RESEARCH ARTICLE

A Comparison of the Costs and Benefits of Bacterial Gene Expression

Morgan N. Price*, Kelly M. Wetmore, Adam M. Deutschbauer, Adam P. Arkin*

Environmental Genomics and Systems Biology, Lawrence Berkeley National Lab, Berkeley, California, United States of America

* morgannprice@yahoo.com (MNP); aparkin@lbl.gov (APA)

OPEN ACCESS

Citation: Price MN, Wetmore KM, Deutschbauer AM, Arkin AP (2016) A Comparison of the Costs and Benefits of Bacterial Gene Expression. PLoS ONE 11(10): e0164314. doi:10.1371/journal.pone.0164314

Editor: Brian Frederick Pfleger, University of Wisconsin Madison, UNITED STATES

Received: July 1, 2016

Accepted: September 22, 2016

Published: October 6, 2016

Data Availability Statement: The gene-level data is contained within the paper and its supporting information files. Tables of barcodes and counts are available at http://genomics.lbl.gov/supplemental/strongselection/. Also, the fitness compendium for E. coli can be browsed at http://fit.genomics.lbl.gov/.

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,

Abstract

To study how a bacterium allocates its resources, we compared the costs and benefits of most (86%) of the proteins in *Escherichia coli* K-12 during growth in minimal glucose medium. The cost or investment in each protein was estimated from ribosomal profiling data, and the benefit of each protein was measured by assaying a library of transposon mutants. We found that 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 (with a benefit of less than 5% per generation) 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 22% 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 at least 13% of the cell's protein is "on standby" in case conditions change.

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 [1–5].

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, 7] or a small multiple of this [8]. In laboratory environments, the measured fitness cost of a useless and harmless protein is about 1–2 times the fraction of protein [8–10].

Bacterial proteins are typically expressed at 3–21 parts per million of the protein mass of a cell (data of [11], 25th-75th percentile). Although a cost of 3 ppm may seem small, it should be significant for evolution. The effective population sizes (N_e) for bacteria are estimated at

Office of Biological & Environmental Research under contract number DE-AC02-05CH11231. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

around 10^6 or 10^7 [12], and under the nearly-neutral theory of molecular evolution, this implies that alleles that increase fitness by just 1 ppm should predominate over evolutionary time ($N_e \cdot s > 1$, where s is the selection coefficient; see [13]).

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. Another approach has been to rely on theoretical models of metabolism and protein production to predict the optimal expression levels [16, 17]. These models include enzymes, ribosomes, and some other steps in gene expression; they include most of the highly expressed proteins [17], but they omit the majority of the genes in the genome. In *Bacillus subtilis*, the concentrations of most enzymes can be explained by their theoretically-optimal flux [16], while in *E. coli*, the models predict that many proteins are expressed when they are not needed or are expressed at unnecessarily high levels [17]. 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 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 [11], which allows us to study weakly-expressed proteins. To measure the benefit of each protein, or its importance for growth, we used a barcoded library of about 150,000 transposon mutants [18] as well as information from individual knockout strains [19, 20]. We found that 96% of protein-coding genes that had mutant phenotypes were expressed at above 10 ppm of protein mass or above 40 monomers per cell, and their median expression was 205 ppm. In contrast, genes that did not have a measurable impact on fitness had a median expression of 13 ppm. 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, we estimate that in this condition, 22% of protein production is unnecessary. Many of these proteins are only expected to be important in other conditions, given their known functions. Indeed, by examining a large compendium of genome-wide fitness assays, we found that the majority of this unnecessary expression, or 13% of total protein, is for proteins that have significant phenotypes in other conditions. We propose that these proteins are "on standby" in case conditions change.

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 [11] 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 [11]. Also, Li and colleagues report that their quantitation is accurate for genes with over 128 reads, which corresponds to roughly 1 ppm of expression. Genes with fewer reads are probably expressed at under 1 ppm. 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 [11].

To measure the importance of each protein for growth, we used a pooled library of about 150,000 transposon mutants with random DNA barcodes [18]. Each insertion contains a kanamycin resistance marker with its own promoter and a random 20-nucleotide "barcode." We previously mapped the location of each insertion as well as the associated barcode, and the barcodes can be amplified by polymerase chain reaction and sequenced to quantify the abundance of each strain [18].

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 [18]. 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 Profiling of *E. coli* Chromosome (PEC) [20] and growth data on individual deletion strains from the Keio collection [19]. From PEC, we identified 286 essential proteins; we classified non-essential genes whose mutants had a significant change in abundance during our pooled assay as either important for fitness (294 genes) or detrimental to fitness (25 genes); we classified 25 other genes with low coverage in our mutant pools as important for fitness because mutants grew poorly in both minimal MOPS media and LB [19]; we classified proteins with non-significant changes in mutant abundance of above 3% per generation (48 genes) or with insufficient coverage (449 genes) as ambiguous; and we classified the remaining 2,944 proteins as having no phenotype.

To determine the statistical significance of each gene's effect on fitness, we used a *t*-like test to quantify how consistently the independent insertion mutants in that gene had altered abundance [18]. 96% of the non-essential non-ambiguous genes have 5 or more independent insertions, and 80% of them have 10 or more independent insertions. 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 significant phenotypes (see Methods). The genes with no phenotype 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 phenotypic classification of the genes is included in S1 Table.

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 (Fig 1A; S1 Fig). 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.

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 gene sizes in the 25th-75th percentile, 10 ppm corresponds to 30–70 monomers per cell.)

All of the essential proteins [20] were expressed at above 5 ppm, except for the putative protein YceQ, which had no ribosomal profiling reads at all. YceQ also lacked ribosomal profiling reads in two other growth conditions [11]; YceQ lacks homology to any other protein; the open reading frame is disrupted in some strains of *E. coli*; and we identified transposon insertions within *yceQ* ([18]; S2 Fig). It appears that *yceQ* does not encode a protein. Deletion of the entire *yceQ* region may not be possible because it contains the promoter of the essential gene *rne* (S2 Fig). 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 [20]; it is not entirely essential but a mutant has severely reduced growth [21]).

Some very weakly-expressed non-essential genes were identified as being important for fitness, including 6 proteins with expression of under 1 ppm of 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 [22]. We are not sure how these proteins could have a measurable effect at 6 copies per cell unless they are regulatory proteins. Another 7 genes were important for fitness despite weak expression of 1–2 ppm of monomers (6–12 copies per cell), including three proteins that are involved in the uptake of iron via the siderophore enterobactin (FepD, FepG, and Fes). Differences between the genetic backgrounds of the two data sets, or subtle differences in growth conditions, might lead to these rare 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).

