Operon formation is driven by co-regulation and not by horizontal gene transfer

Morgan N. Price,[1] Katherine H. Huang,[1] Adam P. Arkin,[1,2] and Eric J. Alm[1,3]

[1]*Lawrence Berkeley Laboratory, Berkeley California, 94720 USA;* [2]*Howard Hughes Medical Institute and the Department of Bioengineering, University of California, Berkeley, California, 94720 USA*

The organization of bacterial genes into operons was originally ascribed to the benefits of co-regulation. More recently, the "selfish operon" model, in which operons are formed by repeated gain and loss of genes, was proposed. Indeed, operons are often subject to horizontal gene transfer (HGT). On the other hand, non-HGT genes are particularly likely to be in operons. To clarify whether HGT is involved in operon formation, we identified recently formed 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 a cause of 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 and infer that operons should be more likely to evolve than independent promoters when regulation is complex. Consistent with this hypothesis, operons have greater amounts of conserved regulatory sequences than do individually transcribed genes.

[Supplemental material is available online at www.genome.org.]

Bacterial genes are often transcribed together in operons, so that several genes are under the control of a single promoter. Genes in operons show coordinated expression (Jacob and Monod 1961; Sabatti et al. 2002), which allows the cell to produce appropriate amounts of the encoded proteins. Disruption of the operon would disrupt this co-regulation and thus, operons should be maintained by purifying selection (Jacob and Monod 1961). Indeed, many conserved clusters of genes can be found within all of the major divisions of prokaryotes (Overbeek et al. 1999; Huynen et al. 2000), and most of these conserved clusters correspond to operons (Ermolaeva et al. 2001; Price et al. 2005). Furthermore, although most operons are shuffled away given sufficient evolutionary time (Itoh et al. 1999), operons containing genes whose protein sequence evolves slowly are more likely to be conserved (de Daruvar et al. 2002). Thus, operons are under strong purifying selection.

Although operons are maintained because of co-regulation, they might form for other reasons. It has been argued that the formation of operons for the purposes of co-regulation is both unnecessary and implausible (Lawrence and Roth 1996; Lawrence and Ochman 1998). Operon formation is unnecessary because independent promoters can evolve to bind the same regulator(s) and provide the benefits of co-regulation. Genes for a metabolic pathway are often found grouped together within an operon, but unlinked enzymes are also common (Lawrence and Roth 1996; Lawrence 1999). Furthermore, because operon structure diverges over time, any pathway that is in a single operon in one genome is likely to be found dispersed in another (Itoh et al. 1999).

[3]Corresponding author.
E-mail ejalm@lbl.gov; fax (510) 486-6059.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3368805. Freely available online through the *Genome Research* Immediate Open Access option.

It has also been suggested that operon formation for the purposes of co-regulation is implausible because two genes need to be rearranged to the correct position out of all the possible sites in the genome, and hence, such rearrangements should be rare (Lawrence and Roth 1996). However, rearrangements occur frequently in cultured *E. coli*, with rates of 10^{-2} to 10^{-4} per generation (Louarn et al. 1985; Papadopoulos et al. 1999). Furthermore, large bacterial population sizes ensure that apparently unlikely rearrangements will take place and have the opportunity to be selected for. Indeed, recombination events that move genes in front of beneficial promoters have been observed experimentally, including apparently implausible double recombinations that move a gene without disrupting the strand bias of the overall chromosome (Konrad 1969; Schmid and Roth 1983). As another example, conserved operons have a surprising 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. Nevertheless, it is not clear why operon formation should be the preferred pathway for evolving co-regulation instead of gradually evolving new regulatory sequences, even if rearrangements to produce operons are possible.

