Random Sequences Rapidly Evolve into *de novo* Promoters

Avihu H. Yona[1,2], Eric J. Alm[2,3,4], Jeff Gore[1#]

[1]Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA
[4]Center for Microbiome, Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[#]Corresponding author gore@mit.edu

Abstract

How do new promoters evolve? The current notion is that new promoters emerge from duplication of existing promoters. To test whether promoters can instead evolve *de novo,* we replaced the lac promoter of *Escherichia coli* with various random sequences and evolved the cells in the presence of lactose. We found that a typical random sequence of ~100 bases requires only one mutation in order to mimic the canonical promoter and to enable growth on lactose. We further found that ~10% of random sequences could serve as active promoters even without any period of evolutionary adaptation. Such a short mutational distance from a random sequence to an active promoter may improve evolvability yet may also lead to undesirable accidental expression. Nevertheless, we found that across the *E. coli* genome, accidental expression is largely avoided by disfavoring codon combinations that resemble canonical promoter motifs. Our results suggest that the promoter recognition machinery has been tuned to allow high accessibility to new promoters, and similar findings might also be observed in higher organisms or in other motif recognition machineries, like transcription factor binding sites or protein-protein interactions.

Introduction

Promoters control the transcription of genes and therefore play a major role in evolutionary adaptation[1]. The extensive study of promoters by genomic analysis[2–4], experimental protein-DNA interactions[5–7] and promoter libraries[8–11] has mostly revolved around highly refined promoters i.e. long-standing wild-type promoters and their derivatives. However, the emergence of new promoters, for example when cells need to activate horizontally transferred genes[12,13], is less understood. Recent studies have demonstrated how new promoters can emerge from duplication of existing promoters via genomic rearrangements[14,15], transposable elements[16,17], or by inter-species mobile elements[18]. Yet, little is known about promoters evolving *de novo*. The canonical promoter of *E. coli* is composed of two six-mer motifs - the 'minus 10' TATAAT and the 'minus 35' TTGACA, which are separated by a spacer of 17±2 bases. The sequence space that encompasses these 12 bases (two six-mer motifs) is composed of ~17 million options (4^12). Due to the extreme size of sequence space, it is typically assumed that starting from a random sequence would require multiple mutations to have any significant level of expression. In cases where multiple mutations are necessary for functionality, the evolutionary search is difficult as the first mutation does not have a selective advantage until the other mutations appear, and evolving a promoter would then require the cell to copy a promoter from elsewhere in the genome.

Exploring the fitness landscape of promoters in order to understand how non-functional sequences turn into functional promoters can be done artificially, by using pooled promoter libraries that allows the measurement of a large number of starting sequences. However, pool competition is less applicable for following an evolutionary process that requires mutational steps from inactive sequences as selection in pool is often dominated by a small fraction of the sequences that exhibit high activity. Therefore, in order to explore the fitness landscape of emerging promoters, in a similar way to evolution in natural ecologies, we utilized lab evolution methods. We evolved parallel populations, each starting with a different random sequence, for their ability to evolve new promoters. Following these evolving populations highlighted an unacknowledged way for new promoters to emerge by stepwise mutations from random sequences rather than by copying an existing promoter. Promoter activity was typically achieved by a single mutation and could be further increased in a stepwise manner by additional mutations that increased the similarity of the random sequence to the canonical promoter (TATAAT and TTGACA motifs). We therefore find a surprising flexibility in the evolution of the bacterial transcription network.

Main Text

To create an ecological scenario that can test how bacteria evolve *de novo* promoters, we sought a beneficial gene in the genome but not yet expressed, similarly to what might occur during horizontal gene transfer with a non-functional promoter. To this end, we modified the lac operon of *E. coli*: the lac metabolic genes (*LacZYA*) remain intact (including their 5'UTR), yet we deleted their promoter and replaced it by a variety of non-functional sequences. To broadly represent the non-functional sequence-space, we used random sequences of 103 bases (same length as the deleted WT lac promoter), which were computer-generated with the typical GC content of the *E. coli* genome (~50.8% GC, see Methods). In addition, the lac repressor (*LacI*) was deleted, and the lactose permease (*LacY*) was fluorescently labeled with YFP[19] for future quantification of expression. To avoid possible artifacts associated with plasmids, all modifications were made on the *E. coli* chromosome[20], so the engineered strains had a single copy of the metabolic genes needed for lactose utilization, yet without a functional promoter (Figure 1A). We began building such strains with random sequences as "promoters", and already observed for the first strains obtained that they could not express the lac genes and thus they could not utilize or grow on lactose. This experimental observation was therefore consistent with the expectation that a random sequence is unlikely to be a functional promoter.

To select for *de novo* lactose utilization we started evolution by serial dilution with the obtained strains, each carrying a different random sequence instead of the WT lac-operon promoter. We first focused on three such strains (termed RandSequence1, 2 and 3) and tested their ability to evolve expression of the lac operon, each in four replicates. As controls, we also evolved a strain carrying the WT lac promoter (termed WTpromoter), and another strain in which the entire lac operon was deleted (termed ΔLacOperon). Before the evolution experiment, only the WTpromoter strain could utilize lactose (Supp. Figure 1). Therefore, to facilitate growth to low population sizes the evolution medium contained glycerol (0.05%) that the cells can utilize and lactose (0.2%) that the cells can only exploit if they express the lac operon. To isolate lactose-utilizing mutants, we routinely plated samples from the evolving populations on plates with lactose as the sole carbon source (M9+Lac) (Figure 1B). Remarkably, within 1-2 weeks of evolution (less than 100 generations), all of these populations exhibited lactose-utilizing abilities, except for the ΔLacOperon population (Supplementary Information). These laboratory evolution results therefore argue that the populations carrying random sequences instead of a promoter can rapidly evolve expression. Next, we addressed the question of whether the solutions found during evolution were mutations in the random sequences or simply copying and pasting of existing promoters from elsewhere in the genome.

Figure 1

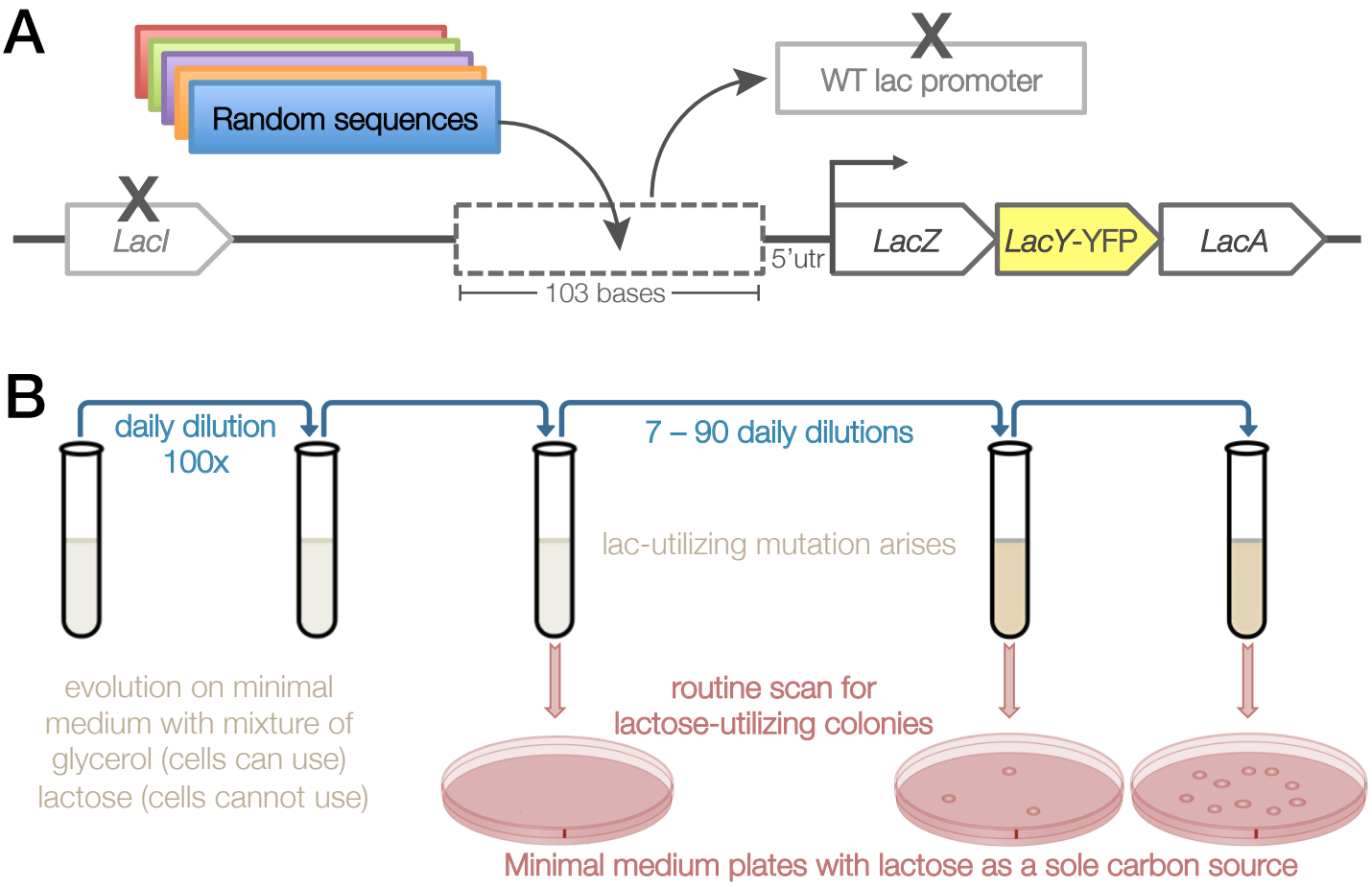
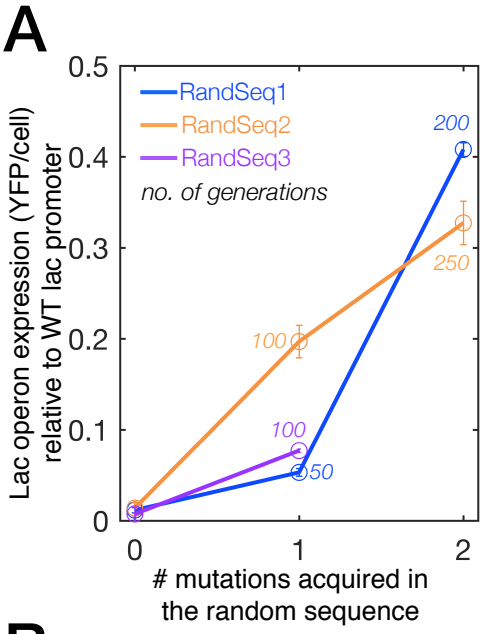


