Research

The source of laterally transferred genes in bacterial genomes

Vincent Daubin[¤*†], Emmanuelle Lerat[¤*] and Guy Perrière[*]

Addresses: [*]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard - Lyon 1, 43 Bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France. [†]Current address: c/o Ochman, Department of Biochemistry and Molecular Biophysics, 233 Life Sciences South, University of Arizona, Tucson, Arizona 85721, USA.

¤ These authors contributed equally to this work.

Correspondence: Vincent Daubin. E-mail: daubin@email.arizona.edu

Published: 21 August 2003

Genome ***Biology*** 2003, **4:**R57

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/R57

Received: 16 April 2003
Revised: 11 June 2003
Accepted: 4 July 2003

Abstract

Background: Laterally transferred genes have often been identified on the basis of compositional features that distinguish them from ancestral genes in the genome. These genes are usually A+T-rich, arguing either that there is a bias towards acquiring genes from donor organisms having low G+C contents or that genes acquired from organisms of similar genomic base compositions go undetected in these analyses.

Results: By examining the genome contents of closely related, fully sequenced bacteria, we uncovered genes confined to a single genome and examined the sequence features of these acquired genes. The analysis shows that few transfer events are overlooked by compositional analyses. Most observed lateral gene transfers do not correspond to free exchange of regular genes among bacterial genomes, but more probably represent the constituents of phages or other selfish elements.

Conclusions: Although bacteria tend to acquire large amounts of DNA, the origin of these genes remains obscure. We have shown that contrary to what is often supposed, their composition cannot be explained by a previous genomic context. In contrast, these genes fit the description of recently described genes in lambdoid phages, named 'morons'. Therefore, results from genome content and compositional approaches to detect lateral transfers should not be cited as evidence for genetic exchange between distantly related bacteria.

Background

The G+C content of a genome and the codon usage of its genes are determined by selection and mutation pressures [1]. Because these evolutionary processes are characteristic of each species, the sequences belonging to a genome share a common pattern of composition of bases, codons and oligonucleotides [2,3], making it possible to identify laterally transferred genes (LTGs) as those whose features are atypical for a particular genome. Thus genes displaying atypical composition or vocabulary are inferred to be alien, and to carry features of their previous genome [4]. However, it is thought that only recently acquired genes would be detected by this approach because sequences quickly adjust to their new genome pattern.

Since the inception of these approaches, it has been noted that alien genes tend to display lower G+C contents than their new host genome [4-7]. Médigue *et al.* [5] analyzed the codon usage of the genes of *Escherichia coli* using a multivariate analysis and found that the genome can be separated into three gene classes according to codon usage. The first class corresponds to highly expressed genes, the second to weakly expressed genes, and the third to genes with unknown function, insertion sequences (IS), phage, and genes possibly related to virulence and antibiotic resistance. Therefore, this last class has been interpreted as the class of genes recently acquired by horizontal transfer.

The fact that recently acquired genes all group together in this analysis implies that they are relatively homogeneous in their codon usage. Although not pointed out by Médigue *et al.* [5], this result is surprising, because these genes are thought to have been acquired through several independent events from different species, and therefore should display very different codon-usage patterns and be separated by the analysis. Other methods based on compositional analysis often show the same result: that is, recently acquired genes tend to share characteristics such as codon usage and G+C content [4,6].

It has thus been argued that the methods used to identify LTGs are unable to detect genes acquired from donors having similar base composition and that the amount of LTGs, although representing a substantial fraction of the genome following their predictions, is yet highly underestimated [4,8]. Moreover, as noted by Lawrence and Ochman [4]: "Since base composition (...) is conserved within and among related lineages, genes with anomalous features are likely to have been acquired recently from distantly related organisms", and genes displaying atypical composition are indeed usually interpreted as such [4,9-12]. In this view, gene exchanges would be very frequent, not only between species but among orders or phyla. Indeed, base composition would hardly allow identification of a gene acquired from *Salmonella* in the *E. coli* genome, despite the fact that these bacteria may have diverged 100 million years ago. Such reasoning relies upon the untested postulate that these genes carry the mark of a previous host genome. It is well known, however, that some elements or regions of bacterial and eukaryotic genomes show systematic, and probably persistent, compositional differences to the rest of the genome. This has been shown for transposable elements, viruses and plasmids [13,14] and for the region of the replication terminus in numerous bacteria [15]. Therefore, the observed peculiarities of LTGs may represent, rather than a previous genome context, the mark of a particular 'lifestyle' or local effect acting on the gene.

Here we address these problems by studying the codon usage and base composition of recently acquired genes detected by an approach based on complete genome comparisons. We show that recently acquired genes tend to have a composition that is shifted toward A+T compared to their hosts, even in A+T-rich genomes. This suggests that LTGs detected by compositional methods are not highly underestimated. We moreover show that the hypothesis of an adaptation to a previous genome context hardly explains the codon and base composition of these alien genes. Therefore, we propose that peculiar evolutionary pressures acting on these genes are responsible for their atypical composition. Hence, the large majority of LTGs detected by compositional approaches do not necessarily originate in distant organisms.

Results

Transfers or losses?

We inferred the numbers of gene acquisitions and losses using the method described in Figure 1 and in the Materials and methods section. Figure 2 shows the number of lost and acquired genes estimated, for each group of genomes considered. Because of the stringency of the BLAST criteria used, these numbers are possibly underestimates. However, they give information about the dynamics of the different genomes. In all cases, the number of acquired genes is higher than the number of genes losses in the branch of the sister grouping. This allows interpretation of genes unique to a lineage as recent gains, rather than two independent losses. In contrast, we cannot exclude the possibility that some inferred gene losses correspond to independent gene acquisition (although the probability of two independent acquisitions of the same gene is difficult to estimate).

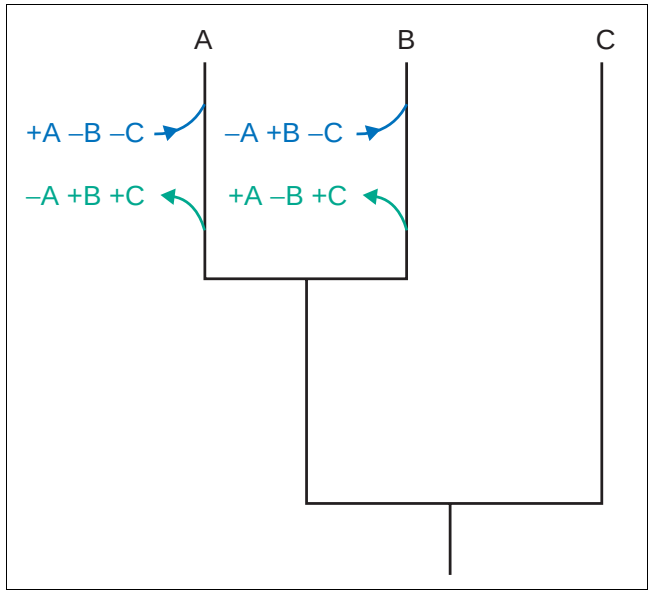

Figure 1
Principle of the detection of recently acquired and lost genes using parsimony. Genes present in species A and absent from species B and C (+A -B -C) are likely to have been acquired recently if the number of lost genes in the sister species (+A -B +C) is relatively small.