Overall, we find that almost all proteins are expressed at above 10 ppm of monomers (or roughly 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).

High expression of many proteins with no expected benefit

31% of total expression (by mass) was due to the 2,944 proteins with no measurable impact on fitness (Fig 1B). As shown in Fig 1A, the distribution of expression is quite skewed, so most of this 31% is due to a few hundred well-expressed genes. We decided to focus on 287 proteins

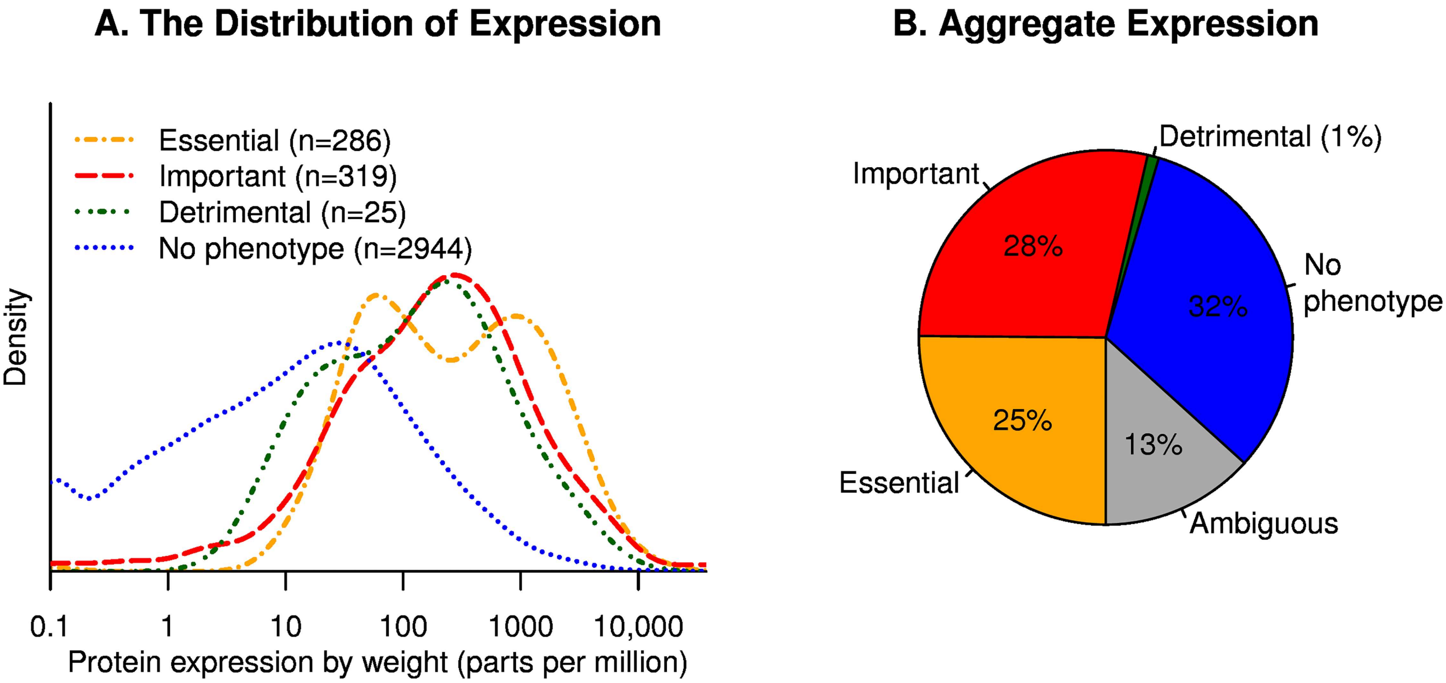

Fig 1. The production of proteins versus their importance for growth. (A) For each class of gene, we show the distribution of protein expression, in parts per million of amino acids (*x* axis, log scale). Proteins with little or no expression are shown at 0.1 ppm. (B) The aggregate expression of each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

doi:10.1371/journal.pone.0164314.g001

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

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

doi:10.1371/journal.pone.0164314.t001

that were expressed at above 200 ppm and had no measurable impact on growth; these account for 24% of total protein production. Proteomics data [15] confirms that most of these "unnecessary" proteins are highly expressed (see Methods).

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 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 [23]. 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 [24]. 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 and AceB are involved in central metabolism—they encode the glyoxylate shunt—but a metabolic model predicts that they are not required for optimal growth on glucose [25], and ^{13}C labeling studies suggest that they do not carry flux in this condition [26, 27]. PntA (the pyridine nucleotide dehydrogenase α subunit) is also involved in central metabolism and is predicted to be dispensible for optimal growth on glucose [25], although it is not clear whether this prediction is correct. (A strain that lacked both subunits (*pntAB*$^{-}$) was reported to have a 30% reduction in growth rate on glucose [26], while we observed a defect of just 2% in mutants of either *pntA* or *pntB*, and this defect was not statistically significant. The discrepancy could be due to differences in growth conditions). Similarly, in *B. subtilis*, some enzymes in central metabolism are highly expressed even when they carry no flux [16]. The highly-expressed *E. coli* proteins 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).

To more systematically examine the highly-expressed and "unnecessary" proteins, we examined the EcoCyc [22] entries for all 287 highly-expressed genes that lack phenotypes (S2 Table). 106 of these 287 proteins are expected to be important in other conditions, as they are involved in utilizing alternate nutrients (66 proteins), stress resistance (24 proteins), parts of central metabolism that are dispensible in our condition (11 proteins), or anaerobic growth (5 proteins). These 106 unnecessary proteins account for 11.4% of total protein production. Another 60 of the 287 proteins are involved in key processes that are expected to be important for growth in our condition, but are redundant because of isozymes (as with SodA and SodB above) or more indirect redundancy (see Appendix 1). These 60 genetically-redundant proteins account for 6.8% of protein production. Another 10 proteins are involved in key processes and have little phenotype, even though they are not redundant as far as we know: these include 4 non-essential components of essential complexes and 2 proteins involved in the modification of tRNA or rRNA (see Appendix 1). These proteins might have fine-tuning or regulatory roles. Just 8 of the 287 proteins are characterized transcriptional regulators [28]. The majority of the remaining genes are poorly characterized: 62/103 (60%) have names that begin with *y*.

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 (6.8/24.1 = 28%), then we estimate that 31% · (1 − 0.28) = 22% of protein production is unnecessary in this growth condition. More conservatively, if we consider only the proteins of known function that are not expected to be important, then we estimate that 11.4/24.1 · 31% = 15% of protein production is unnecessary.

