A Comparison of the Costs and Benefits of Bacterial Gene Expression

by Morgan N. Price, Kelly M. Wetmore, Adam M. Deutschbauer and Adam P. Arkin;
Environmental Genomics and Systems Biology, Lawrence Berkeley National Lab.
February 4, 2016

Abstract

We used genome-wide data on protein expression and mutant phenotypes to compare the cost and the benefit of each protein in *Escherichia coli* K-12 during growth in minimal glucose medium. Proteins that are important for fitness are usually highly expressed, and 95% of these proteins are expressed at above 13 parts per million (ppm). Conversely, proteins that do not measurably benefit the host tend to be weakly expressed, with a median expression of 13 ppm. In aggregate, genes with no detectable benefit account for 31% of protein production, or about 21% if we correct for genetic redundancy. Although some of the apparently unnecessary expression could have subtle benefits in minimal glucose medium, the majority of the burden is due to genes that are important in other conditions. We propose that over 10% of the cell's protein is allocated to preparation for less favorable conditions.

Introduction

The typical bacterial genome encodes thousands of proteins, and many of these proteins are not beneficial for growth at any given time. For example, the model bacterium *Escherichia coli* K-12 preferentially utilizes glucose. Its genome encodes hundreds of genes that enable it to utilize other carbon sources, but these genes will not be beneficial if glucose is available. Furthermore, the activity of many proteins can be detrimental, as the loss of many genes confers a measurable growth advantage in some conditions (1; 2; 3; 4; 5).

Expressing an unnecessary protein should reduce the growth rate even if the protein's activity is harmless. In theoretical models of microbial growth, useless protein causes a reduction in fitness (or the relative growth rate) equal to the fraction of all protein that is useless ((6); M.N.P. et al., in preparation) or a small multiple of this (7). In laboratory environments, the measured fitness cost of a useless and harmless protein is about 1-2 times the fraction of protein (8; 7; 9). This cost could be higher in low-nutrient environments, which are probably common in the wild. For example, in both *E. coli* and *Bacillus subtilis*, highly expressed proteins tend to contain amino acids that cost less energy to synthesize (10). Similarly, sulfur limitation selects against the usage of methionine or cysteine in many proteins (11).

Bacterial proteins are typically expressed at 3-21 parts per million of the protein mass of a cell (data of (12), 25th-75th percentile). Although a cost of 3 ppm may seem small, it should be significant for evolution. The effective population sizes for bacteria are estimated at around 10^6 or 10^7 (13), which implies that alleles that increase fitness by just 1 ppm should predominate over evolutionary time.

Given the high cost of unnecessary expression, bacteria should evolve to allocate their expression of protein to genes that are important for growth or survival. Several recent studies examined the concentrations of proteins in bacteria as a resource allocation problem. In *E. coli*, proteins that are regulated by the growth rate account for about half of the protein mass (14), and the total expression of many functional categories of proteins is correlated with the growth rate (15). However, the importance of these proteins for growth or fitness was not examined. In *B. subtilis*, the concentrations of most enzymes can be explained by the flux that was predicted by a theoretical model (16). This model included metabolic enzymes, ribosomes, and chaperones, but did not include most of the genes in the genome. Also, all of these studies relied on peptide mass-spectrometry ("proteomics"), and so they focused on relatively highly-expressed proteins. These studies reported abundances for just 18-55% of the proteins in the genome.

Here, we compare the costs and benefits of almost all (3,574/4,146, or 86%) of the proteins in *E. coli* K-12 during growth in a minimal glucose medium. To measure protein production or cost, we used ribosomal profiling data (12), which allows us to study weakly-expressed proteins. To measure the benefit of each protein, we used a barcoded library of about 150,000 transposon mutants (17) as well as information from individual knockout strains (18; 19). We found that 97% of protein-coding genes that had mutant phenotypes were expressed at above 5 ppm of protein mass or 30 monomers per cell. Conversely, genes that did not have a measurable impact on fitness had median expression of 13 ppm of protein (by mass) or around 50 monomers per cell. Overall, genes that were not important for fitness accounted for 31% of protein production by mass, but some of these proteins are isozymes or are otherwise expected to be redundant. Once we correct for genetic redundancy, then we estimate that in this condition, 21% of protein production is unnecessary. We also show that the majority of this unnecessary expression is due to proteins that are important for growth in other conditions. So, we propose that most of the 21% burden represents preparation for other conditions, rather than the expression of genes whose benefits are too subtle for us to detect.

Results

Comparison of ribosomal profiling to mutant phenotypes

To compare the costs and benefits of gene expression, we studied *E. coli* K-12 growing at 37°C in MOPS minimal glucose media. We obtained ribosomal profiling data from Li and colleagues (12) and we use the fraction of protein expression (weighted by the length of the protein) to estimate the cost of expression. The ribosomal profiling data should be accurate to within 2-fold for most genes, as the data from two halves of a gene, or for two proteins in an equimolar complex, tend to be consistent within this range (12). Also, Li and colleagues report that their quantitation is accurate for genes with over 128 reads, which corresponds to roughly 1 ppm of expression. For a protein of average size, 1 ppm corresponds to about 6 monomers being produced per cell cycle, as in these conditions there are 5.6 million protein monomers per cell (12).

To measure the importance of each protein for growth, we used a pooled library of about 150,000 transposon mutants with DNA barcodes (17). We grew the pooled mutants in MOPS minimal glucose media for 12 generations and we assayed the abundance of each mutant before and after growth by amplifying the DNA barcodes, sequencing them, and counting them (17). Because mutants that have a strong growth defect in rich media will be missing from our library, we also used a list of 286 essential (or nearly-essential) proteins from PEC, and we classified these genes as essential (19). In our data, mutants in 294 non-essential genes had a statistically significant reduction in abundance after 12 generations, and we classified these as important for fitness. We identified 25 genes whose mutants had a significant increase in abundance: we classified these as detrimental to fitness. Based on a negative control in which we examined the differences between two independently-grown samples of the pool of mutants, we expect around 1 false positive among the genes with phenotypes (see Methods). To our list of important genes, we added 25 non-essential genes that lack representation in our pool of mutants but whose mutants in the Keio collection have strong growth defects in both LB (Luria-Bertani) and minimal MOPS media (18). Most of the remaining genes have little phenotype in MOPS minimal glucose media, but 48 genes had phenotypes of 3% per generation or more but were not statistically significant, and another 449 genes had insufficient coverage; these genes were considered ambiguous. The remaining 2,944 genes have no apparent phenotype. These genes probably affect growth by less than 5% per generation, because of the genes with estimated effects of 4%-5% per generation, 28 of 39 (72%) were statistically significant. The classification of the genes is included in Supplementary Table 1.