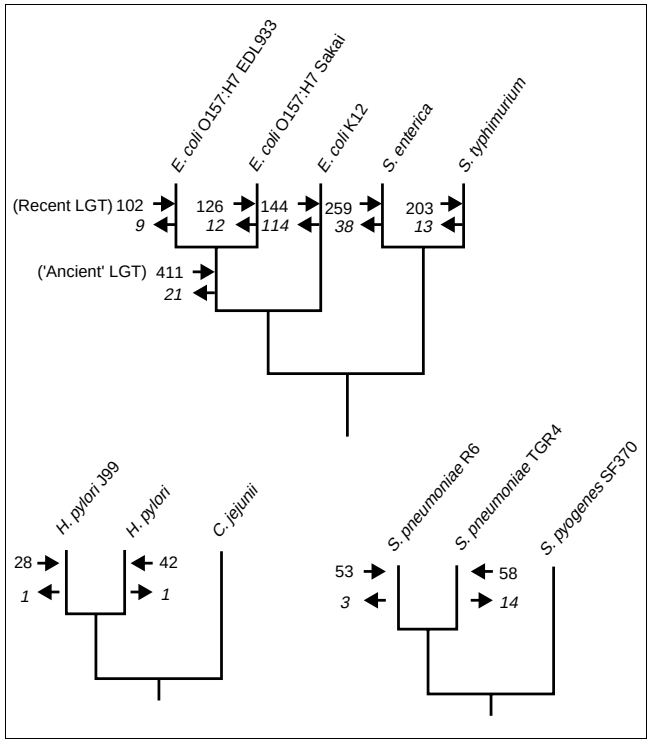

Figure 2
Results of the approach described in Figure 1 in three groups of closely related bacteria. Italic numbers refer to lost genes. A list of the acquired genes is available as an Additional data file.

In most cases, the number of acquired genes is higher than the number of 'lost genes' in the same branch. Two phenomena, not mutually exclusive, may explain these differences: an increase in the size of the genome (this is probably the case for the pathogenic *E. coli* strains, as their genomes contain many more genes than the K12 strain); and a high turnover of acquired genes in the genome. Indeed, the complete sequence represents a 'snapshot' of the genome in which many of the recently acquired genes may be destined to disappear quickly, while the 'lost genes' detected by the method have been conserved during relatively long periods of time in the two other lineages (Figure 3).

Most of the acquired genes have no known functions, though a few are annotated as membrane proteins, phages or IS. In the following results, the genes from these two last classes will appear in the phages and IS classes rather that in the LTG class.

The codon usage of LTGs: comparison with native genes

We computed four independent factorial correspondence analyses (FCA) on the genes of each type (native and transferred genes, IS, and phages) for the four species *E. coli* O157:H7, *Helicobacter pylori*, *Salmonella enterica*, and *Streptococcus pneumoniae*. Figure 4 shows the projection of the genes and the codons on the two first axes for *E. coli*, *Salmonella*, *S. pneumoniae* and *H. pylori*. The codons have been labeled according to their third position. In each case, native genes and LTG form distinct groups (MANOVA test, $p < 10^{-4}$). In *E. coli* and *Salmonella*, the codon projections reveal that the first axis was principally due to G+C content, the laterally transferred genes being A+T-rich. This analysis shows that the criterion used by Médigue *et al.* [5] probably allows identification of most of the recent LTG, as only a few LTG detected independently from codon usage are in the native genes cloud. These last genes may therefore display the codon usage of a closely related species or strain. In *Helicobacter*, the same pattern is observed but with a stronger opposition of A-ending and C-ending codons. In *Streptococcus*, the A+T3/C+G3 separation appeared principally on the second axis. In each case, the ATA codon (isoleucine) is systematically separated from the others on the first axis, suggesting that in all cases, this codon is over-represented in LTGs, compared to native genes. Two other codons show a similar pattern: AGA and - except in *H. pylori* - AGG, the two coding for arginine. These three codons seem to be the principal ones leading to the separation between native and transferred genes in these analyses.

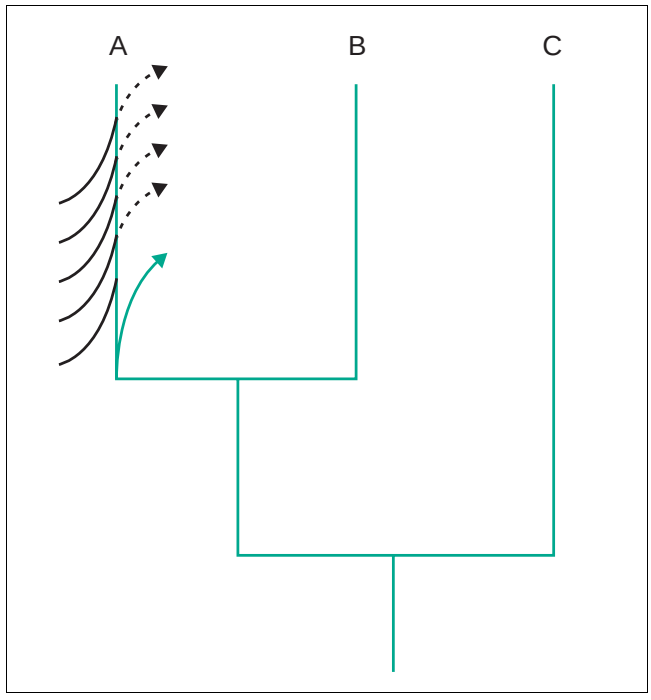

Figure 3
Gene acquisitions and losses. The method described here (see Figure 1) only identifies losses of genes (in genome A) that have been conserved in the two other lineages considered (genomes B and C; in green). If recent acquisitions (in black) are deleted shortly after their integration in the genome, what we observe is a high number of acquisitions compared with losses. This, rather than an increase in genome size, may explain the results presented in Figure 2.

