Duplication is more common among laterally transferred genes than among indigenous genes

Sean D Hooper and Otto G Berg

Address: Department of Molecular Evolution, Uppsala University, Norbyvagen 18C, SE-75236 Uppsala, Sweden.

Correspondence: Otto G Berg. E-mail: otto.berg@ebc.uu.se

Published: 11 July 2003

Genome ***Biology*** 2003, **4:**R48

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/8/R48

Received: 25 March 2003
Revised: 4 June 2003
Accepted: 10 June 2003

Abstract

Background: Recent developments in the understanding of paralogous evolution have prompted a focus not only on obviously advantageous genes, but also on genes that can be considered to have a weak or sporadic impact on the survival of the organism. Here we examine the duplicative behavior of a category of genes that can be considered to be mostly transient in the genome, namely laterally transferred genes. Using both a compositional method and a gene-tree approach, we identify a number of proposed laterally transferred genes and study their nucleotide composition and frequency of duplication.

Results: It is found that duplications are significantly overrepresented among potential laterally transferred genes compared to the indigenous ones. Furthermore, the GC_3 distribution of potential laterally transferred genes was found to be largely uniform in some genomes, suggesting an import from a broad range of donors.

Conclusions: The results are discussed not in a context of strongly optimized established genes, but rather of genes with weak or ancillary functions. The importance of duplication may therefore depend on the variability and availability of weak genes for which novel functions may be discovered. Therefore, lateral transfer may accelerate the evolutionary process of duplication by bringing foreign genes that have mainly weak or no function into the genome.

Background

There are few natural niches left on Earth that have yet to be colonized by microbes. The adaptability of prokaryotic life in particular is arguably a major reason for its success. In order to use new metabolites, or to survive new environments, microbes need genes that code for products that facilitate survival. There are basically two methods by which genomes expand their repertoire of genes: creating them through duplication and adaptation or taking existing genes from outside sources. The method of duplication has been examined extensively, starting as a model of gene evolution where a copy of a gene is free to evolve and diverge, until it may attain a novel function [1]. Recently, several authors have challenged the period of neutrality implied by the old model, and shifted the focus more towards how new paralogs avoid neutrality by gene-amplification effects [2]. A gene amplification will be either selected, neutral or counter-selected, depending on the original gene function. If noncoding DNA is duplicated, then both paralogs are probably transient in the genome. If the paralogs have a weak but slightly selected product, amplification may be selected. However, if gene products are strong, well-established, highly expressed or part of a delicate balance of proteins, then amplification may be counter-selected [3]. Furthermore, fragments of genes can be duplicated, possibly amplifying secondary functions in established genes. Through amplification, new functions can be 'discovered' in extant ones.

As an alternative to developing new gene functions through duplication, an organism can incorporate DNA from outside sources through the process of lateral transfer. This process of gene transfer has been observed not only for eubacteria (for example [4-6]), but also for archaea [7,8] and eukaryotes. Over the years, systematic studies of lateral transfer (for example [9-12]) have sparked an animated debate on the extent of lateral transfer. Some authors propose that it is a major force in genome evolution [13-15], whereas others downplay its role. One factor limiting the impact of lateral transfer may be the high level of complexity in a large number of pathways; a single transferred gene will have difficulty outcompeting an indigenous gene that is part of an adapted system of complex protein interactions [16,17]. Certainly, highly optimized genes are difficult to replace in the genome through lateral transfer, even if they are not a part of a complex pathway. However, duplication of such genes could also be disruptive and thereby reduce the impact of duplication on gene innovation. Of course, clusters of genes as well as single genes can be imported. Some gene families, such as the phosphonate metabolism *phn* group, are widely considered to have been imported into *Escherichia coli* ([9] and references therein), and are organized as clusters of genes. This may, perhaps, be the only way, however improbable, to circumvent conservation caused by complexity. For reasons of complexity and optimization, it may be more interesting from an evolutionary point of view to shift the focus from well-established genes with highly optimized functions to the weaker ones, as it is here that evolutionary mechanisms such as duplication and lateral transfer may have more significant effects.

We previously studied lateral transfer of genes between species in the *Salmonella*/*Escherichia* clade [18], and also the turnover and fixation of duplications within various species [3]. It was found that the rates of both lateral transfer and duplication in for instance *E. coli* are considerable, but that only a few imports and duplications survive deletion. Also, very few recently imported genes seem to have a defined product. However, the results also suggest that the contribution to gene innovation is mainly the result of gene-dosage effects of weak or ancillary gene functions. Even though paralogs can be selected for amplification of strong, established functions, we found that these events rarely result in new gene functions.

Here we examine whether an influx of weak or nonfunctional genes into a genome can actually contribute to gene innovation through duplication. Furthermore, we examine the general patterns of lateral transfer among a selection of organisms of different levels of divergence, including distributions of GC content.

