Title: The Life-cycle of Operons

Authors: Morgan N. Price, Adam P. Arkin, and Eric J. Alm

Author affiliation: Lawrence Berkeley Lab, Berkeley CA, USA and the Virtual Institute for Microbial Stress and Survival. 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 510-486-6899, fax 510-486-6219, address Lawrence Berkeley National Lab, 1 Cyclotron Road, Mailstop 977-152, Berkeley, CA 94720

Abstract:

Operons are a major feature of all prokaryotic genomes, but how and why operon structures vary is not well understood. To elucidate the life-cycle of operons, we compared gene order between *Escherichia coli* K12 and its relatives and identified the recently formed and destroyed operons in *E. coli*. This allowed us to determine how operons form, how they become closely spaced, and how they die. Our findings suggest that operon evolution is driven by selection on gene expression patterns. First, both operon creation and operon destruction lead to large changes in gene expression patterns. For example, the removal of *lysA* and *ruvA* from ancestral operons that contained essential genes allowed their expression to respond to lysine levels and DNA damage, respectively. Second, some operons have undergone accelerated evolution, with multiple new genes being added during a brief period. Third, although most operons are closely spaced because of a neutral bias towards deletion and because of selection against large overlaps, highly expressed operons tend to be widely spaced because of regulatory fine-tuning by intervening sequences. Although operon evolution seems to be adaptive, it need not be optimal: new operons often comprise functionally unrelated genes that were already in proximity before the operon formed.

Introduction

Operons are groups of genes that are transcribed in a single mRNA. Operons are widespread in all bacterial and archaeal genomes (Wolf *et al.* 2001; Ermolaeva *et al.* 2001; Price *et al.* 2005*a*), and in the typical genome, around half of all protein-coding genes are in multi-gene operons. Operons often, but not always, code for genes in the same functional pathway (de Daruvar *et al.* 2002; Rogozin *et al.* 2002). Operons are often conserved across species by vertical inheritance (Overbeek *et al.* 1999; Itoh *et al.* 1999; Ermolaeva *et al.* 2001; Wolf *et al.* 2001) and tend to be quite compact: in most bacteria, genes in the same operon are usually separated by less than 20 base pairs of DNA (Moreno-Hagelsieb and Collado-Vides 2002). Both conservation and close spacing allow for the computational prediction of operons in diverse prokaryotes (Salgado *et al.* 2000; Ermolaeva *et al.* 2001; Wolf *et al.* 2001; Moreno-Hagelsieb and Collado-Vides 2002; Price *et al.* 2005*a*).

Why are operons so prevalent? The traditional explanation is that genes are placed in the same operon so that they will have similar expression patterns (Jacob and Monod 1961). However,

although genes in the same operon do have (mostly) similar expression patterns (Sabatti *et al.* 2002), genes can also be coregulated without being in the same operon. Thus, it has been argued that co-regulation could more easily evolve by modifying two independent promoters rather than by placing two genes in proximity (Lawrence and Roth 1996). In contrast, we argue that for complex regulation, an operon with one complex promoter would arise more rapidly than would two independent complex promoters (Price *et al.* 2005*b*). As predicted by this theory, operons tend to have more complex conserved regulatory sequences than individually transcribed genes (Price *et al.* 2005*b*).

Another popular view has been that operons are selfish: they form because they facilitate the horizontal transfer of metabolic or other capabilities that can be provided by a single operon containing several genes (Lawrence and Roth 1996). This theory is consistent with the compactness of operons and also with the observation that operons often undergo horizontal gene transfer (Lawrence and Roth 1996; Omelchenko *et al.* 2003; Price *et al.* 2005*b*). However, essential and other non-horizontally-transferred (non-HGT) genes are particularly likely to be in operons (Pal and Hurst 2004; Price *et al.* 2005*b*), and non-HGT genes are forming new operons at significant rates (Price *et al.* 2005*b*). The selfish theory also cannot explain why many operons consist of genes with no apparent functional relationship (Rogozin *et al.* 2002; de Daruvar *et al.* 2002). This functional incoherence is particularly common for new operons (Price *et al.* 2005*b*). Thus, it appears that HGT may increase the prevalence of some operons, but that HGT is not involved in operon formation.

Finally, it has been suggested that placing genes that code for multi-subunit protein complexes in the same operon is beneficial because it speeds complex formation and folding (Dandekar *et al.* 1998; Pal and Hurst 2004) or because it reduces wastage due to stochastic fluctuations in gene expression (Swain 2004). Although the most highly conserved operons do tend to code for protein complexes (Dandekar *et al.* 1998), most operons do not, and, vice versa, only a few percent of protein-protein interactions involve genes encoded by the same operon (Butland *et al.* 2005).

Overall, genome-wide studies have supported the traditional view that operons exist because they facilitate co-regulation. However, this does not explain why operon structures change over evolutionary time. Specific questions of interest include: How do operons form? Why are are most operons so closely spaced, while some highly conserved operons are widely spaced? Is operon evolution neutral, as suggested by the loss of most ancestral operons in some genomes (Itoh *et al.* 1999), or is it adaptive? Do changes in operon structure lead to changes in gene expression patterns, or are genes cotranscribed from one promoter in some organisms and co-regulated from distinct promoters in other organisms, without obvious functional consequences?

To address these questions, we examined the newly formed or recently deceased operons of *Escherichia coli* K12. To address the issue of spacing, we compared orthologous operons in *E. coli* K12 and its close relative *Salmonella typhimurium* LT2. To summarize our results, we present a model for the life-cycle of operons (Figure 1).

Results

How Do Operons Form?

It appears that operons containing native genes form without horizontal transfer events (Price *et al.* 2005*b*), but the mechanism is unknown. Because conserved operons often undergo rearrangements or acquire new genes (Itoh *et al.* 1999), we distinguish new operons from modifications to preexisting operons. More precisely, we first examine cases where genes that were not previously co-transcribed are placed next to each other in an operon, and then consider the special case of how new genes are added to pre-existing operons.

New Operons

New operons could form by rearrangement or by deletion. First, genome rearrangements could bring two genes that were not previously near each other into proximity so that they are co-transcribed. Some genomes with large numbers of repetitive elements, such as *Helicobacter pylori* and *Synechocystis* PCC 6803, have lost most of their ancestral operons, presumably because the repetitive elements cause frequent genome rearrangements (Itoh *et al.* 1999). Nevertheless, sequence analysis and expression data suggest that *H. pylori* and *Synechocystis* contain large numbers of operons (Price *et al.* 2005*a*). Although these putative new operons tend not to be functionally coherent, new or poorly conserved operons in *E. coli* also tend not to be functionally coherent (de Daruvar *et al.* 2002; Price *et al.* 2005*b*). Thus, rearrangements may cause the production of new operons as well as the destruction of ancestral operons.

Alternatively, if two genes are near each other and are on the same strand, they could form an operon by deleting the intervening DNA (Lawrence and Roth 1996). However, previous empirical reports discuss only removing genes from operons by deletion (Itoh *et al.* 1999).

To identify the mechanism of operon formation, we examined evolutionarily recent operons in *E. coli* K12. Pairs of adjacent genes were predicted to be in the same operon (or not) from the distance between them on the DNA and the conservation of the putative operon (Price *et al.* 2005*a*). Evolutionarily recent operon pairs were identified as those present only in close relatives (Price *et al.* 2005*b*). In this study, we considered operons that are new to the Enterobacteria or are shared with somewhat more distant relatives (*Haemophilus*, *Pasteurella*, *Vibrio*, or *Shewanella* species). We also classified genes as native, horizontally transferred (HGT), or "ORFan," again based on the presence or absence of the gene in other groups of bacteria (Ragan and Charlebois 2002; Daubin and Ochman 2004). ORFans are genes that lack identifiable homologs outside of a group of closely related bacteria (Fischer and Eisenberg 1999). Most ORFans are functional protein-coding genes that contribute to the fitness of the organism (they are under purifying selection), and they were probably acquired from bacteriophage (Daubin and Ochman 2004).

We found that predicted new operons are highly enriched for ORFan genes (Figure 2A) and often combine an ORFan with a native gene (Figure 2B). The prevalence of ORFans in new operons is somewhat surprising given that ORFans are less likely than native or HGT genes to be in operons (Price *et al.* 2005*b*). The most parsimonious evolutionary scenario for constructing a native-ORFan pair is a single insertion event that transfers the ORFan into the genome and places it adjacent to the native gene. To test this hypothesis, we compared the evolutionary age of the new operon to that of the ORFan. The age was determined from the most distant relative that contained the new operon or ORFan (see Methods). Consistent with the insertion scenario, we found that the estimated evolutionary age of the native-ORFan operon pair often matches the age of the ORFan (Figure 2B). We considered that native-ORFan pairs might be a mechanism for ensuring that the ORFan gene is expressed. Consistent with this view, we found that the native gene is more often the upstream gene in the pair (Figure 2B; $P = 0.03$, binomial test), so that the ORFan gene can be transcribed from a native promoter without perturbing the expression of a native gene.

There are also ORFan-ORFan pairs. The age of these pairs often matches the age of both genes in the pair (Figure 2B), suggesting that the entire operon was imported in a single event. Thus, many of the "new" ORFan-ORFan pairs may actually have been horizontally transfered from an unknown source, such as phage. Because phages have compacted operon-rich genomes, it is surprising that more ORFans are not in such pairs, and that ORFans are less likely to be in operons than other genes (Price *et al.* 2005*b*). Perhaps the phage operon benefits the phage, whereas only one gene in the operon would benefit the host.

Because new operons are, by definition, not conserved across many genomes, these operon predictions may be less reliable. However, new operon pairs of each of the three major types discussed above tend to have strongly correlated expression patterns (Figure 2C). Therefore, most of these predicted new operon pairs are likely to be operons.

Although ORFans tend to be poorly annotated, a few annotated ORFans are in characterized operons. The ORFan *holD*, a regulatory subunit of DNA polymerase III, is co-transcribed with the native gene *rimI*, a ribosomal protein S18 acetyltransferase. The ORFan chromosomal partitioning genes *mukFEB* are co-transcribed with a native methyltransferase (*smtA*). And *flhE* is co-transcribed with native genes *flhAB*. The native genes *flhA* and *flhB* are required for flagellar export, but the ORFan *flhE* is not (Minamino *et al.* 1994; Minamino and Macnab 1999). *flhE* appears to be annotated as a flagellar gene purely because of its location. Because many new operons encode functionally unrelated genes (Price *et al.* 2005*b*), we argue that this annotation is likely to be erroneous. Overall, the genes in these ORFan-native operons do not seem to be functionally related.

If new operons containing ORFans often form by insertion, how do new operon pairs containing only native genes form? Although we cannot examine the ancestor of *E. coli* that formed the new operon, we can examine the gene order in close relatives that lack the operon. Specifically, we examined new operon pairs that were shared by *E. coli* K12, *Salmonella* species, and other Enterobacteria, but had non-adjacent orthologs in *Vibrio* species, which are more distantly related. We considered looking at newer operon pairs, but few of the newer operon pairs consist of two native genes (data

not shown). Because ORFans turn over rapidly in bacterial genomes (Daubin and Ochman 2004), it is not surprising that the older new operons are more likely to contain native genes.