As shown in Figure 1A, genes that have a phenotype are much more likely to

be highly expressed, but even genes with no phenotype are usually expressed at significant levels, with a median expression of 13 ppm. Also note that 13 ppm is far above the level at which the ribosomal profiling data becomes noisy, which is about 1 ppm. 81% of proteins with no phenotype are expressed at above 1 ppm.

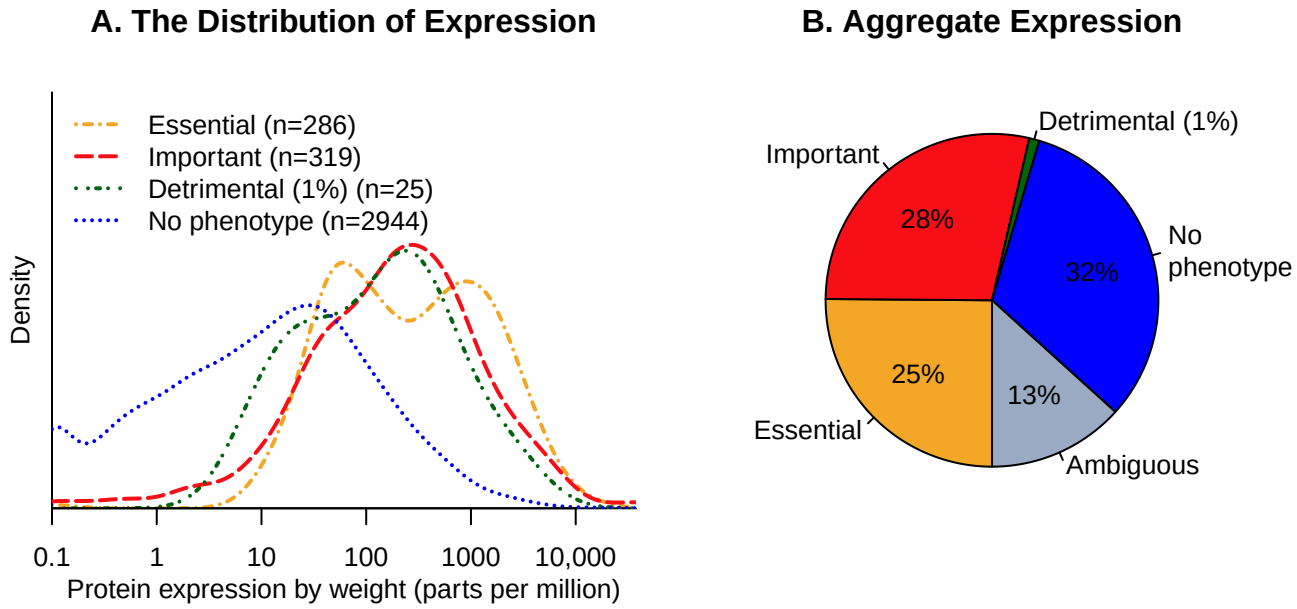

Figure 1: The production of proteins versus their importance for growth. (A) For each class of gene, we show the distribution of protein expression, in parts per million of amino acids (x axis, log scale). Proteins with little or no expression are shown at 0.1 ppm. (B) The aggregate expression of each class of gene.

Most genes that affect fitness are well expressed

Proteins with detectable benefits in minimal glucose medium tended to be well expressed in this condition, with 96% of these genes being expressed at 10 ppm or more. Similarly, 23 of 25 of proteins that were detrimental to fitness (92%) were expressed at 10 ppm or more, which makes sense because they should be expressed at a significant level in order to have a measurable negative impact on the cell. For a protein of average size, 10 ppm corresponds to 50-60 monomers per cell.

All of the essential proteins (19) were expressed at above 5 ppm, except for the putative protein YceQ, which had no ribosomal profiling reads at all. YceQ lacks

homology to any characterized protein or to any protein outside of *Escherichia* or *Salmonella*, and the open reading frame is disrupted in some strains. We infer that *yceQ* does not encode a protein. The other weakly-expressed essential proteins (below 10 ppm) were PrmC and MreD, at around 30 and 60 monomers per cell, respectively. (PrmC, formerly known as HemK, was listed as essential by (19); it is not entirely essential but a mutant has severely reduced growth (20).)

Some very weakly-expressed non-essential genes were identified as being important for fitness, including 6 proteins with expression of under 1 in 10^6 monomers or roughly 6 copies per cell. These proteins were ArpA, WcaE, YahL, YbfK, YdbD, and YnbB. WcaE is believed to be a glycosyltransferase that is involved in the biosynthesis of colanic acid, an exopolysaccharide; little is known about the function of the other proteins (21). We are not sure how these proteins could have a measurable effect at 5 copies per cell unless they are regulatory proteins. Differences between the genetic backgrounds of the two data sets might lead to discrepancies (see Methods). We also wondered if these proteins might be important for the transition to growth in this condition, rather than during exponential growth (which is when protein production was measured), but this does not seem to be the case (see Methods). Three of the proteins (WcaE, YahL, and ArpA) have correlated phenotypes in a compendium of fitness assays in *E. coli* (M. N. Price *et al.*, in preparation; all $r > 0.7$), so we do not think that these discrepancies are just noise in the fitness data.