Results

Summary of gene categories

Genes are grouped by their presence or absence in a selection of organisms (Table 1). In gene group A (Table 2), we expect to find the most recent additions to the genome through contributions from both lateral transfer and duplication. Because these genes are present only in the subject organism, it is probable that lateral transfer occurred after the last divergence. In the case of relatively recent divergence, such as for instance *Salmonella typhimurium* and *S. typhi*, the total number of genes (*N*) in category A is relatively small, whereas organisms which are farther apart (for example, *Bacillus subtilis* and *B. halodurans*) have larger numbers of unique genes. However, there is probably no constant rate at which genomes acquire genes. This is illustrated by the comparison between the two *E. coli* strains, where the A group is almost twice as large in O157:H7 than in K12. This suggests that strain O157:H7 is more active than K12 in gene acquisition either mechanistically or because of added selective pressure. Other pairs of organisms, such as *E. coli* K12 and *S. typhimurium*, have more similar patterns of gene acquisition. Imports in category B (Table 3) are on average older than in category A, and many genes are likely to have been lost in the outgroup genomes as a result of different lifestyles and requirements. Genes in category D are present in all four organisms in the group, and are probably not recently transferred genes.

In previous work [18,19] we have developed a compositional measure (cT^2; see Materials and methods) to find genes with atypical dinucleotide frequencies. This measure is used in combination with the gene categories. Through all categories and in all comparisons, the number of atypical genes is the highest in groups that are associated with recent lateral transfer. The proportion of deviant genes (*Pdev*; $P(cT^2 > 38)$) are consistently higher in categories A and B than in D (Table 4),

Table 1

Overview of gene categories

Category	G	Cr	O1	O2	Suggested content of imports
A	+	-	-	-	Recent imports or loss of imports from B
B	+	+	-	-	Less recent imports
C	+	-	+	+	Loss of genes in Cr, or pervasive imports
D	+	+	+	+	Only highly pervasive imports expected

G, genome under consideration; Cr, paired close relative; O1, first outgroup organism; O2, second outgroup organism. +, Present; -, absent.

Table 2

Category A

Pair	Outgroups	*N*	*Ndev*	*Pdev*	*Ndup*
ecol, edl	stym, klebs	95	17	0.179	7
ecol, stym	kpne, paer	248	43	0.173	25
edl, ecol	stym, klebs	180	43	0.239	51
sfle, ecol	stym, klebs	139	30	0.216	37
stym, ecol	kpne, paer	270	47	0.174	8
stym, styp	ecol, klebs	89	20	0.225	5
styp, stym	ecol, klebs	138	31	0.225	5
bsub, bhal	cace, cper	575	51	0.089	16
bhal, bsub	cace, cper	603	40	0.066	16
cace, cper	bhal, bsub	779	49	0.063	20
samu, samw	bsub, cace	70	9	0.129	6

N, total number of genes; *Ndev*, number of genes with cT^2 scores > 38; *Pdev*, proportion of deviant genes; *Ndup*, number of duplicated genes. See Materials and methods for species abbreviations. All *Pdev* values are significant at 99% when compared to group D (Table 4).

Table 3

Category B

Pair	Outgroups	*N*	*Ndev*	*Pdev*	*Ndup*
ecol, edl	stym, klebs	190	27	0.142	18
ecol, stym	kpne, paer	163	22	0.135	11
edl, ecol	stym, klebs	189	15	0.079	32
sfle, ecol	stym, klebs	169	27	0.160	34
stym, ecol	kpne, paer	171	18	0.105	3
stym, styp	ecol, klebs	249	34	0.137	10
styp, stym	ecol, klebs	233	31	0.133	5
bsub, bhal	cace, cper	582	26	0.085	18
bhal, bsub	cace, cper	602	28	0.047*	6
cace, cper	bhal, bsub	293	19	0.065	8
samu, samw	bsub, cace	489	31	0.063	12

*All *Pdev* values are significant at 99% when compared to group D (Table 3), except for bhal which is not significant. See Materials and methods for species abbreviations.

which would support the notion that there are many candidates for lateral transfer in groups A and B. In *B. subtilis* and *B. halodurans*, which appear to have diverged long ago, *Pdev* in categories A and B is low, although still higher than in D. The numbers of deviant genes in category A are 51 and 40 in *B. subtilis* and *B. halodurans* respectively, similar in number to those of, say, *E. coli* or *S. typhi*. Thus the rates of import may be similar in magnitude and the low *Pdev* scores could be due to category A containing a large number of older genes that have either been ameliorated in *Bacillus* or lost in *Clostridium*. The older the A group, the lower *Pdev* will be, until it approaches the *Pdev* of category D. The comparison between the *Staphylococcus aureus* strains mu50 and mw2 shows an intermediate pattern, with a relatively high *Pdev* in categories A and B compared to category D. This would suggest that there is no great difference between the proteobacteria and the *Bacillus/Clostridium* group in lateral transfer patterns.

