Reconciling the many faces of lateral gene transfer

Jeffrey G. Lawrence and Howard Ochman

The various methods for detecting potential lateral gene transfer events typically uncover different sets of genes. Because the procedures used to recognize transferred genes ask different types of questions, the sets of genes identified by each procedure must be interpreted in the appropriate context. The integration of biological information, along with these analytical procedures, makes it possible to assess the total impact of lateral gene transfer on microbial genomes.

The identification of genes introduced by lateral gene transfer has changed from the occasional pursuit of phylogenetic incongruencies into an integral part of genomic analyses [1–4]. Unfortunately, the processes used to recognize alien genes are complex, methodologically mysterious and often misunderstood or improperly applied. The outcome of such differing methodologies is that two procedures, applied to the same genome, can result in overlapping, but not congruent, sets of putatively transferred genes. For example, Ochman *et al.* [5] identified 132 ‘atypical’ genes (6.4% of the genome) in *Thermotoga maritima* based on their sequence characteristics and posited that these were introduced by lateral transfer; yet Nelson *et al.* [3] predicted, based on BLAST similarity searches, that 25% of the *Thermotoga* genome had been acquired from the Archaea alone. Taken at face value, it is unclear why these estimates should be so different. Even more striking, Ragan [6] found that the sets of putative transferred genes identified by certain methods overlap significantly less than would be expected by chance, raising questions about what features of genes these methods are actually detecting. These apparent conflicts arise because each of the methods used to detect horizontally transferred genes recognizes different features in their target genes and are thus testing different sorts of hypotheses. Because the impact of lateral gene transfer over the entire evolutionary history of a lineage must be inferred from present-day sequences, the different approaches used to recognize horizontally transferred genes must rely on specific models of sequence evolution, and their assumptions must be placed in an appropriate framework.

Approaches for detecting horizontally transferred genes

Phylogenetic methods, which recognize a gene’s unusual similarity or distribution among organisms, are the most intuitive way of identifying horizontally acquired genes. Traditionally, this approach involves comparing phylogenetic trees generated from different genes in the genome, and assessing the significance of any resulting incongruities. Alternative relationship-based tactics have been devised (e.g. Clarke’s phylogenetic discordance test [6] and Lawrence’s rank correlation test [7]), which dispense with phylogenetic reconstruction altogether; in these procedures, gene transfers are recognized by an unusually high level of similarity among genes found in otherwise-unrelated organisms.

These approaches are not without limitations – lateral transfer is not the only mechanism that produces conflicts between phylogenies. Some genes might be coincidentally deleted from multiple lineages, leading to unusual distributions among extant organisms, or sequence similarity can result from convergent evolution. Moreover, the proliferation of gene families can make the identification of orthologous sequences difficult, and rapid sequence evolution makes alignment of homologous sites equivocal. These caveats are exemplified by the seemingly premature speculation that >100 ORFs in the human genome arose by gene transfer from bacteria [4,8,9]. Despite such problems, which are compounded by the limitations of the sequence databases currently available, phylogenetic methods detect many transfer events with high degrees of certainty, including very ancient transfers, whose products might be so widely distributed among recipient hosts (e.g. the mitochondrion) that they might not be suspected of having foreign origins.

In contrast to phylogenetic approaches, there are methods to identify potentially foreign genes that do not rely on comparing genes between organisms; rather, genes that appear atypical in their current genomic context are suspected of having been introduced from a foreign source. The assumption of these methods is that directional mutation pressures within bacterial genomes impart distinctive biases to the composition of long-term residents of the genome [10], such that recently acquired genes will appear aberrant by comparison if they have evolved in a genome with different mutational biases. All such methods rely on a robust description of what defines a ‘typical’ gene, and usually such parameters are based on nucleotide composition [11,12], dinucleotide frequencies [13], codon usage biases [14–16] or patterns inferred by Markov chain analysis [17].

An advantage of these parametric approaches is that putative transferred genes can be identified without relying on comparisons with other organisms and, as a result, such methods provide an independent means of assessing the impact of gene transfer across lineages [5]. A problem associated with these methods is that genes arriving from donor genomes experiencing similar mutational biases will not be detected, because the acquired sequence will not appear unusual in the recipient genome. Moreover, such methods are limited by the amelioration of foreign genes following introduction [11]; that is, newly acquired genes will experience the same mutational biases as long-term residents of the genome and will eventually fail to be recognized as anomalous. As a result, parametric methods will most reliably detect only recently acquired genes, and will underestimate their numbers. Finally, genes might appear atypical owing to stochastic factors (especially if they are short) or to selection for unusual composition.

