

RESEARCH ARTICLES

How Essential Are Nonessential Genes?

Gang Fang,* Eduardo Rocha,*† and Antoine Danchin*

*Unité Génétique des Génomes Bactériens, Institut Pasteur, Paris Cedex, France; and †Atelier de Bioinformatique, Université Pierre et Marie Curie, Paris, France

Gene essentiality in bacteria has been identified in silico, focusing on gene persistence, or experimentally, focusing on the growth of knockouts in rich media. Comparing 55 genomes of Firmicutes and Gamma-proteobacteria to identify the genes which, while persistent among genomes, do not lead to a lethal phenotype when inactivated, we show that the characteristics of persistence, conservation, expression, and location are shared between persistent nonessential (PNE) genes and experimentally essential genes. PNE genes show an overrepresentation of genes related to maintenance and stress response. This outlines the limits of current experimental techniques to define gene essentiality and highlights the essential role of genes implicated in maintenance which, although dispensable for growth, are not dispensable from an evolutionary point of view. Firmicutes and Gamma-proteobacteria are mostly differing in the construction of the cell envelope, DNA replication and proofreading, and RNA degradation. In addition to suggesting functions for persistent genes that had until now resisted identification, we show that these genes have many characters in common with experimentally identified essential genes. They should then be regarded as truly essential genes.

Introduction

Bacterial genome sequences differ considerably in their fine structures. Rearrangements shuffle the order of genes (Rocha 2003; Sankoff 2003) and often delete genetic information (Mira, Ochman, and Moran 2001). Horizontal gene transfer further contributes to genome evolution, often increasing the genomes' gene content (Lawrence and Hendrickson 2003). The quantity of DNA lost and gained is at an equilibrium which can be shifted according to the evolutionary and ecological constraints imposed by the lifestyle of the organisms. Genomes with highly restricted ecological niches, lacking selection on the maintenance of unused functions, tend to have smaller genomes, while most free-living bacteria tend to have larger genomes (Stepkowski and Legocki 2001; Bentley and Parkhill 2004; Klasson and Andersson 2004). However, a subset of genes is rarely lost or gained because it codes for functions that are necessary for growth in almost any condition. These commonly named "essential" genes have sparked significant interest (Koonin 2003). Being essential for growth in media where most nutrients are provided, they are good targets for broad-spectrum antibacterial drugs (Chalker and Lunsford 2002). These genes are of particular interest also because they represent the basal bioprocesses required for life (Maniloff 1996). Finally, they constitute a set of persistent genes, expected to be less subject to loss and horizontal transfer (Lerat, Daubin, and Moran 2003; Lerat et al. 2005).

A seminal work analyzed the complete genomes of *Mycoplasma genitalium* and *Haemophilus influenzae* and proposed 256 genes as forming the essential gene pool (Mushegian and Koonin 1996). Most essential pathways appeared to be conserved, but many were fulfilled by non-orthologous genes, after gene recruitment (Jensen 1976). "Acquisitive evolution" has long been substantiated by experiments showing functions achieved by quite different

structures (Hegeman and Rosenberg 1970). If this were the rule rather than the exception, then it would be difficult, if not impossible, to uncover common properties in distant genomes. The variation uncovered in experimentally defined essential gene pools more or less substantiated this concern, although the different experimental methods certainly contributed to the incongruence. For instance, after transposon mutagenesis, 265–350 of the 480 genes from *M. genitalium* (Hutchison et al. 1999), 670 from the total of 1,700 genes of *H. influenzae* (Akerley et al. 2002), approximately 400 genes from *Pseudomonas aeruginosa* (Jacobs et al. 2003), and 620 genes from *Escherichia coli* (Gerdes et al. 2003) were identified as essential. Other approaches, such as translation inhibition with antisense RNA, predicted 658 essential genes from *Staphylococcus aureus*, among which only 168 were conserved in *M. genitalium* (Forsyth et al. 2002). It was argued that these experimental setups were misestimating the number of essential genes (Kobayashi et al. 2003). Gene disruption with a vector meant to alleviate polarity in transcription was applied to the model gram-positive bacterium *Bacillus subtilis* and led to the identification of 271 essential genes (Kobayashi et al. 2003).

The essentiality of a gene is relative to a set of experimental conditions and to an output variable, typically sustained growth on solid media (colony formation). It is, however, quite different for a cell to survive in a laboratory setting, with plenty of supplied metabolites, and to thrive in the wild, competing with other organisms (or mutants of the same organism) for limited resources. Theoretical models showed that in yeast most metabolic genes are essential under certain experimental conditions (Papp, Pal, and Hurst 2004). Growth is not the steady-state situation of cells. Starvation or stresses are omnipresent, and the fitness effect of mutating genes essential for survival under transition from one environmental condition to another has not yet been assessed in a large-scale experimental setup. Finally, a gene may not be essential for growth but its loss may lead to such a lower fitness that its deletion will never be fixed in natural populations.

Key words: bacterial evolution, essentiality, gene expression, gene strand bias, protein evolution, gene content.

E-mail: adanchin@pasteur.fr.

Mol. Biol. Evol. 22(11):2147–2156. 2005

doi:10.1093/molbev/msi211

Advance Access publication July 13, 2005

This prompted us to explore gene essentiality by approaching it from the perspective of persistence. In bacteria, essential genes appear to be more conserved than nonessential genes (Jordan et al. 2002). This probably results only from their higher expression level (Rocha and Danchin 2004). If orthologous genes can be found in the majority of a certain bacterial phylogenetic group, this informs about the contribution of such genes to fitness. Furthermore, combining miscellaneous biological clues, persistent genes uncover some yet undiscovered functions (Galperin and Koonin 2004). Previous work has shown that in Gamma-proteobacteria and Firmicutes, several types of selection pressure constrain the genome structure. Genes are distributed differently in the leading and lagging strand of the genomes, and those making the core of essential genes are preferentially located on the leading strand (Rocha and Danchin 2003). Starting with the properties uncovered for the “laboratory-essential” genes, we looked for widespread genes that share similar properties. This uncovered a large set of persistent genes connected to the general selection pressures leading to their conservation and organization within genomes. A small category of experimentally identified essential genes does not belong to the persistent class. We further discovered a gene set common to Gamma-proteobacteria and Firmicutes and identified sets of genes specific to these large domains of the Bacteria kingdom. When unknown, the corresponding functions are discussed together with experimental predictions.