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* ([18]; 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 Fig 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%).

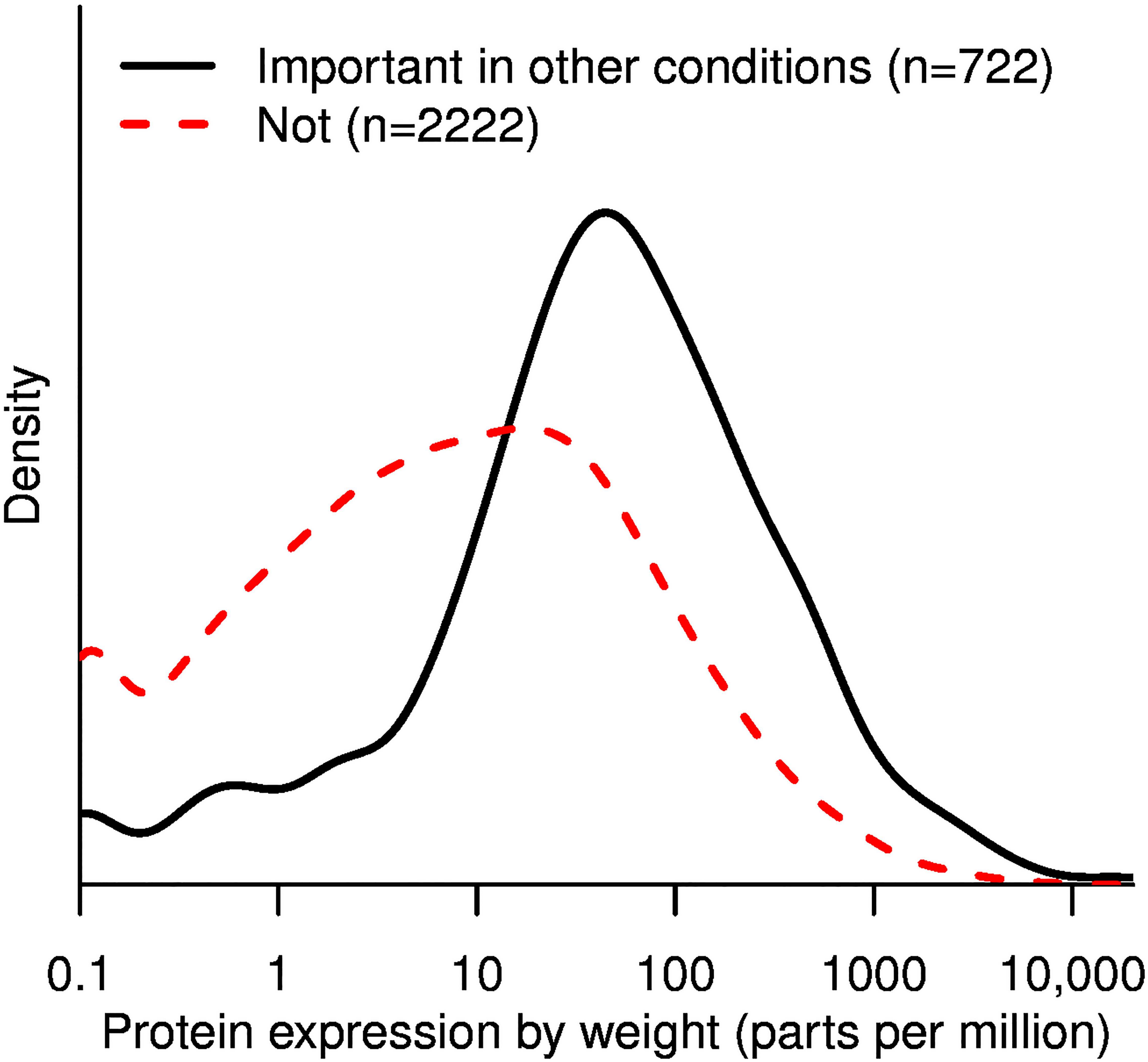

Fig 2. The production of "unnecessary" proteins versus their importance in other conditions. Only genes that were not important for fitness in minimal glucose media are included. The *x* axis is as in Fig 1A.

doi:10.1371/journal.pone.0164314.g002

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. We thought that this caveat would be less likely to be relevant for carbon sources or nitrogen sources: *E. coli* would probably not be able to consume these nutrients unless they were sometimes important in nature. So we considered each carbon or nitrogen source experiment individually, and asked whether proteins that were important for utilizing that nutrient (but not glucose) where more highly expressed than the typical protein (in minimal glucose media). Also, we considered only experiments with 5 or more important proteins. In 82 of 96 experiments, the median protein that was important for fitness (in the condition but not in glucose) was expressed at least 3-fold higher than the median protein (in glucose). Thus, we propose that much of the "unnecessary" expression represents preparation for changing conditions.

As discussed above, if we correct for genetic redundancy then about 22% of expression seems to be unnecessary. And of the unnecessary expression, 57% is important for fitness in another condition. So we estimate that *E. coli* invests at least 22% · 57% = 13% of protein to prepare for other conditions.

We also note that 22 of the 25 the proteins that were detrimental to fitness in glucose media had significant benefits in at least one other condition. This kind of conflict between adaptation to different conditions ("antagonistic pleiotropy") has been reported in diverse microorganisms [1–5].

Little of the unnecessary expression is due to operon structure

As an alternative explanation for the "unnecessary" expression, we considered whether some of these proteins might be highly expressed because they are cotranscribed with important genes. Of the 287 proteins, just 28 (10%) are known or predicted to be in operons with essential or important genes (RegulonDB 9.1; [28]). Furthermore, when we examined these 28 cases in more detail, 11 of the "unnecessary" proteins are involved in key processes, and in another 5 cases, the operon has internal promoters or small RNA regulation such that the "unnecessary" gene could be expressed independently of the nearby important gene. For example, the *yjeF-tsaE-amiB-mutL-miaA-hfq-hflXKC* superoperon includes the essential protein MiaA, which is moderately expressed (32 ppm), and two well-expressed "unnecessary" proteins HflC and HflK (both over 200 ppm). However, there are several promoters between *miaA* and *hfq* that could drive higher expression of *hfq-hflCK* [29]. After excluding these cases, just 12 of 287 "unnecessary" proteins (4%) might be highly expressed because of operon structure.