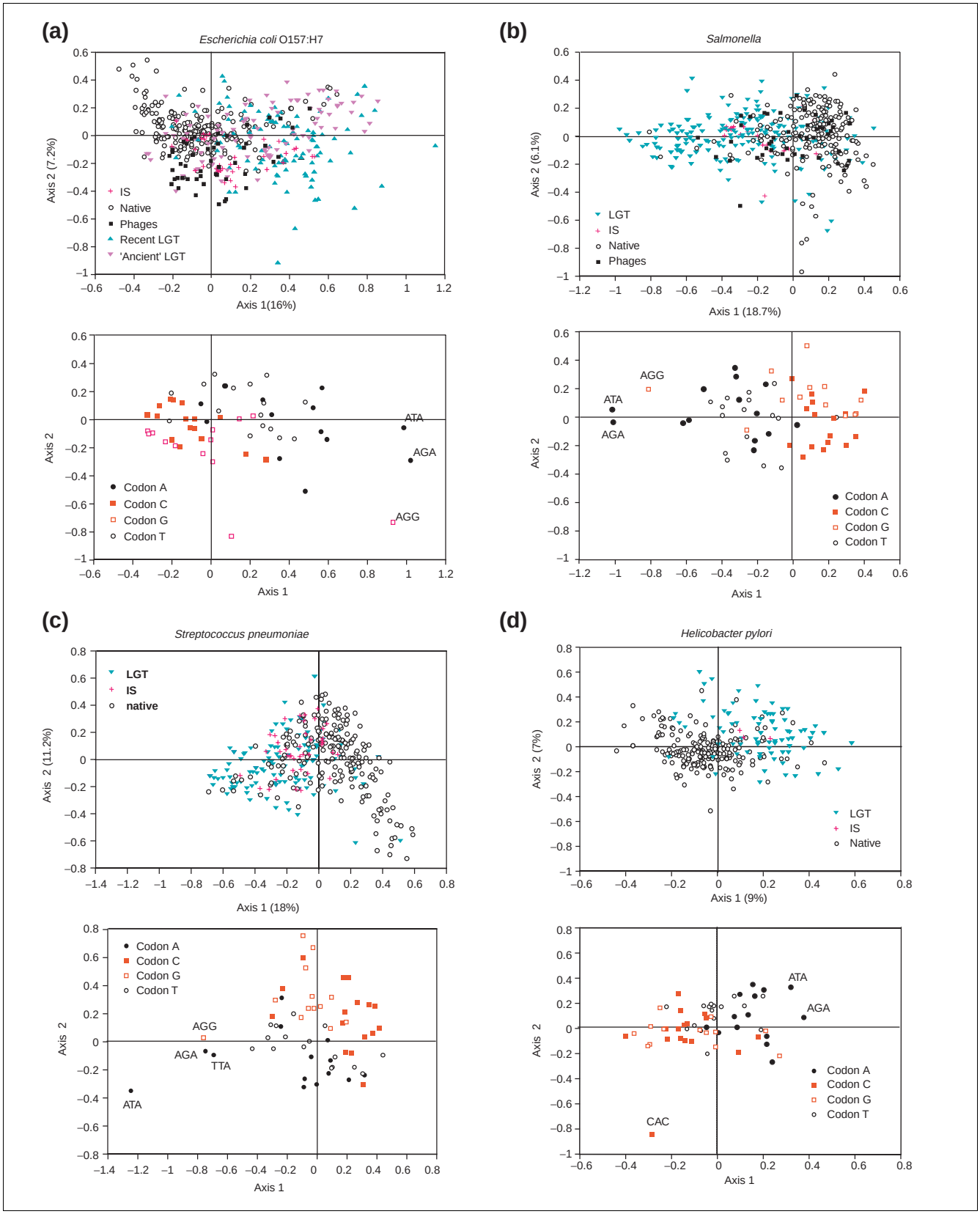

Figure 4 *(see legend on next page)*

Figure 4
Intraspecies FCA. **(a)** *E. coli*; **(b)** *Salmonella enterica*; **(c)** *S. pneumoniae*; and **(d)** *H. pylori*. Both genes (top) and codons (bottom) are plotted on the two first axes of the FCA. Codons are labeled according to the nature of the base at the third position. The percentages of variability explained by the axes are shown between brackets.

Table 1 shows the relative frequencies of the codons of isoleucine (I) and arginine (R) for all the native and transferred genes, IS, and phages for each species. In enterobacteria, the native genes generally avoid the three codons ATA, AGA, and AGG, while the transferred genes show little or no codon bias for the corresponding amino acids. This is also true in *Streptococcus*, although the AGA codon seems to be rather overrepresented in LTGs. This codon is even more frequent in *Helicobacter* LTGs.

A few genes undetected as LTGs by our method, are, however, localized in the cloud of points of the transferred genes. The functions of these genes indicated that they could indeed be transferred genes, acquired before the divergence of the genomes considered. We found, for example, membrane proteins related to the virulence or secretory systems. In *Streptococcus* and *Helicobacter*, we found restriction enzymes and transcription regulators. Interestingly, among these genes we identified a gene coding a ribosomal protein (RPS14) in *Helicobacter*. On the basis of phylogenetic analysis this peculiar gene has been shown to be subject to extensive lateral gene transfer in the proteobacteria group, and could be involved in antibiotic resistance [16].

Comparisons among species

Figure 5 shows the first two axes of an FCA performed on the four species. All the figures can be superimposed, but they have been separated according to gene classes (native genes, transferred genes, and IS). Figure 5d shows the projection of the codons on the same axes. Phages are not represented because they are absent from the *Helicobacter* and *Streptococcus* genomes. For a given species, each class of genes is represented by ellipses that enclose 90% of the points. A MANOVA shows that all groups are significantly coherent and different from each other ($p < 10^{-4}$). This confirms that LTGs in a species tend to use a relatively similar codon usage. The separation on the first axis is mainly due to the base composition, that is, A+T-rich and G+C-rich codons (Figure 5d).

The center of each ellipse is indicated by a color point. The arrows show the displacement observed relative to the position of the native gene ellipses. Transferred genes are systematically displaced in the direction of A+T-rich codons. The IS ellipses display a similar shift toward A+T-rich codons. However, although LTGs from different species show a tendency to group together, they tend to have codon usages comparable to their host genomes.

The base composition of LTGs

We computed the base composition of each gene class for the different species. The G+C content of LTGs is significantly lower than the native genes (Mann-Whitney test, $p < 10^{-4}$) at each codon position, and particularly at the third (G+C3) (Figure 6). This result is unexpected, especially for *Streptococcus* and *Helicobacter*, which have low G+C content (35% and 41% G+C3 respectively). Thus, whatever the base composition of a genome, the acquired genes are more A+T-rich than their host genome. Moreover, when it was possible to measure the amount of lost genes (that is, in enterobacteria), we have found that they also tend to be more A+T-rich than the genome (Mann-Whitney test, $p < 10^{-4}$; results not shown), suggesting a greater turnover of A+T-rich genes.