Figure 1: Experimental setup for evolving promoters from random sequences
(A) We modified the chromosomal copy of the lac operon by replacing the WT lac promoter with a random sequence of the same length (103 bases) that abolished the cells' ability to utilize lactose. In addition, the lacY was tagged with YFP and the lac repressor (lacI) was deleted. **(B)** Cells were evolved by serial dilution in minimal medium containing both glycerol (0.05%), which the cells can utilize, and lactose (0.2%), which the cells cannot utilize unless they evolve de novo expression from the lac operon. During evolution, samples were routinely plated on minimal medium plates with lactose as a sole carbon source, for isolation of lactose-utilizing mutants.

To determine the molecular nature of the evolutionary adaptation, we sequenced the region upstream to the lac operon (from the beginning of the lac operon through the random sequence that replaced the WT lac promoter and up to the neighboring gene upstream). Strikingly, within each of the different random sequences a single mutation occurred, and continued evolution yielded additional mutations within the random sequences that further increased expression from the emerging promoters. All replicates showed the same mutations, yet sometimes in different order (Supp. Table 1). In order to confirm that the evolved ability to utilize lactose is because of the observed mutations, each mutation was inserted back into its relevant ancestral strain. Then, we assessed the lac-operon expression by YFP measurements (thanks to the *LacY*-YFP labeling) (Figure 2A). This experimental evolution demonstrates how non-functional sequences can rapidly become active promoters, in a stepwise manner, by acquiring successive mutations that gradually increase expression. Next, we aimed to determine the mechanism by which these mutations induced *de novo* expression from a random sequence.

Looking at the context of the emerging mutations showed that expression was achieved by mimicking the canonical promoter motifs of *E. coli*[21], which is responsible for transcribing the majority of the genes in a growing *E. coli*. (i.e. the 'minus 10' TATAAT and the 'minus 35' TTGACA, separated by a spacer of 17±2 bases). Each of the five mutations found during evolution of the three random sequences contributed for better capturing of the canonical promoter motifs (Figure 2B). The emerging promoters seem to comply with the higher importance of the TATAAT motif to promoter strength. Randseq1 and Randseq2 both captured 5 out of 6 bases, and RandSeq3 captured the full 6 bases, while for the TTGACA motif they all captured 3 out of the 6 motif bases. Interestingly, although before evolution Randseq3 already captured 3/6 bases of the TTGACA motif plus 5/6 of the TATAAT motif, it was not sufficiently strong to induce expression. Randseq3 was not an active promoter before evolution presumably due to a short spacer (14 bases, compared with the ideal 17 bases spacer), which creates significant torsion of the DNA[22] and thus reduced attachment of the transcription machinery. Nevertheless, a single mutation in Randseq3 induced expression as it both allowed perfect capturing of the TATAAT motif (together with a preexisting TGn motif[23]). Therefore, *de novo* promoters are highly accessible because the different features that make a promoter, like sequence motifs and spacer size, can be compromised and still function.

Figure 2



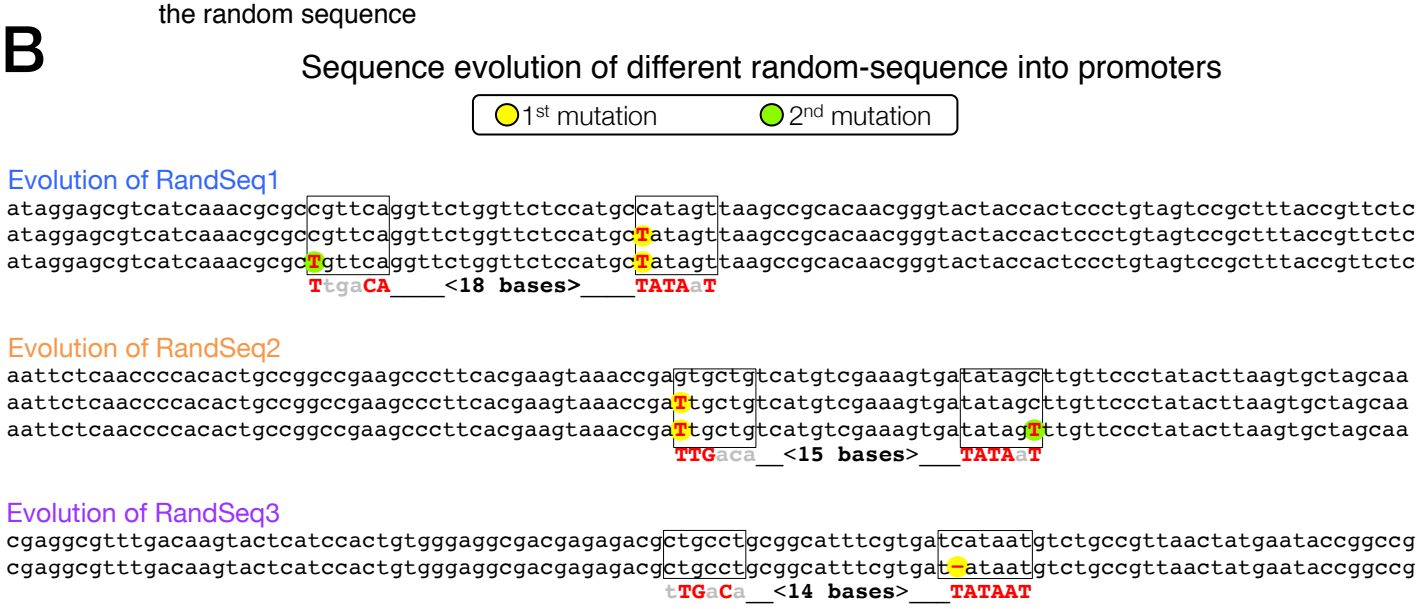


Figure 2: From a random sequence to an active promoter by stepwise mutations that build a canonical promoter (example from three random sequences)

(A) Evolved expression levels of the lac operon are plotted for three strains that carry different random sequences (blue, orange and purple) as a function of the number of acquired mutations. Expression level of 1 is defined as the expression measured from the WT lac promoter, and 0 is defined as the background read of the control strain ΔLacOperon (in which the lac operon is deleted and no YFP gene was integrated). After accumulation of mutations, de novo expression is observed (as well as the ability to utilize lactose). The number of generations is indicated near each mutation. Mutations shown were verified by reinsertion into their non-evolved ancestors. **(B)** Sequences of the evolving promoters. For each strain, the top sequence is the random sequence before evolution, 2nd and 3rd lines are the random sequence with the evolved mutations (1st and 2nd mutations respectively). Increasing similarity to the canonical E. coli promoter motifs can be observed by the different mutations. For each evolving promoter the canonical promoter is shown as the bottom line where capital bases indicate a match.