GC_3 distributions in gene-categories A and D

The distributions of guanine and cytosine at the third codon position (GC_3) in categories A and D are summarized in Figure 1 and Table 5. Category A is generally 'flatter' than the GC_3 distribution of category D. In some genomes, particularly in the *Escherichia/Salmonella* clade, GC_3 is almost uniformly distributed. Given that genes in category A are good candidates for lateral transfer, a flat distribution would suggest that the GC_3 range of imports is wide. Accordingly, the number of donor genomes must be large enough to incorporate a broad spectrum of gene GC_3 content. In the *Bacillus/ Clostridium* clade, the GC_3 pattern of the A groups resembles that of the D groups to varying degrees. Particularly *B. halodurans* appears to have a narrower GC_3 range of potential imports, which could be explained by one of three reasons. First, *B. subtilis* and *B. halodurans* have been diverging for so long that the A group could be dominated by old imports that have had time to ameliorate to the genomic GC_3 distribution. Second, it is possible that the number of donor organisms is lower in the halophilic environment of *B. halodurans*. If so, there could be less variation in the GC_3 range of imports. Third, genes in category A could be primarily genes that have been lost in *B. subtilis* but present in the ancestor, although this explanation also means that the same genes have been lost in *Clostridium*.

To focus on the genes that were likely to be more recent imports in *Clostridium acetobutylicum*, *B. subtilis* and *B. halodurans*, which have long divergence times, we studied the GC_3 distribution of the subset of A genes with high cT^2 values. Genes with cT^2 scores grater than 30 were compared to the rest of the set (Figure 2). The numbers of genes with scores over 30 were 91, 97 and 78 for *C. acetobutylicum*, *B. subtilis* and *B. halodurans* respectively. The GC_3 distribution of the high-cT^2 genes in *C. acetobutylicum* differs neither from the rest of the genes in the A group, nor from those in the D group (Figure 1h). This could either be due to a lower rate of import (all imports have ameliorated), or a consequence of an already low GC content. *B. subtilis* and *B. halodurans*, on the other hand, have a wider GC_3 range in the subsets of high-cT^2 genes relative to the rest of the A genes. *B. subtilis* has a significant overrepresentation of high-cT^2 genes in the GC-poor terminus region, which could be an effect of either a higher recombination rate in this area [9] or local nucleotide variations, or both.

Table 4

Category D

Pair	Outgroups	*N*	*Ndev*	*Pdev*	*Ndup*	*Ndev* and *dup**
ecol, edl	stym, klebs	2,823	91	0.032	46	3
ecol, stym	kpne, paer	2,163	49	0.023	34	2
edl, ecol	stym, klebs	2,865	87	0.030	71	8
sfle, ecol	stym, klebs	2,684	113	0.042	214	20
stym, ecol	kpne, paer	2,129	74	0.035	35	14
stym, styp	ecol, klebs	2,752	95	0.035	35	13
styp, stym	ecol, klebs	2,662	96	0.036	29	15
bsub, bhal	cace, cper	1,329	32	0.024	10	0
bhal, bsub	cace, cper	1,354	51	0.038	15	2
cace, cper	bhal, bsub	1,290	47	0.036	11	0
samu, samw	bsub, cace	1,055	42	0.040	9	0

* The number of duplicated genes with $cT^2 > 38$. See Materials and methods for species abbreviations.

In *B. subtilis* group D, there are a number of genes with high cT^2 and low GC_3 around the *ori* region. These are most likely not imports, as they have defined (and often ribosomal) functions and a high codon adaptation index (CAI) [20], indicating that there is translational selection on these genes. Genes in group A around the terminus region also have high CAI, but this is probably due to the fact that low GC_3 coincides with the choice of major codons in *B. subtilis* [21].

Comparing GC_3 skews of category D in Figure 1, we note that all *Escherichia/Salmonella* clade genomes are skewed towards low GC_3. This relationship could be due to one or both of the following explanations: there may be local, although small, regions that are biased towards lower GC_3; or it may suggest that there are a number of pervasive low-GC_3 imports into the *Escherichia/Salmonella* clade.

Both *E. coli* and *B. subtilis* have distinct regions around the terminus with low GC content [19,22]. This regional bias around the terminus is not a general feature. The genome of *B. halodurans* also has a low-GC region around the terminus, although not as distinct as that of *B. subtilis*. The regional bias of *S. aureus* mu50 is intermediate between *B. halodurans* and *B. subtilis*, whereas *Clostridium* has no low-GC region around the terminus.

Recent duplications

The fate of a recent duplication is selection or redundancy followed by deletion, with the vast majority of recent duplications being deleted [3]. In several organisms it was found that the persistence of neutral material is extremely low. Few neutral paralogs survive long enough to undergo any significant mutational degradation. In all cases considered we find that the total number of duplications in the A groups are significantly overrepresented compared to category D (Table 6).