Selection on the different classes of genes

Figure 7 shows the relative neutrality plot (RNP) for each gene class in *E. coli* O157:H7. As expected for genes undergoing strong selection pressures, native genes show a low slope in the regression plots (0.241; $r^2 = 0.212$). The most recent LTGs display the highest slope (0.568; $r^2 = 0.446$), followed by more ancient LTGs (0.451; $r^2 = 0.553$), suggesting that the base composition of nonsynonymous sites in LTG is mainly the result of mutational pressures, and hence that their amino-acid composition is exceptionally affected by the constraints acting on the nucleotide sequence. It is interesting to note that in native genes, A+T-rich genes tend to show a higher slope for the regression plot, suggesting that these genes might be LTGs acquired before the divergence of the considered genomes. Phages display a correlation slope close to that of native genes (0.3; $r^2 = 0.392$), indicating that they are undergoing stronger selective pressure than LTGs. The correlation for plasmid genes of *E. coli* available in GenBank [17] shows a slope similar to that of phages (0.288; $r^2 = 0.301$). Surprisingly, the IS showed no correlation ($r^2 = 0.001$), indicating that G+C3 is independent of G+C1 and G+C2 in the IS.

We carried out the same analysis on other species and found similar results (data not shown). In *Salmonella*, the higher slope of the correlation for LTGs (0.408; $r^2 = 0.593$ compared with 0.269; $r^2 = 0.288$ for native genes) as well as the absence of correlation for IS was also found, indicating that these results are neither artifacts nor limited to *E. coli*. The same tendencies are observed in *Helicobacter* and *Streptococcus*, although not always significant as a result of the low number of LTGs detected.

Table 1

Relative frequencies of the codons coding isoleucine (I) and arginine (R) in the different classes of genes

		Helicobacter			*Salmonella*				*Escherichia*					*Streptococcus*		
Amino acid	Codon	Natives	LTG	IS	Natives	LTG	IS	Phages	Natives	Recent LTG	Ancient LTG	IS	Phages	Natives	LTG	IS
I	**ATA**	**0.12**	**0.26**	**0.27**	**0.08**	**0.23**	**0.29**	**0.12**	**0.06**	**0.32**	**0.25**	**0.23**	**0.14**	**0.08**	**0.25**	**0.12**
	ATT	0.50	0.50	0.36	0.49	0.48	0.33	0.47	0.51	0.37	0.46	0.34	0.46	0.54	0.57	0.48
	ATC	0.38	0.24	0.37	0.43	0.29	0.38	0.41	0.43	0.31	0.29	0.43	0.40	0.38	0.18	0.39
R	**AGA**	**0.26**	**0.45**	**0.58**	**0.03**	**0.14**	**0.16**	**0.07**	**0.03**	**0.18**	**0.15**	**0.08**	**0.09**	**0.14**	**0.36**	**0.25**
	AGG	**0.25**	**0.18**	**0.21**	**0.02**	**0.10**	**0.16**	**0.05**	**0.02**	**0.16**	**0.08**	**0.07**	**0.07**	**0.04**	**0.11**	**0.07**
	CGA	0.07	0.07	0.04	0.06	0.11	0.19	0.08	0.06	0.10	0.11	0.12	0.08	0.11	0.11	0.20
	CGT	0.14	0.14	0.06	0.35	0.26	0.24	0.30	0.39	0.17	0.26	0.30	0.30	0.50	0.28	0.27
	CGC	0.25	0.14	0.06	0.43	0.23	0.13	0.36	0.41	0.23	0.25	0.26	0.26	0.17	0.10	0.16
	CGG	0.03	0.02	0.04	0.11	0.15	0.11	0.13	0.09	0.17	0.15	0.16	0.20	0.04	0.04	0.07

Underlined numbers refer to the frequency in laterally transferred genes (LTG). Bold numbers refer to codons that are overexpressed in LTG.

Sueoka [18,19] has computed the RNP for a representative sample of bacteria, all species together, and showed that G+C12 and G+C3 are correlated among bacterial genomes with a slope of 0.25. Hence, the slope for genes recently acquired from indiscriminate bacterial species is expected to be 0.25. We have confirmed this prediction by randomly selecting bacterial genes in GenBank [17], release 130. From the RNP, the slope of the correlation is always close to 0.3 (data not shown), even when filtering for A+T-rich sequences. It thus appears that the correlation observed in the LTGs detected by our method is incompatible with the hypothesis that these genes display the compositional features of typical components from other bacterial genomes. In particular, the amino-acid composition of LTG appears to be anomalously determined by base composition, even in comparison to genes of organisms having extreme A+T bias.

Discussion

The A+T richness of the transferred genes

The tendency of LTGs to be A+T-rich has already been noted by several authors in species having intermediate G+C contents [4,5,7]. However, these results were based on compositional analysis and have been interpreted as a limitation of the methods. Our results clearly show that LTG tend to be more A+T-rich than their new host genomes and that the compositional methods do not overlook many of them. The same phenomenon is observed for species having medium (enterobacteria) and low (*H. pylori* and *S. pneumoniae*) G+C content. This striking pattern raises questions about the nature and the source of these LTGs. For example, Lawrence and Ochman [4] hypothesized that the recently transferred genes were adapted to the genomic context of other distant species; however, our results would suggest either that the donor genomes are always more A+T rich than the acceptor genomes or that there is a bias toward the internalization of A+T-rich exogenous DNA in the genome.

Foreign DNA may indeed encounter a physical barrier when entering the cell if, for example, restriction enzymes tend to have G+C-rich target sites. When analyzing the base content of restriction enzyme target sites in REBASE [20] we have found that, after removing redundancy, they indeed present a G+C content higher than 70% on average (data not shown). Since G+C-poor genomes have LTGs with lower G+C content than G+C-rich genomes, this predicts a positive correlation between the G+C content of a genome and of the target sites of its restriction enzymes. Only a few species have sufficient fully characterized restriction enzymes to test this hypothesis. However, *E. coli* and *H. pylori* each contain about 200 fully annotated restriction enzymes, and the average G+C content of their target sites is 73% and 60% respectively. It is therefore possible that restriction enzymes have a role in determining the G+C content of LTGs. But this hypothesis is not sufficient to explain the observed pattern of base composition across species.

The mechanism of gene transfer often implicates the intervention of IS and phages, which are known to be biased towards A+T [13]. It is therefore possible that the use of such vectors influences the base composition of the LTGs. However, both the FCA and the G+C content analyses suggest that IS and phages are less biased in their base composition than other LTGs.