As an alternative to the co-regulation theory of operon formation, Lawrence and Roth (1996) proposed that "selfish operons" form by a process involving horizontal gene transfer (HGT). 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, but only if all of the genes can be acquired. Even if the genes are not yet in an operon, genes that are near each other can still be transferred together. Occasionally, these genes will move still closer together

by a random rearrangement and then be transferred. Moving the genes closer together will increase the likelihood of their simultaneous transfer to another host. After a few cycles of this process, the genes will be tightly clustered, and an operon can form by deletion of the intervening DNA. Thus, during the process of operon formation, the incipient operon is selfish, and the clustering benefits the genes and their further propagation rather than the host. However, once the operon has formed, co-regulation may still play a role in maintaining the operon (Lawrence and Roth 1996; Lawrence 1999).

The selfish operon model is appealing because it provides an intermediate step to operon formation, without requiring rare beneficial rearrangements. The theory also explains why many operons have been acquired by HGT (Lawrence and Roth 1996), a finding that was confirmed by a recent analysis across many complete genomes (Omelchenko et al. 2003). Finally, the theory makes the testable prediction that essential genes should not be in operons, because these genes cannot undergo the cycles of gene loss and gain in the model (Lawrence and Roth 1996; Lawrence 1999). Furthermore, essential genes are often conserved without paralogs across many species, and such genes rarely undergo HGT (Lerat et al. 2003).

At the time the selfish operon model was proposed, it was believed that operons tended to contain non-essential genes, with the exception of a few broadly conserved operons such as the ribosomal operons. Lawrence and Roth (1996) proposed that these operons were ancient and had formed by some other mechanism. However, more recent analyses of essential genes and operons, using genome-wide essentiality data and larger data sets of operons, found that essential genes are preferentially found in operons, even when ribosomal proteins are excluded from consideration (de Daruvar et al. 2002; Pal and Hurst 2004). These findings call into question the selfish operon model (Pal and Hurst 2004) but do not rule out the possibility that these essential operons are also ancient, while new operons are forming selfishly. Furthermore, it is unclear why operons should so often be subject to HGT, as confirmed by Omelchenko et al. (2003).

To clarify the relationship between HGT and operons, we distinguished 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 classified genes as being native or HGT or "ORFan" (lacking homologs outside of one phylogenetic group of bacteria), and similarly classified operons as being ancestral or HGT or newly formed. We then compared the histories of the operons 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. Instead, HGT operates on pre-existing operons.

Having shown that HGT does not explain the creation of new operons, we revisited the co-regulation theory and, in particular, the question of why operon formation is preferred over the evolution of coregulation by independent promoters. We hypothesized that, as the amount of regulatory information required to specify the optimal expression pattern increases, evolving the optimal expression profile separately for each gene becomes more difficult, while creating an operon does not. Thus, we predicted that operons would 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 determined 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 hypothesized to be transferred from phages, where they can evolve rapidly, 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 phage-related genes and transposases (89 genes total). 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),

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 and E.J. Alm, unpubl.). "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.

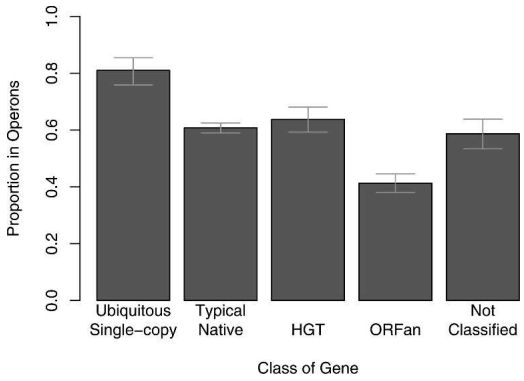

Figure 2. HGT genes are not particularly likely to be in operons. For each class of gene, solid bars show the proportion that are in predicted operons. Error bars show 90% confidence intervals from the binomial test; if two error bars do not overlap, then the corresponding classes have significantly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

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 (Fig. 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 (Supplemental Fig. 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 (Fig. 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." We examined all pairs, regardless of whether the genes in the pair could be classified as native, HGT, or ORFan. 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. (For non-HGT genes and operon pairs, age was defined as the oldest outgroup that contained the gene or pair. For HGT genes and imported operon pairs, the more distant outgroups that were the source of the putative transfer event were excluded when calculating the age, so that the age corresponds to the common ancestor that was the recipient of the putative transfer.) 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 model 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

Table 1. Proportions of native and HGT genes that formed new operons or were imported as operons

	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)

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).

