionality. Many missing metabolic functions can be recovered by overexpression of other genes, which is mainly achieved by nucleotide changes in the sigma binding sites.

Chapter 2

Limited regulatory accuracy implies selection for noisy gene expression

Luise Wolf, Olin K. Silander*, and Erik van Nimwegen*
Biozentrum, University of Basel, and Swiss Insitute of Bioinformatics, Basel, Switzerland
*to whom correspondence should be addressed: olin.silander@unibas.ch,
erik.vannimwegen@unibas.ch

Manuscript in preparation

2.1 Abstract

Although it is often tacitly assumed that gene regulatory interactions are finely tuned, how accurate gene regulation could evolve from a state without regulation is unclear. Moreover, gene expression noise would seem to impede the evolution of accurate gene regulation, and previous investigations have provided circumstantial evidence that natural selection has acted to lower noise levels. By evolving synthetic *E.coli* promoters *de novo*, we here show that, contrary to expectations, promoters exhibit low noise by default. Instead, selection must have acted to increase the noise levels of highly regulated *E.coli* promoters. We present a general theory of the interplay between gene expression noise and gene regulation that explains these observations. The theory shows that propagation of expression noise from regulators to their targets is not an unwanted side-effect of regulation, but rather acts as a rudimentary form of regulation that facilitates the evolution of more accurate regulation.

2.2 Introduction

Studies of gene expression noise in several different model organisms have shown that the promoters of some genes exhibit much more transcriptional noise than others (Newman et al., 2006, Silander et al., 2012, Carey et al., 2013). It is unclear, however, how these differences in noise levels have been shaped by natural selection. On the one hand, it can be argued that in each condition there is an optimal expression level for each protein, such that variations away from this optimal level are detrimental to an organism's fitness, implying that selection will act to minimize noise. Indeed, many studies have used circumstantial evidence to suggest that selection generally acts to minimize noise (Newman et al., 2006, Silander et al., 2012, Lehner, 2010, Lehner, 2008, Barkai & Shilo, 2007). In this interpretation, genes with lowest noise have been most strongly selected against noise, whereas high noise genes have experienced much weaker selection against noise. On the other hand, gene expression noise generates phenotypic diversity between organisms with identical genotypes, and there are well-established theoretical models showing that such phenotypic diversity can be selected for in fluctuating environments (Bull, 1987, Kussell & Leibler, 2005). In addition, there is empirical evidence that selection has acted to increase expression noise in some cases (Blake et al., 2006, Bishop et al., 2007, Ackermann et al., 2008, Zhang et al., 2009). It is thus possible that some of the genes with elevated noise may have been selected for phenotypic diversity.

2.3 Main text

In order to assess how natural selection has acted on the transcriptional noise of promoters, it is critical to determine what default noise levels would be exhibited by promoters that have not been selected for their noise properties. To address this, we evolved a large set of synthetic *E.coli* promoters *de novo* in the laboratory using an experimental protocol in

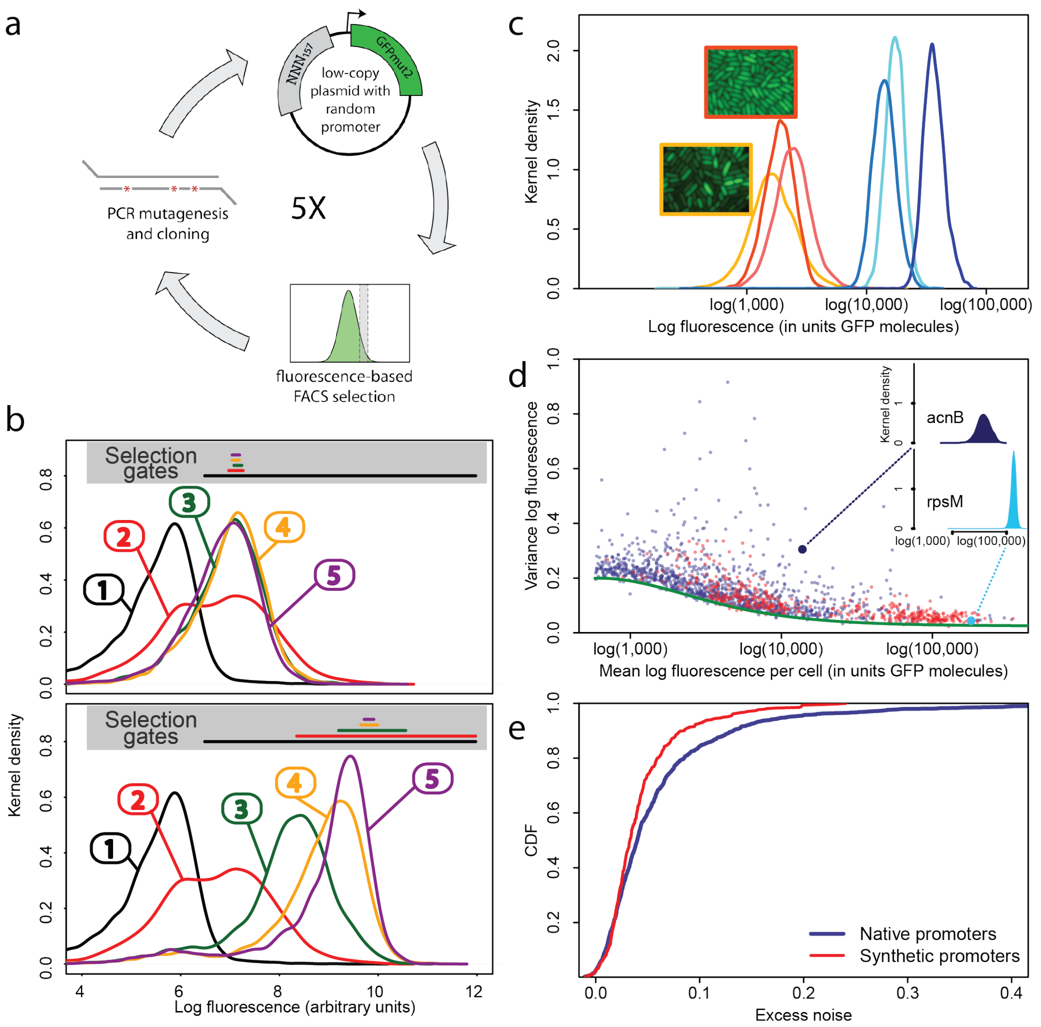

Figure 2.1: **Experimental evolution of functional promoters *de novo*.** **a:** We created an initial library of approximately 10^6 unique synthetic promoters by cloning random nucleotide sequences, of approximately $100 - 150$ base pairs (bp) in length, upstream of a strong ribosomal binding site followed by an open reading frame for GFP as used to quantify the expression of native *E. coli* promoters (Zaslaver et al., 2006), and transformed this library into a population of cells (**Materials and Methods**). We evolved populations of synthetic promoters by performing 5 rounds of selection and mutation on this library. In each round we used fluorescence activated cell sorting (FACS) to select $2 * 10^5$ cells that lie within a gate comprising the 5% of the population closest in fluorescence to a given target level. The plasmids were isolated from the selected cells and PCR mutagenesis was used to introduce new genetic variation into the promoter regions. We then re-cloned the mutated promoters into fresh plasmids, and transformed them into a fresh population of cells. We performed this evolutionary scheme on three replicate populations in which we selected for a target expression level equal to the median expression level (50th percentile) of all native *E. coli* promoters, and three replicate populations in which we selected for a target expression level at the 97.5th percentile of all native promoters (referred to here as medium and high expression levels, respectively). **b:** Changes in the fluorescence distribution for one evolutionary run selecting for medium target expression (top) and one evolutionary run selecting for high target expression (bottom). The curves show the population's expression distributions before selection, with the numbers above each curve indicating the selection round. The colored bars at the top indicate the FACS gates that were used to select cells from the populations at each corresponding round. **c:** Examples of fluorescence distributions for individual clones obtained after five rounds of evolution. Microscopy pictures of two individual clonal promoter populations are shown as insets. **d:** For each native *E. coli* promoter (blue) and synthetic promoter (red) the mean (x-axis) and variance (y-axis) of log-fluorescence intensities across cells were measured using flow cytometry. Fluorescence values are expressed in units of number of GFP molecules. The green curve shows the theoretically predicted minimal variance as a function of mean expression (**Supplementary Text**). The insets show the log-fluorescence distributions for two example promoters (corresponding to the larger dark blue and light blue dots). **e:**, Cumulative distributions of excess noise levels of native (blue) and synthetic (red) promoters.

which promoters were selected on the basis of the mean expression level they conferred, while experiencing virtually no selection on their noise properties (Fig. 2.1 and Supplementary Text). Starting with an initial library of 100-150 base pair long random sequences, we performed five rounds of mutation and selection, resulting in a genetically diverse collection of functional promoters that conferred expression close to a pre-specified target level (Fig. 2.1A-C and Fig. S2.1). We selected a subset of 479 synthetic promoters from the third and fifth rounds, choosing equal numbers of promoters from each of six replicate lineages we evolved (Fig. 2.1; Materials and Methods). We then used flow cytometry, as described previously (Silander et al., 2012), to measure the distribution of fluorescence levels per cell for each synthetic promoter, as well as for all native *E.coli* promoters (Zaslaver et al., 2006).

Observing that the fluorescence distributions across cells were well approximated by log-normal distributions (Fig. 2.1C), we characterized each promoter's distribution by the mean and variance of log-fluorescence, defining the latter as the promoter's noise level (Fig. 2.1D). This definition of noise is equivalent to the square of the coefficient of variation whenever fluctuations are small relative to the mean, which applies to most promoters. Using quantitative Western blotting and qPCR we confirmed that the mean fluorescence levels were directly proportional to GFP molecule numbers and that protein levels were determined primarily by mRNA levels, demonstrating that fluorescence reflected transcriptional activity (Fig. S2.2 and Fig. S2.3 and Supplementary Text). Noise levels were reproducible across biological replicates (Fig. S2.4), and noise levels estimated using microscopy were consistent with those measured by flow cytometry (Fig. S2.5).

As expected (Bar-Even et al., 2006, Newman et al., 2006) we observed a strong relationship between the mean and variance of expression levels of each promoter (Fig. 2.1D). In particular, we observed a strict lower bound on variance as a function of mean expression. This lower bound is well described (Fig. 2.1D green curve) by a simple model that incorporates background fluorescence, and intrinsic and extrinsic noise components (Taniguchi et al., 2010) (Supplementary Text). We defined the *excess noise* of a promoter as its variance above and beyond this lower bound, allowing us to compare the noise levels of promoters with different means (Fig. S2.6). We found, surprisingly, that most of the synthetic promoters exhibited noise levels close to the minimal level exhibited by the native promoters (Fig. 2.1D). Additionally, a substantial fraction of native promoters exhibited excess noise levels significantly greater than the synthetic promoters (Fig. 2.1E and Fig. S2.6 and Fig. S2.7). For example, only 26.1% of the synthetic promoters exhibited excess noise above 0.05, compared to 41.6% of the native *E.coli* promoters ($p < 7.7 * 10^{-10}$, hypergeometric test). Given that the synthetic promoters were evolved from random sequence fragments, and had not been selected on their noise properties (Supplementary Text), we concluded that constitutively expressed *E.coli* promoters exhibit low excess noise levels by default. Importantly, this implies that the native promoters with elevated excess noise must have experienced selective pressures that caused them to increase their noise.

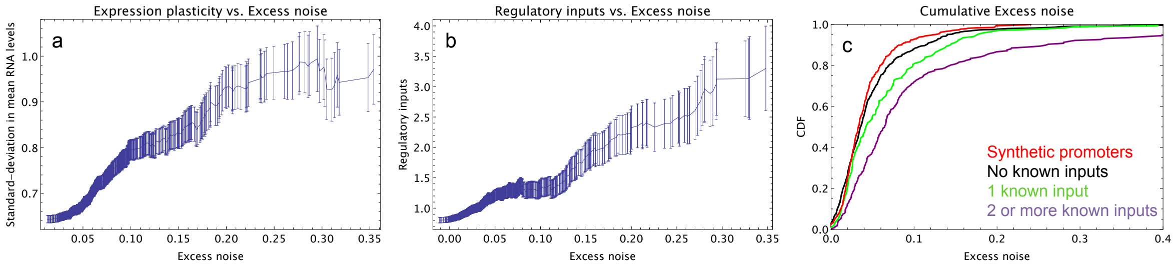

Figure 2.2: **Promoters with elevated noise exhibit high expression plasticity and large numbers of regulatory inputs. a:** Native promoters were sorted by their excess noise x and, as a function of a cut-off on x (horizontal axis), we calculated the mean and standard-error (vertical axis) of the variation in mRNA levels across different experimental conditions (data from http://genexpdb.ou.edu/) of all promoters with excess noise larger than x. **b:** Promoters were sorted by excess noise x as in panel **a**, and mean and standard-error of the number of known regulatory inputs (vertical axis, data from RegulonDB (Salgado et al., 2013)) for promoters with excess noise larger than x is shown. **c:** Cumulative distributions of excess noise levels of synthetic promoters (red) and native promoters without known regulatory inputs (black), with one known regulatory input (green), and with two or more known regulatory inputs (purple).

To understand how selection might have acted to increase noise, we first investigated whether excess noise was associated with other characteristics of the promoters. Previous studies in *S. cerevisiae* have shown that promoters with high noise tend to also show high expression plasticity, i.e. large changes in mean expression level across environments (Newman et al., 2006). Although we did not clearly observe this association in data from our previous study (Silander et al., 2012), a recent re-analysis of this data did uncover a significant association between expression plasticity and noise (Singh, 2013), which we confirmed using our present data (Fig. 2.2A). In addition, we found that there is an equally strong relationship between excess noise and the number of regulators known to target the promoter (Salgado et al., 2013) (Fig. 2.2B). In particular, whereas the excess noise levels of promoters without known regulatory inputs are very similar to those of our synthetic promoters, promoters with one or more regulatory inputs have clearly elevated noise levels (Fig. 2.2C).

We next considered what the origin of this general association between noise and regulation could be. It is important to recognize that, when a promoter couples to a transcription regulator by evolving cognate binding sites, the expression of the associated gene will be affected in two separate ways. First, the gene's mean expression will become correlated with the activity of the regulator in a condition-specific manner. Second, in addition to this 'condition-response' effect, the noise in the expression or activity of the regulator will be propagated to the target gene. This 'noise-propagation' effect will cause an increase in expression noise of the target (Thattai & van Oudenaarden, 2001). Based on this noise-propagation effect, and in analogy with fluctuation-dissipation theorems from physics, it has been proposed that elevated expression noise is simply an unwanted but unavoidable side-effect of regulation (Lehner & Kaneko, 2011).

However, there is no reason to assume that the condition-response and noise-propagation effects must always be in selective conflict with each other. Several theoretical treatments have shown that phenotypic variability may be selectively beneficial when environments change in ways that cannot be accurately sensed or are too rapid for organisms to respond

(Bull, 1987, Haccou & Iwasa, 1995, Kussell & Leibler, 2005). Although such theoretical studies are typically less concerned with the mechanisms by which such increased noise could be genetically encoded, the noise-propagation effect is one obvious candidate mechanism. It would thus seem that, at least in some situations, the condition-response and noise-propagation effects could act in concert. To quantify how selection might act on the combination of these two effects, we developed a general model that considers a gene whose optimal expression levels vary across conditions, and calculated how the condition-response and noise-propagation effects of coupling to a given regulator conspire to affect fitness (Supplementary Text). Although our model applies very generally (Supplementary Text), we illustrate it here using a simple scenario (Fig. 2.3).

The expression of an unregulated promoter is characterized by a distribution with a given mean and variance (Fig. 2.3A, blue curve). We assume that the organism experiences different environments and that, in each environment, cells with expression levels within a certain range are selected. In the simple scenario of Fig. 2.3 there are 3 environments (red, gold, and green), with the green environment requiring up-regulation of the expression and the red environment requiring down-regulation of the expression (Fig. 2.3A). The fitness in each environment corresponds to the fraction of cells with expression levels within the selected range, i.e. the unregulated promoter has reasonably high fitness in the gold environment but very low fitness in the green and red environments. Since the organisms experience all 3 environments, a poor overlap between the expression distribution and the selected range in any one environment leads to low overall fitness.

To improve fitness, a promoter may evolve binding sites for an existing regulator, such that its expression becomes dependent on the activity of this regulator, which will generally vary across environments. Our modeling shows that the resulting fitness depends only on 2 effective parameters of the regulator: The correlation R between the condition-dependent expression levels (or, more generally, activities) of the regulator and the desired levels of the promoter, and the signal-to-noise parameter S that characterizes the accuracy of the regulator's expression.

As intuitively expected, the highest fitness is obtained when coupling to an accurate regulator with high signal-to-noise S, i.e. whose activities correlate precisely with the desired expression levels (cyan dot in Fig. 2.3B and Fig. 2.3C,F). The resulting expression distributions of the promoter accurately track the desired levels, with only moderately increased noise in the promoter's expression (Fig. 2.3F). However, regulators that track the desired expression levels of the promoter with such high accuracy may often not be available. Interestingly, coupling to a noisy regulator whose activity is entirely uncorrelated with the desired expression levels (blue dot in Fig. 2.3B and Fig. 2.3E,H) also substantially increases fitness. In this regime, the increased fitness results exclusively from the noise-propagation mechanism. Surprisingly, coupling to the uncorrelated noisy regulator (blue dot in Fig. 2.3B and Fig. 2.3E,H) outperforms coupling to a moderately correlated regulator (magenta dot in Fig. 2.3B and Fig. 2.3D,G). This is due to the fact that the magenta regulator is not

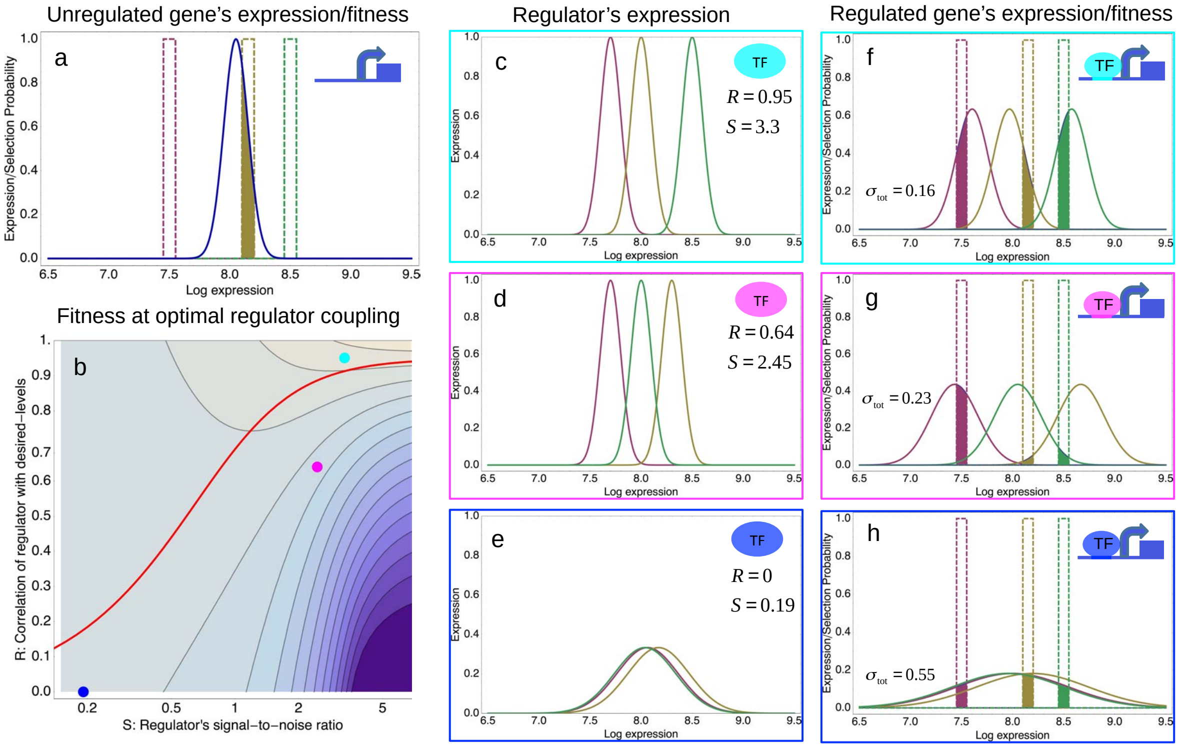

Figure 2.3: **A model of the evolution of gene expression regulation in a variable environment.** **a:** Expression distribution of an unregulated promoter (blue curve) and selected expression ranges in 3 different environments, i.e. the red, gold, and green dashed curves show fitness as a function of expression level in these environments. Although our model applies more generally, for simplicity we here visualize selection as truncation selection (i.e. a rectangular fitness function). The fitness of the promoter in the gold environment is proportional to the shaded area. **b:** Contour plot of the log-fitness of a promoter that is optimally coupled to a transcription factor (TF) with signal-to-noise ratio S and correlation R. Contours run from -0.5 at the top right to -7.5 at the bottom right. The three colored dots correspond to the TFs illustrated in panels **c-h**. The red curve shows optimal S as a function of R. **c-e:** Each panel shows the expression distributions of an example TF across the 3 environments (red, gold, and green curves). The corresponding values of correlation R and signal-to-noise S are indicated in each panel. **f-h:** Each panel shows the expression distributions across the 3 environments for a promoter that is optimally coupled to the TF indicated in the inset. The shaded areas correspond to the fitness in each environment. The total noise levels of the regulated promoters are also indicated in each panel. The unregulated promoter has total noise $\sigma_{\text{tot}} = 0.1$.

noisy enough given its correlation R, i.e. lowering S for this TF would result in an increase of the promoter's noise, and this would lead to an increase in fitness in the green and gold conditions (see Fig. 2.3G). This illustrates that regulators may be under selection to become noisy themselves and the red curve in Fig. 2.3B shows the optimal signal-to-noise S of a regulator as a function of its correlation R. Whereas to the right of this curve the noise-propagation is too large, and too small to the left of it, along the curve the condition-response and noise-propagation effects are optimally acting in concert. This clarifies how accurate gene regulation can evolve smoothly, starting from a noisy regulator with low R and S whose benefits come entirely from the noise-propagation, by increasing both R and S in small steps, until reaching highly accurate regulation with high R and S for which the condition-response effect dominates.

Our model also predicts how the final noise of a promoter depends on the variance in its desired expression levels (Supplementary Text). In particular, assuming the best available regulator in the genome has a given correlation R with the desired levels, there will be a critical variance such that below this variance the final noise will be equal to the noise of the unregulated promoter, and above this critical variance the final noise of the promoter will be proportional to (Fig. S2.8). That is, our model explains the observation that expression noise increases with expression plasticity. Similarly, in our simple model the increase in expression noise is directly due to coupling to regulators, such that our model also explains the observed general association between expression noise and regulatory inputs.

2.4 Discussion

Because genotype-phenotype relationships for complex phenotypic traits are poorly understood, it is often difficult to assess how observable variation in a particular trait has been affected by natural selection. Here we have shown that by comparing naturally observed variation in a particular trait with variation observed in synthetic systems that were evolved under well-controlled selective conditions, definite inferences can be made about the selection pressures that have acted on the natural systems. In particular, by evolving synthetic *E. coli de novo* using a procedure in which promoters are strongly selected on their mean expression and not on their expression noise, we have shown that native promoters must have experienced selective pressures that increased their noise levels. To account for this, we have proposed a theoretical model that provides a simple mechanistic framework for understanding how selection can act to couple transcriptional regulators and target genes, and which quantifies the parameter regimes in which we expect promoters to exhibit high levels of noise. This framework vastly expands the evolutionary conditions under which novel regulatory interactions will be selected for; instead of assuming that the regulators and their targets must evolve in a tightly coordinated fashion, the model shows that genes may often benefit from coupling to regulators whose activities do not correlate with the gene's expression requirements at all. In particular, the condition-response and noise-propagation effects of coupling to a regulator, rather than being in conflict with each other, may often act in concert. Finally, our model shows quite generally that unless regulation is very precise, regulatory interactions that act to increase noise are beneficial. Thus, high levels of expression noise can be expected whenever the accuracy of regulation is limited.

2.5 Materials and Methods

Ab initio promoter library construction from random sequences

We obtained chemically synthesized nucleotide sequences of random nucleotides 200 bp in length (Purimex, Germany). Each sequence had defined 5' and 3' ends to allow PCR amplification. Within these constant regions, restriction sites for BamHI and XhoI were present. The intervening sequence was made up of 157 bp of random nucleotides (5'-CCTTTCGTCTTCACCTCGAG-(N_{157})-GGGATCCTCTGGATGTAAGAAGG-3'). However, as coupling of base pairs during oligonucleotide synthesis is not always successful and strand breaks can frequently occur in long oligonucleotides, many oligonucleotides were shorter than 200 bp in length. We used PCR to generate double stranded DNA from the single stranded oligonucleotides using forward and reverse primers matching the defined 5' and 3' ends. We gel-purified the double-stranded PCR product and double-digested it using BamHI and XhoI. After column-purification, sequences were ligated into a version of the low-copy plasmid pUA66, which contains a *gfpmut2* open reading frame downstream of a strong ribosomal binding site (Zaslaver et al., 2006). The vector was modified to remove a weak σ^{70} binding site present 24 bp upstream of the GFP open reading frame (two point mutations, A→G and T→G, were introduced, changing the putative σ^{70} binding site from TAGATT to TGGATG, with the consensus σ^{70} binding site being TATAAT). The ligation was performed using T4 DNA ligase (NEB) at 16 °C for 24 hours. The ligation product was then column purified and electroporated into *E. coli* DH10B cells. This protocol resulted in extremely high transformation yields (approximately 10^6 individual clones per transformation).

Selection on expression level using flow cytometry

Cultures of transformed cells were regenerated for one hour in 1 mL SOC medium (Super Optimal Broth supplemented with $20mM$ glucose) and afterwards $1mL$ SOC containing $50\mu g/ml$ kanamycin was added for overnight growth, ensuring that only cells containing the plasmid could grow. These cultures were then diluted 500-fold (approximately $5 * 10^6$ cells in total) into M9 minimal media supplemented with 0.2% glucose and grown for 2.5 hours with shaking at 200 rpm. The distribution of GFP fluorescence levels was measured for each culture using fluorescence activated cell sorting (FACS) in a FACSAria IIIu (BD Biosciences), with excitation at 488nm and a 513/17nm bandpass filter used for emission.

We used this distribution of fluorescence values to designate a selection gate. The position of the gate was determined by measuring the mean fluorescence of two reference promoters (Zaslaver et al., 2006): *gyrB* which exhibits a mean expression level that is at the 50th percentile all *E. coli* promoters; and *rpmB*, which exhibits a mean expression level that is at the 97.5th percentile of all *E. coli* promoters (Silander et al., 2012). For each of these reference genes, the mean fluorescence level was measured, and a selection gate was constructed, centered on this mean expression level, such that 5% of all clones in the population fell within the gate. For each round of selection, we sorted $200'000$ cells contained within this gate. Sorted cells were then transferred to $4mL$ Luria Broth (LB) media (containing $50\mu g/ml$ Kanamycin) and grown overnight. These cultures were stored supplemented with 7.5% glycerol at $-80\,°\mathrm{C}$ for subsequent analysis.

For each expression level (i.e. reference gene), we evolved three replicate populations. We refer to these as the medium expressers (those promoters selected based on the *gyrB* reference gate) and high expressers (those promoters selected based on the *rpmB* reference gate).

PCR mutagenesis

Following FACS-based selection on fluorescence, we introduced novel genetic variation into the populations using PCR mutagenesis. We first re-grew the cells overnight and used this culture to prepare plasmid DNA. We amplified the promoter sequences from these plasmids using the GeneMorph II Random Mutagenesis Kit (Stratagene) with the primers referred to previously that matched the defined regions of the promoters. We used 0.01 ng of DNA as starting material and 35 cycles for amplification. This resulted in a mutation rate of around 0.01 per bp (such that we expect that in 200 bp, 95% of the promoters will contain between zero and four mutations). These PCR products were then digested with XhoI and BamHI, ligated back into the vector, and again transformed into DH10B cells. We repeated this entire process (selection, PCR mutagenesis, and transformation) five times in total. At this point, the plasmid libraries of synthetic promoters were isolated and transformed into *E. coli* K12 MG1655 for comparison to a library of native *E. coli* promoters (see below).

Quantification of fluorescence

To quantify fluorescence on a single-cell level, we used flow cytometry with a FACSCanto II (BD Biosciences), with excitation at 488nm and a 513/17nm band-pass filter used for emission. We collected data for at least $50'000$ events. We then gated this data as outlined in (Silander et al., 2012), identifying approximately $5'000$ cells most similar in FSC and SSC. We then calculated the mean and variance in log-fluorescence using these cells, using a Bayesian procedure that accounts for outliers (Supplementary Text). We randomly selected 479 promoters from the evolved set (72 medium expressers and 72 high expressers after 3 rounds of selection; 168 medium expressers and 167 high expressers after 5 rounds of selection) and quantified mean and variance in fluorescence. We used the same measurement procedures to calculate mean and variance for all promoters contained in a library of *E. coli* promoters also placed upstream of the gfpmut2 open reading frame on the pUA66 plasmid (Zaslaver et al., 2006). We refer to the promoters from this library as native *E. coli* promoters. For 288 promoters, we quantified fluorescence in three independent cultures and found that both mean and variance in expression were reproducible across replicate biological experiments (Fig. S2.4). Additionally, we sequenced 378 sequences from our set of 479 promoter sequences, which showed that even after five rounds of selection, the promoters were quite diverse (Fig. S2.1). To confirm the sensitivity and accuracy of the FACS measurements, we selected ten promoters and used fluorescence microscopy to measure their mean and variance in fluorescence. The cells were grown in the same conditions described above, placed on 1% agarose pad, and images were obtained using a CoolSNAP HQ CCD camera (Photometrics) connected to a DeltaVision Core microscope (Applied Precision) with a UPlanSApo 100X/1.40 oil objective (Olympus). Image-processing was done in soft-WoRx v3.3.6 (Applied Precision) and fluorescence values were extracted based on DIC-image mediated cell detection in MicrobeTracker Suite (Sliusarenko et al., 2011). For each cell, we calculated fluorescence per cell volume by summing all pixel values and dividing by the volume of the cell as estimated by MicrobeTracker. Cells undergo substantial phenotypic changes when they are put on agar, including changes in the distribution of cell sizes. Consequently, it is problematic to compare absolute variance measurements directly between FACS and microscope. We therefore compared the relative noise levels of different promoters. The 10 selected native promoters consist of 5 pairs with almost identical mean expression values (as measured by the FACS) but with noise levels that vary by different amounts. For each of the 5 pairs, we calculated the ratio of the noise levels of the higher and the lower noise promoter as measured by both the FACS and the microscope. As shown in Fig. S2.5, with the exception of one pair of promoters that showed almost equal noise levels in the FACS but a 50% difference in noise in the microscope, all other pairs showed good correlation of the relative noise levels in the FACS and in microscope, confirming that relative noise levels are similar in FACS and microscope measurements.

Quantitative Western analysis

To determine the correspondence between fluorescence intensities and absolute GFP numbers per cell, eight individual promoter clones were grown in three biological replicates using the same media conditions as in the experimental evolution. The cells were then re-suspended in SDS sample buffer, heated for 5 minutes at 95 °C, and proteins were resolved by 12% SDS-PAGE. Quantification was done by loading a standard curve consisting of 10, 25, 50, 75, and 100 nanograms of GFP (Clonetech, #632373). Proteins were transferred to a Hybond ECL membrane (GE Healthcare, Life Sciences), which was then blocked in TNT (20 mM Tris pH 7.5, 150 mM NaCl, 0.05% Tween 20) with 1% BSA and 1% milk powder. Detection was performed with the ECL system after incubation with rabbit anti-GFP and polyclonal pig anti-rabbit. Western intensities for each sample were extracted using ImageJ (Fig. S2.2). The number of cells loaded was estimated by calculating the relationship between OD600 and CFU counts. Details of the data analysis procedures are in the Supplementary Text.

Correlating protein and RNA levels per cell by quantitative PCR

Native and evolved single-promoter populations were grown in three biological replicates by diluting overnight LB cultures 500-fold into M9 media supplemented with glucose. These cultures were grown for 2.5 hours, stabilized with an equal volume of RNA Later (Sigma-Aldrich) and RNA was extracted using the Total RNA Purification 96-Well Kit (Norgen Biotek Corp.) with on-column DNAse I digestion. Reverse transcription was done using random hexamers and qPCR with TaqMan probes and performed by Eurofins Medigenomix GmbH (Germany). Three technical replicates were performed. The efficiency of the primers and probes used were validated in a dilution series. Relative RNA levels per cell were obtained by normalizing to the reference gene *ihfB* using a Bayesian procedure for integrating data from the replicates and accounting for failed measurements (Supplementary Text). The primers and probes used were: GFP forward primer: 5'-CCTGTCCTTTTACCAG-ACAA-3'; GFP reverse primer: 5'- GTGGTCTCTCTTTTCGTTGGGAT-3'; GFP probe: 5'-TACCTGTCCACACAATCTGCCCTTTCG-3', ihfB forward primer: 5'-GTTTCGGC-AGTTTCTCTTTG -3', ihfB reverse primer: 5'- ATCGCCAGTCTTCGGATTA-3', ihfB probe: 5'-ACTACCGCGCACCACGTACCGGA-3').

Minimal variance as a function of mean expression and excess noise

In a simple model of gene expression in which there are constant rates of transcription, translation, mRNA decay, and protein decay, the probability distribution for the number of proteins per cell is a negative binomial with variance proportional to the mean $\langle n\rangle$: $\mathrm{var}(n) = (b+1)\langle n\rangle$, where the constant b is the ratio between the mRNA translation rate and the mRNA decay rate, which is often referred to as 'burst size' (Shahrezaei & Swain, 2008). However, in general there are also cell-to-cell fluctuations in the transcription, translation, and decay rates, which are proportional to these rates themselves. These fluctuations lead to an additional term in the variance $\mathrm{var}(n)$ which is proportional to the square of the mean: $\mathrm{var}(n) = \beta\langle n\rangle + \sigma^2_{ab}\langle n\rangle^2$, where β is a renormalized burst size and σ^2_{ab} is the relative variance of the product of transcription, translation, and decay rates across cells (Supplementary Text).

The total fluorescence in a cell (measured in units equivalent to number of GFP proteins) n_{meas} can then generally be written as: $n_{\mathrm{meas}} = n_{\mathrm{bg}} + \epsilon\sqrt{\mathrm{var}(n)}$, where n_{bg} is background fluorescence and ϵ is a fluctuating quantity with mean zero and variance one. Assuming that the fluctuations are small relative to the mean, we then find for the variance of the logarithm of n_{meas}:

$$\mathrm{var}\left(\log[n_{\mathrm{meas}}]\right) = \sigma^2_{ab}\left(1 - \frac{n_{\mathrm{bg}}}{\langle n_{\mathrm{meas}}\rangle}\right)^2 + \frac{\beta}{\langle n_{\mathrm{meas}}\rangle}\left(1 - \frac{n_{\mathrm{bg}}}{\langle n_{\mathrm{meas}}\rangle}\right).$$

We fit this functional form to the minimum variance $\mathrm{var}\left(\log[n_{\mathrm{meas}}]\right)$ as a function of the mean, with $\sigma^2_{ab} = 0.025$ and $\beta = 450$. We defined the excess variance as the difference between the measured variance and this fitted minimal variance. A more detailed derivation is given in the Supplementary Text.

The FACS selection function

By comparing the distributions of the population's expression levels before and after rounds of selection (without intervening mutation of the promoters), we found that the probability that a cell with expression level x is selected by the FACS is well-approximated as $f(x|\mu_*, \tau) = \exp\left[-\frac{(x-\mu_*)^2}{2\tau^2}\right]$, with μ_* the desired expression level and τ the width of the selection window. For the last 3 rounds of selection for medium expression, we estimated $\tau \approx 0.03$ and μ_* fluctuated slightly around an average value of $\mu_* \approx 8.1$.

With this selection function, a promoter genotype that exhibits a distribution of expression values with mean μ and standard-deviation σ has a fitness (fraction of cells selected in the FACS) of

$$f(\mu, \sigma|\mu_*, \tau) = \sqrt{\frac{\tau^2}{\tau^2+\sigma^2}} \exp\left[-\frac{(\mu-\mu_*)^2}{2(\tau^2+\sigma^2)}\right]. \tag{2.1}$$

This estimated fitness function indicated that the fitness of promoter genotypes strongly depends on their mean μ and is almost independent of their excess noise. In addition, applying additional rounds of selection of varying strengths to the population of evolved promoters did not systematically alter their distribution of excess noise levels. Details of the analysis of the FACS selection are in the Supplementary Text.

Model for the evolution of gene regulation in a fluctuating environment

Although the model we present can be extended to include the evolution of gene regulation for multiple genes, for simplicity we focused on the evolution of a single gene and its promoter. We assumed that the population experienced a sequence of different environments and that, in each environment, the fitness of each organism is a function of its gene expression level. We characterized the fitness function in each environment by two parameters: the desired level μ_e that maximizes the fitness and a parameter τ that quantifies how quickly fitness falls away from this optimum and, for simplicity and analytical tractability, we assumed a Gaussian form: $f(x|\mu_e, \tau) = \exp\left[-\frac{(x-\mu_e)^2}{2\tau^2}\right]$. Note that this is the same form as the FACS selection function. Consequently, the fitness $f(\mu, \sigma|\mu_e, \tau)$ of a promoter with mean μ and variance σ^2 is given by equation (2.1) as well, with μ_e replacing μ_*.

The total number of offspring that a promoter will leave behind after experiencing all environments is given by the product of its fitness in each of the environments. Equivalently, the log-fitness of a promoter is proportional to its average log-fitness across all environments. We then find for the log-fitness:

$$\log\left[f(\mu, \sigma)\right] = -\frac{(\mu - \langle\mu_e\rangle)^2 + \mathrm{var}(\mu_e)}{2(\tau^2+\sigma^2)} + \frac{1}{2}\log\left[\frac{\tau^2}{\tau^2+\sigma^2}\right],$$

where $\langle \mu_e \rangle$ is the average of the desired expression levels across environments, and $\text{var}(\mu_e)$ is the variance in the desired expression levels across conditions. If we do not consider gene regulation, but simply optimize the promoter's mean expression and noise level, then we find optimal log-fitness occurs when $\mu = \langle \mu_e \rangle$ and $\sigma^2 = 0$ (when $\text{var}(\mu_e) < \tau^2$) or $\sigma^2 = \text{var}(\mu_e) - \tau^2$ otherwise. That is, when the desired expression level varies more than the width of the selection window, noise is increased so as to ensure the distribution overlaps the desired levels across all conditions. This result is equivalent to previous results on the evolution of phenotypic diversity in fluctuating environments (Bull, 1987).

To increase fitness, a promoter can evolve to become regulated by one of the regulators existing in the genome. Instead of having a constant mean expression μ, the promoter's mean expression will then become a function of the environment e: $\mu(e) = \mu + cr_e$, where r_e is the mean expression (or more generally regulatory activity) of the regulator in environment e, and c is the coupling strength. Since any gene will have some variability in its expression, we assumed that the actual expression/activity of the regulator in environment e is Gaussian distributed with a variance σ_r^2. Consequently, when coupled to the regulator, the promoter's total expression variance will become $\sigma_{\text{tot}}^2 = \sigma^2 + c^2\sigma_r^2$, and the log-fitness of the promoter becomes:

$$\log\left[f(\mu, \sigma, c)\right] = -\frac{\langle (\mu + cr_e - \mu_e)^2 \rangle}{2(\tau^2 + \sigma^2 + c^2\sigma_r^2)} + \frac{1}{2}\log\left[\frac{\tau^2}{\tau^2 + \sigma^2 + c^2\sigma_r^2}\right].$$

Assuming that the basal expression μ is optimized to maximize log-fitness, i.e. $\mu = \langle \mu_e \rangle - c\langle r_e \rangle$, this log-fitness can be rewritten as:

$$\log\left[f(X, Y, S, R)\right] = \text{cons.} - \frac{1}{2}\frac{Y^2(1 - X^2) + (SX - RY)^2}{1 + X^2} - \frac{1}{2}\log[1 + X^2].$$

where X measures the coupling strength ($X^2 = \frac{c^2\sigma_r^2}{\tau^2+\sigma^2}$), Y is the expression mismatch that measures how much the desired expression level varies across environments ($Y^2 = \frac{\text{var}(\mu_e)}{\tau^2+\sigma^2}$), S is the signal-to-noise of the regulator ($S^2 = \frac{\text{var}(r_e)}{\sigma_r^2}$), and R is the Pearson correlation between the desired expression levels μ_e and the activity levels r_e of the regulator. Additional details on this derivation and analysis of the behavior of the fitness function as a function of its parameters are given in the Supplementary Text.

Analysis of excess noise against gene expression variation and regulatory inputs

We re-annotated the promoter fragments of (Zaslaver et al., 2006) by mapping the published primer pairs to the *E. coli* K12 MG1655 genome. Of the 1816 promoter fragments, 1718 could be unambiguously associated with a gene that was immediately downstream, and the 1718 promoter fragments were associated with 1137 different downstream genes (for some genes, there were multiple or repeated upstream promoter fragments). We used the operon annotations of RegulonDB (Salgado et al., 2013) to extract, for each promoter, the set of additional downstream genes that are part of the same operon as the first downstream gene. We obtained known regulatory interactions between transcription factors and genes from RegulonDB and counted, for each *E. coli* gene, the number of transcription factors known to regulate the gene. We defined the number of regulatory inputs of a promoter to equal the average of the number of inputs for all genes in the operon downstream of the promoter. We sorted promoters by their excess noise and, as a function of a cut-off on excess noise level, calculated the mean and standard-error of the number of regulatory inputs for all promoters with excess noise level above the cut-off. We obtained genome-wide gene expression measurements from the Gene Expression Database (`<http://genexpdb.ou.edu/>`). For each *E. coli* gene, we obtained 240 log fold-change values x corresponding to the logarithm of the expression ratio of the gene in a perturbed and a reference condition. We defined the variance in expression of a gene as the average of x^2 across the 240 experiments. We again sorted promoters by their excess noise and, as a function of a cut-off on excess noise level, calculated the mean and standard error of gene expression variances for all promoters with excess noise above the cut-off.

Acknowledgments

We thank D. Blank for performing the Western Blots, D. Bumann for discussions, B. Claudi and J. Zankl for assistance with flow cytometry, and S. Abel and I. de Jong for assistance with microscopy.

OS and EvN designed the study. Experiments were designed by OS and LW and performed by LW. LW, OS, and EvN analyzed the data, and EvN developed the theoretical model. LW, OS, and EvN wrote the paper.

Supplementary Figures

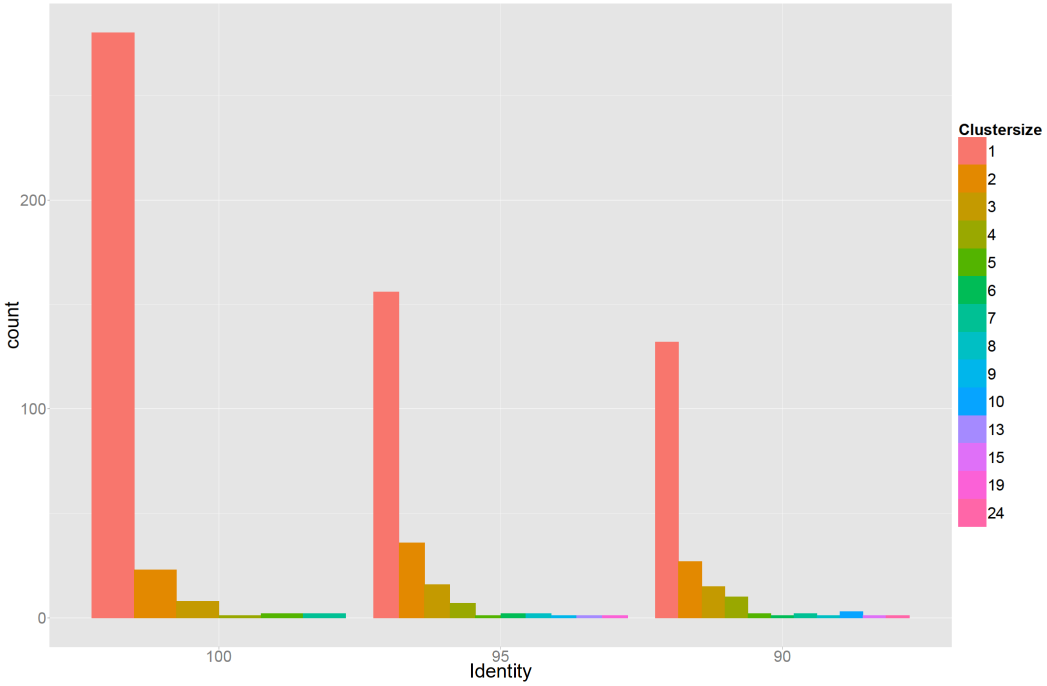

Figure S2.1: Genetic diversity of 378 sequenced promoters, which were extracted from randomly selected clones from the populations that were obtained after 3 and 5 rounds of selection. Sequences were clustered using single-linkage based on 100, 95, or 90 percent sequence identity (left, middle, and right panel) and the bar plots show the corresponding histograms of cluster sizes. The results indicate that the promoters in the populations at the third and fifth rounds are highly diverse, deriving from many different initial random sequences in the initial library.

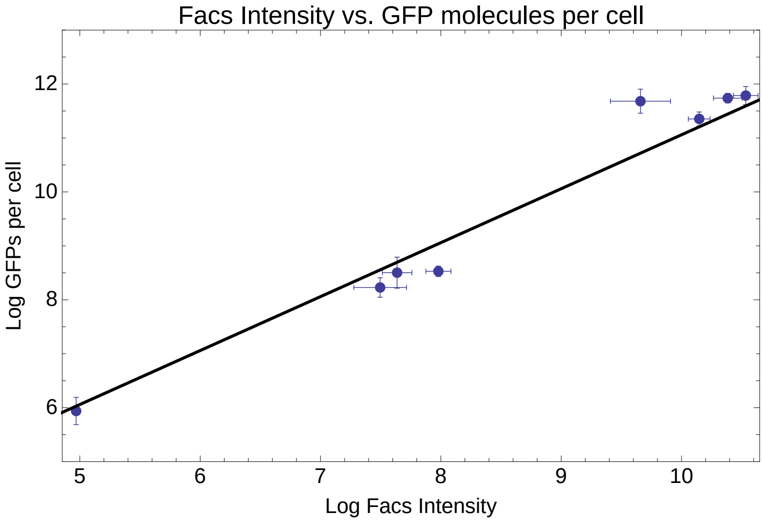

Figure S2.2: Mean log-fluorescence intensities as measured by FACS (horizontal axis) against estimated log GFP molecules per cell (vertical axis) as estimated from quantitative Westerns (see **Supplementary Text**) for 8 selected promoters. Error-bars were estimated from 3 replicates for the FACS measurements and 6 replicates for the GFP levels. The straight line shows the fit $y = x + 1.06$, which is equivalent to: GFP molecules per cell $= 2.88 *$ mean FACS intensity.

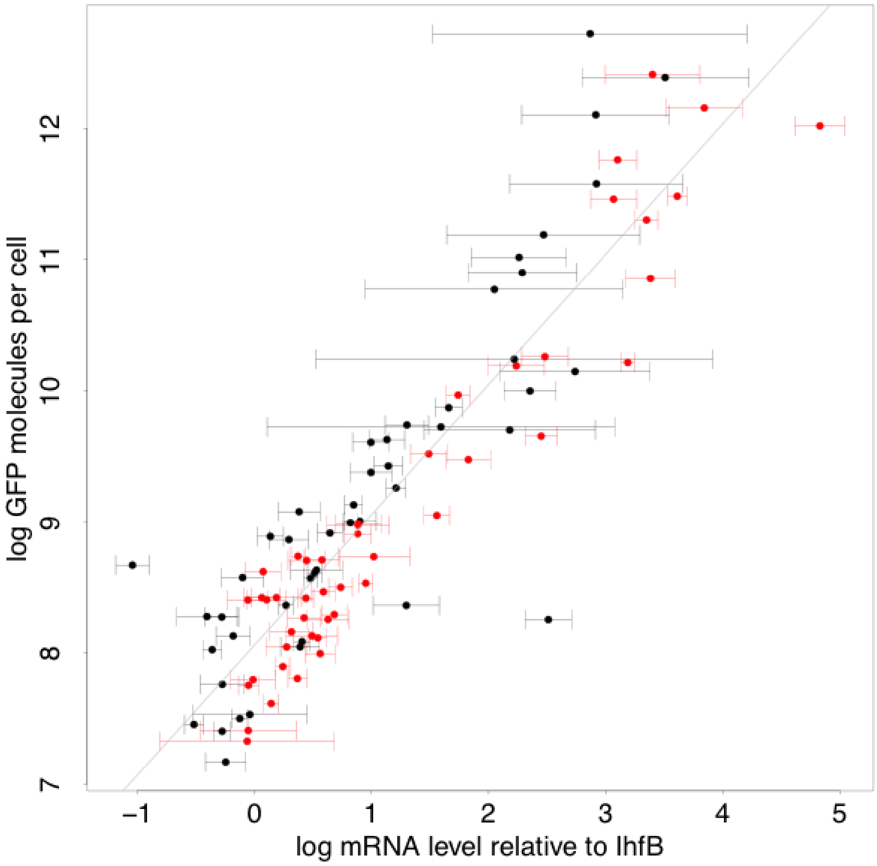

Figure S2.3: Relationship between log protein levels as measured by GFP intensity in FACS (vertical axis) and log mRNA levels (horizontal axis). The mRNA levels are estimated relative to the mRNA level of reference gene IhfB. Error bars show plus and minus one standard-deviation of the posterior probability distribution on mRNA levels (**Supplementary Text**). Black data points correspond to native promoters and red data points to synthetic promoter. The straight line shows a linear fit with slope 1, i.e. the best fit to a model where the protein level p is directly proportional mRNA level m, $\log(p) = c + \log(m)$, with $c = 7.06$ (**Supplementary Text**).