Materials and Methods

Data

Genome sequences and annotations of Firmicutes and Gamma-proteobacteria were taken from the EMBL data bank (see Supplementary Table 1, Supplementary Material online). We did not include the genomes of obligatory endosymbionts (among Enterobacteriaceae) or the Mollicutes. Regarding the species with two or more strains sequenced, we retained one representative genome. We used the refined *B. subtilis* essential list (272 genes) and the Profiling of *E. coli* Chromosome database (PEC, <http://www.shigen.nig.ac.jp/ecoli/pec/>) as our essential gene references. (PEC classifies *E. coli* genes into three groups, i.e., 252 essential, 2,368 nonessential, and 1,789 unknown genes.) Complementarily, we also used the data set of Gerdes et al. (2003) of essential genes. Gene name and annotations were taken from the GenoList database (<http://genolist.pasteur.fr/>); we also referred to the EcoCyc database for *E. coli* genes’ functional classification (<http://ecocyc.org>).

Definition of Orthologs

Orthologs were defined using the bidirectional best hit (BBH) strategy, i.e., if gene *a* of genome *A* is the most similar hit of gene *b* in genome *B* and vice versa, genes *a* and *b* are regarded as a pair of orthologs (Tatusov, Koonin, and Lipman 1997). Furthermore, BBH proteins must show more than 40% similarity and less than 30% difference in length.

Persistent Genes

To represent how persistent a gene is, we created a persistence index (PI) for each gene by dividing the

number of genomes carrying an ortholog ($N(\text{orth})$) by the number of genomes searched ($N(\text{gen})$)

$$\text{PI} = \frac{N(\text{orth})}{N(\text{gen})}$$

Leading-Strand Preference

Because essential genes tend to be coded on the leading strand (Rocha and Danchin 2003), we characterized genes with an ortholog leading-strand index (OLI), describing orthologs leading-strand preference

$$\text{OLI} = \frac{N(\text{orth.leading})}{N(\text{orth.leading} + \text{orth.lagging})}$$

Replication origin and terminus were first roughly identified from the cumulated GC skew curve (Lobry 1996), and then the origin was eventually refined according to the coordinates of the gene *dnaA*.

Protein Sequence Divergence

Sequence conservation integrates a variety of selection pressures to which the protein has been submitted during evolution. Therefore, considering persistent genes within a wide enough phylogenetic branch, sequence conservation results from more important selective constraints, which could suggest that the gene belongs to a structure or process that has some crucial character in the cell’s life cycle. To take this factor into account, we created a third index sequence divergence (SD) for each gene as follows:

$$\text{SD} = \frac{\sum_{i=1}^N S_i/D_i}{N \times 100}$$

where N is the number of orthologs one gene possesses and S_i and D_i are the protein sequence similarity and length difference, respectively, between a gene and its i th ortholog. D_i is the quotient by dividing the length of the longer protein by that of the shorter protein.

Codon Adaptation Index

The codon adaptation index (CAI) monitors gene expression level for fast-growing bacteria (Sharp and Li 1987). In bacteria, a higher expression level is correlated with slower evolutionary speed of protein sequences (Sharp 1991; Rocha and Danchin 2004). Moreover, protein sequence divergence, gene essentiality, and expression level are somewhat correlated in eukaryotes (Krylov et al. 2003). We therefore calculated a CAI value for each gene in *B. subtilis* and *E. coli* using the EMBOSS package (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/cai.html>) and using ribosomal proteins as the set of highly expressed genes.

Results

Persistent Genes Versus Essential Genes

While most essential genes are spread in the genomes we studied, only 34% of the *B. subtilis* essential genes have

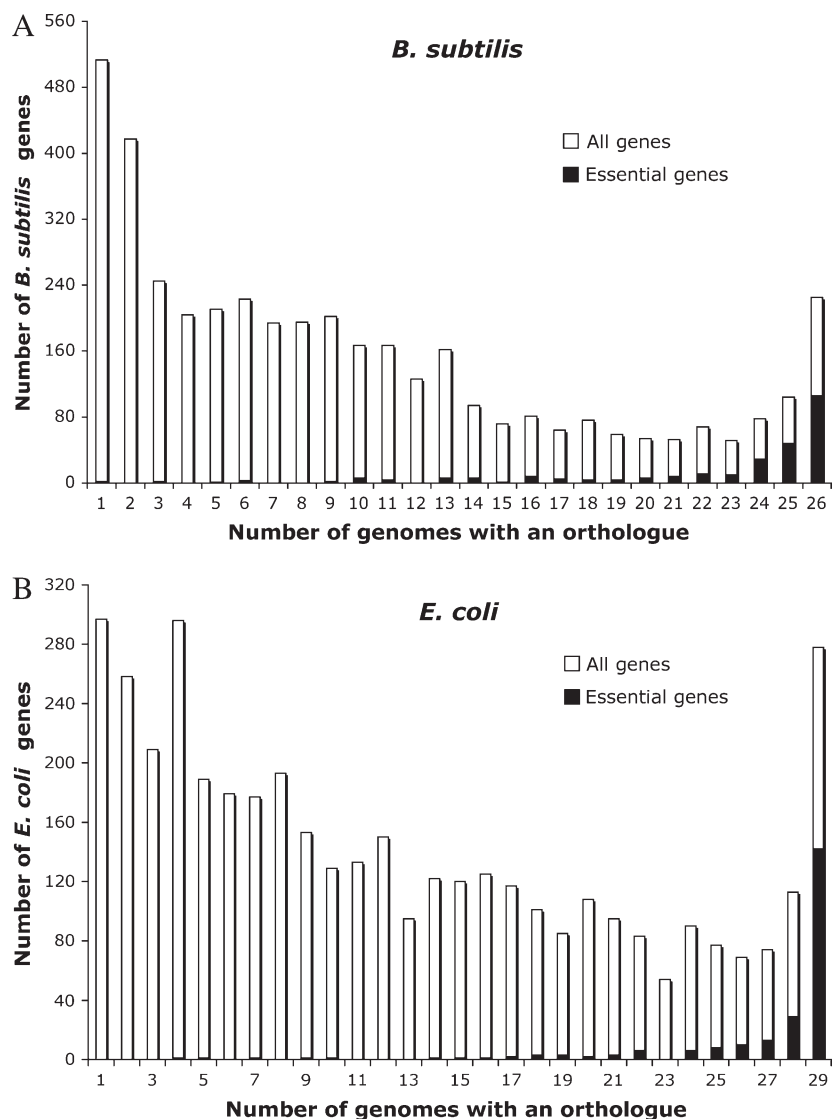


FIG. 1.—Distribution of the number of orthologs for each gene of *Bacillus subtilis* (resp. *Escherichia coli*) among 26 genomes of Firmicutes (respectively 29 of Gamma-proteobacteria).