Another 7 genes were important for fitness despite weak expression of 1-2 per million monomers (6-12 copies per cell). These include three proteins that are involved in the uptake of iron via the siderophore enterobactin (FepD, FepG, and Fes). Again, these could reflect subtle differences in growth conditions.

Overall, we propose that almost all proteins are expressed at above 10 per million monomers (or about 50 monomers per cell) when they have a significant effect on growth. This threshold accounts for 95% of the genes with phenotypes (599 of 630).

Most genes are well expressed in other conditions

During growth in minimal MOPS glucose medium, 34% of the genes in the genome are weakly expressed, with less than 5 monomers per million. We expected that most of these genes would be more highly expressed in other conditions. To test this, we examined a compendium of mRNA expression data from 466 conditions (22). This data is normalized to an arbitrary $\log_2$ scale and includes conditions that are similar to those used for ribosomal profiling (growth in MOPS minimal glucose media and in LB). In those conditions, protein expression of above 5 monomers per million corresponds roughly to mRNA expression of above 8: for proteins with expression of

4-6 monomers per million, the median mRNA expression was 8.0 in minimal MOPS glucose and 8.2 in LB. We found that 96% of *E. coli* mRNAs are expressed above 8.0 in at least one condition. As this proportion might be inflated by experimental noise, we also examined the 99th percentile of expression for each gene, and found that 92% of *E. coli* mRNAs are expressed above 8.0 in some conditions.

We then considered the proteins that have weak expression in both the mRNA compendium (99% of values are under 8.0) and the ribosomal profiling data (under 5 monomers per million in both rich and minimal media). Little is known about most of these 281 weakly-expressed proteins, but 55 have known regulation (23) and another 47 are characterized (their gene names do not start with "y"). We examined the curated descriptions in EcoCyc (21) for a random subset of 10 genes from each group. Of these twenty genes, five are induced by specific conditions that are not included in the mRNA expression compendium (*allB, kilR, mngB, rhsD,* and *xapA*; expression of *kilR* is induced by nalidixic acid stress (24)). Two more genes (*dicB* and *dicC*) are repressed by DicA (25) and their expression increases dramatically in a strain that has a regulatory mutation (26), but the natural signal that controls their expression is not known. Five other genes are functional when overexpressed and conditions that induce their expression are not known (*acrE, htrE, ompN, rihB,* and *smfD*). *GfcC* is a putative recent pseudogene in K-12 strains, with loss of expression due to an IS element inserting into the promoter. Just one gene (*quuD*) has a reported mutant phenotype and its low expression was not explained. We did not find relevant information about the remaining six genes. Overall, 13 of the 14 weakly-expressed proteins are probably expressed more highly when they are functional. These results are consistent with our proposal that functional genes are normally expressed at above 5 monomers per million or above 30 monomers per cell.

High expression of many genes with no measurable benefit

31% of total expression (by mass) was due to the 2,944 proteins with no measurable impact on fitness (Figure 1B). As shown in Figure 1A, the distribution of expression is quite skewed, so most of this 31% is due to a few hundred well-expressed genes. For example, proteins with expression of above 200 ppm comprised 320 genes with phenotypes, 287 genes without a measurable impact on growth, and 79 genes with ambiguous or missing fitness data. The 287 highly-expressed genes with no phenotype account for 24% of total protein production.

We show the 20 most highly-expressed proteins with no mutant phenotype in Table 1. Six of these genes are involved in key processes that are important for growth but the knockout lacks a phenotype because of genetic redundancy. For

Gene	Fraction	Group	Description
ompF	2.0%	Key/redundant	outer membrane porin 1a
ompC	1.5%	Key/redundant	outer membrane porin protein C
livJ	0.6%	Nutrient	leucine/isoleucine/valine transporter subunit
ompT	0.6%	Stress	outer membrane protease VII
ahpC	0.5%	Key/redundant	alkyl hydroperoxide reductase, C22 subunit
aceA	0.5%	Central	isocitrate lyase (glyoxylate shunt)
pflB	0.4%	Anaerobic	pyruvate formate lyase I
aceB	0.3%	Central	malate synthase (glyoxylate shunt)
zinT	0.3%	Stress	cadmium-induced cadmium binding protein
sodA	0.3%	Key/redundant	superoxide dismutase, manganese
adhE	0.3%	Anaerobic	aldehyde-alcohol dehydrogenase
pyrI	0.3%	–	aspartate carbamoyltransferase regulatory subunit
ompX	0.3%	–	outer membrane protein X
oppA	0.3%	Nutrient	oligopeptide transporter subunit
pntA	0.2%	Central	NAD(P) transhydrogenase subunit alpha
sodB	0.2%	Key/redundant	superoxide dismutase, Fe
pykF	0.2%	Key/redundant	pyruvate kinase
metQ	0.2%	Nutrient	DL-methionine transporter subunit
dppA	0.2%	Nutrient	dipeptide transporter
tpx	0.2%	Stress	thiol peroxidase

Table 1: The 20 most highly-expressed genes, by fraction of amino acids, that have no measurable impact on growth.

example, the top two proteins, OmpF and OmpC, are the two major outer membrane porins. Presumably, some expression of porins is necessary for the movement of nutrients through the outer membrane, but deleting either of these individually has little effect. Similarly, AhpC reduces hydrogen peroxide, which is a toxic byproduct of the aerobic electron transport chain, but so do KatE and KatG. If all three genes are disabled, then under aerobic conditions, toxic hydrogen peroxide will accumulate and growth will be inhibited (27). SodA and SodB are redundant isozymes of superoxide dismutase and eliminate another toxic byproduct of oxygen utilization. A strain lacking both SodA and SodB cannot grow aerobically in a minimal glucose medium (28). Finally, PykF is one of two isozymes of pyruvate kinase, which (in reverse) is an ATP-forming step in glycolysis. Again, these isozymes are likely redundant.