Among the 10 *E. coli* operon pairs that have orthologs in *Vibrio* that are not adjacent to each other, we identified six cases where the *Vibrio* genes are near each other (Table 1). In all of these cases, the pair of *Vibrio* genes are on the same strand. In four of these cases, the intervening genes are on the opposite strand, so we are confident that these are not operons in *Vibrio*. (Although an operon could, in principle, contain within it another transcript on the opposite strand, this has not been observed (Salgado *et al.* 2000).) For these six cases, it appears that the operon formed by deleting the intervening genes. Alternatively, the intervening genes could have been inserted into a pre-existing operon in the ancestor of *Vibrio*, but because these operons are unique to the Enterobacteria, deletion is the more parsimonious explanation. Other features of these pairs, such as the absence of homologs for the intervening genes in the Enterobacteria, are consistent with deletion (Supplementary Note 1). In the case of *btuB-murI*, *murI* has over 20 N-terminal amino acids that are encoded by the 3' end of *btuB* and are not present in bacteria that lack the operon (data not shown; evidence that the predicted start codon for *murI* is correct is discussed in Supplementary Note 1). This overlap suggests that the operon formed by deletion in a single event that destroyed the original ribosome binding site of *btuB* as well as other intervening DNA such as promoters and terminators. In general, however, it is possible that the deletion involves several steps (e.g., perhaps the upstream gene's terminator is lost first, and the downstream gene's promoter is lost later).

In another four cases, the *Vibrio* genes were distant from each other, so we suspect that the *E. coli* operons formed by rearrangement (Table 1). In general, we cannot rule out scenarios that involve deletion, such as a rearrangement that placed the genes in proximity that was then followed by a deletion, or a rearrangement in the ancestral *Vibrio* that masked the pre-existing proximity of the genes. However, for *ptr-recB*, we can rule out deletion, as the native gene *ptr* was inserted into and destroyed the ancestral operon *recC-recB*.

We also asked if these new operons might have formed by horizontal gene transfer. All 20 *E. coli* genes listed in Table 1 have phylogenetic trees that are consistent with the species tree of Lerat *et al.* (2003) ($P > 0.2$, Kishino-Hasegawa test; see Methods), so we concluded that HGT was unlikely to be involved, as we found previously in a broader study of HGT and operon formation (Price *et al.* 2005*b*).

In summary, operons containing native genes form both by deleting intervening genes and by rearrangements that bring more distant genes into proximity. In contrast, many new ORFan-native operons probably arise from the insertion of the new gene, and may function to allow the expression of the ORFan gene from a native promoter.

Modifications to Preexisting Operons

We also examined the new operon pairs – adjacent genes that are predicted to be in the same operon in *E. coli* K12 but transcribed separately in related bacteria – for modifications to existing operons (see Methods). Such modifications appear to be much less common than the formation of new operons: we identified 455 new operon pairs but only 81 modification events. However, in a surprisingly large number of cases, two or more new operon pairs are adjacent and of the same age, so that the operon has undergone rapid evolution (Figure 3A). Although it is possible for insertions within pre-existing operons to create two or more new operon pairs with a single event, insertions are much less common than additions at the beginning or end of pre-existing operons (Figure 3B). Also, there is a slight preference for appending a new gene to the end of a preexisting operon instead of pre-pending a gene to the beginning (Figure 3B), so that the majority of genes retain the original promoter instead of acquiring a new one.

To confirm that some operons are undergoing rapid evolution, we manually examined the modified operons. The complete results of this analysis are given in Supplementary Table 1. We found many cases where two or more changes had occurred to the original operon(s). For example, the older operons *yiaMNO* and *sgbHUE* have joined together with several additional genes to give the known *E. coli* operon *yiaKLMNO-lyxK-sgbHUE*. Another striking event is the combination of the ancient *sdhCDAB* and *sucABCD* operons, which code for adjacent steps in the TCA cycle, together with an ORFan gene, to give the experimentally characterized *E. coli* operon *sdhCDAB-b0725-sucABCD*. We also observed several cases where a single gene in an operon has been replaced by a non-homologous gene (Supplementary Table 1). This supports a previous finding that genes in operons are occasionally replaced by horizontally transferred homologs that are too diverged for homologous recombination to occur (Omelchenko *et al.* 2003), although the mechanism by which genes can be replaced or inserted into operons remains unclear. Overall, the rapid evolution of some operons suggests that, even though new operons show limited functional coherence (Price *et al.* 2005*b*), operon formation may be under positive selection.

Operon Spacing

Close Spacings

Most operons are closely spaced, so that adjacent genes in an operon are separated by 20 bases or less of DNA (Moreno-Hagelsieb and Collado-Vides 2002). The stop codon of the upstream gene often overlaps the start codon of the downstream gene, which gives the impression that the genes are packed together as tightly as possible. Close spacing could arise without selection because of the bias of bacterial genomes towards small deletions (Mira *et al.* 2001). Alternatively, close spacings may be preferred because of translational coupling – the ribosome can move directly from the upstream gene's stop codon to a downstream gene's start codon, which can increase translation

from the downstream gene and may also ensure that similar amounts of protein are made from the two genes (reviewed by Kozak 1999; Yu *et al.* 2001). Translational coupling can apply to any close spacing, and does not necessarily explain why the "canonical" overlaps of 1 and 4 bases, in which the start and stop codons overlap, are so common (they account for 24% of known operon pairs in *E. coli* K12).

To study the evolution of close spacing, we first compared the spacing of conserved operons between *E. coli* K12 and its closest relatives. Because spacing is a major factor in operon predictions, we examined only experimentally characterized operons. The spacing within operons evolves very rapidly – in the close relative *Salmonella typhimurium* LT2, a large minority of spacings have changed (Figure 4A). Even in another strain of *E. coli*, O157:H7, 6.4% of spacings have changed. To put this rapid change in perspective, the typical gene that is shared between *E. coli* K12 and *E. coli O157:H7* is 99.5% identical in its protein sequence; for *E. coli* K12 and *Salmonella typhimurium* LT2 the corresponding figure is 90.7%. The changes in spacing are not artifacts from errors in predicted gene starts: if they were, then the change in spacings would often be a multiple of three, but only 34% of changes in spacings between *E. coli* and *Salmonella* are by a multiple of three, which is indistinguishable from the 33% that would be expected by chance. Even canonical spacings are often different between *E. coli* and *Salmonella*, which suggests that canonical spacing may not be under strong selection.

To see how canonical spacings form, we compared the sequences of pairs with canonical spacings in *E. coli* but not *Salmonella*, or vice versa (Table 2). The canonical overlap of the start and stop codons can easily form by deletion (Table 2). Spacing changes are often accompanied by small insertions or deletions at the ends of the protein sequences (e.g., *cysNC*); we speculate that these protein sequence changes are neutral. We also noticed that greater overlaps can form easily (*cysNC* and *cstC-astA*). Because greater overlaps are less common than the canonical overlaps, at least for old operons (Figure 4B), this suggests that there is selection against greater overlaps. Greater overlaps can eliminate translational coupling (reviewed by Yu *et al.* 2001) or they might otherwise interfere with translation. New operons are significantly less likely to be at the canonical spacings than are old operons (Figure 4B; $P < 0.01$, Fisher exact test), which is consistent with the idea that canonical spacings form by deletion after the operon has already formed.

It has also been suggested that the canonical spacing might be common because it stabilizes the transcript – with such close spacings, there is no intergenic region that is free of ribosomes and exposed to RNAses (Moreno-Hagelsieb and Collado-Vides 2002). This hypothesis seems inconsistent with the preference for weakly expressed operons to use the canonical spacings (Figure 4B), as canonical spacings would stabilize the transcript and increase its expression. To test this hypothesis more directly, we examined three genome-wide data sets of mRNA half-lives (Bernstein *et al.* 2002; Selinger *et al.* 2003). Operon pairs with canonical separations tended to have slightly longer half-lives for both downstream and upstream genes in all three data sets, but the effect was not consistently statistically significant (data not shown). We concluded that there is probably a small effect, but that spacing is not a major determinant of mRNA half-lives, and that transcript stability is unlikely to explain the prevalence of overlapping start and stop codons.

Overall, we argue that canonical overlaps form by neutral deletion and are maintained by selection against greater overlaps. However, changes to the spacing are likely accompanied by changes to the translation initiation rates of the downstream gene (e.g., switching to a new Shine-Dalgarno sequence or modifying translational coupling). We would expect these changes to expression levels to be under selection. Indeed, in laboratory experiments, the expression level of the *lac* operon evolves to optimality in a few hundred generations (Dekel and Alon 2005). Thus, changes in operon spacing may well be adaptive. Alternatively, the changes could be slightly deleterious, and other mutations around the start codon might quickly compensate for the effect of changes in spacing. Functional compensation has been observed in other kinds of cis-regulatory sequences (Ludwig *et al.* 2000).

Wide Spacings

Although operons tend to be closely spaced, highly expressed operons, as identified by codon adaptation, tend to be widely spaced (Eyre-Walker 1995; Ma *et al.* 2002). We confirmed with microarray data that highly expressed operons often have wide spacings of over 20 base pairs (see middle of Figure 4B). The correlation of spacings with mRNA levels is stronger than with codon adaptation (data not shown) – we suspect that this is because the empirical mRNA levels are less noisy estimates of expression levels than codon adaptation. The wide spacing of highly expressed operons seems surprising, both because it prevents translational coupling and because the additional RNA in highly expressed transcripts would waste the cell's resources. However, wide separations are particularly common among alternatively transcribed pairs that have internal promoters or terminators (Figure 4B).

To see if the sequences between the widely spaced operon pairs contain functional sequences, we examined phylogenetic footprints (conserved putative regulatory sequences) from McCue *et al.* (2002). 29% of the intergenic regions between known operon pairs that are separated by 50 or more bases contained phylogenetic footprints, which is statistically no different from the proportion of 38% for known alternative transcripts ($P > 0.5$, Fisher exact test). These conserved sequences averaged a total of 37 bases per pair (median 32), which is considerably larger than Shine-Dalgarno sequences. We examined the first 15 pairs with footprints for evidence of function, and found 5 attenuators or partial terminators, 3 internal promoters, 2 translation leader sequences, 1 small RNA not included in our database, 2 conserved REP sequences of unknown function, and only 2 cases with no information in the literature. Thus, most of these footprints correspond to functional regulatory sequences, and by extension, most widely spaced operons are subject to complex regulation. Consistent with this claim, widely spaced operons have significantly less similar expression patterns than do narrowly spaced operons, even if they are not known to be alternatively transcribed (Figure 4C; $P = 0.002$, t test). Instead, the distribution of similarity for widely spaced pairs is statistically indistinguishable from that for those that are known to be alternatively transcribed (Figure 4C; $P > 0.5$, t test). The correlation of complexity of regulation with expression levels suggests regulatory fine-tuning, because making unnecessary proteins would be more costly in materials or energy

or more deleterious in undesired protein activity if the proteins are highly expressed.