Budding yeast also incurs a significant burden due to "unnecessary" expression

To compare the costs and benefits of gene expression in another microorganism, we examined ribosomal profiling data [30] and homozygous mutant data [31] from the budding yeast *Saccharomyces cerevisiae* growing in rich media. As in *E. coli*, most of the proteins that had a significant effect on fitness were well expressed: 89% of them were expressed at 10 ppm or more of protein monomers. Nevertheless, 26% of monomers in *S. cerevisiae* were for proteins that are not important for fitness, and these account for 32% of protein production by weight. We propose that most microbes invest significantly in the expression of proteins that are "on standby" in case conditions change.

Discussion

Do "unnecessary" proteins have subtle benefits?

A major caveat in our results is that the "unnecessary" proteins might have benefits that are too subtle for us to measure. Our assay is sensitive to a loss of fitness due to disabling a gene of about 4% per generation (which corresponds to a 39% reduction in mutant abundance after 12 generations). But natural selection could maintain the expression of proteins with far smaller benefits.

We relied on our understanding of *E. coli*'s physiology to argue that many of these proteins are unlikely to provide a subtle benefit. Highly-expressed proteins that are unlikely to be beneficial were involved in steps in central metabolism that do not carry flux (such as *aceAB*), in anaerobic growth, or in the utilization of nutrients that were not provided in the growth medium. A recent comparison of proteomics data to a model of *E. coli* metabolism and gene expression came to similar conclusions about the high level of "standby" expression in *E. coli* [17], although there are some disagreements between our mutant fitness data and their theoretical model (see Appendix 2).

A few of the "unnecessary" proteins might be involved in the salvage of components that leak out of the cell. For example, of the 66 highly-expressed and "unnecessary" proteins for utilizing alternate nutrients, 16 are involved in the uptake of amino acids or peptides, and these account for 2% of total protein. Although neither amino acids nor peptides were present in our media, they might be released by the cells. For example, peptides might be released after the degradation of misfolded periplasmic proteins by DegP (also known as HtrA) or OmpT, as both of these proteases are highly expressed in our growth condition. But we doubt that the secretion of amino acids or peptides is sufficient to justify an investment of 2% of protein in recovering them. A similar scenario arises with glutathione, which can be secreted to a concentration of over 100 μM by *E. coli* while growing in minimal glucose media [32]. (The secretion of glutathione may be a consequence of the domestication of *E. coli* K-12, as it was not observed with clinical isolates [33].) The GsiB protein (formerly known as YliB) accounts for 0.02% of protein and is involved in the reuptake of the secreted glutathione [32]. Although we did not observe a benefit for *gsiB*—we estimated a +1% change in mutant abundance per generation, which was not statistically significant—it is possible that it has a tiny benefit that is too small for us to observe.

Overall, we cannot be certain of the lack of benefit for any one of these genes, but it does not seem plausible that the expression of most of these genes is beneficial.

Adaptive just-in-case expression of many genes

Instead, we argue that the expression of many of these proteins is due to preparation for other conditions. This was supported by the observation that many of the "unnecessary" and highly-expressed proteins are measurably important for fitness in other conditions.

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 [28]. 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 [34–36]. 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.

The cost of being a generalist

We estimate that 31% of *E. coli*'s protein production in minimal glucose medium is due to genes that are not important for fitness in this condition. Correcting for genetic redundancy reduced this estimate to 22%. Furthermore, the majority of the burden was due to genes that seem to be "on standby", as they are important for fitness in other conditions. Thus, we conservatively estimate that 13% of total protein is on standby. Because the fitness cost of useless protein is at least as large as the fraction of protein [6, 8–10], the burden of preparing for other conditions could reduce *E. coli*'s growth rate by more than 13%. Similarly, metabolic modeling suggests that, across a range of carbon sources, *E. coli* grows more slowly when a higher proportion of protein production is unnecessary [17].

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 [37]. 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.

Another aspect of being a generalist is the need for gene regulation. Among the 186 homomeric transcription factors that have been characterized in *E. coli* [28], the typical expression 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 cause rarely-needed capabilities to be lost from most members of a population [38].

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 *E. coli* 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 were so well expressed. It is possible that a protein with subtle benefits might not require such high expression, but we found that the median *E. coli* 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 [39]. We argue that evolutionary models that rely on such subtle benefits of a gene are not realistic because of the cost of protein production.

The challenge of detecting the cost of unnecessary expression

Although an unnaturally highly-expressed gene can cause a measurable reduction in the growth rate [8–10], it is challenging to measure the reduction in the growth rate due to unnecessary expression at natural levels. In minimal glucose media, many of the highly and "unnecessarily" expressed proteins account for less than 0.1% of protein each, which implies that it would take over 100 generations to see a 10% increase in the relative abundance of a deletion strain. An experiment of this length is risky because of secondary mutations. The most highly-expressed unnecessary (and non-redundant) protein was just 0.6% of protein (LivJ, see Table 1).

Would the cost of expressing unnecessary protein be detectable in laboratory evolution experiments? A cost of 0.6% implies that it would take 115 generations for the abundance of a beneficial mutant to increase by even two-fold. If the mutant is initially very rare, then it still will not be detectable. Furthermore, in laboratory evolution experiments with bacteria, strongly beneficial mutations are common, which leads to "hitchhiking"—if a strongly-beneficial mutation happens to arise in a genetic background that contains mildly-deleterious variants, then those deleterious variants will increase in abundance. Thus, hitchhiking weakens the impact of selection on variants with more subtle effects. In theory, variants become effectively neutral if they are under selection that is an order of magnitude weaker than that of the strongly-beneficial mutations [40]. Because beneficial mutations often have an advantage of several percent per generation or more [41], the loss of an unnecessary protein that is expressed at 0.3% would be effectively neutral.

In practice, it appears that some unnecessary catabolic capabilities are lost during the evolution of *E. coli* in the laboratory in a glucose medium, but most remain, even after 50,000 generations [42]. Losses occur at a higher rate in strains with elevated mutation rates, which suggests that many of the losses are effectively neutral [42]. Also, because many of the losses of catabolic capabilities in the mutator strains are rescued by a reduction in temperature, these losses are probably due to mutations that cause proteins to misfold, rather than mutations that reduce unnecessary expression [42]. A few catabolic capabilities were lost in multiple lines, which seems to reflect selection against the deleterious activity of a protein, rather than selection on the cost of making an unnecessary protein. For example, during the growth of *E. coli* on glucose, there is strong selection (1–2% per generation) for the loss of D-ribose utilization [43]. This seems far too strong to be explained by the cost of expressing the ribose utilization operon, which we estimate at 0.03%. Instead, it appears that the activity of these proteins is detrimental. As another example, O'Brien and collagues [17] compared gene expression in lines of *E. coli* that had evolved in minimal glucose media for around 2,000 generations [44] to a metabolic model. They found that seven out of eight evolved lines allocated a larger fraction of their expression to important proteins than the starting strain did. This change in allocation seems to be due to mutations that affect global regulators *rpoB* and *hns* [44] and increase the expression of proteins that are important in the evolved condition, rather than any specific reduction in expression for the highly-expressed unnecessary proteins (see Appendix 3).