The most surprising aspect of random sequences evolving into functional promoters was the fact that a single mutation was sufficient for turning on expression. Therefore, we predicted that if indeed a single mutation in a 103-base random sequence is often sufficient to generate an active promoter, there might also be a small portion of random sequences that are already active without the need of any mutation. Indeed, when testing all 40 strains (RandSeq1 to 40) for growth on M9+Lac plates before evolution, we observed that four of the strains (10%) formed colonies without acquiring any mutation in their random sequences. We scanned the random sequences of these already-active strains (RandSeq7, 12, 30, 34) and found regions with high similarity to the canonical σ^{70} promoter, equivalent to the similarities caused by the mutations mentioned earlier (Supp. Figure 2). Given that a single mutation might be sufficient to turn expression on, we proceeded with the strains that did not exhibit lac-operon activity, by putting them under selection for lactose utilization both by the abovementioned daily-dilution routine (in M9+GlyLac) and by directly screening for mutants that can form colonies on M9+Lac plates (Methods).

Overall, selecting for expression of the lac operon was successful for all but 5% of the random-sequence strains (38/40). Analysis of all forty strains and their lac operon activating mutations showed that: 10±4.7% were already active without any mutation (4/40), 57.5±7.8% found mutations within the 103 bases of the random sequence (23/40), 12.5±5.2% found mutations in the intergenic region just upstream to the random sequence (5/40) and 15±5.7% utilized genomic rearrangements that relocated an existing promoter of genes found upstream to the lac operon (6/40)(Figure 3A). YFP measurements indicate that all strains displayed substantial expression of the lac operon after acquiring the activating mutations (Figure 3B). To confirm that transcriptional read-thought from the selection gene upstream did not facilitate the emergence of *de novo* promoters, we made six strains in a marker-free manner (Methods) and showed that their ability to evolve *de novo* promoters is similar to the rest of the strains. A typical random sequence of ~100 bases is therefore not an active promoter but is frequently only one point mutation away from being an active promoter (For details on all mutations, their verifications and different outcomes between replicates see Supp. Table 1).

Figure 3

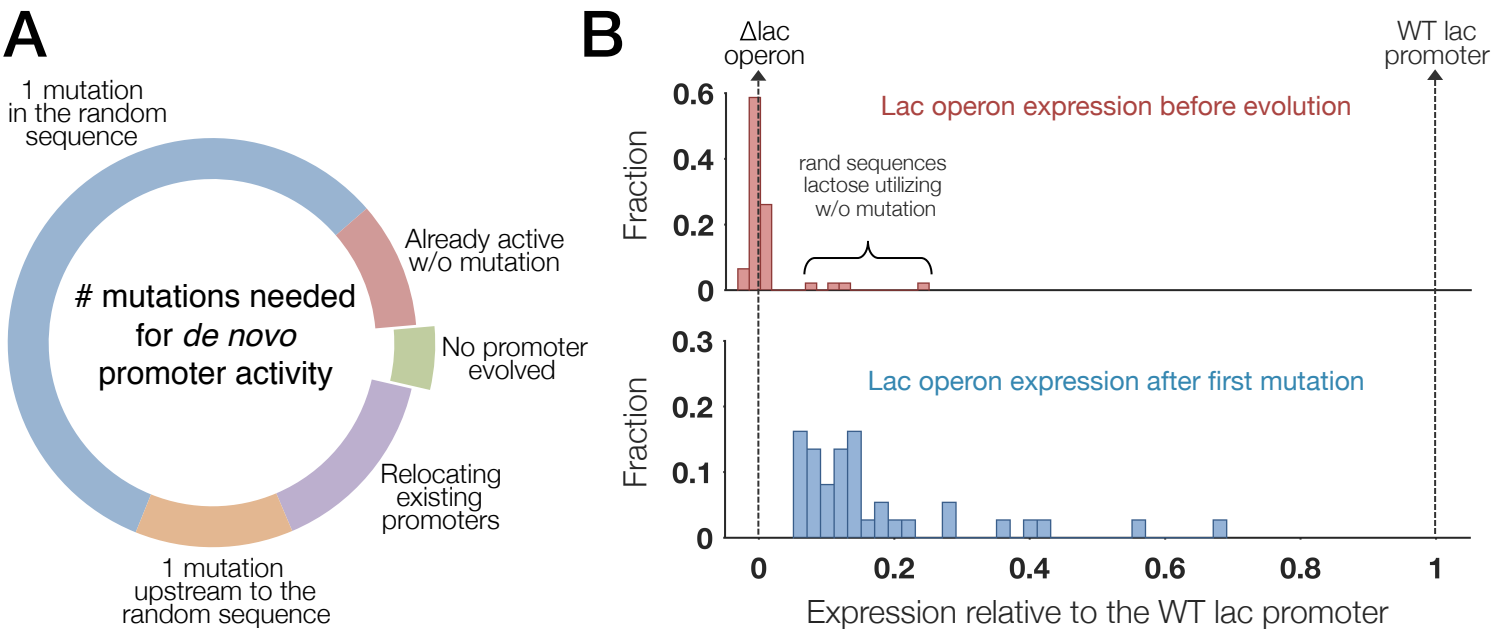


Figure 3: For a typical random sequence of ~100 bases one point mutation is sufficient to function as a promoter

(A) A summary of 40 different random sequences and the different type/number of mutations by which they acquire the ability to express the lac operon and to utilize lactose. ~10% of random sequences require no mutation for such expression of the lac operon that allows growing on lactose as a sole carbon source (red segment). For 57.5% of random sequences a single mutation found within the random sequence enabled expression of the lac operon and growth on lactose (similar to RandSeq 1,2,3 shown earlier)(blue segment). Other strains either relocated an existing promoter from another locus in the genome to be upstream to the lac promoter (15%, purple) or found point mutations in the intergenic region upstream to the random sequence (12.5%, orange). **(B)** Expression of the lac operon before evolution and after the first mutation that was associated with the ability to utilize lactose (upper and lower panel respectively). Measured are YFP reads normalized to OD600 where expression level of 1 is defined as the expression measured from the WT lac promoter (right vertical dashed line), and 0 is defined as the background read of the control strain ΔLacOperon in which the lac operon is deleted and no YFP gene was integrated (left vertical dashed line). The ~10% of random sequences that conferred the ability to utilize lactose even before evolution are found to have significant expression from the lac operon (upper panel).

We performed lab evolution for *de novo* expression by selecting for a functional readout – the ability to grow on lactose. These evolution experiments found that the expression threshold of the lac operon, above which cells can grow on lactose, was often passed by a single mutation. To gain perspective on these surprising findings using a method that is not bound to a specific threshold, we calculated the mutational distance of random sequences from the canonical promoter of *E. coli*. We computationally scanned 10 millions random sequences (of 103 bases) against the canonical promoter motifs and observed that a typical random sequence is likely to match 8 out of the 12 possible matches (of the two six-mers TTGACA and TATAAT, with spacing of 17±2). Interestingly, similar analysis performed on *E. coli*'s constitutive promotes showed that the majority of them have 9 out of 12 matches – only one more than the number of matches observed in random sequences of ~100 bases. Our experimental claim is therefore strengthened, as a random sequence typically requires only one mutation in order to reach the number of matches that characterize naturally occurring constitutive promoters. Furthermore, our computational analysis of random sequences implies that some random sequences may be active already as ~10% of random sequences have 9 or more matches to the canonical promoter sequence (Figure 4).