reliable orthologs in all Firmicutes and 61% of the *E. coli* essential genes have reliable orthologs in all Gamma-proteobacteria. In contrast, the orthologous genes shared by most of the genomes are much more abundant than the number of essential genes (fig. 1). Essential genes account for 60% (respectively 46%) of the orthologs present in all the Firmicutes (respectively Gamma-proteobacteria). Hence, a large fraction of persistent genes are not currently considered essential, while their persistence suggests that their loss is either lethal or extremely deleterious. Here, we define persistent genes as the genes present in more than 85% of the genomes of the clade. We obtained 475 and 611 instances of these genes for *B. subtilis* and *E. coli*, respectively (Supplementary Table 2A and B, Supplementary Material online). We classified genes into five classes according to persistence and essentiality: persistent genes, essential genes, persistent nonessential (PNE) genes (276 in *B. subtilis*, 409 in *E. coli*), essential nonpersistent (ENP) genes (73 in *B. subtilis*, 33 in *E. coli*),

and nonpersistent nonessential (NPNE) genes (3,558 in *B. subtilis*, 3,525 in *E. coli*).

We then characterized the five gene classes in terms of persistence (PI), leading-strand preference (OLI), sequence divergence (SD), and codon usage (CAI) (see *Materials and Methods*). To validate statistically our observations, we compared the four attributes for each pair of classes employing *t*-tests (table 1): both essential and persistent genes are highly biased in terms of these characteristics. In *B. subtilis*, the leading-strand preferences of essential and PNE genes are similar, whereas PNE genes are even more biased than ENP genes. In *E. coli*, essential genes have a higher leading-strand preference and show higher sequence conservation than PNE genes. The high relative frequency of lagging-strand ENP genes may be attributable to errors in the definition of essentiality and the low number of ENP genes. In *B. subtilis*, it is mostly attributable to the presence in the lagging strand of the *mrp* operon, probably not essential (Ito et al. 2000). In *B. subtilis* as in *E. coli*,

Table 1
Summary of PI, OLI, SD, and CAI Comparison Between Different Gene Sets

| | | P Value <i>t</i> -Test | | | | |
|--------------------------|------------|------------------------|------------|-----------|--------|--------|
| | | Mean | Persistent | Essential | PNE | ENP |
| <i>Bacillus subtilis</i> | | | | | | |
| PI | Persistent | 0.946 | — | — | — | — |
| | Essential | 0.851 | <0.001 | — | — | — |
| | PNE | 0.936 | 0.009 | <0.001 | — | — |
| | ENP | 0.557 | — | <0.001 | <0.001 | — |
| | NPNE | 0.278 | — | <0.001 | <0.001 | <0.001 |
| OLI | Persistent | 0.924 | — | — | — | — |
| | Essential | 0.918 | 0.691 | — | — | — |
| | PNE | 0.900 | 0.108 | 0.284 | — | — |
| | ENP | 0.815 | <0.001 | <0.001 | 0.003 | — |
| | NPNE | 0.720 | <0.001 | <0.001 | <0.001 | 0.017 |
| SD | Persistent | 0.678 | — | — | — | — |
| | Essential | 0.692 | 0.058 | — | — | — |
| | PNE | 0.653 | <0.001 | <0.001 | — | — |
| | ENP | 0.633 | <0.001 | <0.001 | 0.133 | — |
| | NPNE | 0.589 | <0.001 | <0.001 | <0.001 | 0.002 |
| CAI ^a | Persistent | 0.500 | — | — | — | — |
| | Essential | 0.515 | 0.022 | — | — | — |
| | PNE | 0.485 | 0.011 | <0.001 | — | — |
| | ENP | 0.489 | 0.280 | 0.017 | 0.662 | — |
| | NPNE | 0.463 | <0.001 | <0.001 | <0.001 | <0.001 |
| <i>Escherichia coli</i> | | | | | | |
| PI | Persistent | 0.956 | — | — | — | — |
| | Essential | 0.931 | 0.010 | — | — | — |
| | PNE | 0.945 | <0.001 | 0.171 | — | — |
| | ENP | 0.631 | — | <0.001 | <0.001 | — |
| | NPNE | 0.335 | — | <0.001 | <0.001 | <0.001 |
| OLI | Persistent | 0.669 | — | — | — | — |
| | Essential | 0.714 | 0.052 | — | — | — |
| | PNE | 0.625 | 0.023 | <0.001 | — | — |
| | ENP | 0.457 | <0.001 | <0.001 | 0.002 | — |
| | NPNE | 0.532 | <0.001 | <0.001 | <0.001 | 0.222 |
| SD | Persistent | 0.731 | — | — | — | — |
| | Essential | 0.755 | <0.001 | — | — | — |
| | PNE | 0.717 | 0.004 | <0.001 | — | — |
| | ENP | 0.719 | 0.417 | 0.038 | 0.902 | — |
| | NPNE | 0.701 | <0.001 | <0.001 | <0.001 | 0.423 |
| CAI ^a | Persistent | 0.486 | — | — | — | — |
| | Essential | 0.503 | 0.066 | — | — | — |
| | PNE | 0.478 | 0.249 | 0.009 | — | — |
| | ENP | 0.482 | 0.829 | 0.336 | 0.848 | — |
| | NPNE | 0.395 | <0.001 | <0.001 | <0.001 | <0.001 |

^a Ribosomal proteins were removed when performing *t*-test to avoid circularity (they are used to define CAI).

essential genes tend to have higher CAI and sequence conservation than PNE genes, but there is no significant difference between PNE and ENP genes.

Essential genes are highly conserved mostly because they also tend to be more expressed than average (Rocha and Danchin 2004). We tested if the same held true for PNE genes, using the previously published stepwise multiple regression technique where the rate of nonsynonymous substitution is modeled, taking into account the CAI and the PNE character of the gene (Rocha and Danchin 2004). In *B. subtilis* (respectively *E. coli*), regression of the two variables explained 34% (respectively 30%) of the variance. However, the inclusion of CAI alone explained 93% (respectively 98%) of the total variance (i.e., 93% and 98% of the 34% and 30% of the variance that is explained by the model). Hence, as for essentiality, persistence is not a very important determinant of nonsynon-

ymous substitution rates. The important sequence conservation of PNE genes stems probably from their high expression level. In summary, all observations converge to suggest that persistent genes and essential genes share very similar properties, in opposition to NPNE genes.