Death of Conserved Operons

Because few operons are conserved across all or even most bacteria (Itoh *et al.* 1999), it is clear that after operons form, many of them die. Operons could be lost by the deletion of one or both genes or else by splitting the operon apart. Here, we focus on cases where a conserved operon has split apart, so that *E. coli* retains both genes but they are not in the same operon. In particular, we ask by what mechanisms the operons "die," and whether certain types of operons are more likely to die.

Operon death by insertion, rearrangement, and replacement

To identify dead operons in *E. coli* K12, we first analyzed the predicted operons in its relatives. We considered conserved operon pairs that were predicted in more than one group of related bacteria and for which orthologous genes were present in *E. coli* (see Methods). To avoid cases of unclear orthology, we required both *E. coli* K12 genes to be the only members of their respective COGs (conserved orthologous groups, Tatusov *et al.* (2001)) in that genome. We then asked whether the *E. coli* K12 genes were in the same operon. Using these criteria, we identified 66 dead operon pairs that were split apart and 334 live operon pairs that were still co-transcribed.

When we categorized these dead operon pairs by their functional relatedness, we found 15 functionally related dead operons and 6 functionally unrelated genes that are probably growth-rate regulated (Table 3). Growth-related genes are often found together in operons even when there is no close functional relationship (Rogozin *et al.* 2002). Of the remaining dead operon pairs, 16 are functionally unrelated and 29 contain uncharacterized genes.

For 11 of the 66 dead operon pairs, the genes are still near each other on the chromosome. In these cases, the operon was probably destroyed by an insertion event. For example, the insertion of *ptr* discussed in a previous section appears to have both created the new *ptr-recB* operon pair and destroyed the ancestral *recCBD* operon. In the other 55 cases, the operon may have been destroyed by genome rearrangements. For example, the dead operon pair *yebI-yebL* is divergently transcribed in *E. coli*, which strongly suggests that the operon was destroyed by a local inversion.

When we investigated *lysA-dapF* and *ribD-ribE* in detail, we discovered another mechanism of operon death, which we term "replacement." *dapF* and *lysA* encode the final two steps of lysine synthesis, but *dapF*'s product is also essential for cell wall synthesis. In *E. coli* and some other species, *lysA* expression responds to lysine levels via a repressor that is encoded by the adjacent gene *lysR* (Karp *et al.* 2002). In many of its relatives, *lysA* is in an operon with *dapF* (see Figure 6A) and is not regulated by lysine (Martin *et al.* 1986). In phylogenetic analyses, *lysR*-associated *lysA* from diverse species constitute a distinct clade (data not shown), which we term *lysA2*. This

suggests horizontal transfer, as does the presence of both *dapF-lysA* and *lysR-lysA2* in some species. Thus, the parsimonious reconstruction is that *E. coli* acquired *lysR-lysA2* by horizontal gene transfer and then deleted *lysA* (Figure 6A). Consistent with deletion of *lysA*, the *E. coli dapF* operon, which has been experimentally characterized (Karp *et al.* 2002), retains genes on both sides of the missing *lysA*.

Similarly, *ribD* and *ribE* encode enzymes for the synthesis of riboflavin. As discussed by Vitreschak *et al.* (2002), many genomes have a second copy of *ribE* that lies outside of the ancestral operon, which we term *ribE2* (see Figure 6B). These *ribE2* genes form a distinct clade (data not shown), and *E. coli* has only *ribE2*. Again, the parsimonious reconstruction is that *ribD-ribE* died when *ribE* was replaced by the horizontally acquired *ribE2*.

Given the distinction between *lysA* and *lysA2*, or between *ribE* and *ribE2*, are these genuine dead operons or are they errors in our automated analysis? We feel that the choice is somewhat arbitrary. Because *lysA/lysA2* and *ribE/ribE2* are believed to have the same function, we prefer to consider *lysA-dapF* and *ribD-ribE* as dead operons. We also note that these HGT events required detailed phylogenetic analysis to uncover, and hence that previous analyses of operon destruction, which examined events across much larger phylogenetic distances (e.g., Itoh *et al.* (1999)), probably included similar cases.

Is operon death by replacement a common mechanism? To study this question systematically, we asked whether genes in dead operons were more likely than genes in live operons to have paralogs or to show evidence of HGT. We identified paralogs across 61 completely sequenced γ-Proteobacteria by using the COG database (Tatusov *et al.* 2001). Although we required all genes in both the dead and the live operons to lack paralogs in *E. coli*, we can still ask if paralogs are common in other organisms. On average, genes in dead operons had paralogs in 10.2% of the genomes, which is statistically indistinguishable from the rate of 9.4% for genes in live operons ($P > 0.5$, t test). We also built phylogenetic trees for all 118 genes in dead operons and compared the resulting trees to the species tree of Lerat *et al.* (2003) (see Methods). We found no evidence of HGT for most of these genes ($P > 0.05$ for 90.0% of genes, Kishino-Hasegawa test). Thus, we suspect that operon death generally occurs by genome rearrangements, or perhaps by insertions that are masked by later rearrangements, and not by replacement.

Rapid death of new and functionally coherent operons

Why do operons die? As a first step towards answering this question, we compared the death rates of different types of operons. For example, do operons that contain genes in different COG functional categories have different likelihoods of dying? For each of the 14 functional categories with at least 10 genes in the combined data set of live and dead operons, we performed a Fisher exact test, and to correct for multiple testing we used the false discovery rate with a cutoff of 0.05. We found unusually high survival rates for energy production and conversion operons (31 genes in surviving operons

vs. 0 in dead operons). We found unusually low survival rates for coenzyme metabolism operons (20 genes in surviving operons vs. 19 in dead operons) and for amino acid transport/metabolism operons (24 genes in surviving operons vs. 16 in dead operons). We speculate that the regulation of amino acid and coenzyme metabolism might evolve quickly because some bacteria, depending on their environmental niche, can import or metabolize these substances, whereas other bacteria must synthesize them.

We also found that new operons are much more likely to die than are older operons (Figure 5). However, even among ancient operons that are conserved between the β- and γ-Proteobacteria, 14% are shuffled apart in *E. coli* K12. Not surprisingly, operon pairs with conflicting COG function codes (Tatusov *et al.* 2001) are more likely to die (23% vs. 10%, $p < 0.005$, Fisher exact test). Ancient operons are also more likely to die if they are functionally incoherent (29% vs. 9%, $p < 0.005$, Fisher exact test). These results raise the question of why these functionally incoherent operons arose in the first place.

Operon Evolution Alters Gene Regulation

If operon formation is driven by gene expression, then operon formation should be associated with changes in the expression patterns of the constituent genes. Although the evolutionary ancestors of *E. coli* are not available for examination, we can study the expression patterns of orthologous genes in a related bacterium that diverged before the operon formed. We examined operon pairs that formed in the *E. coli* lineage soon after its divergence from *Shewanella oneidensis* MR-1, which we refer to as "not yet" operons in *Shewanella*. We compared the coexpression of these pairs to that of pairs that formed new operons just *before* the divergence (pairs that are "already" in operons in *Shewanella*). In *Shewanella*, the "not-yet" operon pairs are not coexpressed, while the "already" operon pairs are, not surprisingly, coexpressed (Figure 7A). Hence, operon formation has a major effect on gene expression patterns. Because bacterial gene regulation is complex and is generally believed to be under strong selection, this suggests that operon formation may be adaptive.

Operon death could also be adaptive. If it is, then the genes in the dead operons should have different expression patterns than they would if they were still co-transcribed. To see if this is the case, we compared the coexpression of conserved operon pairs to that of "dead" operons, of the same evolutionary age, that are split apart in *E. coli* K12 (see Methods). We found that dead operons were significantly less coexpressed than operons that were still alive, but significantly more coexpressed than random pairs (Figure 7B). The (modest) coexpression of dead operons might seem asymmetric or contradictory given our finding that not-yet operons are not coexpressed, but here we considered only conserved operons, as only conserved dead operons can be identified with confidence. As the more conserved operons tend to be more functionally coherent (de Daruvar *et al.* 2002), it is not surprising that these conserved operons retained some coexpression after the death of the operon. However, even dead operons that were functionally related had little coexpression in *E. coli*: the mean coexpression was 0.36, which was not significantly higher than the mean of 0.22 for the other

dead operon pairs ($P > 0.1$, t test).

Adaptive destruction of operons

We identified two dead operons that encode well-characterized protein complexes – *recC-recB* and *ruvC-ruvA*. Perhaps coincidentally, both RecBCD and RuvABC act in the repair of double-stranded DNA breaks by homologous recombination. We expected that both *recC-recB* and *ruvC-ruvA* would be tightly coregulated, but instead we found that neither pair is strongly coexpressed in *E. coli* ($r = 0.29$ and 0.27, respectively). In contrast, both pairs are strongly coexpressed in *Shewanella*, where each pair's genes are adjacent and are predicted to be in the same operon ($r = 0.80$ and 0.79, respectively). Because the regulation of *recC* and *recB-recD* in *E. coli* has not been characterized, we do not know the regulatory consequence of operon death in that case, but the regulation of *ruvA* has been studied.

As shown in Figure 6C, in *E. coli* and close relatives, *ruvAB* is repressed by LexA, so that expression is induced by DNA damage (Shinagawa *et al.* 1988; Merlin *et al.* 2002; Erill *et al.* 2003, 2004). In contrast, *E. coli ruvC* is not induced by DNA damage and is not predicted to bind LexA (Takahagi *et al.* 1991; Erill *et al.* 2003). *E. coli*'s closest relatives appear to have the same regulation (Erill *et al.* 2003). In more distant relatives, none of these genes appears to be regulated by LexA (Erill *et al.* (2003) and our analysis, data not shown). However, in the α-Proteobacterium *S. meliloti*, the *ruvABC* operon is repressed by LexA (Erill *et al.* 2004). The parsimonious reconstruction is that ancient γ-Proteobacteria had *ruvCAB* in a single operon that was not regulated by LexA. Then, in the *E. coli* lineage, *ruvA* acquired its own LexA-regulated promoter and the *ruvC-ruvA* operon pair died. This change appears to have been adaptive, both because inducing *ruvAB* in response to DNA damage makes biological sense and because the regulation has been conserved.

However, it is not clear why *E. coli ruvC* is not induced by DNA damage. RuvC acts together with RuvAB and all three are regulated by LexA in other organisms. It is possible that RuvC has some other function that is constitutively required, but all three *ruv* genes are reported to give the same phenotype when knocked out (Kuzminov 1999). Therefore, it is tempting to speculate that regulating *ruvC* by LexA would be adaptive for *E. coli*. If this is correct, then the death of *ruvC-ruvA* was adaptive, but placing a new LexA-regulated promoter upstream of *ruvC* and maintaining the operon would have been more adaptive.