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, so we propose that they are expressed because of the possibility of a rapid change in conditions. In aggregate, this preparation accounts for at least 13% of the protein in *E. coli* growing in minimal glucose media. 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 by using a Tn5 transposase [18]. 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 [45] with Illumina. Specifically, we extracted genomic DNA and performed PCR using the 98°C protocol [18].

We sequenced these three samples using Illumina HiSeq and 50-nucleotide reads. 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 with 50-nucleotide reads, 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 [18]. Briefly, the fitness of a strain is the normalized $\log_2$ ratio of the number of reads. 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 and a sufficient number of reads in the initial sample. Strains with relatively few reads (under 20 per sample) are weighted lower because their fitness values are noisier. For the HiSeq data, our analysis of gene fitness relied on the most abundant 88,889 strains with insertions in the central 10%–90% of 3,478 genes, and less than 10% of these strains were down-weighted. The coverage of the median strain ranged from 96–114 reads per strain per sample and the coverage of the median gene ranged from 1,986–2,582 reads per gene per sample. Such high coverage implies that the noise in the fitness values due to the limited number of reads is quite modest: about ±0.2 for the typical strain and just ±0.05 for the typical gene. (The standard deviation of the noise due to the limited number of reads should be roughly two divided by the square root of the number of reads per sample, see [18].) The gene fitness values were then normalized to remove a bias due to variation in gene copy number along the chromosome [18]. The bias can be estimated accurately by using the running median of the gene fitness values across the chromosome, with a window size of 251 genes [18]. This estimate of the bias had the expected "V" shape with the minimum being at the origin of replication (S3 Fig). The maximum estimated bias for the 12 generation samples was 0.64 in $\log_2$ units, but the uncertainty in the normalization should be far less (S3 Fig). The gene fitness values were also normalized to set the mode (instead of the median) to be at zero, but for this experiment, this only altered the gene fitness values by 0.01.

The two replicate cultures yielded similar gene fitness values (r = 0.995; Fig 3A). Fitness at 12 generations (from HiSeq) was strongly correlated with fitness at 5 generations (r = 0.936, Fig 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 [18]).

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 (Fig 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 genes that were significantly important for fitness (as described below) despite weak expression of under 2 ppm of monomers were also below the line (Fig 3B). (None of the genes that were detrimental to fitness were so weakly expressed.)

As another way to check that the fitness data represents the impact of disabling each protein, we estimated the fitness at 15 generations separately for the first and second half of each gene, and then averaged the two independent replicate experiments (Fig 3C). We considered the 3,189 non-essential proteins that were included in the comparison to the ribosomal profiling data and have sufficient coverage of both halves of the gene. If the change in abundance of the strains is due to disabling the protein, then insertions at different locations in the gene should have similar fitness values, while if the change in strain abundance is due to some other