Clusters of Essential and PNE Genes

The colocalization of genes in bacterial genomes is important for their coexpression, proper localization, and function (Nitschke et al. 1998; Korbel et al. 2004; Reams and Neidle 2004). We explored whether essential and PNE genes were randomly distributed in the chromosome. To this aim, we computed the distribution of the shortest distances of consecutive genes of a given set in the genome and then compared it to the expected random distribution (fig. 2). A serial randomness test (Zar 1996) showed very

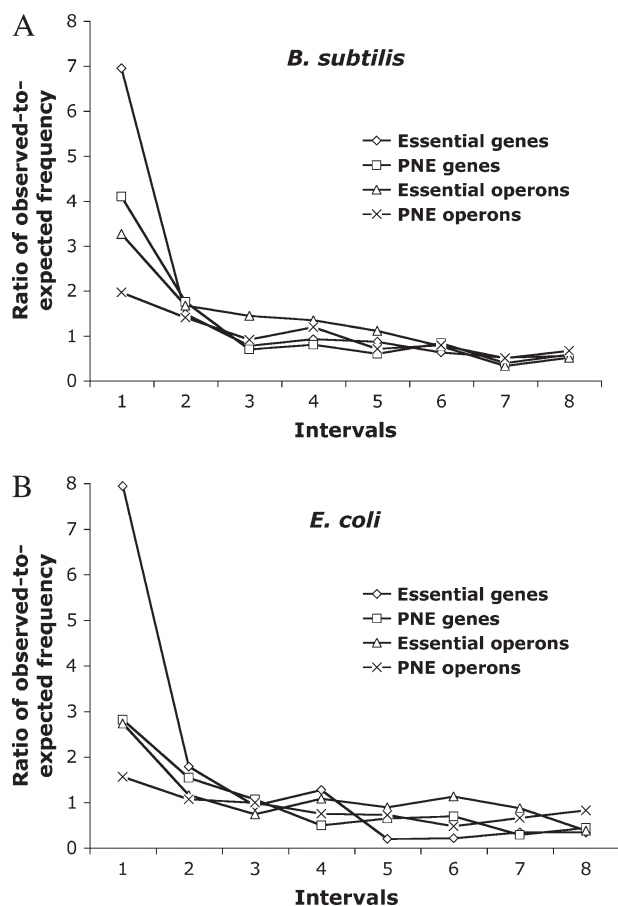


FIG. 2.—Essential and PNE genes cluster together. We calculated the intervals between every two adjacent experimentally defined essential genes. The most frequent interval was 1 (>45%). PNE genes, essential operons (operons with at least an essential gene), and essential PNE operons were analyzed in the same way. As a control, we randomized the distribution of the genes of each group 200 times, and the mean was used as expected value. We also plot the ratio of observed-to-expected frequency for each interval.

significant aggregation of essential and PNE genes both for *B. subtilis* and for *E. coli* ($P < 0.0001$ in the four tests). Hence, both gene sets are very significantly clustered, although the trend is slightly less important for PNE genes than for essential genes (fig. 2). Naturally, such genes can be close together in the chromosome simply because they are cotranscribed. To test if the operons containing essential and PNE genes were also clustered, we classified the genes into putative operons, using information on the gene orientation with respect to the origin of replication and on the presence of rho-independent terminators, identified by TransTerm (Ermolaeva et al. 2000), as in Rocha and Danchin (2003): 1,475 *E. coli* and 1,600 *B. subtilis* operons were predicted. Comparing these putative *E. coli* operons with RegulonDB (http://www.cifn.unam.mx/Computational_Genomics/regulondb/), which stores 2,328 operons, 2,153 (92.5%) were present in our data set (Salgado et al. 2004). Our prediction introduced a slight bias in classing sometimes two or more adjacent operons into a single one, but this renders our analysis more conservative. We considered an operon as essential if it contained an essential gene and an operon as PNE if it contained a PNE gene but no

essential gene. Thus, we obtained 126 (respectively 118) *B. subtilis* (respectively *E. coli*) essential operons and 182 (respectively 299) PNE operons. Both types of operons were significantly clustered in both genomes ($P \ll 0.0001$, serial randomness test). Based on this list, we conclude that although operons are responsible for part of the aggregation of essential and PNE genes, different operons containing such genes also tend to cluster together (fig. 2).

Functional Analysis of Persistent Genes

The previous analyses started from two model genomes, those of *B. subtilis* and *E. coli*. To evaluate the extent of the limitation introduced by using reference genomes and in an effort to retrieve the persistent genes from all the genomes that might be absent in *B. subtilis* or *E. coli*, we searched for orthologs for each pair of all 55 genomes included in this study. From each genome, we picked up genes with orthologs present in more than 95% of the genomes (this very stringent threshold is meant to limit the number of genes in the final “persistent” gene pool). We then pooled together the genes thus identified. We observed 257 persistent genes, which exist both in *B. subtilis* (for the Firmicutes) and *E. coli* (for the Gamma-proteobacteria). Among these, 144 (respectively 139) were previously identified as essential in *B. subtilis* (respectively *E. coli*) and 25 (respectively 18) of the 257 genes are not present in the 475 *B. subtilis* (respectively 611 *E. coli*) persistent genes. All the other members of the pool are PNE genes. These genes’ leading-strand preference (OLI), protein sequence similarity (SD), and codon usage (CAI) were qualitatively similar to the ones of the persistent genes identified using the *E. coli* and *B. subtilis* viewpoints (data not shown).

The functional classification of the foremost 257 persistent genes from all genomes is summarized in Supplementary Table 3 (Supplementary Material online). The proportions of essential and nonessential genes are presented in figure 3. Genes from the basal information transfer machinery compose the largest class: DNA replication, transcription, and translation (e.g., 14 of the 20 amino acid tRNA ligases in *E. coli*). Most of these genes are essential under laboratory growth conditions: 87 (72%) are identified as essential in both bacteria and a further 6 are specific to *E. coli* and 7 to *B. subtilis*. Essential genes also belong to a category implicated in compartmentalization. A third class, energy management and intermediary metabolism, is evenly split into essential and PNE genes. A fourth class, dominated by PNE genes, encompasses maintenance and stress response (here, maintenance genes refer to the genes which preserve the integrity of their product, such as genes involved in altered DNA, RNA, and protein degradation, proofreading, and mismatch repair).

Comparing the Two Clades

We subsequently looked for the genes classified as persistent in Firmicutes or in Gamma-proteobacteria but not in both. A total of 62 Firmicutes genes are present in more than 85% of the Firmicutes genomes and in less than 5% of Gamma-proteobacteria; all exist in *B. subtilis*.

