Title: Operon Formation is Driven by Co-Regulation and Not by Horizontal Gene Transfer

Authors: Morgan N. Price, Katherine H. Huang, Eric Alm, and Adam P. Arkin

Author affiliation: Lawrence Berkeley Lab, Berkeley CA, USA. A.P.A. is also affiliated with the Howard Hughes Medical Institute and the UC Berkeley Dept. of Bioengineering.

Corresponding author: Eric Alm, ejalm@lbl.gov, phone 510-843-1794, fax 510-486-6059, address Lawrence Berkeley National Lab, 1 Cyclotron Road, Mailstop 939R704, Berkeley, CA 94720

Abstract:

Although operons are often subject to horizontal gene transfer (HGT), non-HGT genes are particularly likely to be in operons. To resolve this apparent discrepancy and to determine whether HGT is involved in operon formation, we examined the evolutionary history of the genes and operons in *Escherichia coli* K12. We show that genes that have homologs in distantly related bacteria but not in close relatives of *E. coli* – indicating HGT – form new operons at about the same rates as native genes. Furthermore, genes in new operons are no more likely than other genes to have phylogenetic trees that are inconsistent with the species tree. In contrast, essential genes and ubiquitous genes without paralogs – genes believed to undergo HGT rarely – often form new operons. We conclude that HGT is not associated with operon formation, but instead promotes the prevalence of pre-existing operons. To explain operon formation, we propose that new operons reduce the amount of regulatory information required to specify optimal expression patterns. Consistent with this hypothesis, operons have greater amounts of conserved regulatory sequences than do individually transcribed genes.

Introduction

Bacterial genes are often transcribed together in operons, so that several genes are under the control of a single promoter. Although the study of operons has traditionally focused on gene regulation, from comparing complete genome sequences it has become clear that operons are often associated with horizontal gene transfer (HGT) (Lawrence and Ochman 1998; Omelchenko *et al.* 2003). According to the popular "selfish operon" theory, operons evolve so that organisms can acquire several functionally related genes – genes that together provide a useful capability – with a single transfer event (Lawrence and Roth 1996; Lawrence 1999). More specifically, capabilities that are only occasionally useful are often lost by random deletion of one gene. This is followed by the loss of the other genes in the same pathway, as they are now useless. Such pathways can then be regained by HGT of an entire operon. This theory is appealing for several reasons: it explains why operons are often transferred, it provides a biologically plausible mechanism for operon formation, and it makes testable predictions.

The selfish operon theory has been called into question by the finding that essential genes in *Escherichia coli* are particularly likely to be in operons (Pal and Hurst 2004; de Daruvar *et al.* 2002). As essential genes, by definition, cannot be lost, there is no need for them to be regained, so the selfish operon theory cannot explain these operons. Furthermore, many essential genes are conserved single-copy genes, and such genes rarely undergo HGT (Lerat *et al.* 2003). Thus, there is a

discrepancy: operons are associated with HGT, but essential non-HGT genes are more likely to be in operons.

To explain this discrepancy and to clarify the relationship between HGT and operons, we needed to distinguish between the transfer of existing operons, which appears to be common, and the invention of new operons, which may or may not be associated with HGT. To do this, we investigated the evolutionary history of the operons in *E. coli K12*, and compared them to the histories of the genes in those operons. As we will show, HGT is not associated with the formation of new operons: HGT genes and native genes form new operons at similar rates, and ubiquitous and essential genes – genes believed not to be subject to HGT – often form new operons.

Having shown that HGT does not explain the creation of new operons, we asked whether co-regulation might. Given several genes whose optimal expression pattern is the same, a bacterium can either evolve several independent promoters to the optimal pattern, or it can evolve one optimal promoter and place the genes in an operon. As the amount of regulatory sequence required to specify the optimal expression pattern increases, evolving the optimal expression profile separately for each gene should become more difficult, while creating an operon should not. Thus, the co-regulation theory predicts that operons will have more complex upstream regulatory sequences than individually transcribed genes. We will present evidence from comparative genomics that this is indeed the case.

Results

HGT Genes Are Not Particularly Likely to be in Operons

To test the relationship between HGT and operons, we first needed to identify horizontally transferred genes. We used a presence/absence approach (Ragan and Charlebois 2002) together with a simplified phylogeny of *E. coli* K12 and its relatives, as described by Daubin and Ochman (2004). As shown in Figure 1, we examined which genomes contained potential orthologs for each *E. coli* K12 gene. We refer to each group of genomes at a similar phylogenetic distance from *E. coli* K12 as an outgroup. If a gene had potential orthologs in every outgroup going back to the Proteobacterial outgroup, we classified the gene as "native." If a gene lacked homologs in two or more consecutive outgroups, and then contained a homolog in more distantly related bacteria, we classified the gene as "HGT." Although it is possible that such genes were propagated to *E. coli* K12 by vertical descent from a common ancestor and were then lost twice or more independently, the more parsimonious explanation is that such genes were transferred. To allow us to distinguish paralogs from orthologs and to detect distant homologs, we further required such genes to be present in a database of conserved orthologous groups (COGs, Tatusov *et al.* 2001). Finally, we also required such genes to be present in every outgroup after the putative transfer event. This allowed us to use the outgroup into which the putative transfer event occurred as a measure of the gene's "age," or how long ago the gene came into the *E. coli* K12 lineage. If a gene was neither native nor HGT and lacked homologs outside of the Proteobacteria, we classified the gene as an "ORFan." These are (relatively) new

genes that are believed to be transferred from phage rather than from other bacteria (Daubin and Ochman 2004). Some genes did not fit any of these categories and were excluded from analysis. As recommended by Daubin and Ochman (2004), we also excluded prophages and transposons. We further subdivided the native genes into a "ubiquitous single-copy" set of genes that are present and do not have paralogs in each of 13 diverse γ-Proteobacteria (the genes analyzed by Lerat *et al.* 2003), and classified the remaining native genes as "typical." The 200 ubiquitous single-copy genes rarely undergo horizontal transfer (Lerat *et al.* 2003) – we excluded the two known exceptions to this rule (*bioB* and *mviN*) so that we could treat the ubiquitous single-copy genes as a non-HGT set.

Using this presence/absence approach, we found that HGT genes are about as likely as typical native genes to be in predicted operons (Figure 2). In contrast, ORFans were much less likely to be in predicted operons. As both HGT and ORFan genes tend to be AT-rich (Daubin and Ochman 2004), compositional approaches to studying HGT (e.g., Lawrence and Ochman 1998) would have difficulty distinguishing the two kinds of genes. The differing tendencies of these genes to be in operons extends previous observations of major differences between these two classes of genes (Daubin and Ochman 2004) and validates the use of presence/absence to study the relationship between HGT and operons. We also found that single-copy ubiquitous genes were particularly likely to be in operons. As many of the single-copy ubiquitous genes are essential, this is consistent with a previous report that most essential genes are in operons (Pal and Hurst 2004). To ensure that errors in operon predictions were not biasing our estimates of how often different types of genes were in operons, we also asked how often the different types of genes were adjacent to genes on the same strand: because all operon pairs are same-strand pairs, the frequency of same-strand pairs is a reliable indicator of the number of operons (Ermolaeva *et al.* 2001; Cherry 2003). The analysis of same-strand pairs confirmed that many HGT genes are in operons, but that HGT genes are not particularly likely to be in operons (Supplementary Figure 1).