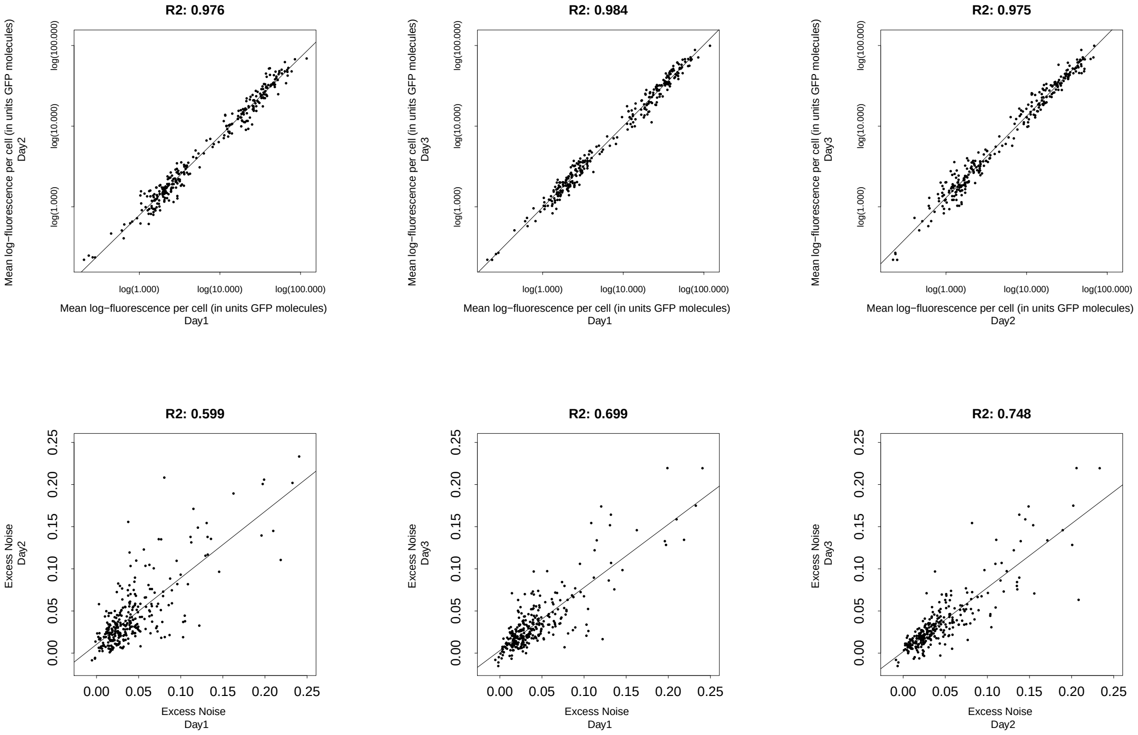

Figure S2.4: Comparison of three biological replicate FACS measurements of means and excess noise of log-fluorescence for evolved *E. coli* promoters. The top 3 panels compare mean log-fluorescences across 3 replicates and the bottom 3 panels compare excess noise in log-fluorescences across 3 replicates. The Pearson squared-correlation coefficients between pairs of replicate measurements are indicated at the top of each panel.

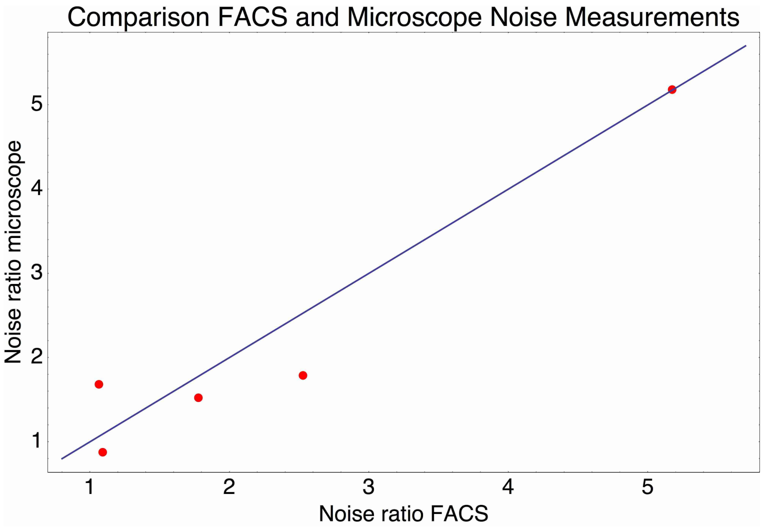

Figure S2.5: Relative noise levels (variance of the log-expression distribution) of 5 pairs of native promoters that have very similar mean expression levels. Each dot correspond to one of the pairs of promoters and shows the ratio of the noise level of the highest noise promoter to that of the lower noise promoter as measured by FACS (horizontal axis) and by microscope (vertical axis). The blue line shows the line $y = x$.

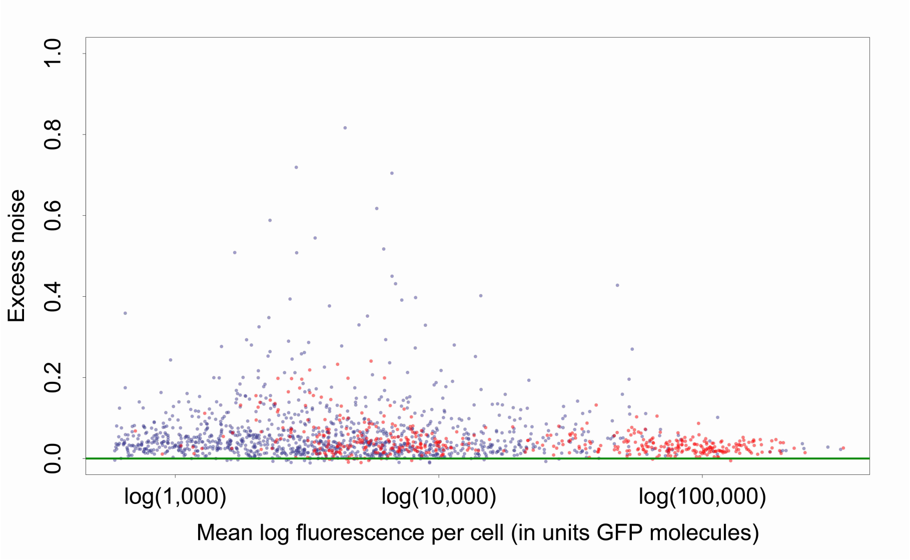

Figure S2.6: Mean log-fluorescence (horizontal axis) and excess noise levels (vertical axis), i.e. the different between variance of log-fluorescence levels and the minimal variance at the corresponding mean, for all native (blue dots) and synthetic (red dots) promoters. Both axes are in units of number of GFP molecules. Note that, in contrast to raw variances in log-fluorescence, that show a clear dependence on mean log-fluorescence, the excess noise levels show no dependence on mean.

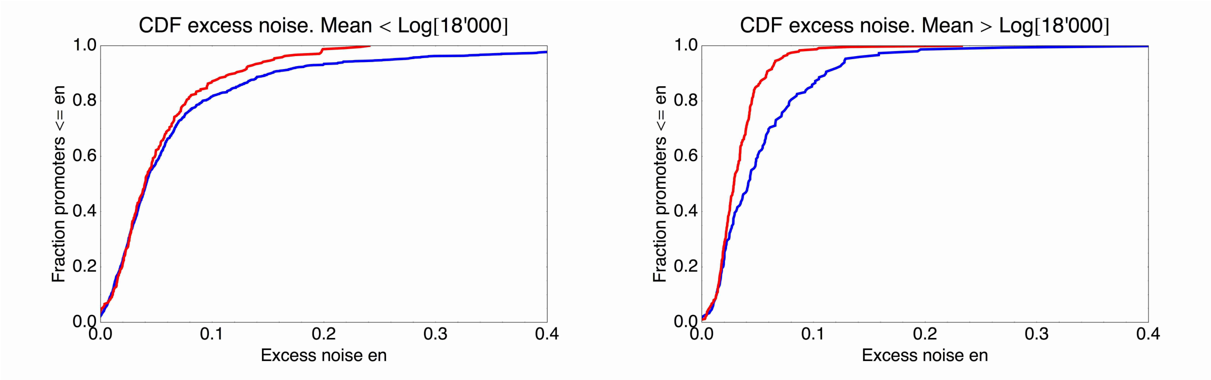

Figure S2.7: Cumulative distributions of excess noise levels for the native (blue) and synthetic promoters (red). The left panel shows the cumulative distribution of excess noise for promoters whose mean log-expression was less than log(18000) (corresponding to the medium expressing synthetic promoters), and the right panel for promoters with mean log-expression more than log(18000) (corresponding to the high expressing synthetic promoters). High noise promoters are clearly enriched among native promoters for both medium and high expressing promoters.

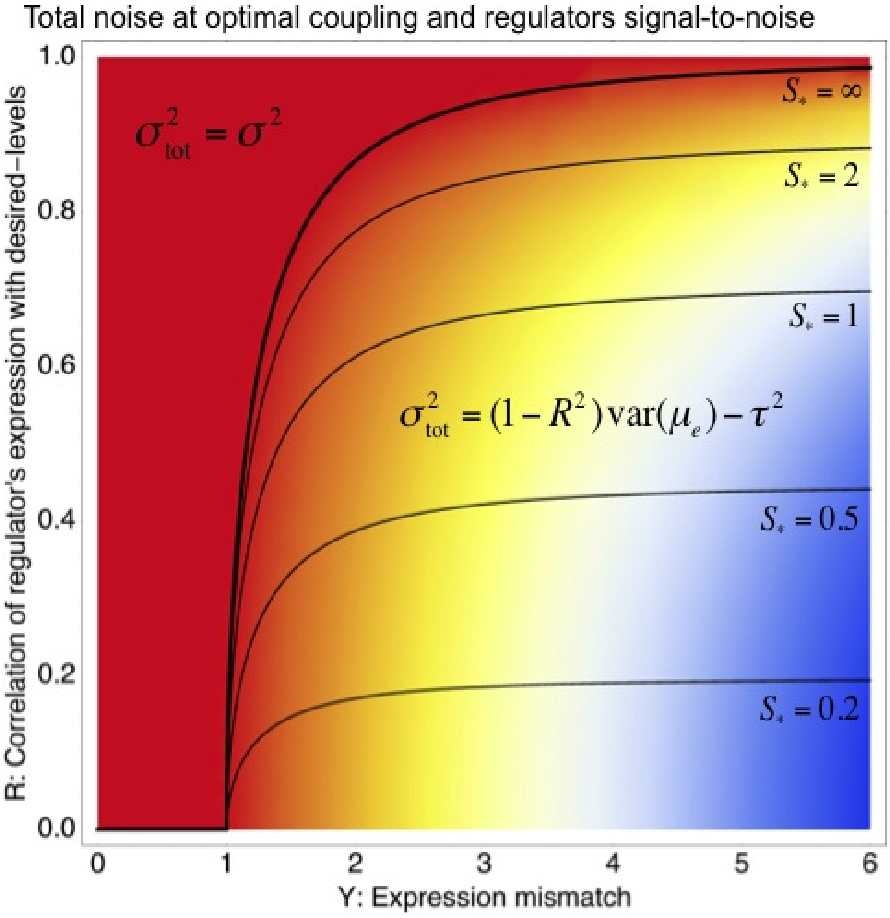

Figure S2.8: Phase diagram of the total noise σ_{tot} of a promoter with expression mismatch Y (horizontal axis) that is coupled (at optimal coupling strength) to a regulator whose regulatory activities have correlation R with the desired expression levels (vertical axis) and whose signal-to-noise ratio S has also been optimized. The colors indicate the value of σ_{tot}, running from σ_{tot} equal to the noise σ of the *unregulated* promoter (red) to $\sigma_{\text{tot}} = 6\sigma$. A phase boundary (thick black curve) separates solutions in a 'basal noise regime' at the top left, where the total noise equals the minimal noise σ^2, and solutions in an 'environment-driven noise regime' at the bottom right, where the total noise matches the variance in desired levels that is not tracked by the regulation, i.e. $\sigma_{\text{tot}}^2 = (1 - R^2)\text{var}(\mu_e) - \tau^2$. The contours show optimal signal-to-noise ratios S_* as a function of Y and R. Note that S_* diverges at the phase boundary.

2.6 Supplementary Text

Estimating the mean and variance of log-fluorescence levels from FACS data

Visual inspection of the distributions of fluorescence intensities for individual cells containing the same promoter construct shows that almost all of these distributions can be well approximated by a log-normal distribution. We thus chose to characterize the distribution of expression levels of each promoter by the mean and variance of log-fluorescence intensities across cells. Visual inspection of the distributions also indicated that, for almost all promoters, there is a small number of measurements with aberrantly high or low values that are likely due to some measurement artefact, and we designed a Bayesian procedure for automatically discounting these aberrant measurements.

For each clone we typically have around $N = 5000$ independent FACS intensities measured. We denote by x the log-intensity (using natural logs) of an individual cell. We first calculate the mean and variance without taking outliers into account, i.e.

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{2.2}$$

and

$$\mathrm{var}(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2, \tag{2.3}$$

where x_i is the log-intensity of cell i. We call these the ‘original’ mean and variance.

Next we take outliers into account. We assume that, of all N measurements, only a fraction ρ are ‘correct’ measurements, and the other $(1 - \rho)$ are ‘outliers’, meaning that these are erroneous measurements. We assume that these ‘outliers’ derive from a uniform distribution that spans the range of measurements $R = (x_{\max} - x_{\min})$. Finally, we assume that the distribution of ‘correct’ measurements is approximately Gaussian with (unknown) mean μ and variance σ^2. Under these assumptions, the probability of a measurement of log-intensity x is given by

$$P(x|\mu, \sigma^2, \rho) = \frac{\rho}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] + \frac{1-\rho}{R}. \tag{2.4}$$

The probability of the entire data-set for a clone is then simply given by

$$P(D|\mu, \sigma^2, \rho) = \prod_{i=1}^{N} P(x_i|\mu, \sigma^2, \rho). \tag{2.5}$$

We then maximize this probability with respect to μ, σ^2, and ρ. This can be easily done using Expectation-Maximization. The resulting mean μ, and variance σ^2 are corrected for outliers.

Inferring the relation between FACS intensity and GFP molecules per cell

To infer the relationship between FACS intensity per cell and GFP molecules per cell we used quantitative Westerns. For each of 8 strains of known FACS intensities, we extracted the protein contents from a fixed number of cells and quantified total GFP intensity. In the same experiment the GFP intensities were measured for known amounts of GFP ranging from 10 to 100 nanograms. We performed 3 replicate experiments. In each replicate we measured the GFP intensity of the 8 strains, as well as 'reference' intensities of bands loaded with 10, 25, 50, 75, and 100 nanograms of GFP. We measured intensities from these gels using both 10 second and 20 second exposure times, giving a total of 6 replicate measurements of the reference amounts and the 8 strains.

Figure S2.9 shows the measured GFP intensities I as a function of the amount of GFP w (weight in grams) for the reference bands, in each of the 6 replicate experiments. Note that there are 5 points, corresponding to weights of 10, 25, 50, 75, and 100 nanograms in each curve.

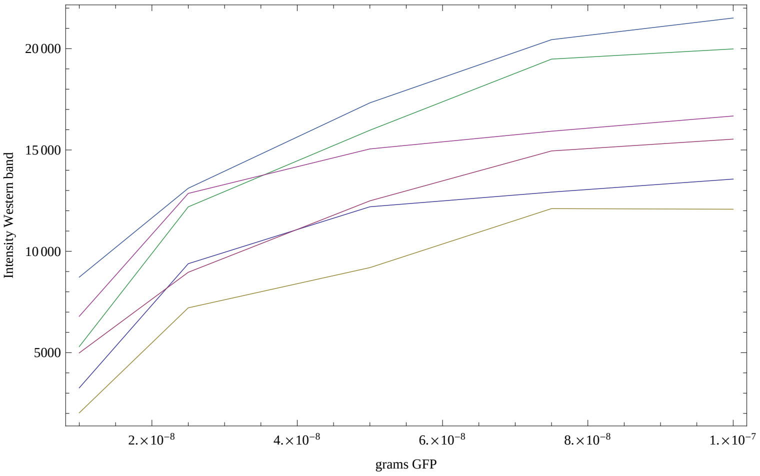

Figure S2.9: Measured intensities of the GFP reference bands as a function of the amount of GFP (in grams) loaded on each band. Each curve corresponds to one replicate (shown in a separate color), and each curve has 5 data-points.

The curves show that the measured intensities are saturating as the amount of GFP increases. Second, the intensity scale varies significantly from replicate to replicate. The simplest linear relationship between I and w that includes saturation is of the form

$$I = I_{\max} \frac{w}{w + w_0}, \tag{2.6}$$

and inspection of the curves shows that each of them can be reasonably well fitted to this functional form. To infer the amount of GFP corresponding for a particular strain in a particular replicate, we need to infer w as a function of the measured value I. We thus invert the relationship and find the general form

$$w = w_0 \frac{I}{I_{\max} - I}. \tag{2.7}$$

In other words, our functional form assumes that for a suitably chosen value $I_{\max}$, the weight w becomes directly proportional to the transformed variable $I/(I_{\max} - I)$. As an example, **Figure S2.10** shows that for the first replicate, when plotting w as a function of $I/(15631 - I)$, i.e. with a value of $I_{\max} = 15'631$, we obtain an approximately linear relationship. Similar approximately linear relationships are observed for the other replicates as well.

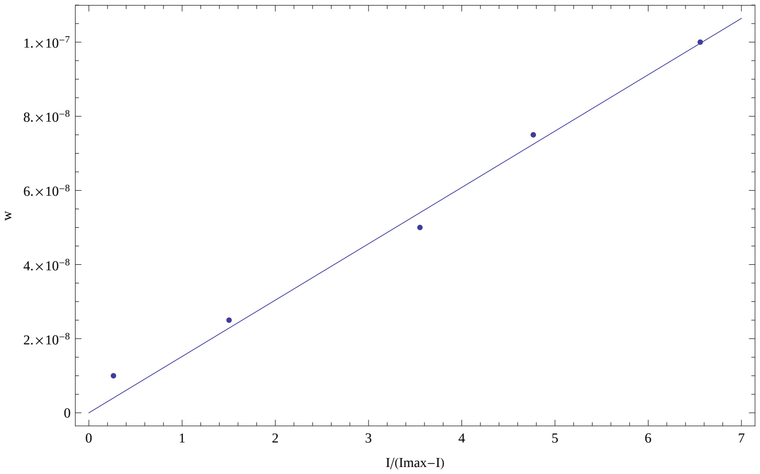

Figure S2.10: For the first replicate, we inferred a saturation value $I_{\max} = 15'631$. Plotting w as a function of $I/(I_{\max} - I)$ we obtain an approximately linear relationship that also approximately goes through the origin $(0,0)$ (as it should).

To fit w as a function of I for each replicate, we assume that the difference between w and $I/(I_{\max} - I)$ is Gaussian distributed with unknown variance σ^2. That is, for each data-point i in a replicate, the weight w_i and its intensity I_i are related through

$$w_i = \epsilon_i + w_0 \frac{I_i}{I_{\max} - I_i}, \tag{2.8}$$

with ϵ_i the noise, which is Gaussian distributed with unknown variance σ^2, i.e.

$$P(\epsilon|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\epsilon^2}{2\sigma^2}\right]. \tag{2.9}$$

Using this, the probability of the observed data in a titration curve, given parameters $I_{\max}$, w_0, and σ is:

$$P(\{w\}|\{I\}, I_{\max}, w_0, \sigma) \propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(w_i - w_0 \frac{I_i}{I_{\max} - I_i}\right)^2\right], \tag{2.10}$$

where $n = 5$ is the number of points in a titration curve, and we have ignoted factors of $\sqrt{2\pi}$ for convenience.

Imagine that we augment our data-set $(\{w\}, \{I\})$ with a single data-point (w_s, I_s), where I_s is the measured intensity of strain s and w_s is a hypothesized amount of GFP for this strain. The probability of this entire data-set is given by

$$P(\{w\}, w_s|\{I\}, I_s, I_{\max}, w_0, \sigma) = \frac{1}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(w_i - w_0 \frac{I_i}{I_{\max} - I_i}\right)^2 + \left(w_s - w_0 \frac{I_s}{I_{\max} - I_s}\right)^2\right], \tag{2.11}$$

Formally, we now need to specify a prior $P(I_{\max}, w_0, \sigma)$ and integrate over these unknown parameters. We will use a uniform prior over $I_{\max}$ and w_0, and a scale prior $1/\sigma$ for σ. That is, formally we want to calculate

$$P(\{w\}, w_s|\{I\}, I_s) = \int dI_{\max} dw_0 \frac{d\sigma}{\sigma} P(\{w\}, w_s|\{I\}, I_s, I_{\max}, w_0, \sigma). \tag{2.12}$$

Note, if we additionally integrate over w_s we obtain

$$P(\{w\}|\{I\}, I_s) = \int dw_s P(\{w\}, w_s|\{I\}, I_s) \tag{2.13}$$

and dividing by this we obtain the posterior distribution of w_s:

$$P(w_s|\{w\}, \{I\}, I_s) = \frac{P(\{w\}, w_s|\{I\}, I_s)}{P(\{w\}|\{I\}, I_s)}. \tag{2.14}$$

To perform the integrals in (2.12), we first simplify the notation by denoting the new data-point (w_s, I_s) as (w_{n+1}, I_{n+1}), i.e. as if it was the $(n+1)$st data-point. The integrand now takes the form

$$P(\{w\}|\{I\}, I_{\max}, w_0, \sigma) = \frac{1}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n+1} \left(w_i - w_0 \frac{I_i}{I_{\max} - I_i}\right)^2\right]. \tag{2.15}$$

To further simplify notation, we write $y_i = I_i/(I_{\max} - I_i)$, keeping in mind that the values of y depend on $I_{\max}$. Further, for any quantity x that takes on values x_i over the 5 titration points and the added point, we write averages like

$$\langle x^2 \rangle = \frac{1}{n+1} \sum_{i=1}^{n+1} (x_i)^2, \tag{2.16}$$

and so on. The integrand can then be rewritten as

$$P(\{w\}|\{I\}, I_{\max}, w_0, \sigma) = \frac{1}{\sigma^{n+1}} \exp\left(-\frac{n+1}{2\sigma^2}\left[\langle w^2 \rangle - 2w_0 \langle wy \rangle + w_0^2 \langle y^2 \rangle\right]\right). \tag{2.17}$$

Performing the integral over w_0 we obtain

$$P(\{w\}|\{I\}, I_{\max}, \sigma) = \frac{1}{\sigma^n \sqrt{\langle y^2 \rangle}} \exp\left[-\frac{(n+1)\langle w^2 \rangle}{2\sigma^2}\left(1 - \frac{\langle wy \rangle^2}{\langle w^2 \rangle \langle y^2 \rangle}\right)\right], \tag{2.18}$$

where we have again ignored prefactors that cancel in the final posterior for w_s, i.e. equation (2.14). We next integrate over σ. Performing this integral we obtain

$$P(\{w\}|\{I\}, I_{\max}) = \left(1 - \frac{\langle wy \rangle^2}{\langle w^2 \rangle \langle y^2 \rangle}\right)^{-n/2} \langle y^2 \rangle^{-1/2}. \tag{2.19}$$

Notice that the key expression in parentheses is simply one minus the squared correlation coefficient between the variables w and y, i.e.

$$r^2(y, w) = \frac{\langle wy \rangle^2}{\langle w^2 \rangle \langle y^2 \rangle}. \tag{2.20}$$

In other words, the values of w_s and $I_{\max}$ that maximize the probability are those such that the linear correlation between the resulting values of y and the w values is maximal.

We next need to perform the integral over $I_{\max}$. Since this integral cannot be performed analytically, we performed the integrals over $I_{\max}$ numerically, separately for each strain s and each of the 6 replicates. Finally, for each replicate r and strain s, we determined the values w_{rs} that has maximal posterior probability. These are our estimated GFP amounts (in grams) for each strain and replicate (**Fig. S2.11**).

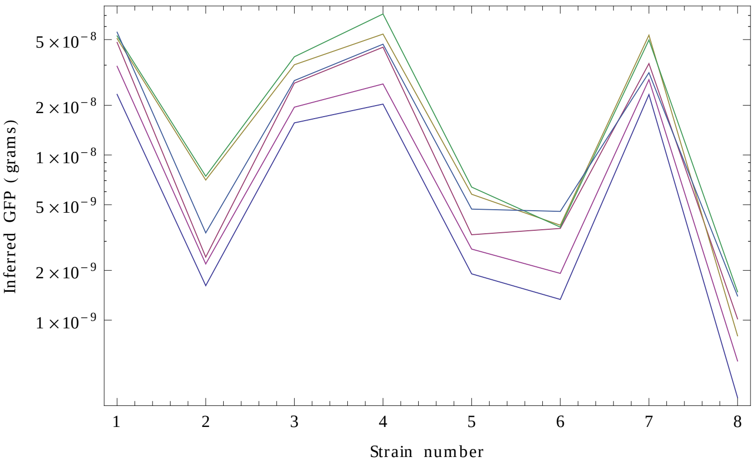

Figure S2.11: Inferred GFP amounts (in grams, vertical axis) for the 8 strains (strain numbers shown along the horizontal axis) using the reference data from each replicate. Each color corresponds to a replicate. The vertical axis is shown on a logarithmic scale.

Although the inference clearly separates the high expressed from the low expressed clones, curves from the different replicates seem to be separated by constant shifts from each other. Since the vertical axis is shown on a logarithmic scale, this means that the curves differ by common multiplicative factors. This difference in scale is almost certainly due to an experimental artefact and we will thus normalize for it.

Let $w_s(i)$ be the inferred amount of strain s in replicate i. To account for the variability in overall scale, we normalize the inferred log-weights in each replicate by calculating the average log-weight in the replicate, i.e.

$$\mu_i = \frac{1}{8} \sum_{s=1}^{8} \log \left[w_s(i) \right], \tag{2.21}$$

and a total average scale of the replicates

$$\mu = \frac{1}{6} \sum_{i=1}^{6} \mu_i, \tag{2.22}$$

and then transforming the estimated shifts as follows:

$$w_s(i) \to \tilde{w}_s(i) = w_s(i) e^{\mu - \mu_i}. \tag{2.23}$$

In addition, dividing the weight $\tilde{w}_s(i)$ by the known weight of a single GFP molecule (4.48210^{-20} grams), we get an estimate of the number of GFP molecules in the bands for each strain. Finally, we used OD measurements to estimate the number of cells loaded on each band, and divided by these to obtain an estimate of the number of GFP molecules per cell for each of the strains across each of the replicates. **Figure S2.12** shows the inferred GFP molecules per cell for each strain after normalization, which indeed show much less variation across the replicates.

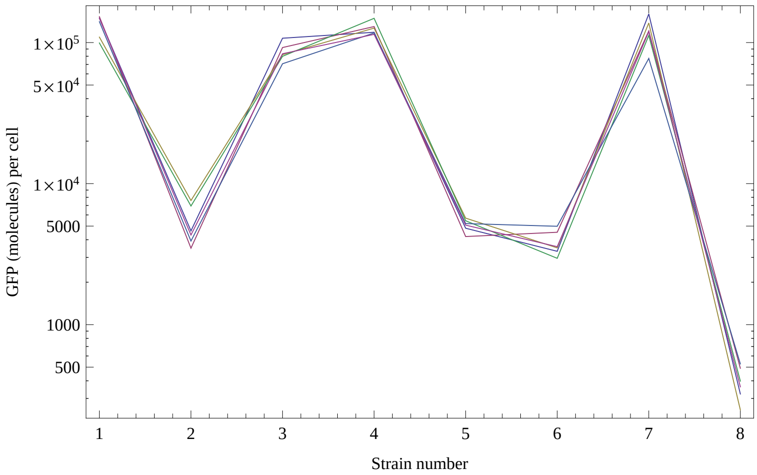

Figure S2.12: Normalized inferred GFP amounts (molecules per cell, vertical axis) for the 8 strains (strain numbers shown along the horizontal axis) using the reference data from each replicate. Each color corresponds to a replicate. The vertical axis is shown on a logarithmic scale.

Finally, we compare the inferred GFP amounts for each strain, with the FACS intensities measured for that strain. Observing that the variation in both estimated FACS intensities and GFP molecules per cell increases with the mean, it is most natural to compare GFP and FACS levels on a logarithmic scale. Let f_s denote the true log-FACS intensity and g_s denote the true log-GFP molecules per cell. Assuming that GFP molecules per cell and (background corrected) FACS intensity are directly proportional to each other, the log-levels are related through

$$g_s = f_s + c, \tag{2.24}$$

with c a constant. For each strain s, we calculated the mean log-FACS intensity $\langle f_s \rangle$ and its variance $\mathrm{var}(f_s)$ across replicate FACS , as well as the mean log-GFP molecules per cell $\langle g_s \rangle$ and its variance $\mathrm{var}(g_s)$ across the quantitative Westerns as described above. Assuming Gaussian deviations between the true and observed levels, the probability of the data given c is given by

$$P(D|c) \propto \prod_s \exp\left[-\frac{(\langle g_s \rangle - \langle f_s \rangle - c)^2}{2(\mathrm{var}(f_s) + \mathrm{var}(g_s))}\right]. \tag{2.25}$$

We thus find for the optimal value of c:

$$c_* = \sum_s \frac{\langle g_s \rangle - \langle f_s \rangle}{\mathrm{var}(f_s) + \mathrm{var}(g_s)} \left[\sum_s \frac{1}{\mathrm{var}(f_s) + \mathrm{var}(g_s)}\right]^{-1}. \tag{2.26}$$

Figure S2.2 shows the estimated log-FACS and log-GFP levels including their error bars, together with the optimal fit $c_* = 1.06$.

Consequently, if F is the FACS intensity of a strain (non-log), then the estimated number of GFP molecules per cell G is equal to $G = e^{1.06}F = 2.88F$. Note that, with these estimates, the highest expressed strain, with an average FACS intensity of 37′500, would have about 108′000 molecules of GFP per cell. The lowest expressed strain (with FACS intensity 143) would have 415 molecules per cell. From now on, we will multiply all FACS intensities by 2.88 so that a FACS intensity of I automatically corresponds to the fluorescence of I GFP molecules, i.e. we express FACS fluorescence intensities in units of GFP molecules per cell.

Comparing mRNA and protein levels

For 94 clones, we quantified mRNA levels using qPCR. The qPCR procedure uses a standard reference curve which allows it to infer the absolute number of molecules of the mRNA of interest in the input sample. Each input sample is created by extracting RNA from a certain number of cells (which we can estimate approximately), and reverse transcribing this RNA into cDNA. Unfortunately, both the total amount of cells used, as well as the efficiency of the reverse transcription can fluctuate significantly outside of our control, and this will make the total number of molecules detected fluctuate as well. To control for this, we always quantify the absolute number of molecules of two types of mRNAs in parallel for each sample; the mRNAs of the gene of interest, and the mRNAs of a reference gene which we are confident is constantly expressed. The reference gene we used was *ihfB*.

For each promoter of interest p, we obtained measured mRNA molecule numbers together with mRNA molecule numbers for the reference gene, in 3 separate biological replicates, and in 3 technical replicates for each biological replicate, i.e. 9 pairs of measurements in total. We denote the log-quantity of the mRNA of promoter p in biological replicate r and technical replicate i as x_{pri}, and the log-quantity of the reference gene in the same replicate as y_{pri}

(note that this depends on the promoter p because these quantities come from a common sample). To estimate a single log-ratio $x_p - y_p$ between the expression of the gene of interest and the reference gene, we will proceed as follows. First, we will integrate the data from the technical replicates to obtain biological replicate expression x_{pr} and y_{pr}. We then combine the differences $d_{pr} = x_{pr} - y_{pr}$ across the biological replicates, to obtain the final $d_p = x_p - y_p$.

The statistical model that we use assumes that the difference between the value x_{pri} measured in technical replicate i, and the true expression x_{pr} is Gaussian distributed with mean zero and an unknown variance σ_r^2. Note that we assume that this 'noise' is the same for all promoters p, but may fluctuate between biological replicates. We similarly assume the difference between y_{pri} and y_{pr} is Gaussian distributed with variance $\tilde{\sigma}_r^2$. We noted that there is a small fraction of measurements that deviate by large amounts from the measurements in other replicates. We assume that there is a small fraction of measurements that failed in some way, giving erroneous measurement values. To take this into account we will use a mixture model that asssumes a small fraction of the measurements come from a uniform distribution that spans the observed range of the data.

Let $R_r = \max_{p,i}(x_{pri}) - \min_{p,i}(x_{pri})$ denote the range of observed values in biological replicate r, and let ρ_r denote the fraction of measurements in replicate r that are meaningful, i.e. not erroneous. The probability of a single measurement x_{pri} given x_{pr}, the variance σ_r^2 and fraction ρ_r is given by

$$P(x_{pri}|x_{pr}, \sigma_r^2, \rho_r) = \frac{\rho_r}{\sqrt{2\pi}\sigma_r} \exp\left[-\frac{1}{2}\frac{(x_{pri} - x_{pr})^2}{2\sigma_r^2}\right] + \frac{1-\rho_r}{R_r}. \tag{2.27}$$

The probability of all technical replicates for all promoters is then simply given by the product over all promoters p and technical replicates i:

$$P(\{x_{pri}\}|\{x_{pr}\}, \sigma_r^2, \rho_r) = \prod_{p,i} P(x_{pri}|x_{pr}, \sigma_r^2, \rho_r). \tag{2.28}$$

We next maximize this likelihood with respect to the fraction ρ_r, the variance σ_r^2, and the expression levels x_{pr} for all promoters p. This optimization can be done using a straightforward Expectation Maximization scheme.

Expectation Maximization

Given a current estimate of x_{pr}, of the variance σ_r^2 and the fraction ρ_r, the posterior probability that the technical replicate with value x_{pri} was a meaningful measurement is given by

$$p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r) = \frac{\frac{\rho_r}{\sqrt{2\pi}\sigma_r}\exp\left[-\frac{(x_{pri}-x_{pr})^2}{2\sigma_r^2}\right]}{\frac{\rho_r}{\sqrt{2\pi}\sigma_r}\exp\left[-\frac{(x_{pri}-x_{pr})^2}{2\sigma_r^2}\right] + \frac{1-\rho_r}{R_r}}. \tag{2.29}$$

Using these posteriors, the updated value x'_{pr} is given by the mean of the technical replicate measurements, weighted by their posteriors

$$x'_{pr} = \frac{\sum_i x_{pri} p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r)}{\sum_i p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r)}. \tag{2.30}$$

Given current values of ρ_r and σ_r^2 we use these equations to iteratively update all x_{pr} until they converge. We then update the values of ρ_r and σ_r^2 using the following equations

$$\rho'_r = \frac{\sum_{p,i} p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r)}{\sum_{p,i} 1}, \tag{2.31}$$

i.e. the updated ρ'_r is the average of the current posteriors over all promoters and technical replicates. The update equation for the variance is given by

$$\sigma'^2_r = \frac{\sum_{p,i} (x_{pri} - x_{pr})^2 p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r)}{\sum_{p,i} p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r)}. \tag{2.32}$$

After each update of σ_r^2 and ρ_r all x_{pr} are updated until convergence again, and this is iterated until the σ_r^2 and ρ_r converge. Exactly analogous expectation-maximization equations are used to optimize the values $\tilde{\sigma}_r$, $\tilde{\rho}_r$, and all y_{pr} of the reference gene measurements.

Table S2.1 shows the fitted fractions and variances for each of the replicates. We see that for the majority of replicates the noise level lies around 0.01 (meaning a measurement error-bar of about 0.1 on log-expression), but it is two and three-fold higher for measurements of the genes of interest in two replicates. The fraction of correct measurements ranges from about 93% to almost 100% across the replicates.

replicate	σ^2	ρ	$\tilde{\sigma}^2$	$\tilde{\rho}$
1	0.0252	1.0	0.0116	0.934
2	0.0113	0.981	0.0118	0.988
3	0.0329	0.956	0.0072	0.955

Table S2.1: Fitted variances and fractions of meaningful measurements for the genes of interest (σ^2, ρ) as well as for the reference gene measurements ($\tilde{\sigma}^2$, $\tilde{\rho}$) for each of the three biological replicates.

When the variances and fractions have been optimized, we obtain the final technical replicate-averaged quantities x_{pr} and y_{pr} and we determine final variances σ_{pr}^2 and $\tilde{\sigma}_{pr}^2$ for each of these averages. These final variances are calculated as follows. For each promoter p and each biological replicate r, we determine the effective number of correct measurements as

$$n_{pr} = \sum_i p(i|x_{pri}, x_{pr}, \sigma_r^2, \rho_r), \tag{2.33}$$

and the final variance is then given by

$$\sigma_{pr}^2 = \frac{\sigma_r^2}{n_{pr}}. \tag{2.34}$$

Analogously, for the reference gene measurements we have

$$\tilde{n}_{pr} = \sum_i p(i|y_{pri}, y_{pr}, \tilde{\sigma}_r^2, \tilde{\rho}_r), \tag{2.35}$$

and the final variance

$$\tilde{\sigma}_{pr}^2 = \frac{\tilde{\sigma}_r^2}{\tilde{n}_{pr}}. \tag{2.36}$$

Combining the biological replicates

For each promoter p, we want to estimate the log-expression ratio $x_p - y_p$ by combining the estimated values x_{pr} and y_{pr} from each of the replicates, taking into account that these values have different variances for different replicates. For a protein p and replicate r, the estimated log-expression difference d_{pr} is

$$d_{pr} = x_{pr} - y_{pr}. \tag{2.37}$$

The variance τ_{pr}^2 associated with that estimated difference is

$$\tau_{pr}^2 = \sigma_{pr}^2 + \tilde{\sigma}_{pr}^2. \tag{2.38}$$

Inspection of the variation in d_{pr} across biological replicates, relative to their uncertainties τ_{pr}, makes it clear that, in addition to the uncertainty in each of the estimates d_{pr}, there is substantial variation in d_{pr} across the biological replicates which is quite different for different promoters. That is, for some promoters the biological replicates give very consistent d_{pr}, lying within the error-bars τ_{pr}, whereas for other promoters the variation in d_{pr} is much larger than the error-bars τ_{pr}, indicating that there must be additional variance across replicates.

We will assume that the true value d_{pr}^t is given by the mean d_p for the promoter plus a biological replicate variation δ_{pr}

$$d_{pr}^t = d_p + \delta_{pr}, \tag{2.39}$$

and we will assume that the deviation δ_{pr} is Gaussian distributed with mean zero and unknown variance τ_p^2. The probability of the estimate d_{pr} given its variance τ_{pr}^2, the true value d_p, and the biological replicate variance τ_p^2 is given by

$$P(d_{pr}|d_p, \tau_{pr}^2, \tau_p^2) = \frac{1}{\sqrt{2\pi(\tau_{pr}^2 + \tau_p^2)}} \exp\left[-\frac{(d_{pr} - d_p)^2}{2(\tau_{pr}^2 + \tau_p^2)}\right]. \tag{2.40}$$

The probability of the data combining all biological replicates for the promoter is simply

$$P(d_{p1}, d_{p2}, d_{p3}|d_p, \tau_p^2, \tau_{p1}^2, \tau_{p2}^2, \tau_{p3}^2) = \prod_{r=1}^{3} P(d_{pr}|d_p, \tau_{pr}^2, \tau_p^2) \tag{2.41}$$

For each promoter p, we now maximize this probability with respect to both τ_p^2 and d_p. Given a fixed value of τ_p^2, the optimal value of d_p is given by the weighted sum

$$d_p = \frac{\sum_r \frac{d_{pr}}{\tau_p^2+\tau_{pr}^2}}{\sum_r \frac{1}{\tau_p^2+\tau_{pr}^2}}. \tag{2.42}$$

Substituting this optimal d_p into the probability (2.41), the expression becomes a function of the variance τ_p^2 only, and we numerically determine the optimal value of τ_p^2 for each p. In this way we obtain a final estimate d_p for each promoter. The variance σ_p^2 associated with this final estimate is given by

$$\sigma_p^2 = \left[\sum_r \frac{1}{\tau_p^2+\tau_{pr}^2}\right]^{-1} \tag{2.43}$$

Suppl. Figure S2.3 shows the relationship between protein levels (estimated by FACS) and estimated mRNA levels for the 94 strains for which we measured mRNA levels using qPCR. We see there is a very good correlation between protein and mRNA levels (Pearson correlation $r^2 \approx 0.82$). Note that, for a given promoter, the average protein level p is related to average mRNA level m by the ratio of the translation rate λ and protein decay rate μ. That is,

$$p = \frac{\lambda}{\mu} m. \tag{2.44}$$

Since GFP is very stable compared to the duplication rate of our cells, for our system the protein decay rate μ is approximately equal to the growth rate of the cells, and thus constant across the promoters. Consequently, the fact that 82% of the variation in protein levels is explained by variations in mRNA levels, suggest that the translation rate λ shows relatively small variations across the strains. Below, we use this data to more rigorously estimate variation in translation rates across the strains.

Estimating relative translation rates

As before, we denote by d_p the relative (to *ihfB*) log-mRNA level of promoter p, and we will denote by y_p the log protein number per cell (as measured by FACS) for promoter p. Denoting by m the absolute number of mRNAs per cell for the reference gene *ihfB*, by λ the average translation rate, and by μ the protein decay rate (as a consequence of cell growth), y_p and d_p are related through

$$e^{y_p} = \frac{\lambda e^{\delta_p}}{\mu} e^{d_p} m, \tag{2.45}$$

where we have written the translation rate λ_p of promoter p in terms of the average translation rate λ, and a promoter specific deviation δ_p, i.e. $\lambda_p = \lambda e^{\delta_p}$. Defining $e^c = \lambda m/\mu$ we have

$$y_p = d_p + c + \delta_p. \tag{2.46}$$

Using that our estimate of d_p is Gaussian distributed with standard-deviation σ_p, and assuming that δ_p is Gaussian distributed with mean 0 and standard-deviation τ, the probability of our data given the σ_p, c, and τ is

$$P(\{d_p, y_p, \delta_p\}|c, \{\sigma_p\}, \tau) = \prod_p \frac{1}{2\pi\sigma_p\tau} \exp\left[-\frac{1}{2}\left(\frac{y_p - d_p - c - \delta_p}{\sigma_p}\right)^2 - \frac{\delta_p^2}{2\tau^2}\right]. \tag{2.47}$$

To estimate the variance τ^2 we integrate over all δ_p and c (using a uniform prior). To simplify the notation of the result we write $w_p = 1/(\sigma_p^2 + \tau^2)$, and $\Delta_p = y_p - d_p$. We then have

$$P(\{d_p, y_p\}|\{\sigma_p\}, \tau) \propto \frac{\prod_p \sqrt{w_p}}{\sqrt{\sum_p w_p}} \exp\left[-\frac{1}{2}\left(\sum_p w_p\Delta_p^2 - \frac{(\sum_p w_p\Delta_p)^2}{\sum_p w_p}\right)\right]. \tag{2.48}$$

We numerically determine the value of τ that maximizes this likelihood and find $\tau_* = 0.47$. Using this maximum likelihood value of τ, the maximum likelihood value of c is given by

$$c_* = \frac{\sum_p w_p\Delta_p}{\sum_p w_p} = 7.06. \tag{2.49}$$

The fit $y = c + d$ is shown as the black line in Fig. S S2.3.

Finally, using τ_* and c_*, we determine the most likely values of the δ_p. We find

$$\delta_p = \frac{\Delta_p - c_*}{1 + \sigma_p^2/\tau_*^2}, \tag{2.50}$$

with a standard-deviation of

$$\sigma(\delta_p) = \left(\frac{1}{\tau_*^2} + \frac{1}{\sigma_p^2}\right)^{-1/2}. \tag{2.51}$$

Figure S2.13 shows the estimated values of δ_p, together with their error bars $\sigma(\delta_p)$, as a function of the log protein level y_p. We see that, for the large majority of promoters, the estimated translation rate is within $2 - 3$ fold of the average translation rate (i.e. $|d_p| < 1$), confirming that there is relatively little variation in translation rates. For the most extreme example, the translation rate is approximate $e^{1.9} = 6.6$ fold lower than the average translation rate.

The figure also shows that there is no correlation between the relative translation rate δ_p and the log mRNA level d_p. We also find no correlation of δ_p with either log protein level y_p, or the variance of the log protein level (data not shown).

Minimal expression noise as a function of mean expression

To model the noise distribution of our promoters we start with the simple case in which there are constant rates of transcription, translation, mRNA decay, and protein decay. Let λ_m be the rate of transcription per unit time, μ_m the rate of mRNA decay (per mRNA per unit time), λ_p the rate of translation (per mRNA per unit time), and μ_p the rate of protein decay (per protein per unit time). Note that in our case all proteins decay at the same rate

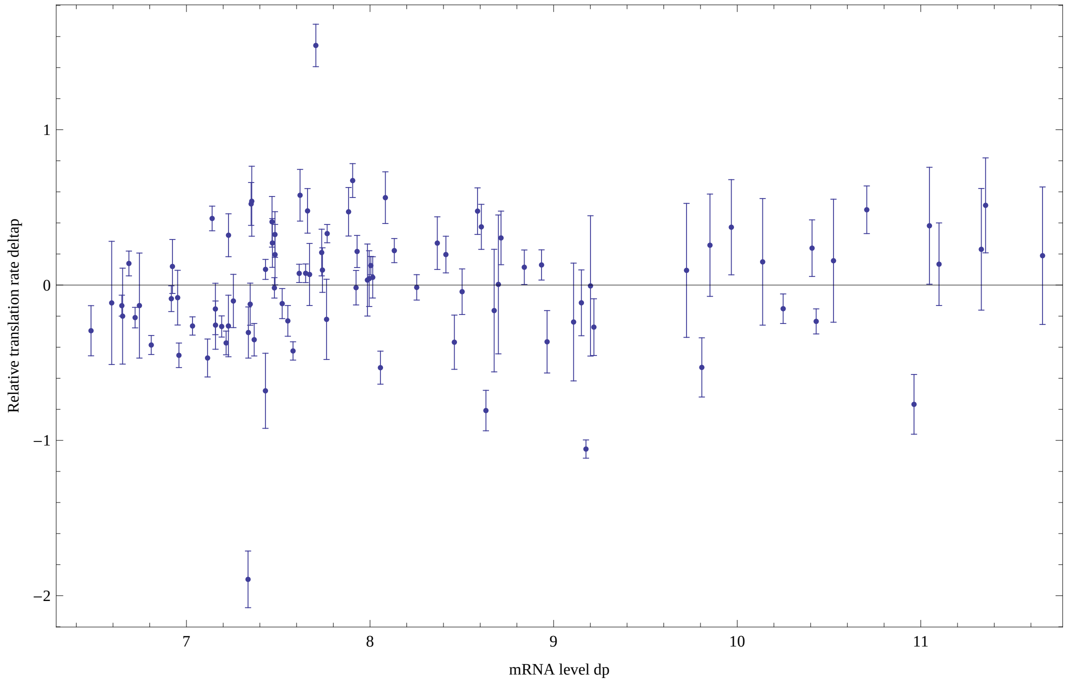

Figure S2.13: Estimated relative log-translation rates δ_p and their error bars $\sigma(\delta_p)$ (vertical axis) as a function of the log-mRNA level relative to *ihfB*, d_P, for each promoter p.

and, because the decay rate of GFP is relatively small compared to the dilution rate as a consequence of cell growth, the rate μ_p is effectively given by the growth rate of the cells.

In (Shahrezaei & Swain, 2008) an analytical expression was derived for the distribution $P(n|\lambda_m, \mu_m, \lambda_p, \mu_p)$ under the assumption that the rate of protein decay is *small* compared the rate of mRNA decay. In *E. coli* the typical mRNA decay rate is on the order of 5 minutes (Bernstein et al., 2002). In the minimal media with glucose in which our cells are grown, the doubling time is more than half an hour, so that the protein decay is indeed smaller than the mRNA decay rate by a factor of approximately 6. Since this is not a very large factor, one may worry that for stable mRNAs the approximation breaks down. Fortunately, in (Shahrezaei & Swain, 2008) it was also shown (by simulation) that as long as the mRNA decay rate is *at least* as large as the protein decay rate, then the approximation still is quite accurate. We will thus assume that we can use this approximation.

Under this approximation the stationary distribution of the number of proteins per cell depends only on the following two ratios:

$$a = \frac{\lambda_m}{\mu_p}, \tag{2.52}$$

and

$$b = \frac{\lambda_p}{\mu_m}. \tag{2.53}$$

The ratio a gives the expected number of transcripts that are produced during the life-time of a single protein, which in our case effectively means the doubling time of the cells, i.e. a is the expected number of transcription events per cell cycle. The ratio b gives the expected number of proteins that are produced from a single mRNA during its life-time. This is sometimes referred to as the 'burst size'. That is, typically one assumes $b > 1$ and given that mRNA decay is faster than protein decay, the proteins are produced 'fast' from a single mRNA in comparison to the life-time of a typical protein, i.e. in a burst.

The limit distribution $P(n|a, b)$ is given by a negative binomial

$$P(n|a,b) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)} \left(\frac{b}{b+1}\right)^n \left(1 - \frac{b}{b+1}\right)^a . \tag{2.54}$$

This distribution has a mean

$$\langle n \rangle = ab, \tag{2.55}$$

and variance

$$\text{var}(n) = ab(1+b) = \langle n \rangle (1+b). \tag{2.56}$$

We extend this simple model by assuming the ratios a and b fluctuate themselves (most likely on a somewhat slower time scale). Although we will not attempt to specify the molecular origins of these fluctuations in a and b, they likely include fluctuations in the concentrations of polymerases, ribosomes, and transcription factors that regulate the promoter in question. Such fluctuations would contribute to the extrinsic noise of the promoters, since they would equally effect two copies of the same promoter in the same cell. However, they may also include fluctuations in the state of the promoter itself and such fluctuations would contribute to the intrinsic noise.

We will assume that the fluctuations in these ratios of rates are multiplicative, i.e. proportional to the means $\langle a \rangle$ and $\langle b \rangle$:

$$\text{var}(a) = \langle a \rangle^2 \sigma_a^2, \tag{2.57}$$

and

$$\text{var}(b) = \langle b \rangle^2 \sigma_b^2. \tag{2.58}$$

We then find for the total variance of n

$$\text{var}(n) = \langle n \rangle^2 (\sigma_a^2 + \sigma_b^2 + \sigma_a^2 \sigma_b^2) + \langle n \rangle \left[1 + \langle b \rangle (1 + \sigma_b^2)\right]. \tag{2.59}$$

To simplify notation, we introduce the variable

$$\sigma_{ab}^2 = \sigma_a^2 + \sigma_b^2 + \sigma_a^2 \sigma_b^2, \tag{2.60}$$

and the renormalized burst-size

$$\beta = 1 + \langle b \rangle (1 + \sigma_b^2). \tag{2.61}$$

With these definitions we have

$$\text{var}(n) = \langle n \rangle^2 \sigma_{ab}^2 + \beta \langle n \rangle, \tag{2.62}$$

which brings out most clearly that there is a term proportional to $\langle n \rangle^2$ that results from fluctuations in a and b, and a term proportional to $\langle n \rangle$ that results from Poisson fluctuations in mRNA and protein production, and is proportional to the burst-size.

We now want to relate this expression to variations in log-fluorescence intensities as measured using FACS. Here it is important to note that the log-fluorescence intensity per cell is the result of a combination of fluorescence coming from GFP proteins and *background* fluorescence of the cell.

Background fluorescence

To estimate background fluorescence, we performed 3 replicate measurements of populations of cells without any plasmid, and 3 replicate measurements of populations of cells containing an empty plasmid (not containing a GFP gene). **Figure S2.14** shows the reverse cumulative distributions of observed intensities in these control populations (colored lines).