Figure 4

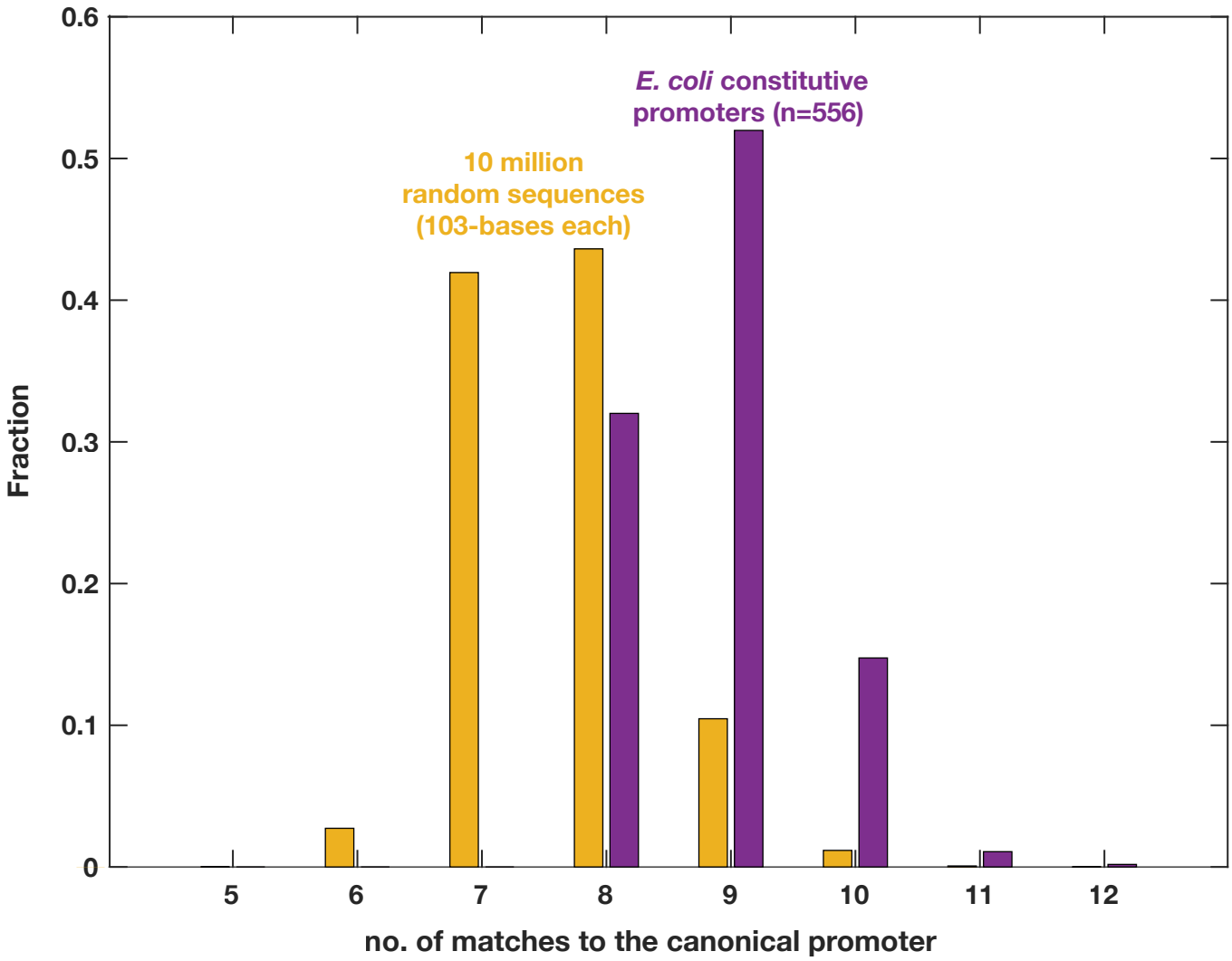


Figure 4:
Mutational distance to the canonical promoter - random sequences are one mutational step behind *E.coli* constitutive promoters
The distribution of the number of matches to the canonical promoter (defined as TTGACA, a spacer of 17±2 bases, and TATAAT) is shown for 10 million random sequences (103 bases each) (orange), alongside the matches found for the 556 E. coli constitutive promoters (purple). The one mutation shift that separates the two distributions suggests that for many of the random sequences a single mutation can increase the number of matches to the number that characterize constitutive promoters in E. coli.

The short mutational distance from random sequences to active promoters may act as a double-edged sword. On the one hand, the ability to rapidly "turn on" expression may provide plasticity and high evolvability to the transcriptional network. On the other hand, this ability may also impose substantial costs, as such a promiscuous transcription machinery is prone to expression of unnecessary gene fragments[24]. Such accidental expression is not only wasteful but can also be harmful as it may interfere with the normal expression of the genes within which it occurs[25,26]. Our data suggest that ~10% of 100-base sequences are an active promoter, meaning that a typical ~1kb gene might naturally contain an accidental promoter inside its coding sequence. Therefore, we looked for strategies that *E. coli* might have taken to minimize accidental expression. Normal promoters typically occur in the intergenic region between genes and not within the coding region. We assessed the occurrence of accidental promoters in the middle of *E. coli* genes (i.e. between the start codon of each gene till its stop codon). This coding region composes 88% of the *E. coli* genome. Since each amino acid can be encoded by multiple synonymous codons, every gene in the genome can be encoded in many alternative ways. We hypothesized that the *E. coli* genome avoids codon combinations that create promoter motifs in the middle of genes. Using promoter prediction software[27,28], we found that the WT *E. coli* genome has much less accidental expression than what would be expected based on a random choice of codons to encode the same amino acids (while preserving the overall codon bias[29], Figure 5A). The *E. coli* genome has therefore likely been under selection to avoid this accidental expression within the coding region of genes.

To assess the optimization level of each gene separately, we compared the accidental expression score of each WT gene to the scores of a thousand alternative recoded versions. Remarkably, we found that ~40% of WT genes had accidental expression as low as the lowest decile of their recoded versions. Our data indicated that some *E. coli* genes minimize accidental expression more than others. Essential genes, for example, exhibit an even stronger signal of optimization compared to the general signal obtained for all genes together (Figure 5B). Essential genes are under stronger selective pressure to mitigate interference[30,31] and therefore they better avoid accidental expression presumably because it leads to collisions with RNA polymerases that transcribe them[32–34]. We observed similar results when we used a recoding method in which we just shuffled the codons of each gene, again indicating that the *E. coli* genome has been under selection to minimize accidental expression (Supp. Figure 3, Methods). To further validate that the WT *E. coli* has depleted promoter motifs within its coding region, we performed a straightforward analysis by unbiased counting of motif occurrences across the genome. The analysis showed that promoter motifs are depleted from the middle of genes, especially the TATAAT motif (Methods, Supp. Table 2). Reassuringly, among this group of depleted motifs we also found the Shine-Dalgarno sequence (ribosome binding site)[35]. Therefore, evolution may have acted to minimize accidental expression by avoiding codon combinations with similarity to promoter motifs, thereby allowing *E. coli* to benefit from flexible transcription machinery while counteracting its detrimental consequences.