HGT Genes Are Not Particularly Likely to be in New Operons

To identify new operons that were invented in the *E. coli* K12 lineage and also operons that were imported into the *E. coli* lineage from other bacteria, we applied the presence/absence method to the history of the operons (Figure 1). Although operons are often rearranged during evolution (Itoh *et al.* 1999), we focused on the creation of new operons – the placement of two separately transcribed genes into the same transcription unit – and ignored rearrangements of existing operons. Specifically, we examined pairs of adjacent *E. coli* K12 genes that were predicted to be in the same operon, and for each pair, we recorded which genomes contained homologs that were in the same predicted operon. Operon pairs that were missing from two consecutive outgroups and then present in a more distantly related bacterium were classified as "imported." Otherwise, operon pairs that were present in the non-Proteobacterial outgroup were classified as "ancestral," and newer operon pairs were classified as "new." HGT genes were far more likely to be in imported operons than were typical native genes (Table 1). Because the histories of the genes and the operons were arrived at independently, this result validated our method. As almost half of all HGT genes were in imported operons, this analysis confirmed that HGT often involves pre-existing operons (Lawrence and Ochman 1998; Omelchenko

et al. 2003).

We then asked whether HGT genes formed new operons more rapidly than other genes. As shown in Table 1, HGT genes were about as likely to be in new operons as typical native genes. However, given that HGT genes have been in the *E. coli* lineage for less evolutionary time than the native genes, HGT genes might be forming operons at high rates in the short time available. To account for this, we analyzed how often HGT genes formed new operons at the time of transfer and also how often they formed new operons after transfer.

To determine whether HGT genes formed new operons at the time of transfer, we asked if the age of each HGT gene matched the age of a new operon pair containing that gene. Only 9% of HGT genes were in new operons of the same age as the gene (see Table 1). In contrast, of the 165 HGT genes that were in imported operons, 90% were in operons of the same age as the gene. Thus, the modest rate of operon formation at the time of HGT was not due to errors in the ages. Furthermore, the selfish operon theory predicts that HGT genes would form operons with other HGT genes, as repeated HGT of both genes is required to drive them together. When an HGT gene did form an operon pair at the time of transfer, the other gene in the pair was not particularly likely to be an HGT gene: 3 out of 35, or 8.6%, were HGT genes, whereas 8.9% of the genes in all classified operon pairs were HGT by our stringent criteria. We concluded that HGT genes formed operons at the time of transfer at modest rates and without a strong preference for other HGT genes.

To determine whether HGT genes formed new operons at high rates after being transferred into the *E. coli* lineage, we restricted our analysis to 138 older HGT genes – those imported before the divergence of *E. coli* and *Salmonella* from other Enterobacteria. To perform a fair comparison on the number of new operons for these genes and for native genes, we considered only the new operons that formed after this divergence. As shown in Table 1, older HGT genes were no more likely to be in the newest operons than were typical native genes. Thus, HGT genes do not form new operons at elevated rates, either at the time of transfer or after transfer.

The presence/absence analysis identified transfer events between distant organisms, but may have missed transfer events between close relatives. To see if transfer events within the *E. coli* lineage were correlated with new operons, regardless of how the genes came into the lineage, we compared gene trees from protein sequence alignments to a fully resolved species tree of 13 γ-Proteobacteria given by Lerat *et al.* (2003). To reduce problems due to paralogs, we used only COGs present as a single copy in *E. coli* K12. Similarly, we only included a homolog in a tree if that homolog was the only copy of the COG in its genome. Based on these criteria we built 1,128 alignments and gene trees (see Methods). To determine whether to accept the hypothesis that the phylogeny of the gene matches the phylogeny of the species, for each tree we performed a one-sided Kishino-Hasegawa test with a cutoff of $p > 0.05$ (Goldman *et al.* 2000). As shown in Table 2, most genes in new operons had trees that were consistent with the species tree. Furthermore, the rates of discordant trees were no higher for genes in new operons than for other genes – instead there was a modest and statistically insignificant effect in the opposite direction. The proportion of genes identified as HGT by this test might be biased by the number of homologs available: trees that contained more homologs were more likely to reject the species tree (not shown). However, this cannot explain why most genes in new operons accepted the species tree, as genes in new operons tended to have more

homologs (an average of 8.0 homologs for genes in new operons versus 7.5 for other genes, $p = 0.01$, t test). Thus, phylogenetic trees confirmed that genes in new operons are no more likely than other genes to be horizontally transferred.

Non-HGT Genes Often Form New Operons

In defense of the selfish operon theory, which predicts that essential genes should not be in operons, it has been suggested that the operons containing essential genes are ancient, and that these ancient operons were formed by distinct mechanisms from newer operons (Lawrence and Roth 1996). We asked whether essential genes formed new operons. Of the 521 essential genes identified in *E. coli* (Gerdes *et al.* 2003) and successfully classified by our method, 398 were native genes (including 117 of the 200 ubiquitous native genes), 74 were ORFans, and 49 were HGT. Of the native essential genes, 32% were in new operons that formed after the divergence of the $\beta\gamma$-Proteobacteria. To validate this finding that essential genes often form new operons, we focused on 58 new operon pairs involving the 200 ubiquitous single-copy genes: the majority of these operon pairs contain essential genes, and the ubiquitous single-copy genes have been shown not to undergo HGT, at least not since the divergence of the γ-Proteobacteria (Lerat *et al.* 2003). As shown in Table 3, 38 of these predicted new operon pairs were previously identified experimentally (Karp *et al.* 2002) or show strong similarity of expression patterns in microarray data (Gollub *et al.* 2003). As the operon predictions were based only on sequence, the microarray data provides an independent confirmation of the predictions (Sabatti *et al.* 2002). Thus, essential and other ubiquitous genes form new operons at significant rates, and the concept of ancient operons is not sufficient to explain why these genes are in operons.

Operons Save Information if Regulation is Complex

As an alternative to the theory that HGT promotes the formation of selfish operons, we considered co-regulation. More specifically, we considered that an operon might reduce the amount of information required to specify the expression patterns of several genes. To see if operons can in fact save information, we compared the information required to place two genes in an operon to the information required to specify one or more transcription factor (TF) binding sites. In both cases, information can be quantified as how unlikely it would be for the operon or binding site(s) to arise by chance, as measured in log-base-two units, or bits. For example, to form a two-gene operon, the first gene can be placed anywhere (0 bits), and the second gene must be placed downstream of it. If a typical genome has 4,000 genes, then $\log_2(4,000) \approx 12$ bits of information are required to specify this placement, and one more bit is required to place the second gene on the same strand as the first gene, for a total of 13 bits. However, for the operon to function correctly there may be restrictions on spacing between the genes, either to avoid polarity or to maintain the translation control sequences of both genes. If a moderately accurate spacing of within 100 base pairs is required, and the genome is 4 megabases in size, then the information required to place the genes near each other rises to $\log_2(4 \cdot 10^6/100) \approx 15$ bits, or 16 bits to form the two-gene operon. To form a larger operon

requires an additional 16 bits for each additional gene.