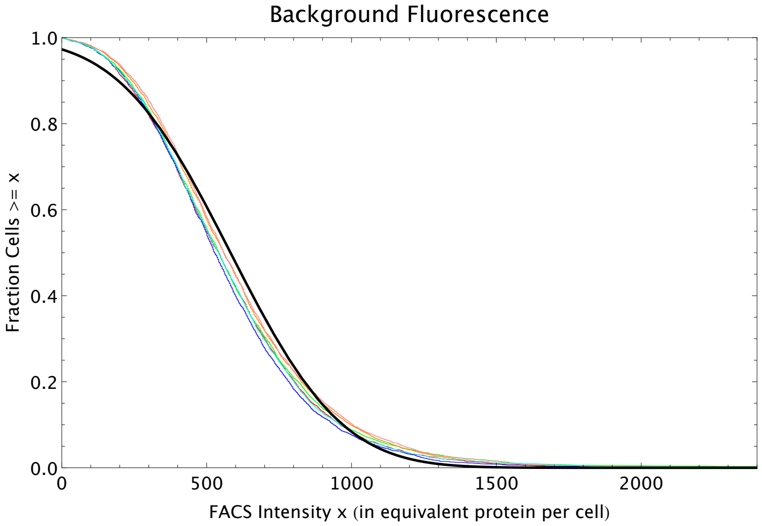

Figure S2.14: Reverse cumulative distributions of the FACS intensities per cell (multiplied by 2.88 so as to correspond to the equivalent of GFP proteins per cell) for MG1655 cells without a plasmid (red, blue and green curves) and MG1655 cells with an empty plasmid (orange, pink and cyan curves). The black line shows a Gaussian distribution with matching mean and variance.

The curves show that each replicate shows a highly similar distribution of fluorescence levels, and pooling the data from all replicates we find a mean background fluorescence of $n_{\text{bg}} = 582.3$ with a standard-deviation of $\sigma_{\text{bg}} = 302.9$. As shown by the black curve in **figure S2.14**, the distribution of background fluorescences is reasonably approximated by a Gaussian with the same mean and standard-deviation.

Relating measured variations to the theoretical expression

Let n_{meas} denote the measured FACS intensity of a cell. We will write this measured intensity as the sum of an average background fluorescence n_{bg}, the average number of proteins $\langle n\rangle$, and a fluctuation of size $\epsilon\sqrt{\text{var}(n)}$:

$$n_{\text{meas}} = n_{\text{bg}} + \langle n\rangle + \epsilon\sqrt{\text{var}(n)}, \tag{2.63}$$

Here ϵ is a quantity that fluctuates from cell to cell, which has mean zero $\langle\epsilon\rangle = 0$, and variance one, i.e. $\langle\epsilon^2\rangle = 1$.

We will assume that the fluctuations $\epsilon\sqrt{\text{var}(n)}$ are small relative to the mean $\langle n_{\text{meas}}\rangle = n_{\text{bg}} + \langle n\rangle$. We can then write for the logarithm of the measured FACS intensity

$$\log[n_{\text{meas}}] \approx \log[\langle n_{\text{meas}}\rangle] + \epsilon\frac{\sqrt{\text{var}(n)}}{\langle n_{\text{meas}}\rangle}. \tag{2.64}$$

We then find for the variance in log-scale of the measured FACS intensities

$$\text{var}\left(\log[n_{\text{meas}}]\right) = \frac{\text{var}(n)}{\langle n_{\text{meas}}\rangle^2}. \tag{2.65}$$

If we substitute the expression (2.62) for the numerator, we obtain

$$\text{var}\left(\log[n_{\text{meas}}]\right) = \sigma^2_{ab}\left(1 - \frac{n_{\text{bg}}}{\langle n_{\text{meas}}\rangle}\right)^2 + \frac{\beta}{\langle n_{\text{meas}}\rangle}\left(1 - \frac{n_{\text{bg}}}{\langle n_{\text{meas}}\rangle}\right). \tag{2.66}$$

The left panel of **figure S2.15** shows the mean and variances of the log-FACS intensities of all native promoters. This scatter shows that, as a function of the mean FACS intensity, there is a sharp lower bound on the observed variances. The red curve shows that this lower bound can be well-fitted by a function of the form (2.66), where we used parameters $\sigma^2_{ab} = 0.025$ and $\beta = 450$. Note that the value of σ^2_{ab} determines the variance in the limit of large means, whereas β controls the curvature at lower means. We fitted these two parameters by hand. Their interpretation is that, σ^2_{ab} corresponds to the minimal amount of cell-to-cell variation in the product ab that is possible for any promoter architecture. The variable $\beta = 450$ roughly corresponds to the burst-size.

Note that the log-fluorescence on the horizontal axis corresponds to the sum of fluorescence resulting from GFP molecules and the background fluorescence. The estimated background level $n_{\text{bg}} = 582.3$ corresponds to a log-fluorescence of 6.37. The region on the horizontal axis between 6.37 and $7 \approx \log(2 * 582.3)$, thus corresponds to cells where the fluorescence due to GFP molecules is less than the background fluorescence. In this regime the noise distribution results from a combination of fluctuations in background fluorescence and in protein numbers (which may be correlated because part of these fluctuations may result from fluctuations in cell size) and our noise model (2.66) breaks down. In the following we will focus on those

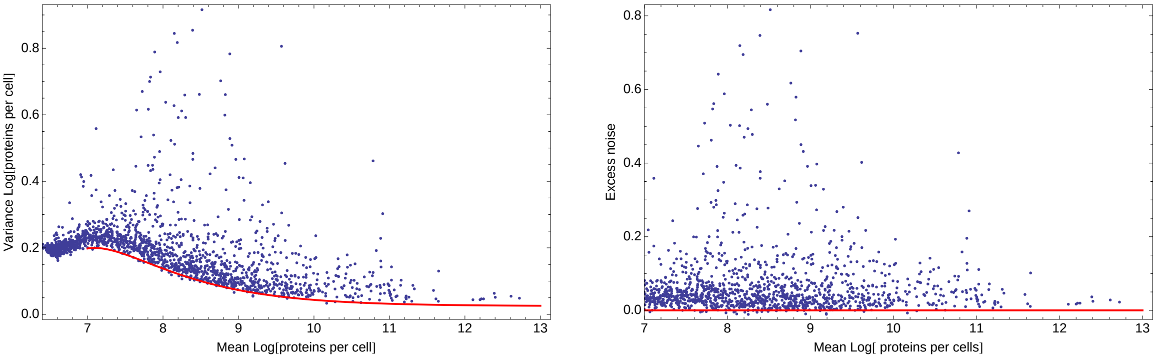

Figure S2.15: Dependence between mean and variance of log FACS intensities. **Left panel:** Means and variances of log-FACS intensities of all native promoters (blue dots) together with a fitted lower bound on the variance as given by equation (2.66) using $\sigma^2_{ab} = 0.025$ and $\beta = 450$ (red curve). **Right panel:** Excess noise (obtained by subtracting the fitted lower bound from the variance) as a function of mean log-FACS intensity for all native promoters (blue dots). The red line shows the x-axis.

promoters with fluorescence due to GFP at least as large as the background fluorescence, i.e. with mean log-fluorescence larger than 7.

To obtain a deviation of each promoter's variance from the minimal variance that is possible at its expression level, we define the *excess noise* η as the difference between a promoter's variance and its the fitted minimal variance $\sigma^2_{\text{min}}(\mu)$ as given by equation (2.66) with $\beta = 450$ and $\sigma^2_{ab} = 0.025$:

$$\eta = \sigma^2 - \sigma^2_{\text{min}}(\mu). \tag{2.67}$$

The right panel of **figure S2.15** shows the excess noise levels of all native promoters as a function of their means. The figure shows that, with this correction, there is no longer any systematic dependence between mean expression levels and noise. Therefore, we can use excess noise as a measure of transcriptional noise that allows us to compare noise levels of promoters with different mean expression levels.

FACS Selection

As explained in the Materials and Methods, for both the medium and high expression evolutionary runs, the desired expression level μ_* is taken from the expression level of a reference promoter from the library of E. coli promoters. At each selection round we measure the expression μ_* of the reference promoter, and set the center of the FACS's selection window to μ_*. We then set the width of the selection window such that 5% of the cells have expression levels within the selection window.

Although, in principle, the FACS's selection should work such that a cell with expression level anywhere within the selection window has 100% probability to be selected, and 0% probability to be selected if the cell's expression is anywhere outside the selection window, it is unrealistic to assume that the boundaries of the selection window are so precisely defined in practice. As illustrated below, comparison of the population's expression levels before and after selection shows that the probability for a cell with log-fluorescence x to be selected can be well-approximated as

$$f(x|\mu_*, \tau) = \exp\left[-\frac{(x-\mu_*)^2}{2\tau^2}\right], \tag{2.68}$$

where μ_* is the desired expression level and τ corresponds to the width of the selection window.

Note that, for a promoter with mean expression μ and variance σ^2, the fraction $P(x|\mu,\sigma)$ of its cells that have expression level x is given by

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}]\right]. \tag{2.69}$$

Consequently, the 'fitness' of this promoter, i.e. the fraction of its cells that are selected in the FACS, is given by

$$f(\mu,\sigma|\mu_*,\tau) = \int dx f(x|\mu_*,\tau) P(x|\mu,\sigma) = \sqrt{\frac{\tau^2}{\tau^2+\sigma^2}} \exp\left[-\frac{(\mu-\mu_*)^2}{2(\tau^2+\sigma^2)}]\right]. \tag{2.70}$$

To infer the values of μ_* and τ that apply to our evolutionary runs, we performed a number of experiments in which we:

1. Took a population from one of the rounds of our evolutionary runs.
2. Measured its distribution of log-fluorescence levels.
3. Set the selection window $[\mu_* - \delta, \mu_* + \delta]$ such that a percentage p of the population has log-fluorescence levels within this selection window.
4. Performed selection and re-measured the log-fluorescence levels of the selected population.

As shown in Fig. 2.1 of the main paper, in the evolutionary runs which are selecting for high expression, the selection window is changing at every round of the evolutionary run. In contrast, in the evolutionary runs selecting for medium expression, the selection window is essentially constant from round 3 through round 5 of the run. We thus decided to focus on inferring the precise fitness function that acted during these 3 rounds of selection.

We took the evolved populations from the third and fifth round of the evolutionary runs selecting for medium expression, and performed another round of selection on them, selecting 5% of the population closest to the desired log-fluorescence μ_*. In addition, we also performed a round of less stringent selection on these populations, selecting 25% closest to the desired level, and a round of more stringent selection, selecting only the 1% of the population closest to the desired level. Besides measuring the log-fluorescence levels of the population both before and after the round of selection, we also selected dozens of clones from the populations before and after the selection, and measured the entire distribution of log-fluorescence levels for these clones. **Figure S2.16** shows the means and variances of the log-fluorescence distributions of these clones.

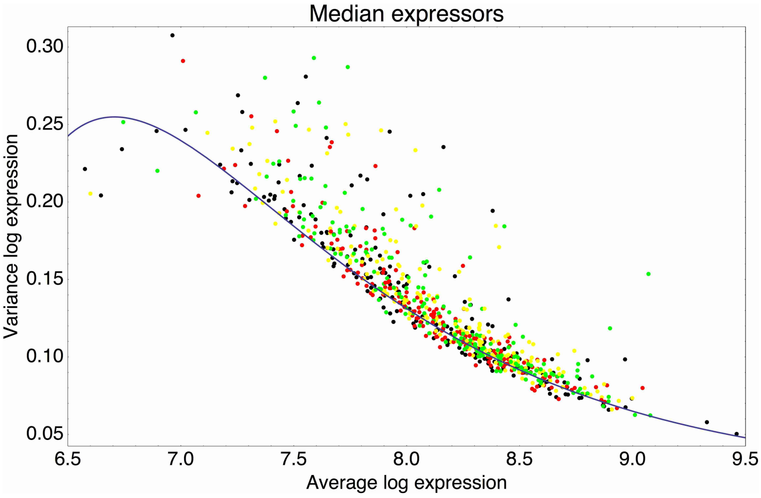

Figure S2.16: Means and variances of the log-fluorescence levels of clones from the third and fifth rounds of the evolutionary runs in which we selected for medium expression (black dots), and clones obtained after performing another round of selection on these populations, selecting either 1% (red), 5% (yellow), or 25% of the population closest to the desired log-fluorescence μ_*. The blue curve shows an approximate fit of the typical variance σ^2 as a function of the mean μ: $\sigma^2(\mu) = 0.02 + 384e^{-\mu} - 156'915e^{-2\mu}$.

Intuitively, one might think that the relative fitness $f(x)$ of each log-fluorescence level x could be easily estimated by simply measuring the ratio of the fraction of the population $p'(x)$ with log-fluorescence level x after selection and the fraction $p(x)$ with log-fluorescence x before selection. However, the single cells that were selected in the FACS each grow into an entire population of cells before the 'after selection' population is measured again. Thus, a selected cell containing a promoter with a given mean μ and variance σ^2 will contribute an entire population of cells with this distribution, even though the individual cell may itself have had a log-fluorescence that was in one of the tails of this distribution. Thus, in general the distribution of log-fluorescence levels in the population after selection may be much wider than the actual selection window itself.

Before selection, the population consisted of a mixture containing (unknown) fractions $\rho(\mu, \sigma)$ of cells containing promoters with mean μ and variance σ^2. This gives rise to an overall distribution $p(x)$ of expression levels given by

$$p(x) = \int d\mu d\sigma \frac{\rho(\mu, \sigma)}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{2.71}$$

Unfortunately we cannot uniquely infer $\rho(\mu, \sigma)$ from knowing only the distribution $p(x)$. However, as shown in **Fig. S2.16**, for the clones in these populations, the large majority of promoters have variances σ^2 lying in a narrow band as a function of μ. We thus chose to make the approximation that *all* promoters in the population have variances σ^2 that are uniquely determined by their mean expression μ, and we used the fit $\sigma^2(\mu)$ shown as the blue curve in **Fig. S2.16**. This simplifies the problem from inferring a two-dimensional distribution $\rho(\mu, \sigma)$ to inferring a one-dimensional distribution $\rho(\mu)$, i.e.

$$p(x) = \int d\mu \frac{\rho(\mu)}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2(\mu)}\right]. \tag{2.72}$$

Applying selection with target log-fluorescence μ_* and width τ to this distribution, we obtain a new distribution

$$\rho'(\mu) = C\rho(\mu)\sqrt{\frac{\tau^2}{\tau^2+\sigma^2(\mu)}} \exp\left[-\frac{(\mu-\mu_*)^2}{2(\tau^2+\sigma(\mu)^2)}\right], \tag{2.73}$$

where C is a normalization constant that ensures $\int d\mu \rho'(\mu) = 1$. The population distribution of log-fluorescence levels after selection is then

$$p'(x) = \int d\mu \frac{\rho'(\mu)}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2(\mu)}\right]. \tag{2.74}$$

To infer the parameters of the fitness-function for each selection that we performed, we fit the distribution $\rho(\mu)$, and parameters μ_* and τ that lead to an optimal fit to the observed distributions $p(x)$ and $p'(x)$. **Figure S2.17** shows the inferred and observed distributions, as well as the inferred fitness function, for each of the 6 selection experiments that we performed.

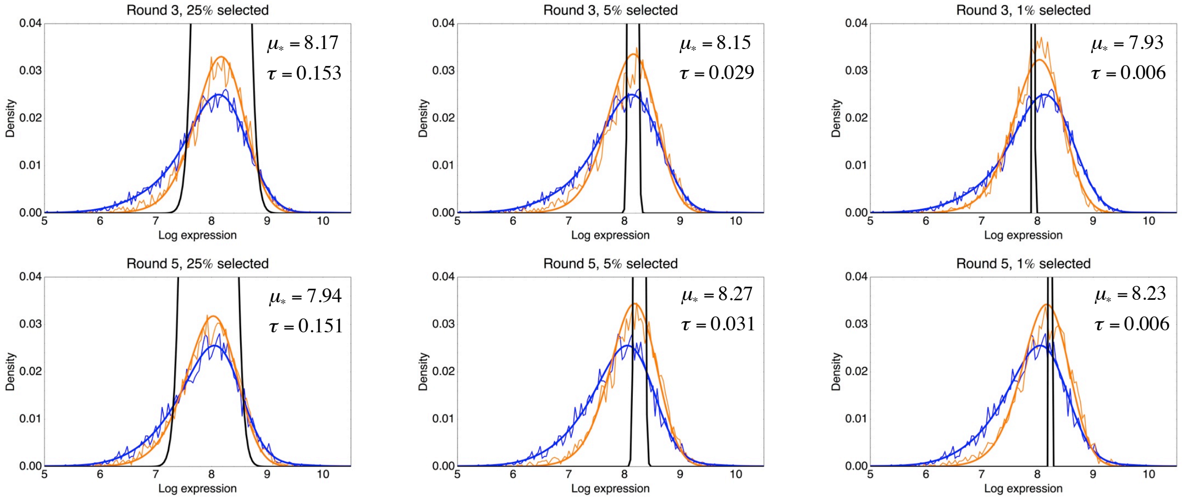

Figure S2.17: Inference of the fitness function from the observed log-fluorescence distribution before and after a round of selection. Each panel corresponds to one selection experiment with the title indicating on which population an extra round of selection was performed, i.e. a population either from the third or fifth round of the evolutionary run for medium expression. The thin blue line indicates the observed log-fluorescence distribution $p(x)$ before selection, and the thin orange line the observed distribution $p'(x)$ after selection. The thick lines show the corresponding fitted distributions. The inferred selection window $f(x|\mu_*, \tau)$, i.e. equation (2.68), is indicated in black, and its parameters μ_* and τ are indicated in each panel as well.

The figure shows that the distributions $p(x)$ and $p'(x)$ can be well fit by this model, illustrating that the form of the selection window, equation (2.68), can well describe the effects of selection in the FACS machine. Moreover, we see that the distributions $p(x)$ and $p'(x)$ are typically significantly wider than the selection window $f(x|\mu_*, \tau)$. Moreover, the fitted values of τ are almost perfectly proportional to the fraction of the population that was selected, with a value of $\tau \approx 0.03$ corresponding to the selection of 5% of the population that was used during the evolutionary runs. The fits also show that, although we in each experiment determine μ_* from the expression of the same reference promoter, there is some variability in μ_* from one experiment to the next. From these 6 experiments, we find that on average $\langle \mu_* \rangle = 8.115$ and $\sigma(\mu_*) = 0.133$.

Thus, in each selection round the fitness of a promoter with mean μ and variance σ^2 is given by expression (2.70), where $\tau = 0.03$ and μ_* fluctuates around $\langle \mu_* \rangle = 8.115$. The effective fitness experienced by this promoter is thus given by the *geometric* average of equation (2.70) with fluctuating values of μ_* and this is given by

$$f(\mu, \sigma | \langle \mu_* \rangle, \sigma(\mu_*), \tau) = \sqrt{\frac{\tau^2}{\tau^2 + \sigma^2}} \exp\left[-\frac{(\mu - \langle \mu_* \rangle)^2 + \sigma(\mu_*)^2}{2(\tau^2 + \sigma^2)} \right]. \tag{2.75}$$

Figure S2.18 shows a contour plot of this fitness function with the inferred parameters of $\langle \mu_* \rangle$, $\sigma(\mu_*)$ and τ as a function of the mean fitness μ, and the excess noise level $\eta = \sigma^2 - \sigma^2_{\text{min}}(\mu)$, where $\sigma^2_{\text{min}}(\mu)$ is the minimal variance as a function of mean expression level μ, equation (2.66). Note that for these measurements the plasmids were transformed into a different strain than those used to compare with the native *E. coli* promoters. We noticed that the minimal noise level $\sigma^2_{\text{min}}(\mu)$ as a function of mean μ is slightly different. Although the background fluorescence and burst-size parameter $\beta = 450$ are the same, the parameter σ^2_{ab} is smaller, i.e. $\sigma^2_{ab} = 0.006$ instead of $\sigma^2_{ab} = 0.025$.

Figure S2.18 clearly shows that fitness drops far more dramatically as a function of mean μ than as a function of excess variance η, i.e. except for right at the optimal mean μ_*, the contours are running almost vertically in the plot. We do not observe promoters with high excess noise, even though their fitness would easily allow it. For example, a promoter with mean expression near the optimum 8.11 but excess noise as high as 0.25 (i.e. significantly higher than observed for any of the clones) would have higher fitness than any promoter with mean less than $\mu = 7.76$ or larger than $\mu = 8.47$ (independent of their noise), even though we observe many promoters with means that deviate this far from the optimum. The fact that we do not observe high excess noise promoters, even though they would not be selected against, strongly suggests that such high noise promoters are uncommon *a priori*, i.e. among all random sequences that drive expression at a medium level, the large majority have low excess noise levels and high noise promoters are rare. Moreover, note also that for promoters that are not near the optimum, the optimal excess noise level is typically *larger* than those of the observed clones, e.g. the optimal excess noise for a promoter with mean $\mu = 7.5$ is $\eta = 0.22$. These observations all suggest that promoters have not experienced significant selection on their noise levels.

To further support this conclusion, **Figure S2.19** shows the inferred fitness values for the observed clones both as a function of their mean expression (left panel) and as a function of their excess noise (right) panel. The figure shows that, whereas the fitness of a clone can be

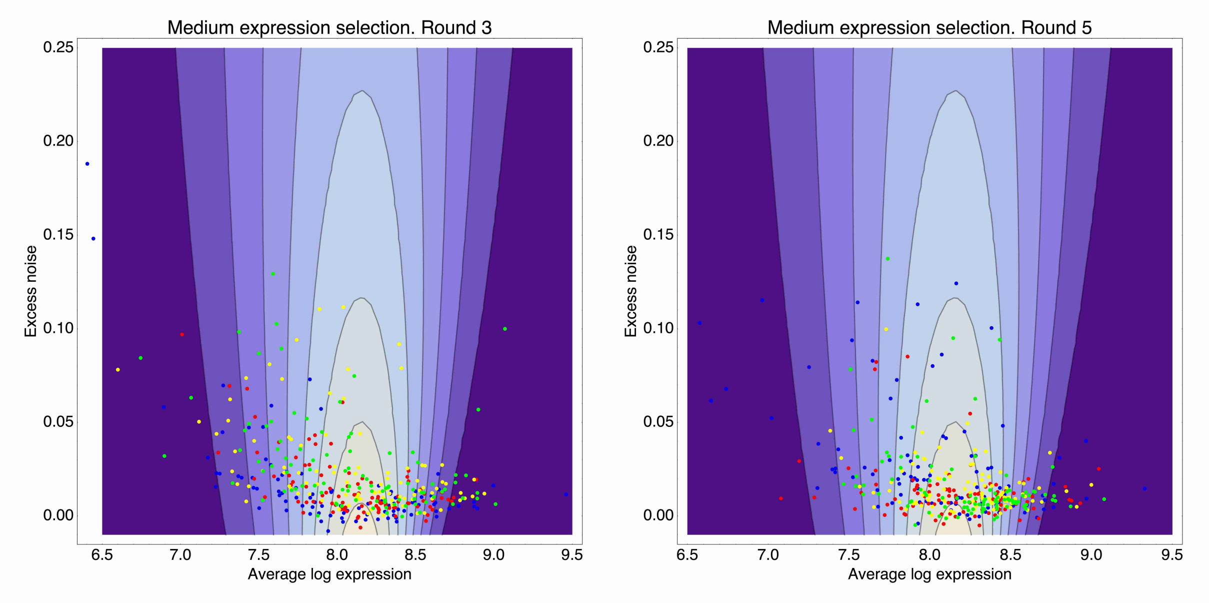

Figure S2.18: Contour plot of the inferred fitness function (2.75) as a function of mean expression μ (horizontal axis) and excess noise (vertical axis), that acts on the population from rounds 3 through 5 of the evolutionary runs for medium expression. The contours correspond to fitness values (fraction of cells selected) of 0.01, 0.02, 0.03, through 0.08. **Left panel**: In addition to the fitness function (contours) the panel shows the means and excess noise levels of a selection of clones from the third round of the evolutionary run (blue dots), and clones that resulted from subjecting this population to another round of selection, selecting either for the 1% (red dots), 5% (yellow dots), or 25% (green dots) of cells with expression closest to the desired expression level. **Right panel**: As in the left panel, but with the dots corresponding to clones from the 5th round of the evolutionary run, and clones resulting from additional rounds of selection on this population (colors as in the left panel).

accurately predicted from its mean μ, fitness is almost entirely uncorrelated to a promoter's excess noise η.

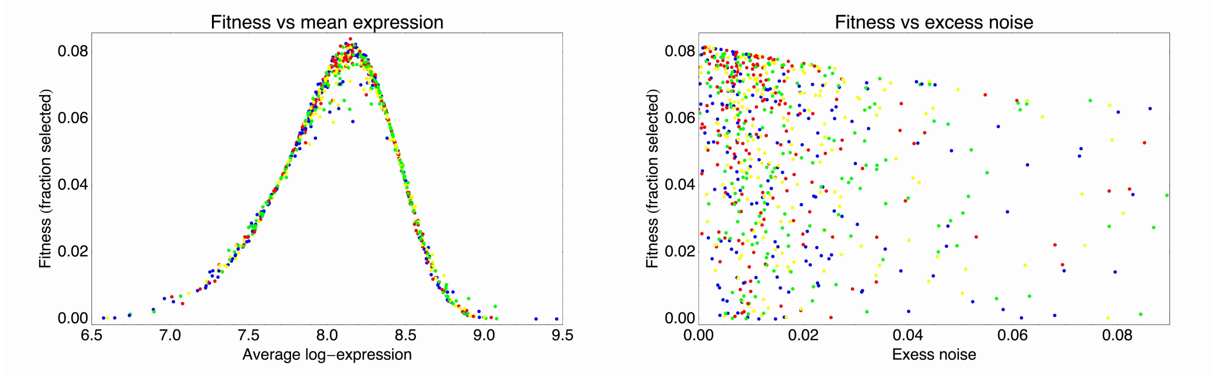

Figure S2.19: Fitness of the observed clones, as given by equation (2.75), as a function of their mean expression μ (left panel) and their excess noise η (right panel). As in figure S2.18, the blue dots correspond to clones from the third and fifth round of the evolutionary run, the red dots result from another round of stringent selection (top 1%), the yellow dots from another round of standard selection (top 5%), and the green dots from a round of weaker selection (top 25%).

Finally, if there was significant selection on noise levels, then we expect noise levels to systematically shift under selection. **Figure S2.20** shows cumulative distributions of excess noise levels for clones obtained from different populations of cells.

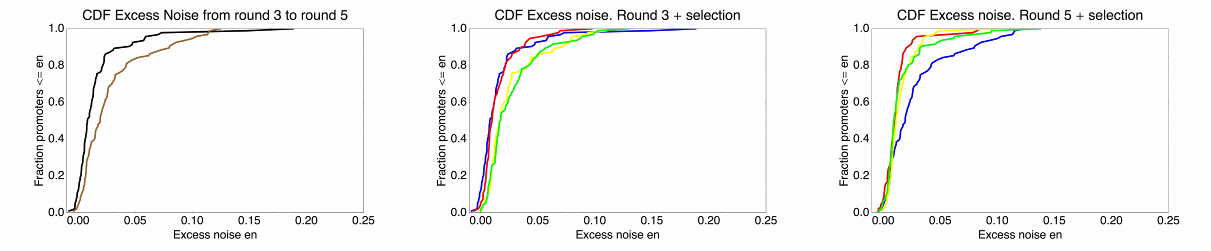

Figure S2.20: Cumulative distribution functions of excess noise levels for the promoters extracted from different populations. **Left panel**: Excess noise levels of promoters from the 3rd (black) and 5th (brown) round of the evolutionary run. **Middle panel**: Excess noise levels of promoters from the 3rd round of the evolutionary run (blue), and from clones that resulted from another round of either stringent (red), normal (yellow), or weak (green) selection. **Right panel:** As in the middle panel but now for clones from the 5th round and clones resulting from another round of selection on this population.

The figure shows that, surprisingly, the excess noise levels seem to increase from the third to the fifth round in the evolutionary run. However, given the limited number of clones involved, the change in excess noise levels is only marginally significant ($p = 0.004$ in a t-test). Similarly, the effect of selection on excess noise levels seems to be opposite on the populations of round 3 and round 5 (center and right panels in **Fig. S2.20**). We suspect that there are some systematic experimental fluctuations that make measured excess noise levels vary across days, and that the observed distributions of excess noise levels are more a reflection of experimental variability than of true shifts in the distribution. Importantly, excess noise levels larger than 0.1, which are observed for a substantial fraction of native promoters, are very rare for all these populations.

In summary, our in depth analysis of the FACS fitness function and the effects of selection show that the noise properties of the synthetic promoters have not been significantly shaped by selection. Already the synthetic promoters at the third round of selection are tightly concentrated in a low noise band, even though selection does not select against low noise, and promoters with mean away from the optimum would benefit from having higher noise. Two additional rounds of mutation and selection on the promoters from the third round do not substantially change the distribution of noise levels confirming that there are no substantial fitness differences among promoters with different noise. Similarly, performing additional rounds of selection (be it very stringent, normal, or lenient) also do not substantially change the observed noise levels of the selected promoters. Thus, our results show that promoters selected from a large collection of random sequences naturally display low noise levels. Importantly, this implies that the native promoters with substantially higher noise levels must have experienced some selective pressures that caused them to increase their noise.

A simple model for the evolution of gene regulation and expression noise

Given a particular environment, the fitness, e.g. growth-rate or survival probability of a cell, depends on the expression level of its genes. Note that the fact that gene regulatory mechanisms have evolved already demonstrates that different environments require different gene expression patterns, i.e. expressing a gene at the 'wrong' level for a given environment has negative effects on fitness/growth-rate. For simplicity, we will focus on a single gene. We assume that, in a given environment, there is an optimal expression μ_* level. Given that, as we have seen, expression levels are roughly log-normally distributed, we will express expression levels in log-space, i.e. the logarithm of the number of proteins per cell. We define that fitness at the optimal expression level μ_* as f_o. Fitness will fall as the expression level x moves away from this optimum. In this simple conceptual model, we will assume that, like in our FACS selection, the fitness $f(x)$ falls approximately Gaussian away from the optimum, i.e.

$$f(x|\mu_*, \tau) = f_o \exp\left[-\frac{1}{2}\left(\frac{x-\mu_*}{\tau}\right)^2\right], \tag{2.76}$$

where τ is again a parameter that determines how fast the fitness falls when the expression x moves away from the optimum μ_*.

To justify the Gaussian form of the fitness function, assume that the fitness is determined by the growth of the population over some characteristic time t. That is, if cells grow at rate ρ, then the fitness is $f = e^{\rho t}$. The growth rate ρ is optimal at $x = \mu_*$, and to second order in the difference between x and μ_*, we can write

$$\rho = \rho_o - \frac{1}{2}(x-\mu_*)^2 \left\| \frac{d^2\rho}{dx^2} \right\|_{x=\mu_*}. \tag{2.77}$$

Defining $f_o = e^{\rho t}$ and $1/\tau^2 = t \left\| \frac{d^2\rho}{dx^2} \right\|_{x=\mu_*}$, we obtain the fitness function defined above.

In complete analogy with the FACS selection case, the fitness $f(\mu,\sigma|\mu_*,\tau)$ of a promoter with mean μ and variance σ^2 in an environment with optimal expression level μ_* and width τ is given by the integral of the product of the fitness function (2.76) and the Gaussian distribution of expression levels, giving

$$f(\mu,\sigma|\mu_*,\tau) = \sqrt{\frac{\tau^2}{\tau^2+\sigma^2}} \exp\left[-\frac{1}{2}\frac{(\mu-\mu_*)^2}{\tau^2+\sigma^2}\right]. \tag{2.78}$$

Note that this functional form is a reasonable approximation to the fitness function as long as expression levels are roughly log-normally distributed, and as long as the integral of expression levels and fitness function can be approximated using the standard Laplace approximation, i.e. expanding the logarithm of fitness to second order around its maximum.

We now extent this simple situation in two respects. First, instead of assuming that the optimal level μ_* is fixed, we imagine that the population of cells has gone through several different 'environments', where in each environment e there was an optimal expression level μ_e. For simplicity we assume τ is the same in each environment.

Let's first consider what this situation implies for the fitness of a promoter expressing at mean level μ with variance σ^2. The number of offspring that a strain with mean μ and variance σ^2 produces (or leaves behind) after experiencing environment e is proportional to $f(\mu,\sigma|\mu_e,\tau)$. Consequently, the final number of offspring produced after experiencing all environments is given by the product $\prod_e f(\mu,\sigma|\mu_e,\tau)$. We define the overall log-fitness $\log[f(\mu,\sigma)]$ as the average of the log-fitness across all environments:

$$\log[f(\mu,\sigma)] = \langle\log[f(\mu,\sigma|\mu_e,\tau)\rangle_e, \tag{2.79}$$

where the subscript e indicates that we are averaging over all environments e (which we drop for convenience from here on). Using the expression (2.78) we obtain

$$\log[f(\mu,\sigma)] = -\frac{\mathrm{var}(\mu_e)+(\langle\mu_e\rangle-\mu)^2}{2(\sigma^2+\tau^2)} + \frac{1}{2}\log\left[\frac{\tau^2}{\sigma^2+\tau^2}\right], \tag{2.80}$$

where $\langle\mu_e\rangle$ is the average 'desired' expression level, and $\mathrm{var}(\mu_e)$ is the variance in desired expression levels across the environments. It is immediately clear from equation (2.80) that, as a function of the mean expression μ, optimal fitness is obtained when $\mu = \langle\mu_e\rangle$. Substituting this optimal mean level, we find that optimal variance is given by

$$\sigma^2 = \mathrm{var}(\mu_e) - \tau^2, \tag{2.81}$$

when $\mathrm{var}(\mu_e) \geq \tau^2$, and $\sigma = 0$ otherwise. That is, when the variance in desired expression levels is larger than the width of the selection window τ, then a strain can increase its fitness by raising the noise-level σ of the promoter. This result is equivalent to results on selection for phenotypic variance obtained previously, e.g. (Bull, 1987, Haccou & Iwasa, 1995). However, in these previous models that more abstractly considered 'phenotypic traits', it was assumed that both the mean and variance of the phenotypic trait were not only directly encoded by the genotype, but could also be independently altered through mutations, without explicitly considering how mean and variance would be encoded in the genotype. In our case, where the ‘trait’ under study is the transcription rate of a promoter, it is a priori quite clear how mutations may alter mean levels, e.g. through changes in the affinity of the sigma-factor binding site, but much less clear how the variance is encoded in the genotype. Moreover, rather than simply increasing its noise, we would naturally expect that promoters would evolve *gene regulation* in order to deal with different required expression levels across different environments.

Including gene regulation

We now further extend the model by considering that there are various transcriptional regulators in the cell whose activities may vary across the different environments e. By evolving binding sites for a transcription factor, the promoter becomes regulated by it and, consequently, the mean expression μ becomes a function $\mu(e)$ of the environment e.

For simplicity we consider the case of a single regulator whose mean activity (i.e. concentration of the DNA-binding version of the regulator) r_e is a function of the environment e. Since the transcriptional regulator's expression will itself also be subject to gene expression noise, the activity of the regulator varies from cell to cell. We will assume that, in each environment, the activity of the regulator varies from cell-to-cell in a roughly Gaussian manner, with variance σ_r^2, i.e. the probability to find a cell in environment e with regulator level r is

$$P(r|r_e,\sigma_r^2) = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left[-\frac{(r-r_e)^2}{2\sigma_r^2}\right]. \tag{2.82}$$

We characterize the regulation of the promoter by the regulator through a single *coupling constant* c such that, in cells with regulator level r, the distribution of expression levels is a Gaussian with mean $\mu + cr$ and variance σ^2, i.e.

$$P(x|\mu,\sigma^2,r) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu-cr)^2}{2\sigma^2}\right]. \tag{2.83}$$

Integrating over the distribution of regulator levels $P(r|r_e,\sigma_r^2)$, the final distribution of expression levels is given by a Gaussian with mean $\mu(e) = \mu + cr_e$ and variance $\sigma^2 + c^2\sigma_r^2$, i.e.

$$P(x|\mu,\sigma^2,c) = \frac{1}{\sqrt{2\pi(\sigma^2+c^2\sigma_r^2)}} \exp\left[-\frac{(x-\mu-cr_e)^2}{2(\sigma^2+c^2\sigma_r^2)}\right]. \tag{2.84}$$

In environment e, with desired level μ_e, the fitness of a promoter with coupling c is then given by

$$f(\mu,\sigma,c|\mu_e,\tau) = \sqrt{\frac{\tau^2}{\sigma^2+c^2\sigma_r^2+\tau^2}} \exp\left[-\frac{(\mu+cr_e-\mu_e)^2}{2(\sigma^2+c^2\sigma_r^2+\tau^2)}\right]. \tag{2.85}$$

The log-fitness, averaged over all environments e, is given by

$$\log[f(\mu,\sigma,c)] = -\frac{1}{2}\frac{\langle(\mu+cr_e-\mu_e)^2\rangle}{\tau^2+\sigma^2+c^2\sigma_r^2} + \frac{1}{2}\log\left[\frac{\tau^2}{\tau^2+\sigma^2+c^2\sigma_r^2}\right]. \tag{2.86}$$

It is again easy to see that, with respect to the mean expression μ, fitness is optimized when $\mu = \langle\mu_e\rangle - c\langle r_e\rangle$. In the following we will assume that the mean expression μ matches this optimal value.

We can rewrite the expression for the average log-fitness in a simpler form by introducing the following set of effective parameters. First, the variable

$$Y^2 = \frac{\mathrm{var}(\mu_e)}{(\tau^2+\sigma^2)}, \tag{2.87}$$

measures the variance in desired expression levels μ_e relative to the sum of the variances associated with the width of selection τ^2 and the noise of the unregulated promoter σ^2. The variable Y quantifies the 'expression mismatch' between the promoters average expression μ and the (varying) desired expression levels μ_e. The second effective parameter

$$X^2 = \frac{c^2\sigma_r^2}{(\tau^2+\sigma^2)} \tag{2.88}$$

measures the strength of the regulator coupling constant c. More precisely, it quantifies the contribution $c^2\sigma_r^2$ to the promoter's variance in gene expression, again relative to $(\sigma^2+\tau^2)$. We will refer to X as the coupling constant. Third, the parameter

$$S^2 = \frac{\mathrm{var}(r_e)}{\sigma_r^2} \tag{2.89}$$

measures the 'signal-to-noise' ratio of the regulator, i.e. the variance $\mathrm{var}(r_e)$ of its mean level across conditions, relative to its variance σ_r^2 within each condition. A regulator with large S varies a lot in activity across environments and has relatively little noise in each, whereas a regulator with small S varies little across environments relative to its noise level. Finally, we have the correlation R between the desired expression levels μ_e and the regulator's activities r_e, i.e.

$$R = \frac{\langle\mu_e r_e\rangle - \langle\mu_e\rangle\langle r_e\rangle}{\sqrt{\mathrm{var}(r_e)\mathrm{var}(\mu_e)}}. \tag{2.90}$$

In terms of these parameters we have for the average log-fitness

$$\log[f(X,Y,S,R)] = -\frac{1}{2}\frac{Y^2(1-R^2)+(SX-RY)^2}{1+X^2} - \frac{1}{2}\log[1+X^2] + \frac{1}{2}\log[\frac{\tau^2}{\tau^2+\sigma^2}]. \tag{2.91}$$

The last term is a constant that does not depend on our effective parameters and we will ignore it from now on.

We intuitively expect that the promoter's fitness would benefit most from coupling to a regulator that is perfectly correlated with the environment's requirements, i.e. at $R = 1$. Indeed we find that the derivative $\log[f(X)]$ with respect to R is given by

$$\frac{d\log[f(X,Y,S,R)]}{dR} = \frac{SXY}{1+X^2}, \tag{2.92}$$

which is positive as long as the desired levels vary ($Y > 0$), the regulator has some variation across environments ($S > 0$) and there is positive coupling ($X > 0$). Thus, in general, if we keep all other variables fixed, an increase in the regulator's correlation R is always beneficial.

We now consider the case in which a regulator with a given correlation R and signal-to-noise rate S is given, and we want to determine the optimal coupling X_* that maximizes $\log[f(X,Y,S,R)]$ as a function of the expression mismatch Y. The derivative of $\log[f(X,Y,S,R)]$ with respect to X is given by

$$\frac{d\log[f(X,Y,S,R)]}{dX} = \frac{XY^2 - X(1+X^2+S^2) + SR(1-X^2)}{(1+X^2)^2}. \tag{2.93}$$

At $X = 0$, this derivative equals SR. Thus, whenever $R > 0$, the derivative is positive at $X = 0$. Because, as can be easily seen from equation (2.93), the derivative is guaranteed to be negative for large X, this implies that, whenever $R > 0$, there is an optimal coupling X_* that is positive, i.e. $X_* > 0$. Thus, whenever $R > 0$, the promoter is guaranteed to increase its fitness by evolving a nonzero coupling to the regulator.

The optimal coupling X_* is given by the positive solution of the third order polynomial in the numerator of (2.93). In general we find that, when Y is small, the optimal coupling is given by

$$X_* = \alpha_0 Y = \frac{SR}{1+S^2}Y, \tag{2.94}$$

and that when Y is very large X_* obeys

$$X_* = \alpha_\infty Y = \left(\sqrt{\left(\frac{SR}{2}\right)^2 + 1} - \frac{SR}{2}\right)Y. \tag{2.95}$$

That is, both for very small and very large Y, the optimal coupling X_* is directly proportional to Y, with proportionality constants α_0 and α_∞, respectively. Moreover, $\alpha_\infty \geq \alpha_0$. The behavior of X_* as a function of Y, for different values of R and S is illustrated in **Fig. S2.21**.

The figure shows that, as Y increases the ratio X_*/Y switches from the lower α_0 to the higher α_∞. Whenever both the correlation R and the signal-to-noise S are high (the orange and red curves in the top two panels), there is only a small difference between α_0 and α_∞. That is, X_* increases roughly linearly with Y when there is a well-correlated regulator with high signal-to-noise.

In contrast, when the correlation R is low or the regulator is noisy, there is a large difference between α_∞ and α_0. Moreover, the optimal coupling shows a sharp transition

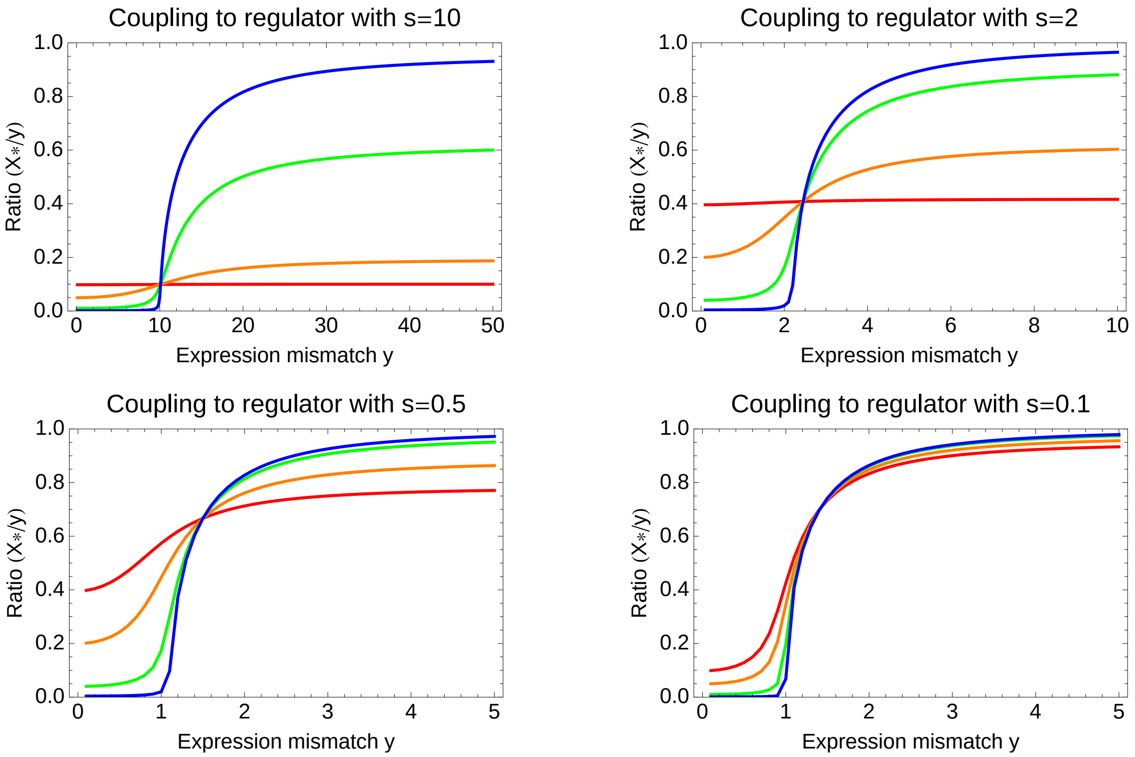

Figure S2.21: The ratio X_*/Y between optimal coupling X_* and expression mismatch Y as a function of Y,for different values of the regulator's signal to noise ratio S and the correlation between regulator and environment R. Each panel corresponds to a different signal to noise ratio S, from a high signal regulator in the top left, to a noisy regulator at the bottom right. In each panel, the different colored lines correspond to different correlations R, i.e. $R = 0.01$ (blue), $R = 0.1$ (green), $R = 0.5$ (orange), and $R = 0.99$ (red).

from low values to much higher values at a 'critical' value of Y. This critical value of Y occurs at $Y = \sqrt{1+S^2}$ when R is low (the blue and green curves), and slightly earlier when R increases (orange and red curves). When the regulator is very noisy (bottom right panel of **Fig. S2.21**) the behavior of X_* becomes almost independent of the correlation R, showing a sharp transition from almost no coupling to strong coupling when $Y \approx 1$. This behavior even extends to the case where there is no correlation whatsoever between the regulator and the environment, i.e. $R = 0$.

Coupling to an uncorrelated regulator

When $R = 0$ the optimal coupling is given by

$$X_*^2 = \max\left[0, Y^2 - 1 - S^2\right]. \tag{2.96}$$

That is, at the critical value $Y = \sqrt{1+S^2}$ the coupling goes from zero to a positive value. For large Y the optimal coupling is simply Y. The behavior of optimal coupling X_* as a function of Y is shown in **figure S2.22**.

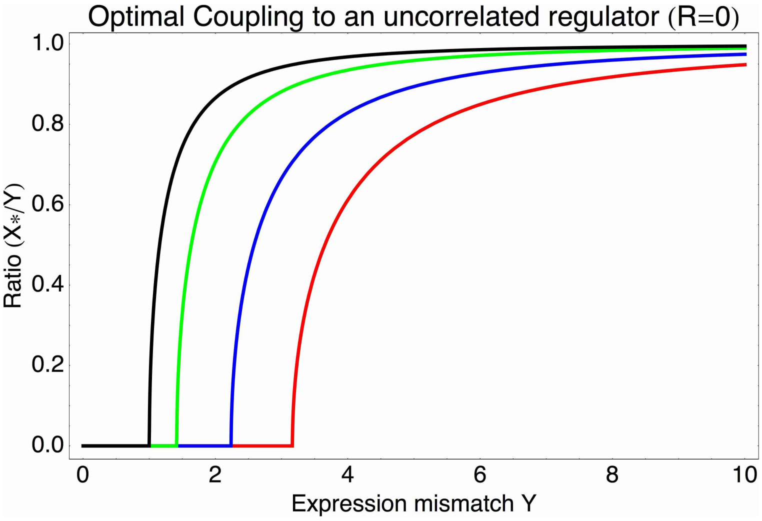

Figure S2.22: Optimal coupling X_* as a function of the expression mismatch Y for different values of the signal-to-noise ratio S, i.e. $S = 0$ (black), $S = 1$ (green), $S = 2$ (blue), and $S = 3$ (red).

Log-fitness at optimal coupling X_*

We next consider the case in which the promoter has a certain expression mismatch Y, and we calculate the log-fitness that it can obtain by *optimally* coupling to a regulator that has a certain signal-to-noise S and correlation R. **Figure S2.23** shows the resulting log-fitness values as a function of S and R for 4 different values of the expression mismatch Y: $Y = 1$ in the top-left panel, $Y = 2$ in the top-right panel, $Y = 4$ in the bottom left panel, and $Y = 8$ in the bottom right panel.

When the expression mismatch is small, i.e. $Y = 1$ corresponding to a variance in μ_e that matches $\sigma^2 + \tau^2$, then fitness generally increases with increasing R and S. However, the

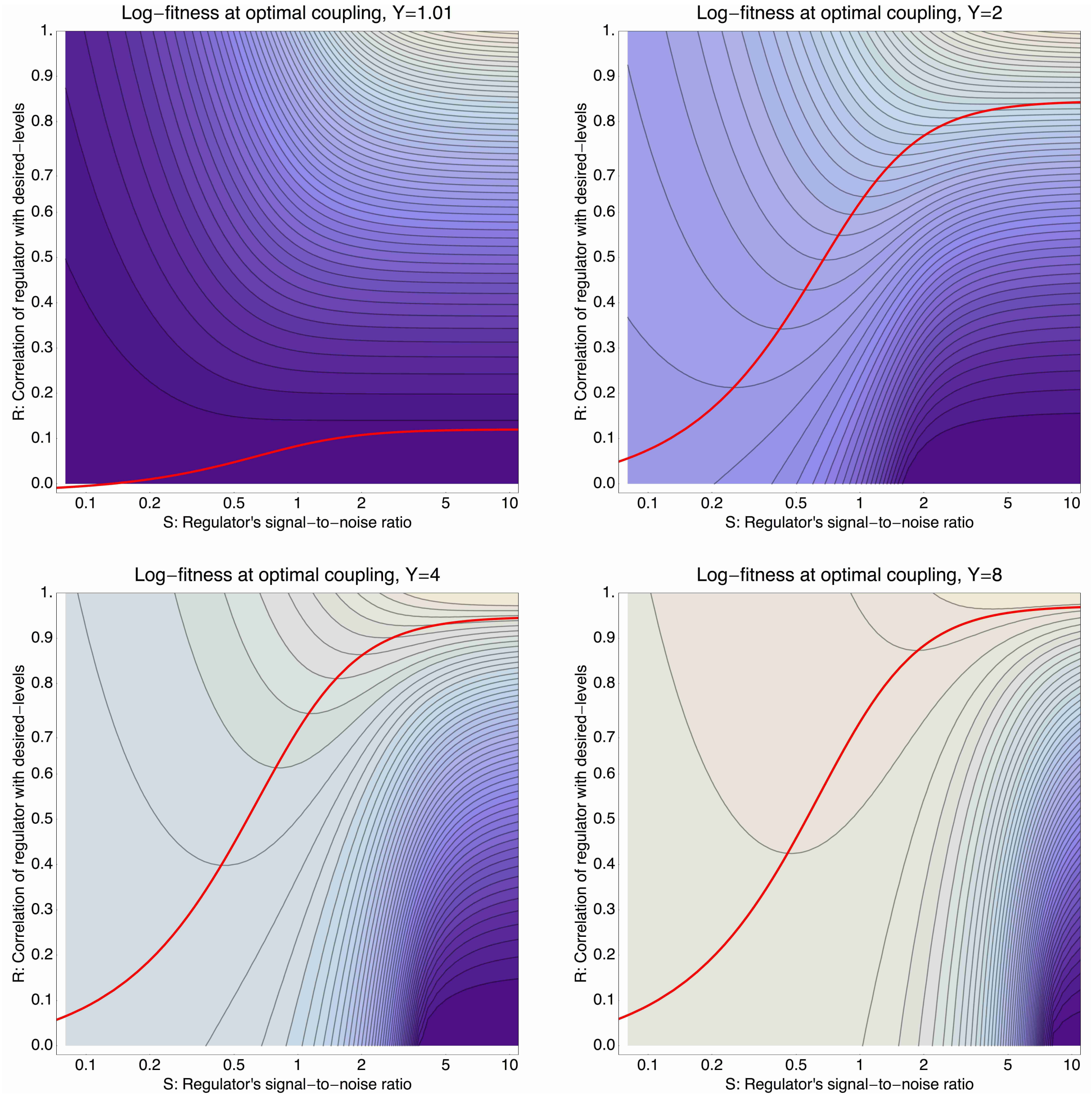

Figure S2.23: Log-fitness as a function of the signal-to-noise ratio S (horizontal axis) and correlation R of the regulator (vertical axis), for a promoter that is optimally coupled ($X = X_*$) to the regulator. The different panels correspond to log-fitnesses that are obtained for different values of the expression mismatch Y (indicated in the title of each panel). The contours run from -0.04 to -0.5 in the top-left panel, from -0.3 to -1.9 in the top-right panel, from -1 to -12 in the bottom-left panel, and from -2 to -30 in the bottom-right panel. The red curves show optimal signal-to-noise S as a function of the correlation R.

absolute value of the fitness increase is small. That is, for small Y promoters already have reasonably high fitness without regulation. As the value of Y increases, the log-fitness values start varying more dramatically as a function of R and S. Independent of the value of Y, the optimal fitness is always obtained for very high R and high S. However, when Y is large, an almost equally high fitness can be obtained by coupling to a 'noisy' regulator with low S and $R = 0$. In particular, when Y is large only regulators with very high correlation R and large signal-to-noise S can outcompete coupling to an entirely noisy regulator with $R = 0$ and small S. That is, in any situation where the desired expression levels vary significantly more than the width σ of the expression distribution, i.e. when $Y > 1$, promoters can substantially increase their fitness by coupling to a regulator that only acts to increase their noise. Moreover, to improve on this coupling to a 'random' regulator, a regulator has to be available with very high correlation R and large signal-to-noise. In other words, unless a regulator is available that very precisely regulates the promoter to attain its desired expression levels, best fitness can often be obtained by increasing the noise in the promoter's expression.