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 attributable 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 1128 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 (KH) 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.

Table 2. Genes in new operons are no more likely than other genes to have trees that are discordant with the species tree

	In a new operon?	
	Yes	No
# Concordant	345	692
# Discordant	22	69
% Concordant	94.0%	90.9%

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).

Non-HGT genes often form new operons

In defense of the selfish operon model, 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 (these included 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 βγ-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

Table 3. Validated new operon pairs containing ubiquitous single-copy (non-HGT) genes

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	βγ-Proteo.
rpsF	*priB**	Yes	0.92	βγ-Proteo.
nlpB	***dapA****	Yes	0.87	βγ-Proteo.
atpI	***atpB****	Yes	0.85	βγ-Proteo.
ftsJ	*hflB*	Yes	0.72	βγ-Proteo.
yacE*	*yacF*		0.69	βγ-Proteo.
folC	*dedD*		0.64	βγ-Proteo.
*yffB**	***dapE****		0.54	βγ-Proteo.
slpA	***lytB****	Yes	—	βγ-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.

Ubiquitous single-copy genes are in bold face type. 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.

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 reconsidered the traditional co-regulation theory. More specifically, we considered that optimal co-regulation can be achieved by forming an operon or by modifying two or more independent promoters. Given sufficient evolutionary time, co-regulation could evolve by either mechanism. We will argue that operon formation will be the preferred pathway when the regulation is complex.

Consider how likely it is for an operon to form by chance. To form a two-gene operon, the first gene can be placed anywhere, and the second gene must be placed between the first gene and the next gene downstream, and also on the same strand. For the operon to function correctly, there might also be restrictions on the spacing between the genes: For example, the second gene might need to be inserted before a pre-existing transcription terminator for the first gene.

To modify the expression pattern of the second gene to match that of the first gene without forming an operon, one or more new transcription factor (TF) binding sites would be required. Most TFs are only moderately specific, so that new binding sites can easily arise by chance. Once a weak but functioning site exists, it can easily be optimized by single base changes, each of which might have a selective advantage. This contrasts to the all-or-nothing benefits of operon formation.

However, most genes are probably regulated by more than one TF binding site. Indeed, of the characterized *E. coli* K12 promoters in EcoCyc (Karp et al. 2002), 64% have more than one known site, and 27% have four or more known sites. As more TF binding sites are required, co-regulation without operons would become progressively more difficult to evolve. Furthermore, there might also be constraints in the spacings between the TFs. A recent study of the 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 spacings of an exact number of base pairs and spacings of within a few dozen base pairs. These constraints represent additional information that each promoter would need to evolve.

In contrast, the likelihood of forming an operon by chance is not affected by the complexity of regulation. Thus, if the regulation is complex, operons give a large savings in the amount of regulatory information that must be encoded in the genome, and operon formation should be preferred. Finally, if operon formation is more strongly preferred when regulation is more complex, then operons should 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 with 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 6595 footprinted sites upstream of 2047 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: Typical native genes at the head of an operon had an average of 3.50 sites upstream, while typical native genes transcribed individually averaged 3.26 sites upstream ($P < 10^{-4}$, *t*-test).

To determine whether this correlation between operons and the amount of regulatory sequences was causal—with complex regulation driving operon formation—rather than being due to some 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 purifying 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

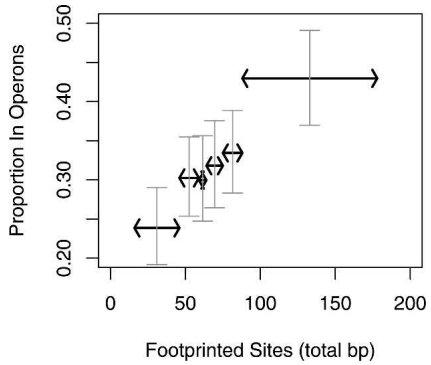

Figure 3. Genes with more conserved upstream sequences are more likely to be in operons. For each *E. coli* gene with one or more sites from phylogenetic footprinting (McCue et al. 2002), we asked whether it was predicted to be at the beginning of a multi-gene operon or to be transcribed by itself. For 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.