The case of *lysA-dapF* discussed in a previous section appears to be another case of adaptive operon destruction (Figure 6A). We argue that *lysA2* was acquired so that it could be regulated by LysR, and that the Enterobacteria then lost the original *lysA*. Because *dapF* is essential (Gerdes *et al.* 2003), the ancestral *dapF-lysA* operon could not have been not lysine regulated. Indeed, in a species that retains the *dapF-lysA* operon, *lysA* activity does not respond to lysine levels (Martin *et al.* 1986).

However, it is not clear why some species maintain both the constitutive *lysA* and the lysine-regulated *lysA2*. We also wonder why *lysA* became part of the *dapF* operon in an ancient γ-Proteobacterium. The close functional relationship of the two genes suggests that this served some purpose, and was not simply to regulate both genes by growth rate, as proposed to explain the conservation of some functionally incoherent operons (Rogozin *et al.* 2002). It is possible is that the organism was not exposed to external lysine and did not need to regulate *lysA* and *dapF* independently.

Discussion

Adaptive evolution of operons

Several of our findings suggest that the evolution of operons is adaptive. First, both the birth and the death of operons lead to large changes in expression patterns. Gene expression is believed to be under strong selection in *E. coli*: the majority of known regulatory sequences are highly conserved (McCue *et al.* 2002), genes are often regulated by multiple transcription factors (Karp *et al.* 2002), gene expression patterns show convergent evolution in the wild (Gall *et al.* 2005) and in laboratory experiments (Cooper *et al.* 2003), and gene expression levels can evolve to optimality in laboratory experiments (Dekel and Alon 2005). Thus, we argue that these changes in operon structure are also under strong selection. Second, some operons acquire several new genes in a relatively short period of evolutionary time; this accelerated evolution suggests positive Darwinian selection. Third, highly expressed operons are particularly likely to use wide spacings with complex regulation; this can be explained by strong selection to avoid making large amounts of unnecessary protein. Finally, many new operons contain ORFan genes, which may be a mechanism for allowing the expression of newly imported genes.

These results contrast a previous suggestion that selection to maintain operon structure is weak, so that genome rearrangements cause neutral or slightly deleterious turnover of operon structure (Itoh *et al.* 1999). The two explanations of neutral and adaptive evolution are not exclusive – the formation and death of operons could be nearly neutral in some cases and highly adaptive in others. Intensive analysis of specific operons will be required to distinguish these possibilities. We have discussed two examples, the replacement of *lysA* in the *dapF* operon with a LysR-regulated *lysA* and the regulation of *ruvA-ruvB* by LexA, in which the change in regulation appears to be adaptive.

Both *lysA* and *ruvA-ruvB* were probably in ancestral operons that contained essential genes (Figure 6). Similarly, in the second case of operon death by replacement that we identified, *ribD-ribE*, it again appears that the ancestral operon contained an essential gene (*nusB*; see Figure 6B) and hence must have been constitutive or growth-regulated. Thus, the turnover of operon structure may accompany switching between constitutive and inducible expression. Although constitutive expression may seem deleterious, it could be neutral if the capability is often required, and could

be adaptive if lack of the protein would create delays in growth until large amounts of new protein were synthesized. Such "just in case" or "standby" expression of proteins that are not required for rapid growth appears to be be common in the soil bacterium *Bacillus subtilis* (Fischer and Sauer 2005).

Non-optimal evolution of operons

If operon evolution is adaptive, then do operons reach an optimal arrangement? In general, we do not know what would make a gene regulatory system optimal, and we often do not know what the criteria are. However, for inducible biosynthetic capabilities such as amino acid synthesis, a plausible design goal is to produce product quickly, so that growth can resume, while also minimizing the amounts of enzyme synthesized. For simple (linear) metabolic pathways, optimal design by this criterion requires differential timing of gene expression, with earlier induction for genes that encode the first steps in the pathway (Zaslaver *et al.* 2004). This suggests that placing genes in operons may prevent fine-tuning of the timing of induction, and could be inherently suboptimal. Operons might exist despite this disadvantage because they facilitate the evolution of co-regulation (Price *et al.* 2005*b*).

Constraints on how operon evolve also likely lead to non-optimal operons. We have already discussed how the introduction of a LexA-regulated promoter between *ruvC* and *ruvA-ruvB* was adaptive, but the regulation of all three genes by LexA would, we imagine, be more adaptive. More broadly, operons containing native genes often from by deletion, so that the two genes in the new operon need to have been near each other and on the same strand. Although there is some tendency for genes with similar functions or expression patterns to cluster together on a larger scale than operons (Pal and Hurst 2004; Allen *et al.* 2003), it seems unlikely that the optimal partners for new operons will be found near each other. Thus, we would not expect new operons that formed by deletion to be optimal.

Similarly, the formation of new native-ORFan operon pairs may be driven by selection for the presence of the ORFan rather than for optimal regulation. This is because selection for the presence or absence of a gene should be much stronger than selection on its regulation. As the insertion of any particular ORFan is probably very rare, the operon that forms and becomes fixed in the population might not be the optimal one. Furthermore, optimal regulation of the ORFan may not be available from any of the pre-existing native promoters.

Consequences for genome annotation

On a practical note, the lack of coexpression of "not-yet" operons extends previous observations that many new operons are functionally not coherent (de Daruvar *et al.* 2002; Price *et al.* 2005*b*). Our

observation that newer operons have high death rates also confirms that the genes in them may not be functionally related. Thus, we caution that the presence of a gene in an operon is not a strong indicator of its function unless the operon is well-conserved. Of the new operon pairs that are new to the Enterobacteria and contain two annotated genes, only four out of nine have related functions according to COG (Tatusov *et al.* 2001). This statistic probably overstates the chance of two genes in a new operon being related, as there are many uncharacterized genes, and it is more likely that both genes in an operon will be characterized if they have closely related functions.

As examples of how over-reliance on new operons can lead to incorrect annotation, consider *flhE* and *btuE*. We previously mentioned the Enterobacterial operon *flhBAE*, which appears to us have led to the unwarranted annotation of *flhE* as a flagellar protein. Another new operon unique to the Enterobacteria, *btuCED*, includes two components of the vitamin B12 ABC transporter and also *btuE*, which is not required for vitamin B12 transport (Rioux and Kadner 1989). Indeed, *btuE* belongs to the glutathione peroxidase family and is not homologous to ABC transporters (data not shown). Nevertheless, *btuE* is consistently mis-annotated as a vitamin B12 ABC transporter in sequence databases.

Most automated predictions of gene function are not affected by these issues because they use only highly conserved operons (Overbeek *et al.* 1999; Huynen *et al.* 2000), but operon predictions based on the distance between adjacent genes have been used to aid in function prediction (Strong *et al.* 2003). The latter method was validated by testing it against textual gene annotations, but the over-annotation of new operons, as with *flhE* and *btuE*, could possibly have exaggerated its benefit. In any case, we suspect that automated function predictions could be improved by down-weighting evidence from the newest operons.

Methods

Operons

For over 100 genomes, we predicted whether pairs of adjacent genes that are on the same strand are co-transcribed based on the intergenic distance between them, whether orthologs of the genes are near each other in other genomes, and the genes' predicted functions (Price *et al.* 2005*a*). Both the predictions and the underlying features are available at http://www.microbesonline.org/operons (Alm *et al.* 2005). These operon predictions are over 80% accurate on pairs of genes in diverse prokaryotes, based on databases of known operons and on analysis of microarray data. For analyses of operon spacing, we used a database of known *E. coli* K12 operons (Karp *et al.* 2002) instead of predictions.

Evolutionary History and Ages of Genes and of Operon Pairs

We used the evolutionary analysis of Price *et al.* (2005*b*). Briefly, we divided the sequenced prokaryotes into groups at varying evolutionary distances from *E. coli* K12 – (1) other strains of *E. coli* and *Shigella*, (2) *Salmonella* species, (3) other Enterobacteria, (4) allied γ-Proteobacteria (*Haemophilus*, *Pasteurella*, *Vibrio*, and *Shewanella* species), (5) distant γ-Proteobacteria (*Pseudomonas*, *Xanthomonas*, and *Xylella* species), (6) β-Proteobacteria, (7) other Proteobacteria, and (8) non-Proteobacteria, including Archaea. *E. coli* K12 genes that had at least one homolog from BLASTp or from COG (Tatusov *et al.* 2001) in each of the groups 1-7 were considered native. (Genes were assigned to COGs by reverse position-specific BLAST (Schaffer *et al.* 2001) against CDD (Marchler-Bauer *et al.* 2003).) Genes that had no homologs any of groups 6-8 were considered ORFans, and the most distant group that did contain a homolog of each ORFan was used an estimate of the ORFan's evolutionary age. Similarly, for each pair of adjacent *E. coli* K12 genes that were predicted to be in the same operon, we asked which groups of genomes contained homologous operons. To account for the frequent reordering of genes in operons (Itoh *et al.* 1999), we did not require the homologs to be adjacent, but only that they be in the same predicted operon. Operons of age 4 or less were considered new, and operons present in each of groups 1-7 were considered old.

For the *E. coli* genes in new native-native operons (Table 1) and for genes in dead operons, we also tested the protein sequences for evidence of horizontal gene transfer. Specifically, we compared the phylogenetic tree inferred from the protein sequences to the species tree of Lerat *et al.* (2003). From orthologs (bidirectional best BLASTp hits) among the species in the species tree, we constructed protein sequence alignments with ClustalW (Thompson *et al.* 1994) and the BLOSUM80 matrix, we removed columns containing gaps, and we constructed phylogenetic trees with TreePuzzle 5.1 (Schmidt *et al.* 2002). To see if the resulting tree was consistent with the species tree, we used the one-sided Kishino-Hasegawa test recommended by Goldman *et al.* (2000). High p-values indicate accepting the species tree.

To identify dead operons, we first enumerated all pairs of *E. coli* K12 genes that were orthologous to predicted operon pairs from any other genome. Here for orthologs we used either bi-directional BLASTp hits or genes in the same COG. We retained pairs that were predicted to be in an operon in two consecutive groups (e.g., both a group 4 genome and a group 5 genome). Of these pairs, those that were adjacent in *E. coli* K12 and predicted to be in the same *E. coli* operon were considered "live" operons; pairs that were far apart were considered to be "dead" operons; and other pairs were considered ambiguous and discarded. We further required that both genes be a unique member of their COG in *E. coli* K12. This requirement was necessary because if the ancient operon AB died, and gene B had a paralog B', then both AB and AB' would otherwise appear to be dead operons.

For the coexpression analysis of dead operons (Figure 7B), we further required that both *E. coli* genes have bidirectional best hits in *S. oneidensis* MR-1, that those bidirectional best hits be adjacent, and that the orthologs be predicted to be in the same operon. We wished to compare dead and live operons of the same evolutionary age, so for both dead and live operons we used only those present

in groups 4 (which includes *Shewanella*) and 5 but not in more distant groups.