As a result of both stoichiometric and expression-level constraints on gene duplication, we expect the number of recent duplications among genes that are either highly expressed or involved in complex protein interactions to be low. While the interaction level can be hard to assess, expression can be estimated from the codon bias [20]. In *E. coli* K12, we see that the paralogs in category A are all hypothetical proteins with low codon bias, CAI < 0.3. In category D, 33 of 47 recent duplications are hypothetical or putative genes, with a low average CAI score. Two exceptions are the *tuf* copies. In *E. coli* O157:H7, hypothetical genes dominate the duplications in category A and none of the paralogs has a known function. The duplications in category D of O157:H7 are similar to those in category D of K12, with mostly hypothetical genes. *S. typhimurium* has only five recent duplications in category A, and all are putative. In category D, we find 15 copies of a number of ABC transporter-related genes, along with putative genes and two *tuf* copies. *C. acetobutylicum* also follows this pattern, with mostly hypothetical genes, along with ABC transporters. In *B. subtilis*, all 16 duplications in category A are hypothetical. Ten of sixteen have CAI scores lower than average. Genes are better defined in category D, with only 6 of 10 duplicated genes having CAI scores lower than average. *S. aureus* mu50 has few recent paralogs in category D, and only one is annotated as hypothetical. In category A, four out of six are hypothetical. If duplication is mainly dependent on function, then it may in part explain the overrepresentation of duplications in category A for almost all organisms. This category should contain the highest proportion of weakly functional or nonfunctional genes, which are less disruptive to amplify.

In general, duplicated genes seem more likely to have high cT^2 scores. This correlation is dominated by the D category (Table 4) where there is a higher proportion of deviant genes

comment
reviews
reports
deposited research
refereed research
interactions
information