The source of LTGs in bacteria

LTGs possess a composition that seems principally determined by mutation, as shown by the RNP. This bias is not

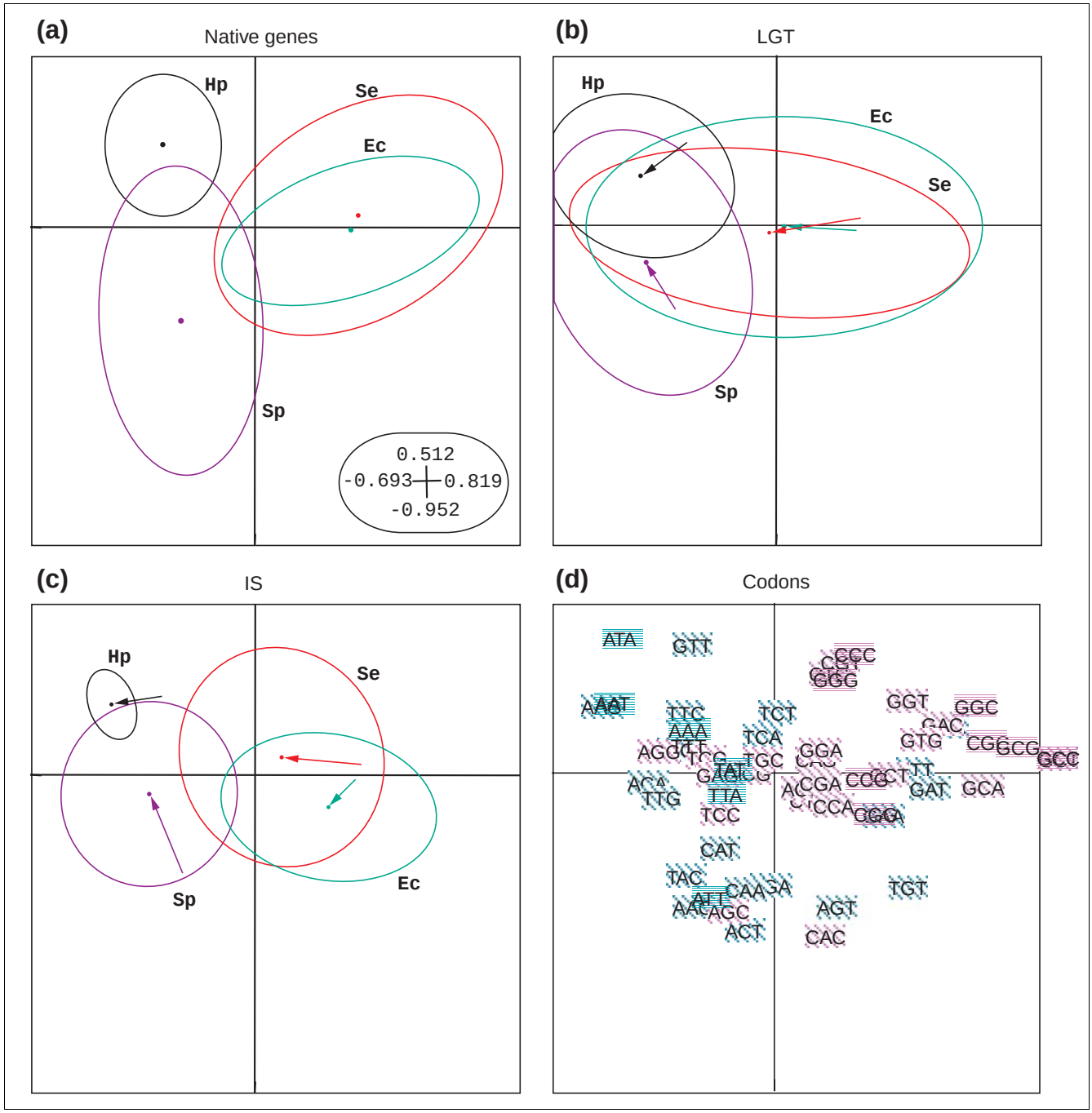

Figure 5
Interspecies FCA for the four groups of species considered. **(a)** Native genes; **(b)** LTG; **(c)** IS; and **(d)** codons are presented separately in superimposable figures. The first two axes, which represent 22.98% and 7.29%,, respectively, of the total variability, are shown. Ellipses represent 90% of the points of each cloud. The arrows in (b) and (c) represent the displacement of the center of the ellipse relative to that of the native genes. Phages were not included in the present analysis because no sequences were found in the *H. pylori* and the *S. pneumoniae* genomes. In (d), green squares represent A+T-rich codons and purple squares G+C-rich codons.

found in other classes of gene such as native, IS, or phage genes. This might suggest that LTGs are not true open reading frames (ORFs). However, even if most of these genes have no known functions or homologs, we find that their codon usage is close to genes implicated in virulence, antibiotic resistance and secretory systems, implying that they may be functional. Moreover, Alimi *et al.* [21] have shown that at least some of the orphan genes in *E. coli* are indeed transcribed.