For comparison, characterized binding sites for TFs generally contain 12-20 bits of information: 50% of the *E. coli* TFs with known binding sites in DPInteract (Robison *et al.* 1998), analyzed for information content by MEME (Bailey and Elkan 1995), are within this range. For example, binding sites for CRP are estimated to have 13.2 bits of information, so that CRP will bind at one in every $2^{13.2} = 9,410$ positions in a random sequence. This analysis measures the information required to specify binding at a precise location, but if binding at any of n different positions sufficed to specify the optimal expression pattern, then the information required to specify the site would be lowered by $\log_2(n)$. For example, the binding site for a transcriptional repressor often overlaps the -10 site of RNA polymerase binding, but the precise location may not be important ($n \approx 16$, or 4 bits of flexibility). There can also be constraints in the spacing between TFs – a study of spacings between predicted TF binding sites in *E. coli* (Bulyk *et al.* 2004) found a statistical excess of certain spacings between many different pairs of TFs at several scales, including precise positioning ($n = 1$, or 0 bits of flexibility) and constrained positioning ($n \approx 30$, or about 5 bits of flexibility). At the other extreme, any position near the promoter might suffice to affect transcriptional activity ($n \approx 200$, or 7-8 bits of flexibility). Thus, a range of 0-8 bits of flexibility is plausible, and the amount of flexibility probably depends on the biological context. With flexibility factored into consideration, it might require as little as $12 - 8 = 4$ bits to specify a TF binding site, but we doubt that this would be biologically useful. For example, five bits information can only specify a binding site for a global regulator that binds to one in every $2^5 = 32$ genes, or over 100 genes in total, without any positional constraints. Overall, we argue that 6-20 bits are usually required to specify a new TF binding site.

Although the information required to specify an operon (13-16 bits) is generally greater than the information required to specify a TF binding site (6-20 bits), many genes are regulated by several TFs, each with their own binding sites (Robison *et al.* 1998). The information required to specify several TF binding sites will often be much greater than the information required to specify an operon. For example, specifying three TF binding sites of moderate complexity and positional constraint requires $3 \cdot 12 = 36$ bits, while moving the gene into an operon downstream of another gene which already has the desired regulatory information requires only 16 bits. Thus, operons give a large savings in regulatory information if the regulation is complex.

Because operons save regulatory information, it should be easier to optimize the coordinated expression of several genes to a situation of complex regulation by placing them in an operon instead of by evolving new TF binding sites independently for each gene. More precisely, given enough evolutionary time, either method of achieving the desired expression pattern would evolve, but the operon should evolve more rapidly. Comparing the amount of information may overstate the relative ease of evolving an operon, as a weak binding site can gradually evolve into a strong site by several single-base mutations, with each individual mutation being fixed by selection. In contrast, the benefit of placing two genes in an operon (or placing a gene downstream of an existing operon) is all-or-nothing. Nevertheless, the dramatic difference in information content suggests that evolving a complex regulation pattern by forming an operon would be easier than by forming several new TF binding sites. Thus, the co-regulation theory predicts that operons will have more complex upstream regulatory sequences than individually transcribed genes.

Operons Have More Conserved Regulatory Sequences

To test this prediction, we examined regulatory sequences that were identified by comparative genomics. The conservation of upstream sequences over hundreds of millions of years of evolution is strong evidence that they are functional, and these "phylogenetic footprints" often correspond to experimentally identified TF binding sites (Terai *et al.* 2001; McCue *et al.* 2002). Specifically, we counted the number of base pairs of conserved sequences found upstream of genes in a genome-wide phylogenetic footprinting analysis of *E. coli* K12 (McCue *et al.* 2002). Because genes that are insufficiently conserved cannot have footprints, regardless of how much regulatory information they contain, we considered only genes with at least one footprinted site. The data set contained 6,595 footprinted sites upstream of 2,047 genes with an average site length of 20.4 base pairs. As shown in Figure 3, genes with larger amounts of conserved regulatory sequence were more likely to be genes at the start of predicted operons, rather than being genes transcribed individually. This relationship between footprinted base pairs and operons was statistically significant, even when considering only the typical native genes ($p < 10^{-4}$, Wilcoxon rank sum test). Analyzing the number of footprinted sites instead of the total base pairs yielded a similar result: the average typical gene at the head of an operon had 3.50 sites upstream, while typical genes transcribed individually averaged 3.26 sites upstream ($p < 10^{-4}$, t test).

To determine whether this correlation reflected a causal relationship, rather than correlation through an intermediate factor, we performed several controls. First, the protein sequences of genes in operons showed greater conservation between *E. coli* K12 and *Salmonella enterica Typhi* than other genes: the median %identity was 91.9% for genes in operons and 90.3% for genes not in operons ($p < 10^{-5}$, Wilcoxon rank sum test). This conservation reflects negative selection that could also operate on regulatory sequences, so that larger footprints would be found for these genes even if the amount of regulatory information were similar. We used the partial Spearman correlation to test if operons were significantly correlated with greater amounts of phylogenetic footprints after taking into account the correlation of both operons and phylogenetic footprints with the conservation of the gene (see Methods). We found that, after controlling for protein sequence conservation, the amount of footprinted sequence upstream of operons remained significantly greater than for other genes (partial Spearman correlation 0.09, $p < 0.001$).

Second, operons tend to have more sequence between them and the next gene upstream than do single-gene transcripts (the averages were 208 and 181 base pairs, respectively; $p < 10^{-4}$, t test). This could reflect the more complex regulatory sequences of operons, but it could also be due to an unknown cause. In the latter case, as intergenic sequences are the input to phylogenetic footprinting, operons might show more false positive footprints simply because there was more input. However, the greater footprint of operons remained significant after controlling for the size of the upstream region (partial Spearman correlation 0.10, $p < 10^{-4}$). Furthermore, a much smaller high-confidence subset of the footprinted sites, which contained 878 individual sites upstream of 581 genes with an average size of 20.5 base pairs, also showed a significant relationship between operons and the number of footprinted base pairs (Wilcoxon rank sum test, $p = 0.04$).

Third, we attempted to confirm the relationship between operons and regulatory sequences by counting the number of experimentally verified TF binding sites (Robison *et al.* 1998) upstream of

operons and other genes, but we did not see any statistically significant differences (not shown). Because verified sites may not be a uniform sample of all sites, this discrepancy could be an artefact. Nevertheless, we were concerned that the observed relationship between operons and phylogenetic footprints might not be due to TF binding sites and might instead reflect other types of conserved sites, such as Shine-Dalgarno sequences.

To resolve this question we examined phylogenetic footprints from *B. subtilis* that were generated with a somewhat different method and from which Shine-Dalgarno sequences were removed (Terai *et al.* 2001). We also examined known TF binding sites for this organism (Makita *et al.* 2004). Similar to our results for *E. coli*, we found that *B. subtilis* operons have significantly more conserved sites upstream than other genes (Supplementary Table 1; $p = 0.04$, t test), yet there was no significant relationship between operons and verified TF binding sites (not shown).

Because Terai *et al.* (2001) clustered the *B. subtilis* phylogenetic footprints into groups of similar sites which should have similar function, we could test whether these phylogenetic footprints predicted the gene's expression patterns. We found that these footprints were strong predictors of whether two genes would have similar expression patterns, and were better predictors than whether the two genes shared a verified TF binding site (Supplementary Figure 2). Thus, the *B. subtilis* phylogenetic footprints do consist largely of genuine regulatory sequences. We concluded that operons have larger amounts of conserved upstream regulatory sequences than other genes.

Discussion

We have shown that HGT is not associated with the formation of new operons:

- HGT genes formed new operons at similar rates as typical genes.
  - At the time of transfer, HGT genes formed new operons at modest rates, and the genes they formed operons with were not enriched in other HGT genes.
  - After transfer, HGT genes formed new operons at the same rate as typical native genes.
- Genes in new operons were no more likely than other genes to have phylogenetic trees that were significantly different from the species tree.
- Essential genes and ubiquitous single-copy genes, which are believed to undergo HGT rarely, formed many new operons.