Note also that, whenever Y is larger than 1 and the correlation R is not very high, fitness generally decreases rapidly with the signal-to-noise of the regulator. That is, when a regulator has only moderate correlation with the desired expression levels of its target, low signal-to-noise is preferred. This suggests that regulators that are regulating targets whose desired expression levels correlate only moderately with the regulator's activity may be under selection for *lowering* their signal to noise ratios.

These considerations suggest different possible scenarios for the joint evolution of promoters and their regulators. On the one hand, when a regulator is coupled to a single promoter, or a set of promoters whose desired expression levels are perfectly correlated across environments, then a regulator can increase overall fitness by increasing the correlation R between the regulator's activities and the desired expression levels of its targets. In this way, regulation may evolve to become more precise over time. On the other hand, promoters often have an incentive to couple to regulators who only moderately correlate with their desired levels. Once a regulator is coupled to multiple promoters that have *different* desired expression levels, there is no way that the regulator can adapt its activities to correlate highly with the desired levels of all its targets, and such regulators will experience selection to become more noisy.

Final noise levels under optimal coupling and signal-to-noise

We next consider what final noise levels $\sigma_{\text{tot}}^2 = \sigma^2 + c^2\sigma_r^2$ result when a promoter, with a certain expression mismatch Y, couples optimally to a regulator which has a certain correlation R, and whose signal-to-noise level has been optimized as well.

To this end we need to determine the jointly optimal coupling X_* and signal-to-noise S_* given a certain expression mismatch Y and correlation R. From equation (2.91) it is easy to see that fitness is maximized with respect to the signal-to-noise level S when

$$S_* = \frac{RY}{X}. \tag{2.97}$$

If we substitute this value back into the equation (2.91), we find that the optimal coupling X_* is now given by

$$X_*^2 = \max\left[0, (1-R^2)Y^2 - 1\right]. \tag{2.98}$$

Note that, Y^2 is the variation in desired expression levels, and R^2Y^2 is the amount of this variation that the promoter manages to 'track' when it is regulated by the regulator. Thus, $(1-R^2)Y^2$ is precisely the remaining variance in desired expression levels that the promoter is unable to track. To bring this out more clearly, we substitute back our original parameters. We then find that when $(1-R^2)Y^2 < 1$:

$$\sigma_{\text{tot}}^2 = \sigma^2, \tag{2.99}$$

and when $(1-R^2)Y^2 > 1$

$$\sigma_{\text{tot}}^2 = (1-R^2)\text{var}(\mu_e) - \tau^2. \tag{2.100}$$

This brings out most clearly that, when the regulation is imprecise ($(1-R^2)Y^2 > 1$), the final noise level that is evolved matches the fraction of the variance in desired expression levels $\text{var}(\mu_e)$ that is not tracked by the regulation. In other words, the evolved transcriptional noise level of a promoter precisely reflects to what extent the promoter's regulation is not able to track the expression levels desired by the environment.

Fig. S2.8 shows the total noise level σ_{tot} as a function of Y and R when the promoter is optimally coupled to a regulator with optimal signal-to-noise. The figure illustrates that there are two regimes of solutions ('phases') separated by a phase boundary (thick black curve) that occurs at $(1-R^2)Y^2 = 1$. On one side of this boundary, in the top-left of the figure, the final noise level σ_{tot} is essentially not different from the original noise level σ. This occurs either when $Y < 1$, i.e. when no regulation is necessary, or when very accurate regulation is available. Note that, similarly to what we saw in the last section, very high correlations R are necessary to realize this regime at larger values of Y. We call this the 'basal noise regime'.

The largest part of the parameter space occurs on the other side of the phase boundary, which we call the 'environment-driven noise regime'. Here the final noise level σ_{tot} becomes independent of the original noise σ, but is instead determined by the fraction of variation in desired expression levels $\text{var}(\mu_e)$ that is not tracked by the regulation, i.e. by $(1-R^2)\text{var}(\mu_e)$. The figure also indicates the optimal values of S_* as a function of Y and R. The optimal signal-to-noise S_* diverges at the phase boundary. That is, in the 'basal noise regime' regulators are preferred with signal-to-noise that is as high as possible. In contrast, for the majority of the parameter space in the environment driven noise regime signal-to-noise levels of 1 or less are preferred. That is, unless regulation is very precise, noisy regulators are typically preferred over precise regulators.

The figure also demonstrates that, unless R is close to one, the final noise increases with the variance in desired expression levels $\mathrm{var}(\mu_e)$. Thus, unless there is a systematic correlation between the expression mismatch Y of a promoter, and the correlation of the regulator with highest available correlation, noise levels are expected to increase with the 'plasticity' $\mathrm{var}(\mu_e)$ that the environment requires of the promoters.

Similarly, the larger Y, the larger the remaining variance $Y' = (1 - R^2)Y^2$ tends to be after coupling to the regulator with the highest available correlation R. Whenever this remaining expression mismatch is Y' large, the promoter will have an incentive to couple to further regulators. That is, the theory also generally predicts that promoters with high $\mathrm{var}(\mu_e)$ tend to couple to more regulators.

Finally, we note that this theoretical model can easily be extended to the case of multiple promoters and regulators. In particular, because in our model the promoter's expression is a linear function of the regulatory inputs, the theory extends easily to promoters coupling to multiple regulators with different coupling constants. However, the regulatory network structure that will evolve in this general case will depend crucially on the correlation structure of the desired expression levels across all the promoters. Moreover, there might be many environmental changes that affect the optimal expression levels, but that cannot be sensed by any of the regulators, and this will constrain the extent to which regulators can optimize their activities to match the desired levels of their targets.

Chapter 3

σ^{70} binding is a prerequisite for expression but not predictive for transcript levels

The previous Chapter 2 described the evolution of expression levels towards predefined expression levels on the phenotypic level. In this Chapter, we are zooming into the sequence features that give rise to the expression levels observed.

3.1 Introduction

A very fundamental question in molecular biology is what sequence properties separate regulatory sequences from coding regions. Here we investigate the minimal requirements for a sequence to function as a promoter sequence.

A gene can only be of benefit for an organism if it can be expressed and thus be seen by selection. New genes integrate into the genome by different mechanisms, mainly gene duplications (Serres et al., 2009) and horizontal gene transfer (Ochman et al., 2000, Treangen & Rocha, 2011) across species borders. In both cases, genes may be transferred including regulatory sequences upstream of the genes or not. If a gene is not transcribed at all, regulation has to evolve from the non-functional sequence upstream in order to be of benefit for the organism. If a gene is transcribed, but protein levels have to be adjusted, transcriptional regulation has to evolve so as to reach the correct protein amounts in the cell.

What are the outstanding sequence features that characterize a promoter sequence? Functional promoter sequences have the ability to recruit the RNA polymerase to their sequence. This marks the first step in transcription initiation, which takes place upstream of the coding region of the regulated gene. In bacteria, promoter sequences provide binding sites for a particular subunit of the RNA polymerase, the sigma factor (Helmann & Chamberlin, 1988, Burgess & Anthony, 2001). Each bacterial species encodes a set of sigma factors, where each sigma factor recognises a subset of all promoter sequences in the genome (Helmann & Chamberlin, 1988). Each subset contains a sequence in the regulatory region of the gene that allows interactions between the sigma factor and its target gene (Gruber & Gross, 2003).

The affinity of the RNA polymerase to a promoter sequence is thus a function of the activity of the sigma factors in the cell and their binding sites. Strength of σ^{70} binding is indeed a well-known predictor for transcript level (Blank et al., 2013, Brewster et al., 2012, Kiryu et al., 2005, Szoke et al., 1987) and has also been shown to correlate for σ^{24} (Rhodius & Mutalik, 2010). Small sequence changes in the -10 or -35 region of the σ^{70} site have predictable changes on the expression level when keeping the surrounding sequence constant (Brewster et al., 2012). Other sequence features around the -10 and -35 region in the bacterial promoter region can also impact expression levels, like the spacer length between those regions (McLean et al., 1997). Nucleotides upstream of the -10 region, called extended -10 region, impact promoter strength (Burr et al., 2000, Voskuil & Chambliss, 1998, Voskuil & Chambliss, 2002) as well as the so-called UP element upstream of the -35 region (Ross et al., 2001, Espinosa et al., 2005, McCracken et al., 2000, Rhodius et al., 2012).

Besides sigma factors, bacterial promoter sequences also provide binding sites for more specific transcription factors. The combinatorial design of promoters in terms of transcription factor binding sites allows differential gene expression in varying conditions. Selection favors incorporating binding sites for specific transcription factors whose activities are in accordance with the desired expression levels in the environments selected. Each bacterial cell has a set of transcription factors it can make use of. In *E.coli* , there are 271 transcription factors known (Babu & Teichmann, 2003) that may show differential activities over varying conditions. By altering the binding affinity for a factor active in a certain environment, expression levels for a particular gene can be changed on the transcriptional level already (Lloyd et al., 2001).

After binding of the RNA polymerase holoenzyme complex to the promoter sequence, the DNA around the -10 region has to be melted in order to form the open complex. The rate of open complex formation theoretically depends on the melting temperature of the DNA double-strand (Djordjevic & Bundschuh, 2008) and is helped by region 1.2 in the sigma factor (Bochkareva & Zenkin, 2013, Revyakin et al., 2004, Haugen et al., 2008). Native promoter sequences in *E.coli* (*Escherichia coli*) are over 10% more AT rich than coding regions (Blank et al., 2013), facilitating possibly unwinding of the DNA duplex. To what extent unwinding influences transcription levels is poorly understood, similar as the last step during transcription initiation which involves the release of the RNA polymerase holoenzyme from its binding site (Revyakin et al., 2004).

Prediction and modification of expression levels from regulatory sequences has been attempted on various control levels. Translation rates can be tuned by changes in the ribosomal binding sites (Lee et al., 2013) and their surrounding sequence affecting secondary structure of the RNA molecule (Salis et al., 2009). RNA stability is also influenced by the sequence itself (Carrier & Keasling, 1997, Carrier & Keasling, 1999). Transcription rates of individual genes differ over two orders in magnitude, with many genes showing similar expression levels in a given environment as shown in Chapter 2. The differences in expression levels are only explained by the variable transcriptional activities and these phenotypic differences are encoded in the promoter sequences. Prediction of expression levels based on σ^{70} binding strength is possible in otherwise constant promoter sequences (Brewster et al., 2012). Native promoter sequences harbor many binding sites for transcription factors, and different combinations of these factors do not always allow reliable prediction of the expression level (Kosuri et al., 2013).

Evolution of a functional promoter sequence, starting from a random sequence has already been shown in a directed evolution experiment, where σ^{70} did indeed evolve after only few rounds of selection (Liu & Libchaber, 2006). Here, we extend this experiment by starting artificial selection on the expression level from around a million of random sequences. Phenotyping expression levels in combination with genotyping gives us insights into the sequence features that determine whether a sequence is a functional regulatory sequence.

3.2 Main part

Genotypic characterization of promoter pools obtained after each round of selection or mutation and selection was done by deep sequencing. This allows tracking of evolutionary fates of individual promoter sequences over rounds and furthermore enables classification of expression levels in later rounds. The latter can be refined with single sequence sequencing in combination with phenotypic characterization of expression levels of cells in clonal populations. Those obtained expression distributions are characterized by their means and variances and can be readily compared to promoter genotypes of native genes in *E.coli* .

3.2.1 Characterization of the initial library

For comparison of the initial promoter library to functional promoter sequences, we first characterized the initial sequence pool.

The starting pool of promoter sequences consisted of random oligonucleotides with flanking restriction sites, PCR-doublestranded and cloned in front of GFP on the low-copy plasmid pUA66 (Zaslaver et al., 2006) as described in Chapter 2. The distribution of expression levels found in cells with these random promoter sequences was close to the expression levels found in cells without the plasmid (wildtype strain DH10B) or cells without promoter sequence but empty plasmid (Fig. 3.1A). As cellular background fluorescence in wildtype cells was comparable to cells with vector but without promoter sequence, we concluded that the plasmid itself did not contribute to the fluorescence detected. The initial library pool already contains functional promoter sequences (Fig. 3.1A) with measurable transcriptional activity above background indicated by a tail in the single cell expression distribution towards higher expression levels.

Initially, the planned size of promoter sequences should have spanned 157 bp but was shortened in most promoter sequences due to inefficient nucleotide coupling and strand breaks during primer synthesis. Promoter sequences were deep sequenced from one end,

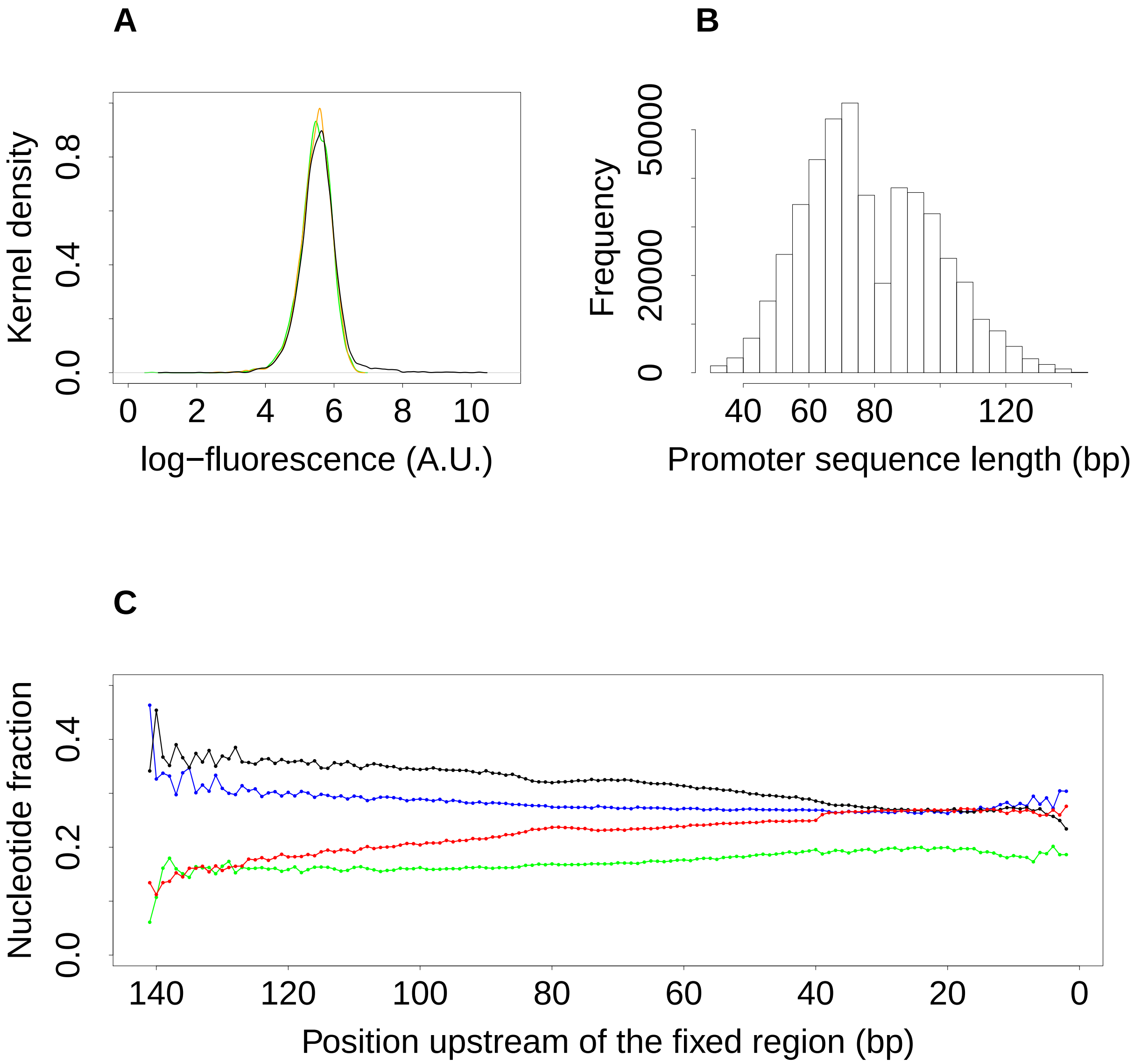

Figure 3.1: **The initial promoter library.** (A) Measured distributions of GFP fluorescence from single cells. Background fluorescence from wildtype cells (orange), wildtype cells with modified vector pUA66 (green) and fluorescence of wildtype cells with the initial library of promoter sequences upstream of GFP (black). (B) Histogram of lengths of random promoter sequences in the initial library. (C) Nucleotide frequencies upstream of the fixed promoter region, taken from unique sequence reads. Bases A, C, G and T are colored in green, blue, black and red respectively.

including promoter sequences spanning a size-range of 30-141 bp (Fig. 3.1B, Methods Section 3.4.2). In total, 890'019 sequences were sequenced from the initial library, of which 472'160 sequences were unique sequences (100% sequence identity). From the overall transformation efficiency we concluded that around 1 million unique promoter sequences were present in the initial library pool. A biased nucleotide distribution was observed for the unique sequences of the initial library, containing 18% A, 25% T, 30% G, and 27% C. Additionally, there are nucleotide biases observed in the unique promoter sequences showing increasing GC content with increasing distance from the translational start site (Fig. 3.1C).

3.2.2 Similar sequences exhibit similar expression

If sequence information is predictive for expression level and variations within, similar sequences should exhibit comparable expression levels. To address this, single promoter sequences were clustered based on their genotypes and phenotyped based on their expression levels.

Out of 479 individual promoter sequences after 3 and 5 rounds of selection that were phenotyped, 378 sequences were Sanger sequenced individually. 316 of these sequences were unique promoter sequences, that could be mapped to 166 clusters based on sequence identity (see Methods section 3.4.5)). Sequences that were mapped to the same cluster were considered to be derived from a common ancestor sequence. For a sequence of length 157 bp, 4^{157} possible sequences could had been synthesized; but with only 1 million sequences successfully transformed, chances that a sequence derived from a few-bp neighbor were negligible. Mapping of the similar genotypes to their respective phenotypes (Fig. 3.2) shows that the phenotype depends strongly on the genotype.

Most sequences within clusters show very similar expression, indicating that small changes in the genotype do not impact the phenotype dramatically in many cases. Generally, promoter sequences are mutationally robust, meaning that many mutations do not impact expression levels. However, some of the clusters show outliers that do not follow this general trend. These can be separated into clusters containing sequences that were selected for both medium and high expression or were only selected for one expression level but containing outliers. Some particular mutations can have severe impact on the expression levels, which allows selection of new phenotypes in the case of changing environments. Besides introduction of novel binding sites for transcription factors to bind to and allow differential expression in varying environments, populations can also undergo changes in the promoter sequence that alter their expression levels possibly independent of specific factors. The mean levels of expression can be better separated than the variance in expression levels across cells (excess noise, see Chapter 2), where only two clusters that were selected for medium expression show clearly elevated levels of excess noise.

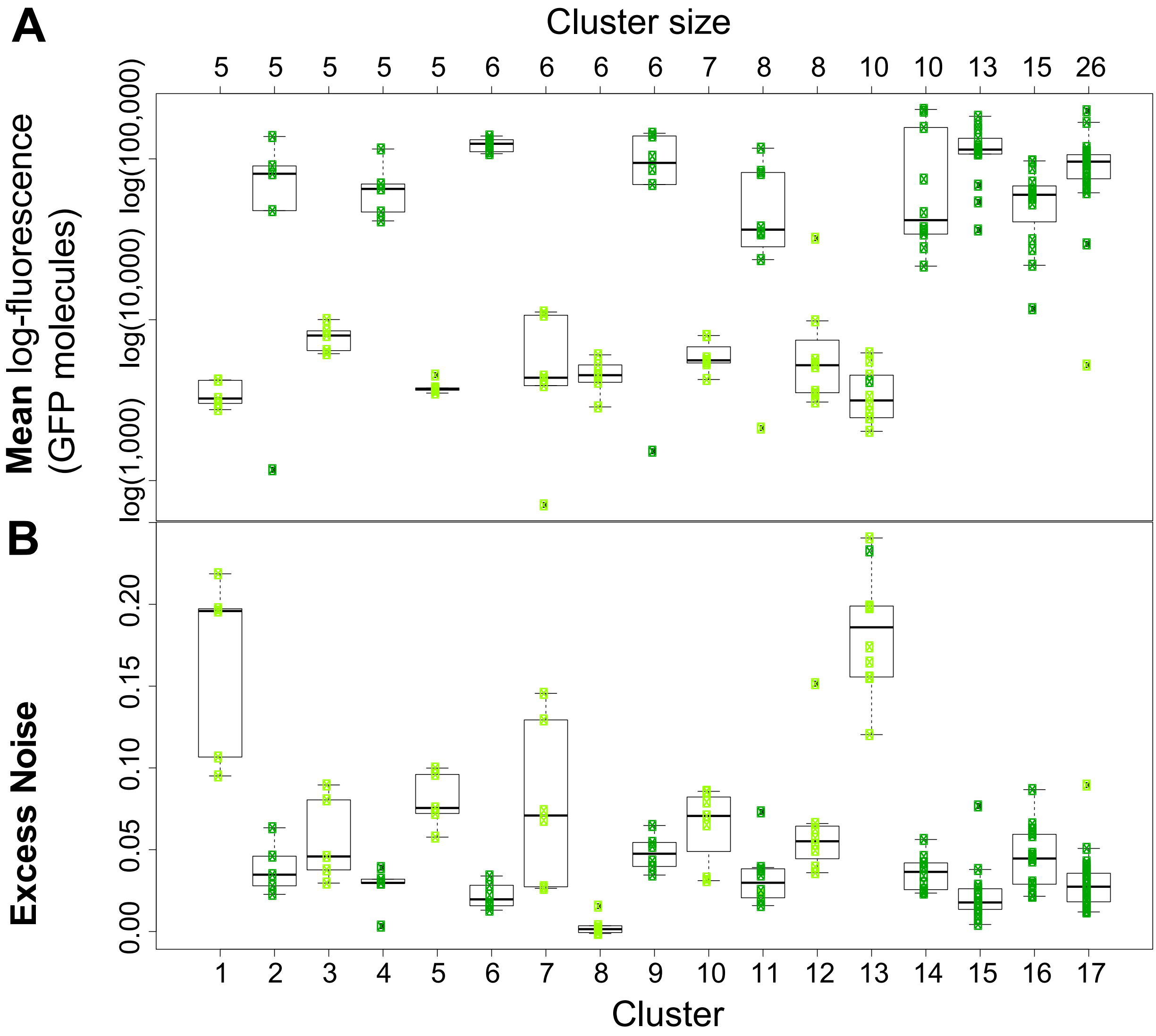

Figure 3.2: **Expression characteristics of promoter sequences with high sequence similarity.** Phenotype measures for similar sequences in clusters with cluster size bigger than four. Sequences that were selected for medium expression levels are plotted in light green, sequences selected for high expression levels are in dark green. Cluster sizes indicated at the top. The box in the boxplots represents the first and third quartiles with the band in the box representing the second quartile (the median). The lower end of the whiskers represent the lowest datum found within 1.5 IQR (interquartile range) distance from the first quartile and the upper end of the whiskers represent the highest datum found within 1.5 IQR from the third quartile. (A) Mean number of GFP molecules per cell as inferred from the FACS measurement for each sequence cluster. (B) The excess noise associated with members in each cluster.

3.2.3 Evolution of transcription factor binding sites

Recruitment of the RNA polymerase to the promoter sequence is supported by transcription factors, that bind in the promoter region or close by. To test if a certain fraction of promoter sequences has evolved transcription factor binding after three and five rounds of selection, transcription factor binding sites were predicted (Chapter 3.4.6) for native *E.coli* promoter sequences with known experimentally validated binding sites (Salgado et al., 2013) using sequence motifs from these sites. The same prediction was performed on the evolved sequences which were Sanger sequenced (1 representative sequence from each of the 166 clusters). Sequence composition biases that could lead to predicted binding sites were excluded by testing the predictions on the synthetic sequences that were shuffled, maintaining ATGC content and sequence length. For the 39 transcription factors where we obtained sequence motifs and promoter sequences with predicted sites in native sequences (Fig. 3.3), some synthetic promoter sequences reached the predicted binding strength of the strongest predicted TFBS that had been observed in native promoters. However, binding strengths between the original synthetic and shuffled synthetic sequences were comparable, not showing statistically significant differences (the lowest two-sample Kolmogorov-Smirnov test p-value observed was 0.03 for factor CpxR).

These results suggest that for the specific transcription factors we analyzed, strong transcription factor binding sites have not been evolved. For the sigma factors, that are thought to be essential for binding of the RNA polymerase to the promoter sequence, binding sites were predicted especially for the housekeeping sigma factor σ^{70} (Fig. 3.4). The difference between the maximum weight matrix scores predicted in native and synthetic promoter sequences was not significant, while the distributions of scores between synthetic and synthetic shuffled sequences were statistically significant dissimilar (two-sample Kolmogorov-Smirnov test, $p = 6.5 * 10^{-1}$ and $p = 3.8 * 10^{-11}$). Binding strengths of σ^{38} and σ^{32} in synthetic sequences were also closer to native sites (two-sample Kolmogorov-Smirnov test, $p = 1.3 * 10^{-3}$ and $p = 1.7 * 10^{-2}$) than synthetic shuffled sequences, but the clear separation between original synthetic promoter sequences and the shuffled versions was less strong (two-sample Kolmogorov-Smirnov test, $p = 1.9 * 10^{-3}$ and $p = 4.2 * 10^{-3}$).

Native *E.coli* promoters evolve binding sites for specific transcription factors, allowing them to express their genes at different levels in changing conditions. However, the synthetic promoter sequences have been selected for one defined expression level in only one given condition. In this condition, only a subset of the 271 specific transcription factors (Babu & Teichmann, 2003) will show activity that might be favorable for their expression levels. Generally, a sequence that has no activity will evolve sites only for transcription factors that positively impact their expression levels, i.e. activating transcription initiation. Second, if sigma factor binding in the RNA polymerase complex is a pre-requisite for transcriptional activity at a promoter sequence, than there are additional size restrictions, that would maybe not allow the promoter sequences to evolve additional sites for transcription factors on such a short promoter sequence in synthetic promoters. Evolution of two functional sites, namely

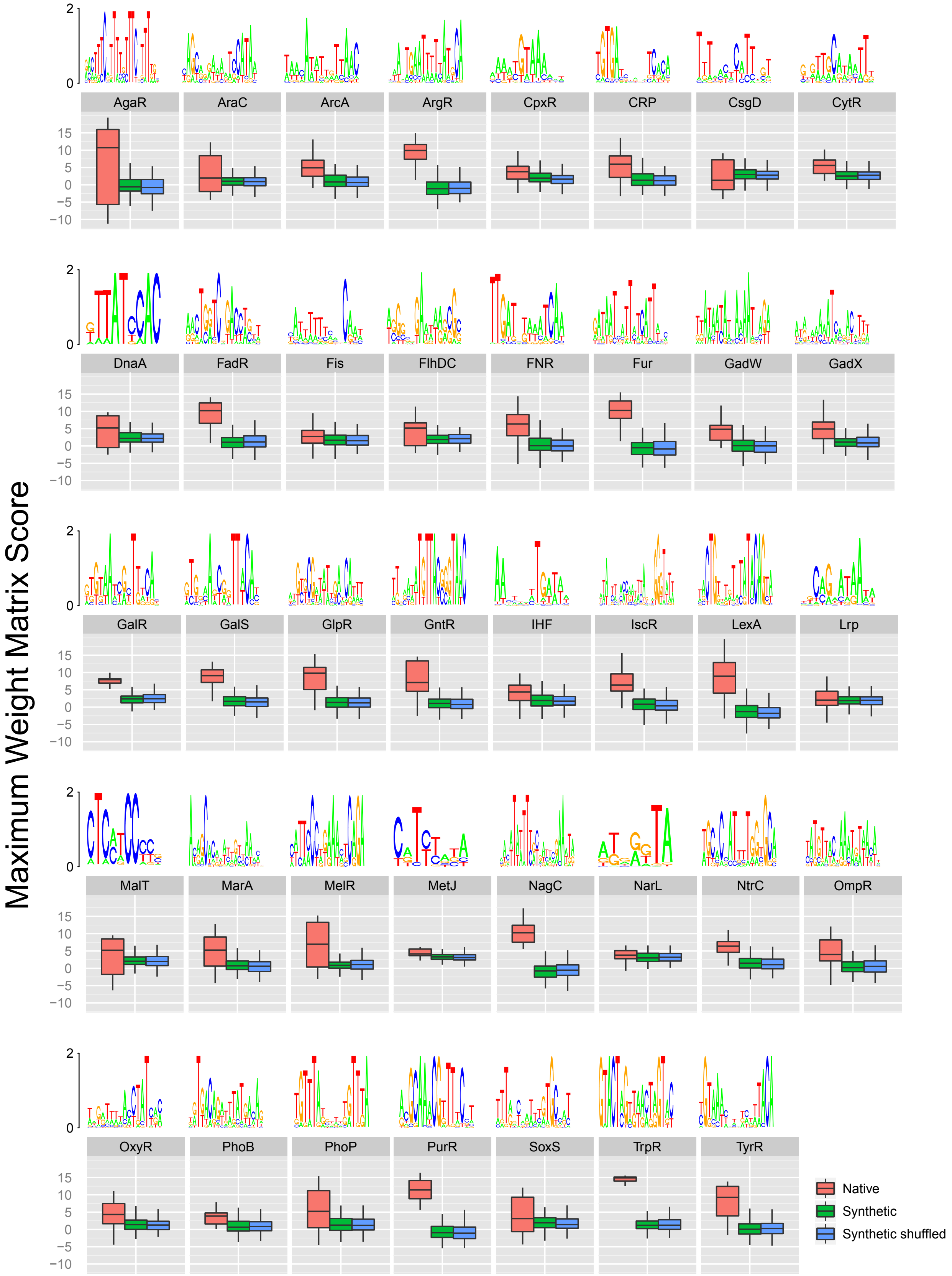

Figure 3.3: **Strength of the strongest predicted transcription factor binding site for 39 transcription factors in *E.coli*.** For each transcription factor, native promoter sequences with known sites are compared to synthetic evolved promoter sequences and their shuffled derivatives. Weight matrices used for prediction are shown above each factor, the height of the letters representing the information content in bits.

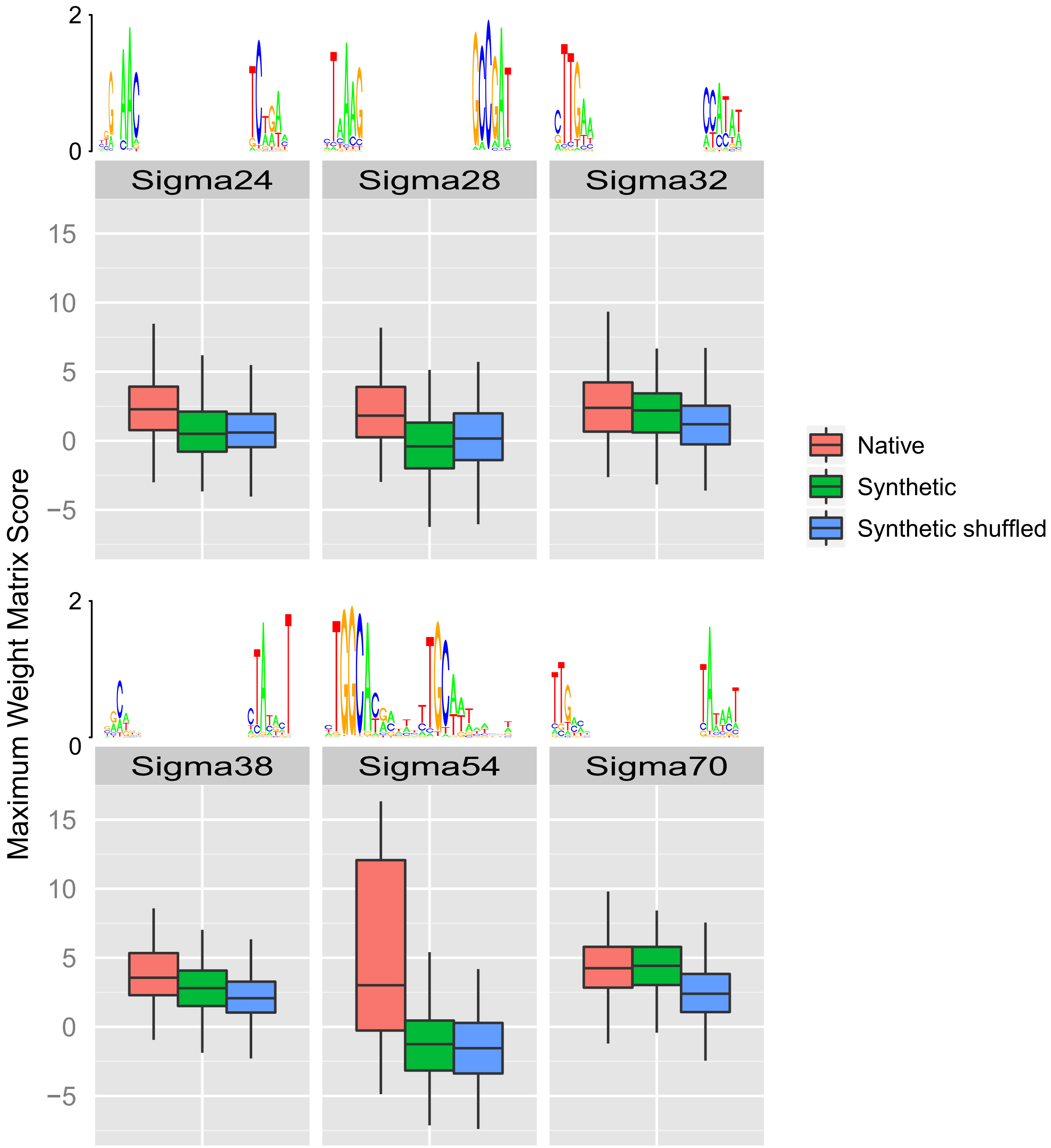

Figure 3.4: **Strength of the strongest predicted transcription factor binding site for 6 out of 7 sigma factors in *E.coli.*** For each sigma factor, native promoter sequences with known sites are compared to synthetic evolved promoter sequences and their shuffled derivatives. Weight matrices used for prediction are shown above each factor, the height of the letters representing the information content in bits.

for a sigma factor and another specific transcription factor, at the same time over a short evolutionary timescale seems to be a much harder task than evolution of only a sigma site. Small changes in the promoter sequence apart from transcription factor sites are able to change expression level.

The activity of sigma factors in the cell is known to be regulated on multiple levels and changes dependent on the state of the cell (Lange & Hengge-Aronis, 1994). Having evolved mainly the sigma housekeeping factor σ^{70} indicates that this was the predominant sigma factor present in the conditions selected in. Weak signals could also be obtained for σ^{38} involved in the stress response in *E.coli* (Battesti et al., 2011), or the prediction is simply a side-effect of the closeness of the consensus sequences for σ^{70} and σ^{38} binding, especially in the -10 region.

3.2.4 Evolutionary dynamics of sequence features during promoter sequence evolution

Promoter sequences that give rise to expression levels close to the selected level have a higher chance to be maintained during experimental evolution. Tracking sequence features from the initial library to medium expression reveals important sequence features in the transition of random to functional promoter sequences. Underlying genotypes from medium to high expression levels give insight into sequence properties important for high transcriptional activities.

Some sequence clusters take over during evolution

Successful promoter sequences should contribute more copies of their sequences and sequences deriving therefrom after several rounds of selection. We tested if only a few sequences took over during the selection process or if a large number of sequences were similarly fit. Random promoter sequences were selected based on their expression level, resulting in expression distributions after only five rounds of selection that are close to the target level (Fig. S3.1). Going through multiple rounds of selection provides a bottleneck which sequences make it to the next rounds or not. On the other hand, mutation steps introduce greater variability of the sequences present, expanding sequence diversity. The sequences that are not the same, but vary by only a few base pairs can be considered as arising from the same ancestor. These sequence clusters can be tracked from round-to-round, allowing a picture of what sequences were preferred and to which degree the mutations were allowed during evolution. Deep sequencing allowed an insight into the sequence variability observed after each step of selection or selection and mutation. As expected, the number of clusters observed was biggest in in the initial library with 426'092 clusters, with a downtrend up to round 5 which ended up with 2'974 clusters for the high expression line and 1'994 cluster for the medium expression line. Interestingly, 876 sequence clusters were shared between medium and high expressors, although most sequences observed in those clusters belonged to either medium or high expression. Small differences in the sequences can thus have a big impact on the expression level, but in the main expression levels stayed at a similar level (Fig. 3.2A). The phenomenon that many clusters between medium and high expression were shared was due to the fact that both populations derived from the same pool of sequences and the expression levels of the populations after five rounds were still partly overlapping (Fig. S3.1). Selection in combination with mutation removed most of the clusters observed in early rounds of evolution, but allowed some successful clusters to expand dramatically in the promoter population (Fig. 3.5).

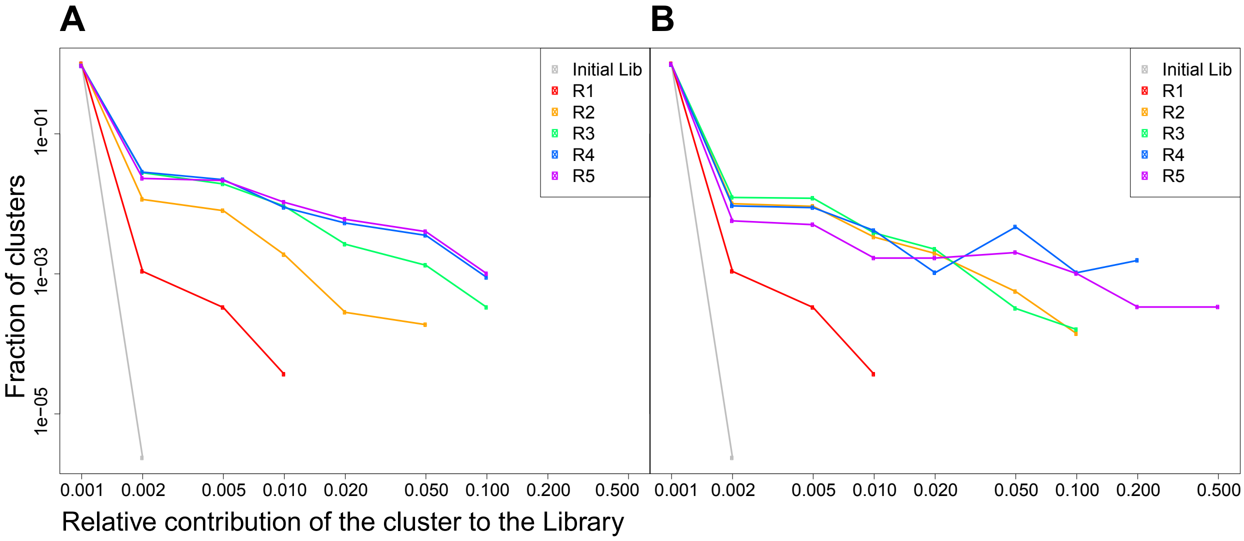

Figure 3.5: **Evolution of cluster contributions after each round of selection or selection/mutation over evolutionary rounds.** Cluster contributions are binned (up to the number plotted). Colors represent the different libraries sequenced, the initial library in grey and proceeding libraries as depicted in the legend for both medium (left panel) and high expression (right panel).

In the initial library, most clusters apart from few exceptions contribute very little to the entire library. This fraction does only marginally decreases over evolutionary time, but some clusters expand and contribute more in terms of sequence numbers to the libraries in later rounds. In medium expression lines (Fig. 3.5A), this trend seems to be saturated after 4 rounds of evolution. Evolution of promoter sequences to high expression (Fig. 3.5B) follows a similar trend, with two clusters that contribute more than 34% to the fifth round.

The contribution of the sequence cluster reflects the fitness of the sequences within that cluster. Thus, taking into consideration all sequences that were found within each library shows which genotypes were favored during sequence evolution.

Selection on expression level favors longer promoter sequences

Promoter lengths found in the initial library were quite diverse (Fig. 3.1). The longer a sequence gets, the higher the probability that functional sites that enable expression are present. On the other hand, functional sites may have to be within a certain distance to the start of transcription or start of translation. That is why the length distribution of later rounds in evolution give an idea about what optimal lengths may be given to drive transcription in the space of sequence lengths available. During the first two rounds of selection, longer sequences were generally preferred in medium and high expressing lines (Fig. 3.6). Selection for high expression favors even longer sequences than medium expression.

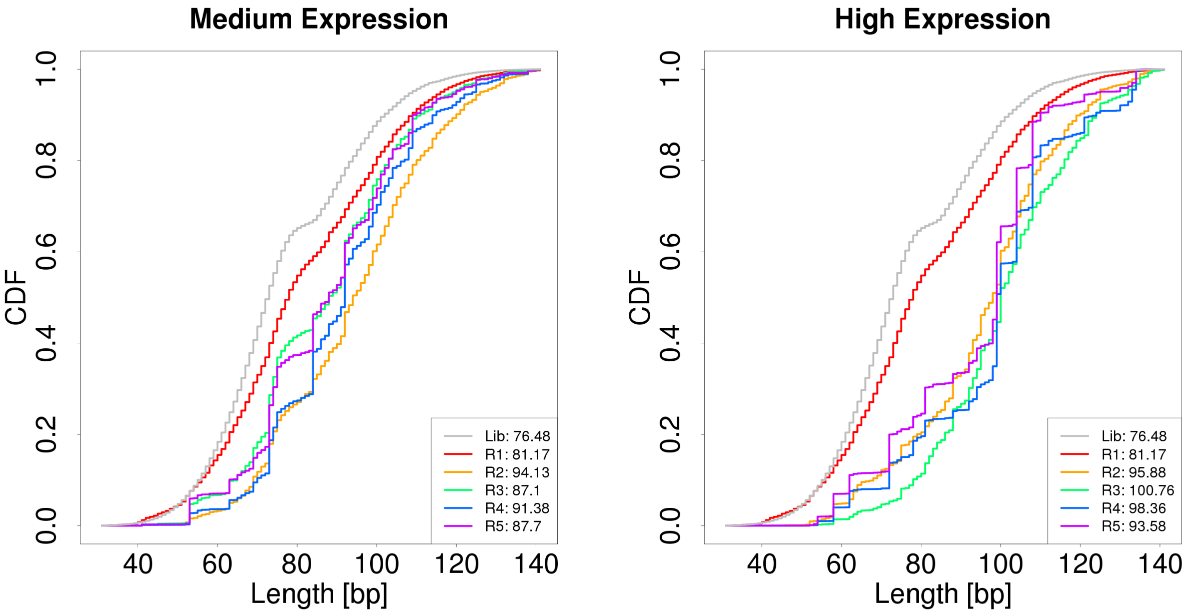

Figure 3.6: **Evolution of promoter sequence lengths (the variable part) over evolutionary rounds.** Cumulative distributions of lengths in bp content found in each of the libraries for medium (left side) and high (right side) expression. Libraries are colored as depicted in the legend and the mean lengths are given.

Selection on expression level favors AT-rich sequences

Given that the initial promoter region library started with a biased nucleotide composition towards GC richness over all positions, the nucleotide composition was tracked over all rounds of evolution. As native promoter sequences are generally rich in AT, we wanted to investigate the importance of nucleotide frequencies in the synthetic promoter sequences. It appeared that the frequency of bases observed was similar between medium and high expression lines, although small differences emerged. The general trend was lowering of C content while increasing A and T content (Fig. 3.7). In particular, A content increased for two rounds of selection with high expressing promoters on top elevated their A content further in later rounds. For the C content, the exact opposite trend was found. Although G was the dominant nucleotide observed in the initial library (Fig. 3.1C), high C content was selected against stronger than high G content, especially in the high expression lines. Percentage of T's raised until the second round while % G was reduced.

The differences in nucleotide content selected between medium and high expression lines might be an indicator for expression levels. As high expression lines in the last two rounds are dominated by a few clusters, expression differences might also be explained by other features, and the nucleotide distribution might be a by-product of those features. Generally, evolution of higher AT content from the initial library to later rounds of selection support the trend observed in native *E.coli* promoters, which are shown to have higher AT content

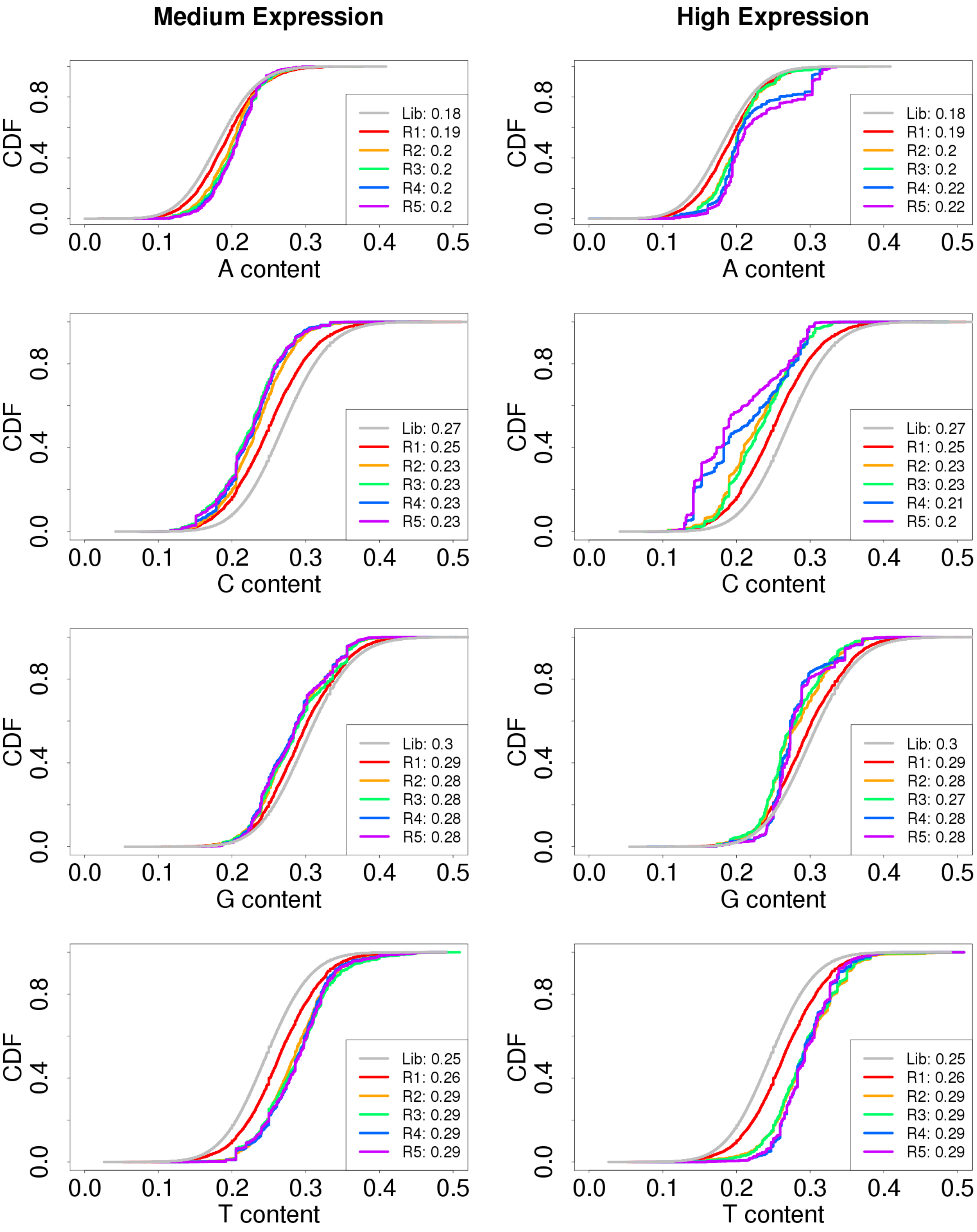

Figure 3.7: **Evolution of promoter sequence composition over evolutionary rounds.** Nucleotide fractions are shown for each type of nucleotide over all evolutionary rounds, medium (left panel) and high (right panel) expression. Mean ACGT content for each library is depicted in the legend.

than coding regions (Blank et al., 2013). Strongest selection for elevated AT content takes place in the first few rounds, suggesting that a raised AT content gives a higher probability of hitting a functional sequence from a random distribution.

Low free energies of σ^{70} binding as a recognition feature for functional promoter sequences

We tested the hypothesis that σ^{70} binding strength can be used as a predictor for expression levels. Upon binding of the sigma factor to its binding site, energy is released. The greater the difference in energy between the bound and the unbound state, the more likely the sigma factor remains bound to the particular site. Free binding energies were calculated for each sequence (Chapter 3.4.7) that appears in each evolutionary round and the distribution of σ^{70} free binding energies observed are shown in Figure 3.8 along with the shifts in expression distributions. Before and after the first step of selection, promoters with higher expression levels are selected, shifting the overall fluorescence distribution. A similar trend is observed for more negative free binding energies of σ^{70} binding, selecting for promoter sequences that have more negative free binding energies for σ^{70} binding predicted. This confirms that expression levels of promoters with σ^{70} binding sites is believed to increase with stronger binding (Brewster et al., 2012). From evolutionary rounds two to three, increase in expression levels was correlated with more negative free binding energies in both medium and high expression. From round three to five, expression levels stay the same for medium expressors, as free binding energies do. High expressors deviate from round three to five from this general behavior, showing elevated expression with unchanged free binding energies. This suggests that in a regime ranging from low or no expression to medium expression, better binding of σ^{70} is required. In the rounds where free binding increased with expression level, also shuffled versions would increase their binding energies, but not as strong. Sequences in those round transitions were also selected for longer sequences (Fig. 3.6) and higher AT content (Fig. 3.7). In the transition from medium to high expression, binding strength cannot account any more for the expression differences observed. Changes must have arisen from other sequence features.