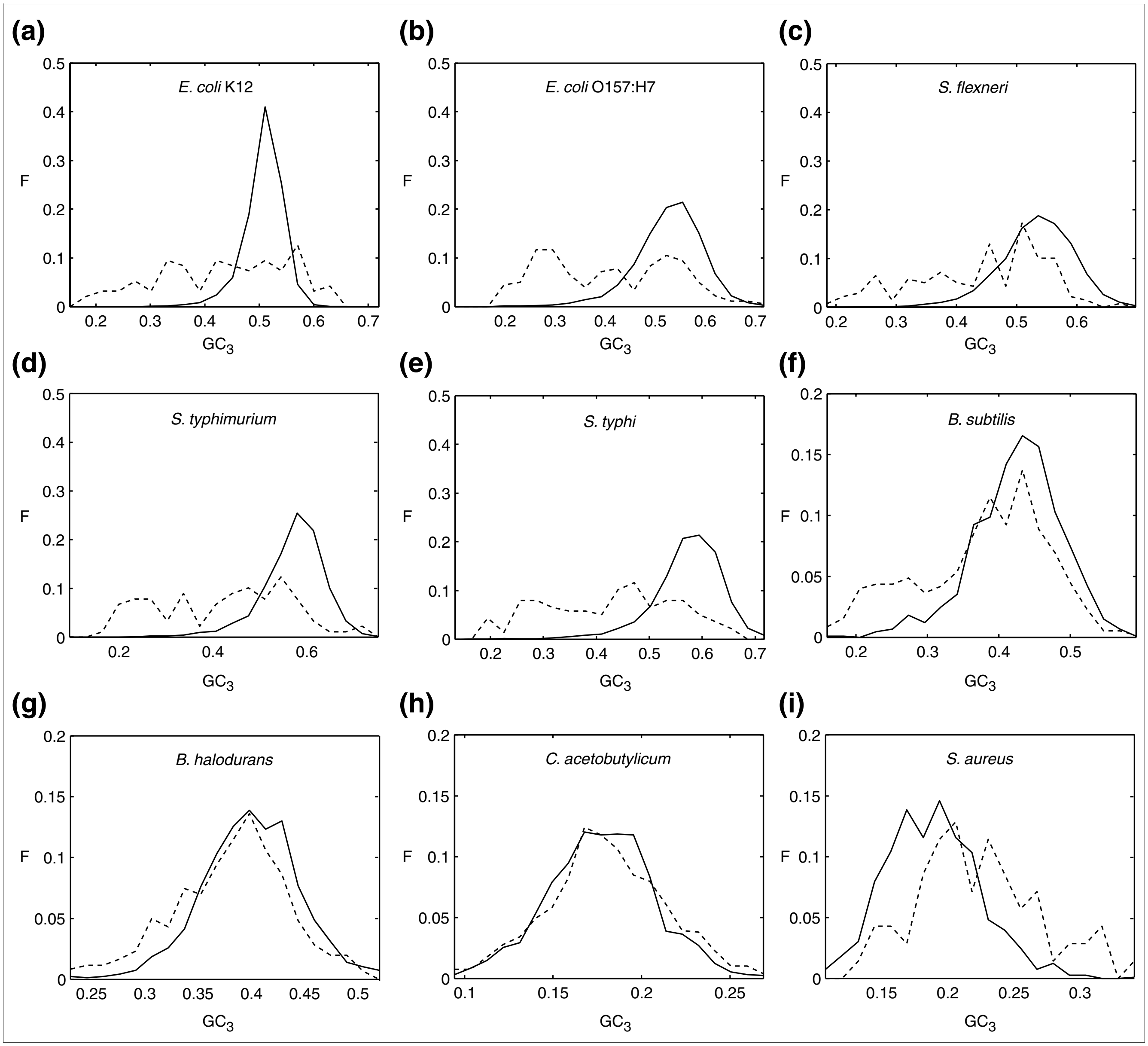

Figure 1
Distribution of GC_3 content of genes in categories A (dashed line) and D (solid line). **(a)** *E. coli* K12; **(b)** *E. coli* O157:H7; **(c)** *S. flexneri*; **(d)** *S. typhimurium*; **(e)** *S. typhi*; **(f)** *B. subtilis*; **(g)** *B. halodurans*; **(h)** *C. acetobutylicum*; **(i)** *S. aureus*.

among the duplications in all but three cases. This would suggest a general link between atypical sequences and duplication, which may be due to unusual DNA sequences such as transposon-like structures. Even though annotated transposons have been removed from the datasets, some such sequences may still persist. Furthermore, the overrepresentation could also be due to higher rates of recombination in certain chromosomal regions. Finally, it is possible that the paralogs were imported as a cluster - the duplication therefore taking place before the transfer. This is not supported by the chromosomal positions of paralogs in category A of the paired organisms in Table 6, suggesting that duplication is likely to have taken place after transfer. Another possibility could be a multiple and probably simultaneous import of the same gene from a single source. It would be difficult to tell this event apart from an indigenous duplication event.

Discussion

Patterns of lateral transfer

The context bias (cT^2) has been shown to be a useful measure to identify recently imported genes [18,19]. This is further corroborated in this study where the gene categories that are expected to contain more recently imported genes always

Table 5

GC_3 % distribution data

Pair	Outgroups	m(A)	SD(A)	m(D)	SD(D)
ecol, edl	stym, klebs	0.451	0.119	0.523	0.035
edl, ecol	stym, klebs	0.425	0.130	0.544	0.068
sfle, ecol	stym, klebs	0.455	0.108	0.546	0.062
stym, styp,	ecol, klebs	0.441	0.138	0.589	0.069
styp, stym	ecol, klebs	0.441	0.123	0.589	0.069
bsub, bhal	cace, cper	0.387	0.091	0.437	0.061
bhal, bsub	cace, cper	0.383	0.061	0.408	0.049
cace, cper	bhal, bsub	0.186	0.037	0.182	0.032
samu, samw	bsub, cace	0.231	0.049	0.195	0.035

m, Mean; SD, standard deviation. A, category A; D, category D. See Materials and methods for species abbreviations.

have a larger fraction of deviant genes (high cT^2). The connection is strengthened further by the observation that the deviant genes have a broader and more uniform GC_3 distribution in keeping with the expectation that recent imports have not had time to ameliorate to the genomic bias. Thus, the number of recent imports in each genome can be assumed to be related to the number of deviant genes in category A. The absolute numbers of deviant genes in the A and B groups are very similar for all genomes considered (Tables 2,3). A possible exception is *S. aureus*, which seems to have a lower number of deviant genes in category A. This suggests that the intensity of lateral transfer is roughly of the same magnitude in all cases.

GC_3 distribution of proposed imports

The general flatness of the distributions of GC_3 content in category A may support the notion that recently laterally transferred genes dominate in this group. This is the kind of distribution we would expect under the assumption that the multitude of potential donors have a wide range of GC_3 content. The distribution of GC_3 content in category D is almost always more peaked and has more weight in its tails, reflecting a non-Gaussian probability distribution. Among the organisms with approximately 50% GC_3 content, there is a slight skew to low GC_3. Furthermore, in the two organisms with low genomic GC_3 content, both categories A and D are skewed slightly to higher GC_3.

Whereas recent imports generally seem to have broad, nearly uniform, GC_3 distributions, there seem to be few genes with high GC_3 content. This is particularly evident in the genomes with low overall GC_3 content. Thus, the recent imports in the *Escherichia/Salmonella* clade with intermediate overall GC_3 content have broad tails to low GC_3, but the low-GC_3 genomes have no corresponding tails to high GC_3 (Figure 1). Either the potential donors for these genomes are all of low GC_3 content or low-GC_3 genes are more easily exchanged. *E. coli* has regions in the genome with low GC_3 content [19,22] as does *B. subtilis*. Possibly such regions are more likely to accept imports of similar GC content.

In the *Escherichia/Salmonella* clade, the chromosomal position of category A genes was studied. For *E. coli* K12 and *S. typhimurium*, there is a tendency to have a high number of A genes at gene position 2,000 to 3,000. *E. coli* O157:H7 has a peak at 1,000 to 2,000. In general, there is no overrepresentation of potential laterally transferred genes in the immediate vicinity of the low-GC terminus region in contrast to previous proposals [10,18].