One potential limitation of the presence/absence analysis is that our method can only discover HGT between relatively distant organisms. Transfer between closely related organisms might suffice to create operons, and such transfer events would not be detected. However, the phylogenetic trees should be able to detect transfers between close relatives, and the trees did not show any tendency for HGT genes to be in new operons. Moreover, given that essential genes are forming new operons,

and that the selfish theory requires genes to be lost and then regained to drive operon formation, it seems unlikely that HGT between closely related organisms is a major factor in operon formation.

Although HGT is not associated with operon formation, many HGT genes are in operons. This seems to reflect a high rate of transfer of operons: almost half of HGT genes were transferred into the *E. coli* K12 lineage as operons. The reason for HGT genes to be transferred at high rates in operons is presumably that originally given to justify the selfish operon theory – genes in operons tend to be functionally related, and transferring an entire operon allows an organism to acquire a useful new capability (Lawrence and Roth 1996; Lawrence 1999). As operons often "die" by being shuffled apart (Itoh *et al.* 1999), we infer that HGT extends the lifetime of individual operons, and that HGT increases the prevalence of operons in bacterial genomes, even though HGT does not contribute to operon formation.

As an alternative to the selfish theory of operon formation, we proposed that operons reduce the information required to specify optimal expression patterns for several co-regulated genes. This theory predicts that as the amount of regulatory information increases, genes should be more likely to be in operons. Indeed, we found that operons in both *E. coli* and *B. subtilis* tend to have more conserved regulatory sequences than other genes. This effect remained significant after controlling for the greater protein sequence conservation of genes in operons, which might plausibly correlate with stronger purifying selection on operonic regulatory sequences and hence larger footprints. Yet another explanation for why operons would have more conserved regulatory sequences is that the regulatory sequences of operons are under stronger purifying selection because they control the expression of more genes. We were not able to test this hypothesis, but most operons are small: the average size of our predicted operons is 3.1 genes. Because the selection pressure on regulatory sequences should depend on the total level of expression of the genes in the operon, and because the expression levels of individual genes varies by orders of magnitude, we doubt that the effect of operon size would be significant.

Although we were not able to confirm the relationship between operons and regulatory sequences with databases of verified TF binding sites, the phylogenetic footprints in *B. subtilis* were strong predictors of the expression patterns of the downstream genes. Thus, we do not believe that the difference in findings for phylogenetic footprints compared to that for verified sites was due to errors in the phylogenetic footprints. Biases in the databases of verified sites might be skewing the results. Alternatively, Terai *et al.* (2001) observed that a significant number of the *B. subtilis* phylogenetic footprints were attenuators – sequences that regulate gene expression by forming structures in the nascent mRNA instead of by binding to TFs as DNA. The tendency of operons to have more regulatory sequences may be particularly strong for attenuators. Because attenuators are larger and more complex than individual TF binding sites, it should be much more difficult to evolve a new attenuator from scratch than to evolve a new TF binding site, so a strong preference for attenuators to be located upstream of operons would be consistent with the co-regulation theory.

We are not aware of any previous work with direct evidence for the co-regulation theory, but the theory is consistent with the existence of conserved operons containing genes that are not functionally related (Rogozin *et al.* 2002). This "genomic hitchhiking" is believed to reflect both serendipity and the existence of genes with similar expression patterns – for example, perhaps both

genes are regulated by growth rate – even if they are in quite distinct pathways.

One attractive feature of the selfish operon theory was that it provided an intermediate state to operon formation: if two functionally related genes are near each other, they may be likely to be transferred together, even if they are not directly adjacent (Lawrence and Roth 1996; Lawrence 1999). In contrast, the co-regulation theory requires two genes with similar optimal expression patterns to be placed directly adjacent, which appears implausible. However, genome rearrangements are quite common in culture, occurring at rates of around 10^{-4} per generation (Papadopoulos *et al.* 1999), so that potentially advantageous rearrangements, including double rearrangements, should be sampled at significant rates during evolutionary timescales. Evidence for apparently implausible rearrangements is indeed available. For example, comparison of conserved gene order across bacteria has identified a number of cases of xenologous displacement, whereby some – but not all – of the genes in an operon have been replaced by distant homologs (Omelchenko *et al.* 2003). As these homologs are too diverged for homologous recombination to take place, it appears that the foreign genes have been acquired and furthermore shuffled to the correct location to maintain the original operon. Thus, we argue that two genes with similar optimal expression patterns will often be shuffled and selected to be directly adjacent.

Alternatively, it has been observed that essential genes tend to cluster together over distances of up to 30 genes or roughly 30 kilobases (Pal and Hurst 2004), and that regions of the genome over 100 kilobases in size tend to have similar expression patterns (Allen *et al.* 2003). Both of these effects appear to occur over a larger scale than operons, which average about 4-5 genes or 4-5 kb. Thus, an intermediate form of genomic hitchhiking may exist, whereby certain regions of the genome have a bias towards different expression patterns. This might help drive functionally related genes closer together, so that operon formation is more likely.

Although our results support the co-regulation theory for operon formation, other alternatives to the selfish operon theory have been proposed, based on the observation that many highly conserved operons code for multi-protein complexes (Dandekar *et al.* 1998). We argue that this observation is consistent with the co-regulation theory: genes with weaker functional links would have similar optimal expression patterns in only a restricted group of organisms, and such operons would be less conserved. Furthermore, although the strong conservation of operons that code for complexes may reflect factors besides conserved regulation, such as co-translational folding (Dandekar *et al.* 1998) or minimizing the half-life of toxic monomers (Pal and Hurst 2004), these factors cannot explain the frequent formation of operons that do not contain physically interacting genes. For example, many metabolic operons contain proteins that are believed not to interact physically (Lawrence and Roth 1996), and physical interaction is unlikely to explain the "genomic hitchhiking" phenomenon discussed above. Overall, we argue that selection for co-regulation may be a dominant force in the formation of operons, as well as in the maintenance of existing operons. Further research into the evolution of gene regulation in prokaryotes will be required to confirm this hypothesis.

Methods

HGT Genes

A major challenge in studying the evolutionary history of genes is to identify distant orthologs and to distinguish orthologs from paralogs. To assist in both problems, we used clusters of orthologous groups (COGs, Tatusov *et al.* 2001) as well as BLAST hits (see below). In contrast, a recent study of HGT and ORFan genes in *E. coli* (Daubin and Ochman 2004) relied on BLAST hits, and used a more relaxed E-value cutoff when determining the absence of a homolog than when determining the presence of a homolog. Compared to the previous study, we identified additional HGT genes because COG allowed us to distinguish paralogs from orthologs with confidence, but missed other HGT genes because they were not in COG. However, we obtained very similar results on the relationship between HGT and operon formation with both classifications (data not shown).

To describe our method in more detail, we considered a gene to be a "good homolog" if it was either a putative ortholog or in the same COG. We defined putative orthologs as bidirectional best BLAST hits with 75% coverage both ways. BLAST hits were identified with an E-value cutoff of 10^{-5} and an effective database size of 10^8. We assigned genes to COGs via reverse position-specific BLAST (Schaffer *et al.* 2001) against CDD (Marchler-Bauer *et al.* 2003) with an E-value cutoff of 10^{-3}, again using an effective database size of 10^8. To identify good homologs when determining HGT genes, we required them to be in COG, and measured the presence of either the COG or an ortholog in each genome. However, in a few cases, COG assignments were obviously inconsistent. For example, a COG might be present in the Enterobacteria, missing from the two older outgroups (HPVS and γ-Proteobacteria), and present in distantly related organisms, yet the best BLAST hit of the *E. coli* gene outside of the Enterobacteria might be to a γ-Proteobacterial gene. To overcome this limitation of COG, before we classified a gene as HGT, we checked that there were no good BLAST hits (better than any of the older outgroups) in the two consecutive outgroups that were missing the COG. For identifying ORFans, we relied on the gene not being classified as either native or HGT and on the absence of BLAST hits to genes outside of the Proteobacteria. We used the complete genome sequences of 28 γ-Proteobacteria, 24 other Proteobacteria, 63 other Bacteria, and 16 Archaea.