Smaller entropy in binding energies are favored

Promoter sequences may achieve one strong site for σ^{70} binding, resulting in one main transcriptional start site from which transcription is initiated. However, multiple weak binding sites may also be present for recruitment of the RNA polymerase to the promoter region. We calculated the entropy of σ^{70} binding predicted and found that entropy of promoter sequences is smaller in comparison to their shuffled relatives. Entropies for medium (Fig. 3.9A) and high (Fig. 3.9C) expression lines decline up to round three and then do not change any more. The shuffled versions of the sequences (Fig. 3.9B and D) change in the other direction towards higher entropy, more strongly for the high expressors. Although shuffled sequences increase the number of potential σ^{70} binding sites in their promoters, real sequences focus on providing few binding sites.

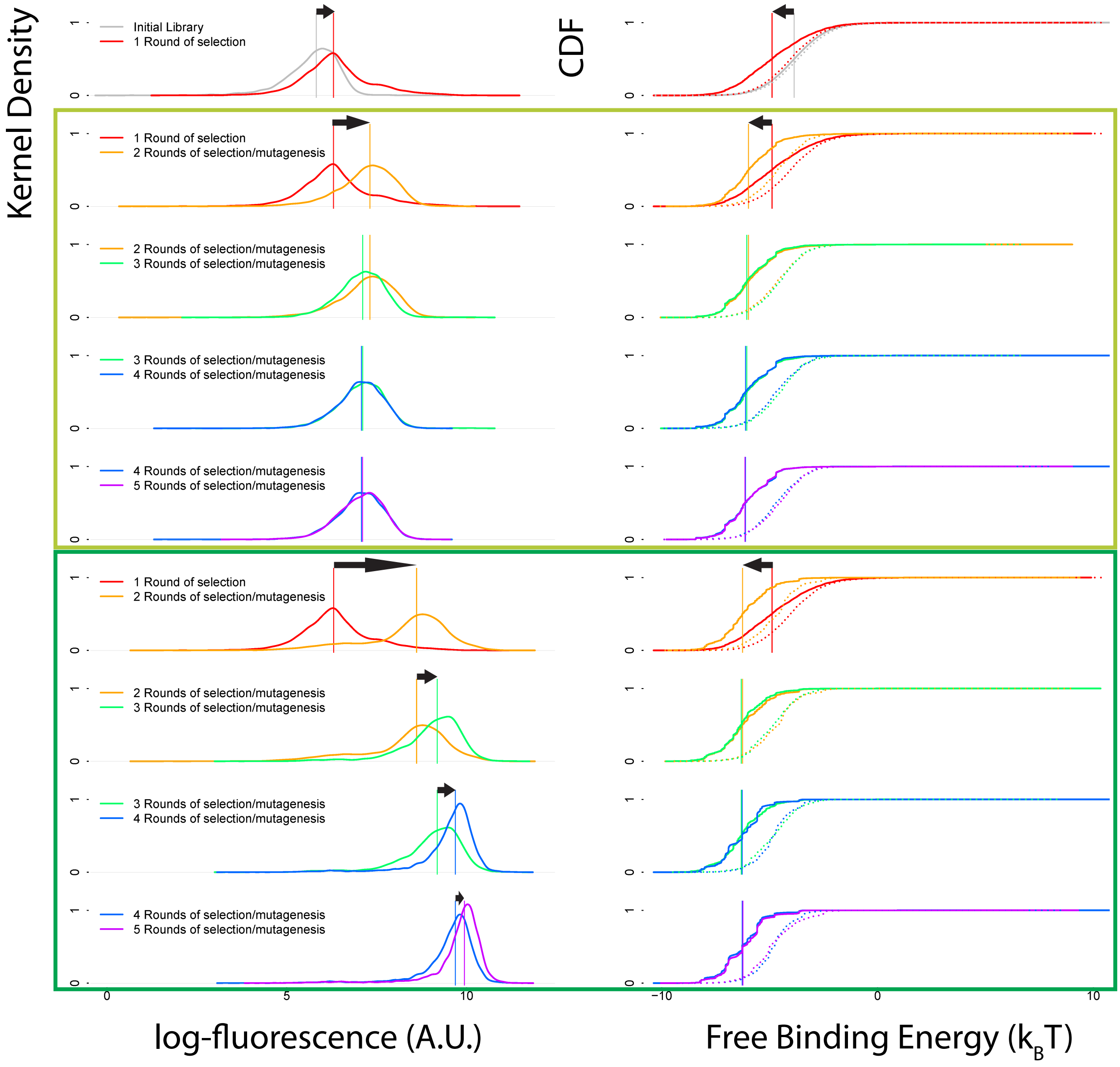

Figure 3.8: **Evolution of expression distributions and free binding energies for σ^{70} binding over evolutionary rounds.** Kernel density distributions of log-fluorescence (in arbitrary units, left side) are shown along with predicted cumulative density functions of free binding energies (in K_BT units, right side) for all sequence libraries from the initial library to after the last round selection. The first row shows the shift in distributions from the initial library to after first round of selection, which was the same for medium and high expression lines. Vertical lines specify the median values of the distributions, the arrows show the direction of the shifts. Dotted lines on the right side display free binding energies of the respective libraries, but drawn for shuffled sequences. The following four rows framed in light green show the distribution for the medium expression line. High expression line is framed in dark green.

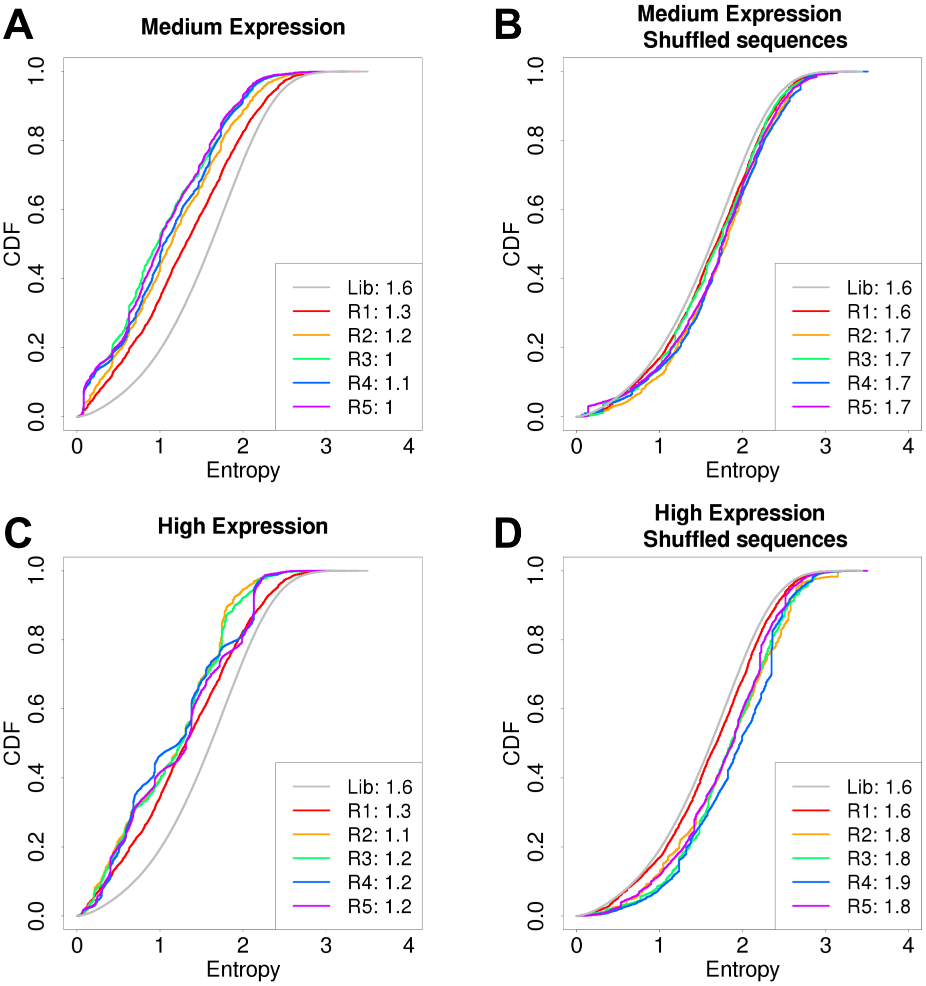

Figure 3.9: **Evolution of entropies of free binding energies for σ^{70} binding over evolutionary rounds.** Cumulative density functions of the entropy are shown for all sequence libraries from the initial library to after the last round of selection. The values in the legend behind the name of the libraries gives the calculated mean entropy in binding energies of the respective library. (A) and (C) show the distributions for the medium and high expressing line, (B) and (D) the distributions for the sequences from the same rounds, but shuffled.

Evolution of the initial library to medium expression includes preference of promoter sequences that show smaller entropies in their σ^{70} binding, meaning that few strong sites are selected over a large number of weak sites. Differences in expression levels between medium and high expression do not go along with a decrease in binding entropy, as entropies of binding do not decline any more from round three to five.

Short 5'UTRs are more successful

Positioning of σ^{70} binding sets the start of transcription and thus defines the length of the 5' untranslated region (UTR) that is transcribed. We wanted to test if the length of the 5'UTR may have an effect on the expression level. Based on the σ^{70} strengths predicted, the average position of the σ^{70} binding site was calculated (Chapter 3.4.7). Length of the 5'UTR is the average position plus 39 bp. Generally, the average length of the 5'UTR (Fig. 3.10A and C) increases from the initial library to later rounds for both medium and high expression, but less strongly as expected from the length and nucleotide distribution (Fig. 3.10B and D). Importantly, the length of the 5'UTR seems to be under selection towards smaller sequence lengths in comparison to their shuffled variants. As differences in expression levels are explained by transcriptional activity changes and the 5'UTR of the construct provides a strong ribosomal binding site, small 5'UTRs may help initiation of transcription, at least at the transition from low to medium expression.

3.3 Discussion

Our work illustrates sequence features that separate a random from a functional promoter sequence. Chances for a random sequence to provide functionality are quite significant.

As the initial library was sequenced with only two times coverage and the number of individual promoter sequences transformed was estimated from cell counting, we expect that the initial library contained at least one million individual sequences. Given that around 4'000 sequence clusters were present after five rounds of selection for medium and high expression, the chance of hitting a functional sequence under the given length and ATGC distribution was around 1:250. Hitting a functional coding sequence that gives rise to a particular function seems much more unlikely than hitting the right expression level in a given condition.

There are many ways for promoters to encode for a certain expression level in a given condition. The ability to change expression levels across conditions may limit the combinations possible to encode for all the different levels. Most changes in the promoter sequences did not impact the expression levels in the majority of cases. Exceptions from this general pattern are important- they allow expression modifications even after small changes enabling fast adaptation to changing environments. Variations associated with the expression levels of single promoter sequences were also dependent on the sequence, but absolute differences

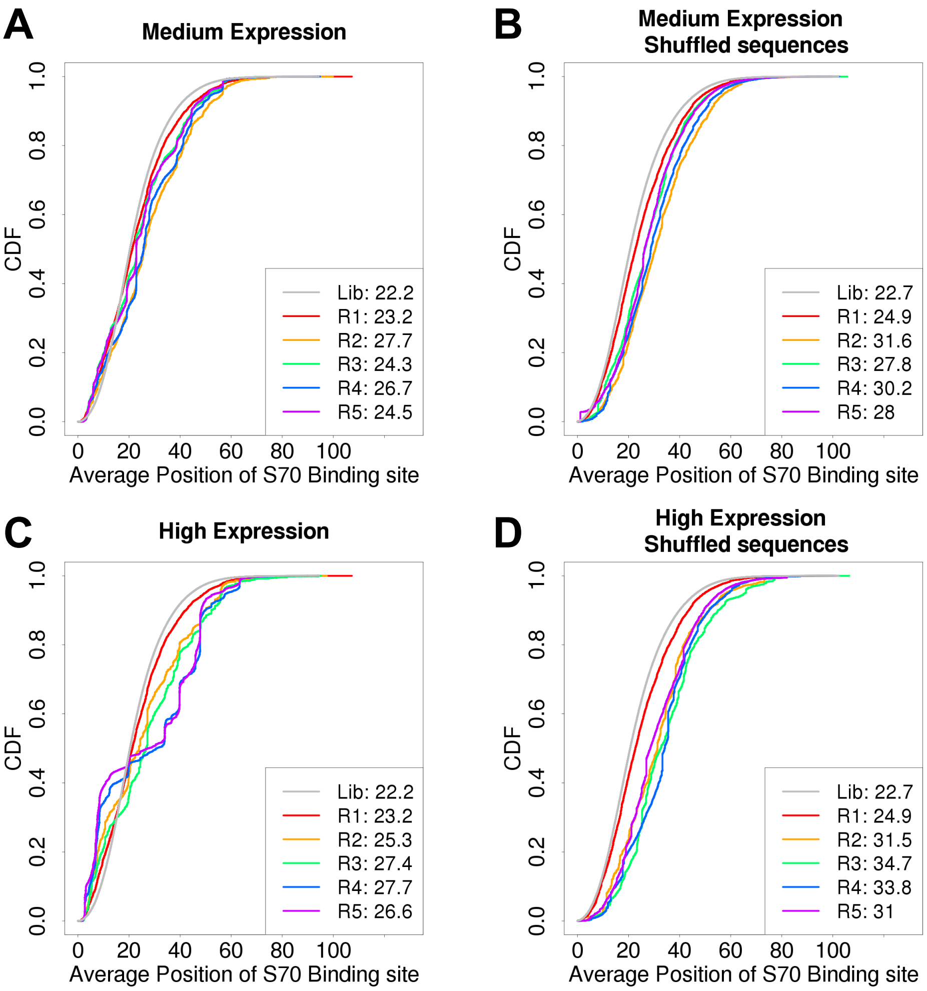

Figure 3.10: **Evolution of the average position of σ^{70} binding over evolutionary rounds.** Cumulative density functions of the average position in bp of σ^{70} binding upstream of the fixed promoter region are shown for all sequence libraries from the initial library to after the last round of selection. The values in the legend behind the name of the libraries gives the calculated mean average position in binding energies of the respective library. (A) and (C) show the distributions for the medium and high expressing line, (B) and (D) the distributions for the sequences from the same rounds, but shuffled.

observed between promoters do span only a small range. Few basepair mutations did not impact the noise levels fundamentally.

Sequence composition that goes along with sequence functionality are AT richness, and in particular avoidance of C nucleotides. Why there is a imbalance between incorporation of G's and C's remains not understood. High AT content in the region of σ^{70} decreases the melting temperature required for unwinding the doublestranded DNA. As the binding motif of σ^{70} is rich in % AT, promoters with higher AT content might also provide better binding of the sigma factor. The second feature, namely the length of the inserted promoter region, shows that longer promoter sequences have a higher chance of showing transcriptional activity. Both features give hints on what distinguishes a functional from a non-functional promoter sequence.

Another minimal requirement for a promoter sequence seems to be a sigma binding site, in the conditions synthetic promoters were evolved, σ^{70} appear as the main active sigma factor present in the cell. Thus, synthetic promoters exhibit similar binding strengths for σ^{70} as native promoter sequences that are known to be bound by σ^{70}. Expression levels towards higher levels cannot be explained by the binding strength. Besides σ^{70} binding strength, synthetic promoter sequences show characteristic features of the binding sites in terms of position relative to the translation start site as well as the entropy of binding strengths observed. Closeness of the binding site to the translation start and generally smaller entropies are observed in the transition from low to medium expression, but are not clear indicators for high expression levels. As we have shown in previous studies (Chapter 2), the expression activity measured reflects the transcriptional activity quite well.

In the transcription initiation process, besides binding of the RNA polymerase with its sigma subunit, two other processes play an important role. These involve unwinding as well as clearance of the promoter region (Revyakin et al., 2004). Unwinding is helped by interaction of the 1.2 σ^{70} region, but the interaction itself is unspecific around 4 bp upstream of the transcription start site (Bochkareva & Zenkin, 2013). Prediction of the unwinding capabilities of our promoters is not solved, but selection for low GC content might be a hint towards better unwinding for low melting temperatures. Promoter clearance is the last step of the transcription initiation process, allowing the RNA polymerase to transit from the bound into the unbound state, elongating the transcript until it's 3'end. This last step is known to fail from time to time, releasing unfinished RNA transcripts and at the same time occupying the region around the transcription start site (Hsu, 2002). As the competence depends on the promoter sequence, differences in expression levels between promoter sequences might well be explained by promoter clearance properties. Given that the strength of sigma binding correlates negatively with promoter clearance (Hsu, 2002), failing in prediction of expression levels based on sigma binding affinity may not be surprising. Strong binding may even be counterproductive, leading to many abortive transcripts. Closeness of the sigma binding site to the RBS and translation start site are also indicators that a short untranslated region at the 5'end facilitates transcription initiation or that it stabilizes the RNA in a way that helps translation initiation.

In previous studies (Liu & Libchaber, 2006, Brewster et al., 2012), expression levels were predicted based on the closeness to the σ^{70} consensus sequence. Here we reported binding affinities for a large number of sequences, but comparison of σ^{70} binding strength across different promoter backgrounds seems to represent a much more difficult and complex problem. Context of the binding site may have more far-reaching consequences on expression levels than the binding sites themselves.

3.4 Methods

3.4.1 Library preparation of promoter sequences for deep sequencing

Libraries from each generation after selection were sequenced on an Illumina HiSeq2000. Custom design of adapter Sequences allowed usage of restriction cut sites for ligation of the adapters. Adapter sequences were ordered as HPLC-purified oligonucleotides (Microsynth, Switzerland) with Phosphothioate Oligonucleotide bonds marked by '*', phosphorylated by T4-Polynucleotide Kinase (NEB, MA, USA) and annealed by a slow temperature decrease. 10^9 mol of each primer P5.1 (5'-A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*-T*C*T-3') and P5.2 (5'-T*C*G*A*GATCGGAAGAGCGTCGTGTAGGGAAA*G*A*G*-T*G*T-3') were incubated for 30 min at 37°C in 1x T4 DNA Ligase ligation buffer (NEB) with 12.5 units of T4-Polynucleotide Kinase. After a heat-inactivation step for 2 minutes at 95°C, the temperature was slowly ramped down to 12°C at a rate of 0.05°C/second. The same annealing procedure was applied to the other adapter pair P7.1 (5'-G*T*G*A*-CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T-3') and P7.1 (5'-G*A*T*C*G*A*-TCGGAAGAGCACACGTCTGAACTCCAG*T*C*A*C-3'), both primer pairs were mixed and adjusted to a concentration of 80 μM each. Plasmids from each library were prepared from overnight cultures and 300 ng of plasmids were amplified in a 20 cycle PCR reaction for enrichment of promoter sequences with oligonucleotides LigpUA66Pro-F1 (5'-CCTTTCGTCTTCACCTCGAG-3') and LigpUA66Pro-R1 (5'-CCTTCTTACATCCAGA-GGATCCC-3'). Afterwards PCR products between 100-300 bp were gel size-selected from a 2.5% agarose gel. After purification from the gel, PCR products were double-digested for 3 hours with XhoI and BamHI and then column-purified. Adapters with overhangs for the cut sites were added in a ligation reaction to the digested PCR products. The ligation reaction was performed at 20°C for 45 minutes in 1x Quick ligase buffer with 1.25 μl of Quick T4 DNA Ligase (NEB) with equimolar ratios of adapters and digested PCR product. Ligation products were again gel-purified and sizes selected between 100-350 bp. From 25 μl column-purified eluate, 10 μl were used in the Index PCR step. A 50 μl PCR reaction with Phusion$^{\text{TM}}$ High-Fidelity DNA Polymerase (NEB) was used to amplify ligation products with both adapters attached and to add specific barcodes to each library. Each reaction was supplemented with 1 μl primer IS4 (5'-AATGATACGGCGACCACCGAGATCTACACT-

CTTTCCCTACACGACGCTCTT-3') and the primer with the specific barcode. Barcode combinations were chosen based on considerations made in (Meyer & Kircher, 2010) like equal usage of lasers at each position. Python code for choosing those barcode sets could be found under `https://bioinf.eva.mpg.de/multiplex/`. The set of barcodes used with moderate secondary structure formation can be found in section 3.6. The program for the Indexing PCR consisted of an initial denaturation step at 98°C for 30 seconds, followed by 18 cycles with 10 seconds denaturation at 98°C, 20 seconds primer annealing at 60°C and 1 minute elongation at 72°C. The final elongation step was performed at 72°C for 10 minutes. A last gel-purification step with size-selection between 100-350 bp on a 3% agarose gel and column-purification were conducted. Each library was quality checked on a NanoDrop 2000 Spectrophotometer (Thermo Scientific, Switzerland) and libraries were pooled to a final concentration of 10 nM according to their individual concentrations estimated with a Qubit $^{\text{TM}}$Fluorometer (Invitrogen, CA, USA). 150 cycles with a modified version of the Sequencing Primer (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGAG-3') masking the XhoI cut side were used on the Illumina HiSeq 2000 Sequencer to obtain the promoter sequences.

3.4.2 Pre-processing of Sequencing reads

Raw Sequencing reads were quality-checked and only sequences where all nucleotides had Phred Quality Scores of at least 20 (99% Base Call Accuracy) were considered in further analyses. Different evolutionary rounds were separated by their assigned barcode. Sequences flanked by both restriction sites were considered, yielding a set of promoter sequences with lengths ranging from 30-141 bp.

3.4.3 Genotyping individual sequences by Sanger Sequencing

From the population of promoter constructs after round 3 and 5 of selection, single promoters were isolated by colony-picking. 479 of those colonies were collected into 96 well plates and their promoters were PCR-amplified and sequenced using Sanger-Sequencing (Sanger et al., 1977) with primers (5'-CCTTCTTACATCCAGAGGATCCC-3') and (5'-GGCTTCCCAACCTTACCAGAGG-3'). 378 of the clones sequenced were unique.

3.4.4 Phenotyping individual sequences by flow cytometry

For all clones that were sequenced, expression of each promoter population was measured using FACS (fluorescence activated cell Sorting) as described in the Methods section of Chapter 2.

3.4.5 Unification and clustering of sequences based on similarity

First, all promoter sequences obtained by deep sequencing where made unique (based on 100% similarity over the entire sequence length) with keeping the number of occurrences.

The second step involved clustering of all unique sequences based on their similarity. CD-HIT, a greedy incremental clustering algorithm (Fu et al., 2012, Li & Godzik, 2006) clustered all sequences. Parameters that were used to control similarity were -c 0.8 -n 5 -s 0.9 -aL 0.9 -aS 0.9, clustering sequences with at least 80% identity over the entire alignment length.

3.4.6 Prediction of binding sites for transcription factors in *E.coli*

Retrieval of Weight Matrices Weight matrices were calculated from sequences with experimentally supported binding sites in *E.coli* using the RegulonDB (Salgado et al., 2013). Potential transcription factor binding regions were extended by 5 bp and putative binding sites for each factor were calculated using Phylogibbs (Siddharthan et al., 2005). Obtained matrices were refined based on the input sequences using MotEvo (Arnold et al., 2012). This yielded 41 weight matrices for specific transcription factors in *E.coli* as well as binding site predictions for six sigma factors, with spacer lengths of 17 bp for all of them except σ^{54} which deviates from this common pattern.

***Ab-initio* TFBS prediction** Putative binding sites for TFs were predicted using MotEvo (Arnold et al., 2012), an algorithm that calculates the probability of a particular factor given its weight matrix to bind at a certain position in the sequence under a competition required for unbinding of the factor. Considering only the highest weight matrix score allows to compare sequences of varying lengths. All sites with a minimum posterior of 0 were included for extraction of the maximum weight matrix score.

3.4.7 Prediction of σ^{70} free binding energies

Free binding energies for binding of σ^{70} to individual sequences were calculated as described in (Blank et al., 2013), making use of annotated σ^{70} binding sites (Djordjevic, 2011) to infer a position weight matrix (Schneider, 1997). The total binding energy for each sequence was calculated by sliding the weight matrix in k_BT units over the entire sequence, summing up the weight matrix scores over all possible windows. The probability for each sequence window to be bound $\exp(-S_i) = \sum_s \exp(-S_{i,s})$ was calculated as the summed probability over all spacer lengths considered (15-19 base pairs). The sum of the individual window probabilities $\exp(-S_j) = \sum_i \exp(-S_i)$ is the probability for the sequence to be bound by a σ^{70} factor. The free binding energy for each sequence is $S = -log \sum_i \exp(-S_i)$. The probability for each sequence to be bound at position i in the sequence is $P_i = \exp(-(S_i - S))$. On average, the σ^{70} factor is bound at position $\langle i \rangle = \sum_i i * P_i = \sum_i i * \exp(-(S_i - S))$. The entropy $H = -\sum_i log(P_i) * P_i = \sum_i (S_i - S) * \exp(-(S_i - S))$ describes if a promoter provides few strong binding sites (small entropy) or many weak binding sites (large entropy).

3.5 Acknowledgements

We would like to thank Martin Kircher and Matthias Meyer for their great support on designing primers and barcodes used for deep sequencing of the promoter sequences. Thanks to Philippe Demougin for assistance in the Sequencing library quality control and Christian Beisel, Ina Nissen and Manuel Kohler for realization of the deep sequencing and data management. Thanks to Phil Arnold for modification of MotEvo to run it on a large number of sequences.

3.6 Supplementary information

3.6.1 Supplementary figures

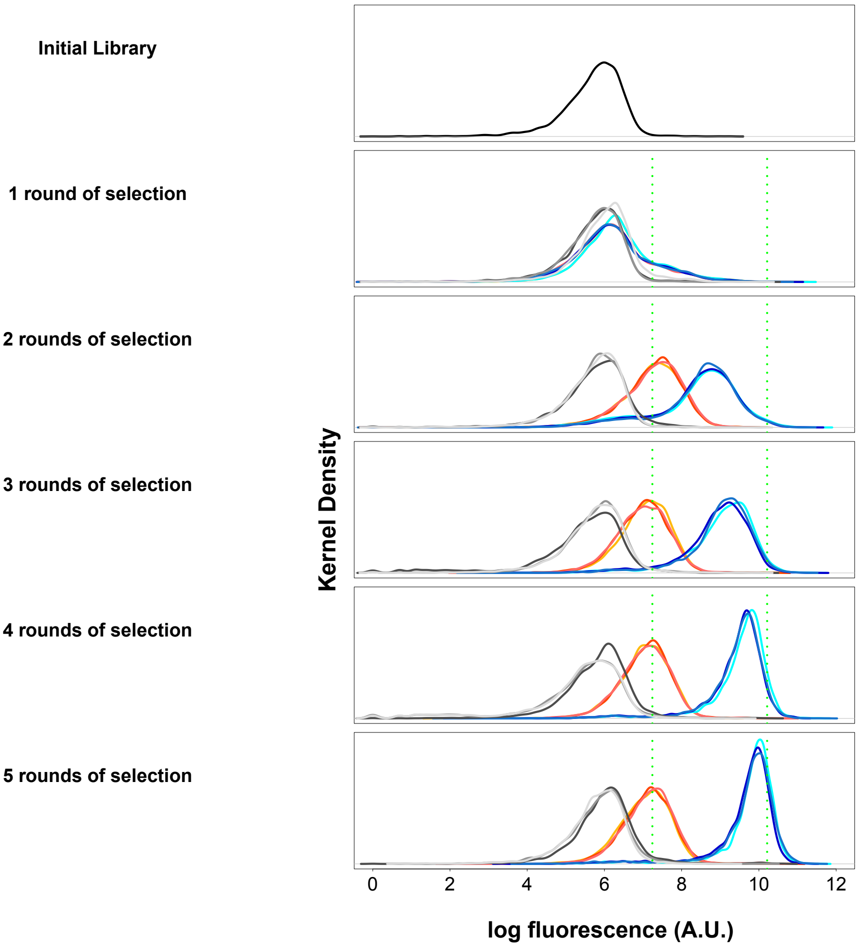

Figure S3.1: **Evolution of target expression values over rounds of evolution is highly replicable.** Target expression values are depicted with green dotted lines, based on the expression of the native *E.coli* reference genes *gyrB* for medium and *rpmB* for high expression. The fluorescence distribution of cells from the initial library is drawn in black, with grey lines in later rounds from populations that were not selected based on their expression, but subject to mutation. The three biological replicate lines for medium expression are drawn in red colors, for high expression in blue colors.

3.6.2 Set of barcodes for multiplexed deep sequencing

Table S3.1: Set of barcode sequences chosen for deep sequencing

Index	Name	Oligonucleotide
AACCAG	index_6nt_1	CAAGCAGAAGACGGCATACGAGATctggttGTGACTGGAGTTCAGACGTGT
AAGACT	index_6nt_2	CAAGCAGAAGACGGCATACGAGATagtcttGTGACTGGAGTTCAGACGTGT
AAGGAC	index_6nt_3	CAAGCAGAAGACGGCATACGAGATgtccttGTGACTGGAGTTCAGACGTGT
AATGCG	index_6nt_4	CAAGCAGAAGACGGCATACGAGATcgcattGTGACTGGAGTTCAGACGTGT
ACCGAT	index_6nt_5	CAAGCAGAAGACGGCATACGAGATatcggtGTGACTGGAGTTCAGACGTGT
ACGCGT	index_6nt_7	CAAGCAGAAGACGGCATACGAGATacgcgtGTGACTGGAGTTCAGACGTGT
ACTCTC	index_6nt_9	CAAGCAGAAGACGGCATACGAGATgagagtGTGACTGGAGTTCAGACGTGT
ACTTCA	index_6nt_10	CAAGCAGAAGACGGCATACGAGATtgaagtGTGACTGGAGTTCAGACGTGT
AGACCG	index_6nt_11	CAAGCAGAAGACGGCATACGAGATcggtctGTGACTGGAGTTCAGACGTGT
AGAGTT	index_6nt_12	CAAGCAGAAGACGGCATACGAGATaactctGTGACTGGAGTTCAGACGTGT
AGATAC	index_6nt_13	CAAGCAGAAGACGGCATACGAGATgtatctGTGACTGGAGTTCAGACGTGT
AGCATA	index_6nt_14	CAAGCAGAAGACGGCATACGAGATtatgctGTGACTGGAGTTCAGACGTGT
AGGTTG	index_6nt_17	CAAGCAGAAGACGGCATACGAGATcaacctGTGACTGGAGTTCAGACGTGT
AGTACC	index_6nt_18	CAAGCAGAAGACGGCATACGAGATggtactGTGACTGGAGTTCAGACGTGT
ATAATG	index_6nt_19	CAAGCAGAAGACGGCATACGAGATcattatGTGACTGGAGTTCAGACGTGT
ATACGC	index_6nt_20	CAAGCAGAAGACGGCATACGAGATgcgtatGTGACTGGAGTTCAGACGTGT
ATATCT	index_6nt_21	CAAGCAGAAGACGGCATACGAGATagatatGTGACTGGAGTTCAGACGTGT
ATCAGT	index_6nt_22	CAAGCAGAAGACGGCATACGAGATactgatGTGACTGGAGTTCAGACGTGT
ATTCAT	index_6nt_23	CAAGCAGAAGACGGCATACGAGATatgaatGTGACTGGAGTTCAGACGTGT
ATTGGA	index_6nt_24	CAAGCAGAAGACGGCATACGAGATtccaatGTGACTGGAGTTCAGACGTGT
CAACCT	index_6nt_25	CAAGCAGAAGACGGCATACGAGATaggttgGTGACTGGAGTTCAGACGTGT
CAAGTA	index_6nt_26	CAAGCAGAAGACGGCATACGAGATtacttgGTGACTGGAGTTCAGACGTGT
CAATAG	index_6nt_27	CAAGCAGAAGACGGCATACGAGATctattgGTGACTGGAGTTCAGACGTGT
CAGATG	index_6nt_28	CAAGCAGAAGACGGCATACGAGATcatctgGTGACTGGAGTTCAGACGTGT
CAGCGC	index_6nt_29	CAAGCAGAAGACGGCATACGAGATgcgctgGTGACTGGAGTTCAGACGTGT
CATTCC	index_6nt_30	CAAGCAGAAGACGGCATACGAGATggaatgGTGACTGGAGTTCAGACGTGT
CCAACG	index_6nt_31	CAAGCAGAAGACGGCATACGAGATcgttggGTGACTGGAGTTCAGACGTGT
CCAGAC	index_6nt_32	CAAGCAGAAGACGGCATACGAGATgtctggGTGACTGGAGTTCAGACGTGT
CCATGA	index_6nt_33	CAAGCAGAAGACGGCATACGAGATtcatggGTGACTGGAGTTCAGACGTGT
CCGAAT	index_6nt_34	CAAGCAGAAGACGGCATACGAGATattcggGTGACTGGAGTTCAGACGTGT
CCGCCA	index_6nt_35	CAAGCAGAAGACGGCATACGAGATtggcggGTGACTGGAGTTCAGACGTGT
CCGTTC	index_6nt_36	CAAGCAGAAGACGGCATACGAGATgaacggGTGACTGGAGTTCAGACGTGT
CCTAGC	index_6nt_37	CAAGCAGAAGACGGCATACGAGATgctaggGTGACTGGAGTTCAGACGTGT
CCTCAG	index_6nt_38	CAAGCAGAAGACGGCATACGAGATctgaggGTGACTGGAGTTCAGACGTGT
CCTGCT	index_6nt_39	CAAGCAGAAGACGGCATACGAGATagcaggGTGACTGGAGTTCAGACGTGT
CGACTC	index_6nt_40	CAAGCAGAAGACGGCATACGAGATgagtcgGTGACTGGAGTTCAGACGTGT
CGCAAC	index_6nt_41	CAAGCAGAAGACGGCATACGAGATgttgcgGTGACTGGAGTTCAGACGTGT
CGCCGT	index_6nt_42	CAAGCAGAAGACGGCATACGAGATacggcgGTGACTGGAGTTCAGACGTGT
CGCTCG	index_6nt_43	CAAGCAGAAGACGGCATACGAGATcgagcgGTGACTGGAGTTCAGACGTGT
CGTATT	index_6nt_45	CAAGCAGAAGACGGCATACGAGATaatacgGTGACTGGAGTTCAGACGTGT
CTAGGT	index_6nt_46	CAAGCAGAAGACGGCATACGAGATacctagGTGACTGGAGTTCAGACGTGT
CTCCTG	index_6nt_47	CAAGCAGAAGACGGCATACGAGATcaggagGTGACTGGAGTTCAGACGTGT
CTCGAA	index_6nt_48	CAAGCAGAAGACGGCATACGAGATttcgagGTGACTGGAGTTCAGACGTGT
CTCTGC	index_6nt_49	CAAGCAGAAGACGGCATACGAGATgcagagGTGACTGGAGTTCAGACGTGT
CTGACC	index_6nt_50	CAAGCAGAAGACGGCATACGAGATggtcagGTGACTGGAGTTCAGACGTGT
GAAGCC	index_6nt_52	CAAGCAGAAGACGGCATACGAGATggcttcGTGACTGGAGTTCAGACGTGT

Continued on next page

Table S3.1 – *Continued from previous page*

Index	**Name**	**Oligonucleotide**
GACCTT	index_6nt_53	CAAGCAGAAGACGGCATACGAGATaaggtcGTGACTGGAGTTCAGACGTGT
GATAAC	index_6nt_55	CAAGCAGAAGACGGCATACGAGATgttatcGTGACTGGAGTTCAGACGTGT
GCAATC	index_6nt_56	CAAGCAGAAGACGGCATACGAGATgattgcGTGACTGGAGTTCAGACGTGT
GCCTCT	index_6nt_57	CAAGCAGAAGACGGCATACGAGATagaggcGTGACTGGAGTTCAGACGTGT
GCGCTG	index_6nt_58	CAAGCAGAAGACGGCATACGAGATcagcgcGTGACTGGAGTTCAGACGTGT
GCTGAA	index_6nt_59	CAAGCAGAAGACGGCATACGAGATttcagcGTGACTGGAGTTCAGACGTGT
GTCGCG	index_6nt_64	CAAGCAGAAGACGGCATACGAGATcgcgacGTGACTGGAGTTCAGACGTGT
TAAGAT	index_6nt_65	CAAGCAGAAGACGGCATACGAGATatcttaGTGACTGGAGTTCAGACGTGT
TATCGT	index_6nt_66	CAAGCAGAAGACGGCATACGAGATacgataGTGACTGGAGTTCAGACGTGT
TCATTG	index_6nt_67	CAAGCAGAAGACGGCATACGAGATcaatgaGTGACTGGAGTTCAGACGTGT
TCCTAC	index_6nt_68	CAAGCAGAAGACGGCATACGAGATgtaggaGTGACTGGAGTTCAGACGTGT
TCTATA	index_6nt_70	CAAGCAGAAGACGGCATACGAGATtatagaGTGACTGGAGTTCAGACGTGT
TTACTT	index_6nt_74	CAAGCAGAAGACGGCATACGAGATaagtaaGTGACTGGAGTTCAGACGTGT
TTCCGA	index_6nt_75	CAAGCAGAAGACGGCATACGAGATtcggaaGTGACTGGAGTTCAGACGTGT
TTCGTC	index_6nt_76	CAAGCAGAAGACGGCATACGAGATgacgaaGTGACTGGAGTTCAGACGTGT

Chapter 4

The predictability of molecular evolution during functional innovation

Published in PNAS, February 25, 2014, vol. 111, no. 8, 3044–3049

Diana Blank[a,1], Luise Wolf[a,1], Martin Ackermann[b,c], and Olin K. Silander[a,2]

a Computational and Systems Biology, Biozentrum, University of Basel, 4056 Basel, Switzerland;

b Department of Environmental Systems Science, ETH Zürich, 8092 Zürich, Switzerland;

and c Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland

Edited by Richard E. Lenski, Michigan State University, East Lansing, MI, and approved January 7, 2014 (received for review October 9, 2013)

Author contributions: O.K.S. designed research; D.B., L.W., and O.K.S. performed research; L.W. and M.A. contributed new reagents/analytic tools; O.K.S. analyzed data; and M.A. and O.K.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession no. $PRJNA$234028).

1 D.B. and L.W. contributed equally to this work.

2 To whom correspondence should be addressed. E-mail: olinsilander@gmail.com.

This article contains supporting information online at
`www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318797111/-/DCSupplemental`.

adaptation — transcription — compensatory mutation — biosynthesis

4.1 Abstract

Determining the molecular changes that give rise to functional innovations is a major unresolved problem in biology. The paucity of examples has served as a significant hindrance in furthering our understanding of this process. Here we used experimental evolution with the bacterium *Escherichia coli* to quantify the molecular changes underlying functional innovation in 68 independent instances ranging over 22 different metabolic functions. Using whole-genome sequencing, we show that the relative contribution of regulatory and structural mutations depends on the cellular context of the metabolic function. In addition, we find that regulatory mutations affect genes that act in pathways relevant to the novel function, whereas structural mutations affect genes that act in unrelated pathways. Finally, we use population genetic modeling to show that the relative contributions of regulatory and structural mutations during functional innovation may be affected by population size. These results provide a predictive framework for the molecular basis of evolutionary innovation, which is essential for anticipating future evolutionary trajectories in the face of rapid environmental change.

4.2 Significance

Understanding the genetic changes that underlie phenotypic functional innovations is a fundamental goal in evolutionary biology, giving insight into species' past, present, and future evolutionary trajectories. One important unresolved question is whether such genetic changes typically affect protein expression or protein structure. Here we use large-scale laboratory evolution with bacteria to quantify the types of genetic changes that occur during functional innovation. We show that whether these changes affect protein expression or protein structure depends on which cellular functions are being selected upon. We then show that changes affecting protein expression occur in qualitatively different sets of genes from changes affecting protein structure. These results show that using functional knowledge it is possible to predict the course of evolution.

4.3 Introduction

One of the most important questions in evolutionary biology concerns the molecular mechanisms that underlie functional innovations. These changes are often polarized into two classes: those that affect protein structure and those that affect protein expression level. Both of these classes have been shown to play important roles across a wide range of taxa, from vertebrates (Jones et al., 2012, Zhang et al., 2002) to bacteria (Ando et al., 2013, Lieberman et al., 2011), and their relative importance has been the topic of considerable discussion (Wray, 2007, Hoekstra & Coyne, 2007, Stern & Orgogozo, 2008, Jacob, 1977, Mutero et al., 1994, Hoekstra et al., 2006, Shapiro et al., 2004, Stern & Orgogozo, 2009). Significantly, many previous studies have addressed these questions by focusing on single instances of functional innovation (Kasak et al., 1997, Dabizzi et al., 2001, Brilli & Fani, 2004, Lin & Hacking, 1976) or selective regimes (Blount et al., 2012, Meyer et al., 2012, Beaumont et al., 2009, Lindsey et al., 2013, Barrick et al., 2009, Tenaillon et al., 2012). However, to identify general principles, it is necessary to study evolutionary innovation for a large number of different functions in parallel. Indeed, the fact that only a small number of examples exist has resulted in few hypotheses being put forth that identify general characteristics of the molecular changes underlying functional innovation. One prominent hypothesis states that if the development of a novel trait is spatially or temporally limited, then innovation frequently occurs through changes in regulation (Haag & Lenski, 2011, Stern, 2000). Whether there are general patterns beyond this is not well-established.

Here we used an experimental system that allows the analysis of a large number of independent cases of evolutionary innovation and investigation of the underlying genetic changes. We worked with a collection of 87 strains of *Escherichia coli* that each had a deletion of one gene encoding a different metabolic function (SI Appendix, Table S4.1). Each of these deletions resulted in an inability to grow in minimal glucose media. Then, for each of these 87 deleted metabolic functions, we used experimental evolution to select for novel functionality that could replace the functionality that was lost through gene deletion. As an example, one of the deleted genes was *serB*, a phosphoserine phosphatase that catalyzes the final step in serine biosynthesis. Regaining the ability to grow in the absence of this gene requires the evolution of a new function that allows sufficient amounts of serine to be made to support cell growth. Although this experimental system does not necessarily recapitulate natural evolutionary scenarios, many aspects of this design are reflective of ecological and evolutionary features found in more natural circumstances. For example, the loss of specific genes or functions may occur through drift in small populations or through selection in the face of antagonistic pleiotropy (Elena & Lenski, 2003). One well-established example of this in *E. coli* is that many natural isolates have null alleles at the locus for the stress response sigma factor *rpoS*; this is thought to be due to tradeoff between stress resistance and growth rate (Ferenci, 2003). Similarly, the experimental design here provides an evolutionary scenario very similar to that experienced by microbes during the evolution of abilities that allow the degradation of nonnatural compounds, such as organic chlorides (Bosma et al., 2002) (SI Appendix).

Our experimental design provides two significant advantages that cannot easily be realized in settings outside of the laboratory. First, we can study how a large number of different types of metabolic functions arise, and thus gain general insights into the process of evolutionary innovation. Second, we have information on the cellular context of the evolved novel metabolic functionality, and can thus analyze whether the mechanisms of evolutionary innovation depend in a predictable manner on the specific characteristics of the pathway or interaction networks in which the function acts (Stern & Orgogozo, 2009).

4.4 Results and Discussion

We began the experimental evolution by establishing five replicate cultures of each of the 87 deletion genotypes, yielding a total of 435 independent cultures. We grew large populations of cells in rich media, and used serial transfer in glucose minimal media to evolve these populations for 28 transfers, or approximately 145 generations (Materials and Methods). To ensure the transfer of cells that evolved even low levels of novel metabolic activity, we severely limited the rate of serial dilution, imposing only 5-fold dilutions for the first 10 transfers and 100-fold dilutions for the following 18 transfers. We allowed 48 h of growth in-between transfers.

At the end of this period, 68 out of the 435 populations recovered the ability to grow in glucose minimal media. These 68 populations encompassed 22 out of the 87 different deletion genotypes, and the functions encoded by these 22 deleted genes were distributed throughout the *E. coli* metabolic network (SI Appendix, Fig. S4.1). For a small number of deletion genotypes, all five replicate populations regained the ability to grow. However, for the majority of deletion genotypes in which growth recovered, between one and four replicate populations regained growth ability (Fig. 4.1A). The pattern of recovery that we observed was consistent with a scenario in which a small number of functions are easy to recover, whereas a much larger number of functions are difficult or perhaps impossible to evolve. Under such a scenario, it is possible that increases in the population size or mutation rate might result in more functions being recovered.

We found, furthermore, that the probability of growth recovery depended on the metabolic function that had been deleted. We classified the 87 deleted genes as acting in one or more of four metabolic functional categories (Serres & Riley, 2000): carbon compound utilization, energy metabolism, central intermediary metabolism, and building block biosynthesis. The functions of proteins that acted in building block biosynthesis were much less likely to be replaced than the other types of metabolic functions (Fisher's exact test, $P = 0.0002$, odds ratio 0.09; Fig. 4.1B), indicating that new functionality in building block biosynthesis was more difficult to evolve than other types of new functionality.

We selected one clone from each of the 68 populations in which growth recovered and determined its maximum growth rate. Fifty-seven of these clones exhibited detectable growth as assayed by changes in optical density (Fig. 4.2 and SI Appendix, Materials and Methods); in the remaining 11 populations, growth could be detected only as changes in colony-forming units over time. Although many clones exhibited growth rates similar to that of the wild type, lag times were considerably longer (SI Appendix, Fig. S4.2), suggesting that in most cases the functionality of the deleted gene had not been fully replaced. Notably, we found a striking level of parallelism in growth rates, with clones from replicate populations evolving similar growth rates (Fig. 4.2). This set of 68 clones comprises a large set of independent instances in which functional innovation has evolved to confer novel growth abilities. We propose that this provides a model system for investigating the molecular mechanisms underlying the

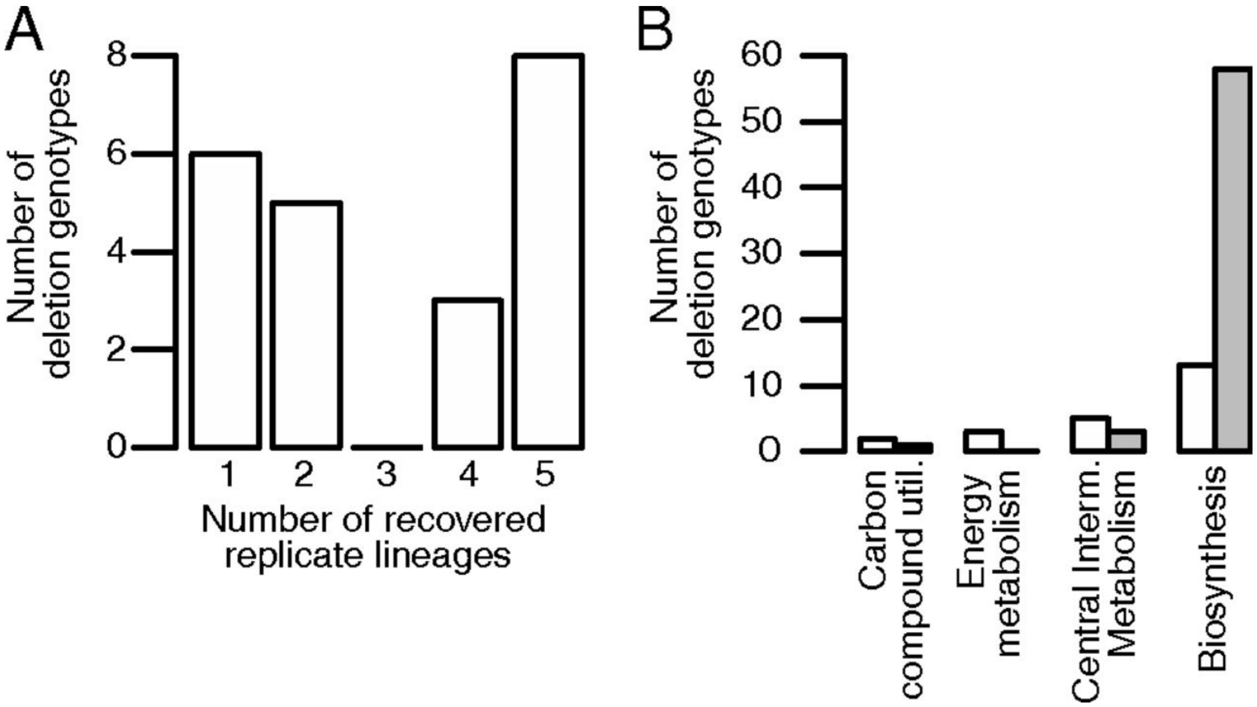

Figure 4.1: **Sixty-eight out of 435 populations evolved the ability to compensate for the function of the deleted gene.** (A) Growth recovery was not deterministic. For some deletion genotypes, five out of five replicates recovered growth; for the majority, between one and four replicates recovered growth. Three hundred sixty-seven populations went extinct during the evolutionary process; these are not shown. (B) Novel functions that were related to building block biosynthesis were more difficult to evolve. The white bars indicate those deletion genotypes in which novel functionality evolved; in gray are those for which no novel function was evolved. For all categories except building block biosynthesis, novel functions evolved that compensated for the majority of deleted functions.

evolution of new functionality and for identifying whether the genetic changes are predictably regulatory or structural in their nature.

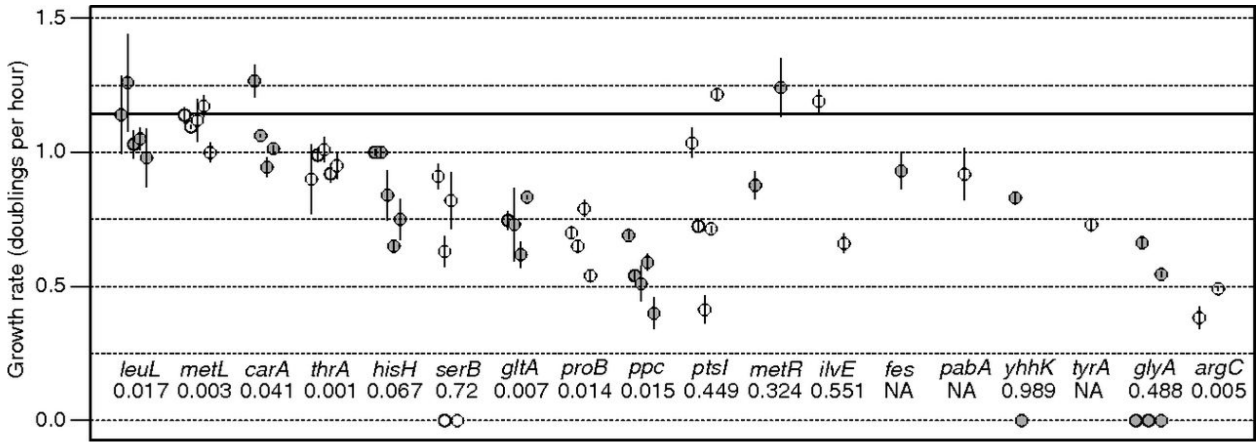

Figure 4.2: **Growth rates of recovered clones were more similar for lineages derived from the same deletion genotypes.** Each point shows the mean estimated doublings per hour for each clone (± SEM); clones are grouped by genotype (x axis) and colored gray and white to emphasize the groupings. The black solid line indicates the growth rate of the ancestral BW25113 strain. Six recovered clones did not exhibit detectable growth as assayed by OD_{600} and are indicated as having growth rates of 0. For four deletion genotypes, no clones exhibit detectable growth as assayed by OD_{600}; these deletion genotypes are not shown here. The number below each genotype indicates the probability of observing a set of clones with growth rates at least as clustered as those that we observe (SI Appendix). Cases in which only one lineage recovered for a genotype are indicated with NA, as no clustering probability could be calculated. Each point is based on three biological replicates, except for 11 cases in which one replicate was excluded due to no growth being observed (SI Appendix).