Microarray Data

To quantify the similarity of two gene's expression patterns, we used the Pearson correlation of their normalized log ratios across microarray experiments. For *E. coli* K12, 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. For *S. oneidensis* MR-1, we used data on salt stress (Liu *et al.* 2005), heat shock (Gao *et al.* 2004), cold shock, strontium stress, and high and low pH stress (Z. He, Q. He, and J. Zhou, unpublished data).

To quantify gene expression (mRNA) levels in *E. coli* K12, we used the average foreground intensity across arrays and across both red and green channels. We used intensities rather than more direct measures of expression levels, which can be obtained from microarray experiments where an mRNA sample is compared to genomic DNA, because only a few of the experiments were of that type. Within the "genomic control" experiments, the average across replicates of the intensity in the mRNA channel was highly correlated with the average log-ratio between the mRNA and genomic DNA channels (the Spearman rank correlation was 0.84).

Testing for Accelerated Evolution

As shown in Figure 3, new operon pairs are often adjacent to other new operon pairs of the same age. To see how often this would occur under completely random evolution, we used the fraction p_i of operon pairs that have age i, the fraction q of operon pairs that are adjacent to the next (downstream) operon pair, and the total number N of operon pairs. (Operon pairs AB and BC within the operon ABC are adjacent, while the standalone operon AB is not adjacent to another operon pair.) Under random evolution, the number of adjacent new operon pairs of the same age would be $N \cdot q \cdot \sum_{i=0}^{4} p_i^2$, and the number of adjacent new operon pairs would be $N \cdot q \cdot \sum_{i=0}^{4} p_i$.

Modifications to Pre-existing Operons

To identify and classify the new operon pairs that arose by modification to pre-existing operons, we performed an automated analysis (shown in Figure 3B) and also inspected the results manually (Supplementary Table 1). The automated analysis relied on comparing the ages of the new operon pair to the age of adjacent or surrounding operon pairs. For example, if the operon pair AB was prepended to the preexisting operon BC, then AB should be newer than BC. If the operon ABC arose from inserting the gene B into the preexisting operon AC (or from replacing gene D in the operon ADC), then both AB and BC should be newer than AC. If the operon ABCD formed by joining two

preexisting operons AB and CD, then BC should be newer than either AB or CD. To avoid confusion due to paralogs, we only considered pairs where the age using homologs from COG matched the age using putative orthologs (bidirectional BLASTp hits). Manual inspection was performed with the MicrobesOnline comparative genomics browser at http://microbesonline.org (Alm *et al.* 2005), with careful attention to cases where potential orthologs were not identified automatically.

Statistics

Statistical tests were conducted with the R open-source statistics package (http://r-project.org).

Acknowledgments

We thank Zhili He, Qiang He, and Jizhong Zhou for pre-publication access to microarray data. This work was supported by a grant from the DOE Genomics:GTL program (DE-AC03-76SF00098). A.P.A. would also like to acknowledge the support of the Howard Hughes Medical Institute.

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.

Alm,E.J., Huang,K.H., Price,M.N., Koche,R.P., Keller,K., Dubchak,I.L. and Arkin,A.P. (2005) The MicrobesOnline web site for comparative genomics. *Genome Res.,* **15**, 1015–22.

Amundsen,S.K., Taylor,A.F., Chaudhury,A.M. and Smith,G.R. (1986) recD: the gene for an essential third subunit of exonuclease V. *Proc. Natl. Acad. Sci. USA,* **83**, 5558–62.

Bernstein,J.A., Khodursky,A.B., Lin,P.H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. USA,* **99**, 9697–702.

Butland,G., Peregrin-Alvarez,J.M., Li,J., Yang,W., Yang,X., Canadien,V., Starostine,A., Richards,D., Beattie,B., Krogan,N., Davey,M., Parkinson,J., Greenblatt,J. and Emili,A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature,* **433**, 531–7.

Conlin,C.A. and Miller,C.G. (2000) opdA, a Salmonella enterica serovar Typhimurium gene encoding a protease, is part of an operon regulated by heat shock. *J. Bacteriol.,* **182**, 518–21.

Cooper,T.F., Rozen,D..E. and Lenski,R.E. (2003) Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. *Proc. Natl. Acad. Sci. USA,* **100**, 1072–7.

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.

Dekel,E. and Alon,U. (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature,* **436**, 588–92.

Erill,I., Escribano,M., Campoy,S. and Barbe,J. (2003) In silico analysis reveals substantial variability in the gene contents of the gamma proteobacteria LexA-regulon. *Bioinformatics,* **19**, 2225–36.

Erill,I., Jara,M., Salvador,N., Escribano,M., Campoy,S. and Barbe,J. (2004) Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. *Nucleic Acids Res.,* **32**, 6617–26.

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

Eyre-Walker,A. (1995) The distance between Escherichia coli genes is related to gene expression levels. *J. Bacteriol.,* **177**, 5368–9.

Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics,* **15**, 759–62.

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

Gall,T.L., P., Escobar-Paramo,P., Picard,B. and Denamur,E. (2005) Selection-driven transcriptome polymorphism in Escherichia coli/Shigella species. *Genome Res.,* **15**, 260–8.

Gao,H., Wang,Y., Liu,X., Yan,T., Wu,L., Alm,E., Arkin,A., Thompson,D.K. and Zhou,J. (2004) Global transcriptome analysis of the heat shock response of Shewanella oneidensis. *J. Bacteriol.,* **186**, 7796–803.

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.

Huynen,M., Snel,B., 3rd,W.L. and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.,* **10**, 1204–10.

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.

Jacob,F. and Monod,J. (1961) On the regulation of gene activity. *Cold Spring Harbor Symp. Quant. Biol.,* **26**, 193–211.

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.

Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene,* **234**, 187–208.

Kuzminov,A. (1999) Recombinational repair of dna damage in escherichia coli and bacteriophage lambda. *Microbiol. Mol. Biol. Rev.,* **63**, 751–813.

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.

Liu,Y., Gao,W., Wang,Y., Wu,L., Liu,X., Yan,T., Alm,E., Arkin,A., Thompson,D.K., Fields,M.W. and Zhou,J. (2005) Transcriptome analysis of Shewanella oneidensis MR-1 in response to elevated salt conditions. *J. Bacteriol.,* **187**, 2501–7.

Ludwig,M.Z., Bergman,C., Patel,N.H. and Kreitman,M. (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature,* **403**, 564–7.

Ma,J., Campbell,A. and Karlin,S. (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.,* **184**, 5733–45.

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.

Martin,C., Cami,B., Borne,F., Jeenes,D.J., Haas,D. and Patte,J.C. (1986) Heterologous expression and regulation of the lysA genes of Pseudomonas aeruginosa and Escherichia coli. *Mol. Gen. Genet.,* **203**, 430–4.

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.

Merlin,C., McAteer,S. and Masters,M. (2002) Tools for characterization of Escherichia coli genes of unknown function. *J. Bacteriol.,* **184**, 4573–81.

Minamino,T., Iino,T. and Kutuskake,K. (1994) Molecular characterization of the Salmonella typhimurium flhB operon and its protein products. *J. Bacteriol.,* **176**, 7630–7.

Minamino,T. and Macnab,R.M. (1999) Components of the Salmonella flagellar export apparatus and classification of export substrates. *J. Bacteriol.,* **181**, 1388–1394.

Mira,A., Ochman,H. and Moran,N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.,* **17**, 589–96.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics,* **18 Suppl. 1**, S329–36.

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.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA,* **96**, 2896–901.

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

Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005*a*) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.,* **33**, 880–92.

Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005*b*) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.,* **15**, 809–19.

Ragan,M.A. 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.

Rioux,C.R. and Kadner,R.J. (1989) Vitamin B12 transport in Escherichia coli K12 does not require the btuE gene of the btuCED operon. *Mol. Gen. Genet.,* **217**, 301–8.

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.

Selinger,D.W., Saxena,R..M., Cheung,K..J., Church,G.M. and Rosenow,C. (2003) Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res.,* **13**, 216–23.

Shinagawa,H., Makino,K., Amemura,M., Kimura,S., Iwasaki,H. and Nakata,A. (1988) Structure and regulation of the escherichia coli ruv operon involved in DNA repair and recombination. *J. Bacteriol.,* **170**, 4322–9.

Strong,M., Mallick,P., Pellegrini,M., Thompson,M.J. and Eisenberg,D. (2003) Inference of protein function and protein linkages in mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biol.,* **4**, R59.

Swain,P.S. (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J. Mol. Biol.,* **344**, 965–76.

Takahagi,M., Iwasaki,H., Nakata,A. and Shinagawa,H. (1991) Molecular analysis of the Escherichia coli ruvC gene, which encodes a Holliday junction-specific endonuclease. *J. Bacteriol.,* **173**, 747–53.

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.

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.

Vitreschak,A.G., Rodionov,D.A., Mironov,A.A. and Gelfand,M.S. (2002) Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.,* **30**, 3141–51.

Wolf,Y., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.,* **11**, 356–72.

Yoshikawa,A., Isono,S., Sheback,A. and Isono,K. (1987) Cloning and nucleotide sequencing of the genes rimI and rimJ which encode enzymes acetylating ribosomal proteins S18 and S5 of Escherichia coli K12. *Mol. Gen Genet.,* **209**, 481–8.

Yu,J.S., Madison-Antenucci,S. and Steege,D.A. (2001) Translation at higher than an optimal level interferes with coupling at an intercistronic junction. *Mol. Microbiol.,* **42**, 821–34.

Zaslaver,A., Mayo,A.E., Rosenberg,R., Bashkin,P., Sberro,H., Tsalyuk,M., Surette,M.G. and Alon,U. (2004) Just-in-time transcription program in metabolic pathways. *Nat. Genet.,* **36**, 486–91.

Table 1: Mechanism of formation for new native-native operon pairs. For each operon pair that is unique to the Enterobacteria and has non-adjacent orthologs in two or more species of *Vibrio*, we classified the pair as arising by deletion of intervening genes or by a rearrangement. On inspection, three additional pairs (not shown) arose by insertion of a horizontally transferred gene next to a native gene. The gene names for the new operon pair are in bold; the numbers indicate the spacing between the genes. For each pair, we also show the gene order in a representative member of the *Vibrio* (Vc = *V. cholerae*; Vp = *V. parahaemolyticus* RIMD 2210633; Vv = *V. vulnificus CMCP6*). Parentheses indicate genes on the opposite strand, numbers again indicate spacing, and ellipses (. . .) indicate separation by > 20 kb or placement on another chromosome and do not imply ordering. For some pairs, there is evidence that they are co-transcribed in *E. coli* or in its close relative *Salmonella typhimurium* (Karp *et al.* 2002; Conlin and Miller 2000). The co-transcription of *recB-ptr* is unclear because *ptr* may have its own promoter (Amundsen *et al.* 1986), but the genes overlap and have similar expression patterns. The co-transcription of *rimJ-yceH* is likely but not certain because only part of *yceH* was present in the clone from which the *rimJ* promoter was studied (Yoshikawa *et al.* 1987). Finally, the rightmost column shows the similarity of expression patterns for the putative new operon pair, as measured by the Pearson correlation coefficient on microarray data from *E. coli* K12 (see Methods).