We also confirmed that the genes that we classified as HGT were imported into the *E. coli* lineage from distant bacteria, and not the other way round. Although the requirement that a gene be absent from two consecutive outgroups is intended to ensure that the gene was imported into *E. coli* (Daubin and Ochman 2004), it is also possible that such genes are ORFans that were later exported to other bacteria. The two scenarios lead to different predictions of how diverse the bacteria containing the gene would be. In the import scenario, the gene could be very old, and could be present in diverse bacteria, while in the second scenario, the gene must be new, and should be restricted to two or three closely related groups of bacteria representing one or two export events. (Multiple export events are also possible, but seem much less likely than multiple import events, as the import scenario does not restrict the time to perform these transfers.) To distinguish between import and export, we chose 10 HGT genes at random and examined the diversity of bacteria that contained that gene. In all ten cases, the genes were present in highly diverse bacteria and were

not consistent with recent export to one or two lineages. As a typical example, *yjgK* (COG2731) is present in *Vibrio* and closer relatives of *E. coli* and also in *Clostridium*, *Fusobacterium*, *Bacteroides*, and δ-Proteobacteria. Thus, we believe that most of the HGT genes were imported into the E. coli lineage, rather than being ORFans that were then transferred out.

Operon Pairs

To identify imported or new operon pairs, we examined which genomes contained homologous pairs of genes in the same predicted operons. We used both COG and BLAST hits to identify homologous pairs – as paralogous operons are less common than paralogous genes, we did not use the best-hit rule that was used for classifying genes as HGT. We predicted operons in each genome by examining adjacent pairs: every adjacent pair of genes on the same strand was predicted to be in the same operon or not based on the distance between the genes (in base pairs) and the conservation of the potential operon. At the level of pairs, these predictions are estimated to be 84% accurate in *E. coli* K12 and at least 82% accurate in most prokaryotes (M.N.P, K.H.H., E.J.A., & A.P.A., submitted). The effect of false negatives in these operon predictions was minimal because we examined several genomes within each outgroup and because we required imported pairs to be absent from two consecutive outgroups. Manual examination of the new operon pairs shown in Table 3 and of some of the imported pairs confirmed that false negative operon predictions in other genomes did not create spurious new or imported operons in *E. coli*. The effect of false positive operon predictions was minimized by considering only homologs of adjacent genes predicted to be in the same operon in *E. coli*.

We validated the new operon pairs shown in Table 3 to verify that they were in fact operon pairs. To do this we compared them to a database of known transcripts in *E. coli* (Karp *et al.* 2002), or, when such information was not available, we examined their expression patterns. To quantify the similarity of two gene's expression profiles, we used the Pearson correlation of their normalized log ratios across microarray experiments. We used the normalized log-ratios given in the Stanford Microarray Database (Gollub *et al.* 2003), except that we subtracted the mean from each experiment before computing the correlation coefficient for two genes. Overall, we confirmed 38 of the 58 predicted new operon pairs containing ubiquitous single-copy genes as being known operon pairs or having similar expression patterns. We also looked for further information in the literature about these pairs, and identified one difficult case, the predicted new operon pair *holC-valS*. *valS* has its own promoter, located in the middle of the *holC* gene (Heck and Hatfield 1988), and we did not find any information about the transcription of *holC*. Nevertheless, these genes overlap by one base pair, which is a strong indicator of operons (Salgado *et al.* 2000), they are in the same predicted operon in most of the γ-Proteobacteria, and they have similar expression patterns (the Pearson correlation coefficient is 0.74). Thus, we think it likely that *valS* can be transcribed with *holC* as well as from its own promoter.

As mentioned in Table 1, our method classified three single-copy ubiquitous genes as being in imported operon pairs. These genes were in two operon pars: *yabC-ftsL* and *glmU-glmS*. First, the single-copy ubiquitous gene *yabC* is in an operon with the rapidly evolving gene *ftsL*. COG

incorrectly classified some γ-proteobacterial homologs of *ftsL* as not being in the same family, and furthermore, *ftsL* appears to have been transferred from the δ-proteobacteria to *Thermoanaerobacter tengcongensis*. Together these gave the false impression the operon pair is present in *T. tengcongensis* but not in the distant γ-proteobacteria. Second, *glmU* and *glmS* are both single-copy ubiquitous genes and are in an ancient operon. Around the time that the common ancestor of *Pseudomonas* and *E. coli* diverged from other γ-Proteobacteria, this operon was apparently split into two operons by the insertion of a transcription factor (TF), giving *glmU* and TF-*glmS*. Although *glmU* and the TF might still be an operon pair, our method predicted that it was not. Because the pair is widely spaced (up to 300 bp) and was independently disrupted in several species (either by shuffling the genes apart or by inserting another gene), this prediction seems likely to be correct. Then, after the divergence of the Enterobacteria, the TF was deleted, thus reviving the ancient operon. Phylogenetic trees for *glmU* and *glmS* do not support the alternative hypothesis that the *glmU-glmS* operon was transferred into the ancestor of the Enterobacteria (data not shown). The errors in classifying *yabC-ftsL* and *glmU-glmS* illustrate the challenges of automatically inferring the history of genes and operons. However, errors appear to be rare: of 178 operon pairs containing single-copy ubiquitous genes that are believed not to be subject to HGT, only 2 were classified as imported. Thus, these errors do not affect the reliability of our conclusions.

Gene Trees

To test whether genes in new operons had sequence evidence for HGT, we built phylogenetic trees. We examined each gene in *E. coli* K12 that is present as a single-copy COG in four or more of 13 γ-Proteobacteria with a fully resolved species tree (Lerat *et al.* 2003). (Four genomes is the minimum number of nodes required to distinguish different topologies for unrooted trees.) We used single-copy COGs to reduce the prevalence of paralogs. Although paralogous duplication followed by gene loss in several species can never be ruled out entirely, similar results were obtained when analyzing COGs that were never present more than once in these 13 genomes (data not shown). Given protein sequences for a gene and its single-copy homologs, we created multiple sequence alignments with ClustalW (Thompson *et al.* 1994), using the BLOSUM-80 matrix, and then removed columns containing gaps. Phylogenetic trees were created from these trimmed alignments with TreePuzzle 5.1 (Schmidt *et al.* 2002). To reduce the computation time when computing so many trees, we used TreePuzzle's default assumption of uniform evolutionary rates across sites instead of the more biological assumption of gamma-distributed rates. (Using uniform rates caused a few of the genes classified as non-HGT by Lerat *et al.* (2003) to reject the species tree.) To determine whether the maximum likelihood gene tree was consistent with the species tree, we used the one-sided Kishino-Hasegawa (KH) test implemented by TreePuzzle, instead of building trees on resampled data sets or conducting the Shimodaira-Hasegawa test on every possible tree. These alternatives were computationally impractical for over 1,000 trees. The one-sided KH test is too aggressive in rejecting the pre-given (species) tree and in accepting the maximum likelihood tree, and strictly speaking it should be used only to accept the species tree (Goldman *et al.* 2000). Nevertheless, even this test accepted the species tree for over 90% of the genes in new operons. Furthermore, we were investigating the relative level of HGT in genes that formed new operons versus other genes, rather than making determinations about any specific gene, and we controlled for the increasing sensitivity

of the KH test as the number of homologs in the tree increased (see Results). Thus, the details of alignment and tree construction and statistical testing should not affect our conclusions.