We determined the genetic changes that occurred during experimental evolution using whole-genome sequencing. Using the same set of 68 clones used in the quantification of growth rates, we identified 238 genomic changes in total (SI Appendix and Dataset S1). We focus here on those mutations that are most likely to provide phenotype-specific novel functionality. For this reason, we excluded from all subsequent analyses those mutations that have been observed in other laboratory evolution studies or in the ancestral deletion genotypes (these are likely to be general laboratory adaptations; SI Appendix, Tables S4.2 and S4.3). Excluding this class yielded a total of 210 mutations (Fig. 4.3A and 4.3B) that may have specifically been responsible for novel functionality, compensating for the role of the deleted genes. We term the genes affected by these mutations the “recruited” genes (Dataset S1), and propose that by changing protein expression or structure they confer crucial functional innovations that are necessary for growth recovery on minimal media. Notably, we found that clones isolated from populations in which a large number of replicate lineages recovered contained fewer mutations (Spearman’s rho $= -0.38$, $P = 0.001$; Fig. 4.3C). One interpretation of this observation is that novel functions that required fewer mutational steps evolved with higher probability. Alternatively, it is possible that certain types of deletions lead to increased mutation rates, and as a consequence to a higher number of fixed mutations.

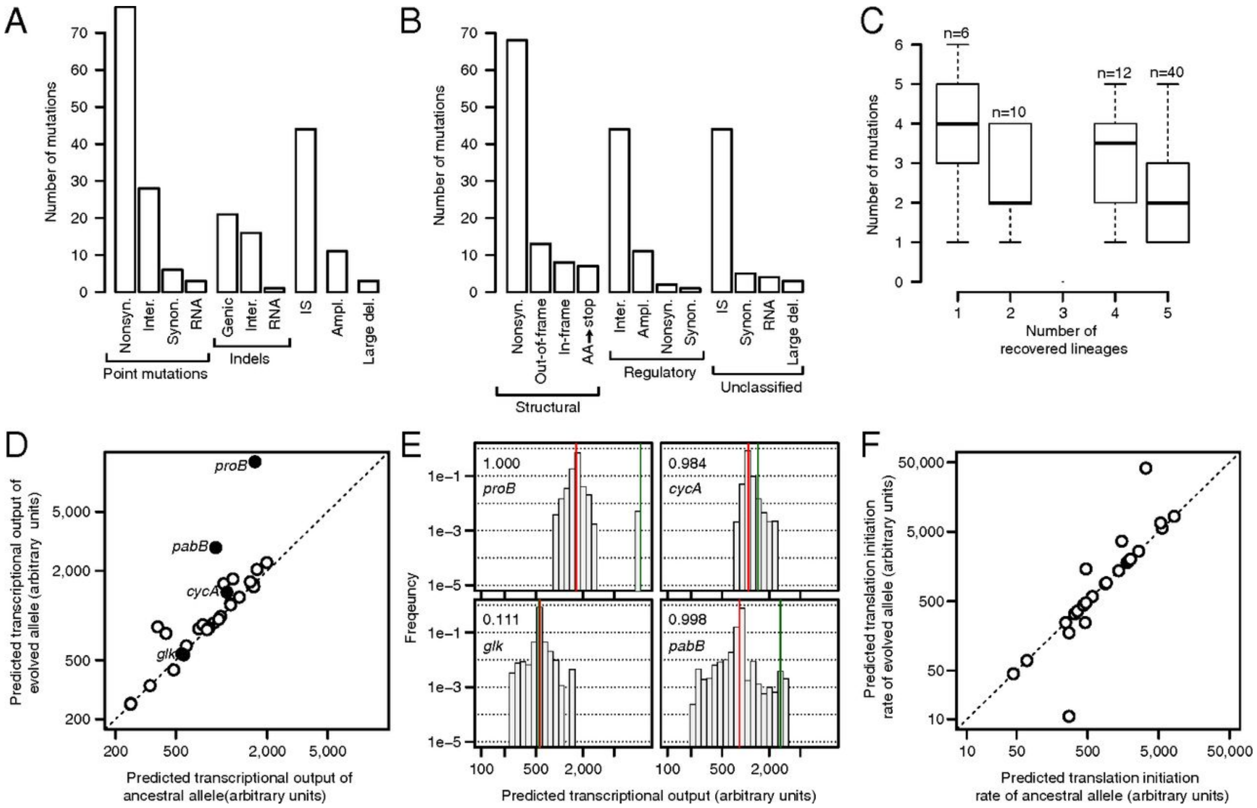

Figure 4.3: **Mutational events that occurred during the evolution of novel functionality affected both protein structure and expression level.** (A) Mutations classified by type. The overwhelming majority of changes were point mutations (Left), followed by insertions sequence (IS) element-mediated changes, small indels, large amplifications (larger than 200 bp), and large deletions (larger than 200 bp). (B) Mutations classified by functional effect (SI Appendix). We inferred that the majority of changes affected protein structure, although more than a fifth of these resulted in altered reading frames or the incorporation of premature stop codons. Almost 40% of changes were inferred to be regulatory, that is, as directly affecting protein expression (in contrast to indirect effects, which may occur via structural changes in transcription factors or other mechanisms). (C) Deletion genotypes in which few replicates recovered tended to contain clones with more mutations. Each box shows the numbers of mutations found within clones, classified by the number of replicate lineages that recovered (e.g., for three deletion genotypes, four replicate lineages recovered; the number of mutations in each of the 12 sequenced clones is shown). The boxplots indicate the median, first and third quartiles, and the extreme values within the category. (D) Mutations in evolved clones often increased predicted transcriptional output due to changes in σ^{70} binding. For each unique intergenic mutation, we predicted the transcriptional output for the ancestral sequence and the evolved sequence (SI Appendix, Materials and Methods). The dotted black line indicates unchanged transcriptional output. The annotated black points are the promoters shown in E. (E) Random mutations that result in increased transcriptional output are rare. We predicted transcription (σ^{70} binding) for all point mutations and 1-bp indels in the promoter region surrounding the intergenic mutations plotted in D. Four examples are shown here. The predicted transcriptional output of the ancestor is shown as a red line; that of the same promoter region with the evolved mutation is shown as a green line. Most random mutations have little effect on transcription; however, in several of the evolved clones, the observed mutation was among those mutations with the largest possible predicted effect on transcription (SI Appendix, Fig. S4.4). Clockwise from the top left, the deletion genotypes and recruited genes are $\Delta argC$ and *proB*; $\Delta glyA$ and *cycA*; $\Delta pabA$ and *pabB*; and $\Delta ptsI$ and *glk*. The numbers in the top left of each panel indicate the fraction of all one-mutant neighboring promoters that have a predicted transcriptional output that is equal to or lower than the observed mutant. Note that both the x and y axes are on a log scale. (F) Mutations that affect translation both increase and decrease the predicted translation initiation rate. Translation initiation rates were predicted using a biophysical model (Salis et al., 2009) for the ancestral and derived alleles for all intergenic mutations, with the black dotted line indicating no change.

The relative numbers of mutational types that we observed suggested that the vast majority were positively selected. Within coding regions, the ratio of nonsynonymous mutations per nonsynonymous site to synonymous mutations per synonymous site (Ka/Ks) was 6.0 (4.3A and SI Appendix, Materials and Methods); the analogous ratio for intergenic sites (Ki/Ks) was 11.8 (Fig. 4.3A). In both cases, if the two classes of sites had evolved neutrally, the expected ratio is 1. This suggested that there was pervasive strong positive selection for point mutations that caused amino acid changes, which likely lead to structural changes in proteins (SI Appendix, S4.6). Furthermore, this provides evidence that positive selection for point mutations at intergenic sites was even more prevalent; such mutations are likely to lead to regulatory changes. We also found evidence for positive selection on other types of mutations that likely affect protein levels through regulatory changes, including a 3.5-fold enrichment of indels in intergenic regions compared with coding regions, four mutations in RNA molecules (two small RNAs, and two in the tRNA-processing *rnpB*), amplifications, and transposon insertions (Aronson et al., 1989) (Fig. 4.3A and 4.3B and SI Appendix, Fig. S4.7, and Table S4.4). Finally, we found that parallel changes were common. Certain amplifications occurred in all clones that recovered for certain deletion genotypes (SI Appendix, Fig. S4.3). In some cases, these amplifications increased the genome size by more than 106 bp (approximately 25% of the genome), a change that is likely to be deleterious unless mitigated by considerable beneficial effects. Such extensive parallelism was also observed for point mutations. For example, all four lineages that recovered functionality for a deletion of *carA* contained a mutation 16 bp upstream of the start codon of *carB*, a mutation that is likely to increase protein expression level by affecting translation initiation (Fig. 4.4).

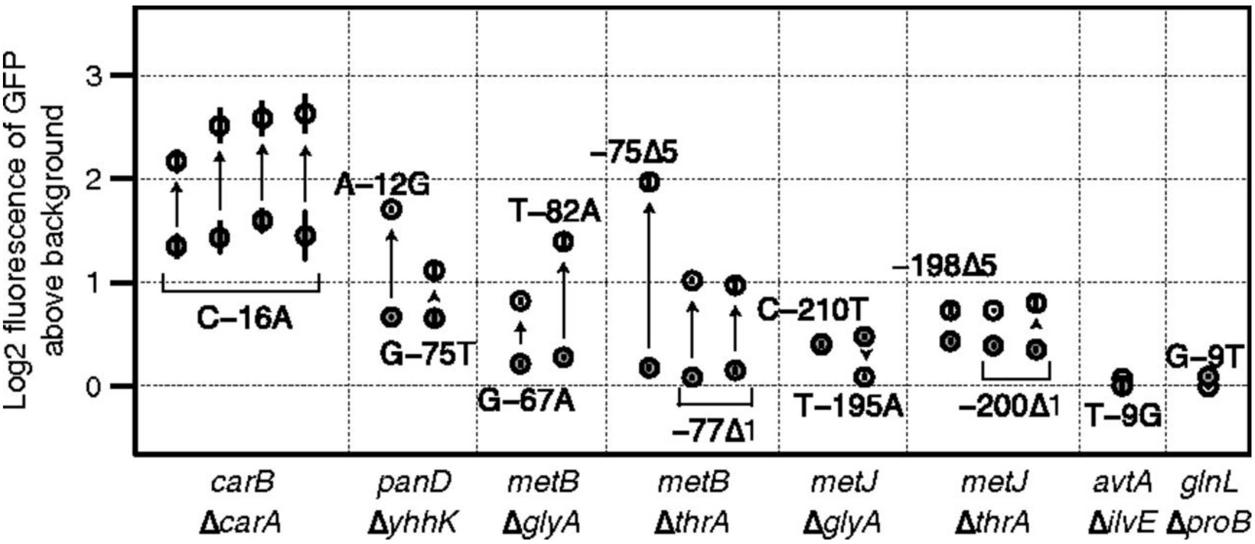

Figure 4.4: (Continued on the following page.)

Figure 4.4: **Intergenic mutations confer only moderate changes in protein expression.** Mean fluorescence levels (± SEM) conferred by chromosomal copies of intergenic regions containing the ancestral (gray points) and evolved alleles (white points). Each pair of alleles is annotated with the mutational change that occurred, with the number indicating the position, in base pairs, from the first base pair of the downstream ORF. The x axis is annotated with the recruited genes whose promoters were affected by the mutation (first row) and the deletion genotype in which the mutation arose (second row). The ORFs of both *metJ* and *metB* are downstream of a single intergenic region (in opposite directions). Thus, the sequence contained in these constructs is identical, but GFP expression is driven by promoters on opposite strands. The arrows emphasize the direction of expression change. We predicted significant expression changes in *carB*, *panD* A-12G, *avtA*, and *glnL* of 12.4-, 2.7-, 2.4-, and 0.64-fold, respectively. For all other genotypes, we predicted no significant changes based on changes in σ^{70} binding or changes in ribosome binding. Note that the sensitivity of the assay means that very low expression levels (i.e., the *avtA* and *glnL* alleles) cannot be accurately measured. Thus, the fold change in expression, particularly for these strains, is likely larger than what we measured.

These genomic analyses yielded the following insights into the relative contribution of structural and regulatory changes during the early stages of functional innovation in bacteria: We found that during the early stages of functional innovation, structural mutations are more common. Sixty-one percent of all observed point mutations led to amino acid changes, and thus likely to structural change (Fig. 4.3B). Although regulatory mutations were less common, they were strongly overrepresented. Whereas only 12.2% of the genome of *E. coli* is intergenic, 25% of the point mutations and 43% of all indels were located in intergenic sites. This means that mutations that occurred in intergenic regions, and which potentially change protein expression levels, were approximately three times more likely to increase in frequency compared with mutations in coding regions. To further test the adaptive nature of these potential regulatory mutations and to understand their molecular consequences, we examined how they affected transcriptional and posttranscriptional processes.

Computational analyses showed that many of these potential regulatory mutations increased transcriptional output or translation initiation rates. First, we analyzed changes in transcription using an approach based on information theory (Schneider, 1997, Berg & von Hippel, 1987, Shultzaberger et al., 2010) to predict how the regulatory mutations (point mutations and indels) affected binding of the housekeeping sigma factor σ^{70} (SI Appendix, Materials and Methods). In 11 out of 35 cases (30.5%), we found that the mutation increased σ^{70} binding strength, by 1.1- to 6.5-fold (Fig. 4.3D and 4.3E and Materials and Methods). This fraction is much higher than what is expected by chance: Only 3.6% of all random mutations are predicted to increase binding by more than 10% (Fig. 4.3E and SI Appendix, Fig. S4.4 and Table S4.6)). As binding is directly proportional to transcriptional output (Brewster et al., 2012), this supports the hypothesis that many of the intergenic mutations resulted in increased protein expression levels. We next used a biophysical model to predict changes in translation initiation rates. Of the eight mutations that occurred within 35 bp upstream of the start codon, and which may have affected translation (Salis et al., 2009), four were predicted to increase translation initiation, whereas three were predicted to decrease it (Fig. 4.3E). This suggests that mutations affecting protein expression often incurred their beneficial effects through increasing transcriptional output, and less often through increasing the rate of translation initiation.

We then experimentally quantified the effects of these mutations on protein expression. We placed single copies of the intergenic region containing the ancestral and evolved alleles as translational fusions with GFP in a neutral location in the chromosome (McKenzie & Craig, 2006) (Materials and Methods). In the four cases for which we also computationally predicted a change in protein level, expression changed in the direction anticipated. Surprisingly, we found that in many cases, expression levels increased by only moderate levels (Fig. 4.4). Despite these weak effects on protein expression level, as pointed out above, there was strong evidence that many of these mutations were adaptive. Together, these data suggested that the regulatory mutations that occurred during the evolution of novel functionality often conferred adaptive benefits through the creation of novel regulatory elements, but that these elements frequently effected only moderate changes in protein expression level.

To test whether changes in protein expression alone could provide novel functionality that rescues the lethal phenotypes and to understand the extent of expression change needed, we selected eight deletion genotypes (SI Appendix, Materials and Methods) in which several evolved clones appeared to have single large-effect regulatory mutations, suggesting that the lethal phenotypes could be rescued by increased expression of a single ORF; this criterion largely excluded evolved clones with structural changes that were highly paralleled in other lineages and evolved clones that contained two or more regulatory changes. We then transformed the eight ancestral deletion genotypes with plasmids containing the ORFs of the recruited gene under an Isopropyl $\beta - D - 1-$thiogalactopyranoside (IPTG)$-$regulated promoter (Kitagawa et al., 2005) (Materials and Methods) and observed growth for 72 h. In four out of the eight cases, high growth rates were observed when expression of the ORF was induced using 100μM IPTG (SI Appendix, Fig. S4.5), showing that increased expression of this single ORF was sufficient to provide the new functionality. In three cases, growth rates displayed threshold behavior, with high growth rates occurring at one concentration of IPTG and small decreases in IPTG level resulting in little or no growth at all. This suggests that subtle changes in protein expression level can have profound effects on growth (Ando et al., 2013), a result that provides an interesting corollary to previous data showing that small changes in protein structure can exert large effects on cell growth (Walkiewicz et al., 2012).

Our results indicate that structural and regulatory changes are both important for the evolution of new functions, but show that regulatory mutations are consistently more likely to be positively selected. We next sought to develop a more refined predictive framework to understand how the genetic changes that occur during evolutionary innovation depend on functional or demographic parameters. First, we asked whether the relative importance of structural and regulatory mutations depends on cellular contexts or pathways. Using the same categorical metabolic functional classifications outlined above (carbon compound utilization, energy metabolism, central intermediary metabolism, and building block biosynthesis), we found that deletion genotypes involved in building block biosynthesis were more likely to be compensated for by regulatory mutations (Fisher's exact test, P = 0.006, odds ratio 2.5; Fig. 4.5A). Thus, we found that the enrichment of regulatory mutations described above was dependent on cellular context, revealing our first predictive pattern in the path of molecular evolution during functional innovation.

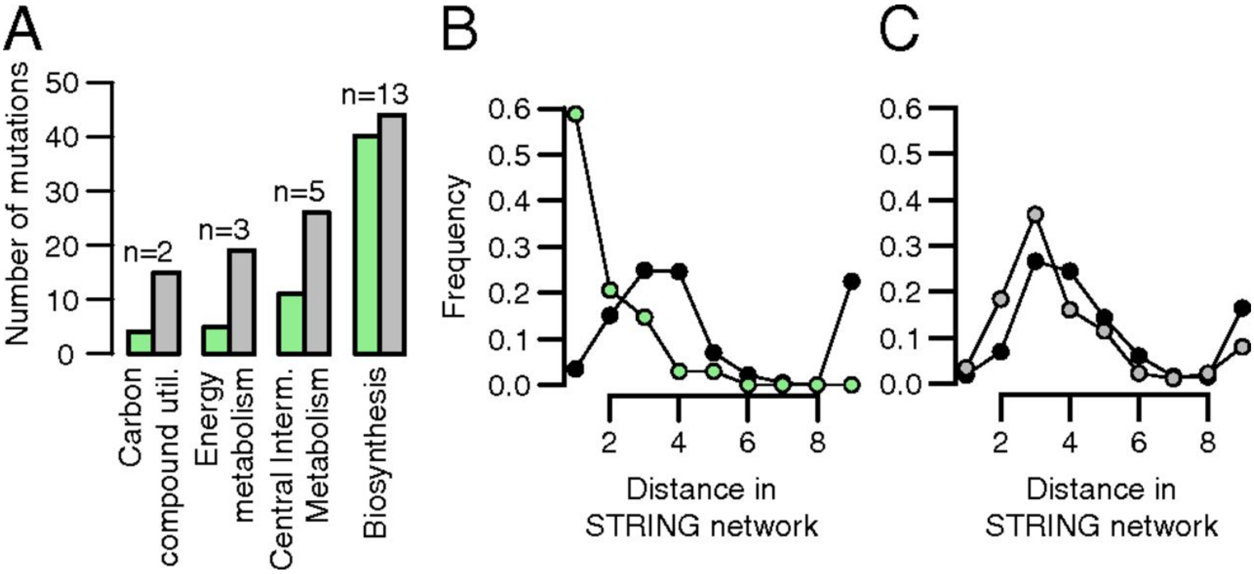

Figure 4.5: (Continued on the following page.)

Figure 4.5: **Mechanisms promoting novel functionality are dependent on cellular context.** (A) The relative enrichment of regulatory and structural mutations is dependent on cellular function. Mutations that contribute toward novel functionality related to building block biosynthesis are more enriched for regulatory mutations, with 48% of all mutations being regulatory. In contrast, in other pathways, only 17–23% are regulatory. Green bars indicate regulatory mutations; gray bars indicate structural mutations. The numbers above the bars indicate the number of deletion genotypes within the category. (B) Regulatory mutations recruit proteins that act in functions related to the missing function. We calculated the shortest network distance between pairs of genes from high-confidence links in the STRING database (Jensen et al., 2009). Green points indicate the network distances between the deleted gene and the genes recruited for functional compensation. Black points indicate the expected network distance between the set of deleted genes and a randomly selected recruited gene based on 5,000 randomizations of protein pairs. The last bin includes all gene pairs with a distance of nine or more, or which are not connected in the network. Genes recruited via regulatory mutations are on average more than three network links closer than expected by chance (Wilcoxon rank-sum test between observed and randomized network distances; n = 34; P = 2.5e-15). (C) Structural mutations recruit proteins that act in functions unrelated to the missing function. Gray points indicate the network distances between the deleted gene and the genes recruited for functional compensation. Black points indicate the expected network distance between the set of deleted genes and a randomly selected recruited gene based on 5,000 randomizations of protein pairs; genes recruited via structural change mutations are on average only 0.6 network links closer than expected by chance (Wilcoxon rank-sum test; n = 85; P = 4.0e-5).

Having established that the types of mutations that occur are affected by the nature of the novel metabolic function that is required, we asked whether these mutations themselves affect predictable cellular functions. We measured the shortest physical and functional proximity [network distances (Jensen et al., 2009)] between a deleted gene and the genes recruited to compensate for its function (Materials and Methods), and found that genes recruited via regulatory changes were, on average, more than 3 network links closer to the deleted genes than would be expected in a randomized network (Fig. 4.5B; P = 2.5e-15, n = 34, Wilcoxon rank-sum test). In contrast, genes recruited through structural changes were, on average, only 0.6 links closer to the deleted genes than in a randomized network (Fig. 4.5C; P = 4.0e-5, n = 85, Wilcoxon rank-sum test). Thus, proteins that confer novel functionality through regulatory change tended to affect proteins that function within the pathways that are most relevant to the missing (deleted) function, whereas proteins that are affected by structural mutations tended to be located in unrelated pathways.

Finally, we asked how demographic parameters might influence the relative contribution of regulatory and structural mutations. We have shown that structural mutations were more frequently selected for during functional innovation but regulatory mutations were more strongly enriched than expected on the basis that intergenic regions provide a relatively small mutational target. This differential enrichment yields a general insight into the nature of regulatory and structural mutations: Either regulatory mutations have a higher probability of being beneficial than structural mutations, or regulatory mutations have larger beneficial effects. A simple population genetic model suggests that it should be possible to disentangle these two hypotheses by testing how changes in population size affect the fraction of regulatory mutations that are observed (Fig. 4.6). If regulatory mutations have a higher probability of being beneficial, then their relative numbers will be enriched relative to structural mutations (compare Fig. 4.6A and Fig. 4.6B), and this enrichment will be independent of population size (Fig. 4.6B). In contrast, if regulatory mutations have larger beneficial effects than coding mutations, the level of enrichment will be dependent on population size: Larger populations will fix a greater fraction of regulatory mutations (Fig. 4.6C). These results show that population size can impact the relative contribution of regulatory and structural mutations, and emphasize that a predictive framework of the molecular basis of evolutionary innovation should take into account demographic parameters.

Figure 4.6: **Population genetic modeling (SI Appendix, Materials and Methods) shows that the relative numbers of regulatory (green points) and structural mutations (gray points) that contribute to novel function can depend on demographic parameters.** (A) When the proportion of structural and regulatory mutations is 0.85 and 0.15, respectively (similar to the ratio of nonsynonymous to intergenic sites in the *E. coli* genome), and the distribution of selective effect sizes is identical, the ratio of the average number of structural and regulatory mutations within an individual is approximately independent of population size (white points; Lower). (B) If the number of structural sites at which structural mutations are beneficial is halved, the ratio again remains independent of population size, but the fraction of regulatory mutations approximately doubles. (C) If the mean and variance of the effects of structural mutations on fitness are half that of regulatory mutations, the ratio is dependent on population size, with individuals in larger populations containing larger relative numbers of regulatory mutations. (Insets) The shape of the distribution of mutational effects for structural (black) and regulatory (green) mutations. In A, the two distributions are identical. The results shown here correspond to 150 generations of evolution. All points shown are the means of at least 50 independent simulations.

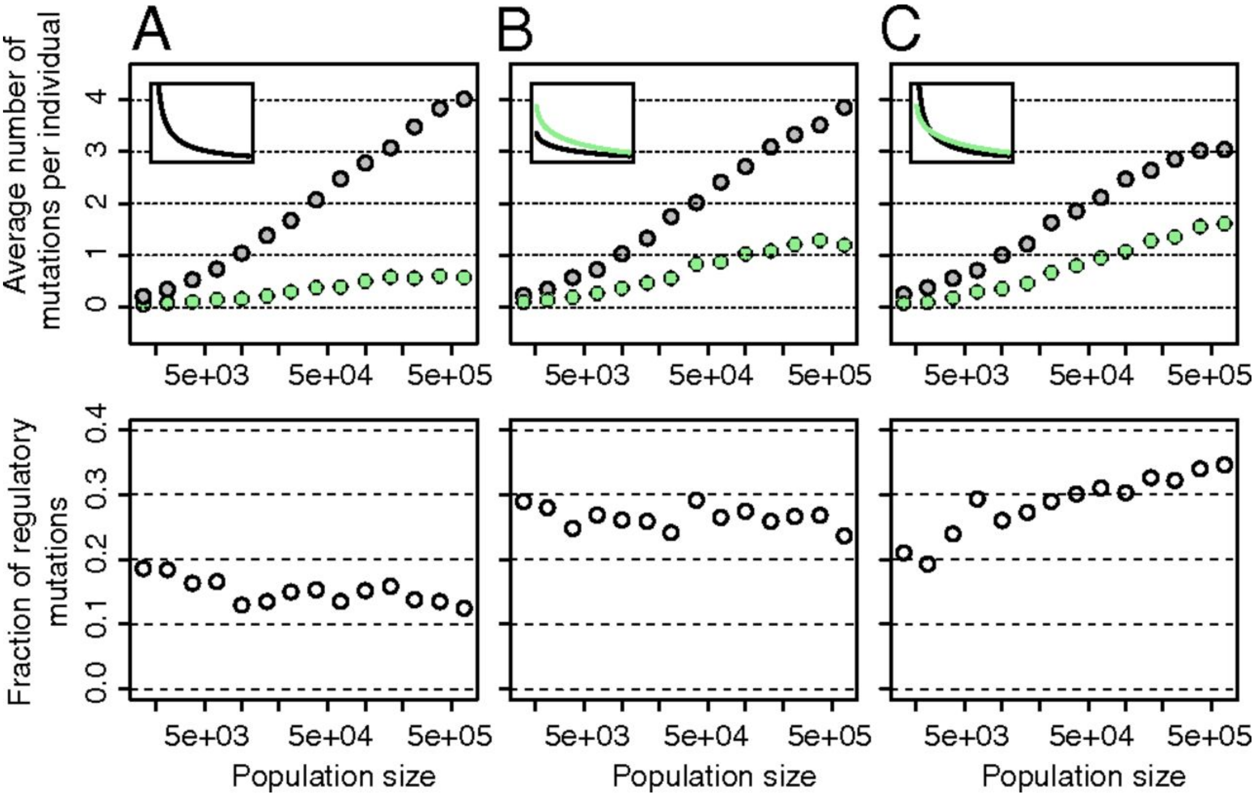

Figure 4.6: (Continued on the following page.)

By using a large number of independent instances of functional innovation and comprehensively characterizing their molecular effects, we have been able to develop a predictive framework for the genetic basis of evolutionary innovation. This framework provides insight into both when and via what mechanisms functional innovation will occur. We have shown that coding mutations tend to be numerically dominant overall, but regulatory mutations are much more common than expected based on the small fraction of the genome that does not encode proteins. One possible explanation for this observation is that coding mutations are more likely to incur antagonistic pleiotropic effects. For example, several of the regulatory mutations we observed affected genes that are essential for growth in minimal glucose media [e.g., *proB* (a glutamyl kinase), *pabB* (an aminodeoxychorismate synthase), and *metE* (a homocysteine transmethylase)]. Changes in the coding regions of these proteins may have detrimental effects on these genes' native functions, whereas regulatory mutations may tend to have less of a deleterious effect. The regulatory mutations we observed frequently resulted in the creation of new transcriptional control elements. They were particularly frequent in novel functions related to building block biosynthesis—functions which were hard to evolve—and generally affected genes that act in cellular contexts closely related to the novel function that was evolved.

It has previously been proposed that gene duplications play a critical role in functional innovation (Bergthorsson et al., 2007, Ohno, 1970); more recently, this phenomenon has been observed in a laboratory setting (Näsvall et al., 2012). However, we found that gene amplifications did not dominate during the early stages of the evolution of functional innovation.

This is perhaps due to deleterious pleiotropic effects that manifest when large parts of the genome are amplified, and the rarity with which smaller and less deleterious duplications occur (e.g., (Blount et al., 2012)). At the same time, our results suggest that during the early stages of functional innovation, some genes may take on dual roles in the cell (Carroll, 2008), and these dual roles may be facilitated by overexpression. Later, rare duplication of these loci may allow their specific enzymatic activities to diverge (Näsvall et al., 2012).

The predictive framework we present here contributes to our fundamental understanding of the evolutionary process, and at the same time provides insights into how populations can respond to rapid environmental change.

4.5 Materials and Methods

4.5.1 Gene deletion strains

All deletion strains (SI Appendix, Table S4.1) were taken from the Keio collection (Baba et al., 2006). Growth phenotypes were confirmed through liquid culture (SI Appendix, Materials and Methods).

4.5.2 Experimental evolution

We evolved five replicate 1−mL cultures of 0.2% glucose minimal media (M9) for each deletion line for 28 transfers (SI Appendix, Materials and Methods).

4.5.3 Sequence analysis

Sequencing was performed on an Illumina HiSeq to a median of at least 50x depth per genome. All mutation data are available in Dataset S1.

4.5.4 Assaying the effects of intergenic mutations on protein expression

We used a transposon-mediated method (McKenzie & Craig, 2006) to integrate translational fusions consisting of the promoter with either the evolved or ancestral alleles, 20 amino acids of the upstream and downstream ORFs, and GFP into the chromosome at a defined neutral locus (SI Appendix, Materials and Methods).

4.5.5 Plasmid rescue of deletion genotypes

We transformed the original deletion genotypes from the Keio library with the respective complementary IPTG-inducible plasmids from the ASKA+ library (Kitagawa et al., 2005). See SI Appendix, Materials and Methods for deletion genotype–ORF pairs.

4.5.6 Analysis of network distances

We used the STRING database (Jensen et al., 2009) to find network distances between proteins and used a randomization test to calculate a null distribution.

4.5.7 Statistics

All statistical analyses were performed in R 3.0.1 (Team, 2007).

4.6 Acknowledgements

We thank C. Bigosch for genotyping assistance; E. van Nimwegen for input on calculating transcriptional output; F. Bertels, C. Field, F. Geier, and O. Tenaillon for comments and suggestions; and T. Bollenbach for sharing the pNDL1 plasmid. D.B. and O.K.S. were supported by grants from the Swiss National Foundation ($PZ00P3_126617$ and $PZ00P3_144605$). M.A. was supported by Swiss National Foundation Grant $31003A_130735$.

4.7 SI Appendix

4.7.1 SI Appendix, Materials and Methods

Confirmation of conditional lethal growth phenotypes

Some of the deletion genotypes (Table S4.1) that we used have been variably classified as exhibiting conditionally lethal phenotypes in glucose minimal media or not (Baba et al., 2006, Patrick et al., 2007, Feist et al., 2007). This may be due to different methodologies or to the ability of these strains to grow for a short period of time after dilution from rich media into minimal media due to low-levels of residual nutrients present in the cell. Thus, we confirmed the conditional lethal growth phenotypes in the following manner: we grew deletion mutants overnight in 200 μl of LB media and then seeded wells containing 1mL of 0.2% glucose minimal media with approximately 100 cells. We allowed these cultures to grow for 48 hours, diluted 1:1000 into minimal media, allowed these cultures to grow for another 48 hours, again diluted 1:1000, and again allowed the cultures to grow for 48 hours. A small number of the deletion lines exhibited visible growth after the first dilution into minimal media (visible growth at this point would require on the order of 16 doublings); diluting less at the first step resulted in visible growth from more cultures. However, after the additional second and third serial dilutions, the vast majority of genotypes no longer exhibited growth, having no measurable OD nor exhibiting any viable colonies after plating on rich media.

Using this above dilution protocol, we divided the deletion mutants into two classes: strains in which we observed no sustained growth (fewer than $5 * 10^1$ cells per mL after serial dilution and regrowth) and those that exhibited prolonged growth (greater than $5 * 10^5$ cells per mL after serial dilution and regrowth). The vast majority of deletion strains that we examined were in the first class, with a small number in the second class ($\Delta cysC$, $\Delta cysD$,

$\Delta cysG$, $\Delta cysH$, $\Delta cysI$, $\Delta cysJ$, and $\Delta cysN$); we did not examine these latter strains further. In these cases, growth may have been due to the non-lethality of the deletion, to low levels of contaminating nutrients in our media, or to rapid evolution of suppressor mutations. Two strains exhibited an intermediate number of viable cells after these three dilutions. These were *argC* and *argB*, for which the ancestral strains exhibited between $5 * 10^1$ and 10^3 per mL after the above transfer regime. However, for these strains, only two and one lineages, respectively, recovered function, indicating either that prolonged growth could not occur, or that this growth was too slow for the cultures to maintain viability. We thus included the evolved lineages descending from these deletion strains in all our analyses.

Experimental Evolution

Five replicate 200 μl LB cultures were grown overnight for each deletion line. Each culture was then used to inoculate an 800 μl culture of 0.2% glucose minimal media (M9) with 50 μg/ml kanamycin. These were grown for 48 hours and diluted five-fold into 800 μl of fresh M9 and grown for 48 hours. This process was repeated ten times, after which the cultures were diluted 1:100 into 990 μl M9 and allowed to grow for 48 hours. This was repeated 18 times. During evolution, all wells contained a single glass bead for turbulent mixing. Plates were incubated at 37°C, shaken at 600 rpm and were covered with plastic foil to prevent evaporation.

Measurement of growth rates of evolved clones

All evolved clones were stored in LB glycerol stocks to ensure that sufficient numbers of cells were frozen. In order to ensure that no residual nutrients from LB were present during growth rate measurements, cultures from frozen stocks were passaged by growing overnight in LB, and diluting 10^7-fold and 10^3-fold in M9, allowing 48 hours of growth after each dilution. These cultures were then diluted 10^4-fold into M9 and OD_600 was measured every 40 minutes for 48 hours. We used an oil overlay to prevent evaporation. This oil overlay had no measurable effect on the growth rates. Maximum growth rates were calculated using an exponential fit over a sliding window of six time points (approximately 200 minutes), with the requirement that the fit have an r^2 greater than 0.98. All measurements were performed in triplicate. Occasionally no growth was observed in one well. This was most likely due to the large variability in lag times (see SI Appendix), which sometimes resulted in very few or no cells being transferred between cultures. We excluded these cases in which no growth was observed from the calculations of the means and standard deviations in growth rates.

To test if growth rates of clones with the same deletion genotype were more similar than expected, we used a randomization test. For each deletion genotype in which n lineages recovered, we selected n growth rates without replacement from the set of all measured growth rates and quantified the standard deviation of these growth rates. We repeated this selection process 10,000 times, and then compared the standard deviation of the observed standard growth rates for each genotype to the set of randomized growth rates to calculate the probability of observing a standard deviation in growth rates as small or smaller than the observed standard deviation.

Sequencing

Genomic DNA was prepped using the GenElute Bacterial Genomic DNA Kit (Sigma). Samples were sonicated, and size restricted to isolate fragments between 300 bp and 500 bp in size. A PCR step was used to index the fragments. Single-end 50 bp reads were obtained using an Illumina HiSeq, with between eight and 24 libraries per lane. All lines were sequenced to a median depth of at least 50−fold. Mutations were identified using Breseq 0.21 (Barrick et al., 2009). We confirmed a subset of the identified mutations using Sanger sequencing and in all cases confirmed the mutations. Due to variation in read coverage across the genome, deletions and amplification were identified by calculating average coverage within a sliding window of 25 bp across the genome and looking for decreases (greater than four-fold) in coverage between neighboring windows. In all cases in which amplifications were identified, we further examined these genomic regions for polymorphisms using SAMtools 0.1.18 (Li et al., 2009). In no cases were any polymorphisms identified.

Exclusion of ancestral and previously observed mutations

For any mutations that were shared across all evolved lineages of a single deletion strain, the ancestral deletion strain was Sanger sequenced at this locus to check whether the mutation was independently evolved in these lineages, or whether it was present in the ancestor. In the majority of cases, these shared mutations were present in the ancestral strains (Table S4.2). These mutations were excluded from all analyses that we present. As noted in the main text, we also excluded any mutations that have been observed previously in other experimental evolution studies (Schneider et al., 2000, Notley-McRobb et al., 2002, Charusanti et al., 2011, Jensen, 1993, Conrad et al., 2009, Rath & Jawali, 2006) (Table S4.3).

Derivation of $\mathbf{K}_a$, $\mathbf{K}_s$, and $\mathbf{K}_i$

We used data from Hershberg and Petrov (Hershberg & Petrov, 2010) on synonymous mutations to estimate mutational biases in *E. coli*, approximating mutation rates as 66% GC:AT; 17% AT:GC; 10% GC:TA; 4% AT:CG; 2% AT:TA; and 1% GC:CG. Thus, mutations at GC sites comprise 77% of all mutations, while mutations at AT sites comprise only 23%. Intergenic regions in *E. coli* are 41.9% GC, while coding regions are 52.0% GC, excluding RNA genes. The *E. coli* genome is 12.2% intergenic (again, excluding RNA genes). We then use these numbers to calculate the fractions of mutations that we expect to be synonymous, nonsysnonymous, and intergenic. With the above data, we then expect that 11.1% of all point mutations will occur in intergenic regions, 60.7% will be nonsynonymous, and 28.1% will be synonymous. However, we found that 25.2% of all point mutations were intergenic, 69.4% were nonsynonymous, and 5.4% were synonymous.

Calculation of σ^{70} binding and transcription initiation rates

We used alignments of annotated σ^{70} sites (Djordjevic, 2011) to infer a weight matrix (Schneider, 1997) including the -35 and an extended -10 (Djordjevic, 2011) (SI Appendix).

The 35 mutations investigated here were all the unique point mutations (26 in total), unique small insertions (four insertions of 1 bp), and unique small deletions (one 5 bp deletion, one 4 bp deletion, one 3 bp deletion, and two 1 bp deletions). Other larger deletions almost certainly affected terminator structures or the binding of multiple transcription factors (SI Appendix and Table S4.6). We did not consider IS element changes or amplifications in calculating changes in σ^{70} or translation initiation rates.

We calculated promoter binding energy as the summed weight matrix scores over all possible windows and all possible spacers in a window running from 150 bp upstream of the mutation to either 150 bp downstream of the evolved mutation or to 10 bp before the first codon of the downstream open reading frame (to account for the 5' UTR required for ribosomal binding), whichever was shorter. In two cases, intergenic mutations occurred directly downstream of the deleted gene. In these cases, the upstream region was substituted with that of the Kanamycin resistance locus and FRT site that was present in the genome, and the weight matrix score was calculated from this sequence.

We calculated the binding energy of a promoter as:

$$E_{pr} = \left(\frac{1}{\beta}\right) * \log\left(\sum_{s} \exp(\beta * E(S))\right)$$

In which E$_{pr}$ is the energy of the promoter; β is a scaling factor (set to one in this case as we express energies in k$_B$T units); and E(S) is the energy of site S (or weight matrix score), defined as (Brewster et al., 2012):

$$E(S) = \sum_{bi} E_{bi} * S_{bi}$$

In which E is a 4 x *i* matrix with element E$_{b,i}$ specifying the energy contributed by base *b* at position *i*, and S is a 4 x textiti matrix in which element *b,i* is 1 if the observed base at position *i* in site S is *b*, and 0 otherwise.

Energy matrix elements E$_{b,i}$ were derived as (Kinney et al., 2010):

$$E_{b,i} = -\log\left(\frac{f_i(b)}{p(b)}\right), f_i(b) = \frac{c_i(b)+1}{N+4}$$

In which ci(b) is the number of sites with base *b* at position *i*, N is the total number of sites, and p(b) is the frequency of base *b* in the intergenic DNA of E. coli (e.g. for guanine and cytosine, 41.9%, as indicated above). We used an annotated set of σ^{70} binding sites (Djordjevic, 2011) to infer the energy matrix elements.

The probability that a promoter is bound by a sigma factor is (assuming that the concentration of sigma factor is not saturating) (Brewster et al., 2012):

$$C * \exp(\beta * E_{pr})$$

In which C is a constant specifying the transcription expected from non-specific binding, and which we interpret to be similar for the ancestral and evolved strain. Finally, we assume that this probability is proportional to the level of gene expression (Brewster et al., 2012), which is what we plot in Fig. 4.3D and E and Fig. S4.4 (expressed in arbitrary units).

These calculations were performed for both the ancestral and derived promoter regions. We also calculated changes in the binding energy for all possible mutations in this region, weighted by the probabilities of each type of point mutation as approximated above (66% GC:AT; 17% AT:GC; 10% GC:TA; 4% AT:CG; 2% AT:TA; and 1% GC:CG), and with the weight of 1 bp insertions and deletions equal to each other and consisting of 10% of all mutations.

We used the online RBS calculator (Salis, 2010) to calculate translation initiation rates for the ancestral and derived alleles.

We list the predicted transcriptional and translational effects for all intergenic point mutations and small indels in Table S4.6.

Assaying the effects of intergenic mutation on protein expression

Transposon-mediated integration of translational fusions was performed in the following manner: the region between AttR1 and AttR2 from plasmid pNDL1 (Bollenbach & Kishony, 2011) was replaced with GFPmut2 (Zaslaver et al., 2006) using the ApaI and XhoI sites. At the same time, XmaI and SacII cut sites were added between the GFP start codon and the XhoI site. This resulted in an MCS consisting of XhoI, SacII, and XmaI followed immediately by the start codon of GFP. This plasmid was then used to clone the translational fusions. In the case of the mutation upstream of *carB*, which was observed in a *carA* deletion line, the entire region encompassing the 20AA from the gene upstream of *carA* (*dapB*) to the first 20AA of *carB* was cloned, such that the promoter of *carA* as well as the Kanamycin resistance gene, was present in the construct.

The region between the attR1 and attR2 sites in these plasmids was then integrated into the chromosome at the attTn7 site of the evolved strain (McKenzie & Craig, 2006), resulting in two strains for each intergenic region, one with the ancestral intergenic allele in the evolved strain background, and a second with the evolved intergenic allele in the evolved strain background. All integrated alleles were sequenced to confirm the expected mutations.

We grew these strains overnight in glucose minimal media, diluted 1:100 in minimal media, and grew for four hours (such that early exponential phase was reached). We assayed expression level using flow cytometry, with excitation at 488nm and an emission filter of 513/17. We collected data for at least 50,000 cells, gated on a subset of approximately 5,000 cells having similar FSC and SSC values (Silander et al., 2012), and calculated the median fluorescence level of these cells. We also calculated the median fluorescence of cells without the integrated construct, and used this as the background fluorescence level (caused by cellular autofluorescence). Each construct was measured in triplicate.

Quantifying growth during plasmid rescue of deletion genotypes

We selected eight deletion genotypes for which evolved clones contained mutations that appeared to possibly contain large effect regulatory mutations allowing rescue. These eight deletion genotypes and open reading frames were: Δ*metR* and *metE*; Δ*thrA* and *metL*; Δ*carA* and *carB*; Δ*ilvE* and *avtA*; Δ*leuL* and *leuA*; Δ*cayB* and *cysP*; Δ*ilvE* and *ydeM*; and Δ*gltA* and *yebY*.

The latter two deletion genotypes contained large deletions, which we hypothesized might result in novel promoter-open reading frame combinations. All deletion genotypes were also transformed with a randomly selected open reading frame from this same set, but which we did not expect to allow rescue. As expected, none of these resulted in rescue. Growth curves were performed by growing strains overnight in LB containing kanamycin and chloramphenicol, diluting 10^6-fold into M9 minimal media containing 0.2% glucose, 50μg/mL kanamycin, and 34 μg/mL chloramphenicol, and monitoring OD_600 every 20 minutes for 72 hours (although see below). As above, an oil overlay was used. Over this period of time, no non-complementary plasmid resulted in growth. We note that two of these reading frames (*carB* and *avtA*) have been shown previously to complement *carA* and *ilvE* deletion genotype, respectively (Patrick et al., 2007). In one case (Δ*leuL* and *leuA*), we observed robust growth (approximately one hour doubling time) in one well of one replicate at 50μM IPTG, but in no others. This growth occurred after an approximately 65 hour lag phase. We thus continued to monitor all wells for another 24 hours. After this period, we observed growth in one additional well of this deletion genotype - open reading frame combination (at 100μM IPTG). We are not sure whether this growth, which occurred after an extremely long lag phase, was due to rescue via the plasmid, or to the appearance of compensatory mutations in these replicates. We thus did not include these in Fig. S4.5. In addition, none of the latter three deletion genotype-open reading frame combinations listed above (Δ*cysB* and *cysP*; Δ*ilvE* and *ydeM*; and Δ*gltA* and *yebY*) resulted in rescue, and thus are also not included in Fig. S4.5.

Analysis of network distances

We used the STRING database to find network distances between proteins, including only those interactions with a score of 0.800 or above. To calculate a null distribution of shortest distances, we randomized the set of recruited genes among the set of deleted genes and recalculated the shortest distance. We repeated this randomization 5'000 times, doing this separately for the pairs of deleted genes and recruited genes that affected protein structure, and the pairs that affected protein expression.

Modelling

We modeled an asexual population of N individuals that divide by binary fission. Generations were discrete. During each generation, N/2 individuals were selected to divide based on their relative fitness. These individuals produced two offspring; each offspring had a probability $U_r eg$ of gaining a mutation at a regulatory site, and a probability $U_s tr$ of mutating at a structural site. The size of both regulatory and structural mutational effects was gamma distributed and mutations were multiplicative in their effects on fitness. This scheme we use is somewhat similar to that outlined in (Fogle et al., 2008), and, as noted there, ignores deleterious mutations, which are not thought to play an important role in large populations in which beneficial mutations are relatively common (Desai & Fisher, 2007), which is the regime in which we are most likely to be operating, as is evidenced by the large number of adaptive mutations that we have observed.

The beneficial mutation rate at regulatory sites was $1.5e^{-6}$ for all simulations and $8.5e^{-6}$ at structural sites except those presented in Fig. 4.6B, in which structural mutations have a 2−fold lower probability of being beneficial, decreasing the rate to $4.25e^{-6}$. For all simulations, for both regulatory and structural mutations, the scale parameter of the gamma distribution was 0.5. In Fig. 4.6A the shape parameter was 0.4 for both structural and mutations; in Fig. 4.6B, 0.8 for both; and in Fig. 4.6C, 0.4 for structural mutations and 0.8 for regulatory mutations. Each simulation consisted of 150 generations of evolution, after which the mean numbers and ratios of regulatory and structural mutations were enumerated for each individual in the population.

4.7.2 SI Discussion

Laboratory evolution as a method to understand functional innovation

Here we have utilized a laboratory evolution scheme, as it is a tractable method that allows systematic investigation of the molecular mechanisms underlying functional innovation. We briefly discuss the applicability of such a method to understand functional innovation. We begin by discussing the notion of the evolution of functional innovation.

It is difficult to establish a simple and objective definition of functional innovation. Pigliucci (2008) describes it as “a necessarily fuzzy concept.” (Pigliucci, 2008). The most commonly used definition that we are aware is: the development of an ability that allow the colonization of novel ecological niches, originally put forward by Ernst Mayr in 1963 (Mayr, 1963). This definition has been utilized in other laboratory evolution experiments (Meyer et al., 2012, Blount et al., 2012). The traits that we investigate here satisfy this definition in that they enable the bacterial strains that we used to grow in an environment they were not previously capable of growing in – specifically, minimal media. From a genetic standpoint, we can divide phenotypic changes into two types: “loss of function” alleles - molecular changes that result in phenotypes that are equivalent to those when the relevant locus is deleted, and “gain of function,” which are all other phenotypes that differ from the ancestral phenotype. By this definition, many "loss of function" alleles would include evolutionary changes frequently considered evolutionary innovations (e.g. coat color changes (Hoekstra et al., 2006)), suggesting that it is less appealing. Despite this caveat, many of the changes that we observe here (such as increased sigma factor binding) fall clearly into the category of "gain of function."

Thus, from both an evolutionary and mechanistic (genetic) point of view, laboratory evolution provides a valid method to investigate the process of functional innovation. A possible critique is that in the experimental design we have used here, we have relied on a set of deletion mutants to create the initial conditions under which we select for novel function. There is one notable limitation to such an approach: the mechanisms involved in re-evolution of a lost function may not be similar to those found during functional innovation in more natural circumstances. However, as mentioned in the main text, there are many reasons to expect that gene inactivation is a common occurrence in natural systems. This may occur through point mutations or deletions; such inactivations might often require re-evolution of this function when ecological circumstances change. Even for cases in which novel functions evolve that are not compensating for previously lost functions, there are many similarities to the circumstances of the experimental design we present here. For example, many trait innovations, ranging from toxin and antibiotic resistance to changes in the light wavelength absorbance of photoreceptors to changes in oxygen affinity of globins, involve single amino acid changes that affect binding, many of which are similar to those we have observed here. Finally, as mentioned in the main text, the changes we have observed here, and the novel functions that we select for, are very similar, both mechanistically and evolutionarily, to the changes that occur when bacteria evolve the ability to degrade novel compounds, such as organic chlorides. This often occurs through changing both the expression level and binding specificity of genes already present that have low-level activity toward the novel substrate (Bosma et al., 2002).

Evidence for adaptive evolution: parallel substitutions across lineages

In a large number of cases, we found precisely paralleled changes, both regulatory and structural. The full list of mutations is contained in the supplementary dataset. We highlight two here.

(1) In all five recovered lineages that were deleted for *ppc* (phosphoenolpyruvate carboxylase), a single mutation occurred in *icd* (isocitrate dehydrogenase), showing strong parallelism at the genic level. In addition, two lineages shared identical changes at position 302 (Met -¿ Ile).

(2) In all five recovered lineages that were deleted for *thrA* (aspartate kinase), a single regulatory mutation affected the *metBL* operon (*metB* is also an aspartate kinase). One of these occurred in the repressor, *metJ*; the other four occurred upstream of *metL*, and two were precisely convergent (a 1 bp deletion 77 bp upstream of the *metBL* operon). Again, this emphasizes the strong parallelism at both the genic and molecular level.