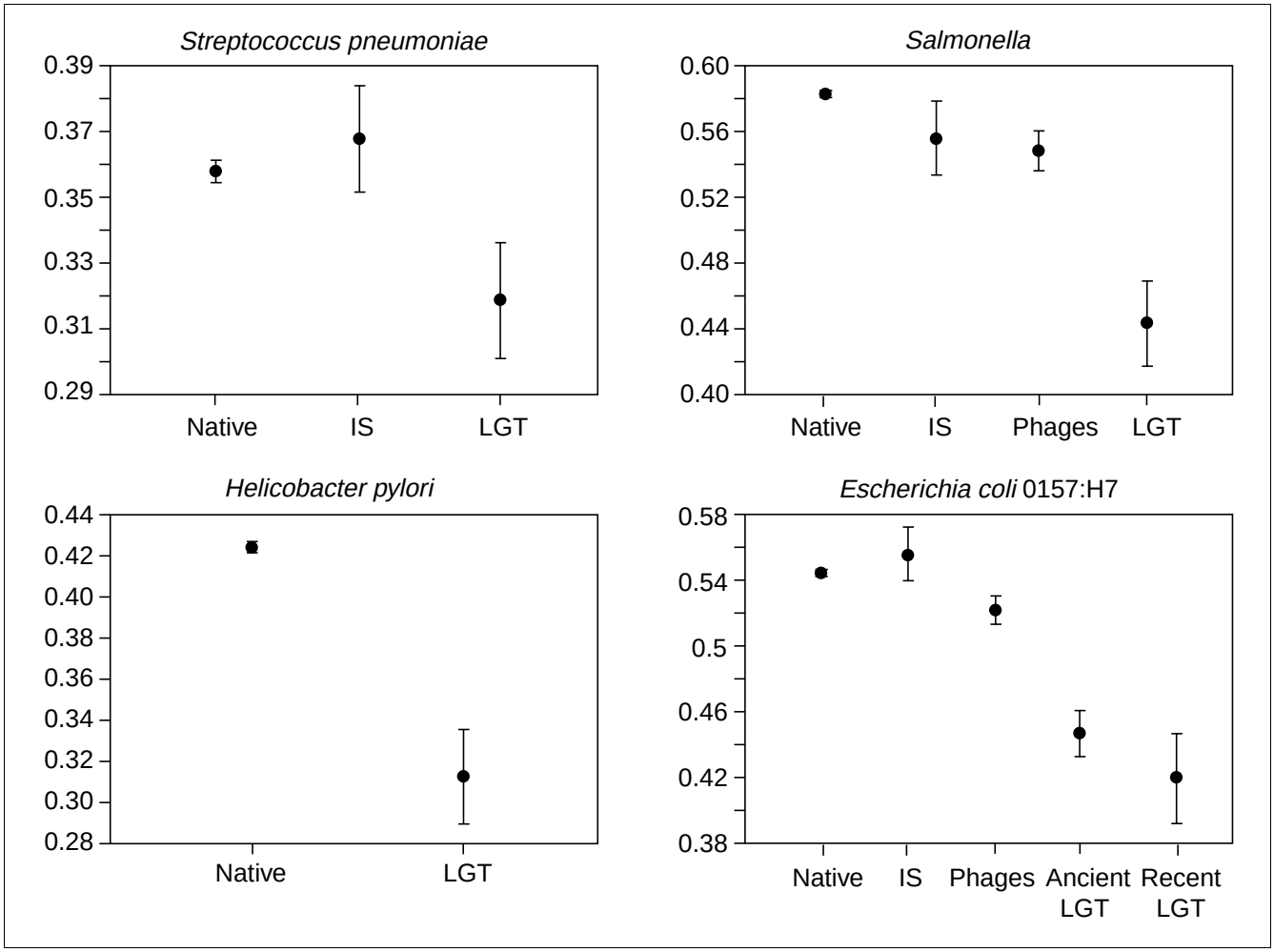

Figure 6
G+C content at the third position of codons in the different classes of gene. IS and phages are absent from certain species because their numbers were insufficient. Bars represent 95% of confidence interval.

The comparisons with randomly selected genes in GenBank using RNP shows that LTGs do not have the expected characteristics of genes adapted to previous genome contexts, even if only A+T-rich sequences are able to enter the cell. Moreover, it is very unlikely that these characteristics emerged since their insertion in their new genome. Indeed, while showing differences in gene content, the two *E. coli* O157:H7 strains, for instance, are virtually identical at the nucleotide level for the remainder of their genomes. The LTGs could not have undergone sufficient mutational pressure in such a short period of time. They more likely represent genes that are either adapted to or carry the marks of frequent lateral transfers. Their A+T-richness tends to classify them with phages and other mobile elements [13]. However, the RNP suggests that LTGs undergo low selection pressure on the protein sequence compared to these elements. Interestingly, phages have been shown to carry ORFs named 'morons' (because they add more DNA to the phage genome), which often have unknown functions, but are thought to occasionally confer benefit to the host when the prophage is integrated in its genome [22]. These genes undergo high mutation and nonhomologous recombination rates, and often display high A+T-content, even in comparison to the phage itself [22,23]. Most LTGs fit this description and may therefore have been

Figure 7
Relative neutrality plots for the different classes of gene in *E. coli* O157:H7. GC12 is plotted as a function of GC3 and the slope of the correlation (bold line) is computed.

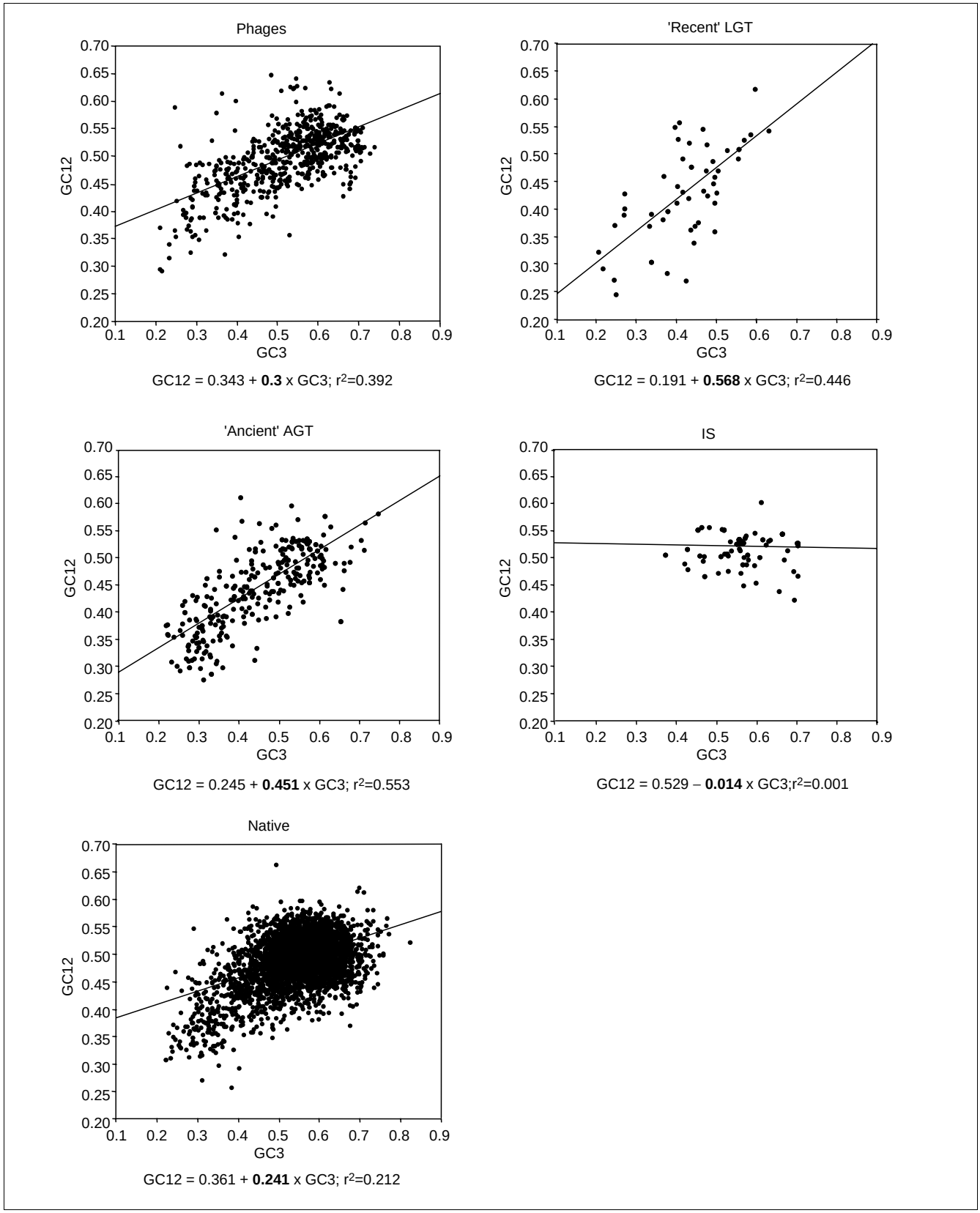

Figure 7 *(see legend on previous page)*

introduced into the genome by phages. Moreover, this may explain why most of these genes are orphans, as frequent nonhomologous recombination may preclude the recognition of homologs. The current knowledge of bacteriophage diversity is still extremely limited [24], and this lack may also explain our failure to find homologs of these genes. Indeed, the vast majority of bacteriophages being still unknown, they might represent an enormous reservoir of such genes.