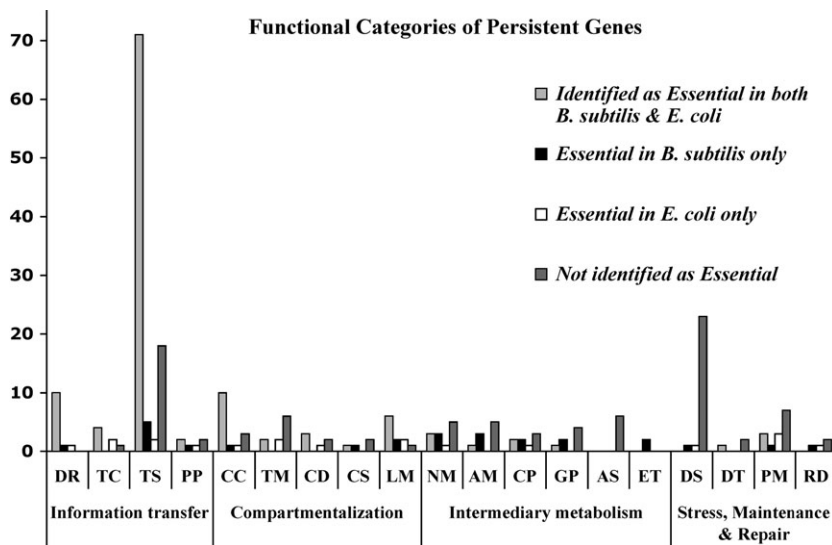


FIG. 3.—Functional categories of persistent genes. Abbreviations: DR, DNA replication; TC, transcription; TS, translation; PP, protein processing and modification; CC, cell wall and cell surface; TM, transports and membrane function; CD, cell division; CS, cell shaping; LM, lipid and membrane biosynthesis; NM, nucleotides metabolism; AM, amino acid and related molecules; CP, coenzymes and prosthetic groups; GP, glycolytic pathway; AS, ATP synthetase; ET, electron transfer; DS, DNA maintenance and other stress response; DT, DNA packaging and/or transcription regulation; PM, protein maintenance; and RD, RNA degradation.

Conversely, 124 genes from Gamma-proteobacteria belong to more than 85% of the genomes of the clade and to less than 5% of Firmicutes; they are all present in *E. coli* (Supplementary Table 3, Supplementary Material online).

Part of these differences may be attributable to sequence divergence, which can prevent the detection of orthologs. However, the analysis of this list highlights major differences between the two clades. For example, five Firmicutes-specific (*dnaI*, *parC*, *polC*, *yqeN*, and *yloA*) and four Gamma-proteobacteria-specific (*parC*, *parE*, *hola*, and *holC*) DNA replication-related genes are clade specific, confirming that the replication mechanisms adopted by Firmicutes and Gamma-proteobacteria differ considerably (Dervyn et al. 2001). A previous study already pointed out that *polC*, coding for the alpha subunit of DNA polymerase, may be a key character leading to the stronger strand bias in Firmicutes genes (Rocha 2002). Another interesting example is the ribosome: *rpsN* of *B. subtilis*, coding for ribosomal protein S14, is specific to Firmicutes. Its counterpart in *E. coli*, also named *rpsN*, has a different conserved homolog in *B. subtilis*, *yhza*. This suggests that in most Firmicutes an ancient duplication led to two S14 homologs.

Comparing Different Essentiality Data Sets

Genes experimentally defined as essential are contingent to a certain bacterium and set of methods used in the laboratory. For example, of the 277 *B. subtilis* and 235 *E. coli* essential genes, only 144 are reliable orthologs. About one-third (respectively two-third) of the *B. subtilis* (respectively *E. coli*) essential genes have reliable orthologs in all Firmicutes (respectively Gamma-proteobacteria). Considering the biases introduced by laboratory experiments (such as polarity in transcription when transposons are used for gene inactivation) (Salama, Shepherd, and Falkow 2004), the incongruence between essential gene

sets identified in different experiments would create havoc if the conclusions drawn using them were extended to other genomes. In PEC, essential genes were compiled using biological expertise. We illustrate in table 2 the comparison of PEC gene sets with another exploration of *E. coli* essential genes by a whole-genome-scale transposon gene inactivation experiment (Gerdes et al. 2003). The difference is remarkable because the latter essential genes are almost randomly separated into PEC essential, unknown, and nonessential categories. We evaluated the robustness of our results to changes in the classification of essential genes by analyzing the Gerdes classification (table 3). Despite the noteworthy incongruence between the two sets, the major conclusions remain the same: though Gerdes essential genes are less persistent, both essential and PNE genes have a more biased tendency to be coded in the leading strand compared to NPNE genes; PNE genes have a more conserved protein sequence than NPNE genes, but the SD difference between essential and NPNE genes is not significant; PNE genes have higher CAI values than those of essential and NPNE genes. Hence, the main conclusion that PNE genes have a prominent role in the cell and share many characteristics of essential genes is robust to the variations in the way essential genes are identified.

Table 2
Comparison of PEC and Gerdes' Essential *Escherichia coli* Genes

| Gerdes' | PEC | | |
|----------------------|-----------------|-----------------|----------------------|
| | Essential (235) | Unknown (1,675) | Nonessential (2,259) |
| Essential (601) | 158 | 175 | 268 |
| Nonessential (3,010) | 33 | 1,246 | 1,737 |
| Noncovered (306) | 23 | 157 | 126 |
| Ambiguous (187) | 21 | 87 | 79 |

Table 3
Comparison of Different Sets by Using
Gerdes' Essential Genes

| | | P Value <i>t</i> -Test | | | | |
|------------------|------------|------------------------|------------|-----------|--------|--------|
| | | Mean | Persistent | Essential | PNE | ENP |
| PI | Persistent | 0.956 | — | — | — | — |
| | Essential | 0.629 | <0.001 | — | — | — |
| | PNE | 0.947 | 0.007 | <0.001 | — | — |
| | ENP | 0.381 | <0.001 | <0.001 | <0.001 | — |
| | NPNE | 0.334 | <0.001 | <0.001 | <0.001 | <0.001 |
| OLI | Persistent | 0.671 | — | — | — | — |
| | Essential | 0.618 | 0.005 | — | — | — |
| | PNE | 0.642 | 0.162 | 0.277 | — | — |
| | ENP | 0.546 | <0.001 | 0.002 | <0.001 | — |
| | NPNE | 0.537 | <0.001 | <0.001 | <0.001 | 0.68 |
| SD | Persistent | 0.738 | — | — | — | — |
| | Essential | 0.714 | <0.001 | — | — | — |
| | PNE | 0.736 | 0.715 | <0.001 | — | — |
| | ENP | 0.693 | <0.001 | 0.013 | <0.001 | — |
| | NPNE | 0.707 | <0.001 | 0.144 | <0.001 | 0.078 |
| CAI ^a | Persistent | 0.486 | — | — | — | — |
| | Essential | 0.439 | <0.001 | — | — | — |
| | PNE | 0.479 | 0.346 | <0.001 | — | — |
| | ENP | 0.402 | <0.001 | <0.001 | <0.001 | — |
| | NPNE | 0.396 | <0.001 | <0.001 | <0.001 | 0.266 |

^a Ribosomal proteins were removed when performing *t*-test to avoid circularity (they are used to define CAI).