Evidence for adaptive evolution: changes in substrate binding

In more than one case, mutations occurred that affected genes that have been observed previously to provide compensatory function. For example, overexpression of *hisB* has been shown to compensate for a *serB* deletion. We found three cases in which *hisB* was mutated in a *serB* deletion line. Two of these were identical changes (D57N), with the third affecting a different amino acid (Q23K). Notably, both of these fall close to the active site of the protein, and provide evidence for adaptation not only because of the molecular parallelism, but also because of their hypothetical functional effects (Fig. S4.6). One of these (D57N) has recently been shown to provide novel functionality to HisB allowing it to function in place of SerB (Yip & Matsumura, 2013).

Evidence for adaptive evolution: amplifications and IS element-mediated changes

The variability in copy number of the 11 large-scale amplifications that we observed (Fig. S4.3) suggests that they exert their beneficial effects through changing protein levels; in addition, the location of hypothetical compensating genes suggests that this occurred through increasing copy number and not through promoter capture (Blount et al., 2012). If such amplifications were advantageous solely because of their effect on protein structure, we would not expect that more than a single rearrangement would be necessary.

The functional effects (i.e. regulatory or structural) of the IS element-mediated changes are more difficult to infer. However, in almost one third of all cases, these insertions occurred outside of reading frames, implying that they have a regulatory effect; this appears to be particularly true for those that exhibit strong adaptive signatures (i.e. identical substitutions across lineages; Fig. S4.7, Table S4.4).

Inferring ambiguous recruited genes and functional consequences

The majority of coding mutations could be unambiguously inferred as affecting the structure of a single gene.

Inferring which mutations affected protein level was more difficult. We included all large-scale amplifications and all intergenic mutations in the class of mutations affecting protein level (Fig. S4.3B), although this former class was excluded from all functional analyses we present.

Many of the intergenic mutations we observed had ambiguous effects, as they occurred upstream of two reading frames (in only one case did a mutation occur downstream of two reading frames, and this mutation was the removal of a REP element, a high frequency mutational event). We resolved such ambiguous mutations on a case-by-case basis using evidence from the literature, computational predictions (e.g. effects on sigma factor binding), and experimental measurements (Table S4.5).

Evidence for regulatory effects of mutations in coding regions

Based on circumstantial evidence, in the functional analyses we included three mutations that occurred in coding regions as affecting protein levels.

In the first case, a region upstream of pabB contained two mutations: one intergenic mutation 19 bp upstream of the start codon, and a second within the reading frame of the upstream gene (*yoaH*, a gene of unknown function), causing a nonsynonymous change. This change was 83 bp upstream of the *pabB* start codon. We surmised that this change was selectively advantageous through its affect on *pabB* expression. Further supporting this is that these two changes occurred in a *pabA* deletion line, and overexpression of *pabB* is known to rescue *pabA* mutants (Patrick et al., 2007). Finally, computational inference suggested that the nonsynonymous change in the *yoaH* reading frame would strongly increase the binding energy of σ^{70} to the promoter region (Fig. 4.3E and Fig. S4.4).

In the second case, a strain deleted for *ptsI* contained a mutation in *yfeO* (a predicted ion channel protein), which lies upstream of *glk*. Overexpression of both Glk and GalP are known to increase growth in PTS mutants (Hernández-Montalvo et al., 2003); this mutation also increased σ^{70} binding (Fig. S4.4). Three other mutations occurred in *galR* or upstream of *galP* in other Δ*ptsI* lineages, suggesting that this PTS mutant could indeed increase growth by changing *galP* expression; in addition, a second mutation upstream of *glk* also occurred in one of these lineages. The occurrence of these other mutations strongly suggested that the nonsynonymous mutation in *yfeO* exerted a beneficial effect through changing the protein level of *glk*.

In the final case, a strain deleted for *thrA* contained a synonymous mutation in *metB*, which is contained in the *metBL* operon and is upstream of *metL*; this was the only mutation that we found in this line. In all four other Δ*thrA* deletion lines we also found only a single mutation, and these mutations occurred in either the intergenic region upstream of the *metBL* operon, or in the transcriptional repressor of the *metBL* operon, *metJ*. As in the above instances, the mutation increased σ^{70} binding energy, although only marginally. In addition, *metL* is an extremely close homologue of *thrA*, suggesting a specific functional connection. Finally, IPTG-mediated expression of the *metL* ORF rescues the lethal Δ*thrA* phenotype. We surmised, then, that the synonymous mutation in *metB* most likely affected the expression of *metL* through an unknown mechanism.

Besides the above synonymous mutation we could not infer with any certainty functional effects for the other five synonymous mutations (Fig. 4.3B). These may have been neutral in their phenotypic effect, although in at least one case, there is a suggestion of a functional connection. In one *glyA* deletion line, a synonymous mutation occurred in *pepN*. The STRING database (Jensen et al., 2009) has very high confidence in the interaction between these two proteins (0.8; *glyA* is in the top ten most likely interacting proteins for *pepN*) due to the pathways of these two genes being closely connected (cyanoamino acid metabolism and glutathione metabolism). This suggests a functional link between these two genes. However, this evidence is not strong.

Four changes occurred in genes encoding RNA molecules. Although we did not include these in any of our functional analyses, one almost certainly conferred a beneficial effect through changing protein levels. In a *fes* deletion line, a mutation occurred in *ryhB*. *fes* is essential for the release of iron from siderophores (Pettis et al., 1988); *ryhB* acts to reduce iron consumption by reducing the expression of mRNAs that encode proteins that utilize iron (Massé et al., 2005) and by increasing the expression of *shiA*, which is involved in siderophore production; this mutation alters a hairpin structure, perhaps allowing for increased binding to *shiA* or the target genes that it down-regulates (see (Fröhlich & Vogel, 2009, Prévost et al., 2007)).

Five of the mutations we observed were mid-sized indels (between 20 and 100 bp). The majority of these acted through incurring clear losses-of-function: Three removed terminator structures; one removed several activator sites from a promoter; and one removed a REP element.

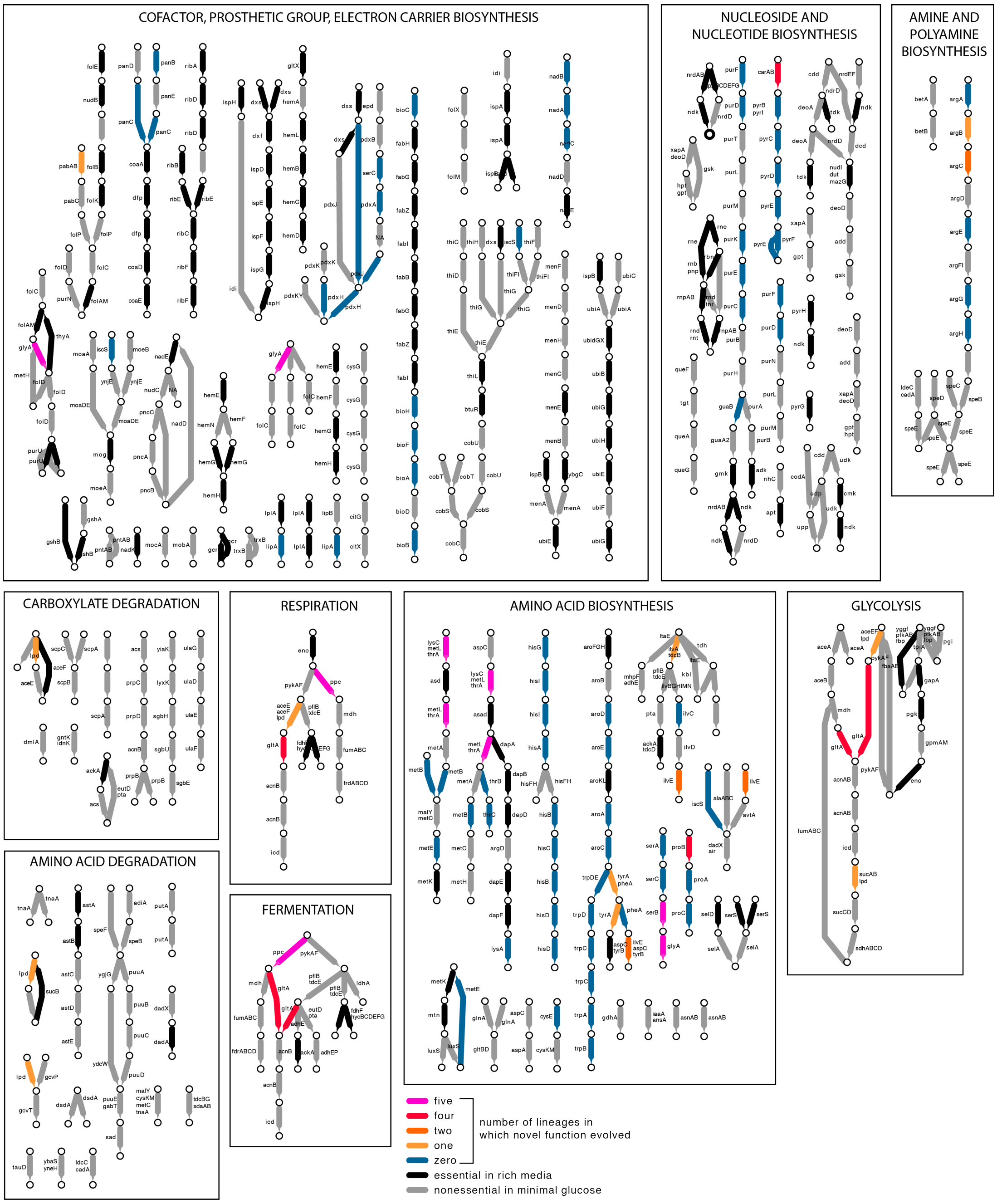
COFACTOR, PROSTHETIC GROUP, ELECTRON CARRIER BIOSYNTHESIS
NUCLEOSIDE AND NUCLEOTIDE BIOSYNTHESIS
AMINE AND POLYAMINE BIOSYNTHESIS
CARBOXYLATE DEGRADATION
RESPIRATION
AMINO ACID BIOSYNTHESIS
GLYCOLYSIS
AMINO ACID DEGRADATION
FERMENTATION
five
four
two
one
zero
number of lineages in which novel function evolved
essential in rich media
nonessential in minimal glucose

Figure S4.1: Metabolic network context of the 22 deleted functions for which one or more lineage evolved novel functionality (Keseler et al., 2011). Reactions are color-coded according to the number of lineages that evolved a new function that compensated for the deleted function. Reactions that are catalyzed by more than one enzyme are colored if at least one of the enzymes was deleted and the function was recovered. Reactions colored in grey are not essential in minimal glucose media, and thus were not investigated; reactions colored in black are essential in rich media (and minimal glucose media) and thus could not be investigated in the manner used in this study. Reactions colored in light blue are essential in minimal glucose media (Patrick et al., 2007), but were not investigated in this study. The figure was created using Cytoscape (Shannon et al., 2003).

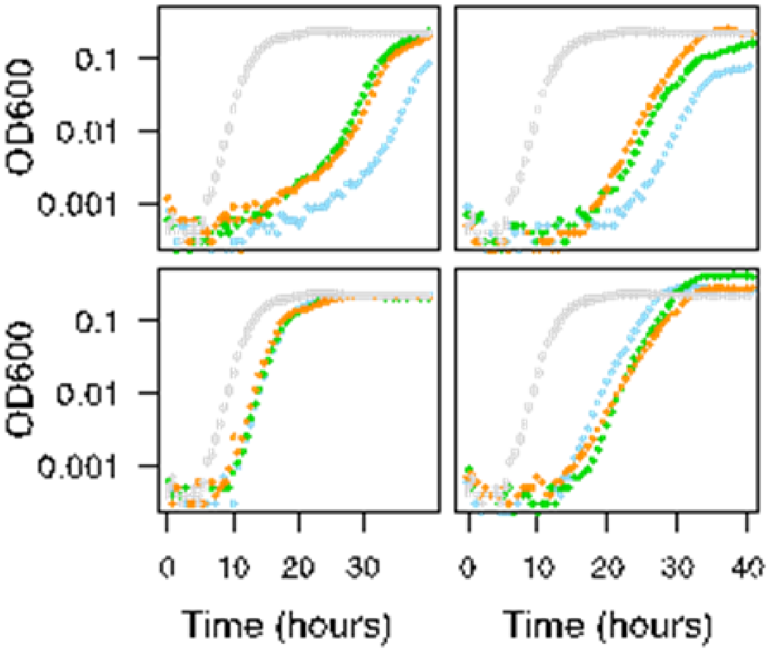

Figure S4.2: The lag times in recovered clones were substantially longer than in the wildtype, even for those lineages that recovered near-wildtype maximum growth rates. The growth curves are from clones within populations having (clockwise from top left, with the population replicate in subscript): $\Delta glyA_B$, $\Delta hisH_D$, $\Delta hisH_E$ and $\Delta carA_C$ deletion genotypes. Growth curves for the ancestral clone (BW25113) harboring no deletion are shown in grey. Three biological replicates are shown for each clone.

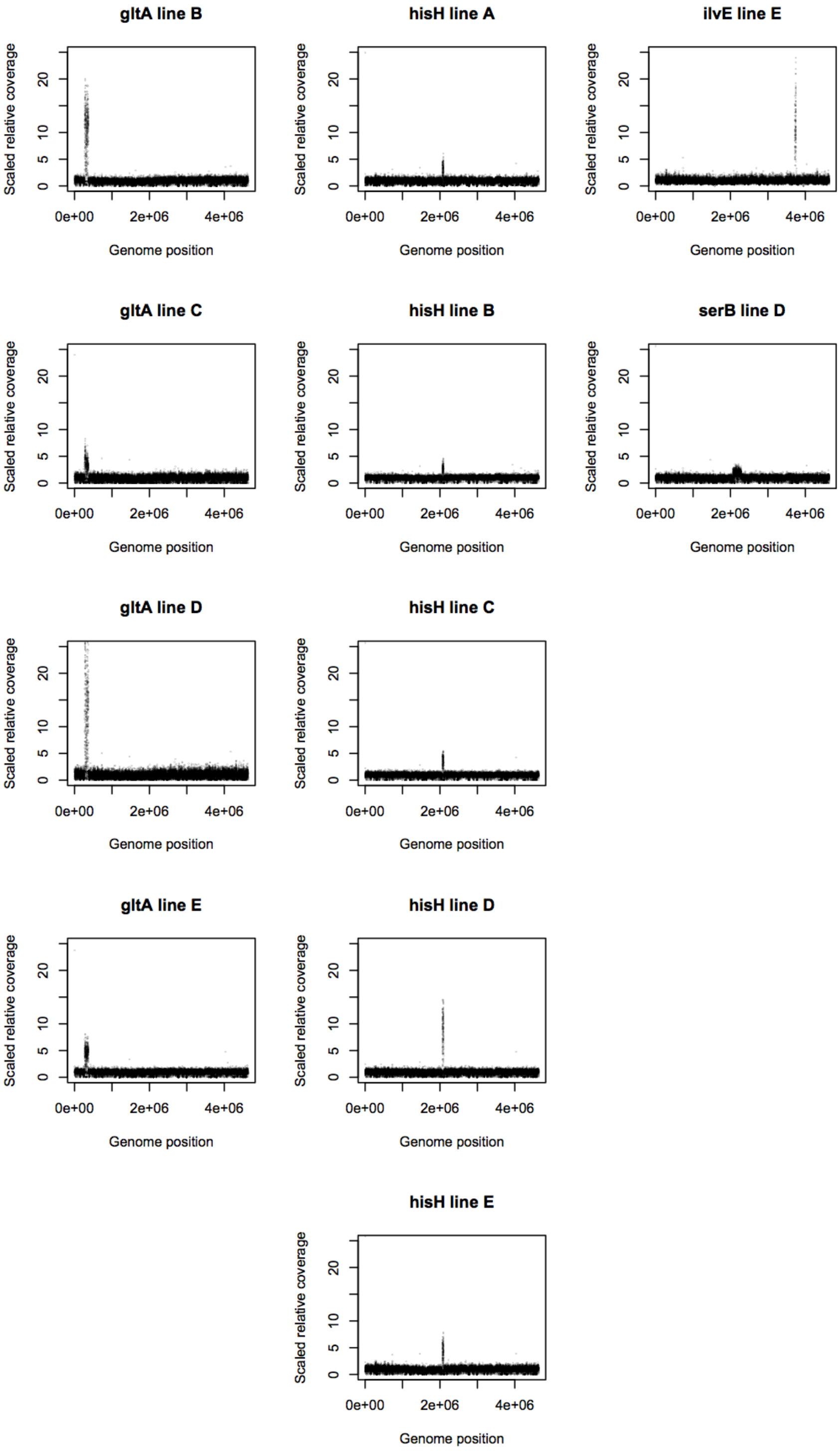

Figure S4.3: Large-scale genomic amplifications occurred frequently. Coverage for uniquely mapping reads is plotted. Each point represents the coverage at one base pair; only one in 200 base pairs are plotted. The coverage is shown relative to the coverage for a genome lacking any amplifications, and is scaled such that the median coverage is one. All amplifications observed in the $\Delta gltA$ lineages occurred between base pairs 270,987 and 370,982, a 100 Kbp amplification. In one case ($\Delta gltA$ line D) a greater than 10- fold amplification occurred, increasing the genome size by more than 106 base pairs (on the order of 25% of the genome). All amplifications observed in the $\Delta hisH$ lines occurred between base pairs 2,064,329 and 2,100,935, a 36.6 Kbp amplification. The amplifications in the *ilvE* and serB deletion genotypes were 27 Kbp and 220 Kbp, respectively.

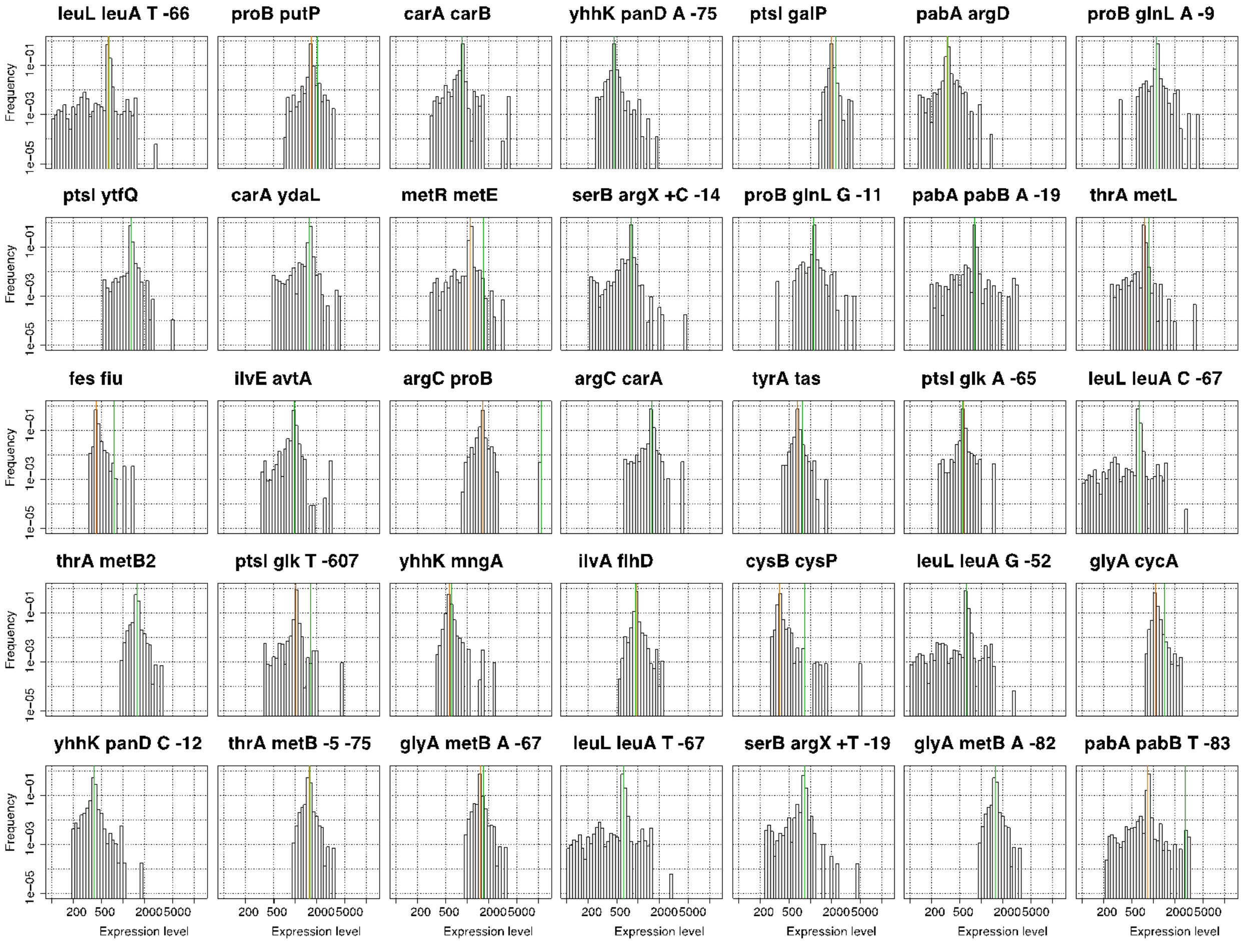

Figure S4.4: Many of the intergenic mutations we observe are among those that have the largest effect on predicted expression levels. **Above:** Predicted expression levels for all possible mutations in each promoter region for all of the unique intergenic point mutations and small indels. The orange vertical line indicates the predicted expression of the ancestor; the green vertical line indicates the predicted expression level of the evolved promoter. The deletion genotype and the recruited gene, respectively, are indicated above each plot, with the specific mutation listed when multiple different mutations were observed within the same pairs of genes. In many cases, the predicted expression levels do not change, such that the green line completely obscures the orange line. **Following page:** The same data shown in the above set of plots, but as reverse cumulatives. The orange points indicate the predicted expression of the ancestor; the green points indicate the predicted expression level of the evolved promoter. Note that both the x- and y-axes are on log scales.

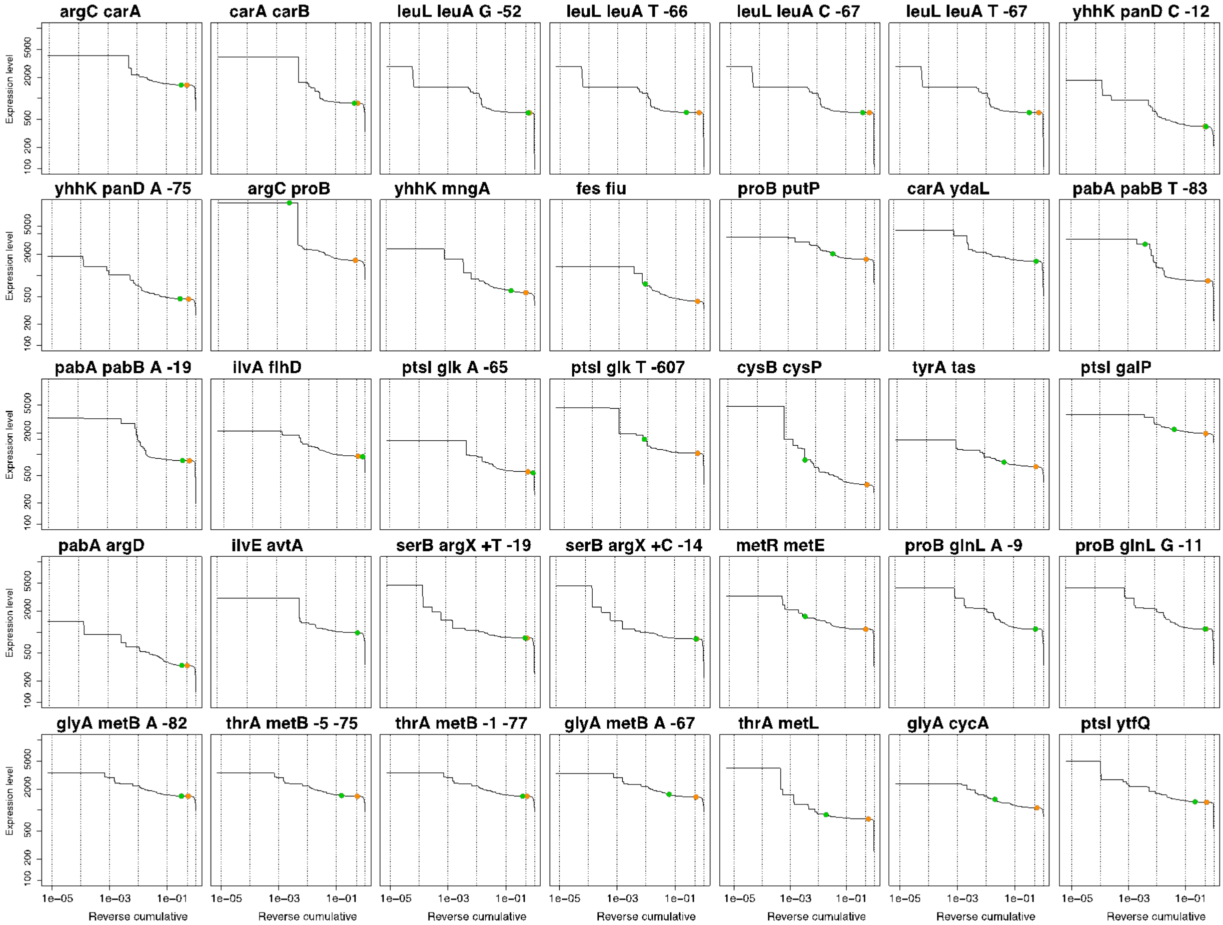

Figure S4.4: cont.

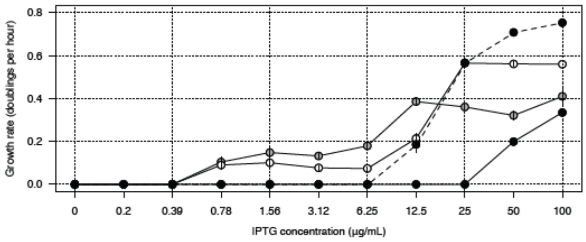

Figure S4.5: Changes in protein expression level alone can rescue the lethal effects of the deletion genotypes. We induced expression of a single open reading frame using IPTG. Each point shows the mean estimated doubling time (± s.e.m). In three cases, growth exhibits a threshold-like behavior, with strains growing robustly at one expression level, but failing to grow at slightly lower expression levels, suggesting that small changes in expression can have profound effects on growth. The deletion genotypes and open reading frames are: Δ*metR* and *metE* (grey); Δ*thrA* and *metL* (black); Δ*carA* and *carB* (white); Δ*ilvE* and *avtA* (black with a dashed line).

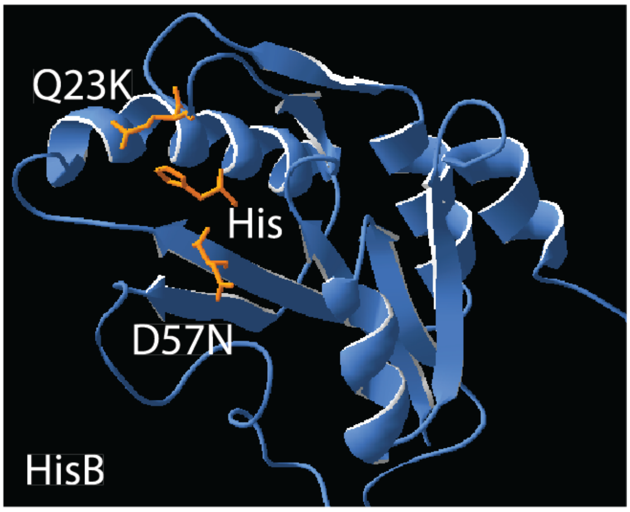

Figure S4.6: Changes in HisB (Rangarajan et al., 2006) were paralleled across several *serB* deletion lineages. In each of these cases, the mutations affected sites proximate to the binding site in the protein (changes are shown in orange above). These are likely to cause increased binding for the substrate of SerB; overexpression of *hisB* has been proposed previously to allow compensation for *serB* deletion (Patrick et al., 2007); D57N has recently been shown to provide specific compensatory activity for *serB* deletion (Yip & Matsumura, 2013).)

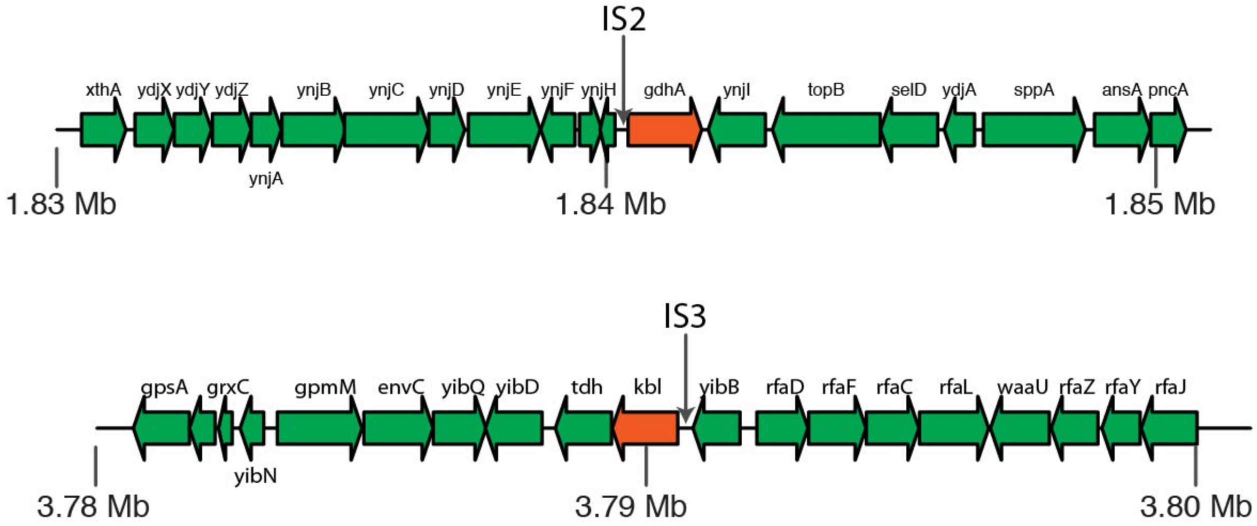

Figure S4.7: Examples of IS element insertions that are likely to be adaptive. The insertion locations are indicated with an arrow. **Top:** an IS2 insertion occurred in one of four Δ*carA* clones, upstream of *gdhA*. *carA* hydrolyzes glutamine to glutamate; *gdhA* acts in the glutamate biosynthesis pathway. **Bottom:** an IS3 insertion occurred in three of five Δ*glyA* clones. It appeared upstream of *kbl*. *glyA* converts serine to glycine; *kbl* acts in the threonine to glycine pathway.

Deletion	**Recovered Replicates**	**Deletion**	**Recovered Replicates**	**Deletion**	**Recovered Replicates**
Δ*argA*	0	Δ*hisH*	5	Δ*pheA*	0
Δ*argB*	1	Δ*hisI*	0	Δ*ppc*	5
Δ*argC*	2	Δ*ilvA*	1	Δ*proA*	0
Δ*argE*	0	Δ*ilvC*	0	Δ*proB*	4
Δ*argG*	0	Δ*ilvE*	2	Δ*proC*	0
Δ*argH*	0	Δ*iscS*	0	Δ*ptsI*	5
Δ*aroA*	0	Δ*leuA*	0	Δ*purC*	0
Δ*aroC*	0	Δ*leuB*	0	Δ*purD*	0
Δ*aroD*	0	Δ*leuC*	0	Δ*purE*	0
Δ*aroE*	0	Δ*leuD*	0	Δ*purF*	0
Δ*bioA*	0	Δ*leuL*	5	Δ*purK*	0
Δ*bioB*	0	Δ*lipA*	0	Δ*pyrB*	0
Δ*bioC*	0	Δ*lpd*	1	Δ*pyrC*	0
Δ*bioF*	0	Δ*lysA*	0	Δ*pyrD*	0
Δ*bioH*	0	Δ*metB*	0	Δ*pyrE*	0
Δ*carA*	4	Δ*metE*	0	Δ*pyrF*	0
Δ*carB*	0	Δ*metF*	0	Δ*serA*	0
Δ*cysB*	2	Δ*metL*	5	Δ*serB*	5
Δ*cysE*	0	Δ*metR*	2	Δ*serC*	0
Δ*fes*	1	Δ*nadA*	0	Δ*thrA*	5
Δ*gltA*	4	Δ*nadB*	0	Δ*thrB*	0
Δ*glyA*	5	Δ*nadC*	0	Δ*thrC*	0
Δ*guaB*	0	Δ*pabA*	1	Δ*trpA*	0
Δ*hisA*	0	Δ*pabB*	0	Δ*trpB*	0
Δ*hisB*	0	Δ*panB*	0	Δ*trpC*	0
Δ*hisC*	0	Δ*panC*	0	Δ*trpD*	0
Δ*hisD*	0	Δ*pdxA*	0	Δ*trpE*	0
Δ*hisF*	0	Δ*pdxH*	0	Δ*tyrA*	1
Δ*hisG*	0	Δ*pdxJ*	0	Δ*yhhK*	2

Table S4.1: The 87 deletion genotypes used for experimental evolution. The first column indicates the deletion genotype; the second column indicates the number of population replicates that were recovered, out of five.

Deletion strain	**Gene**	**Location**	**Mutation**
Δ*glyA*	*aceE*	125,336	R774C CGC→TGC
Δ*ppc*	*hemL*	174,831	G18C GGT→TGT
Δ*yhhK*	*hemB*	388,339	R205S CGT→AGT
Δ*metR*	*cyoE*	446,083	F283L TTT→CTT
Δ*hisH*	*cyoE*	446,680	Δ10 bp (241 250/891 coding)
Δ*proB*	*cyoB*	449,021	Δ1 bp (845/1992 coding)
Δ*ptsI*	*trpL/yciV*	1,321,160	A→G (54/ 84)
Δ*leuL*	*mrp/metG*	2,192,197	T→G (7/ 125)
Δ*metR*	*mqo*	2,304,371	Δ1 bp (406/1647 coding)
Δ*ptsI*	*lrhA*	2,404,225	Q147* CAG→TAG
Δ*glyA*	*lrhA*	2,404,563	Δ9 bp (93 101/939 coding)
Δ*yhhK*	*lrhA*	2,404,563	Δ9 bp (93 101/939 coding)
Δ*leuL*	*lrhA*	2,404,630	6 bp x 2 coding
Δ*cysB*	*eutB*	2,556,085	L206S TTG→TCG
Δ*yhhK*	*ung*	2,715,066	E97D GAA→GAC
Δ*gltA*/Δ*metR*/Δ*ppc*	*kgtP/rrfG*	2,724,089	T→C (-321/+2)
Δ*hisH*	*ygcW/yqcE*	2,898,328	Δ1 bp (33/ 286)
Δ*argC*	*ubiH*	3,050,678	A288E GCG→GAG
Δ*ilvE*	*ilvY*	3,955,330	A172S GCG→TCG
Δ*ptsI*	*cyaA*	3,990,603	Δ7 bp (1428 1434/2547 coding)
Δ*metR*	*yigM*	4,009,898	Δ2 bp (800 801/900 coding)
Δ*metL*	*metB*	4,127,836	Δ1 bp (1142/1161 coding)

Table S4.2: List of 22 mutations present in ancestral deletion clones and which were excluded from all molecular and functional analyses.

Gene	**Number of clones**	**Location**	**Mutation**	**Explanation**
cyoB	1	448,997	L290Q	cyoB inactivation ancestral in Δ*proB* ancestor (Table S4.2)
cyoB	1	449,376	E164*	See Above
cyoB	1	449,574	H98N	See Above
e14 phage	2	1,195,432	15,214bp deletion	Large deletion; observed previously (Charusanti et al., 2010)
fadR - *dadA*	1	1,234,354	2,811 bp deletion	Large deletion
ydeN	1	1,579,737	789 bp deletion	Large deletion
pykF	1	1,754,478	7bp duplication	*pykF* inactivation observed previously (Schneider et al., 2000)
pykF	1	1,754,852	1bp deletion	See Above
cspC	1	1,905,451	1bp deletion	Observed previously; known growth advantage (Rath & Jawali, 2006, Tenaillon et al., 2012)
yebZ	1	1,921,844	303 bp deletion	Large deletion
flhE - *flhD*	1	1,960,762	16,555 bp deletion	Large deletion
cheW - *flhD*	1	1,970,924	5,578 bp deletion	Large deletion
lrhA	1	2,404,563	9bp deletion	Probable *lrhA* inactivation in several ancestral lines (Table S4.2)
kgtP/*rrfG*	1	2,724,089	T→C -321/+2	rrfG mutation in Δ*gltA*, Δ*ppc*, and Δ*metR* ancestors; observed previously(Shiomi & Niki, 2011)
rpoS	1	2,864,823	Q251*	Inactivation observed previously (Notley-McRobb et al., 2002, Tenaillon et al., 2012)
rpoS	1	2,864,858	8bp deletion	See above
rpoS	1	2,865,080	1bp insertion	See above
rpoS	1	2,865,131	W148*	See above
rpoS	1	2,865,192	1bp deletion	See above
rpoS	1	2,865,389	1bp deletion	See above
pyrE/*rph*	1	3,813,830	2bp deletion	Observed previously; known mis-regulation (Jensen, 1993, Conrad et al., 2009)
pyrE/*rph*	1	3,813,831	1bp deletion	See above
pyrE/*rph*	1	3,813,833	1bp deletion	See above
pyrE/*rph*	1	3,813,834	C→T -43/+52	See above
pyrE/*rph*	1	3,813,848	C→T -57/+38	See above
rph	4	3,813,882	82 bp deletion	See above

Table S4.3: List of non-specific laboratory mutations and large deletions that occurred in evolved clones and were excluded from all molecular and functional analyses.

Table S4.4: List of IS element mediated changes found in the evolved lineages. The first column contains the IS element; the second column the location; in the third column are the deletion lines(s) that contained the mutation; in the fourth column is the genic location(s) of the mutation; in the fifth column are the genes that are affected; in the last column we list hypothetical functional and adaptive connections between the location of the IS insertion and the deleted gene.

IS element	**Location**	**Deletion line (number of lineages)**	**Genic location**	**Affected gene(s)**	**Hypothetical functional connection or adaptive effect**
IS150	606,988	Δ*metL* (1/5)	5.536231884	*ybdK* / *hokE*	Unknown
IS5	611,151	Δ*fes* (1/1)	535 / 882 bp	*fepA*	disrupts *fepE*; 1840 bp upstream of *entD* and 887 bp upstream of *fes-ybdZ-entF* operon
IS2	1,220,286	Δ*carA* (1/4)	1463 / 2648 bp	*ycgH*	Unknown
IS2	1,234,293	Δ*lpd* (1/1)	133 / 720 bp	*fadR*	Unknown
IS1	1,651,251	Δ*metR* (2/2)	681 / 1020 bp	*rspB*	Unknown

Continued on next page

Table S4.4 – *Continued from previous page*

IS element	**Location**	**Deletion line (number of lineages)**	**Genic location**	**Affected gene(s)**	**Hypothetical functional connection or adaptive effect**
IS2	1,840,348	Δ*carA* (1/4)	4.395348837	*ynjH/gdhA*	*carA* hydrolyzes glutamine to glutamate; *gdhA* acts in glutamate biosynthesis
IS1	1,868,139	Δ*metR* (1/2)	1161 / 1284 bp	*yeaH*	Unknown
IS186	1,877,853	Δ*lpd* (1/1)	115 / 360 bp	*yeaR*	Unknown
IS1	1,976,183	Δ*ppc* (1/5)	31 / 351 bp	*flhD*	10AA after start codon of *flhD*; probable effect on expression of *flhDC*, most likely due to selection on motility (Barker et al., 2004) or on starvation (Zhong et al., 2009)
IS2	1,976,527	Δ*metL* (2/5)/ Δ*thrA* (1/5) / Δ*gltA* (1/4) / Δ*serB* (1/5)	776 bp IS element deletion	*insA* / *insB*	Probable effect on expression of *flhDC*, most likely due to selection on motility or on starvation (Zhong et al., 2009)
IS1	1,977,402	Δ*hisH* (1/5)	0.444141689	*insA* / *uspC*	See above
IS5	1,977,510	Δ*metL* (1/5)	1.026515152	*insA* / *uspC*	See above
IS30	1,977,533	Δ*thrA* (1/5)	1.209876543	*insA* / *uspC*	See above
IS5	2,404,462	Δ*metR* (1/2)	199 / 939 bp	*lrhA*	Probable *lhrA* inactivation; observed in several ancestral clones, most likely due to selection on motility
IS1	2,434,496	Δ*serB* (1/5)	169 / 1014 bp	*usg*	Unknown
IS186	2,534,334	Δ*ptsI* (4/5)	479 / 510 bp	*crr*	Occurs 10AA before stop codon of *crr* and upstream of a terminator; *ptsI* effects phosphotransfer cascade; *crr* transports glucose for phosphotransfer
IS2	2,770,075	Δ*serB* (4/5)/ Δ*gltA* (1/4)	98 / 153 bp	*yfjU*	Unknown
IS3	3,174,745	Δ*ptsI* (1/5)	109 / 828 bp	*cpdA*	Unknown
IS1	3,725,837	Δ*ilvE* (1/2)	-0.694736842	*yiaB* / *xylB*	Unknown
IS3	3,790,617	Δ*glyA* (3/5)	-0.186956522	*kbl* / *yibB*	*glyA* converts serine to glycine; *kbl* acts in threonine to glycine pathway
IS5	3,936,748	Δ*ptsI* (1/5)	499 / 993 bp	*rbsR*	Unknown
IS1	4,127,837	Δ*metL* (1/5)	1143 / 1161 bp	*metB*	7 AA before stop codon of *metB*; ∼ 350bp upstream of *metF*
IS2	4,540,110	Δ*ilvE* (2/2)	51 / 597 bp	*fimE*	Probable *fimE* inactivation; observed in several ancestral clones, most likely due to selection on motility
IS5	4,540,184	Δ*gltA* (1/4) serB (1/5)	125 / 597 bp	*fimE*	See above
IS5	4,540,331	Δ*ppc* (1/5)	272 / 597 bp	*fimE*	See above
IS186	4,541,631	Δ*serB*(1/5) leuL (1/5)	494 / 549 bp	*fimA*	See above
IS5	4,584,845	Δ*gltA* (2/4)	0.491935484	*hsdR* / *mrr*	Unknown

Table S4.5: List of mutations with ambiguous phenotypic effects (amplifications, mutations located upstream or downstream of two reading frames, and exceptional mutations in coding regions).

Deletion lineage	**Location**	**Mutation**	**Effect**	**Affected gene(s)**	**Inferred functional effect conferring innovation**
Δ*yhhK*	146,706	T→C	0.045801527	*panD*/*yadD*	Changes *panD* protein level: additional coding mutation in *panD*; experimental confirmation of expression change
Δ*yhhK*	146,769	C→A	0.376884422	*panD*/*yadD*	See above
Δ*gltA*	270,987	amplification	0.271-0.371 Mb	100 Kbp ampl.	Changes *prpC* protein level: highly homologous to *gltA*; *prpR* likely mutated to constitutive activity (Palacios & Escalante-Semerena, 2004)
Δ*yhhK*	765,187	C→A	4.45	*mngR*/*mngA*	Changes *mngA* protein level: *mngR* affects only *mngA* (Sampaio et al., 2004)
Δ*proB*	1,078,422	C→T	2.990566038	*putA*/*putP*	Changes *putP* protein level: increases promoter binding energy; *putA* affects *putP* level
Δ*pabA*	1,892,746	C→T	G4S/-83	*yoaH*/ *pabB*	Changes *pabB* protein level: direct functional connection to *pabA*; changes promoter binding energy
Δ*pabA*	1,892,810	G→A	2.894736842	*yoaH*/*pabB*	See above
Δ*hisH*	2,066,000	amplification	2.066 -2.102 Mb	46 Kbp ampl.	Changes *hisF* protein level: previous observation of Δ*hisH* rescue (Patrick et al., 2007); experimental complementation
Δ*serB*	2,069,000	amplification	2.069 -2.285 Mb	216 Kbp ampl.	Changes *hisB* protein level: second line contains a coding mutation in *hisB*; *hisB* overexpression rescues Δ*serB*(Patrick et al., 2007)
Δ*ptsI*	2,507,513	C→A	0.467625899	*glk*/*yfeO*	Changes *glk* protein level: *glk* is known to compensate for *ptsI* mutant (Hernández-Montalvo et al., 2003)
Δ*ptsI*	2,507,513	C→A	-607/G135V	*glk*/*yfeO*	See above; increases promoter binding energy for *glk*
Δ*proB*	3,201,180	Δ91 bp	0.467741935	*cca*/*bacA*	Affects REP element downstream of two reading frames; unknown phenotypic effect; excluded from functional analyses
Δ*ilvE*	3,724,000	amplification	3.724 -3.751 Mb	27 Kbp ampl.	Changes *avtA* protein level: additional coding mutation in *avtA*; *avtA* overexpression rescues Δ*ilvE*
Δ*metR*	4,010,897	A→T	0.324022346	*metR*/*metE*	Changes metE protein level: metR is deleted; functional connection

Continued on next page

Table S4.5 – *Continued from previous page*

Deletion lineage	**Location**	**Mutation**	**Effect**	**Affected gene(s)**	**Inferred functional effect conferring innovation**
Δ*glyA*	4,126,613	T→A	2.37804878	*metJ*/*metB*	Changes both *metBL* protein (operon) level and *metJ*; *metJ* represses *metBL*(Keseler et al., 2011); experimental confirmation of expression change in both
Δ*thrA*	4,126,616	Δ5 bp	2.64	*metJ*/*metB*	See above
Δ*thrA*	4,126,618	Δ1 bp	2.597402597	*metJ*/*metB*	See above
Δ*glyA*	4,126,628	G→A	3.134328358	*metJ*/*metB*	See above
Δ*thrA*	4,127,273	G→A	L193L/-585	*metB*/*metL*	Changes *metL* protein level: *metL* is close paralogue of *thrA*; all four other *thrA* clones contained a single mutation affecting the *metBL* operon; increases promoter binding energy
Δ*ptsI*	4,447,815	C→A	0.823529412	*ppa*/*ytfQ*	Changes *ytfQ* (galactose transport) protein level: three other mutations affect galactose transport in *ptsI* deletion lines

Table S4.6: Predicted (using σ^{70} promoter binding energies and RBS calculator) and measured changes in expression, together with predicted functional effects caused by mutations. For cases in which there is no predicted change in σ^{70} binding or translation initiation, we list the nearest annotated functional element (e.g. the transcription start site (TSS) or repressor binding site (Gama-Castro et al., 2008)). A.U.: arbitrary units. NA: not applicable because change in expression was not quantified.