Some of the morons have been shown to be related to other bacterial genes, suggesting that they may at first have been host genes integrated in the phage genome [22]. These genes may then have diverged rapidly because phages are known to have evolutionary rates orders of magnitude higher than those of bacteria [25]. Morons seem rarely to confer a direct advantage on the phage, but rather stabilize the host-prophage interaction by slightly increasing host fitness [20]. Therefore, they may undergo weaker selection pressure than genes directly involved in the phage life-cycle and be more sensitive to the mutational bias inherent to parasitic sequences [13]. From our results, it appears that whatever the nature of the organism in which the gene was first recruited, its compositional characteristics no longer represent its previous genome context. Although these genes seem to have a high turnover in the genome, it is likely that, when proved useful to the cell, they establish a long-lasting association with their new host. Thus, while 'moron accretion' has been proposed as a key mechanism of phage evolution [22], this process may also contribute to some extent to the evolution of bacterial genomes and to their adaptation to new habitats. In this view, phages could be considered as a powerful way of inventing new genes potentially beneficial to their hosts.

Thus, although bacterial genomes tend to acquire large amounts of DNA, we have shown that those transferred genes have very peculiar features that do not denote a previous genomic context but connect them with parasitic sequences such as phages. The genes involved in such lateral transfer obviously do not belong to classes of genes that encode typical cellular pathways. Hence, though the differences in content between closely related genomes have been extensively cited as evidence for constant exchanges with distant relatives, these sequences carry no evidence for such exchanges. Therefore, attempts to use codon usage of an LTG as an indication of their phylogenetic origin should be considered with caution.

Materials and methods

All genome sequences and annotations were extracted from the EMGLib database [26] using the Query retrieval system [27].

Inferring recent acquisitions and losses by parsimony

The availability of sequenced bacterial genomes allows comparison of the gene content between closely related species, and thus the finding of very recently acquired genes in the genomes using parsimony analysis. Figure 1 describes the ideal case of three closely related species, A, B, and C, for which the genomes are sequenced and the phylogenetic relationships known. Three scenarios can explain the presence in species A of a gene which is absent in species B and C: first, the gene was present in the common ancestor of the species A, B, and C, and has then been independently lost in species B and C; second, the gene has been acquired by the common ancestor of species A and B, and then lost in species B; and third, the gene has been recently acquired by species A. The last hypothesis is the most parsimonious explanation if one considers that the acquisition of a gene is at least as probable as a loss. A possible verification that this hypothesis is realistic is to estimate the number of apparent gene losses. Indeed, the absence in species A of a gene present in species B and C can be interpreted as the loss of the gene in species A or two independent acquisitions in species B and C. Note that apparent gene losses in A may be overestimated if recombination occur frequently between B and C (that is, acquisition of genes by B and C is not independent), however the effect of such a recombination event is probably low in the present cases (see 'Genomes' section). These estimations of acquisition and losses can be made for species A and B.

We identified genes acquired by species A after the divergence of species A and B (case +A -B -C in Figure 1), by making a BLASTP [28] query of the protein sequences more than 50 amino acids long in genome A against those in B and C. Proteins having no match with a bit score >10% of the bit score of the query protein against itself were considered as being recently acquired in species A.

We identified gene losses by species A after the divergence of species A and B (case -A +B +C in Figure 1), by making a BLASTP query of the protein sequences more than 50 amino acids long in genome B against those in A and C (Figure 1). A protein was considered as recently lost in species A if it had no match in species A (same criterion as before) and at least one match in species C (bit score higher than 50% of the bit score of the protein against itself). To avoid problems due to possible gene misannotations, these results were verified using a BLASTN query.

Genomes

To use the method described above, it is important to have at least three complete sequenced genomes that are closely related with unambiguous phylogenetic relationships. For this purpose, we used five closely related genomes in the enterobacteria group: *E. coli* O157:H7 EDL933 [11], *E. coli* O157:H7 Sakaï [29], *E. coli* K12 [30], *Salmonella enterica* [31], and *S. typhimurium* LT2 [10]; three closely related genomes in the alpha-proteobacteria group: *Helicobacter pylori* J99 [32], *H. pylori* 26695 [33], and *Campylobacter jejunii* [34]; and three closely related genomes in the *Streptococcus* genus: *S. pneumoniae* R6 [35], *S. pneumoniae* TIGR4

[36], and *S. pyogenes* [12]. We considered as unambiguous the relationships between these bacteria because, for example, the orthologous genes of the two strains of *E. coli* O157:H7 are almost identical at the nucleotide level, while they show noticeable differences from *E. coli* K12 (data not shown). This suggests that, since their divergence, the two strains of *E. coli* O157:H7 have undergone only a few recombination events with more distant strains. The same reasoning has been applied in the other cases. In the group of enterobacteria, it was possible to classify transferred genes relative to their date of acquisition in the three strains of *E. coli*. Thus, we identified transferred genes acquired before the separation of the two strains O157:H7 ('ancient transfers') and those acquired in one of the two strains O157:H7 after their separation ('recent transfers').

Gene classes

On the basis of sequence annotations, we have removed genes related to IS and prophages from the different classes of genes defined using the method described below. Genes from each class, that is, native genes, potentially transferred genes (LTG), IS and phages, of the four groups of bacterial genomes (*Escherichia*, *Salmonella*, *Helicobacter*, and *Streptococcus*) were then retained for codon-usage analysis when their lengths were greater than 150 base-pairs (bp) to avoid artifacts linked to stochastic variations that might happen in shorter genes.

Factorial correspondence analysis on codon usage

To compute our FCA on gene codon composition, we used absolute codon frequencies, without considering the three stop codons or the ATG and TGG codons, which are not degenerate. We thus obtained a matrix consisting of 59 columns (corresponding to the 59 degenerate codons) and as many rows as sequences analyzed. Such a matrix can be used in a FCA, which is a multivariate analysis often used to study codon usage [2,5,14,37,38]. It allows one to calculate the position of sequences in a multidimensional space with respect to their codon usage and to give a graphical representation of the dimensions maximizing their dispersion. Genes having similar codon usage are hence regrouped. The analysis, being symmetrical, makes it possible to represent the codons in the same space as the one used to visualize the genes, which allows identification of those responsible for the clustering of the genes. We used ADE-4 software package [39] to perform the FCA presented in this study.