Of the remaining 14 highly-expressed proteins with no mutant phenotype, 12 are only expected to be important for growth in other conditions. High expression of these proteins might nevertheless be selected for, just in case growth conditions change (4). For example, AceA, AceB, and PntA are involved in central metabolism

but do not carry flux during growth on glucose (29). Similarly, in *B. subtilis*, some enzymes in central metabolism are highly expressed even when they carry no flux (16). LivJ, OppA, MetQ, and DppA are involved in the uptake of amino acids or short peptides, which are not present in our media. PflB and AdhE are probably important for growth in anaerobic conditions. OmpT, ZinT, and Tpx are involved in responses to stresses that were not present in our experiment. The two remaining proteins are a regulatory subunit of an important enzyme (PyrI) and and an outer membrane protein whose function is not well understood (OmpX).

We then examined the EcoCyc entries for all 287 highly-expressed genes that lack phenotypes (Supplementary Table 2). We identified 70 genes that are involved in key processes that are important for growth, 68 genes involved in utilizing alternate nutrients, 24 genes involved in stress resistance, 9 genes from central metabolism (but not needed for aerobic growth on glucose), and 5 genes involved in anaerobic growth. Regulatory proteins and proteins with poorly-understood functions were not included in any of these categories, and just 8 of the 287 genes are characterized transcriptional regulators (according to RegulonDB 9.0, (30)). The majority of the remaining genes are poorly characterized: 62/103 (60%) have names that begin with *y*.

Among the 70 genes involved in key processes, 60 (86%) are likely to be redundant. Sometimes the redundancy is indirect. For example, SpeD is highly expressed and is required for the synthesis of spermidine, which is one of the major polyamines in *E. coli*. Previous studies found that a strain of *E. coli* that lacks spermidine has a subtle growth defect (around 15%) while a strain that lacks both spermidine and another polyamine, putrescine, has a severe growth defect (around 70%) (31; 32). This indicates that putrescine and spermidine synthesis are partly redundant, and under our conditions, spermidine synthesis may be fully redundant. Alternatively, the effect of the loss of spermidine might be small: SpeE is also required for spermidine synthesis, and we found that knockouts of *speE* had a subtle growth defect (3% per generation) that is near the limit of sensitivity of our fitness assay.

Why doesn't mutating any of the other 10 highly-expressed proteins that are involved in key processes lead to reduced growth? BamB, BamC, SecB, and YajC are non-essential components of essential protein complexes. LepA, MiaB, and RimO are accessory proteins for translation or for the modification of tRNA or rRNA, and might have more subtle advantages. ZapB is required for Z ring placement and mutants have altered size but still grow at about the same rate as wild-type cells (33). TatA is the twin arginine translocase and apparently none of the proteins that it exports are important for growth in our conditions. And MlaC is involved in ensuring the asymmetry of lipids in the outer membrane; this is important for

stabilizing the outer membrane but is not important for fitness in standard growth conditions (34).

Overall, we found that many of the highly-expressed proteins that lack a phenotype are expected to be important for growth only under other conditions. These proteins account for 9.8% of total protein production. Another 8.3% of protein production is accounted for by proteins that lack a phenotype because of genetic redundancy. The roles of most of the remaining highly-expressed proteins with no phenotype (5.9% of protein production) are unclear. If we assume that the proportion of "unnecessary" expression that is due to genetic redundancy is similar for moderately-expressed proteins as it is for highly-expressed proteins ($8.3/24.1 = 34\%$), then we estimate that $31\% \cdot (1 - 0.34) = 21\%$ of protein production is unnecessary in this growth condition.

Highly-expressed genes that are not important for fitness often have phenotypes in other conditions

If much of unnecessary protein production is due to preparation for other conditions, then the highly-expressed yet unnecessary proteins should be important for fitness in other conditions. So, we asked if these genes have phenotypes in a compendium of 162 fitness experiments for *E. coli* ((17); M. N. Price *et al.*, in preparation; http://fit.genomics.lbl.gov/). This compendium includes growth in 29 different carbon sources, growth in 16 different nitrogen sources, growth in the presence of 35 different antibiotics or biocides, and motility on an agar plate. The compendium covers 2,944 proteins that do not have a phenotype in minimal glucose media, and 722 of these are important for growth in other conditions or for motility. As shown in Figure 2, proteins that are not important in minimal glucose media are much more likely to be highly expressed if they have phenotypes in other conditions, with a median expression of 45 ppm instead of 7 ppm ($P < 10^{-15}$, Wilcoxon rank sum test). In total, the 722 proteins that are not important for fitness but have phenotypes in other conditions account for more of protein production in minimal glucose media (18%) than the 2,222 proteins that do not have any phenotypes at all (14%).

A caveat is that some of the genes with measurable phenotypes in artifical conditions might be more subtly useful in nature. For example, some form of cellular damage might occur at low rates under natural conditions, so that the repair genes have subtle benefits. Yet during growth in the presence of an inhibitor that creates this type of damage, these fine-tuning genes could have large benefits. To address this, we looked at each fitness experiment separately. We considered only the experiments with 5 or more proteins that were important for fitness in this experiment

but did not have a measurable phenotype in minimal glucose media. In 134 of 153 experiments, the median important protein was expressed at 3-fold higher than the median protein. Since we found phenotypes for the "unnecessarily" highly-expressed genes in a broad range of conditions, we expect that this holds under natural conditions as well. We also thought that this caveat would be less likely to be relevant for carbon sources or nitrogen sources, as *E. coli* would probably not be able to consume these nutrients unless they were sometimes important in nature. If we considered only the carbon source and nitrogen source experiments, then in 82 of 96 experiments (85%), the median protein that was important for fitness was expressed 3-fold higher in glucose minimal media than the median protein that was not important. Thus, we propose that much of the "unnecessary" expression represents preparation for changing conditions.