Different methods, different assumptions

Ragan [6] observed that different approaches, when applied to the same

http://tim.trends.com 0966-842X/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0966-842X(01)02282-X

Box 1. Consolidating methods for recognizing acquired DNA in the *E. coli* genome

The ancestry of each of the annotated ORFs in the *Escherichia coli* MG1655 genome was ascertained by examining their distributions among other enteric bacteria (*Salmonella enterica* serovars Typhimurium, Typhi and Paratyphi, and *Klebsiella pneumoniae*) and applying parsimony. Two criteria were used to establish orthology: overall level of sequence identity (>60%) and genomic context (the coincidence of adjacent ORFs across genomes). By considering both factors, the ancestry of equivocal cases, such as those in which ORFs displayed a low level of sequence identity and sporadic phylogenetic distribution, could be resolved. The PipMaker program (http://globin.cse.psu.edu/enterix), with *E. coli* MG1655 as the reference genome, greatly facilitated this analysis. Applying these criteria, genes present in two or more taxa are considered ancestral (although these could have been acquired before the divergence of the *Salmonella* lineage), whereas *E. coli* genes lacking positional homologues were considered to be horizontally acquired.

The concordance between phylogenetic and parametric methods of identifying horizontally acquired genes in *E. coli* MG1655 is shown in Fig. I, a linear representation of the chromosome. Within each centisome, each horizontal bar delineates a continuous segment of transferred DNA containing one or more ORFs, and whose length is rounded to the nearest ~500 bp; different colours represent segments of transferred DNA identified by either or both methods. Despite several reasons why these procedures are expected to identify somewhat different sets of genes, the degree of overlap (red bars) is quite good. Among the 755 genes originally identified as being horizontally acquired based on aberrant sequence characteristics, 627 (83%) display a phylogenetic distribution compatible with lateral gene transfer. Importantly, the 128 putatively transferred genes whose phylogenetic distributions did not reveal evidence of lateral transfer (blue bars) had an average length of only 628 bp compared with 1075 bp for genes detected by both methods (black bars). This suggests that stochastic factors contribute to atypical GC contents and codon usage patterns, and, thus, to the incorrect assignment of some shorter genes as having been acquired.

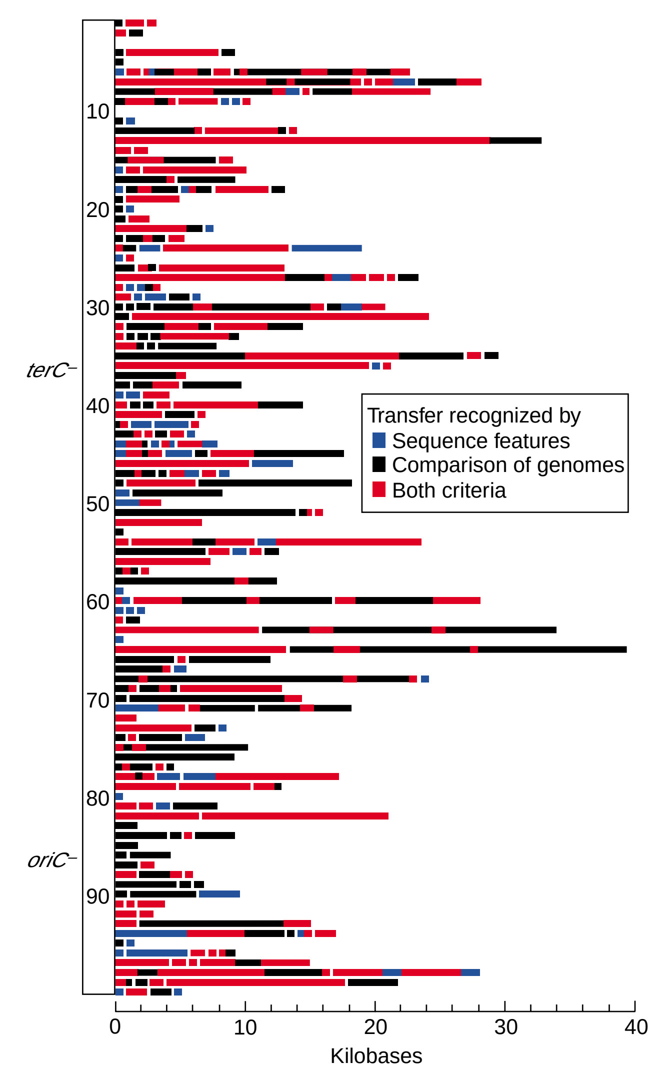

Fig. I

TRENDS in Microbiology

Based on their phylogenetic distributions, a total of 1052 ORFs (combining genes within both red and black bars in Fig. I) were acquired by the *E. coli* lineage leading to MG1655 following its divergence from *Salmonella*. This increase is anticipated given that parametric methods only rarely recognize genes acquired from organisms with similar mutational biases. Indeed, the base composition of the 425 transferred genes that were not detected as atypical (black bars) is 51.2 (± 3.6) %GC, which is very similar to the *E. coli* genome average of 51.0 %GC.