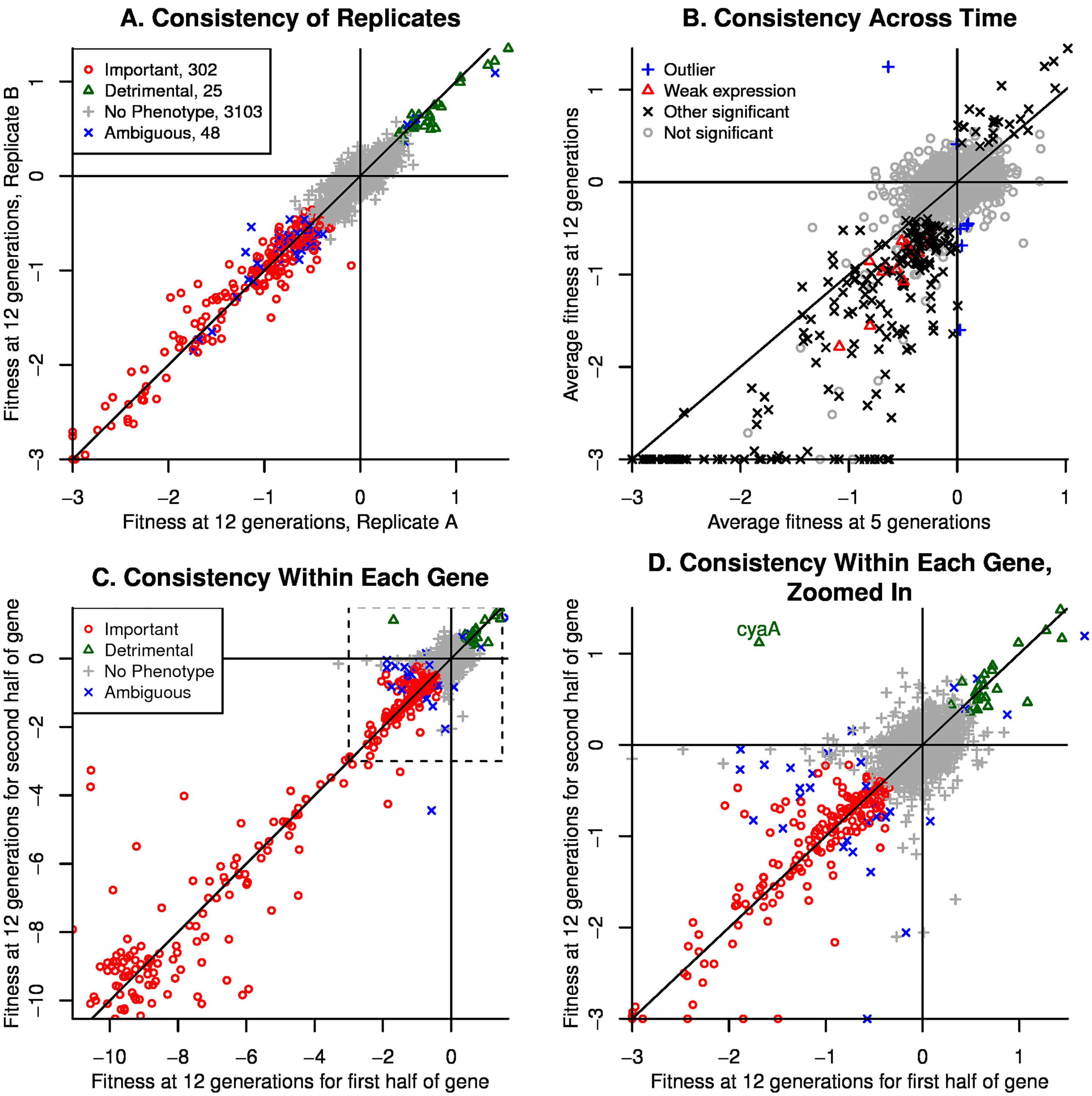

Fig 3. Consistency of gene fitness values in minimal glucose medium. (A) Consistency between replicates at 12 generations. Fitness values less than −3 are shown at −3 so as to focus on the more subtle fitness defects that are more susceptible to noise. 123 genes have fitness under −3 in both replicates. (B) Consistency across time. Genes with significant phenotypes (of either sign) are subdivided into those with weak expression (under 2 ppm of monomers) or above. Fitness values less than −3 are shown at −3. (C) Consistency within each gene. Fitness values at 12 generations were computed separately for the first and second half of each gene that had sufficient coverage. (D) shows the same data as (C), but only for fitness values above −3. In all panels, lines show $x = 0$, $y = 0$, and $x = y$.

doi:10.1371/journal.pone.0164314.g003

factor such as secondary mutations, then the insertions at different locations would have uncorrelated fitness values. The fitness values from the two halves were strongly correlated (r = 0.96), and just one gene had a significant phenotype but had inconsistent signs for the two halves (the adenylate cyclase *cyaA*, see Fig 3D).

A caveat that our analysis did not take into account is polar effects, in which a transposon insertion in one gene leads to reduced expression of downstream genes via rho-dependent termination. Although we do not believe polar effects are common (see [18] for a systematic analysis), they might lead us to underestimate the proportion of "useless" protein.

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 [18]. The *t* value is given by the gene fitness divided by the square root of an estimate of the variance. For most genes, which have many strains, the estimate of the variance is roughly the variance of the strain fitness values. But because estimating the variance from just a few data points is not very reliable, the *t* value also incorporates an estimate of the variance that is based on the number of reads for the gene [18]. We also add a small term to the estimate of the variance to account for other sources of noise that might be shared by all of the strains, such as imperfections in the normalization; this term corresponds to a shared noise in the fitness values of ±0.1.

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 Variance($A - C + B - C$) = Variance(A) + Variance(B) + 4 · Variance(C), while with independent start samples, it is Variance($A - C + B - D$) = Variance(A) + Variance(B) + Variance(C) + Variance(D). (Remember that Variance($A + B$) = 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 Fig 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.

Most genes whose deletion strains have strong growth defects when grown individually in glucose medium also have a significant fitness defect in our assay. Baba and colleagues [19] grew each mutant from the Keio deletion collection in a minimal MOPS glucose media and measured the optical density at 24 and 48 hours. (Although their media was similar to ours, it contained 2 mM instead of 1.32 mM phosphate.) We focused on the measurement at 24 hours as it should be more sensitive to a reduction in the growth rate, and we considered a 2-fold reduction in OD relative to the median as a strong defect. Of the genes that were included in the comparison to the ribosomal profiling data, 140 had such a strong reduction in OD, and in the fitness data, we classified 116 (83%) of these as significantly important for fitness, 13 as ambiguous, 1 as detrimental (*rseA*), and 10 as not important for fitness. *RseA* encodes an anti-sigma factor of σ^E and is not important for fitness in our other defined-media experiments either [18]; the cause of the discrepancy is not clear. The other 10 discordant genes included several genes involved in molybdenum cofactor synthesis (*moeA*, *moeB*, *moaC*) or selenocysteine synthesis (*selD*), but these processes are not expected to be important for growth under aerobic conditions. The discrepancy probably indicates a low concentration of oxygen in the experiment of Baba and colleagues, which was conducted in 96-well microplates without shaking. The other discordant genes were *ubiC*, which is involved in ubiquinone synthesis but is not required for growth on glucose [46]; the DNA repair enzyme *uvrD*; *exoX*, which encodes a genetically redundant DNA repair enzyme [47], so it is not clear why it would be important for fitness in these growth conditions; *metL*, which is expected to be redundant with *thrA* in these conditions; and the poorly characterized genes *hflD* and *ycaL*.

Processing of ribosomal profiling data

For *E. coli* growing in minimal glucose media, we obtained estimates of the abundance of each protein from the supplementary material of [11]. We defined the monomer fraction of a protein as its estimated abundance divided by the total abundance of all proteins. We defined the weight fraction as the estimated abundance times the number of amino acids in the protein, divided by the total of this product for all proteins.

Proteomics confirms the high expression of “unnecessary” proteins

To check that the high expression of genes with no measurable benefit is genuine, we considered proteomic estimates of protein abundance [15]. Unfortunately, the proteomic data with the highest coverage is from cells that were grown in a different media formulation than we used (M9 media with trace elements added instead of MOPS media, and with a higher concentration of glucose). Nevertheless, among the 4,062 proteins that were included in our analysis, 1,909 were quantified by proteomics, and for these proteins, the proteomics data was quite correlated with the ribosomal profiling data. For example, the Spearman rank correlation of the weight fraction, according to the two different methods, was 0.84, and the Pearson linear correlation of the monomer fractions was 0.82. The proportion of expression (by weight) that was due to proteins with no measurable phenotype was 23% according to proteomics, as compared to 29% for these same proteins according to ribosomal profiling. (29% is slightly less than 31%, which was reported above, because of the aggregate expression of the proteins that were not quantified by proteomics.)

We also checked if the 287 highly-expressed “unnecessary” proteins (with a weight fraction of above 200 ppm) were also highly expressed according to the proteomics data. 270 of the 287 proteins (94%) were quantified by proteomics, and of these, 219 (81%) had a weight fraction of above 100 ppm in the proteomics data. Overall, the proteomics data confirmed the high expression of most of these proteins.

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 [48]. 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. Indeed, the proteomics data [15], which gave similar results as the ribosomal profiling data, was obtained using BW25113. 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% [49].

The media formulations for the fitness and ribosomal profiling 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. The proteomics data used a different media and a smaller volume: it was collected using M9 media with 5 g/L glucose and added trace minerals, and with a 50 ml culture in a 500 ml flask shaking at 300 rpm [15].

A fitness compendium for Escherichia coli K-12

This compendium includes previously-described fitness experiments with various carbon sources and M9 minimal media [18]. 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 [18] 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 · 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 $t < -4$.