***E. coli* K12**	***Vibrio* Homolog**	**Known?**	**Similarity**
Deletion of Intervening DNA			
ybbO 25 ***ybbN***	*Vc:* ***ybbO*** 75 (*VC0978*) 172 ***ybbN***	–	0.24
prlC 8 ***yhiQ***	*Vp:* ***prlC*** 67 (*asnC* -43 *GGDEF*) 90 ***yhiQ***	Yes	-0.06
serB 49 ***radA***	*Vc:* ***serB*** 205 (*VC2344*) 128 ***radA***	Yes	–
ygiF 23 ***glnE***	*Vp:* ***ygiF*** 63 (*VP0422*) -24 *MCP* 109 ***glnE***	Yes	0.06
pdlB 8 ***yigL***	*Vv:* ***pldB*** 181 ***yigL***	–	–
btuB -68 ***murI***	*Vp:* ***btuB*** 79 *ATPase* 41 ***murI***	–	0.52
Rearrangement of Two Native Genes			
recC 176 ***ptr*** -8 ***recB***	*Vp: recC* 319 ***recB*** . . . ***ptr***	No?	0.47
malK 72 ***lamB***	*Vc:* ***malK*** . . . ***lamB***	Yes	0.69
rimJ 11 ***yceH***	*Vc:* ***rimJ*** . . . ***yceH***	Yes?	0.60
ybjU -20 ***ybjT***	*Vc:* ***ybjU*** . . . ***ybjT***	–	0.54

Table 2 : Mechanisms for forming the canonical spacing. Known operon pairs that invented or lost the canonical spacing (an overlap of 1 or 4 nucleotides) in the *E. coli* lineage were identified by comparison to *Salmonella* and *Yersinia* species (*Yersinia* is a more distant relative). For each pair, we show an alignment of the DNA sequences around the stop and start codons from *E. coli* K12 ("Ec") and *Salmonella typhimurium* LT2 ("St"). The stop codon of the upstream gene has wavy underlines, the start codon of the downstream gene is underlined, and conserved nucleotides are capitalized. Because *cysNC* and *cstC-astA* have larger separations in other Enterobacteria, we suspect that the common ancestor of *Escherichia* and *Salmonella* formed the canonical separation and that a larger overlap then formed in *Salmonella*. We also identified 9 operon pairs with canonical but different spacings in *E. coli* and *Salmonella*, which are not shown.

Pair	**Separations**	**Alignment**
Loss of canonical spacing in *E. coli*		
rfaF rfaC	Ec: 5	TGAcggaTG
	St: -1	TGA----TG
Creation of canonical spacing in *E. coli*		
xseB ispA	Ec: -1	TA-ATG
	St: 1	TAAATG
dnaN recF	Ec: -1	TA- -148- ATG
	St: 148	TAG 148nt ATG
Creation and then loss in in *S. typhimurium*		
cysN cysC	Ec: -1	AAAtAATGGCGCTGCATGA
	St: -12	AAAc-ATGGCGCTGCATGA
cstC astA	Ec: -3	ATGatGGTcA
	St: -9	ATGcgGGTgA

Table 3: Dead operon pairs comprising functionally related genes or likely growth-regulated genes.

Functionally related pairs (15)	
ribD-ribE	Riboflavin synthesis
lipB-lipA	Lipoate modification
nadA-nadC	Synthesis of NAD
moaA-mobA	Molybdenum cofactor synthesis
flgA-flgM	Flagellar synthesis
ruvC-ruvA	Homologous recombination
thiD-thiE	Thiamin synthesis
tyrA-aroA	Tyrosine synthesis
recC-recB	Homologous recombination
thyA-folA	Synthesis of formyl-THF
lysA-dapF	Lysine synthesis
argG-argH	Arginine synthesis
sbp-cysU	Sulfate transport
infA-rpsM	Protein synthesis
rplY-pth	Protein synthesis
Likely growth-rate regulated pairs (6)	
prsA-pth	Protein synthesis & PRPP synthesis
prsA-rplY	PRPP synthesis & protein synthesis
argS-ftsN	Protein synthesis & cell division
lepB-rnc	Signal peptidase & RNAse
rpoC-rpsL	rRNA synthesis & protein synthesis
rplI-dnaB	Protein synthesis & DNA synthesis

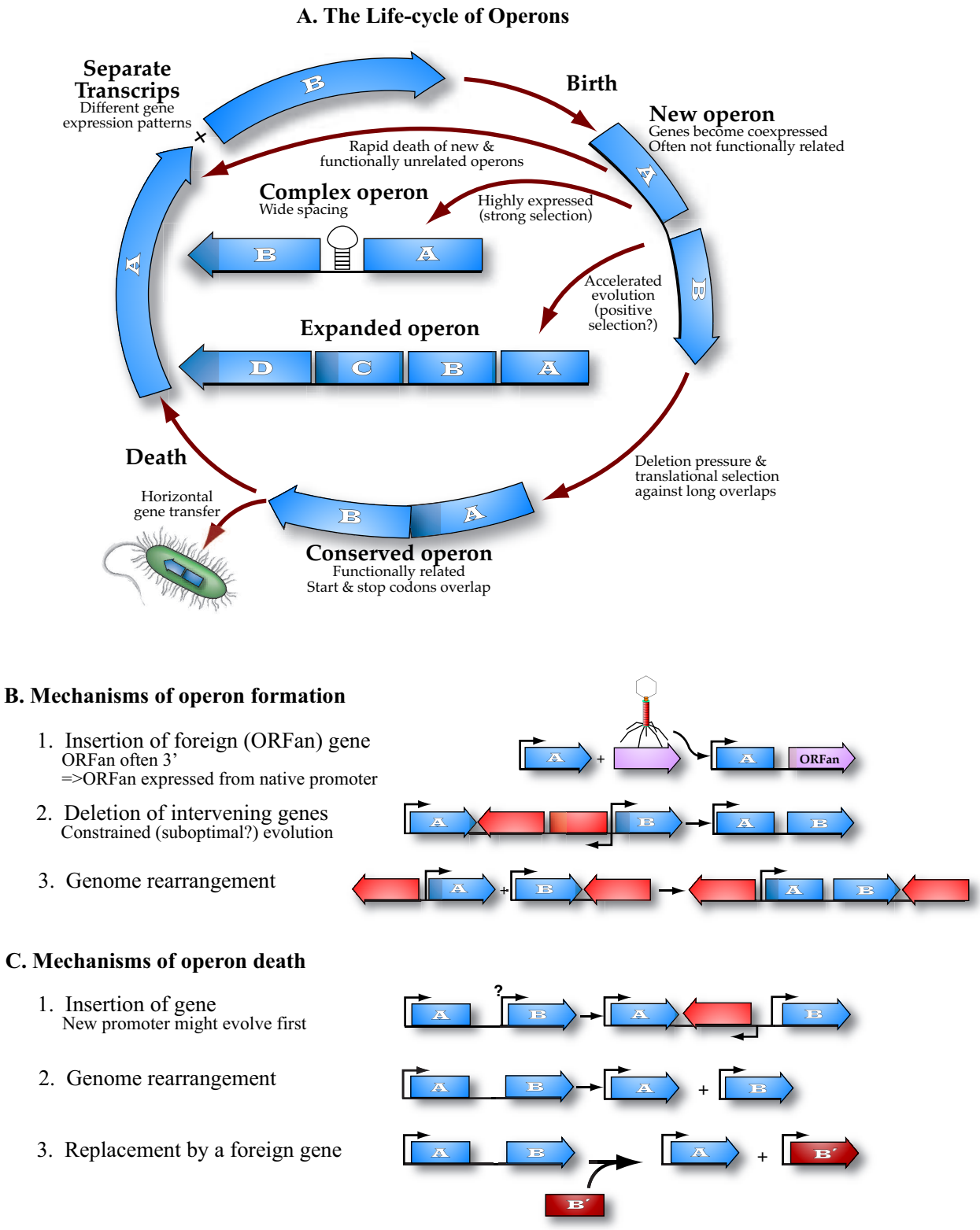

Figure 1: A model for the life-cycle of operons, and the major mechanisms of operon formation and destruction.

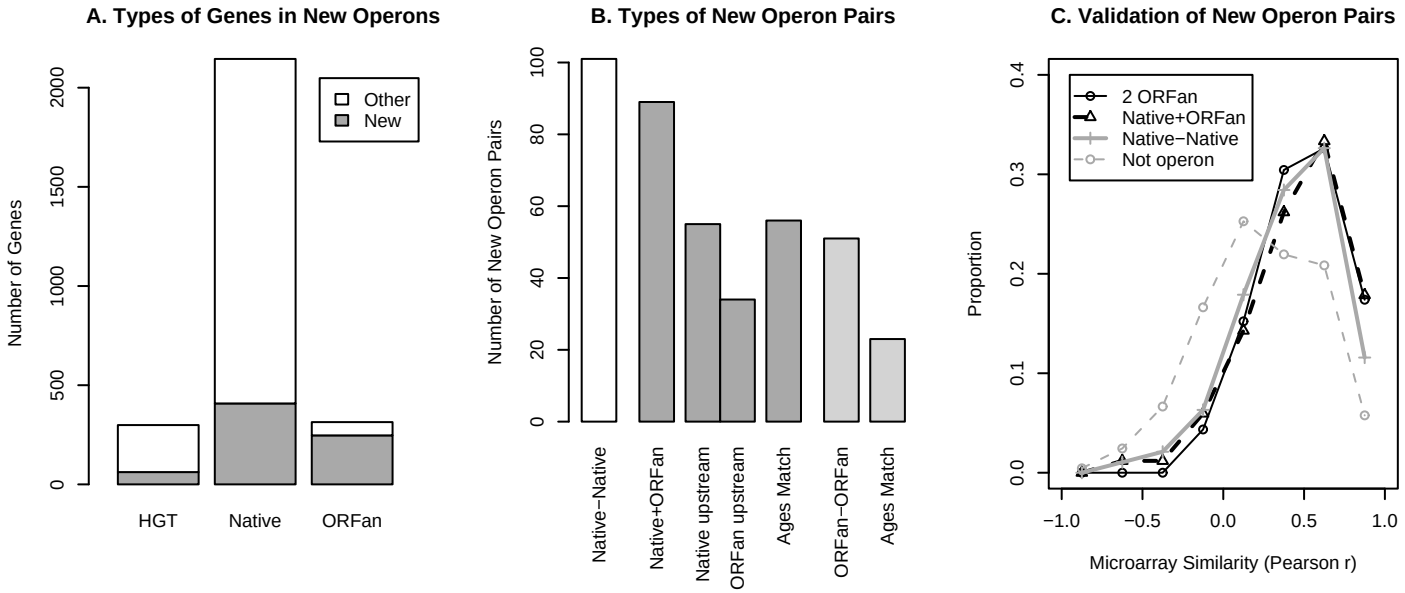

Figure 2: New operons often combine a native gene with an "ORFan" gene that is found only in *E. coli* and close relatives. (A) Types of genes in new operon pairs and in other operon pairs. The enrichment for ORFans in new operon pairs is highly significant ($P < 10^{-15}$, Fisher exact test). (B) Types of new operon pairs. Only new operon pairs involving native and ORFan genes are shown (there are relatively few HGT genes in the new operons). Within the native-ORFan pairs, we show how often the native gene is upstream of the ORFan, or vice versa. For both the native-ORFan and ORFan-ORFan pairs, we show how often the evolutionary age of the ORFan(s) match those of the operon. (C) Validation of predicted new operon pairs of each of the three major types. We quantified the similarity of expression patterns in microarray data using the Pearson correlation. As a negative control, we also tested non-operon pairs (adjacent genes on the same strand that are known not to be cotranscribed) from Karp *et al.* (2002).