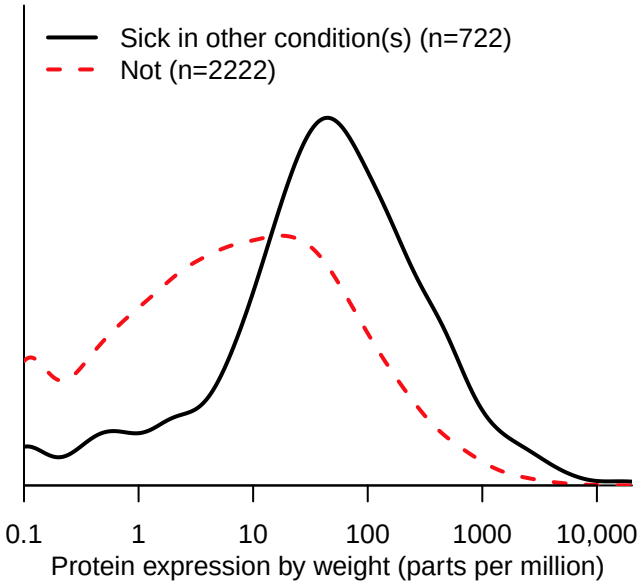

Figure 2: The production of unnecessary proteins versus their importance in other conditions. Only genes that were not important for fitness in minimal glucose media are included. The x axis is as in Figure 1A.

The cost of gene regulation

We have focused on the cost and benefit in a single condition, but regulatory genes need to be expressed across conditions to benefit the cell. Thus, the lack of a phenotype for a regulatory gene is arguably misleading. So, we examined the 186 characterized transcription factors in *E. coli* that are homomeric (30). We identified a phenotype in minimal glucose medium for 15% of regulators, which is about the same rate as for other genes (16%).

The typical expression of these transcription factors was 10 ppm to 60 ppm (25th to 75th percentile), and their total expression is 1.8%. Thus, the total cost of gene regulation does not seem to be that high. However, the cost of an individual regulator would be significant during evolution, which might prevent the maintenance of rarely needed capabilities in most members of a population (35).

Discussion

The cost of being a generalist

We estimate that 21% of *E. coli*'s protein production in minimal glucose medium is due to genes that are not important for fitness in this condition (after correcting for genetic redundancy). Furthermore, the majority of this burden seems to reflect preparation for changing conditions, rather than proteins with benefits that were too subtle for us to detect. This implies that *E. coli* pays a high price, over 10% of protein, for the ability to grow in a wide range of conditions. Because the fitness cost of useless protein is at least as large as the fraction of protein (8; 7; 9; 6), this burden reduces *E. coli*'s growth rate by at least 10%.

To test if this kind of burden occurs in other microbes, we compared ribosomal profiling data (36) and homozygous mutant data (37) for budding yeast *Saccharomyces cerevisiae* growing in rich media. We found that about 25% of protein production is for genes that are not important for fitness. We suspect that most microbes invest significantly in the expression of proteins that are "on standby" in case conditions change.

The high aggregate cost of unnecessary expression suggests that it might be possible to engineer strains with reduced genomes that will grow faster or more efficiently. For example, Posfai and colleagues constructed a strain of *E. coli* K-12 with 42 deletions that removed 14% of the genome (38). We estimate that deleting these genes saved 2.6% of protein production (we ignored any regulatory effects of the deletions). But, Posfai and colleagues also removed six of the genes that we identified as being important for fitness (*yagM*, *ydbD*, *ynbB*, *wcaE*, *yfdI/gtrS*, and *yfjI*), which may explain why they did not observe any improvement in the growth rate.

Adaptive just-in-case expression of many genes

The high expression of genes that are important in other conditions need not imply that those genes are constitutively expressed. Of the 227 highly-expressed genes with no measurable phenotype in glucose minimal media and no expectation of genetic redundancy, 45% are known to be regulated by one or more transcription factors (30). This is similar to the rate for all genes (38%). Because many of the transcription factors in *E. coli* K-12 are still poorly characterized, the true proportion could be much higher.

High unnecessary expression of regulated genes may seem paradoxical, but intuitively, if the good times are not likely to last, then it is adaptive to express these genes at significant (but not maximal) levels: turning them on only when needed could lead to a long lag in growth (39; 40; 41). Alternatively, the high expression of these genes in artificial conditions could reflect their regulation by signals that are not directly related to their function (4), and such high levels of unnecessary expression might not occur in natural conditions.

Might "unnecessary" proteins for nutrient uptake have subtle benefits?

We have already mentioned that stress resistance proteins that have strong phenotypes in the laboratory might have subtle benefits under natural conditions. Similarly, some of the proteins that we identified as being important for utilizing alternate nutrients could actually be important for salvaging nutrients. For example, consider the high and "unnecessary" expression of uptake systems for amino acids and peptides (i.e., LivJ, OppA, MetQ, and DppA in Table 1). If proteins are occasionally degraded in the periplasm, then amino acids and peptides would be released. Importing them back into the cytoplasm instead of letting them diffuse out of the cell would have subtle benefits, especially once nutrient levels drop. In theory, this could be tested by knocking out amino acid or peptide uptake systems (although multiple mutations might be necessary) and checking for extracellular accumulation of amino acids or peptides.

Proteins with tiny benefits will not be maintained

We propose a minimum threshold for the benefit of a protein, on the assumption that it will not be maintained by natural selection unless the benefit exceeds the cost. We found that 96% of proteins with a detectable fitness advantage had a cost of above 10 ppm. Similarly, in *S. cerevisiae* growing in rich media, 89% of proteins with a detectable fitness advantage are expressed at 10 ppm or higher (combining (37; 36)). It is possible that a protein with subtle benefits might not require such high

expression, but we found that the median protein without a measurable phenotype was still expressed at 13 ppm. So, we propose that proteins with benefits of 10 ppm or less will be selected against. Although a benefit of 10 ppm might seem small, such tiny benefits are sometimes considered in evolutionary theory, such as in a recent model of selection for genetic redundancy (42). We argue that evolutionary models that rely on such subtle benefits of a gene are not realistic because of the cost of protein production.