Figure 5

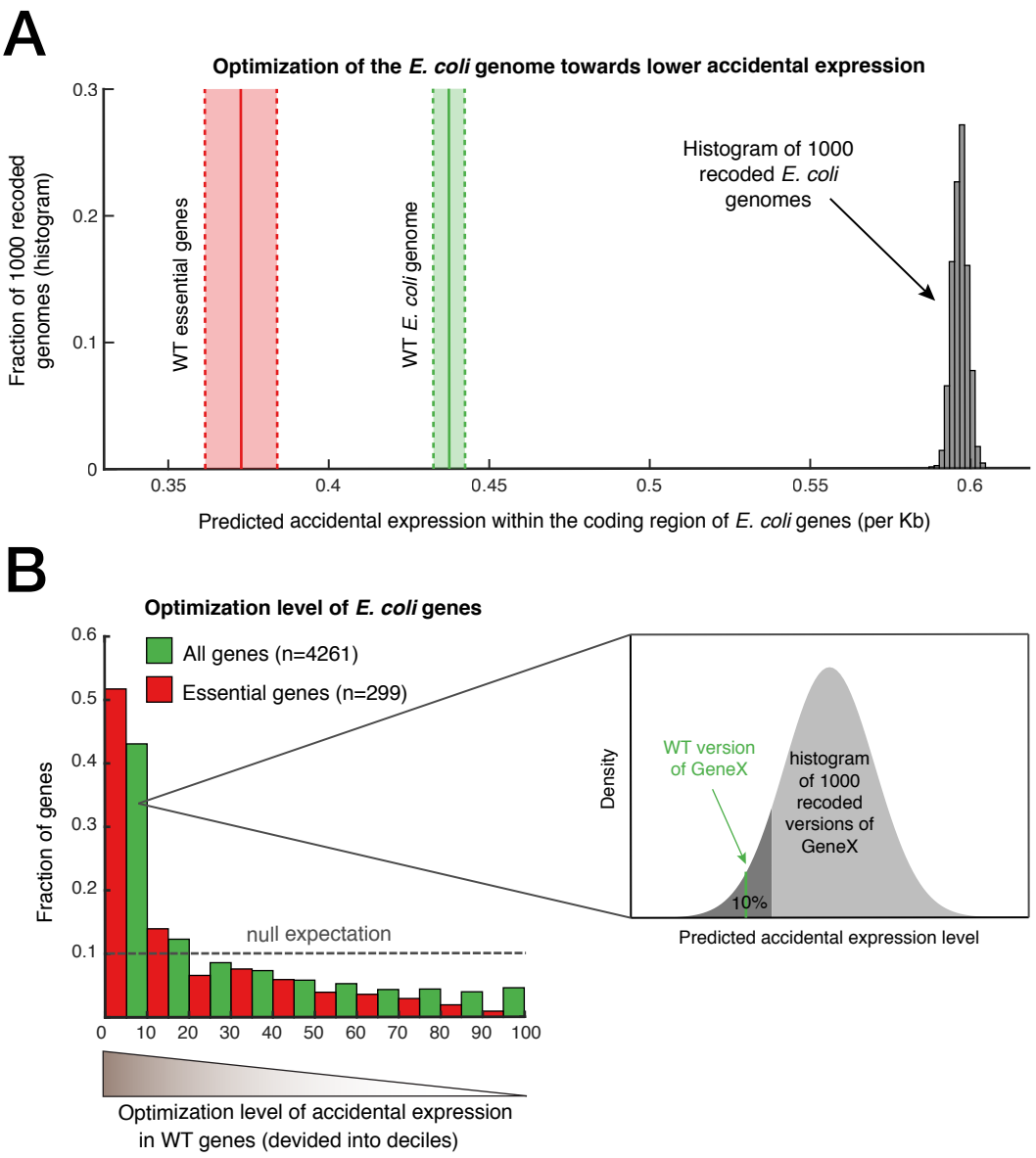


Figure 5: Selection against the occurrence of random promoters in the coding region of genes
We evaluated promoters that accidentally occur across the genome by searching for promoter motifs in the coding region of E. coli. As a reference we did the same evaluation for 1000 alternative versions of the E. coli coding region by recoding each gene with synonymous codons while preserving the amino acid sequence and the codon bias. **(A)** Accidental expression of the thousand recoded versions of E. coli are shown by a histogram (grey), and the accidental expression of the WT E. coli genome is shown by vertical solid lines, for all genes in green, and for the subset of essential genes in red. Shaded areas around the vertical solid lines represent S.E. (delineated by vertical dashed lines). The WT version of the genome is significantly depleted for promoter motifs, indicating genome-wide minimization of accidental expression. **(B)** For each WT gene and its 1000 recoded versions a score for accidental expression was calculated. The WT gene was then ranked in the distribution of its 1000 recoded versions (see inset illustration). Ranking values are divided into deciles, for all WT genes (green), and for the subset of essential genes (red) demonstrating that ~40% of WT genes and more than 50% of essential genes are ranked at the most optimized decile. Dashed line shows expected histogram if WT genes had similar values to their recoded versions.

Discussion

Our study suggests that the sequence recognition of the transcription machinery is rather permissive and not restrictive[36] to the extent that the majority of non-specific sequences are on the verge of operating as active promoters. We found that the typical ~100-base sequence requires only a single mutation to become an active promoter. Consequently, some small portion of non-specific sequences can function as active promoters even without any mutation. This low sequence specificity of the transcription machinery may explain part of the pervasive transcription seen in unexpected locations in bacterial genomes[24] as well as the expression detected in large pools of plasmids that harbor degenerate sequences upstream to a reporter gene[37]. Despite the ability to avoid accidental expression by histone-like proteins[38,39] and by depletion of promoter-like motifs, accidental expression might not always be detrimental and may sometimes be selected for. When we analyzed accidental expression in toxin/antitoxin gene couples[40], we observed higher accidental expression in toxin genes compared with their antitoxin counterparts (Supp. Figure 4, Supplementary Information). Interestingly, when we split the accidental expression score into its 'sense' (same strand as the gene) and 'antisense' (opposite strand) components, we observed that toxins had a much stronger accidental expression in their antisense direction compared to the sense direction. However, in the antitoxins, sense and antisense scores correlated, as largely seen genome-wide (Supp. Figure 5). This leads us to speculate that *E. coli* might have utilized accidental expression as a means to restrain gene expression[41,42] of specific genes, presumably by causing head-to-head collisions of RNA polymerases[32–34].

Our main findings may be relevant to other organisms and to other DNA/RNA binding proteins like transcription factors. The mutational distance between random sequences to any sequence-feature should be considered for possible "accidental recognition" and for the ability of non-functional sequences to mutate into functional ones. We demonstrated that a random sequence is likely to capture 8 out of 12 motif bases of a promoter, while natural constitutive promoters usually capture 9 out of 12. Furthermore, our experiments demonstrated that this "missing" mutation that separates a random sequence from a functional one, is repeatedly found when unutilized lactose is present. Therefore, the implications of this study may also prove useful to synthetic biology designs, as one needs to be aware that spacer sequences might not always be non-functional as assumed. Moreover, spacer sequences can actually be properly designed to have lower probability for accidental functionality, for example a spacer that has particularly low chances of acting as a promoter (or ribosome binding site, or any other sequence feature).

Tuning a recognition system to be in a metastable state so that a minimal step can cause significant changes might serve as a mechanism by which cells increase their adaptability. In our study the minimal evolutionary step (one mutation) was often sufficient to turn the transcription machinery from off to on. If two or more mutations were needed in order to create a promoter from a non-functional sequence, cells would face a much greater fitness-landscape barrier that would drastically reduce the ability to evolve *de novo* promoters.

The rapid rate at which new adaptive traits appear in nature is not always anticipated and the mechanisms underlying this rapid pace are not always clear. As part of the effort to reveal such mechanisms[43] our study suggests that the transcription machinery was tuned to be "probably approximately correct"[44] as means to rapidly evolve *de novo* promoters. Further work will be necessary to determine whether this flexibility in transcription is also present in higher-organisms and in other recognition processes.

Acknowledgments

We thank the Human Frontier Science Program for supporting A.H.Y. Special thanks for Idan Frumkin, Rebecca Herbst and members of the Gorelab and the Almlab for fruitful discussions. We thank the Xie lab for providing strains and Gene-Wei Li, Jean-Benoit Lalanne and Tami Lieberman for their helpful comments on the manuscript.

Methods

Strains – Strains were constructed using the Lambda-Red system[20], including integration of random sequences as promoters by using chloramphenicol resistance selection gene. Yet, for the strains with RandSeq9, 12, 15, 17, 18, 23, integration was done by the Lambda-Red-CRISPR/Cas9 system without introducing a selection marker, in order to exclude transcriptional read-through due to the expression of an upstream selection gene. The ancestral strain for all 40 random sequence strains, as well as for the control strains (WTpromoter and ΔLacOperon) was SX700[19] in which the *lacY* was tagged with YFP. In addition, the *mutS* gene was deleted (by gentamycin resistance gene) to achieve higher yield in chromosomal integration using the lambda-red system[45] and as a potential accelerator of evolution due to increased mutation rate. For Randseq1, 2 and 40 we created additional strains from an ancestor in which the *mutS* was not deleted and after similar evolution the exact same mutations arise. In all strains, *lacI* was deleted (for all but the CRISPR/Cas9 strains, by spectinomycin resistance gene) and replaced by an extra double terminator (BioBricks BBa_B0015) to prevent transcription read through from upstream genes.

Random sequences – random sequences were generated computationally, 103 bases long (same length as the WT lac promoter they replaced). To prevent deviation from the overall GC content of *E. coli* (50.8%) sequences with GC context lower than 45.6% or higher than 56.0% were excluded. In addition, to avoid sequencing issues, sequences with homo-nucleotide stretches longer than five were excluded.

Selection for lactose utilization – Lab evolution was performed on liquid cultures grown on M9+GlyLac by daily dilution of 1:100 into fresh medium. M9 base medium for 1L included 100uL $CaCl_2$ 1M, 2ml of $MgSO_4$ 1M, 10ml NH_4Cl 2M, 200ml of M9 salts solution 5x (Sigma Aldrich). Concentrations of carbon source were 0.05% for glycerol and 0.2% for lactose for M9+GlyLac, 0.2% lactose for M9+Lac and 0.4% glycerol for M9+Gly (all in w/v). Cultures were routinely checked for increased yield at saturation and samples were plated on M9+Lac plates for isolation of colonies that can utilize lactose as a sole carbon source. In parallel to our liquid M9+GlyLac selection for lactose-utilization we also performed agar-plate selection by growing random-sequence strains on non-selective medium (M9+Gly) and then plated them while in late logarithmic phase on M9+Lac plates to select for lactose-utilizing colonies. All populations were evolved in parallel duplicates, but RandSeq1, 2, 3 had four replicates.

Quantifying growth and expression – Growth curves were obtained by 24h measurements of OD_{600} every 10min. Lac operon expression was quantified by YFP florescence measurements. Both measurements performed by a Tecan M200 plate reader.