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, $P < 0.001$). Analyzing only the typical native genes gave the same result (partial Spearman correlation 0.10, $P < 0.005$).

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, $P = 0.04$).

A third 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. This hypothesis suggests that larger operons would have more conserved regulatory sequences than shorter operons. In contrast, the mean amount of regulatory sequence for predicted operons containing two genes is 67.5 bases, which is *more* than the 65.9 bases for operons predicted to contain three or more genes. Although the difference is not statistically significant ($P = 0.25$, Wilcoxon test), it still casts doubt on this hypothesis.

Fourth, examining the amount of regulatory sequences in all existing operons is an indirect test of the causes of operon formation. One possible explanation of the above results is that operons that contain little regulatory information might disappear more quickly. To perform a more direct test on "new" operons, we examined 157 operons that formed since the latest common ancestor of the βγ-Proteobacteria and whose first gene was a typical native gene. We compared these to the typical native genes that were predicted to be transcribed individually. The mean amount of footprint was 70 bases for new operons and 66 bases for single-gene transcripts. Because of small sample size the effect was not statistically significant ($P = 0.07$, *t*-test), but the effect was similar in size to the difference between all operons and single-gene transcripts.

Fifth, 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 artifact. 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 *Bacillus 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 (Supplemental 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. In fact, they were better predictors of co-expression than whether the two genes shared a verified TF binding site (Supplemental Fig. 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

Consequences for the selfish operon model

We have shown that HGT is not correlated with the formation of new operons. First, genes identified as HGT by their phylogenetic distribution formed new operons at the same rate as typical native genes, both at the time of transfer and after transfer. Furthermore, at the time of transfer, HGT genes did not preferentially form operons with other HGT genes. Second, genes in new operons were no more likely than other genes to have phylogenetic trees that were significantly different from the species tree. Third, 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 more 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 closer 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 operon model requires genes to be lost and then regained to drive operon formation, it seems unlikely that HGT between closely related organisms would be a major factor in operon formation.

Our results do not rule out the possibility that recombination within a population of bacteria of the same species might be involved in operon formation. Such events would probably not be detected by either of our methods, and these recombination events might not require gene loss. However, the selfish operon model as originally described involves HGT from more distant relatives, including those having different base composition (Lawrence and Roth 1996). Furthermore, the selfish operon model requires successive transfers, with a rearrangement after each transfer, to shuffle the genes into closer and closer proximity (see the simulation in Lawrence and Roth 1996). Because of these repeated rearrangements, the organism that contains the fully evolved operon should have substantially different gene order than the original recipient. However, bacteria in a recombining population generally have co-linear chromosomes, with genes being in the same order for hundreds of kilobases except

for small deletions and insertions. Indeed, even *E. coli* K12 and *Salmonella typhimurium*, which diverged perhaps 100 million years ago, have largely co-linear chromosomes. It might yet be possible for operons to form by HGT without large-scale rearrangements, for example by the transfer of a large chromosomal region containing two genes followed by the deletion of the intervening DNA, or by the fortuitous insertion of a single gene in proximity to the other genes in a pathway. However, these scenarios eliminate the major advantage of the selfish operon model, namely the gradual formation of gene clusters. It would seem simpler to form an operon by a rearrangement (or insertion/deletion) followed by selection for co-regulation, without invoking any selfish behavior.