Discussion

Our analyses were based on orthologs identified using BBH. We tried to evaluate the limitations created by this methodology. Firstly, when very similar, duplicated genes may be missed in the BBH analysis. The number of duplicated genes (genes with >80% protein sequence similarity and <1.3 length difference within the same genome) is 109 in *E. coli* and 56 in *B. subtilis*. Taking the duplicated genes separately, we identified the number of putative orthologs in the other genomes. We found 6 genes in *B. subtilis* and 12 in *E. coli* with a homolog in at least 85% of the genomes (Supplementary Table 4A and B, Supplementary Material online). These should be compared with, respectively, 475 and 611 persistent genes for each genome. Secondly, to evaluate the impact of fast evolution in missing orthologs, we compared the bacteria at the farthest 16S rRNA distance (*Clostridium acetobutylicum* for *B. subtilis* and *Francisella tularensis* for *E. coli*) (Supplementary Table 5, Supplementary Material online). We estimate that 2 (0.4%) out of the 475 *B. subtilis* and 3 (0.5%) out of the 611 *E. coli* persistent genes evolve fast enough that some orthologs may have been missed. Among experimentally essential genes, 18 (6.6%) out of the 272 *B. subtilis* and 10 (4.3%) out of the 235 *E. coli* essential genes were lost by fast evolution. In the whole genome, we estimated 863 (21.1%) of *B. subtilis* and 1,066 (25.6%) of *E. coli* genes as fast-evolving genes. Our analyses involve cross-species comparisons of individual genes and infer connections between gene essentiality, persistence, and natural environmental conditions. They do not focus on constraints arising from the genomic environment and may miss essential genes which, because they coevolve as families, would not present a strong sequence conservation.

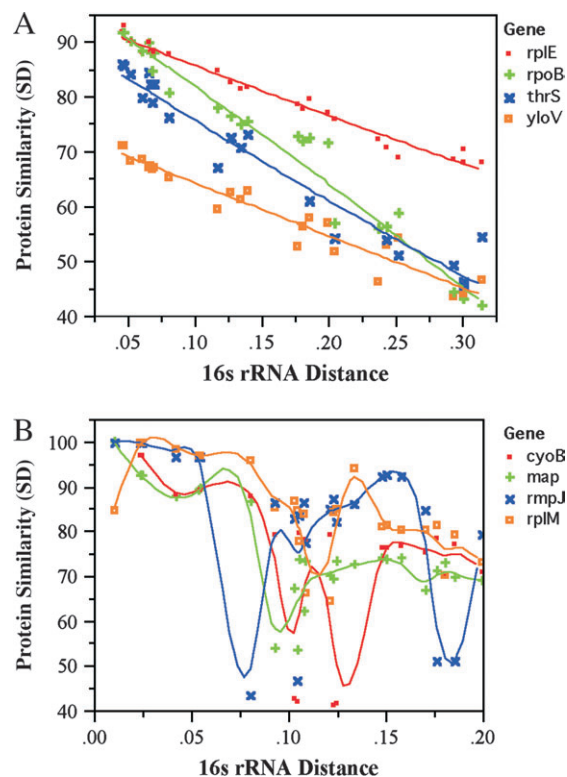


Fig. 4.—Evolutionary divergence of typical genes extracted from the sets identified in this work. The SD of genes usually changes proportionally with the phylogenetic distance (A). Note that their evolution speed differs: this implies that some genes evolving fast may not have been retained in the present study. In contrast, some genes have an erratic evolutionary pattern (B), suggesting horizontal gene transfer, competition with genes recruited for the same function, recruitment for other functions, etc.

Using four semiempirical indices to reflect the evolutionary properties of genes in bacterial genomes, we found that PNE genes, just like essential genes, tend to be coded in the leading strand. If the degree of strand bias is a measure of gene essentiality, this shows how important this set is. We also found that PNE genes are expressed more than average, almost as much as essential genes. Previous studies have pointed out that expression levels correlate with protein amino acid substitution rates (Sharp 1991; Rocha and Danchin 2004). Here we find that PNE genes are more conserved than average but that, as for essential genes, this results mostly from their high degree of expression. Hence, we propose that gene persistence should be considered as the major factor in describing gene's evolutionary attributes.

Some of the genes missing from the list of persistent genes and especially in the comparison between Gamma-proteobacteria and Firmicutes have diverged considerably. To assess the contribution of this effect, we measured for each pair of genomes the correlation between the similarity of orthologous pairs and that of the 16S rRNA. As expected, the correlations were high. For example (fig. 4A), 38% (respectively 48%) of *B. subtilis* (respectively *E. coli*) persistent genes showed a correlation coefficient >0.9 between the sequence similarity of the pair of orthologs and the 16S (Supplementary Table 2A and B, Supplementary

Material online). After removing outliers (two genomes at most), this percentage increased to 62% (*B. subtilis*) and 77% (*E. coli*). In contrast, some genes (fig. 4B) evolve in an erratic way. This may be due to horizontal gene transfer, local adaptations leading to faster or slower evolutionary pace, or simply wrong assignments of orthology. The last can be a significant problem, especially in large protein families. However, the genes presenting such an erratic pattern are rare in the persistent set. For example, only 13 genes out of the 611 *E. coli* persistent genes have correlation coefficients lower than 0.7. Interestingly, four *E. coli* ribosomal proteins belong to this group: *rpmJ*, *rplM*, *rpsJ*, and *rpmB*. *Bacillus subtilis* shows a somewhat larger number of such genes with a correlation coefficient lower than 0.7 (69 out of 475). However, Firmicutes make a broader clade than Gamma-proteobacteria, and the assignment of orthology is thus less accurate.