Duplication bias

Genes that are likely to have been imported into the organism from an outside source are also more likely to be duplicated. This is the overall impression from studying bacteria both

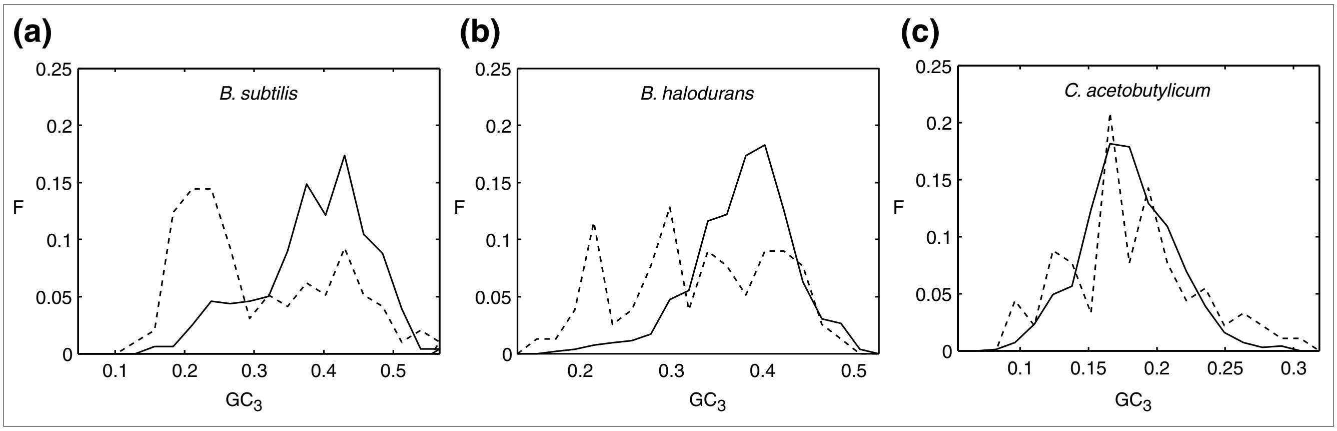

Figure 2
Distribution of GC_3 content of genes of high (dashed line) and low (solid line) cT^2 in category A. **(a)** *B. subtilis*; **(b)** *B. halodurans*; **(c)** *C. acetobutylicum*.

Table 6

Frequency of duplications by category

Pair	Outgroups	*Ndup*(A)/ *N*(A)	*Ndup*(D)/ *N*(D)
ecol, edl	stym, klebs	0.074**	0.016
ecol, stym	kpne, paer	0.101**	0.018
edl, ecol	stym, klebs	0.283**	0.025
sfle, ecol	stym, klebs	0.266**	0.080
stym, ecol	kpne, paer	0.030*	0.015
stym, styp	ecol, klebs	0.056**	0.014
styp, stym	ecol, klebs	0.036*	0.011
bsub, bhal	cace, cper	0.029**	0.008
bhal, bsub	cace, cper	0.027**	0.011
cace, cper	bhal, bsub	0.026**	0.009
samu, samw	bsub, cace	0.086**	0.009

*Significant at 95%; **significant at 99%.

from the proteobacteria group and from the *Bacillus/ Clostridium* clade, although the relation between cT^2 scores and propensity to duplicate is weaker in the latter group. This observed bias may be due to the following:

(1) Genes that have been imported into the genome may by their nature be more mobile or 'promiscuous'.

(2) Imported genes that are required in a new niche may be needed in larger amounts until control of expression levels has become established.

(3) Duplications are more likely to be retained or occur in genes that are poorly optimized. Thus, amplification is more likely to succeed in category A since there is a lower risk of counter-selection and a greater chance of discovering a weak but novel function.

(4) Genes may be imported in multiple copies into a genome. Even though this is not an effect of the duplication mechanism of the new host, the multiple imports still act as if they were actual paralogs, with the same restrictions. This explanation would be largely indistinguishable from the others.

Although all genes or open reading frames (ORFs) annotated as being transposon- or phage-related have been removed from this analysis, the possibility always remains that some such sequences remain that are as yet unrecognized. Therefore, explanation (1) cannot fully be ruled out. However, we believe that explanation (3) is more pertinent from observations in category D, where duplication occurs more often among genes with poorly understood function, and also from previous observations [4] where we find that highly expressed genes are less likely to be duplicated. This suggests that amplification of a gene product with a weak function is generally less disruptive and costly than amplification of a stronger gene product. If this is the case, the bias for duplications in category A need not be explained by areas of high recombination or increased mobility, but simply by virtue of a larger proportion of poorly optimized genes and gene functions among the potential imports.

As virtually all duplicated recent imports are annotated as hypothetical or putative, it is not easy to suggest functions. It is possible, and even likely, that some recent imports have strong gene functions that are required in a new niche and could be duplicated to provide an appropriate gene-dosage effect. However, recent imports that have identifiable functions are rarely duplicated. Thus it seems reasonable to suggest that most of the duplicated genes among the recent imports have poorly optimized functions and that the apparent bias for duplication is due to selection for gene dosage. Nevertheless, we expect that most of these duplicated imports are also transient, as are most other imports and other duplications. The main point is that neutral and near-neutral genes have a limited persistence in the genome and that the pool of weakly functional or nonfunctional genes is dominated by the recent imports. This is where new functions seem most likely to be discovered.