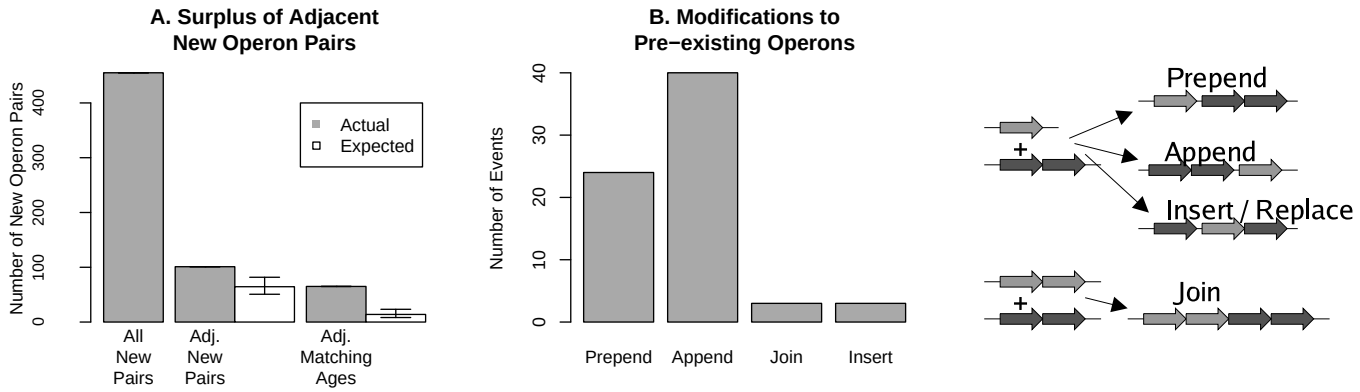

Figure 3: Accelerated evolution of some operons. (A) New operon pairs are more likely to be adjacent to each other than expected by chance. The surplus of adjacent pairs of the same age is particularly striking. The error bars show 95% confidence intervals from a χ^2 test of proportions. The model for neutral expectation is detailed in the Methods. (B) The frequency of different types of modifications to pre-existing operons. The excess of append over prepend pairs is not quite statistically significant ($P = 0.06$, binomial test).

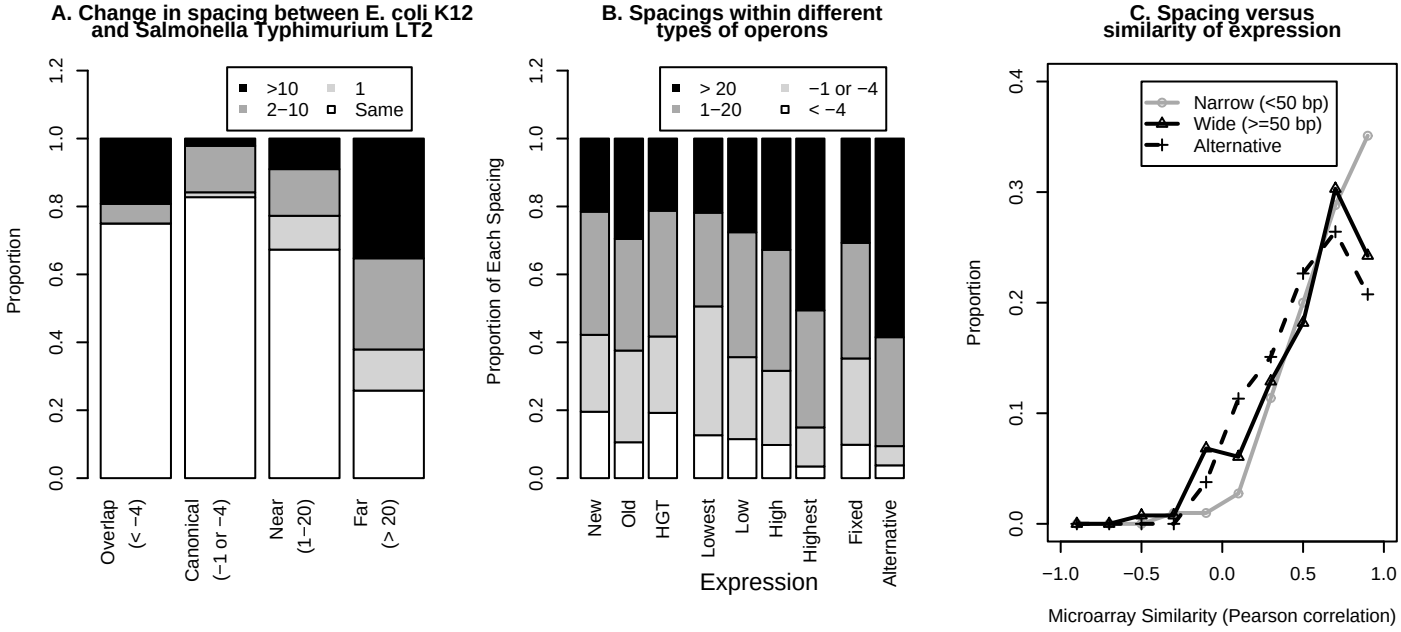

Figure 4: Spacings between adjacent genes in the same operon. (A) Known operon pairs in *E. coli* often have different spacing than the orthologous operon in *Salmonella typhimurium* LT2. For each class of spacing in *E. coli* (x axis), a vertical bar shows the proportion with various amounts of change. (B) The frequency of different types of spacings for operon pairs classified by their evolutionary history (left), their expression level as estimated from microarray data (middle), or whether the operon has an alternative transcript (right). Because operon predictions rely heavily on spacing, only known *E. coli* operons were used. (C) The distribution of microarray similarity for known operon pairs spaced by less than 50 bp or by more than 50 bp and for alternatively transcribed operon pairs. Operons that are known to be alternatively transcribed were excluded from the "narrow" and "wide" sets.

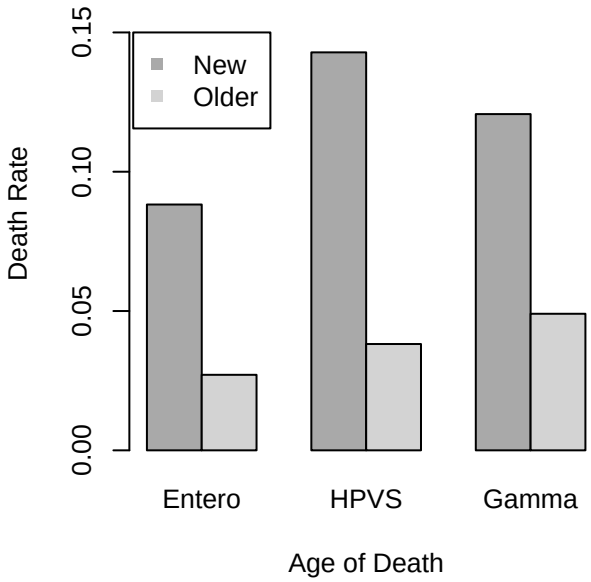

Figure 5: New operons die at faster rates. Ancestral operons were identified by their presence in two or more consecutive groups of relatives, and were considered dead if they were no longer in the same operon in *E. coli* K12. The death rate at a given "age" is the proportion of operons that are present in that group but not in more recent relatives. Here, an operon is considered new at the time of its death if it is present only in the minimum two consecutive groups. In increasing order, the ages are "Entero" – Enterobacteria besides *E. coli* and *Salmonella*; "HPVS" – *Haemophilus*, *Pasteurella*, *Shewanella*, and *Vibrio* species; and "Gamma" – other γ-Proteobacteria. All differences between new and older operons were statistically significant ($P < 0.05$, Fisher exact test).