Exploration of gene persistence unveiled the unexpected importance of many genes asking for experimental answers (Supplementary Table 3, Supplementary Material online). The structure of the ribosome is fairly universal (Moore and Steitz 2003). While its core structure, with its RNA machinery, is widely conserved in bacteria and evolves slowly, some components are more prone to variation. In *E. coli*, 48 of the 54 ribosomal proteins belong to the persistent class. They are generally labeled as essential despite experiments showing the contrary for some of these (Dabbs 1978). The exceptions are genes *rpsA*, *rpsF*, *rpsN*, *rplY*, *rpmF*, and *rpmH*. Protein S1 (RpsA) has a special status. It is loosely bound to the ribosome and does not have a ribosome-bound counterpart in *B. subtilis*. It is involved in the recognition of the ribosome-binding sites in Gamma-proteobacteria mRNAs and has been shown to be part of the degradosome in *E. coli* (Noria and Danchin 2002). In *Lactobacillus plantarum*, *rpsF* has been involved in stress response (De Angelis et al. 2004). Ribosomal protein S14 is exceptional in both Firmicutes and Gamma-proteobacteria. Its evolution pattern is erratic, perhaps as a result of adaptation to antibiotics (Brochier, Philippe, and Moreira 2000). RplY (L25) is coded by a gene associated to a stress response: its counterpart in *B. subtilis* and other Firmicutes belongs to the CTC protein family, a family of general stress proteins (Schmalisch, Langbein, and Stulke 2002; Gongadze et al. 2005). *rpmF*, *rpmG*, and *rpmH* code for ribosomal proteins L32, L33, and L34, respectively. *rpmF* is cotranscribed with a gene encoding a protein involved in membrane lipid synthesis and several fatty acid biosynthetic genes, indicating multiple functions besides translation (Podkovyrov and Larson 1995). L33 is dispensable under laboratory growth conditions (Maguire and Wild 1997). L34 is unusual: in the genus *Thermus*, *rpmH* is a gene within a gene, overlapping with the unusually large *rnpA* (RNase P protein subunit) sequence, suggesting some sort of adaptation (Ellis and Brown 2003). L34 has also been shown to regulate biosynthesis of polyamines, metabolites that play an important role in stress (Cohen 1998).

These observations point to a major role of adaptation to different ecological niches in the phylogenetic divergence of components of the basal biosynthetic machinery. Essentiality and persistence would be separated each time

a component of an otherwise essential product would be recruited in another process, regulation in particular, while keeping its former function. For example, the remarkable translational repression mediated by threonine tRNA ligase in *E. coli* is absent from the close organism *Salmonella typhimurium* (Romby and Springer 2003).

The difference between the very distant Firmicutes and Gamma-proteobacteria is particularly informative in that it provides us with a pair of self-consistent evolutionary diverging processes essential to life in the wild. Three major processes are prominent (Supplementary Table 3, Supplementary Material online): cell envelope biosynthesis, DNA synthesis and maintenance, and RNA metabolism. Firmicutes and Gamma-proteobacteria differ considerably in their envelope: this is reflected in the larger number of Gamma-proteobacteria-specific persistent genes devoted to the outer membrane and lipopolysaccharides synthesis. The organization of DNA polymerase differs in both classes with two DNA polymerase III genes in Firmicutes (Dervyn et al. 2001), and divergent delta subunits (YqeN; Kobayashi et al. 2003) in Firmicutes differ from HoIA in Gamma-proteobacteria. Another remarkable difference between Firmicutes and Proteobacteria is in RNA processing and degradation. For example, *rne* (ribonuclease E), *rnt* (ribonuclease T), and *orn* (oligoribonuclease) are specific to Gamma-proteobacteria. While RNA degradation is essential, their counterparts in Firmicutes have not yet been identified. It is tempting to suggest their counterpart from the list of genes of unknown function: *yazC*, *yhaM*, *ykqC*, *ylbM*, *ymdA*, *yrzB*, and *yvcL*. YhaM is the likely counterpart of Orn (Oussenko, Sanchez, and Bechhofer 2002). *yazC* codes for a protein with features in common with ribonuclease III. Its conserved location in an operon regulated by an S-box riboswitch suggests that it is involved in processing them. YkqC codes for a metallohydrolase that could fit one of the missing functions. YlbM has no significant similarity with known functions. YmdA is a putative hydrolase: we have thus two putative candidates for the counterpart of ribonuclease E, YkqC and YmdA, that could fit one of the missing functions. *yrzB* is in an operon with a possible nuclease; it could code one of the missing functions. Finally, we remark the presence of an unexpected set of persistent genes involved in the metabolism of serine or serine-like amino acids: *sdaAA* and *sdaAB* in Gamma-proteobacteria and *sdaA*, *sdaB*, and *tdcG* in Firmicutes, with *ydfG* present in both classes. This may be related to the ubiquitous "serine effect" (growth inhibition of bacteria when serine is added under particular conditions) (Uzan and Danchin 1976), still not understood but, as emphasized by the presence of these genes here, probably corresponding to a universal and essential feature of bacterial metabolism (Inoue et al. 2002).

Most of the persistent essential genes belong to the information transfer class (fig. 3), while the majority of genes involved in maintenance and stress response are PNE genes (Supplementary Table 3, Supplementary Material online). These figures strongly support the hypothesis that laboratory experiments could not detect genes that are essential in situations of stress or starvation or genes associated with maintenance because their absence still allows residual growth, while they are fundamental in the long term.

Supplementary Material

Supplementary Tables 1–5 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

G.F. was supported by the Friends of the Institute Pasteur in Hong Kong. This work benefited from the Bio-Sapiens European Union grant LSHG-CT-2003-503265 and from the French Ministry of Research ACI IMPBio BlastSets program. We also thank two anonymous referees.