Conclusions

Proteins that are important for fitness are highly expressed, but many of the highly-expressed proteins are not important for fitness. Some of this is due to genetic redundancy, but most of these proteins are important for fitness in other conditions and are expressed because of the possibility of a rapid change in conditions. In aggregate, this preparation accounts for 10-20% of the protein in the cell. The bulk of this investment is due to around 200 highly-expressed proteins, but most other proteins that are not important for fitness are still expressed at detectable levels and have significant costs during evolution.

Materials and Methods

Measuring mutant fitness in minimal glucose medium

We used a collection of 152,018 randomly-barcoded transposon mutants that were derived from *E. coli* strain BW25113 (17). The pool of mutants was recovered from the freezer by growing it in rich medium (LB) until $OD_{600} = 1$ and pelleted, and an initial sample was collected. The remaining cells were washed and inoculated into two different 2 liter flasks, each with 200 mL of MOPS minimal medium (Teknova) supplemented with 2 g/L D-glucose, at an initial OD = 0.02. (MOPS includes inorganic salts as well as 3-(N-morpholino)propanesulfonic acid and tricine as buffering agents, but does not include any vitamins.) The cells grew aerobically at 37°C until late exponential phase (OD = 0.57-0.59), were diluted back to OD = 0.02, and grew again to saturation (OD = 2.8-3.1). Thus the cells grew in minimal media for a total of about 12 generations. To compare the abundance of each strain at the end of each experiment to its abundance at the beginning, we used DNA barcode sequencing (43) with Illumina. Specifically, we extracted genomic DNA and performed PCR using the 98°C protocol (17).

We sequenced these three samples using Illumina HiSeq. For each sample, we obtained 23-26 million reads with barcodes that matched the pool. We also sequenced

these three samples on a MiSeq instrument, along with two additional samples that were collected from the two replicate cultures before the first transfer (at about 5 generations). The MiSeq run had 1.7-3.0 million reads per sample.

We computed gene fitness values as described previously (17). Briefly, the fitness of a strain is the normalized $\log_2$ ratio of the number of reads, and the fitness of a gene is the weighted average of the fitness values for strains with insertions in the central 10-90% of the gene. In each HiSeq sample, the median gene had around 2,000 reads for relevant strains. The two replicate cultures yielded similar gene fitness values (r = 0.995; Figure 3A). Fitness at 12 generations (from HiSeq) was strongly correlated with fitness at 5 generations (r = 0.936, Figure 3B). Gene fitness at 12 generations was also similar (r = 0.91) to the results of an independent experiment on a different day with the same pool of mutants, a higher concentration of D-glucose (3.96 g/L), a smaller volume (10 mL), and fewer generations of growth (about 6.9; experiment set2IT096 of (17)).

Gene fitness values at 12 generations were usually more extreme than those at 5 generations, which indicates that the abundance of the mutants continued to change in the same direction from 5-12 generations as they had during 0-5 generations (Figure 3B). This shows that most of these genes were important during exponential growth, which is when the ribosomal profiling data was collected, and not just for the transition to growth in this medium. All 13 of genes that were significantly important for fitness (as described below) despite weak expression of under 2 monomers per million were also below the line (Figure 3B). (None of the genes that were detrimental to fitness were so weakly expressed.)

Identifying significant phenotypes in minimal glucose medium

For each replicate and for each gene, we computed a t-like test statistic that takes into account the variability of the fitness values for the strains for each gene (17). To identify genes with statistically significant changes, we wanted to combine the two t values, but they are not independent as they both used the same data for the initial sample. So, instead of using the usual way of combining t values of $(t_A + t_B)/\sqrt{2}$, we used $t_{comb} = (t_A + t_B)/\sqrt{3}$. The increased denominator makes up for the partial non-independence. To see why 3 instead of 2 is correct, consider the expected variance of the sum of the two fitness values, and remember that fitness is the difference of log abundances. With non-independent start samples, the variance is $\text{Variance}(A - C + B - C) = \text{Variance}(A) + \text{Variance}(B) + 4 \cdot \text{Variance}(\text{C})$, while with independent start samples, it is $\text{Variance}(A - C + B - D) = \text{Variance}(A) + \text{Variance}(B) + \text{Variance}(\text{C}) + \text{Variance}(\text{D})$. (Remember that $\text{Variance}(A + B) =$