Data for budding yeast

Ribosomal profiling data for *S. cerevisiae* growing in YPD media [30] was obtained from accession GSM346111 at the Gene Expression Omnibus. The monomer fraction of each protein was estimated as the normalized abundance divided by the total normalized abundance. (The normalization was performed by [30].) Similarly, the weight fraction was estimated as the abundance times the protein's length, divided by the total of this product.

Genes that are essential or important for fitness in YPD media were obtained from a study of barcoded deletion strains [31]. Specifically, we downloaded their results from http://chemogenomics.stanford.edu/supplements/01yfh/files/OrfGeneData.txt and classified genes by their homozygous mutant phenotypes. Proteins that are represented in the fitness data account for 83% of protein expression by monomer or 92% by weight. To estimate the fraction of "unnecessary" expression, we conservatively assumed that the other proteins and translated regions were also important for fitness.

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.

Appendix 1: Comments on the functions of highly-expressed and "unnecessary" proteins

70 of the 287 proteins are involved in key processes that we expected would be important for growth in minimal glucose media. For 60 of them, we can explain the lack of phenotype due to redundancy. The redundancy may be due to isozymes, but 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%) [50, 51]. 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 [52]. 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 [53].

Appendix 2: Comparison of mutant fitness to metabolic model of O'Brien et al

We noted some disagreements between our mutant fitness data and the theoretical predictions of [17]. First, 264 of the proteins that we identified as essential or important for fitness are not included in their models or are not predicted to be important in any carbon source. (Their supplementary information does not include predictions for glucose specifically, so we used their combined predictions across 180 carbon sources instead.) According to the ribosomal profiling data, these 264 proteins account for 13% of protein expression.

We also examined the 469 "core" genes that are predicted to be utilized in all defined media. The model implies that these core genes should be important for fitness (or perhaps genetically redundant), as they carry flux during optimal growth. However, we found that six of the core genes were detrimental to fitness in glucose minimal media: *fabF*, *greA*, *rplI*, *rpoS*, *rsmA*, and *truB*. We checked the per-strain data and confirmed that there was a consistent increase in abundance for mutants in these six genes after 12 generations of growth in glucose media. Also, this increase in abundance was observed for insertions in both strand orientations, which suggests that the phenotypes are due to the disruption of the gene rather than alterations in the expression of nearby genes (i.e., polar effects). To check the theoretical plausibility of our experimental observations, we examined the gene summaries in EcoCyc [22]. FabF or β-ketoacyl-ACP synthetase II is one of three isozymes and is reported to be particularly important for synthesis of unsaturated fatty acids. Unsaturated fats increase the fluidity of the membrane and are important for adaptation to low temperature, but they might be deleterious at 37°C. RplI or L9 is a subunit of the ribosome, but it is not essential, and mutants do not have a strong phenotype in standard growth conditions. RpoS is a major alternate sigma factor and many studies have reported that it is detrimental to fitness in some conditions. RsmA (also known as KsgA because mutants are resistant to kasugamycin) and TruB are involved in the modification of rRNA or tRNA, but they are not essential, and mutants do not have a strong phenotype in standard growth conditions. Overall, our observation that FabF and RpoS are detrimental for fitness is not surprising, and while our observations for the other four genes are not supported by the literature, neither are the model's predictions that these genes should be required for optimal growth in minimal glucose media.

Appendix 3: Expression of highly-expressed and "unnecessary" genes after laboratory evolution

O'Brien and colleagues reported that investment in "utilized" proteins (which corresponds roughly to our essential or important proteins) often increased after laboratory evolution of *E. coli* in glucose minimal media [17]. This implies that investment in unimportant genes went down. To confirm this, we estimated the proportion of total mRNA expression for each gene as RPKM (reads per kilobase per million) [44] times the length in nucleotides. (To compute the total mRNA expression, we included only the protein-coding genes.) Total expression of important or essential genes was 49–51% of mRNA in the two wild type samples. In 7/8 evolved lines, this proportion increased, to 54–61%. In the other evolved line (strain 8), 48% of mRNA was for important proteins, or slightly less than in wild type strains. Strain 8 was also an outlier in the analysis of O'Brien and colleagues. Not surprisingly, in the seven evolved lines with increased expression of important proteins, the total expression of all "unnecessary" proteins was reduced, from 36–38% in wild type to 23–31%. If we focus on just the 106 highly-expressed proteins that we believe are on standby, then their expression was 11–13% of mRNA in the wild type samples; this decreased in the seven lines, to 6–8%.

Although the mRNA expression of the "standby" proteins went down, we do not believe that there was selection specifically to reduce the expression of these genes. In particular, the standby genes do not seem to be downregulated more than other non-important proteins. If we compute a $\log_2$ fold change for each protein, and we normalize these so that the median protein has a value of zero, then the 106 standby genes were, on average, downregulated relative to the median gene in just 2 of 8 lines ($P < 0.05$, Wilcoxon rank sum test; median $\log_2$ fold changes were −0.40 or −0.26 in these lines). And the 106 genes were significantly upregulated in 1 of the 8 lines ($P < 0.05$; median $\log_2$ change of +0.22 in strain 7B). In other words, the highly-expressed "standby" genes were down-regulated about as much as other unnecessary genes. This suggests the importance of more global regulatory changes rather than mutations that affect the expression of individual genes. Indeed, two of the three most common mutations in these evolved lines affected global regulators: these were misssense mutations in *rpoB* and an insertion element that increased the expression of *hns* [44].

Supporting Information

S1 Fig. Protein abundance versus gene fitness in *E. coli*.
(PDF)

S2 Fig. Locations of transposon insertions in and around *yceQ*.
(PDF)

S3 Fig. Normalization of gene fitness values by location on the chromosome.
(PDF)

S1 Table. Comparison of fitness data and ribosomal profiling data.
(XLS)

S2 Table. Manual classification of highly expressed genes with no phenotype in minimal glucose.
(XLS)

Acknowledgments

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

Author Contributions

Conceptualization: MNP AMD APA.

Formal analysis: MNP.

Funding acquisition: AMD APA.

Investigation: KMW.

Resources: AMD.

Supervision: APA.

Writing – original draft: MNP.

Writing – review & editing: MNP AMD APA.