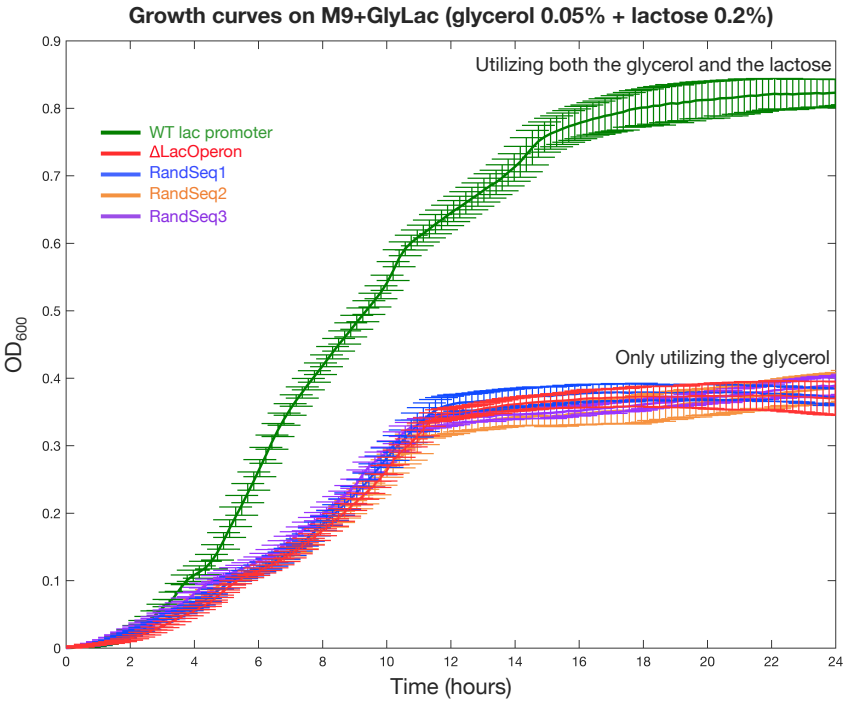
***E. coli* genomic data** - Lists of essential genes and prophage genes were downloaded from EcoGene[46], a list of toxin-antitoxin gene couples was obtained from Ecocyc[40], coding sequences of genes were downloaded from GeneBank (K-12 substr. MG1655, U00096).

Recoding the coding sequence of *E. coli* genes – To create alternative versions of the coding region we recoded all translated genes in *E. coli* (n=4261) 1000 different times while preserving the amino acid sequence and codon bias. As another null model we also shuffled the codons of each gene in 1000 permutations. Although a shuffled version of a gene does not preserve the amino acid sequence, it exactly preserves the GC content of each gene, and thus it controls for another aspect that may result in accidental expression.

Promoter prediction – Using the output from BPROM[27,28] we obtained predicted expression scores by combining the scores of the minus-10 site and the minus-35 site and factoring in the prediction score (LDF) from the output by multiplying. In addition, we scanned sequences for promoters by running a sliding window with the canonical motif and identified regions with maximal agreement.

Six-mer analysis - Looking for depleted and over represented motifs we counted the occurrences of all six-mers within the coding region of *E. coli*. We compiled a list of all 4096 possible six-mers and counted how many times each six-mer occurs in all WT coding region compared to the 1000 recoded versions. Then, we focused on six-mers that are significantly rare/abundant in WT version compared with their counting in the recoded versions.

Supplementary Figure 1



Supplementary Figure 1:
Replacing the WT lac promoter with a random sequence typically abolishes the ability to utilize lactose

Growth curve measurements of WTpromoter (green), ΔLacOperon (red) and RandSequence1, 2, 3 (blue, orange and purple respectively). Shown in values of optical density (OD600) over time during continuous growth on minimal medium (M9+GlyLac, glycerol 0.05% plus lactose 0.2%) at 37°C. The random sequence strains can only utilize the glycerol in the medium and show a growth curve very similar to the Δ LacOperon strain in which the lac genes were deleted. The difference in growth curves between the random sequence strains to WTpromoter reflects the adaptive potential for de novo expression of the lac operon.

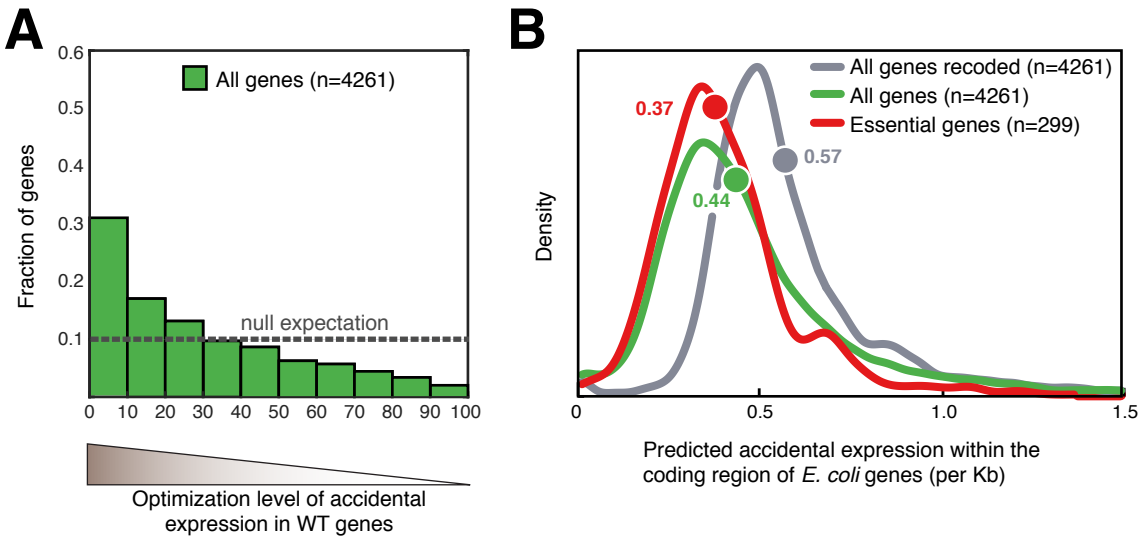
Supplementary Figure 2

Locating the promoter motifs of random sequences that enabled lactose utilization before evolution

Supplementary Figure 2:
Realizing promoter motifs in the random sequences that were already active promoters before evolution

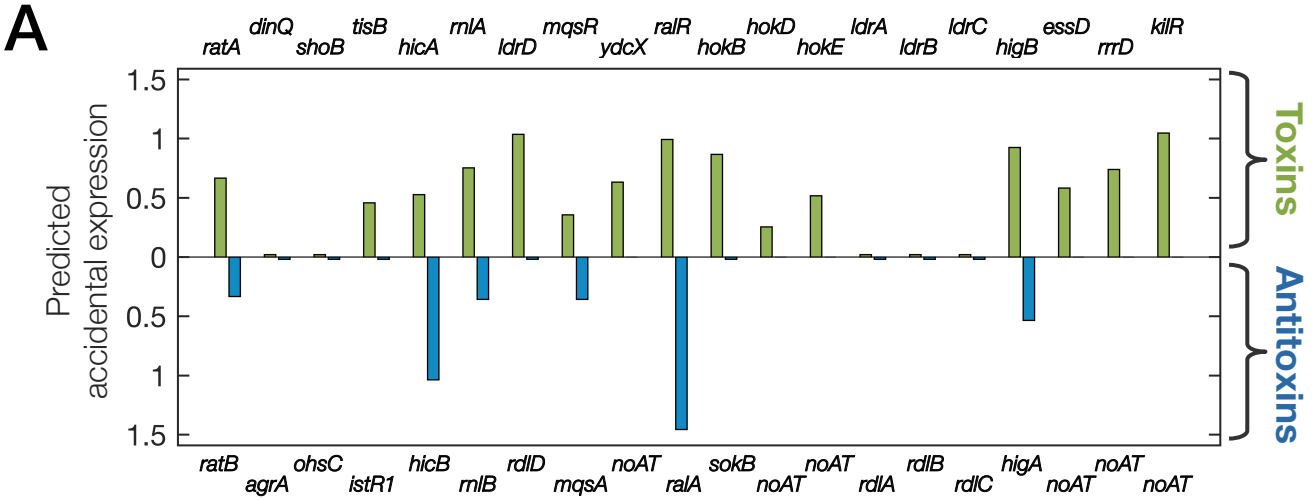
Shown are the sequences of RandSequences7, 12, 30, 34 and the locations of promoter motifs in the random sequences. For these four strains, we observed the ability of cells to grow on lactose-only plates (M9+Lac) without any adaptation. Below each random sequence the canonical promoter is shown where capital bases indicate a match to the canonical motifs TTGACA and TATAAT.