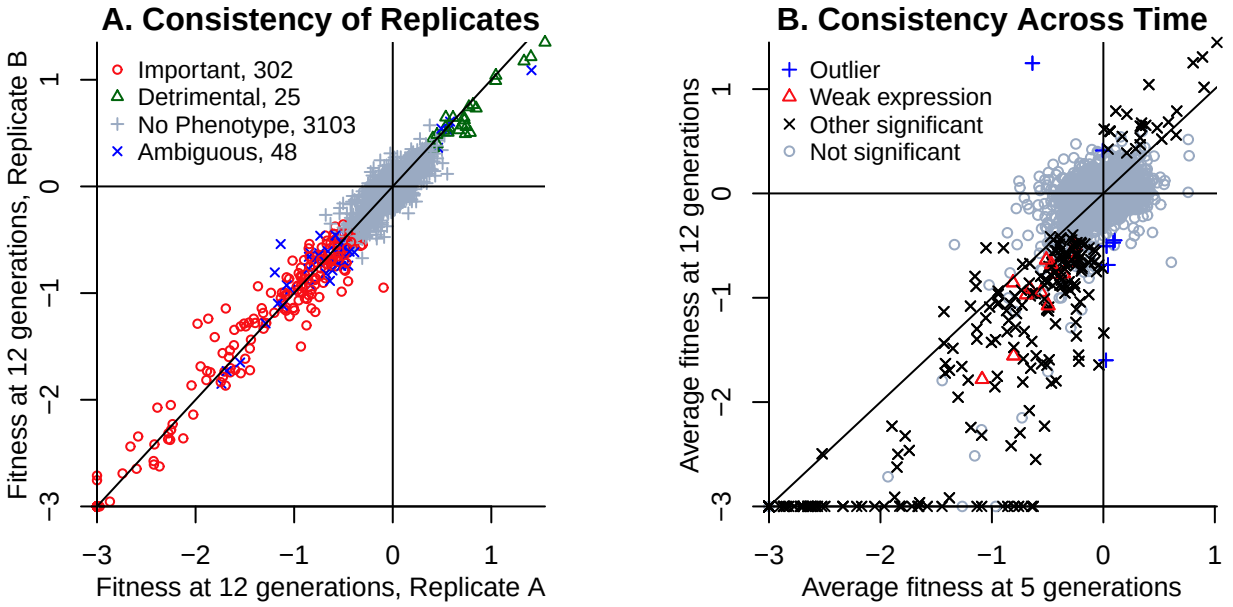

Figure 3: Consistency of gene fitness values in minimal glucose medium. In both panels and on both axes, values less than -3 are shown at -3. (At 12 generations, a fitness of -3 corresponds to a 16% reduction in growth rate; at 5 generations, it corresponds to a 34% reduction in growth rate.) In panel A, 123 genes are at the bottom left corner; in panel B, 78 genes are at the bottom left corner. The lines show $x = 0$, $y = 0$, and $x = y$. In panel B, the genes with significant phenotypes (of either sign) are subdivided into those with weak expression (under 2 monomers per million) or above.

Variance(A) + Variance(B) if A and B are independent.) If each component has about the same variance, then the non-independence increases the variance by a factor of $6/4 = 3/2$.

Genes were considered to significantly affect fitness if $|t_{comb}| > 4$. If t follows the standard normal distribution then we expect 0.2 false positives. As a control, we compared the two 12-generation samples to each other and identified just 1 gene with $|t| > 4$.

Seven genes had strains with statistically significant changes at 12 generations but had a different sign after 5 generations (these can be seen as outliers in Figure 3B). Another 39 genes had mutants that showed non-significant changes in abundance of 3% per generation or more. These 48 genes were classified as ambiguous.

Experimental differences between the fitness data and the ribosomal profiling data

The ribosomal profiling data was from strain MG1655, while our transposon mutants were made from strain BW25113. The genomes of these strains were compared by (44). BW25133 lacks *araBAD*, *rhaDAB*, or *valX*, has a truncated and modified *lacZ*, has a frameshift in *hsdR*, and has a premature stop codon in *yjjP*. BW25133 also lacks 110 nt in an intergenic region. MG1655 (but not BW25113) has mobile element insertions in *crl*, in *mhpC*, and in two intergenic regions, and has frameshifts in *glpR* and in *gatC*. The strains also differ at a 3 nt stretch in *rrlD* and 13 other single-nucleotide substitutions. We do not expect these differences to lead to global changes in gene expression or in growth. Incidentally, both strains have a frameshift mutation in *rph* that reduces the expression of the downstream gene *pyrE*, which is required for pyrimidine biosynthesis; this mutation reduces the growth rate in minimal media by around 10% (45).

The media formulations for the two types of experiments were identical, and both experiments used a culture volume of 200 mL at 37°C and shaking at 180 rpm. However, the ribosomal profiling experiments used 2.8 L flasks, while for mutant fitness experiments we used 2.0 L flasks, so the concentration of oxygen might not have been identical.

A fitness compendium for Escherichia coli K-12

This compendium includes previously-described fitness experiments with various carbon sources and M9 minimal media (17). Nitrogen source experiments were conducted similarly, with D-glucose as the carbon source. Stress experiments were conducted in LB, at 28°C instead of 37°C, and in a 48-well microplate. Inhibitors were added at a concentration that would reduce the growth rate by about 2-fold. All of these experiments were inoculated at OD = 0.02 and grown aerobically until saturation. Only experiments that met standards for internal and biological consistency (17) were retained for analysis.

Fitness values and t scores were obtained as described above. t values from replicate experiments were combined as described above if they shared a control. If there were 3 or 4 replicates with shared controls, then the variance was reduced by 2 or 2.5 instead of by 1.5 fold. If the replicates were fully independent, then the t values were combined in the traditional way ($t_{comb} = \sum t/\sqrt{n}$, where n is the number of replicates). For stress experiments, only independent samples grown at the same concentration were considered to be replicates.

Across the compendium, genes were considered to have a significant phenotype

if, in any condition, average fitness was under -0.5 and $t_{comb} < -4$. There were 1,107 such genes. In 13 control comparisons between independent samples from the same culture, this threshold was never reached. (Given $13 \cdot 3,789$ values from the standard normal distribution, the expected number of values under -4 would be 1.6.) Based on the standard normal distribution, we would expect 13 false positives in this data set, or a false discovery rate of about 1%.

For the analysis of individual experiments, a phenotype was considered significant if fitness was under -1 and $t < -4$.

Software

The *E. coli* fitness experiments were analyzed using FEBA statistics version 1.0.1 (https://bitbucket.org/berkeleylab/feba). Statistical analyses were conducted in R 2.15.0.

Data Availability

Tables of counts per barcode, fitness values, and t values are available at http://genomics.lbl.gov/strongselection/, as are supplementary tables 1 and 2. Also, the *E. coli* fitness compendium can be browsed at http://fit.genomics.lbl.gov/.

Supplementary Tables

Supplementary Table 1: Comparison of fitness data and ribosomal profiling data.

Supplementary Table 2: Manual classification of highly expressed genes with no phenotype in minimal glucose.

Acknowledgements

We thank Mark Callaghan for technical assistance with the *E. coli* fitness compendium.

Funding

This material by ENIGMA - Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory, is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231.

References

1. Fischer E, Sauer U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. Nat Genet. 2005;37:636–40.

2. Qian W, Ma D, Xiao C, Wang Z, Zhang J. The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. Cell reports. 2012;2(5):1399–1410.

3. Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. Bacterial adaptation through loss of function. PLoS Genetics. 2013;9(7):e1003617.

4. Price MN, Deutschbauer AM, Skerker JM, Wetmore KM, Ruths T, Mar JS, et al. Indirect and suboptimal control of gene expression is widespread in bacteria. Molecular Systems Biology. 2013;9(1).