Deletion	**Recruited gene**	**Location**	**Predicted ancestral σ^{70} binding energy (A.U.)**	**Predicted evolved σ^{70} binding energy (A.U.)**	**Predicted ancestral translation initiation rate (Salis** et al., **2009)**	**Predicted evolved translation initiation rate**	**Predicted fold-change in protein level**	**Measured fold-change in expression**	**Primary functional effect**
Δ*argC*	*proB*	-45	1662	10880	45	45	6.55	∼ 4 (McLoughlin & Copley, 2008)	σ^{70}
Δ*pabA*	*pabB*	-83	915	2869	462	462	3.14	NA	σ^{70}
Δ*fes*	*fiu*	-142	430	758	328	328	1.76	NA	σ^{70}
Δ*ptsI*	*glk*	-607	1033	1640	437	437	1.59	NA	σ^{70}
Δ*metR*	*metE*	-58	1189	1766	5619	5619	1.49	NA	σ^{70}
Δ*proB*	*putP*	-106	1713	2045	447	447	1.19	NA	σ^{70}
Δ*tyrA*	*tas*	-44	707	814	1800	1800	1.15	∼ 2.7 (Johnson et al., 2001)	σ^{70}

Continued on next page

Table S4.6 – *Continued from previous page*

Deletion	**Recruited gene**	**Location**	**Predicted ancestral σ^{70} binding energy (A.U.)**	**Predicted evolved σ^{70} binding energy (A.U.)**	**Predicted ancestral translation initiation rate (Salis** et al., **2009)**	**Predicted evolved translation initiation rate**	**Predicted fold-change in protein level**	**Measured fold-change in expression**	**Primary functional effect**
Δ*glyA*	*cycA*	-87	1085	1435	889	889	1.32	NA	σ^{70}
Δ*ptsI*	*galP*	-161	1988	2275	1374	1374	1.14	NA	σ^{70}
Δ*thrA*	*metL*	-585	753	867	8313	8313	1.15	NA	σ^{70}
Δ*cysB*	*cysP*	-46	379	839	1887	1887	2.22	NA	σ^{70}
Δ*carA*	*carB*	-16	251	252	3309	41138	12.44	~ 1.8 - 2.3	RBS
Δ*yhhK*	*panD*	-12	484	430	474	1459	2.74	~ 2.1	RBS
Δ*ilvE*	*avtA*	-9	987	986	1515	3660	2.41	>1.1	RBS
Δ*carA*	*ydaL*	-1	1626	1626	5342	6690	1.25	NA	RBS
Δ*proB*	*glnL*	-9	1145	1203	274	175	0.67	<0.9	RBS
Δ*pabA*	*pabB*	-19	890	892	462	246	0.53	NA	RBS
Δ*proB*	*glnL*	-11	1148	1180	274	11	0.04	NA	RBS
Δ*leuL*	*leuA*	-52	157	156	918	918	0.99	NA	Disrupts terminator
Δ*leuL*	*leuA*	-66	252	256	918	918	1.02	NA	Disrupts terminator
Δ*leuL*	*leuA*	-67	252	253	918	918	1.01	NA	Disrupts terminator
Δ*leuL*	*leuA*	-67	252	254	918	918	1.01	NA	Disrupts terminator
Δ*yhhK*	*mngA*	-20	586	625	70	70	1.07	NA	Unknown (σ^{70})
Δ*ilvA*	*flhD*	-261	958	943	246	246	0.98	NA	Disrupts Crp activator site
Δ*pabA*	*argD*	-776	338	337	359	359	1	NA	2bp upstream of TSS of pabA (deleted)
Δ*serB*	*argX*	-19	827	827	NA	NA	1	NA	6bp upstream of TSS
Δ*serB*	*argX*	-14	802	802	NA	NA	1	NA	1bp upstream of TSS
Δ*glyA*	*metB*	-82	1589	1595	2619	2619	1	NA	Disrupts MetJ repressor site

Continued on next page

Table S4.6 – *Continued from previous page*

Deletion	**Recruited gene**	**Location**	**Predicted ancestral σ^{70} binding energy (A.U.)**	**Predicted evolved σ^{70} binding energy (A.U.)**	**Predicted ancestral translation initiation rate (Salis** et al.**, 2009)**	**Predicted evolved translation initiation rate**	**Predicted fold-change in protein level**	**Measured fold-change in expression**	**Primary functional effect**
Δ*thrA*	*metB*	-75	1583	1631	2619	2619	1.03	NA	Disrupts MetJ repressor site
Δ*thrA*	*metB*	-77	1587	1596	2619	2619	1.01	NA	Disrupts MetJ repressor site
Δ*ptsI*	*ytfQ*	-140	1310	1330	584	584	1.02	NA	5bp downstream of TSS
Δ*ptsI*	*glk*	-65	563	543	437	437	0.97	NA	unknown
Δ*argC*	*carA*	-42	1632	1564	2014	2014	0.96	NA	unknown
Δ*glyA*	*metB*	-67	1548	1687	2619	2619	1.09	NA	Disrupts MetJ repressor site
Δ*yhhK*	*panD*	-75	546	552	474	474	1.01	$\sim$ 1.4	24bp upstream of TSS

Dataset S1

The tab-delimited file contains information on all the mutations that we observed excluding general lab adaptation, IS element changes, and large deletions (see Tables S4.3 and S4.4). The file contains 13 columns in which the following data are indicated, for each mutation that we observed:
(1) The name of the deletion genotype in which the mutation was observed
(2) The replicate lineage in which the mutation was observed
(3) The Blattner number of the deleted gene
(4) The genomic location, in MG1655, of the mutation
(5) The ancestral allele
(6) The evolved allele
(7) TThe genic location, in MG1655, of the mutation (e.g. the position within the ORF and the amino acid affected, when applicable)
(8) The codon affected, in relevant cases
(9) The recruited gene(s)
(10) The resolved recruited gene, in ambiguous instances
(11) The Blattner number of the resolved recruited gene
(12) The strand of the recruited gene
(13) TThe inferred effect of the mutation (structural, regulatory, synonymous, RNA, or unclassified)

Supportingdata.txt

```
######################
### READ ME
### The tab-delimited file contains information on all the mutations that we observed that are
### likely to confer effects that are relevant to the specific functions that are being
### compensated e.g. they are unlikely to be general lab adaptations.
### The file contains 13 columns in which the following data are indicated, for each mutation
### that we observed:
### 1 The the deletion genotype in which the mutation was observed
### 2 The replicate lineage in which the mutation was observed
### 3 The Blattner number of the deletion gene
### 4 The genomic lacation, in MG1655, of the mutation
### 5 The ancestral allele
### 6 The evolved allele
### 7 The genic location, in MG1655, of the mutation e.g. the position within the ORF and the amino acid affected
### 8 The codon affected, in relevant cases
### 9 The recruited genes
### 10 The resolved recruited gene, in ambiguous instances
### 11 The Blattner number of the resolved recruited gene
### 12 The strand of the recruited gene
### 13 The inferred effect of the mutation structural, regulatory, synonymous, or RNA
######################

deletion	lineage	bnum	genomic_location	ancestral_allele	derived_allele	genic_location	codon	gene_s	aff_gene	bnum	strand	struct_reg
argB	C	b3959	1990832	T	C	N4D	AAT4GAT	pgsA	pgsA	b1912	-1	STR
argC	A	b3958	29606	NA	-4	+411/-42	NA	dapB/carA	carA	b0032	1	REG
argC	A	b3958	259567	C	T	-243/-45	NA	phoE/proB	proB	b0242	1	REG
argC	A	b3958	261874	A	C	E383A	GAA383GCA	proA	proA	b0243	1	STR
argC	B	b3958	259567	C	T	-243/-45	NA	phoE/proB	proB	b0242	1	REG
argC	B	b3958	261874	A	C	E383A	GAA383GCA	proA	proA	b0243	1	STR
carA	B	b0032	30801	C	A	+2/-16	NA	carA/carB	carB	b0033	1	REG
carA	B	b0032	3990075	T	A	D300E	GAT300GAA	cyaA	cyaA	b3806	1	STR
carA	C	b0032	30801	C	A	+2/-16	NA	carA/carB	carB	b0033	1	REG
carA	C	b0032	888013	C	A	D344Y	GAT344TAT	ybjL	ybjL	b0847	-1	STR
carA	C	b0032	1404002	C	A	+329/-1	NA	abgR/ydaL	ydaL	b1340	1	REG
carA	C	b0032	3991043	NA	+A	1868/2547	NA	cyaA	cyaA	b3806	1	STR
carA	D	b0032	30801	C	A	+2/-16	NA	carA/carB	carB	b0033	1	REG
carA	D	b0032	2001617	NA	+TAG	14/1497	NA	fliC	fliC	b1923	-1	STR
carA	E	b0032	30801	C	A	+2/-16	NA	carA/carB	carB	b0033	1	REG
carA	E	b0032	3484742	G	C	G201R	GGT201CGT	crp	crp	b3357	1	STR
cysB	B	b1275	457431	G	C	R261P	CGT261CCT	clpX	clpX	b0438	1	STR
cysB	B	b1275	1157361	NA	-1	270/1434	NA	ptsG	ptsG	b1101	1	STR
cysB	B	b1275	1911452	NA	-1	1389/2049	NA	prc	prc	b1830	-1	STR
cysB	B	b1275	2541596	NA	-1	-46/+258	NA	cysP/ucpA	cysP	b2425	-1	REG
cysB	D	b1275	1718557	G	T	P98T	CCG98ACG	slyA	slyA	b1642	-1	STR
cysB	D	b1275	2541596	NA	-1	-46/+258	NA	cysP/ucpA	cysP	b2425	-1	REG
fes	D	b0585	625494	C	T	Q68fes	D	b0585	839664	G	A	T364M	ACG364ATG	fiu	fiu	b0805	-1	STR
fes	D	b0585	840896	T	A	-142/+123	NA	fiu/mcbA	fiu	b0805	-1	REG
fes	D	b0585	3176378	G	T	G81V	GGC81GTC	tolC	tolC	b3035	1	STR
fes	D	b0585	3578954	G	T	86/90	NA	ryhB	ryhB	b4451	-1	RNA
fes	D	b0585	4020699	T	A	A153A	GCT153GCA	tatB	tatB	b3838	1	SYN
gltA	B	b0720	125514	A	G	D833G	GAC833GGC	aceE	aceE	b0114	1	STR
gltA	B	b0720	270987	NA	99995bp x 12	NA	NA	NA	prpC	b0333	1	REG
gltA	C	b0720	270987	NA	99995bp x 3	NA	NA	NA	prpC	b0333	1	REG
gltA	C	b0720	1236407	A	T	W20R	TGG20AGG	ycgB	ycgB	b1188	-1	STR
gltA	C	b0720	3849010	C	A	V36F	GTT36TTT	ilvN	ilvN	b3670	-1	STR
gltA	C	b1497	1921844	NA	-303	NA	NA	yebY	yebY	b1839	-1	UNK
gltA	D	b0720	270987	NA	99995bp x 12	NA	NA	NA	prpC	b0333	1	REG
gltA	D	b0720	347465	G	T	A68D	GCT68GAT	prpR	prpR	b0330	1	STR
gltA	D	b0720	1142579	C	T	D338N	GAC338AAC	rne	rne	b1084	-1	STR
gltA	D	b0720	4083074	NA	-1	772/3051	NA	fdoG	fdoG	b3894	-1	STR
gltA	E	b0720	87447	G	A	E31K	GAA31AAA	ilvH	ilvH	b0078	1	STR
gltA	E	b0720	270987	NA	99995bp x 3	NA	NA	NA	prpC	b0333	1	REG
gltA	E	b0720	347193	C	T	A159T	GCA159ACA	prpR	prpR	b0330	-1	STR
gltA	E	b0720	3350180	T	G	I290L	ATC290CTC	arcB	arcB	b3210	-1	STR
gltA	E	b0720	4083074	NA	-1	772/3051	NA	fdoG	fdoG	b3894	-1	STR
glyA	A	b2551	254	NA	-49	NA	NA	thrA	thrA	b0002	1	REG
glyA	A	b2551	2449421	A	C	D62E	GAT62GAG	yfcR	yfcR	b2335	-1	STR
glyA	B	b2551	125559	A	T	Y848F	TAT848TTT	aceE	aceE	b0114	1	STR
glyA	B	b2551	1510856	G	C	E6Q	GAG6CAG	ydcT	ydcT	b1441	1	STR
glyA	B	b2551	4126628	G	A	-210/-67	NA	metJ/metB	metB	b3939	1	REG
glyA	C	b2551	992106	C	A	A754A	GCC754GCA	pepN	pepN	b0932	1	SYN
glyA	C	b2551	4427800	NA	+T	+222/-87	NA	fklB/cycA	cycA	b4208	1	REG
glyA	D	b2551	281	NA	-39	+26/-18	NA	thrL/thrA	thrA	b0002	1	REG
glyA	D	b2551	4126613	T	A	-195/-82	NA	metJ/metB	metB	b3939	1	REG
glyA	E	b2551	123617	G	A	G201S	GGT201AGT	aceE	aceE	b0114	1	STR
hisH	A	b2023	2064329	NA	36606bp x 3	NA	NA	NA	hisF	b2025	1	REG
hisH	B	b2023	2064329	NA	36606bp x 3	NA	NA	NA	hisF	b2025	1	REG
hisH	C	b2023	2064329	NA	36606bp x 3	NA	NA	NA	hisF	b2025	1	REG
hisH	C	b2023	3275289	NA	7bp x 2	NA	NA	sohA	sohA	b3129	1	STR
hisH	D	b2023	1446887	C	T	R449C	CGC449TGC	feaB	feaB	b1385	1	STR
hisH	D	b2023	2064329	NA	36606bp x 8	NA	NA	NA	hisF	b2025	1	REG
hisH	D	b2023	3231852	NA	-1	315/417	NA	higA	higA	#N/A	-1	STR
hisH	E	b2023	2064329	NA	36606bp x 5	NA	NA	NA	hisF	b2025	1	REG
hisH	E	b2023	3275289	NA	7bp x 2	NA	NA	sohA	sohA	b3129	1	STR
ilvA	C	b3772	221760	NA	-3	884-886/1032	NA	metN	metN	b0199	-1	STR
ilvA	C	b3772	1332145	G	A	W89ilvA	C	b3772	1976482	G	T	-261/+60	NA	flhD/insB	flhD	b1892	-1	REG
ilvA	C	b3772	4083054	NA	3bp x 2	NA	NA	fdoG	fdoG	b3894	-1	STR
ilvA	C	b3772	4126827	G	C	D45H	GAT45CAT	metB	metB	b3939	1	STR
ilvE	D	b3770	713143	T	A	P121P	CCT121CCA	pgm	pgm	b0688	1	SYN
ilvE	D	b3770	3737719	T	G	+169/-9	NA	malS/avtA	avtA	b3572	1	REG
ilvE	E	b3770	669058	T	G	T25P	ACC25CCC	cobC	cobC	b0638	-1	STR
ilvE	E	b3770	3724467	NA	26992bp x 10	NA	NA	NA	avtA	b3572	1	REG
ilvE	E	b3770	1579737	NA	-789	NA	NA	ydeM	ydeM	b1497	-1	UNK
leuL	A	b0075	83567	NA	-27	-38/+55	NA	leuA/leuL	leuA	b0074	-1	REG
```

```
leuL	A	b0075	3990977	C	A	S601leuL	B	b0075	83596	G	T	-67/+26	NA	leuA/leuL	leuA	b0074	-1	REG
leuL	C	b0075	83596	G	C	-67/+26	NA	leuA/leuL	leuA	b0074	-1	REG
leuL	D	b0075	83581	C	G	-52/+41	NA	leuA/leuL	leuA	b0074	-1	REG
leuL	E	b0075	83595	G	T	-66/+27	NA	leuA/leuL	leuA	b0074	-1	REG
lpd	B	b0116	457101	C	A	A151D	GCC151GAC	clpX	clpX	b0438	1	STR
lpd	B	b0116	756738	C	T	R537C	CGT537TGT	sdhA	sdhA	b0723	1	STR
lpd	B	b0116	1268473	A	G	I9T	ATT9ACT	ldrA	ldrA	b4419	-1	STR
metL	A	b3940	3057931	C	T	R53C	CGT53TGT	argP	argP	b2916	1	STR
metL	A	b3940	3990703	A	T	K510metL	B	b3940	1550192	NA	-178	NA	NA	C0362	C0362	NA	1	RNA
metL	B	b3940	1748359	T	C	Q5R	CAA5CGA	ydhU	ydhU	b1670	-1	STR
metL	B	b3940	3057971	T	G	L66R	CTG66CGG	argP	argP	b2916	1	STR
metL	B	b3940	3484772	T	G	metL	C	b3940	305121	C	A	P307P	CCG307CCT	yagW	yagW	b0290	-1	SYN
metL	C	b3940	1058164	A	C	K286N	AAA286AAC	torC	torC	b0996	1	STR
metL	C	b3940	2387297	C	T	P55S	CCT55TCT	yfbP	yfbP	b2275	1	STR
metL	C	b3940	3057883	NA	-1	109/894	NA	argP	argP	b2916	1	STR
metL	C	b3940	3990611	NA	7bp x 2	NA	NA	cyaA	cyaA	b3806	1	STR
metL	D	b3940	3991450	NA	-3	2275-2277/2547	NA	cyaA	cyaA	b3806	1	STR
metL	E	b3940	2387297	C	T	P55S	CCT55TCT	yfbP	yfbP	b2275	1	STR
metL	E	b3940	3484574	G	C	A145P	GCA145CCA	crp	crp	b3357	1	STR
metR	A	b3828	4010897	A	T	-58/-179	NA	metR/metE	metE	b3829	1	REG
metR	E	b3828	4010897	A	T	-58/-179	NA	metR/metE	metE	b3829	1	REG
pabA	D	b3360	1892746	C	T	G4S/-83	GGT4AGT	pabB/yoaH	pabB	b1812	1	REG
pabA	D	b3360	1892810	G	A	-55/-19	NA	yoaH/pabB	pabB	b1812	1	REG
pabA	D	b3360	3488876	NA	-3	-25/+5	NA	pabA/fic	argD	b3359	-1	REG
ppc	A	b3956	1195251	G	C	M302I	ATG302ATC	icd	icd	b1136	1	STR
ppc	B	b3956	1195325	A	G	N327S	AAC327AGC	icd	icd	b1136	1	STR
ppc	B	b3956	1234354	NA	-2811	NA	NA	NA	NA	NA	NA	UNK
ppc	C	b3956	1194823	T	G	Y160D	TAT160GAT	icd	icd	b1136	1	STR
ppc	D	b3956	1195251	G	A	M302I	ATG302ATA	icd	icd	b1136	1	STR
ppc	E	b3956	1156938	NA	-64	+141/-155	NA	ycfH/ptsG	ptsG	b1101	1	REG
ppc	E	b3956	1195316	C	T	P324L	CCT324CTT	icd	icd	b1136	1	STR
ppc	E	b3956	1327731	G	A	A126T	GCT126ACT	sohB	sohB	b1272	1	STR
proB	B	b0242	1236374	C	A	E31proB	B	b0242	3201180	NA	-91	+29/+62	NA	cca/bacA	NA	NA	NA	REG
proB	B	b0242	4054371	C	A	-9/+277	NA	glnL/glnA	glnL	b3869	-1	REG
proB	B	b0242	4055598	C	T	D154N	GAT154AAT	glnA	glnA	b3870	-1	STR
proB	C	b0242	2015065	T	C	V163A	GTG163GCG	fliI	fliI	b1941	1	STR
proB	C	b0242	2866139	NA	-12	626-637/1140	NA	nlpD	nlpD	b2742	-1	STR
proB	C	b0242	4054373	C	G	-11/+275	NA	glnL/glnA	glnL	b3869	-1	REG
proB	C	b0242	4055405	C	G	G218A	GGT218GCT	glnA	glnA	b3870	-1	STR
proB	C	b0242	4398617	T	G	proB	D	b0242	1078422	C	T	-317/-106	NA	putA/putP	putP	b1015	1	REG
proB	D	b0242	1699063	C	T	A28V	GCT28GTT	malY	malY	b1622	1	STR
proB	D	b0242	3383022	A	C	T100P	ACC100CCC	argR	argR	b3237	1	STR
proB	D	b0242	3487423	NA	-3	778-780/1221	NA	argD	argD	b3359	-1	STR
proB	E	b0242	4055615	G	A	S148F	TCC148TTC	glnA	glnA	b3870	-1	STR
ptsI	A	b2416	1861498	C	T	T235I	ACC235ATC	gapA	gapA	b1779	1	STR
ptsI	A	b2416	2507513	C	A	-65/-139	NA	glk/yfeO	glk	b2388	-1	REG
ptsI	A	b2416	3086145	NA	+T	+263/-161	NA	metK/galP	galP	b2943	1	REG
ptsI	A	b2416	3820302	NA	-1	174/276	NA	rpoZ	rpoZ	b3649	1	STR
ptsI	B	b2416	2966703	C	A	W99C	TGG99TGT	rppH	rppH	b2830	-1	STR
ptsI	B	b2416	2974640	T	C	V7A	GTA7GCA	galR	galR	b2837	1	STR
ptsI	B	b2416	3456256	G	T	D620Y	GAC620TAC	gspD	gspD	b3325	1	STR
ptsI	B	b2416	4245227	C	A	R141S	CGT141AGT	malK	malK	b4035	1	STR
ptsI	D	b2416	2924606	T	C	L93L	TTG93CTG	ygdH	ygdH	b2795	1	SYN
ptsI	D	b2416	2974707	C	A	S29R	AGC29AGA	galR	galR	b2837	1	STR
ptsI	D	b2416	3331430	NA	9bp x 2	NA	NA	rplU	rplU	b3186	-1	STR
ptsI	D	b2416	4447815	C	A	-140/-170	NA	ppa/ytfQ	ytfQ	b4226	-1	REG
ptsI	E	b2416	2508055	G	T	-607/G135V	GGC135GTC	glk/yfeO	glk	b2388	-1	REG
serB	A	b4388	2091660	G	A	D57N	GAT57AAT	hisB	hisB	b2022	1	STR
serB	B	b4388	2091558	C	A	Q23K	CAG23AAG	hisB	hisB	b2022	1	STR
serB	B	b4388	2706482	NA	+G	295/957	NA	rseB	rseB	b2571	-1	STR
serB	B	b4388	3268394	G	A	221/377	NA	rnpB	rnpB	b3123	-1	RNA
serB	C	b4388	2091660	G	A	D57N	GAT57AAT	hisB	hisB	b2022	1	STR
serB	C	b4388	3813934	NA	-16	624-639/687	NA	rph	rph	b3643	-1	STR
serB	C	b4388	3980379	NA	+T	+84/-19	NA	yifK/argX	argX	b3796	1	REG
serB	D	b4388	104121	A	G	D47G	GAT47GGT	ftsA	ftsA	b0094	1	STR
serB	D	b4388	2064329	NA	223774bp x 2	NA	NA	hisB	hisB	b2022	1	REG
serB	D	b4388	3268473	G	A	142/377	NA	rnpB	rnpB	b3123	-1	RNA
serB	E	b4388	2094128	NA	6bp x 2	NA	NA	hisF	hisF	b2025	1	STR
serB	E	b4388	3625116	T	G	I816L	ATC816CTC	rbbA	rbbA	b3486	-1	STR
serB	E	b4388	3980384	NA	+C	+89/-14	NA	yifK/argX	argX	b3796	1	REG
thrA	A	b0002	4126616	NA	-5	-198/-75	NA	metJ/metB	metB	b3939	1	REG
thrA	B	b0002	4126618	NA	-1	-200/-77	NA	metJ/metB	metB	b3939	1	REG
thrA	C	b0002	4126618	NA	-1	-200/-77	NA	metJ/metB	metB	b3939	1	REG
thrA	D	b0002	4126249	A	C	L57R	CTG57CGG	metJ	metJ	b3939	1	STR
thrA	E	b0002	4127273	G	A	L193L/-585	TTG193TTA	metB/metL	metL	b3940	1	REG
tyrA	D	b2600	1476839	T	C	V197A	GTC197GCC	ynbB	ynbB	b1409	1	STR
tyrA	D	b2600	1959225	C	A	F380L	TTC380TTA	argS	argS	b1876	1	STR
tyrA	D	b2600	2969575	C	T	+64/-44	NA	ygdR/tas	tas	b2834	1	REG
tyrA	D	b2600	2970547	C	A	T310N	ACC310AAC	tas	tas	b2834	1	STR
tyrA	D	b2600	2970560	G	T	Q314H	CAG314CAT	tas	tas	b2834	1	STR
yhhK	B	b3459	146680	C	A	M5I	ATG5ATT	panD	panD	b0131	-1	STR
yhhK	B	b3459	146769	C	A	-75/-199	NA	panD/yadD	panD	b0131	-1	REG
yhhK	B	b3459	179876	C	A	R214S	CGT214AGT	dgt	dgt	b0160	1	STR
yhhK	B	b3459	4375123	G	C	V76L	GTG76CTG	sugE	sugE	b4148	1	STR
yhhK	C	b3459	146706	T	C	-12/-262	NA	panD/yadD	panD	b0131	-1	REG
yhhK	C	b3459	765187	C	A	-89/-20	NA	mngR/mngA	mngA	b0731	1	REG
yhhK	C	b3459	1920939	T	A	I19F	ATC19TTC	pphA	pphA	b1838	-1	STR
yhhK	C	b3459	3157314	C	T	A619T	GCG619ACG	ygiQ	ygiQ	b4469	-1	STR
```

Chapter 5

Summary and future perspectives

In the chapters presented, questions concerning the evolution of transcriptional regulation in *E.coli* were addressed. The work allows a better understanding of (i) why and (ii) how transcriptional control evolves.

(i) Transcriptional regulation is important to allow different genes in the genome to be expressed at different levels and, moreover, provides the possibility for differential gene expression of single genes in varying conditions. If a promoter sequence is able to establish the most beneficial quantitative transcriptional output, it will most likely not be subject to changes modifying the expression level.

However, if a gene is not expressed at the right level, changes in the genotype of the promoter will be selected upon that give raise to the desired protein amounts in the cell. Subtle changes in the promoter sequence with small effects on the RNA expression level can have a severe impact on the fitness level of the organism, as shown in Chapter 4. Due to the exponential growth of bacterial populations, beneficial mutations can spread easily and take over in the population. To further quantitatively elucidate the fitness effect of expression level of certain genes on a single cell, expression level and growth rate could be directly tracked in a microfluidic device.

Fluctuating environments may ask for diversification of desired expression levels for a single gene, and this information can be encoded in the promoter sequence as well. The activity of a gene upon a stimuli can only be changed, if the environment or internal state of the cell can be sensed and passed on to the gene required in this situation. If either sensory mechanisms are not able to track the changes or transcription factor activity cannot account for the desired levels, a gene can at least broaden its expression distribution across cells. One way of achieving this is the incorporation of binding sites for transcription factors in the promoter sequence, preferentially for factors that have a broad expression distribution themselves as highlighted in detail in Chapter 2. Synthetic promoter sequences did not show as high variation in their expression levels across cells and sites for specific transcription factors were not predicted. To evolve promoter sequences with variations in expression levels comparable to native promoters, promoters may be selected on a selection scheme, where the desired expression level varied from round to round. Phenotypes of those promoters could be determined by flow cytometry, and their sequenced genotype could reveal features in the sequence that explain these variations. Those features may be, for instance, binding sites for transcription factors.

(ii) Transcriptional regulation can evolve in many ways, resulting in multiple outcomes, which is already depicted in the diversity of promoter sequences present in the genome of *E.coli* .

In Chapter 2 and 3, evolution of expression levels selected was possible in short time-scale with high mutation rates and flow cytometry based selection. Starting from a population of random sequences, expression level distributions as observed for native promoter sequences were obtained. Evolution of a functional promoter sequence was possible from a million random sequences, illustrating that regulatory sequences were evolvable *de novo*.

In a step-wise approach, the desired expression level could be reached by changing expression levels of already functional sequences by point mutations. Changing expression levels from medium expression to high expression was possible within only two rounds of evolution. Modifications in the genotype where shown to impact expression levels also in synthetically evolved promoter sequences and that many different sequences resulted in similar expression levels. Some mutations were shown to have a greater impact on the transcription rate than others. Promoter sequence features that were tackled during the artificial selection were the inclusion of σ^{70} binding sites and an increase in AT content. The evolutionary dynamics directed towards the integration of only a few σ^{70} binding sites per promoter sequence. These features can possibly explain what are the minimal requirements for a functional promoter sequence, but not the exact rate of transcription in the cell. The prediction of transcriptional activity for a given promoter sequence does not solely depend on the strength of σ^{70} sites present in the promoter as presented in Chapter 3. Sequence properties that impact other mechanisms during for transcription initiation, like the unwinding of the double strand and clearance of the promoter sequence prior to elongation of the final transcript might have a profound impact on the transcription rate as well. A more detailed analysis of the promoters obtained after five rounds of evolution could be provided by grouping those sequences into more fine grained expression classes than medium and high expression. By experimental examination of the transcription start sites for each promoter, the exact position(s) of sigma site(s) could be identified and not only predicted.

The importance of regulatory changes on the fitness of organisms was reassured in Chapter 4. In addition, changes in the transcription rate were more common than changes in the translation rate, when changes in protein levels accounted for fitness effects. Marginal upregulation of some genes were shown to have a large effect on the growth rates. In some cases, regulatory modifications were obtained by single point mutations, and some of those mutations were affecting the strength of the σ^{70} binding site. Tuning transcription rate by the strength of sigma binding is possible in otherwise fixed sequence backgrounds. However, other mutations can also have large effects on the expression strength as depicted in Chapter 3.

References

Ackermann, M., Stecher, B., Freed, N. E., Songhet, P., Hardt, W.-D. & Doebeli, M. (2008). Nature 454, 987–990.

Ando, H., Miyoshi-Akiyama, T., Watanabe, S. & Kirikae, T. (2013). Mol Microbiol .

Arnold, P., Erb, I., Pachkov, M., Molina, N. & van Nimwegen, E. (2012). Bioinformatics 28, 487–494.

Aronson, B. D., Levinthal, M. & Somerville, R. L. (1989). J Bacteriol 171, 5503–5511.

Arriaga, E. A. (2009). Anal Bioanal Chem 393, 73–80.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. (2006). Mol Syst Biol 2, 2006.0008.

Babu, M. M. & Aravind, L. (2006). Trends Microbiol 14, 11–14.

Babu, M. M., Balaji, S. & Aravind, L. (2007). Genome Dyn 3, 66–80.

Babu, M. M. & Teichmann, S. A. (2003). Nucleic Acids Res 31, 1234–1244.

Bajić, D. & Poyatos, J. F. (2012). BMC Genomics 13, 343.

Balderas-Martínez, Y. I., Savageau, M., Salgado, H., Pérez-Rueda, E., Morett, E. & Collado-Vides, J. (2013). PLoS One 8, e65723.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. & Barkai, N. (2006). Nat Genet 38, 636–643.

Barkai, N. & Shilo, B.-Z. (2007). Mol Cell 28, 755–760.

Barker, C. S., Prüss, B. M. & Matsumura, P. (2004). J Bacteriol 186, 7529–7537.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E. & Kim, J. F. (2009). Nature 461, 1243–1247.

Battesti, A., Majdalani, N. & Gottesman, S. (2011). Annu Rev Microbiol 65, 189–213.

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C. & Rainey, P. B. (2009). Nature 462, 90–93.

Berg, O. G. & von Hippel, P. H. (1987). J Mol Biol 193, 723–750.

Bergthorsson, U., Andersson, D. I. & Roth, J. R. (2007). Proc Natl Acad Sci U S A 104, 17004–17009.

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S. & Cohen, S. N. (2002). Proc Natl Acad Sci U S A 99, 9697–9702.

Bernstein, J. A., Lin, P.-H., Cohen, S. N. & Lin-Chao, S. (2004). Proc Natl Acad Sci U S A 101, 2758–2763.

Bishop, A. L., Rab, F. A., Sumner, E. R. & Avery, S. V. (2007). Mol Microbiol 63, 507–520.

Blake, W. J., Balázsi, G., Kohanski, M. A., Isaacs, F. J., Murphy, K. F., Kuang, Y., Cantor, C. R., Walt, D. R. & Collins, J. J. (2006). Mol Cell 24, 853–865.

Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. (2003). Nature 422, 633–637.

Blank, D., Wolf, L., Ackermann, M. & Silander, O. K. (2013). under review at PNAS .

Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. (2012). Nature 489, 513–518.

Bochkareva, A. & Zenkin, N. (2013). Nucleic Acids Res 41, 4565–4572.

Bollenbach, T. & Kishony, R. (2011). Mol Cell 42, 413–425.

Boor, K. J. (2006). PLoS Biol 4, e23.

Bosma, T., Damborský, J., Stucki, G. & Janssen, D. B. (2002). Appl Environ Microbiol 68, 3582–3587.

Brewster, R. C., Jones, D. L. & Phillips, R. (2012). PLoS Comput Biol 8, e1002811.

Brilli, M. & Fani, R. (2004). Gene 339, 149–160.

Bull, J. (1987). Evolution 41, 303—315.

Burgess, R. R. & Anthony, L. (2001). Curr Opin Microbiol 4, 126–131.

Burr, T., Mitchell, J., Kolb, A., Minchin, S. & Busby, S. (2000). Nucleic Acids Res 28, 1864–1870.

Cai, L., Dalal, C. K. & Elowitz, M. B. (2008). Nature 455, 485–490.

Cai, L., Friedman, N. & Xie, X. S. (2006). Nature 440, 358–362.

Carey, L. B., van Dijk, D., Sloot, P. M. A., Kaandorp, J. A. & Segal, E. (2013). PLoS Biol 11, e1001528.

Carrier, T. A. & Keasling, J. D. (1997). Biotechnol Bioeng 55, 577–580.

Carrier, T. A. & Keasling, J. D. (1999). Biotechnol Prog 15, 58–64.

Carroll, S. B. (2008). Cell 134, 25–36.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E. & Vidal, M. (2012). Nature 487, 370–374.

Casadesús, J. & Low, D. (2006). Microbiol Mol Biol Rev 70, 830–856.

Charusanti, P., Chauhan, S., McAteer, K., Lerman, J. A., Hyduke, D. R., Motin, V. L., Ansong, C., Adkins, J. N. & Palsson, B. O. (2011). BMC Syst Biol 5, 163.

Charusanti, P., Conrad, T. M., Knight, E. M., Venkataraman, K., Fong, N. L., Xie, B., Gao, Y. & Palsson, B. . (2010). PLoS Genet 6, e1001186.

Conrad, T. M., Joyce, A. R., Applebee, M. K., Barrett, C. L., Xie, B., Gao, Y. & Palsson, B. . (2009). Genome Biol 10, R118.

Crick, F. (1970). Nature 227, 561–563.

Crick, F. H. (1958). Symp Soc Exp Biol 12, 138–163.

Dabizzi, S., Ammannato, S. & Fani, R. (2001). Res Microbiol 152, 539–549.

Dekel, E. & Alon, U. (2005). Nature 436, 588–592.

Desai, M. M. & Fisher, D. S. (2007). Genetics 176, 1759–1798.

Djordjevic, M. (2011). J Bacteriol 193, 6305–6314.

Djordjevic, M. & Bundschuh, R. (2008). Biophys J 94, 4233–4248.

Dong, D., Shao, X., Deng, N. & Zhang, Z. (2011). Nucleic Acids Res 39, 403–413.

Elena, S. F. & Lenski, R. E. (2003). Nat Rev Genet 4, 457–469.

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002). Science 297, 1183–1186.

Espinosa, V., González, A. D., Vasconcelos, A. T., Huerta, A. M. & Collado-Vides, J. (2005). J Mol Biol 354, 184–199.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. & Palsson, B. . (2007). Mol Syst Biol 3, 121.

Ferenci, T. (2003). Trends Microbiol 11, 457–461.

Fogle, C. A., Nagle, J. L. & Desai, M. M. (2008). Genetics 180, 2163–2173.

Fong, S. S., Joyce, A. R. & Palsson, B. . (2005). Genome Res 15, 1365–1372.

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen, M. B. (2004). PLoS Biol 2, e137.

Fröhlich, K. S. & Vogel, J. (2009). Curr Opin Microbiol 12, 674–682.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). Bioinformatics 28, 3150–3152.

Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A. M., Treviño-Quintanilla, L. & Collado-Vides, J. (2008). Nucleic Acids Res 36, D120–D124.

Gilad, Y., Oshlack, A. & Rifkin, S. A. (2006). Trends Genet 22, 456–461.

Goldberg, A. L. & John, A. C. S. (1976). Annu Rev Biochem 45, 747–803.

Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. (2005). Cell 123, 1025–1036.

Goodman, D. B., Church, G. M. & Kosuri, S. (2013). Science 342, 475–479.

Gruber, T. M. & Gross, C. A. (2003). Annu Rev Microbiol 57, 441–466.

Haag, E. S. & Lenski, R. E. (2011). Development 138, 2633–2637.

Haccou, P. & Iwasa, Y. (1995). Theoretical Population Biology 47, 212—-243.

Haugen, S. P., Ross, W., Manrique, M. & Gourse, R. L. (2008). Proc Natl Acad Sci U S A 105, 3292–3297.

Helmann, J. D. & Chamberlin, M. J. (1988). Annu Rev Biochem 57, 839–872.

Hernández-Montalvo, V., Martínez, A., Hernández-Chavez, G., Bolivar, F., Valle, F. & Gosset, G. (2003). Biotechnol Bioeng 83, 687–694.

Hershberg, R. & Margalit, H. (2006). Genome Biol 7, R62.

Hershberg, R. & Petrov, D. A. (2010). PLoS Genet 6, e1001115.

Hoekstra, H. E. & Coyne, J. A. (2007). Evolution 61, 995–1016.

Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. (2006). Science 313, 101–104.

Hsu, L. M. (2002). Biochim Biophys Acta 1577, 191–207.

Jacob, F. (1977). Science 196, 1161–1166.

Jacob, F. & Monod, J. (1961). Cold Spring Harb Symp Quant Biol 26, 193—211.

Jensen, K. F. (1993). J Bacteriol 175, 3401–3407.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. & von Mering, C. (2009). Nucleic Acids Res 37, D412–D416.

Johnson, J. M., Ding, W., Henkhaus, J. & Fix, D. (2001). Mutat Res 479, 121–130.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., Brady, S. D., Zhang, H., Pollen, A. A., Howes, T., Amemiya, C., Team, B. I. G. S. P. . W. G. A., Baldwin, J., Bloom, T., Jaffe, D. B., Nicol, R., Wilkinson, J., Lander, E. S., Palma, F. D., Lindblad-Toh, K. & Kingsley, D. M. (2012). Nature 484, 55–61.

Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. (2005). Nat Rev Genet 6, 451–464.

Kaessmann, H. (2010). Genome Res 20, 1313–1326.

Kasak, L., Hõrak, R. & Kivisaar, M. (1997). Proc Natl Acad Sci U S A 94, 3134–3139.

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A. G., Mackie, A., Paulsen, I., Gunsalus, R. P. & Karp, P. D. (2011). Nucleic Acids Res 39, D583–D590.

Kill, K., Binnewies, T. T., Sicheritz-Pontén, T., Willenbrock, H., Hallin, P. F., Wassenaar, T. M. & Ussery, D. W. (2005). Microbiology 151, 3147–3150.

King, M. C. & Wilson, A. C. (1975). Science 188, 107–116.

Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. (2010). Proc Natl Acad Sci U S A 107, 9158–9163.

Kiryu, H., Oshima, T. & Asai, K. (2005). Bioinformatics 21, 1062–1068.

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H. & Mori, H. (2005). DNA Res 12, 291–299.

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D. & Church, G. M. (2013). Proc Natl Acad Sci U S A 110, 14024–14029.

Krell, T., Lacal, J., Busch, A., Silva-Jiménez, H., Guazzaroni, M.-E. & Ramos, J. L. (2010). Annu Rev Microbiol 64, 539–559.

Kussell, E. & Leibler, S. (2005). Science 309, 2075–2078.

Lange, R. & Hengge-Aronis, R. (1994). Genes Dev 8, 1600–1612.

Lee, J. M., Lee, J., Kim, T. & Lee, S. K. (2013). PLoS One 8, e52382.

Lehner, B. (2008). Mol Syst Biol 4, 170.

Lehner, B. (2010). PLoS Genet 6, e1001185.

Lehner, B. & Kaneko, K. (2011). Cell Mol Life Sci 68, 1005–1010.

Lemos, B., Meiklejohn, C. D., CÃ¡ceres, M. & Hartl, D. L. (2005). Evolution 59, 126–137.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Subgroup, . G. P. D. P. (2009). Bioinformatics 25, 2078–2079.

Li, J., Min, R., Vizeacoumar, F. J., Jin, K., Xin, X. & Zhang, Z. (2010). Proc Natl Acad Sci U S A 107, 10472–10477.

Li, W. & Godzik, A. (2006). Bioinformatics 22, 1658–1659.

Lieberman, T. D., Michel, J.-B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M. R., Skurnik, D., Leiby, N., LiPuma, J. J., Goldberg, J. B., McAdam, A. J., Priebe, G. P. & Kishony, R. (2011). Nat Genet 43, 1275–1280.

Lin, E. & Hacking, A. J. A. J. (1976). BioScience 26, 548–555.

Lindsey, H. A., Gallie, J., Taylor, S. & Kerr, B. (2013). Nature 494, 463–467.

Liu, S. & Libchaber, A. (2006). J Mol Evol 62, 536–550.

Lloyd, G., Landini, P. & Busby, S. (2001). Essays Biochem 37, 17–31.

Lozada-Chávez, I., Janga, S. C. & Collado-Vides, J. (2006). Nucleic Acids Res 34, 3434–3445.

Maeda, H., Fujita, N. & Ishihama, A. (2000). Nucleic Acids Res 28, 3497–3503.

Massé, E., Vanderpool, C. K. & Gottesman, S. (2005). J Bacteriol 187, 6962–6971.

Maurizi, M. R. (1992). Experientia 48, 178–201.

Mayr, E. (1963). Animal Species and Evolution. Harvard University Press, Cambridge, MA.

McAdams, H. H. & Arkin, A. (1997). Proc Natl Acad Sci U S A 94, 814–819.

McAdams, H. H., Srinivasan, B. & Arkin, A. P. (2004). Nat Rev Genet 5, 169–178.

McCracken, A., Turner, M. S., Giffard, P., Hafner, L. M. & Timms, P. (2000). Arch Microbiol 173, 383–389.

McKenzie, G. J. & Craig, N. L. (2006). BMC Microbiol 6, 39.

McLean, B. W., Wiseman, S. L. & Kropinski, A. M. (1997). Can J Microbiol 43, 981–985.

McLoughlin, S. Y. & Copley, S. D. (2008). Proc Natl Acad Sci U S A 105, 13497–13502.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T. & Lenski, R. E. (2012). Science 335, 428–432.

Meyer, M. & Kircher, M. (2010). Cold Spring Harb Protoc 2010, pdb.prot5448.

Molina, N. & van Nimwegen, E. (2008). Genome Res 18, 148–160.

Munsky, B., Neuert, G. & van Oudenaarden, A. (2012). Science 336, 183–187.

Mutero, A., Pralavorio, M., Bride, J. M. & Fournier, D. (1994). Proc Natl Acad Sci U S A 91, 5922–5926.

Myers, K. S., Yan, H., Ong, I. M., Chung, D., Liang, K., Tran, F., Keleş, S., Landick, R. & Kiley, P. J. (2013). PLoS Genet 9, e1003565.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L. & Weissman, J. S. (2006). Nature 441, 840–846.

Notley-McRobb, L., King, T. & Ferenci, T. (2002). J Bacteriol 184, 806–811.

Näsvall, J., Sun, L., Roth, J. R. & Andersson, D. I. (2012). Science 338, 384–387.

Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Nature 405, 299–304.

Ohno, S. (1970). Springer, New York .

Palacios, S. & Escalante-Semerena, J. C. (2004). Microbiology 150, 3877–3887.

Patrick, W. M., Quandt, E. M., Swartzlander, D. B. & Matsumura, I. (2007). Mol Biol Evol 24, 2716–2722.

Pettis, G. S., Brickman, T. J. & McIntosh, M. A. (1988). J Biol Chem 263, 18857–18863.

Pigliucci, M. (2008). Philosophy of Science 75, 887—-898.

Prévost, K., Salvail, H., Desnoyers, G., Jacques, J.-F., Phaneuf, E. & Massé, E. (2007). Mol Microbiol 64, 1260–1273.

Price, T. D., Qvarnström, A. & Irwin, D. E. (2003). Proc Biol Sci 270, 1433–1440.

Raj, A. & van Oudenaarden, A. (2008). Cell 135, 216–226.

Rajewsky, N., Socci, N. D., Zapotocky, M. & Siggia, E. D. (2002). Genome Res 12, 298–308.

Rangarajan, E. S., Proteau, A., Wagner, J., Hung, M.-N., Matte, A. & Cygler, M. (2006). J Biol Chem 281, 37930–37941.

Raser, J. M. & O'Shea, E. K. (2004). Science 304, 1811–1814.

Raser, J. M. & O'Shea, E. K. (2005). Science 309, 2010–2013.

Rath, D. & Jawali, N. (2006). J Bacteriol 188, 6780–6785.

Re, S. D., Tolstykh, T., Wolanin, P. M. & Stock, J. B. (2002). Protein Sci 11, 2644–2654.

Revyakin, A., Ebright, R. H. & Strick, T. R. (2004). Proc Natl Acad Sci U S A 101, 4776–4780.

Rhodius, V. A. & Mutalik, V. K. (2010). Proc Natl Acad Sci U S A 107, 2854–2859.

Rhodius, V. A., Mutalik, V. K. & Gross, C. A. (2012). Nucleic Acids Res 40, 2907–2924.

Rifkin, S. A., Kim, J. & White, K. P. (2003). Nat Genet 33, 138–144.

Rolfe, M. D., Ocone, A., Stapleton, M. R., Hall, S., Trotter, E. W., Poole, R. K., Sanguinetti, G., Green, J. & Consortium, S. O.-S. U. M. O. (2012). Open Biol 2, 120091.

Ross, W., Ernst, A. & Gourse, R. L. (2001). Genes Dev 15, 491–506.

Rowley, D. L., Fawcett, W. P. & Wolf, R. E. (1992). J Bacteriol 174, 623–626.

Roymondal, U., Das, S. & Sahoo, S. (2009). DNA Res 16, 13–30.

Salgado, H., Gama-Castro, S., Martínez-Antonio, A., Díaz-Peredo, E., Sánchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jiménez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martínez, C. & Collado-Vides, J. (2004). Nucleic Acids Res 32, D303–D306.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, A., Porrón-Sotelo, L., Huerta, A. M., Bonavides-Martínez, C., Balderas-Martínez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Moral-Chávez, V. D., Hernández-Alvarez, A., Morett, E. & Collado-Vides, J. (2013). Nucleic Acids Res 41, D203–D213.

Salis, H. M. (2010).

Salis, H. M., Mirsky, E. A. & Voigt, C. A. (2009). Nat Biotechnol 27, 946–950.

Sampaio, M.-M., Chevance, F., Dippel, R., Eppler, T., Schlegel, A., Boos, W., Lu, Y.-J. & Rock, C. O. (2004). J Biol Chem 279, 5537–5548.

Sanchez, A., Garcia, H. G., Jones, D., Phillips, R. & Kondev, J. (2011). PLoS Comput Biol 7, e1001100.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). Proc Natl Acad Sci U S A 74, 5463–5467.

Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. (2000). Genetics 156, 477–488.

Schneider, T. D. (1997). J Theor Biol 189, 427–441.

Serres, M. H., Kerr, A. R. W., McCormack, T. J. & Riley, M. (2009). Biol Direct 4, 46.

Serres, M. H. & Riley, M. (2000). Microb Comp Genomics 5, 205–222.

Shachrai, I., Zaslaver, A., Alon, U. & Dekel, E. (2010). Mol Cell 38, 758–767.

Shahrezaei, V. & Swain, P. S. (2008). Proc Natl Acad Sci U S A 105, 17256–17261.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Genome Res 13, 2498–2504.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., Schluter, D. & Kingsley, D. M. (2004). Nature 428, 717–723.

Shiomi, D. & Niki, H. (2011). Microbiol Immunol 55, 885–888.

Shultzaberger, R. K., Malashock, D. S., Kirsch, J. F. & Eisen, M. B. (2010). PLoS Genet 6, e1001042.

Siddharthan, R., Siggia, E. D. & van Nimwegen, E. (2005). PLoS Comput Biol 1, e67.

Silander, O. K., Nikolic, N., Zaslaver, A., Bren, A., Kikoin, I., Alon, U. & Ackermann, M. (2012). PLoS Genet 8, e1002443.

Singh, G. P. (2013). G3 (Bethesda) 3, 2115–2120.

Sliusarenko, O., Heinritz, J., Emonet, T. & Jacobs-Wagner, C. (2011). Mol Microbiol 80, 612–627.

Stern, D. L. (2000). Evolution 54, 1079–1091.

Stern, D. L. & Orgogozo, V. (2008). Evolution 62, 2155–2177.

Stern, D. L. & Orgogozo, V. (2009). Science 323, 746–751.

Stock, A. M., Robinson, V. L. & Goudreau, P. N. (2000). Annu Rev Biochem 69, 183–215.

Swain, P. S., Elowitz, M. B. & Siggia, E. D. (2002). Proc Natl Acad Sci U S A 99, 12795–12800.

Szoke, P. A., Allen, T. L. & deHaseth, P. L. (1987). Biochemistry 26, 6188–6194.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A. & Xie, X. S. (2010). Science 329, 533–538.

Tautz, D. & Domazet-Losǒ, T. (2011). Nat Rev Genet 12, 692–702.

Team, R. D. C. (2007). .

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D. & Gaut, B. S. (2012). Science 335, 457–461.

Thattai, M. & van Oudenaarden, A. (2001). Proc Natl Acad Sci U S A 98, 8614–8619.

Thattai, M. & van Oudenaarden, A. (2004). Genetics 167, 523–530.

Treangen, T. J. & Rocha, E. P. C. (2011). PLoS Genet 7, e1001284.

Tsai, Z. T.-Y., Tsai, H.-K., Cheng, J.-H., Lin, C.-H., Tsai, Y.-F. & Wang, D. (2012). BMC Genomics 13, 717.

van Nimwegen, E. (2003). Trends Genet 19, 479–484.

Vellanoweth, R. L. & Rabinowitz, J. C. (1992). Mol Microbiol 6, 1105–1114.

Vimberg, V., Tats, A., Remm, M. & Tenson, T. (2007). BMC Mol Biol 8, 100.

Voskuil, M. I. & Chambliss, G. H. (1998). Nucleic Acids Res 26, 3584–3590.

Voskuil, M. I. & Chambliss, G. H. (2002). J Mol Biol 322, 521–532.

Wagner, A. (2005). Mol Biol Evol 22, 1365–1374.

Walkiewicz, K., Cardenas, A. S. B., Sun, C., Bacorn, C., Saxer, G. & Shamoo, Y. (2012). Proc Natl Acad Sci U S A 109, 21408–21413.

Wang, P., Robert, L., Pelletier, J., Dang, W. L., Taddei, F., Wright, A. & Jun, S. (2010). Curr Biol 20, 1099–1103.

Wang, Z. & Zhang, J. (2011). Proc Natl Acad Sci U S A 108, E67–E76.

Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J. & Gustafsson, C. (2009). PLoS One 4, e7002.

Woo, Y. H. & Li, W.-H. (2011). Proc Natl Acad Sci U S A 108, 3306–3311.

Wray, G. A. (2007). Nat Rev Genet 8, 206–216.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. & Romano, L. A. (2003). Mol Biol Evol 20, 1377–1419.

Yip, S. H.-C. & Matsumura, I. (2013). Mol Biol Evol 30, 2001–2012.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. G. & Alon, U. (2006). Nat Methods 3, 623–628.

Zhang, J., ping Zhang, Y. & Rosenberg, H. F. (2002). Nat Genet 30, 411–415.

Zhang, Z., Qian, W. & Zhang, J. (2009). Mol Syst Biol 5, 299.

Zhong, S., Miller, S. P., Dykhuizen, D. E. & Dean, A. M. (2009). Mol Biol Evol 26, 2661–2678.

Acknowledgments

Getting a project started is much easier than finishing, and this work would not had been possible without the support of many friendly and helpful people.

First and foremost, I want to thank my co-supervision team Erik van Nimwegen and Olin Silander for their great encouragement during all stages of the project. Erik's advice, guidance and indestructible patience in teaching me sophisticated quantitative principles as well as his generous support in attending numerous scientific meetings within and outside Switzerland set the scene for personal as well as professional development. Olin's daily mentoring, his openness for extensive discussions and his 'evolutionary spirit' helped to push forward and ensure a successful conclusion of this project.

Many thanks to Dirk Bumann and Andreas Wagner for taking part in my PhD Thesis Advisory Committee, accompanying the work progress over years and providing very useful comments.

From the wetlab technical side I am indebted to people not working in my own group, namely Sören Abel, Imke de Jong, Bea Claudi, Janine Zankl, Maxime Québatte, Phil Demougin, Ina Nissen, Christian Beisel and Manuel Kohler.

Specially warm thanks go to my current and former fellow- both wet and dry - labmates from the van Nimwegen and Zavolan group, who made my time in the group especially enjoyable and provided never-ending assistance.

Curriculum Vitae

Personal Data

Place and Date of Birth: Halle (Saale), Germany — 25 April 1984
Address: Im Ettingerhof 2, Basel, Switzerland
email: wluise@gmail.com

Work Experience

Jan 2014-Current	Postdoctoral Research Fellow in the lab of Prof. Dr. Erik van Nimwegen, University of Basel
Jul-Oct 2009	Research Assistant in the group of Prof. Dr. Stefan Rensing, University of Freiburg
Jan-Mar 2008	Teaching Assistant, University of Freiburg
Jun 2006-Dec 2007	Research Assistant at Fraunhofer Institute for Solar Energy Systems (ISE), Freiburg

Education

Nov 2009-Dec 2013	PhD in Computational and Systems Biology, **University of Basel**, Switzerland co-supervised by Prof. Dr. Erik van Nimwegen and Dr. Olin Silander PhD thesis title: Evolution of transcriptional regulation in *Escherichia coli*
Oct 2004-Mai 2009	Diploma in Biology, **University of Freiburg**, Germany co-supervised by Prof. Dr. Stefan Rensing and Prof. Dr. Roman Ulm Diploma thesis title: Molecular analysis of the UV-B response in *Physcomitrella patens*

Publications

Wolf L, Silander OK, van Nimwegen EJ. 2014. Expression noise facilitates the evolution of gene regulation. *Prepared for submission.*

Hiss M, Laule O, Meskauskiene R, Arif MA, Decker E, Erxleben A, Frank W, Hanke S, Lang D, Martin A, Neu C, Reski R, Richardt S, Schallenberg-Rüdinger M, Szövenyi P, Tiko T, Wiedemann G, **Wolf L**, Zimmermann P, Rensing S. 2014. Large scale gene expression profiling data of the model moss *Physcomitrella patens* aid to understand developmental progression, culture and stress conditions. *The Plant Journal* 79 (3), 530-9.

Blank D*, **Wolf L***, Ackermann M, Silander OK. 2014. The predictability of molecular evolution during functional innovation. *PNAS* 11 (8), 3044-3049. *These authors contributed equally to this work.

Wolf L, Rizzini L, Stracke R, Ulm R, Rensing SA. 2010. The molecular and physiological responses of *Physcomitrella patens* to ultraviolet-B radiation. *Plant Physiology* 153, 1123-1134.