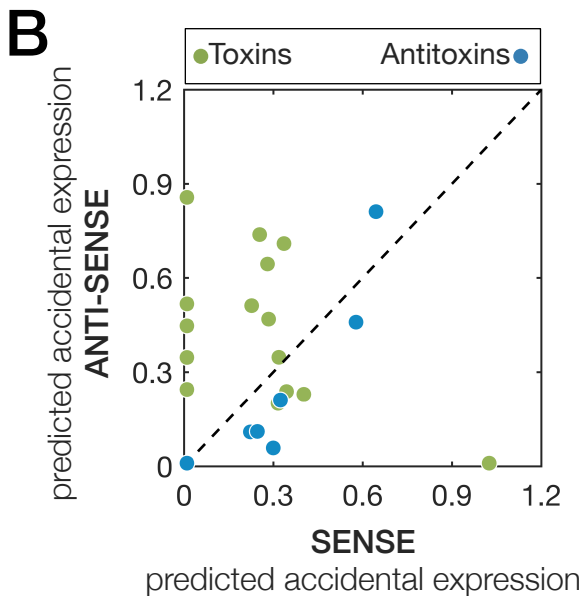
Supplementary Figure 3



Supplementary Figure 3:
Selection against the occurrence of random promoters in the genome – alternative null model
We evaluated promoters that accidentally occur across the genome by searching for promoter motifs in the coding region of E. coli. As a reference we did the same evaluation for 1000 alternative versions of the E. coli coding region by shuffling the codons of each gene, which maintains the GC content and codon bias of each gene. Comparing the WT genes to the 1000 shuffled versions allowed us to look for codon combinations that might have been under negative selection in the WT genome. For example, the shuffled versions can indicate if a combination of two specific codons is avoided in the WT genes because it creates a promoter motif inside a gene. **(A)** A score for accidental expression is calculated for each WT gene and a rank is assigned to each gene by its order in the scores of its 1000 shuffled versions. Shown is the histogram of ranks (divided into deciles) for all WT genes demonstrating that ~30% of WT genes are ranked at the most optimized decile. Dashed line shows expected histogram if WT genes had similar values to their shuffled versions. **(B)** Density plots of accidental expression in the coding sequences of E. coli genes. Distribution of a thousand shuffled versions of E. coli coding region are shown in grey (the value that represent each gene is the median of its 1000 shuffled versions), the accidental expression of the WT E. coli genes is shown in green, and for the subset of essential genes in red. The WT version of the genome is significantly more depleted for promoter motifs, indicating genome-wide minimization of accidental expression. This minimization is further emphasized for the essential genes.

Supplementary Figure 4

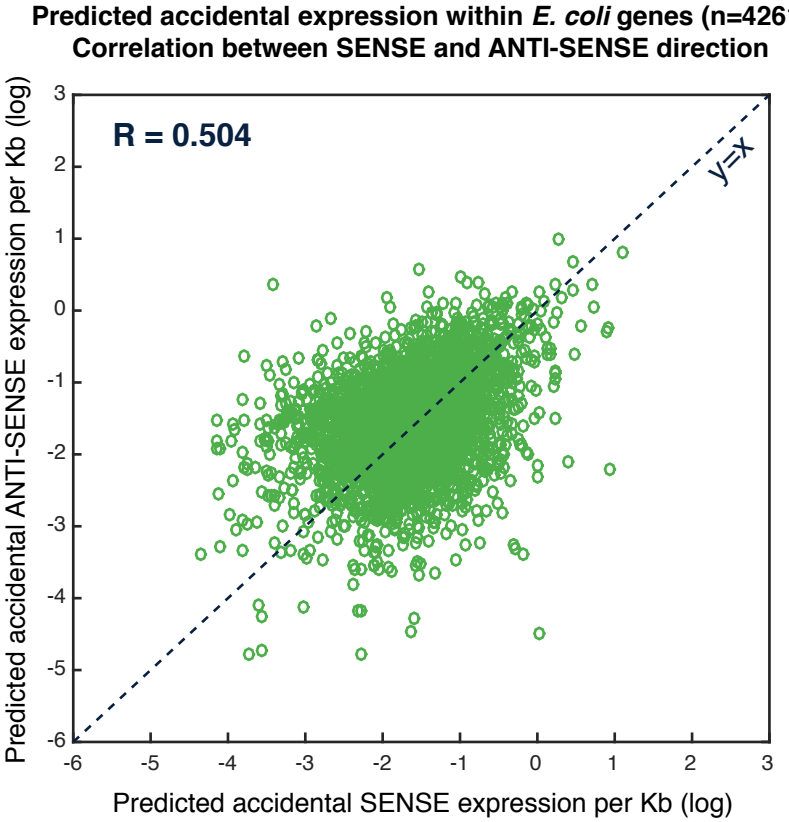




Supplementary Figure 4:
Accidental expression within toxin genes might be selected for as a means to control their expression

For each toxin-antitoxin couple the accidental expression scores examined for differences between toxins genes to their antitoxins and between accidental expressions in 'sense' (with the direction of the gene) compared to the 'antisense' (against the direction of the gene). **(A)** Accidental expression scores are compared between toxins (above the X-axis) and their antitoxin (below the X-axis) showing a tendency of toxins to have higher accidental expression compared with their antitoxin counterparts. **(B)** For both toxin and antitoxin genes the accidental expression was split into 'sense' and 'antisense' direction. While in antitoxin genes the two components tend to correlate (as generally seen in the genome, see Supplementary Figure 5) in the toxins genes the 'antisense' direction is significantly higher, which may imply that E. coli selects for maintaining 'antisense' accidental expression in order to control expression of genes whose higher dosages may harm the cells. Mechanistically, this is presumably due to the fact that antisense transcription collides with the RNA polymerase that expresses the toxin genes.

Supplementary Figure 5



Supplementary Figure 5:
Genome-wide correlation between predicted accidental expression in 'sense' and 'anti-sense' directions
For each WT gene of E. coli we split the score obtained for accidental expression into its two contributing directions (each gene is represented by a green circle). A general correlation (R=0.504) is observed between 'sense' and 'anti-sense' directions.

Supplementary Information

The possibility of evolving lactose utilizing capabilities w/o the lac operon – The fact that the ΔLacOperon strain did not evolve lactose utilizing capabilities indicates that in the random sequence strains lactose utilization arose due to actual activation of the lac operon, by the verified mutations, rather than due to dubious trans–acting mutations. Furthermore, the possibility of activating an *EBG* gene[47] (evolved β-galactosidase) is unlikely as it can only cover for the lack of *lacZ*, but still there is no active permease to replace the function of *lacY*.

Expression activation by capturing an existing promoter or a mutation in the intergenic region upstream to the random sequence –

For the random sequences listed in Extended Data Table.1 as evolved by capturing an existing promoter upstream, we observed various deletions in the intergenic region upstream to the lac operon. All of these deletions placed the lac operon in front of the upstream chloramphenicol selection gene. These deletions also eliminated the termination sequences that separated the lac operon from the genes upstream.

In strains where activating mutations appeared in the intergenic region, just upstream to the random sequence, *de novo* promoters were observed in some cases by mutations, in a similar manner to mutations that created *de novo* promoters in the random sequences (detailed in the mutations table). Yet, there was a group of point mutations, all at the same nucleotide, that occurred within the spacer of a predicted promoter that was experimentally inactive. Nevertheless, one of these mutations was sufficient for expression of the lac operon. The sequence of the predicted, yet inactive, promoter located in the intergenic region was (tcgaaa)gactgggcctttcg(ttttat), where the minus-35 site is TcGAaA and the minus-10 site is TtTtAT. This promoter has a 14-base spacer and the 'g' in the middle of this spacer was mutated multiple times in different strain. In some cases from g to T, in other cases from g to A and once the g was deleted (1 base deletion). It is not clear what was the mechanism by which these mutations activate expression.

We hypothesize that random sequences that evolved expression via mutations in the intergenic region might do so because they could not find an activating mutation in the random sequence. For such sequences, a mutation in the random sequence that can induce expression might not exist. Therefore, we took such a sequence, RandSeq27, and computed mutations that might improve its chances of becoming an active promoter. To this end, we scanned the original RandSeq27 for maximal matches to the canonical promoter. Since there were multiple matches, we chose the maximal match with an optimal spacer of 17 bases. Then, we introduce a point mutation that improved the minus-10 motif. After introducing this mutation into RandSeq27, it did not show promoter activity, yet after applying selection for growth on lactose (like in the library of ransom sequences) the strain found a second mutation that together with the first one we inserted exhibited expression of the lac operon:

RandSeq27 – inactive promoter:

cggtccgtttataacatgcgcagaggaagctgtctgtgcgtcgccagactcagagacccttatactacacccgcctggctgcgaatcatccaccactttaagt

RandSeq27 + 1st mutation (computed) – still inactive promoter:

cggtccgtttataacatgcgcagaggaagctgtctgtgcgtcgccagactcagagacccttatactacacccgcctggctgcg**T**atcatccaccactttaagt

RandSeq27 + 2nd mutation (via selection) – active promoter:

cggtccgtttataacatgcgcagaggaagctgtctgtgcgtcgccagactcagagaccctt-tactacacccgcctggctgcg**T**atcatccaccactttaagt
TTtACt--------17--------TATcAT

This might imply that such sequences are two mutations away from functioning as active promoters.