Statistics

Statistical tests were conducted with the R open-source statistics language (http://www.r-project.org). The partial Spearman correlation between two variables x and y, after controlling for a third variable z, was computed from the pairwise Spearman correlation coefficients by the formula $r_{XY,Z} = (r_{XY} - r_{XZ} \cdot r_{YZ}) \; / \; \sqrt{(1 - r_{XZ}^2) \cdot (1 - r_{XZ}^2)}$. The significance of a partial correlation $r_{XY,Z}$ with n data points was assessed with a two-tailed t-test on $t = r_{XY,Z} \cdot \sqrt{(n-3)/(1 - r_{XY,Z}^2)}$ with $n - 3$ degrees of freedom. We used partial Spearman correlations rather than partial Pearson correlations – which is equivalent to using the ranks of the data instead of the raw values – because the amount of footprinted base pairs has a skewed distribution, as can be seen from the broad right-most arrow in Figure 3.

To compute the protein sequence conservation of genes between *E. coli* and *Salmonella enterica Typhi*, we used the %identity (from BLAST) between putative orthologs. To avoid paralogs, we required the orthologs to have at least 60% identity.

Acknowledgments

We thank Vincent Daubin for providing an alternate classification of *E. coli* K12 genes and Lee-Ann McCue for providing phylogenetic footprints for *E. coli* K12. This work was supported by a grant from the DOE Genomes To Life program (DE-AC03-76SF00098).

References

Allen,T.E., Herrgard,M.J., Liu,M., Qiu,Y., Glasner,J.D., Blattner,F.R. and Palsson,B.O. (2003) Genome-scale analysis of the uses of the escherichia coli genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.,* **185**, 6392–9.

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning,* **21**, 51–80.

Bulyk,M.L., McGuire,A.M., Masuda,N. and Church,G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Gen. Research,* **14**, 201–8.

Cherry,J.L. (2003) Genome size and operon content. *J. Theor. Biol.,* **221**, 401–10.

Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci.,* **23**, 324–8.

Daubin,V. and Ochman,H. (2004) Bacterial genomes as new gene homes: the genealogy of orfans in e. coli. *Genome Res.,* **14**, 1036–42.

de Daruvar,A., Collado-Vides,J. and Valencia,A. (2002) Analysis of the cellular functions of escherichia coli operons and their conservation in bacillus subtilis. *J. Mol. Evol.,* **55**, 211–21.

Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.,* **29**, 1216–21.

Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balazsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S. *et al.* (2003) Experimental determination and system level analysis of essential genes in escherichia coli mg1655. *J. Bacteriol.,* **185**, 5673–84.

Goldman,N., Anderson,J.P. and Rodrigo,A.G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.,* **49**, 652–670.

Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al.* (2003) The stanford microarray database: data access and quality assessment tools. *Nucleic Acids Res.,* **31**, 94–6.

Heck,J.D. and Hatfield,G.W. (1988) Valyl-trna synthetase gene of escherichia coli k12. molecular genetic characterization. *J. Biol. Chem.,* **263**, 857–67.

Itoh,T., Takemoto,K., Mori,H. and Gojobori,T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.,* **16**, 332–46.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The ecocyc database. *Nucleic Acids Res.,* **30**, 56–8.

Lawrence,J.G. (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.,* **9**, 642–8.

Lawrence,J.G. and Ochman,H. (1998) Molecular archaeology of the escherichia coli genome. *Proc. Natl. Acad. Sci. USA,* **95**, 9413–7.

Lawrence,J.G. and Roth,J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics,* **143**, 1843–60.

Lerat,E., Daubin,V. and Moran,N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biol.,* **1**, E19.

Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) Dbtbs: database of transcriptional regulation in bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res.,* **32**, D75–7.

Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y. *et al.* (2003) Cdd: a curated entrez database of conserved domain alignments. *Nucleic Acids Res.,* **31**, 383–7.

McCue,L.A., Thompson,W., Carmack,C.S. and Lawrence,C.E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.,* **12**, 1523–32.

Omelchenko,M.V., Makarova,K.S., Wolf,Y.I., Rogozin,I.B. and Koonin,E.V. (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.,* **4**, R55.

Pal,C. and Hurst,L.D. (2004) Evidence against the selfish operon theory. *Trends Genet.,* **20**, 232–4.

Papadopoulos,D., Schneider,D., Meier-Eiss,J., Arber,W., Lenski,R.E. and Blot,M. (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proc. Natl. Acad. Sci. USA,* **96**, 3807–3812.

Ragan,M.. and Charlebois,R.L. (2002) Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *Int. J. Syst. Evol. Microbiol.,* **52**, 777–87.

Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of dna-binding site matrices for 55 proteins applied to the complete escherichia coli k-12 genome. *J. Mol. Biol.,* **284**, 241–54.

Rogozin,I.B., Makarova,K.S., Murvai,J., Czabarka,E., Wolf,Y.I., Tatusov,R.L., Szekely,L.A. and Koonin,E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.,* **30**, 2212–23.

Sabatti,C., Rohlin,L., Oh,M.K. and Liao,J.C. (2002) Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acids Res.,* **30**, 2886–93.

Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in escherichia coli: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA,* **97**, 6652–7.

Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L. *et al.* (2001) Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.,* **29**, 2994–3005.

Schmidt,H.A., Strimmer,K., Vingron,M. and von Haeseler,A. (2002) Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics,* **18**, 502–504.

Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.,* **29**, 22–8.

Terai,G., Takagi,T. and Nakai,K. (2001) Prediction of co-regulated genes in bacillus subtilis on the basis of upstream elements conserved across three closely related species. *Genome Biol.,* **2**, RESEARCH0048.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.,* **22**, 4673–4680.

Web Site References

http://www.r-project.org/ – the R statistics package

	History of the Gene		
	Native		
	Ubiquitous	**Typical**	**HGT**
Formed a new operon	20% (39/200)	22% (474/2164)	17% (58/345)
At time of HGT	–	–	9% (31/345)
Since Salmonella	2% (3/200)	8% (174/2164)	9% (13/138)
In an imported operon	2% (3/200)	14% (294/2164)	48% (165/345)

Table 1: Proportions of native and HGT genes that formed new operons or were imported as operons. For the analysis of newer operons (since *Salmonella*), we included only HGT genes with older ages, so that all genes were in the *E. coli* lineage for the entire time period analyzed and had equal opportunity to form new operons. The three single-copy ubiquitous genes that are in imported operons reflect rare errors of our automated classification and not HGT of these genes (see Methods).

	In a New Operon?	
	Yes	**No**
# Concordant	345	692
# Discordant	22	69
% Concordant	94.0%	90.9%

Table 2: Genes in new operons are no more likely than other genes to have trees that are discordant with the species tree. As described in the text, we used the one-sided Kishino-Hasegawa test to determine whether genes in new operons had trees that were concordant with the species tree. To avoid discordant trees due to paralogs, only genes present as unique members of a COG were included. The two percentages shown are not significantly different ($p = 0.08$, Fisher exact test).