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. doi: 10.1038/ng1555 PMID: 15880104
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. doi: 10.1016/j.celrep.2012.09.017 PMID: 23103169
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. doi: 10.1371/journal.pgen.1003617 PMID: 23874220
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). doi: 10.1038/msb.2013.16 PMID: 23591776
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. doi: 10.1128/JB.01836-14 PMID: 25112473
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. doi: 10.1073/pnas.1416533112 PMID: 25695966
7. Price MN, Arkin AP. A Theoretical Lower Bound for Selection on the Expression Levels of Proteins. Genome Biology and Evolution. 2016; p. evw126. doi: 10.1093/gbe/evw126 PMID: 27289091
8. 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. doi: 10.1126/science.1192588 PMID: 21097934

9. 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; 38(5):758–767. doi: 10.1016/j.molcel.2010.04.015 PMID: 20434381

10. Tomala K, Korona R. Evaluating the Fitness Cost of Protein Expression in Saccharomyces cerevisiae. Genome biology and evolution. 2013; 5(11):2051–2060. doi: 10.1093/gbe/evt154 PMID: 24128940

11. 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. doi: 10.1016/j.cell.2014.02.033 PMID: 24766808

12. 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. doi: 10.1128/mBio.01302-15 PMID: 26670382

13. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics. 2009; 10(3):195–205. doi: 10.1038/nrg2526 PMID: 19204717

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. doi: 10.15252/msb.20145697 PMID: 25678603

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. doi: 10.1038/nbt.3418 PMID: 26641532

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. doi: 10.1016/j.ymben.2015.10.003 PMID: 26498510

17. O'Brien EJ, Utrilla J, Palsson BO. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. PLoS Comput Biol. 2016; 12(6):e1004998. doi: 10.1371/journal.pcbi.1004998 PMID: 27351952

18. 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. doi: 10.1128/mBio.00306-15 PMID: 25968644

19. 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. doi: 10.1038/msb4100050 PMID: 16738554

20. Kato Ji, Hashimoto M. Construction of consecutive deletions of the Escherichia coli chromosome. Molecular systems biology. 2007; 3(1). doi: 10.1038/msb4100174 PMID: 17700540

21. 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. doi: 10.1073/pnas.032488499 PMID: 11805295

22. 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. doi: 10.1093/nar/gks1027 PMID: 23143106

23. 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. doi: 10.1128/JB.183.24.7173-7181.2001 PMID: 11717276

24. 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. PMID: 3011417

25. 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. doi: 10.1099/mic.0.27481-0 PMID: 15758216

26. Sauer U, Canonaco F, Heri S, Perrenoud A, Fischer E. The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli. Journal of Biological Chemistry. 2004; 279(8):6613–6619. doi: 10.1074/jbc.M311657200 PMID: 14660605

27. Waegeman H, Beauprez J, Moens H, Maertens J, De Mey M, Foulquié-Moreno MR, et al. Effect of iclR and arcA knockouts on biomass formation and metabolic fluxes in Escherichia coli K12 and its implications on understanding the metabolism of Escherichia coli BL21 (DE3). BMC microbiology. 2011; 11(1):1. doi: 10.1186/1471-2180-11-70 PMID: 21481254

28. 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. doi: 10.1093/nar/gkv1156 PMID: 26527724

29. Tsui H, Feng G, Winkler ME. Transcription of the mutL repair, miaA tRNA modification, hfq pleiotropic regulator, and hflA region protease genes of Escherichia coli K-12 from clustered Eσ^{32}—specific promoters during heat shock. Journal of bacteriology. 1996; 178(19):5719–5731. PMID: 8824618

30. 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. doi: 10.1126/science.1168978 PMID: 19213877

31. 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. doi: 10.1534/genetics.104.036871 PMID: 15716499

32. Suzuki H, Koyanagi T, Izuka S, Onishi A, Kumagai H. The yliA, −B, −C, and −D genes of Escherichia coli K-12 encode a novel glutathione importer with an ATP-binding cassette. Journal of bacteriology. 2005; 187(17):5861–5867. doi: 10.1128/JB.187.17.5861-5867.2005 PMID: 16109926

33. Owens RA, Hartman PE. Export of glutathione by some widely used Salmonella typhimurium and Escherichia coli strains. Journal of bacteriology. 1986; 168(1):109–114. PMID: 3531162

34. 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. doi: 10.1371/journal.pone.0061686 PMID: 23637881

35. 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. doi: 10.1371/journal.pbio.1002041 PMID: 25626068

36. 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. doi: 10.1371/journal.pbio.1002042 PMID: 25626086

37. 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. doi: 10.1126/science.1126439 PMID: 16645050

38. Wagner A. Risk management in biological evolution. J Theor Biol. 2003; 225:45–57. doi: 10.1016/S0022-5193(03)00219-4 PMID: 14559058

39. 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. doi: 10.1093/molbev/msu131 PMID: 24723420

40. Good BH, Desai MM. Deleterious passengers in adapting populations. Genetics. 2014; 198:1183–208. doi: 10.1534/genetics.114.170233 PMID: 25194161

41. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, et al. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 2009; 461(7268):1243–1247. doi: 10.1038/nature08480 PMID: 19838166

42. Leiby N, Marx CJ. Metabolic erosion primarily through mutation accumulation, and not tradeoffs, drives limited evolution of substrate specificity in Escherichia coli. PLoS Biol. 2014; 12(2):e1001789. doi: 10.1371/journal.pbio.1001789 PMID: 24558347

43. Cooper VS, Schneider D, Blot M, Lenski RE. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of Escherichia coli B. Journal of Bacteriology. 2001; 183(9):2834–2841. doi: 10.1128/JB.183.9.2834-2841.2001 PMID: 11292803

44. LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, et al. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Applied and environmental microbiology. 2015; 81(1):17–30. doi: 10.1128/AEM.02246-14 PMID: 25304508

45. 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. doi: 10.1101/gr.093955.109 PMID: 19622793

46. Lawrence J, Cox G, Gibson F. Biosynthesis of ubiquinone in Escherichia coli K-12: biochemical and genetic characterization of a mutant unable to convert chorismate into 4-hydroxybenzoate. Journal of bacteriology. 1974; 118(1):41–45. PMID: 4595202

47. Burdett V, Baitinger C, Viswanathan M, Lovett ST, Modrich P. In vivo requirement for RecJ, ExoVII, ExoI, and ExoX in methyl-directed mismatch repair. Proceedings of the National Academy of Sciences. 2001; 98(12):6765–6770. doi: 10.1073/pnas.121183298 PMID: 11381137

48. Grenier F, Matteau D, Baby V, Rodrigue S. Complete genome sequence of Escherichia coli BW25113. Genome announcements. 2014; 2(5):e01038–14. doi: 10.1128/genomeA.01038-14 PMID: 25323716

49. 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. PMID: 8501045

50. 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. doi: 10.1016/0378-1119(93)90598-W PMID: 8472951

51. 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. PMID: 159306

52. 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. doi: 10.1111/j.1365-2958.2008.06190.x PMID: 18394147

53. 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. doi: 10.1073/pnas.0903229106 PMID: 19383799