The different costs of accidental expression and the motivation to focus on toxin-antitoxin gene couples – Accidental expression has a global cost due to waste of resources and occupying cellular machineries. In addition there is also a cost that is due to interference of specific genes. We observed that depletion of accidental expression is more emphasized in essential genes and is less observed in foreign genes like toxin and antitoxin prophage genes. Besides the stronger selective pressure to mitigate interference in essential genes, additional possible reasons for these differences may include: (a) foreign genes have been in the *E. coli* genome for shorter time and thus their expected optimization level is lower, and (b) foreign genes may have lower GC content than *E. coli*, which may affect accidental expression[48] as promoter motifs are AT-rich. To decipher between these potential factors, we therefore focused on toxin/anti-toxin gene couples[40], as for each couple the age in the *E. coli* genome is presumably the same, and they have similar GC content. Nonetheless, the anti-toxin gene is more important to the *E. coli* fitness than its toxin counterpart. Indeed, we observed lower accidental expression in anti-toxin genes compared with toxin genes. This result implies that for each gene the level of avoiding accidental expression is mainly dependent on how important to the fitness it is to have this gene expressed without interference.

Toxin Anti-toxin couples – When analyzing toxin-antitixin gene couples for potential differences in their accidental expression, especially between sense and anti-sense orientations, we excluded gene couples whose orientation in the genome could not lead us to meaningful conclusions. Specifically, we excluded gene couples for the following reasons:

a) Toxin and antitoxin genes were overlapping, hence internal expression affects both (e.g. *ibsA* nad *sibA*).

b) Couples that had this orientation Antitoxin➔ Toxin➔ in which antisense expression from within the toxin gene also influences the adjacent upstream antitoxin (e.g. *yafQ* and *dinJ*).

c) Couples where the annotated promoter of the antitoxin gene is within the toxin gene and thus interference to the toxin is from a canonical functional promoter (e.g. *symE* and *symR*).

References

1. McAdams HH, Srinivasan B, Arkin AP (2004) The evolution of genetic regulatory systems in bacteria. Nat Rev Genet 5: 169–178. doi:10.1038/nrg1292.
2. Manson McGuire A, Church GM (2000) Predicting regulons and their cis-regulatory motifs by comparative genomics. Nucleic Acids Res 28: 4523–4530.
3. Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res 36: D120-4. doi:10.1093/nar/gkm994.
4. Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, et al. (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. Nucleic Acids Res 38: W299-307. doi:10.1093/nar/gkq531.
5. Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, et al. (2009) The transcription unit architecture of the Escherichia coli genome. Nat Biotechnol 27: 1043–1049. doi:10.1038/nbt.1582.
6. Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJW (2005) Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. Proc Natl Acad Sci U S A 102: 17693–17698. doi:10.1073/pnas.0506687102.
7. Wade JT, Castro Roa D, Grainger DC, Hurd D, Busby SJW, et al. (2006) Extensive functional overlap between sigma factors in Escherichia coli. Nat Struct Mol Biol 13: 806–814. doi:10.1038/nsmb1130.
8. Kinney JB, Murugan A, Callan CG, Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A 107: 9158–9163. doi:10.1073/pnas.1004290107.
9. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, et al. (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc Natl Acad Sci U S A 110: 14024–14029. doi:10.1073/pnas.1301301110.
10. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, et al. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol 30: 521–530. doi:10.1038/nbt.2205.
11. Rhodius VA, Segall-Shapiro TH, Sharon BD, Ghodasara A, Orlova E, et al. (2013) Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. Mol Syst Biol 9: 702. doi:10.1038/msb.2013.58.
12. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304. doi:10.1038/35012500.
13. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480: 241–244. doi:10.1038/nature10571.
14. Somvanshi VS, Sloup RE, Crawford JM, Martin AR, Heidt AJ, et al. (2012) A single promoter inversion switches Photorhabdus between pathogenic and mutualistic states. Science 337: 88–93. doi:10.1126/science.1216641.
15. Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature 489: 513–518. doi:10.1038/nature11514.
16. Chu ND, Clarke SA, Timberlake S, Polz MF, Grossman AD, et al. (2017) A Mobile Element in mutS Drives Hypermutation in a Marine Vibrio. MBio 8. doi:10.1128/mBio.02045-16.
17. Matus-Garcia M, Nijveen H, van Passel MWJ (2012) Promoter propagation in prokaryotes. Nucleic Acids Res 40: 10032–10040. doi:10.1093/nar/gks787.
18. Oren Y, Smith MB, Johns NI, Kaplan Zeevi M, Biran D, et al. (2014) Transfer of noncoding DNA drives regulatory rewiring in bacteria. Proc Natl Acad Sci U S A 111: 16112–16117. doi:10.1073/pnas.1413272111.
19. Choi PJ, Cai L, Frieda K, Xie XS (2008) A stochastic single-molecule event triggers phenotype switching of a bacterial cell. Science 322: 442–446. doi:10.1126/science.1161427.
20. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A 97: 6640–6645. doi:10.1073/pnas.120163297.
21. Lisser S, Margalit H (1993) Compilation of E. coli mRNA promoter sequences. Nucleic Acids Res 21: 1507–1516.
22. Gore J, Bryant Z, Nöllmann M, Le MU, Cozzarelli NR, et al. (2006) DNA overwinds when

stretched. Nature 442: 836–839. doi:10.1038/nature04974.
23. Kumar A, Malloch RA, Fujita N, Smillie DA, Ishihama A, et al. (1993) The minus 35-recognition region of Escherichia coli sigma 70 is inessential for initiation of transcription at an "extended minus 10" promoter. J Mol Biol 232: 406–418. doi:10.1006/jmbi.1993.1400.
24. Wade JT, Grainger DC (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol 12: 647–653. doi:10.1038/nrmicro3316.
25. Palmer AC, Ahlgren-Berg A, Egan JB, Dodd IB, Shearwin KE (2009) Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter. Mol Cell 34: 545–555. doi:10.1016/j.molcel.2009.04.018.
26. Shearwin KE, Callen BP, Egan JB (2005) Transcriptional interference--a crash course. Trends Genet 21: 339–345. doi:10.1016/j.tig.2005.04.009.
27. Solovyev V, Salamov A (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In: Robert W. Li, editor. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies. Nova Science Publishers. pp. 61–78.
28. Solovyev V, Salamov A (2015) BPROM - Prediction of bacterial promoters. Available: http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb.
29. Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37: D93-7. doi:10.1093/nar/gkn787.
30. Rocha EPC, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377–378. doi:10.1038/ng1209.
31. Price MN, Alm EJ, Arkin AP (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. Nucleic Acids Res 33: 3224–3234. doi:10.1093/nar/gki638.
32. Hobson DJ, Wei W, Steinmetz LM, Svejstrup JQ (2012) RNA polymerase II collision interrupts convergent transcription. Mol Cell 48: 365–374. doi:10.1016/j.molcel.2012.08.027.
33. Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH (2006) Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. Nucleic Acids Res 34: 5416–5425. doi:10.1093/nar/gkl668.
34. Callen BP, Shearwin KE, Egan JB (2004) Transcriptional interference between convergent promoters caused by elongation over the promoter. Mol Cell 14: 647–656. doi:10.1016/j.molcel.2004.05.010.
35. Li G-W, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541. doi:10.1038/nature10965.
36. Struhl K (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell 98: 1–4. doi:10.1016/S0092-8674(00)80599-1.
37. Wolf L, Silander OK, van Nimwegen E (2015) Expression noise facilitates the evolution of gene regulation. Elife 4. doi:10.7554/eLife.05856.
38. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, et al. (2014) Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev 28: 214–219. doi:10.1101/gad.234336.113.
39. Landick R, Wade JT, Grainger DC (2015) H-NS and RNA polymerase: a love-hate relationship? Curr Opin Microbiol 24: 53–59. doi:10.1016/j.mib.2015.01.009.
40. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, et al. (2013) EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res 41: D605-12. doi:10.1093/nar/gks1027.
41. Dornenburg JE, Devita AM, Palumbo MJ, Wade JT (2010) Widespread antisense transcription in Escherichia coli. MBio 1. doi:10.1128/mBio.00024-10.
42. Brophy JA, Voigt CA (2016) Antisense transcription as a tool to tune gene expression. Mol Syst Biol 12: 854–854. doi:10.15252/msb.20156540.
43. Yona AH, Frumkin I, Pilpel Y (2015) A relay race on the evolutionary adaptation spectrum. Cell 163: 549–559. doi:10.1016/j.cell.2015.10.005.
44. Leslie Valiant (2013) Probably Approximately Correct. Basic Books. 195 p.
45. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. Nature 460: 894–898. doi:10.1038/nature08187.
46. Zhou J, Rudd KE (2013) EcoGene 3.0. Nucleic Acids Res 41: D613-24. doi:10.1093/nar/gks1235.
47. Hall BG (2003) The EBG system of E. coli: origin and evolution of a novel beta-galactosidase for

the metabolism of lactose. Genetica 118: 143–156.

48. Lamberte LE, Baniulyte G, Singh SS, Stringer AM, Bonocora RP, et al. (2017) Horizontally acquired AT-rich genes in Escherichia coli cause toxicity by sequestering RNA polymerase. Nat Microbiol 2: 16249. doi:10.1038/nmicrobiol.2016.249.