5. Deutschbauer A, Price MN, Wetmore KM, Tarjan DR, Xu Z, Shao W, et al. Towards an informative mutant phenotype for every bacterial gene. Journal of bacteriology. 2014;196(20):3643–3655.

6. Weiße AY, Oyarzún DA, Danos V, Swain PS. Mechanistic links between cellular trade-offs, gene expression, and growth. Proceedings of the National Academy of Sciences. 2015;112(9):E1038–E1047.

7. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;330(6007):1099–1102.

8. Shachrai I, Zaslaver A, Alon U, Dekel E. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. Mol Cell. 2010 Jun;38(5):758–767.

9. Tomala K, Korona R. Evaluating the Fitness Cost of Protein Expression in Saccharomyces cerevisiae. Genome biology and evolution. 2013;5(11):2051–2060.

10. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proceedings of the National Academy of Sciences. 2002;99(6):3695–3700.

11. Bragg JG, Wagner A. Protein material costs: single atoms can make an evolutionary difference. Trends in Genetics. 2009;25(1):5–8.

12. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell. 2014;157(3):624–635.

13. Price MN, Arkin AP. Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes. mBio. 2015;6(6):e01302–15.

14. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Molecular systems biology. 2015;11(2):784.

15. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent Escherichia coli proteome. Nature biotechnology. 2016;34(1):104–110.

16. Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, et al. Quantitative prediction of genome-wide resource allocation in bacteria. Metabolic engineering. 2015;32:232–243.

17. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. mBio. 2015;6(3):e00306–15.

18. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2:2006.0008.

19. Kato Ji, Hashimoto M. Construction of consecutive deletions of the Escherichia coli chromosome. Molecular systems biology. 2007;3(1).

20. Nakahigashi K, Kubo N, Narita Si, Shimaoka T, Goto S, Oshima T, et al. HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. Proceedings of the National Academy of Sciences. 2002;99(3):1473–1478.

21. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, et al. EcoCyc: fusing model organism databases with systems biology. Nucleic acids research. 2013;41(D1):D605–D612.

22. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic acids research. 2008;36(suppl 1):D866–D870.

23. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic acids research. 2013;41(D1):D203–D213.

24. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, et al. Cryptic prophages help bacteria cope with adverse environments. Nature communications. 2010;1:147.

25. Béjar S, Bouché F, Bouché JP. Cell division inhibition gene dicB is regulated by a locus similar to lamboid bacteriophage immunity loci. Molecular and General Genetics MGG. 1988;212(1):11–19.

26. Yun SH, Ji SC, Jeon HJ, Wang X, Kim SW, Bak G, et al. The CnuK9E H-NS complex antagonizes DNA binding of DicA and leads to temperature-dependent filamentous growth in E. coli. PloS one. 2012;7(9):e45236.

27. Seaver LC, Imlay JA. Alkyl hydroperoxide reductase is the primary scavenger of endogenous hydrogen peroxide in Escherichia coli. Journal of bacteriology. 2001;183(24):7173–7181.

28. Carlioz A, Touati D. Isolation of superoxide dismutase mutants in Escherichia coli: is superoxide dismutase necessary for aerobic life? The EMBO Journal. 1986;5(3):623.

29. Kayser A, Weber J, Hecht V, Rinas U. Metabolic flux analysis of Escherichia coli in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state. Microbiology. 2005;151(3):693–706.

30. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic acids research. 2015;p. gkv1156.

31. Xie QW, Tabor CW, Tabor H. Deletion mutations in the speED operon: spermidine is not essential for the growth of Escherichia coli. Gene. 1993;126(1):115–117.

32. Hafner EW, Tabor CW, Tabor H. Mutants of Escherichia coli that do not contain 1, 4-diaminobutane (putrescine) or spermidine. Journal of Biological Chemistry. 1979;254(24):12419–12426.

33. Ebersbach G, Galli E, Møller-Jensen J, Löwe J, Gerdes K. Novel coiled-coil cell division factor ZapB stimulates Z ring assembly and cell division. Molecular microbiology. 2008;68(3):720–735.

34. Malinverni JC, Silhavy TJ. An ABC transport system that maintains lipid asymmetry in the Gram-negative outer membrane. Proceedings of the National Academy of Sciences. 2009;106(19):8009–8014.

35. Wagner A. Risk management in biological evolution. J Theor Biol. 2003;225:45–57.

36. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–223.

37. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics. 2005;169(4):1915–1925.

38. Pósfai G, Plunkett G, Fehér T, Frisch D, Keil GM, Umenhoffer K, et al. Emergent properties of reduced-genome Escherichia coli. Science. 2006;312(5776):1044–1046.

39. Boulineau S, Tostevin F, Kiviet DJ, ten Wolde PR, Nghe P, Tans SJ. Single-cell dynamics reveals sustained growth during diauxic shifts. PLoS ONE. 2013;8(4):e61686.

40. Wang J, Atolia E, Hua B, Savir Y, Escalante-Chong R, Springer M. Natural variation in preparation for nutrient depletion reveals a cost-benefit tradeoff. PLoS Biol. 2015;13(1):e1002041.

41. Venturelli OS, Zuleta I, Murray RM, El-Samad H. Population diversification in a yeast metabolic program promotes anticipation of environmental shifts. PLoS Biol. 2015;13(1):e1002042.

42. Ho WC, Zhang J. The Genotype–Phenotype Map of Yeast Complex Traits: Basic Parameters and the Role of Natural Selection. Molecular biology and evolution. 2014;31(6):1568–1580.

43. Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, et al. Quantitative phenotyping via deep barcode sequencing. Genome research. 2009;19(10):1836–1842.

44. Grenier F, Matteau D, Baby V, Rodrigue S. Complete genome sequence of Escherichia coli BW25113. Genome announcements. 2014;2(5):e01038–14.

45. Jensen KF. The Escherichia coli K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. Journal of bacteriology. 1993;175(11):3401–3407.