Lateral transfer and gene innovation

If every duplication event has roughly the same chance of being selected by amplifying a gene product, then the contribution of lateral transfer to gene innovation can be estimated. Here we assume that the observed duplications are essentially neutral or weakly selected, and that deleterious amplifications are quickly purged from the population. In *E. coli* K12 categories A and B, when compared to both *E. coli* O157:H7 and *S. typhimurium*, there are roughly half as many duplication events as in category D. In the corresponding *S. typhimurium* comparisons, a quarter of all duplications are in category A or B. For the proteobacteria, *E. coli* O157:H7 tops the list, with almost half of its recent duplications in category A or B. In the *Bacillus/Clostridium* group, more than half of all duplications are in these categories.

As a conservative estimate, under the assumption that novel gene functions arise through modifying paralogs that are retained through an amplification of primarily weak functions, we propose that at least a quarter of all gene-innovation events are a direct consequence of lateral transfer. The estimate is conservative as there may be duplications in D that are 'stuck' in the amplified function, such as elongation factor and, possibly, ABC transporter genes. In any case, lateral transfer in combination with duplication may have a considerable contribution to gene innovation. This contribution is not always apparent when focusing solely on the transfer of established gene functions.

Materials and methods

Genome data

A number of genomes were downloaded from GenBank, and abbreviated as follows: *Escherichia coli* K12 (ecol) [23]; *E. coli* O157:H7 EDL 933 (edl) [24]; *Shigella flexneri* (sflex) [25]; *Salmonella typhimurium* (stym) [26]; *S. typhi* (styp) [27]; *Bacillus subtilis* (bsub) [28]; *B. halodurans* (bhal) [29]; *Clostridium perfringens* (cper) [30]; *C. acetobutylicum* (cace) [31]; *Staphylococcus aureus* strains mu50 (samu) [32] and mw2 (samw) [33]; and *Pseudomonas aeruginosa* (paer) [34]. Furthermore, contigs from *Klebsiella pneumoniae* (klebs), which is nearing completion, were downloaded from the Genome Sequencing Center [35].

ORFs annotated as related to either transposons or phage sequences were removed, because occurrences of these sequences may prove confounding. Only genes longer than 400 nucleotides were considered, because of small-sample effects when calculating cT^2 scores (see below).

The bacteria studied here fall loosely into two clades: the *Bacillus/Clostridium* and the *Salmonella/Escherichia* clades. Although it is often difficult to produce good estimates of bacterial divergence, 16s RNA trees (data not shown) suggest that the bacteria in the *Salmonella/Escherichia* clade diverged much later than the bacteria in the *Bacillus/Clostridium* clade, with the possible exception of the *S. aureus* strains, which may be comparable to the *Salmonella/Escherichia* clade.

Lateral transfer

As a measure of lateral transfer into various genomes, we use a phylogenetic approach among a group of related bacteria. For instance, as the divergence of *Salmonella* and *Escherichia* is more recent than the divergence of their ancestor to *Klebsiella* and *Pseudomonas*, we assume that a gene present only in *Escherichia* is a likely lateral transfer (gene import) that occurred after the divergence of *Salmonella* and *Escherichia*, or a transfer that occurred before divergence but was subsequently lost in *Salmonella*. This category of genes contains likely candidates for lateral transfer. As support, we use a method developed to detect atypical nucleotide context biases (cT^2 [18,19]). As defined, this context bias is independent of the nucleotide usage of the gene considered and is expected to reflect a mutation bias that deviates from the average of the genome. The values of cT^2 are low when genes appear typical and high when they appear atypical. When a foreign gene is first introduced, it may appear atypical because it has been adapted to a different mutation bias. In time it is expected to ameliorate and approach the context bias of its new host.

Organisms are organized into groups of four. The first group comprises the subject organism (for example, *E. coli* K12) and a close relative (*S. typhimurium*). These are then compared with two organisms that are more distant (*K. pneumoniae* and *P. aeruginosa*). There are four categories of genes that are considered: A, B, C and D, according to Table 1 (compare [18]). Category A is assumed to include very recent gene imports, entering the genome after the last divergence. Category B is presumed to also include imports, although to a lesser extent than category A, as there is the increased probability of genes being lost in the outlier genomes. Category C, which is only briefly addressed in this study, is likely to be composed of genes that have been lost in the paired genome and retained in the others. Category D consists mainly of genes that have been present for a long time and would include imports only if they are very pervasive, that is, they have been lost and reimported or imported several times.

As a threshold for determining presence or absence, we used the *blastx* package [36], which gives six-frame translations of nucleotide sequence to amino acid sequence. An E-value of less than 10^{-10} indicates presence in the studied genome. The value was set to a relatively low level of significance because of the highly varying degrees of divergence of the organisms studied. Furthermore, we chose the same threshold for all organisms, as any variable value would imply that the relative ages of divergence were known.