Literature Cited

- Akerley, B. J., E. J. Rubin, V. L. Novick, K. Amaya, N. Judson, and J. J. Mekalanos. 2002. A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. USA* **99**:966–971.
- Bentley, S. D., and J. Parkhill. 2004. Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* **38**:771–792.
- Brochier, C., H. Philippe, and D. Moreira. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet.* **16**:529–533.
- Chalker, A. F., and R. D. Lunsford. 2002. Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. *Pharmacol. Ther.* **95**:1–20.
- Cohen, S. 1998. A guide to the polyamines. Oxford University Press, Oxford.
- Dabbs, E. R. 1978. Mutational alterations in 50 proteins of the *Escherichia coli* ribosome. *Mol. Gen. Genet.* **165**:73–78.
- De Angelis, M., R. Di Cagno, C. Huet, C. Creccchio, P. F. Fox, and M. Gobetti. 2004. Heat shock response in *Lactobacillus plantarum*. *Appl. Environ. Microbiol.* **70**:1336–1346.
- Dervyn, E., C. Suski, R. Daniel, C. Bruand, J. Chapuis, J. Errington, J. Janniere, and S. D. Ehrlich. 2001. Two essential DNA polymerases at the bacterial replication fork. *Science* **294**:1716–1719.
- Ellis, J. C., and J. W. Brown. 2003. Genes within genes within bacteria. *Trends Biochem. Sci.* **28**:521–523.
- Ermolaeva, M. D., H. G. Khalak, O. White, H. O. Smith, and S. L. Salzberg. 2000. Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.* **301**:27–33.
- Forsyth, R. A., R. J. Haselbeck, K. L. Ohlsen et al. (23 co-authors). 2002. A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**:1387–1400.
- Galperin, M. Y., and E. V. Koonin. 2004. ‘Conserved hypothetical’ proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* **32**:5452–5463.
- Gerdes, S. Y., M. D. Scholle, J. W. Campbell et al. (21 co-authors). 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**:5673–5684.
- Gongadze, G. M., A. P. Korepanov, E. A. Stolboushkina et al. (11 co-authors). 2005. The crucial role of conserved intermolecular H-bonds inaccessible to the solvent in formation and stabilization of the TL5-5S rRNA complex. *J. Biol. Chem.* **280**:16151–16156.
- Hegeman, G. D., and S. L. Rosenberg. 1970. The evolution of bacterial enzyme systems. *Annu. Rev. Microbiol.* **24**:429–462.
- Hutchison, C. A. III, S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, and J. C. Venter. 1999. Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**:2165–2169.
- Inoue, K., J. Chen, I. Kato, and M. Inouye. 2002. Specific growth inhibition by acetate of an *Escherichia coli* strain expressing Era-dE, a dominant negative Era mutant. *J. Mol. Microbiol. Biotechnol.* **4**:379–388.
- Ito, M., A. A. Guffanti, W. Wang, and T. A. Krulwich. 2000. Effects of nonpolar mutations in each of the seven *Bacillus subtilis* mrp genes suggest complex interactions among the gene products in support of Na⁺ and alkali but not cholate resistance. *J. Bacteriol.* **182**:5663–5670.
- Jacobs, M. A., A. Alwood, I. Thaipisuttikul et al. (15 co-authors). 2003. Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* **100**:14339–14344.
- Jensen, R. A. 1976. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**:409–425.
- Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**:962–968.
- Klasson, L., and S. G. E. Andersson. 2004. Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol.* **12**:37–43.
- Kobayashi, K., S. D. Ehrlich, A. Albertini et al. (99 co-authors). 2003. Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* **100**:4678–4683.
- Koonin, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**:127–136.
- Korbel, J. O., L. J. Jensen, C. von Mering, and P. Bork. 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22**:911–917.
- Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**:2229–2235.
- Lawrence, J. G., and H. Hendrickson. 2003. Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50**:739–749.
- Lerat, E., V. Daubin, and N. A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the Gamma-proteobacteria. *PLoS Biol.* **1**:1–9.
- Lerat, E., V. Daubin, H. Ochman, and N. A. Moran. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* **3**:e130.
- Lobry, J. R. 1996. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie* **78**:323–326.
- Maguire, B. A., and D. G. Wild. 1997. The effects of mutations in the rpmB,G operon of *Escherichia coli* on ribosome assembly and ribosomal protein synthesis. *Biochim. Biophys. Acta* **1353**:137–147.
- Maniloff, J. 1996. The minimal cell genome: “on being the right size.” *Proc. Natl. Acad. Sci. USA* **93**:10004–10006.
- Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**:589–596.
- Moore, P. B., and T. A. Steitz. 2003. The structural basis of large ribosomal subunit function. *Annu. Rev. Biochem.* **72**:813–850.
- Mushegian, A. R., and E. V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93**:10268–10273.
- Nitschke, P., P. Guerdoux-Jamet, H. Chiapello, G. Faroux, H. Henaut, A. Henaut, and A. Danchin. 1998. Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol. Rev.* **22**:207–227.
- Noria, S., and A. Danchin. 2002. Just so genome stories: what does my neighbor tell me? Pp. 3–13 in H. Yoshikawa,

- N. Ogasawara, and N. Satoh, eds. Proceedings of the Uehara Memorial Foundation Symposium; Genome Science: towards a new paradigm? Elsevier Science BV, Tokyo.
- Oussenko, I. A., R. Sanchez, and D. H. Bechhofer. 2002. *Bacillus subtilis* YhaM, a member of a new family of 3'-to-5' exonucleases in gram-positive bacteria. *J. Bacteriol.* **184**:6250–6259.
- Papp, B., C. Pal, and L. D. Hurst. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**:661–664.
- Podkovyrov, S., and T. J. Larson. 1995. Lipid biosynthetic genes and a ribosomal protein gene are cotranscribed. *FEBS Lett.* **368**:429–431.
- Reams, A. B., and E. L. Neidle. 2004. Selection for gene clustering by tandem duplication. *Annu. Rev. Microbiol.* **58**:119–142.
- Rocha, E. P. C. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.* **10**:393–395.
- . 2003. DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet.* **19**:600–603.
- Rocha, E. P. C., and A. Danchin. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet.* **34**:377–378.
- . 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21**:108–116.
- Romby, P., and M. Springer. 2003. Bacterial translational control at atomic resolution. *Trends Genet.* **19**:155–161.
- Salama, N. R., B. Shepherd, and S. Falkow. 2004. Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.* **186**:7926–7935.
- Salgado, H., S. Gama-Castro, A. Martinez-Antonio et al. (11 co-authors). 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32**:D303–D306.
- Sankoff, D. 2003. Rearrangements and chromosomal evolution. *Curr. Opin. Genet. Dev.* **13**:583–587.
- Schmalisch, M., I. Langbein, and J. Stulke. 2002. The general stress protein Ctc of *Bacillus subtilis* is a ribosomal protein. *J. Mol. Microbiol. Biotechnol.* **4**:495–501.
- Sharp, P. M. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *J. Mol. Evol.* **33**:23–33.
- Sharp, P. M., and W.-H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.
- Stepkowski, T., and A. B. Legocki. 2001. Reduction of bacterial genome size and expansion resulting from obligate intracellular lifestyle and adaptation to soil habitat. *Acta Biochim. Pol.* **48**:367–381.
- Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. *Science* **278**:631–637.
- Uzan, M., and A. Danchin. 1976. A rapid test for the *relA* mutation in *E. coli*. *Biochem. Biophys. Res. Commun.* **69**:751–758.
- Zar, J. H. 1996. *Biostatistical analysis*. Prentice-Hall International Limited, London.

Edward Holmes, Associate Editor

Accepted June 30, 2005