Our results also do not rule out the possibility that genes are forming functionally related clusters by repeated rounds of gene acquisition and loss, as described in the selfish operon model, without then forming operons. However, most of the clustering of functionally related genes in prokaryotic genomes appears to be due to operons. This is certainly the case for conserved clusters of genes, which overwhelmingly consist of same-strand pairs instead of opposite-strand pairs (Overbeek et al. 1999). The two other major forms of clustering, which we discuss below, are unlikely to be due to selfish gene clusters because they involve essential genes. First, Pal and Hurst (2004) recently reported that essential genes are functionally clustered on a larger scale than operons. Second, although one hallmark of operons is the prevalence of conserved adjacent genes on the same strand, some divergently transcribed pairs of genes are also conserved. These divergent pairs are co-regulated (Korbel et al. 2004) and could plausibly arise by the same mechanisms as operons. Indeed, in *E. coli* K12, these conserved divergent pairs are significantly enriched in essential genes (data not shown). Thus, although selfish gene clusters may exist, they are not a major structural feature of bacterial genomes.

The role of HGT in propagating operons once they have formed

Despite the lack of involvement of HGT in 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 model—(1) genes in operons tend to be functionally related, (2) transferring an entire operon allows an organism to acquire a useful new capability, and (3) the proximity of the genes facilitates transfer by a single HGT event (Lawrence and Roth 1996; Lawrence 1999). Consistent with the hypothesis of functional coherence, of the operon pairs that were imported into *E. coli* K12 and are also annotated, 73% have matching function codes from COG (Tatusov et al. 2001). Because operons often "die" by being shuffled apart (Itoh et al. 1999), we also expect that HGT extends the lifetime of individual operons and that HGT may increase the prevalence of operons in bacterial genomes, even though HGT does not contribute to operon formation.

The role of co-regulation in operon formation

As an alternative to the selfish operon model, we proposed a new interpretation of the traditional co-regulation theory: Operons reduce the information required to specify the 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.

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 with 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. 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. Thus, once an attenuator exists, the pressure to place other genes in the pathway in the same operon behind the attenuator may be particularly strong.

We are not aware of any previous work with direct evidence that co-regulation drives the formation of operons, 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. Consistent with these previous observations, the functional coherence of the new operons that we identified was modest—of the new operon pairs for which both genes had annotations, only 33% had matching COG function codes, whereas 81% of ancestral operon pairs and 73% of imported operon pairs had matching function codes if they were annotated (all differences significant at $P < 0.01$, Fisher exact test). This low level of function coherence does not appear to reflect errors in operon predictions—we obtained similar results (not shown) when examining only the new operon pairs with strongly similar expression patterns in microarray data (Pearson correlation >0.7) or only the new operon pairs that overlapped by 1 or 3 bases, which is a strong indicator of operons (Salgado et al. 2000). The low functional coherence of new operons is further evidence against the selfish operon model, which requires that new operons encode coherent pathways. The much greater functional coherence of older operons, relative to that of new operons, presumably reflects the stronger conservation of functionally coherent operons (de Daruvar et al. 2002).

One attractive feature of the selfish operon model is that it provides 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 im-

plausible. As we have discussed, such rearrangements can be identified in culture (Konrad 1969; Schmid and Roth 1983), and have been identified in a comparative genomics study (Omelchenko et al. 2003). Another factor that might aid such rearrangements is intermediate forms of clustering in the bacterial genome. For example, adjacent operons are sometimes regulated by the same TF (Hershberg et al. 2005). On a larger scale, essential genes with broadly related functions tend to cluster together over distances of up to 30 genes or roughly 30 kilobases (Pal and Hurst 2004), and regions of the genome over 100 kilobases in size tend to have similar expression patterns (Allen et al. 2003). These latter two effects occur on a much larger scale than operons, which typically contain 2–5 genes or 2–5 kb or DNA. 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 model 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), minimizing the half-life of toxic monomers (Pal and Hurst 2004), or reducing stochastic differences in gene expression (Swain 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. Conversely, only a tiny fraction of the protein complexes in *E. coli* consist of genes in the same operon (Butland et al. 2005). 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 with 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. Genes with BLASTp hits that contradicted the COG assignments were excluded from our classification. 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, as described previously (Price et al. 2005). 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. 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*. The effect of false negatives in these operon predictions on our results 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 both new and imported pairs confirmed that false negative operon predictions in one or two genomes was not creating spurious new or imported operons—only in rare cases were the genes near each other and on the same strand in a genome that was predicted to lack the operon, and in these cases, it was plausible that the *E. coli* K12 operon had formed by deleting intervening DNA that was present in the common ancestor.