Upstream Gene	Downstream Gene	Known Operon?	Microarray similarity	Age of Pair
yfhB	**yfhC** *		0.51	Salmonella
yrdC	aroE		0.80	HPVS
yrdD	**yrdC**		0.76	HPVS
yhbE	**yhbZ** *		0.73	HPVS
pyrF *	yciH		0.71	HPVS
murB	birA *		0.67	HPVS
ygiM	**cca** *		0.65	HPVS
kdtA *	**kdtB** *	Yes	0.61	HPVS
yggJ	gshB *		0.58	HPVS
yhhF *	yhhL *		0.54	HPVS
pyrE	rph	Yes	-0.13	HPVS
lgt *	thyA *	Yes	0.81	γ-Proteo.
lspA *	slpA	Yes	0.76	γ-Proteo.
holC *	**valS** *		0.74	γ-Proteo.
rnhB *	**dnaE** *		0.73	γ-Proteo.
ygbB *	ygbO		0.69	γ-Proteo.
ksgA	apaG	Yes	0.59	γ-Proteo.
dapF *	yigA		0.54	γ-Proteo.
ycfC	**purB**	Yes	0.34	γ-Proteo.
priB *	**rpsR** *	Yes	0.93	$\beta\gamma$-Proteo.
rpsF	priB *	Yes	0.92	$\beta\gamma$-Proteo.
nlpB	**dapA** *	Yes	0.87	$\beta\gamma$-Proteo.
atpI	**atpB** *	Yes	0.85	$\beta\gamma$-Proteo.
ftsJ	hflB	Yes	0.72	$\beta\gamma$-Proteo.
yacE *	yacF		0.69	$\beta\gamma$-Proteo.
folC	dedD		0.64	$\beta\gamma$-Proteo.
yffB *	**dapE** *		0.54	$\beta\gamma$-Proteo.
slpA	**lytB** *	Yes	-	$\beta\gamma$-Proteo.
sucB *	sucC	Yes	0.98	Proteo.
rnc *	era *	Yes	0.85	Proteo.
yaeL	**yaeT** *		0.83	Proteo.
ydgQ	**nth**		0.76	Proteo.
ndk	**yfgB**		0.74	Proteo.
pdxA	**ksgA**	Yes	0.73	Proteo.
pepA *	holC *		0.73	Proteo.
pheT	himA	Yes	0.64	Proteo.
ribF *	**ileS** *	Yes	0.63	Proteo.
b2512	**b2511** *		0.52	Proteo.

Table 3: Validated new operon pairs containing ubiquitous single-copy (non-HGT) genes. The ubiquitous single-copy genes are in **bold**, and asterisks (*) mark the genes reported to be essential (Gerdes *et al.* 2003). Known operons are taken from Karp *et al.* (2002). The microarray similarity is the Pearson (linear) correlation of normalized log-ratios across 74 *E. coli* microarray experiments that compared mRNA levels (Gollub *et al.* 2003). We used a microarray similarity of 0.5 or greater as confirmation that the predicted pair is a true operon pair. We validated this threshold against a database of known transcripts (Karp *et al.* 2002): 72% of known operon pairs and only 27% of known not-operon adjacent pairs had correlation coefficients greater than 0.5.

Genes			*Operon Pairs*			
Native	**HGT**	**ORFan**	**Ancestral**	**Imported**	**New**	**Groups of Genomes**
+	+	+	+	+	+	E. coli K12
+	+	+	+	+	+	Other E. colis, Shigellas (5)
+	+	+	+	+	+	Salmonellas (3)
+	-	-	+	-	-	Other enterics (6)
+	-	-	+	-	-	HPVS (6)
+	-	-	+	-	-	Other γ-Proteobacteria (7)
+	+	-	+	+	-	β-Proteobacteria (4)
+	+	-	+	+	-	Other proteobacteria (20)
+	+	-	+	+	-	Other bacteria (79)

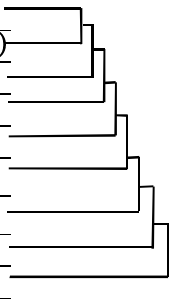


Figure 1: The evolutionary history of genes and operons. For each gene in *E. coli* K12, we determined which groups of genomes contained a potential ortholog of that gene, and classified genes as native, HGT, or ORFan. We performed a similar analysis on each adjacent pair of genes predicted to be in the same operon, and classified pairs as ancestral, imported, or new. Some genes and pairs could not be classified. We show examples of patterns of presence or absence for each class of gene and for each class of operon pair. The placement of the genomes at varying distances from *E. coli* K12 is in accordance with generally accepted phylogenies and with a whole-genome protein sequence tree (P. Dehal & E.J.A., unpublished results). "Other enterics" includes *Yersinia*, *Buchnera*, and *Wigglesworthia* species; "HPVS" includes *Haemophilus*, *Pasteurella*, *Vibrio*, and *Shewanella* species; and "other γ-Proteobacteria" includes *Pseudomonas*, *Xanthomonas*, and *Xylella* species. For the inferred histories to be correct, the union of all groups up to a given age must be monophyletic, but each outgroup need not be. For example, we believe that HPVS and the Enterobacteria together form a monophyletic clade, but not HPVS by themselves.

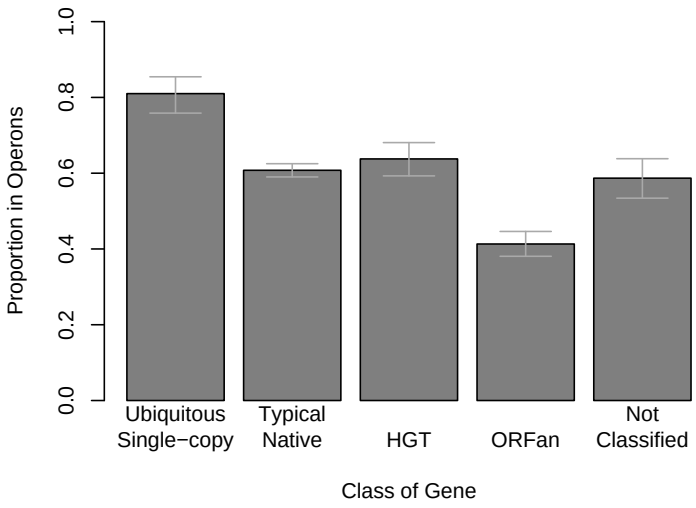


Figure 2: HGT genes are not particularly likely to be in operons. For each class of gene, solid bars show the proportion that are in predicted operons. Error bars show 90% confidence intervals from the binomial test; if two error bars do not overlap, then the corresponding classes have significantly different probabilities of being in operons ($p < 0.05$).