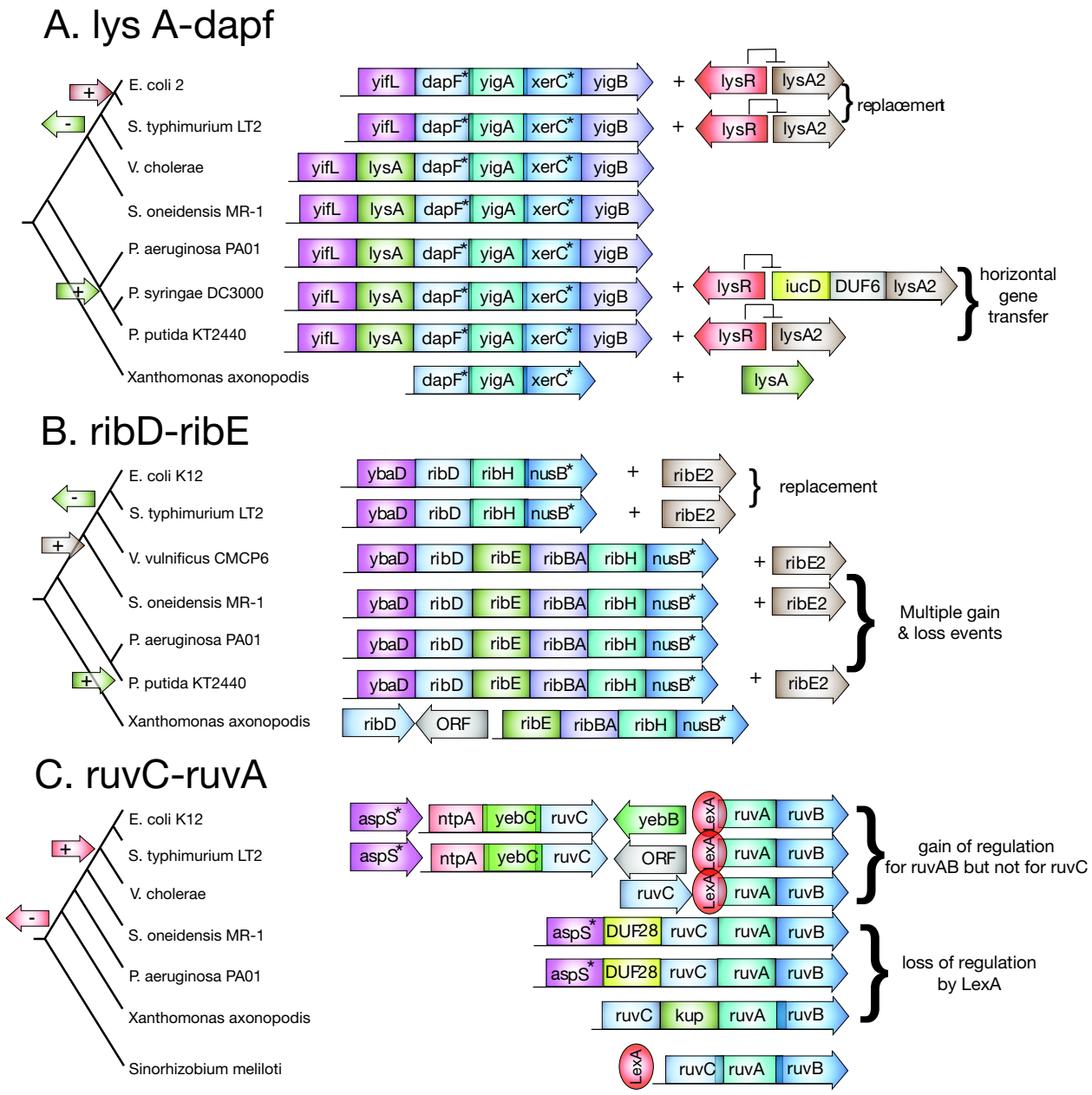

Figure 6: Reconstructed histories of three dead operons. For each dead operon pair, we show the gene order and the predicted or known operon structure in *E. coli* K12 and its relatives. The amount of spacing between genes is not shown. The trees show the branching order of the species according to Lerat *et al.* (2003) and concatenated protein trees (data not shown). We also show a parsimonious reconstruction of events, marked by "+" and "-" on the branches and the labels at right. Genes that are essential for growth in rich media (from Gerdes *et al.* (2003)) are marked with an asterisk (*).

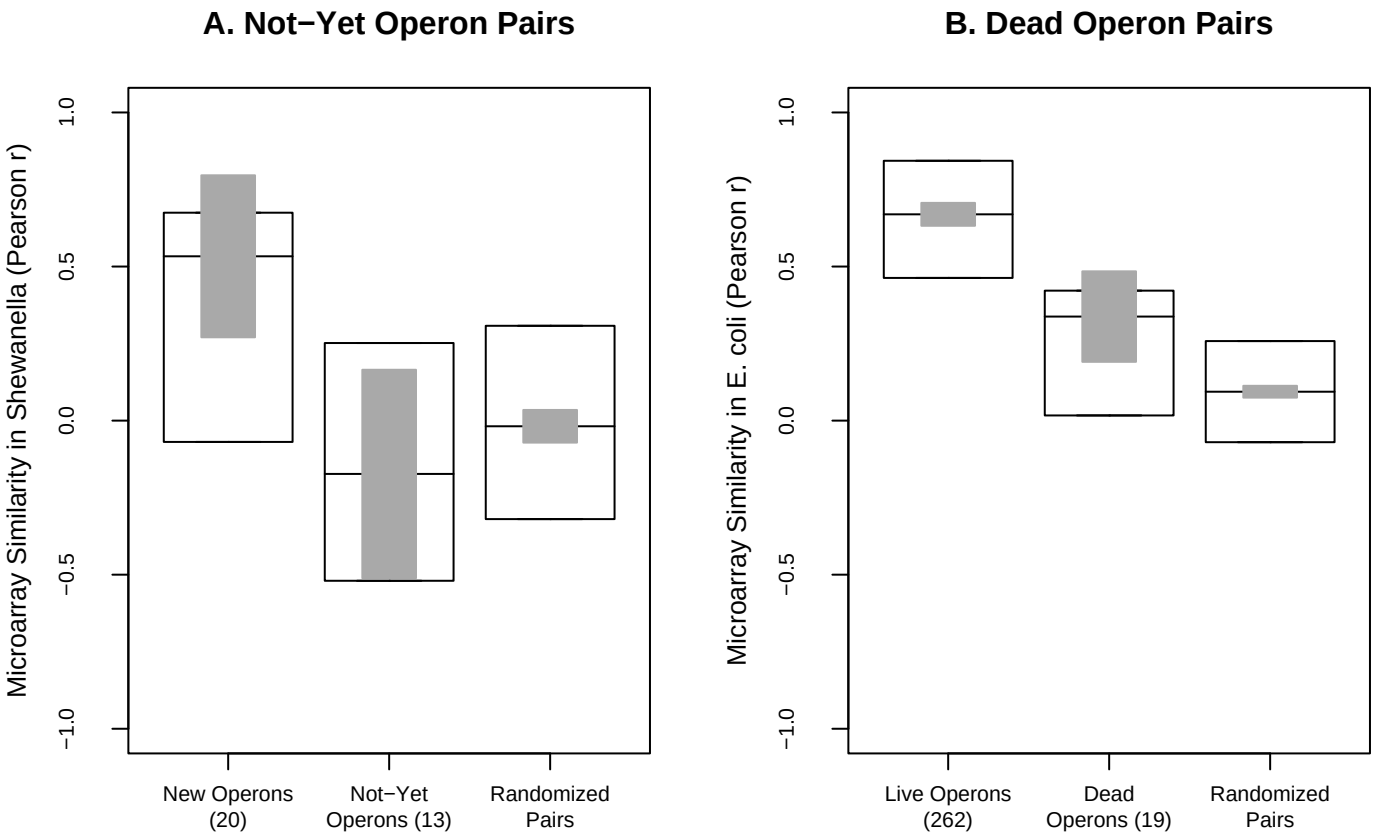

Figure 7: Operon evolution affects the pattern of gene expression. (A) The distribution of microarray similarity in *Shewanella oneidensis* MR-1 for "already" new operon pairs that are shared between *Shewanella* and *E. coli* K12, for "not-yet" pairs that are far apart in *Shewanella* but are in newer operons in *E. coli*, and for randomized pairs of the genes in the latter pairs. For each distribution, the box shows the median and first and third quartiles, and the grey bar shows a 90% confidence interval for the median, so that if two bars do not overlap then the difference in medians is significant ($P < 0.05$). (B) The distribution of microarray similarity in *E. coli* K12 for "live" new operon pairs that are conserved in *Shewanella*, for "dead" operon pairs of similar age that are far apart in *E. coli* K12, and for randomized pairs of the latter genes. For both (A) and (B), t tests gave similar results for significance (data not shown).

Supplementary Table 1: Modifications to pre-existing operons. The modified operons tabulated in Figure 3B were examined by hand. To avoid potential false positives in the operon predictions, only coexpressed pairs (Pearson correlation > 0.5) were included, and known non-operon pairs were excluded. Because of the large number of append and prepend events, only known operons (Karp *et al.* 2002) affected by those events were examined. Preexisting operons are underlined; inserted, replaced, appended, and prepended genes are in bold.

Join Two Preexisting Operons
yggS-yggT-yggU-rdgB-yggW
mhpA-mhpB-mhpC-mhpD-mhpF-mhpE
yiaKL-*yiaMNO*-***lyxK***-*sgbH-sgbU-sgbE*
Internal Replacement
ymcC-***ymcB***-*ymcA*
rpoZ-spoT-***spoU***-*recG*
nirB-nirD-***nirC***-*cysG*
Internal Insertion
celA-celB-celC-***celD***-*celF-ydjC*
Prepend (5 known / 12 total)
mhpA-mhpB-***mhpC***-*mhpD-mhpF-mhpE*
csgD-*csgE-csgF-csgG*
damX-***dam***-*rpe-gph-trpS*
nfrA-***nfrB***-*nfrC-nfrD-nfrE-nfrF-nfrG*
creA-*creB-creC-creD*
Append (5 known / 15 total)
cynT-cynS-***cynX***
pspA-pspB-pspC-***pspD***-*pspE*
fucP-fucI-fucK-fucU-***fucR***
yhdT-panF-***prmA***
mtlA-mtlD-***mtlR***

Supplementary Note 1: New Operons that Formed by Deletion

If two genes that are in the same operon in *E. coli* are near each other but in different operons in *Vibrio* species, then we infer that the operon formed by deleting the intervening genes. A second possible explanation is that the common ancestor of the Enterobacteria and *Vibrio* formed the operon, and that another gene was then inserted in the *Vibrio* lineage. Finally, a third alternative is that the common ancestor formed nearby genes by rearrangement, without forming an operon, and then both insertions in the *Vibrios* and deletions in the *E. coli* lineage occurred. The deletion scenario is more parsimonious than the insertion scenario because it involves a single operon creation/destruction event, instead of operon creation followed by later destruction in the *Vibrios*. The deletion scenario is more parsimonious than the insertion/deletion scenario because fewer events are required.

Conserved proximity in *Shewanella oneidensis* MR-1 occurs for two of the putative deletion events. First, *serB* and *radA* (also known as *sms*) are separated in *S. oneidensis* by a homolog of *VC2344* and one additional protein. Second, *ygiF* and *glnE* are separated by only 8 intervening genes in Shewanella, including an ortholog of *VP0422*. Because *S. oneidensis* probably diverged from *E. coli* before the *Vibrios* (a concatenated protein tree using TreePuzzle (Schmidt *et al.* 2002) gave a puzzling score of 97/100), this shows that the common ancestor of the *Vibrios* and *E. coli* had the intervening genes and not the operon, as in the deletion scenario.

If the deletion scenario is correct then the intervening genes should be absent from the Enterobacteria and sometimes present in more distant relatives of *E. coli*. In the two cases of conserved proximity in *S. oneidensis*, an intervening gene is present in the same location in *S. oneidensis* and absent from Enterobacteria and from other closer relatives of E. coli such as *Haemophilus* and *Pasteurella*. Most of the other intervening genes appear to be horizontally transferred into *Vibrio* from distant bacteria, so that their absence from the Enterobacteria is unsurprising and uninformative. A striking exception is *asnC*, one of the genes that separates *prlC* and *yihQ* in Vibrios: *asnC* has clear orthologs in most Enterobacteria and in *S. oneidensis*. Although this type of deletional rearrangement seems somewhat surprising, it is equivalent to insertional rearrangements (as in the formation of *ptr-recB*, see Table 1) and is arguably more parsimonious than the alternative, which would be rearrangement to form the operon and then an insertion.

Another argument for deletion arises with *btuB* and *murI*: the 68 bp overlap results from the addition of over 20 amino N-terminal amino acids to *murI* that is not present in genes without the operon. These amino acids are encoded by the 3' end of *btuB*. This overlap is probably correct because the predicted molecular weight from the *murI* sequence matches that observed in Western blots (P. Doublet et al., J. Bacteriol. 175:2970-9). Without the overlap the gene product would be 10% too light. The overlap is present in all of the sequenced Enterobacteria, and is probably the ancestral state of the operon. Thus, we speculate that the original start codon was lost during the deletion

event. The original start codon could also have been lost by a rearrangement to create the operon, followed by an insertion in *Vibrio*, but then it would be particularly difficult to insert the *Vibrio* ATPase between *btuB* and *murI*.