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 (see Table 3 for details). 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, which is near the median for known operon pairs, and above the 90th percentile for known non-operon pairs). 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 misclassified 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 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, such 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 γ-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 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 1000 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 rightmost arrow in Figure 3.

To compute the protein sequence conservation of genes between *E. coli* and *S. enterica Typhi*, we used the %identity (from BLASTp) between putative orthologs. To avoid paralogs that are bidirectional best hits because the truly orthologous genes were lost from one or both genomes, we required the putative orthologs to have at least 60% identity (similar in spirit to Lerat et al. 2003).

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 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.Ø. 2003. Genome-scale analysis of the uses of the *Escherichia coli* genome: Model-driven analysis of heterogeneous data sets. *J. Bacteriol.* **185:** 6392–6399.

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*. *Genome Res.* **14:** 201–208.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. 2005. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433:** 531–537.

Cherry, J.L. 2003. Genome size and operon content. *J. Theor. Biol.* **221:** 401–410.

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:** 324–328.

Daubin, V. and Ochman, H. 2004. Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. *Genome Res.* **14:** 1036–1042.

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:** 211–221.

Ermolaeva, M.D., White, O., and Salzberg, S.L. 2001. Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29:** 1216–1221.

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:** 5673–5684.

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:** 94–96.

Heck, J.D. and Hatfield, G.W. 1988. Valyl-tRNA synthetase gene of *Escherichia coli* K12. Molecular genetic characterization. *J. Biol. Chem.* **263:** 857–867.

Hershberg, R., Yeger-Lotem, E., and Margalit, H. 2005. Chromosome organization is shaped by the transcription regulatory network. *Trends Genet.* **21:** 138–142.

Huynen, M., Snel, B., Lathe III, W., and Bork, P. 2000. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res.* **10:** 1204–1210.

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:** 332–346.

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:** 56–58.

Konrad, E.B. 1969. *The genetics of chromosomal duplications*. PhD thesis, Harvard University, Cambridge, MA.

Korbel, J.O., Jensen, L.J., von Mering, C., and Bork, P. 2004. Analysis of genomic context: Prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22:** 911–917.

Lawrence, J.G. 1999. Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.* **9:** 642–648.

Lawrence, J.G., and Ochman, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci.* **95:** 9413–9417.

Lawrence, J.G. and Roth, J.R. 1996. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* **143:** 1843–1860.

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

Louarn, J., Bouche, J., Legendre, F., Louarn, J., and Patte, J. 1985. Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis. *Mol. Gen. Genet.* **201:** 467–476.

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–D77.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., et al. 2003. CDD: A curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31:** 383–387.

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:** 1523–1532.

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.* **96:** 2896–2901.

Pal, C. and Hurst, L.D. 2004. Evidence against the selfish operon theory. *Trends Genet.* **20:** 232–234.

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.* **96:** 3807–3812.

Price, M.N., Huang, K.H., Alm, E.J., and Arkin, A.P. 2005. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* **33:** 880–892.

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:** 777–787.

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:** 241–254.

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:** 2212–2223.

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:** 2886–2893.

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.* **97:** 6652–6657.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29:** 2994–3005.

Schmid, M.B. and Roth, J.R. 1983. Selection and endpoint distribution of bacterial inversion mutations. *Genetics* **105:** 539–557.

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.

Swain, P.S. 2004. Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J. Mol. Biol.* **344:** 965–976.

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:** 22–28.

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.

Received November 15, 2004; accepted in revised form March 16, 2005.