Duplication

As a simple measure of recent duplications, genomes were searched against themselves using *blastn* [36]. Genes that could fulfill both a more stringent E-value of lower than 10^{-20} and a nucleotide identity of ≥ 95% were considered to be recent duplications. There is always the possibility that such genes are not recent paralogs, but highly conserved, through gene conversion for instance. However, if these paralogs are so highly conserved, they would more likely be found in categories C or D.

Acknowledgements

This work was supported by the Swedish Research Council.

References

1. Ohno S: *Evolution by Gene Duplication*. Heidelberg: Springer-Verlag; 1970.
2. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3:**research0008.1-0008.9.
3. Hooper SD, Berg OG: **On the nature of gene innovation: duplication patterns in microbial genomes.** *Mol Biol Evol* 2003, **20:**945-954.
4. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, *et al.*: **Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58.** *Science* 2000, **287:**1809-1815.
5. Moszer I, Rocha EP, Danchin A: **Codon usage and lateral gene transfer in *Bacillus subtilis*.** *Curr Opin Microbiol* 1999, **2:**524-528.
6. Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A: **Evidence for horizontal gene transfer in *Escherichia coli* speciation.** *J Mol Biol* 1991, **222:**851-856.
7. Nesbo CL, L'Haridon S, Stetter KO, Doolittle WF: **Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria.** *Mol Biol Evol* 2001, **18:**362-375.

8. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: **Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles.** *Trends Genet* 1998, **14:**442-444.
9. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44:**383-397.
10. Lawrence JG, Ochman H: **Molecular archaeology of the *Escherichia coli* genome.** *Proc Natl Acad Sci USA* 1998, **95:**9413-9417.
11. Garcia-Vallve S, Romeu A, Palau J: **Horizontal gene transfer in bacterial and archaeal complete genomes.** *Genome Res* 2000, **10:**1719-1725.
12. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12:**17-25.
13. Doolittle WF: **Phylogenetic classification the universal tree.** *Science* 1999, **284:**2124-2129.
14. Lawrence JG: **Gene transfer in bacteria: speciation without species?** *Theor Popul Biol* 2002, **61:**449-460.
15. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer.** *Mol Biol Evol* 2002, **19:**2226-2238.
16. Kurland CG: **Something for everyone: horizontal gene transfer in evolution.** *EMBO Rep* 2000, **1:**92-95.
17. Berg OG, Kurland CG: **Evolution of microbial genomes: sequence acquisition and loss.** *Mol Biol Evol* 2002, **19:**2265-2276.
18. Hooper SD, Berg OG: **Gene import or deletion - a study of the difference genes in *Escherichia coli* strains K12 and O157:H7.** *J Mol Evol* 2002, **55:**734-744.
19. Hooper SD, Berg OG: **Detection of genes with atypical nucleotide sequence in microbial genomes.** *J Mol Evol* 2002, **54:**365-375.
20. Sharp PM, Li W-H: **The Codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.
21. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F: **Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccaromyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity.** *Nucleic Acids Res* 1988, **16:**8207-8211.
22. Guindon S, Perriere G: **Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes.** *Mol Biol Evol* 2001, **18:**1838-1840.
23. 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.
24. Perna NT, Plunkett G III, Burland V, Mau BM, 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.
25. Jin Q, Yuan ZH, Xu JG, Wang Y, Shen Y, Lu WC, Wang JH, Liu H, Yang J, Yang F, *et al.*: **Genome sequence of *Shigella flexneri* 2a, insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157.** *Nucleic Acids Res* 2002, **30:**4432-4441.
26. 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.
27. 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.
28. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390:**249-256.
29. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki Y, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, *et al.*: **Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*.** *Nucleic Acids Res* 2000, **28:**4317-4331.
30. Shimizu T, Ohtani K, Hirakawa H, Ohshima K, Yamashita A, Shiba T, Ogasawara N, Hattori M, Kuhara S, Hayashi H: **Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater.** *Proc Natl Acad Sci USA* 2002, **99:**996-1001.
31. Nolling J, Breton G, Omelchenko MV, Markarova KS, Zeng Q, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J, *et al.*: **Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*.** *J Bacteriol* 2001, **183:**4823-4838.
32. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, *et al.*: **Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*.** *Lancet* 2001, **357:**1225-1240.
33. Baba T, Takeuchi F, Kuroda M, Yuzawa H, Aoki K, Oguchi A, Nagai Y, Iwama N, Asano K, Naimi T, *et al.*: **Genome virulence determinants of high virulence community-acquired MRSA.** *Lancet* 2002, **359:**1819-1827.
34. Stover CK, Pham X-QT, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FSL, Hufnagle WO, Kowalik DJ, Lagrou M, *et al.*: **Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen.** *Nature* 2000, **406:**959-964.
35. **Genome Sequencing Center** [http://genome.wustl.edu]
36. Altschul SF, Gish W, Miller M, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.