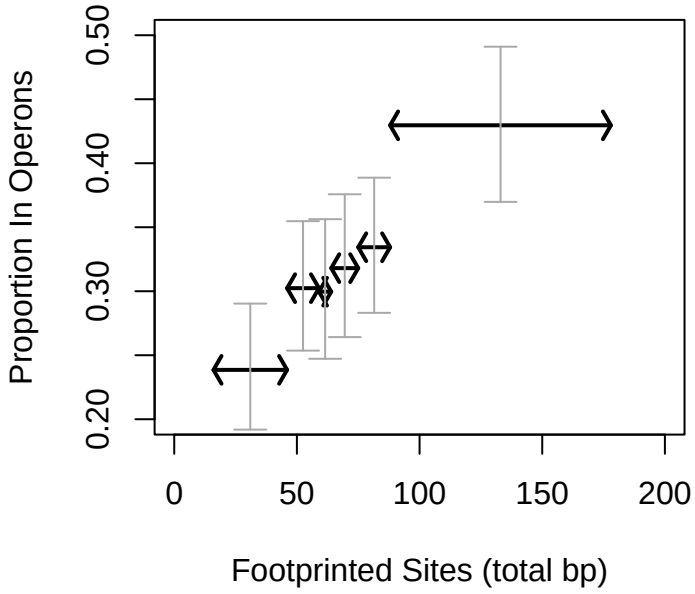
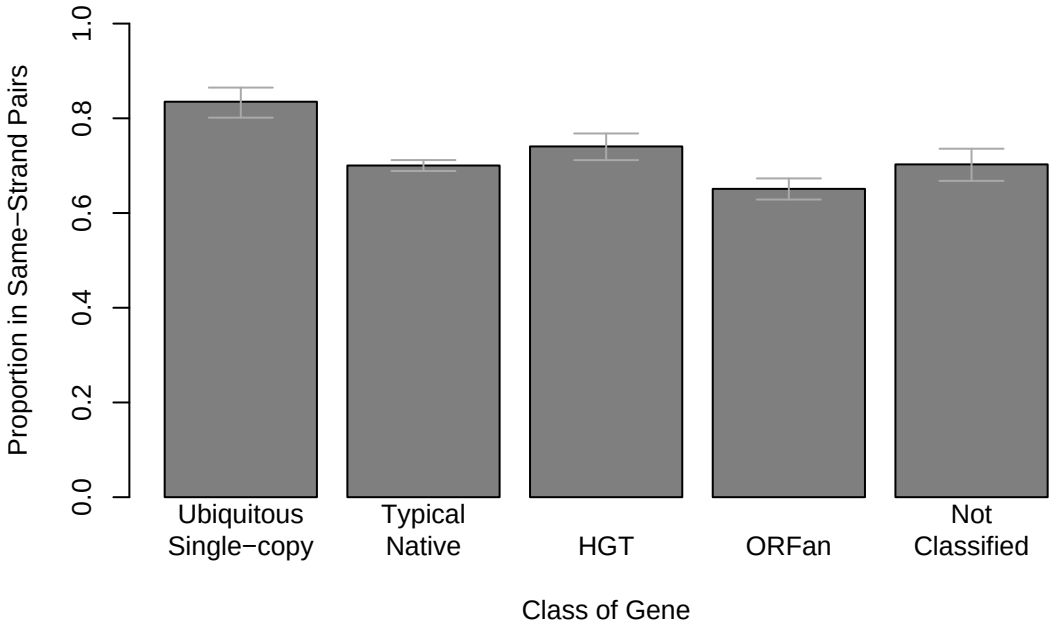


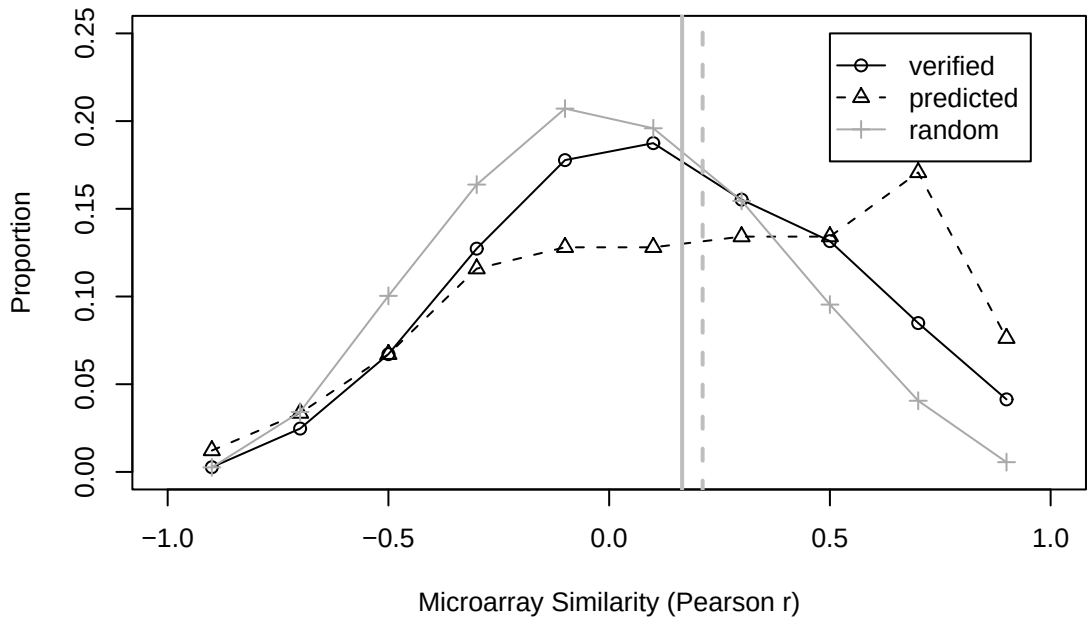
Figure 3: Genes with more conserved upstream sequences are more likely to be in operons. For each *E. coli* gene with one or more sites from phylogenetic footprinting (McCue *et al.* 2002), we asked whether it was predicted to be at the beginning of a multi-gene operon or to be transcribed by itself. For each group of genes with varying amounts of footprinted sequence, as measured in total base pairs and indicated with the horizontal arrows, the y axis shows the proportion of genes that are in operons. (These ranges were chosen to give the same number of genes in each range.) For each range, a vertical bar shows the 90% confidence interval for the proportion (from the binomial test). Genes in the middle or at the end of predicted operons were excluded from this analysis, which is why the proportion of genes in operons is lower than in Figure 2.

	Number of Sites				
	1	2	3	4	5
Number of Genes					
Operons	76	18	9	7	2
Non-Operons	79	17	2	0	3

Supplementary Table 1: Phylogenetic footprints in B. subtilis correlate with operons. We show the distribution of the number of predicted regulatory sites (from Terai *et al.* 2001) upstream of predicted operons and upstream of predicted single-gene transcripts. The greater frequency of operons with three or more sites is statistically significant ($p = 0.01$, Fisher exact test).



Supplementary Figure 1: For each class of gene, the proportion of adjacent pairs that are on the same strand. Each gene is adjacent to two other genes, and in the absence of operons, on average, one of these other genes will be on the same strand, giving a proportion of 0.5. For genes in long operons, the average proportion will be near 1.0. For each class, the solid bar shows this proportion and the error bar shows the 90% confidence interval from the binomial test. If two error bars do not overlap, then the corresponding classes have significantly different probabilities of same-strand pairs ($p < 0.05$).



Supplementary Figure 2: In B. subtilis, predicted regulatory sites from phylogenetic footprinting are as predictive of expression patterns as experimentally verified sites. We show the distribution of microarray similarity for random pairs of genes, for pairs of genes with predicted sites from the same cluster of phylogenetic footprints (Terai *et al.* 2001), and from pairs of genes that are experimentally verified to bind the same TF (Makita *et al.* 2004), with sigma factors excluded. Microarray similarities were computed with Pearson correlation coefficients of normalized log-ratios across 78 *B. subtilis* microarray experiments that compared mRNA levels (Gollub *et al.* 2003). The vertical solid line shows the median over all TFs of the mean similarity of the pairs that are verified to bind that TF. This median is right-shifted from the overall distribution for pairs sharing a verified site because some of the less predictive TFs have many binding sites. The corresponding median for the predicted clusters of sites is also shown as the vertical dashed line. This median is to the right of the median for verified sites, which confirms that the relationship between predicted sites and expression patterns is very strong.