To avoid statistical bias due to the differences in numbers of sequences composing each category, we randomly selected 200 genes among the native genes and 200 among the transferred genes (when their number was greater than this value), for the intraspecific species analysis. When the number of phages and IS was greater than 50, we randomly selected 50 sequences in the phage and IS categories. For the same reason, in analysis gathering the four species, we randomly selected 100 genes among the native genes and 100 among the transferred genes. The numbers of transferred genes in the different strains of *S. pneumoniae* and *H. pylori* were approximately 100, so the entire sets were used. Ten independent selections of genes were performed to guarantee the reproducibility of the results.

Relative neutrality plots (RNPs)

The strength of the selection on a given gene relative to the mutation pressure can be estimated by the method of the relative neutrality plot (RNP), which gives indications on how 'neutral' a coding sequence can be considered [18,19]. The method consists of plotting the G+C content at the constrained (or nonsynonymous) positions (that is, first and second positions) of the codons against the G+C content at the relaxed (or synonymous) position (that is, third position). The slope of the resulting linear correlation gives evidence on how the protein sequence is affected by the mutational bias acting on the nucleotide sequence, and thus on how strongly the selection pressure acting on the protein can counteract this bias. Note that the effect measured is relative to the translational selection acting on the third position of codons and that the strength of this pressure is supposed to be weak compared to the selection on the protein sequence. The slope is expected to be equal to one if the protein sequences are under no selective constraints, and to decrease with the strength of the selection acting at the protein level. Translational selection is also expected to reduce the correlation, though to a lower extent. For this study, we analyzed the correlations according to different gene classes to determine whether there were differences in the relative selection pressures in each of the classes. All the correlations presented here are highly significant ($p < 0.0001$) except when stated.

Additional data files

A list of the acquired genes in three groups of closely related bacteria as estimated by the method in Figure 1 is available as an additional data file (additional data file 1) with the online version of this paper.

Acknowledgements

We would like to thank Laurent Duret, Manolo Gouy and Howard Ochman for comments about the results and manuscript.

References

1. Sueoka N: **Directional mutation pressure and neutral molecular evolution.** *Proc Natl Acad Sci* 1988, **85:**2653-2657.
2. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A: **Codon catalog usage and the genome hypothesis.** *Nucleic Acids Res* 1980, **8:**r49-r62.
3. Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends Genet* 1995, **11:**283-290.
4. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44:**383-397.
5. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A: **Evidence for horizontal gene transfer in *Escherichia coli* speciation.** *J Mol Biol* 1991, **222:**851-856.
6. Moszer I, Rocha EP, Danchin A: **Codon usage and lateral gene**

transfer in *Bacillus subtilis*. *Curr Opin Microbiol* 1999, **2:**524-528.
7. Syvanen M: **Horizontal gene transfer: evidence and possible consequences.** *Annu Rev Genet* 1994, **28:**237-261.
8. Martin W: **Mosaic bacterial chromosomes: a challenge en route to a tree of genomes.** *Bioessays* 1999, **21:**99-104.
9. Lan R, Reeves PR: **Gene transfer is a major factor in bacterial evolution.** *Mol Biol Evol* 1996, **13:**47-55.
10. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, *et al.*: **Complete genome sequence of *Salmonella enterica* serovar *Typhimurium* LT2.** *Nature* 2001, **413:**852-856.
11. Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, *et al.*: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7.** *Nature* 2001, **409:**529-533.
12. Ferretti JJ, McShan WM, Adjic D, Savic D, Savic G, Lyon K, Primeaux C, Sezate SS, Surorov AN, Kenton S, *et al.*: **Complete genome sequence of an M1 strain of *Streptococcus pyogenes*.** *Proc Natl Acad Sci USA* 2001, **98:**4658-4663.
13. Rocha E, Danchin A: **Base composition bias might result from competition for metabolic resources.** *Trends Genet* 2002, **18:**291-294.
14. Lerat E, Capy P, Biémont C: **Codon usage by transposable elements and their host genes in five species.** *J Mol Evol* 2002, **54:**625-637.
15. Daubin V, Perrière G: **G+C3 structuring along the genome: a common feature in prokaryotes.** *Mol Biol Evol* 2003, **20:**471-483.
16. Brochier C, Philippe H, Moreira D: **The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome.** *Trends Genet* 2000, **16:**529-533.
17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2002, **30:**17-20.
18. Sueoka N: **Intrastrand parity rules of DNA base composition and usage biases of synonymous codons.** *J Mol Evol* 1995, **40:**318-325.
19. Sueoka N: **Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C.** *J Mol Evol* 1999, **49:**49-62.
20. Roberts RJ, Macelis D: **REBASE--restriction enzymes and methylases.** *Nucleic Acids Res* 2001, **29:**268-269.
21. Alimi JP, Poirot O, Lopez F, Claverie JM: **Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from *Escherichia coli* K-12 MG1655.** *Genome Res* 2000, **10:**959-966.
22. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S: **The origins and ongoing evolution of viruses.** *Trends Microbiol* 2000, **8:**504-508.
23. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW: **Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages.** *J Mol Biol* 2000, **299:**27-51.
24. Hendrix RW: **Bacteriophages: evolution of the majority.** *Theor Popul Biol* 2002, **61:**471-480.
25. Drake JW: **A constant rate of spontaneous mutation in DNA-based microbes.** *Proc Natl Acad Sci USA* 1991, **88:**7160-7164.
26. Perrière G, Bessières P, Labedan B: **EMGLib: the enhanced microbial genomes library (update 2000).** *Nucleic Acids Res* 2000, **28:**68-71.
27. Gouy M, Gautier C, Attimonelli M, Lavane C, di Paola G: **ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage.** *Comp Appl Biosci* 1985, **1:**167-172.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
29. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han C-G, Ohtsubo E, Nakayama K, Murata T, *et al.*: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Res* 2001, **8:**11-22.
30. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1474.
31. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MTG, *et al.*: **Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar *Typhi* CT18.** *Nature* 2001, **413:**848-852.
32. Alm RA, Ling L-SL, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, *et al.*: **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*.** *Nature* 1999, **397:**176-180.
33. Tomb J-F, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388:**539-547.
34. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, *et al.*: **The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences.** *Nature* 2000, **403:**665-668.
35. Hoskins J, Alborn WE Jr, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, Estrem ST, Fritz L, Fu D-J, Fuller W, *et al.*: **Genome of the bacterium *Streptococcus pneumoniae* strain R6.** *J Bacteriol* 2001, **183:**5709-5717.
36. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, *et al.*: **Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*.** *Science* 2001, **293:**498-506.
37. Shields DC, Sharp PM: **Evidence that mutation patterns vary among *Drosophila* transposable elements.** *J Mol Biol* 1989, **207:**843-846.
38. Perrière G, Thioulouse J: **Use and misuse of correspondence analysis in codon usage studies.** *Nucleic Acids Res* 2002, **30:**4548-4555.
39. Thioulouse J, Chessel D, Dolédec S, Olivier JM: **ADE-4: a multivariate analysis and graphical display software.** *Stat Comput* 1997, **7:**75-83.