Most importantly, the large number of ORFs recognized as being horizontally acquired based only on their phylogenetic distributions (black bars) are not randomly distributed with respect to regions identified as being transferred by other means. Note that in the majority of cases, these DNA segments (black bars) are contiguous with, or link together, the previously recognized, horizontally acquired regions (red or blue bars). Thus the total number of transfer events occurring within this *E. coli* lineage has not increased with the inclusion of phylogenetic information, but the number of acquired genes and the average size of acquired regions are both greatly augmented. Ignoring potentially transferred regions identified solely on the basis of sequence features (blue bars), the total proportion of recently transferred genes in the current MG1655 genome is now estimated to be 24.5%, introduced in at least 221 events.

genomic sequence, sometimes recognized significantly different subsets of genes as being subject to lateral transfer. For example, 69% fewer genes than expected are predicted both by Clarke's phylogenetic discordance test [6] and by methods looking at atypical sequence features; he suggests that this alarming result is rooted in the different null hypotheses that are being tested by each approach. The phylogenetic discordance test identifies genes whose closest homologues are found in taxa not otherwise related to the query genome, and thus it uncovers a set of genes biased towards those which have been transferred across large phylogenetic distances, regardless of their time of arrival in a genome. As mentioned, methods that examine sequence features preferentially identify genes that have recently been introduced into a genome from an organism having different mutational biases, regardless of phylogenetic distance. Assuming that the frequency of transfer between lineages is inversely related to their phylogenetic distance, these two methods would identify quite different sets of genes. Similarly, Nelson *et al.* [3] employed a phylogenetic distance method (BLAST searches) to deduce that a large proportion (25%) of the *T. maritima* genome reflects a history of gene exchange with Archaea – a value expected to be higher than that proposed by Ochman *et al.* [5] to have arrived only recently.

Given these diverse approaches, it is not surprising that parametric and phylogenetic methods delineate somewhat different sets of genes as having been subject to gene transfer [18]. Box 1 compares these procedures for identifying genes potentially transferred into the *Escherichia coli* MG1655 genome. Despite the potential limitations of each of these methods [19,20], most of the genes identified as atypical in *E. coli* MG1655 based on sequence features are not in the genomes of related enteric bacteria (Box 1), and have limited distribution among strains of *E. coli* [21]. And, as expected, genes probably acquired by transfer, but only detected by phylogenetic methods, are not always compositionally atypical (Box 1).

Using all the data

The wealth of data available to interpret gene transfer in *E. coli* illustrates the need for a multifaceted approach to reconstructing the history of gene transfer. Although not all recently acquired genes can be identified with certainty based on sequence features alone, additional information can assist in determining its ancestry. For example, applying composition and codon-usage bias criteria to the nine-gene *fim* cluster (encoding fimbrae) at 4538 kb in the *E. coli* genome, four genes are clearly aberrant, two are clearly normal and three are equivocal. Markov chain analysis [17] also detects some of the *fim* genes as atypical, but not the same set. Yet the translational coupling among these genes makes it highly likely that the entire *fim* region was acquired in a single transfer event, and the absence of a cognate cluster in *Salmonella enterica* supports this hypothesis. Therefore, sets of genes recently acquired by lateral transfer will often include sequences lacking unusual features, and their identification requires scrutiny on the part of the investigator. Importantly, the phylogenetic discordance test fails to recognize any of the *fim* genes as horizontally acquired because related *fim* genes are found at multiple locations in many enteric bacterial genomes.

In addition to gene organization, other biological information can be used to interpret patterns of gene transfer events. For example, whereas only half of the genes in a six-gene cluster at 2464 kb in the *E. coli* MG1655 genome are compositionally unusual, the similarity of some of the genes with bacteriophage-encoded counterparts – including an integrase found adjacent to the tRNA-encoding *argW* gene – provides strong evidence that the entire region is a remnant of a bacteriophage. Here, gene identity plays a role in proposing that the entire region was recently acquired. In an analogous manner, a nine-gene cluster inserted into the *E. coli eut* operon is readily recognized as a likely interloper because genome alignments reveal that the *S. enterica eut* operon lacks such an insertion. In itself, this distribution does not fully support acquisition by *E. coli* (the genes might have been ancestral and subsequently deleted from the *Salmonella* chromosome); however, the atypical features of these genes suggest that acquisition is the more likely event. As illustrated by these three examples, the boundaries of gene transfer events are more robustly delineated when all of the available data are used.

Conclusions

Comparisons of different methods for detecting potential lateral gene transfer events in microbial genomes provide several valuable lessons: (1) a substantial fraction of recently acquired genes are insufficiently atypical to be detected by most published methods examining sequence features alone, although more sophisticated analyses perform better; (2) parsimony methods can be employed if genome sequences are available from three or more closely related taxa, but such comparisons are limited to identifying only the subset of potential transfer events defined by the taxa compared; (3) short sequences (<500 bp) often appear atypical for stochastic reasons and might be misidentified as having been transferred; and (4) phylogenetic analyses using the entire sequence database can detect ancient transfer events, but might fail to detect transfers between more closely related organisms. Whenever possible, application of a variety of methods provides the best information about the scope of gene transfer across broad timescales.

Acknowledgements

We thank Mark Ragan for sharing his data on the phylogenetic discordance test. This work was supported by a grant from the David and Lucille Packard Foundation to J.G.L., and by grants from the National Institutes of Health and Department of Energy to H.O.

References

1 Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815

2 Ruepp, A. *et al.* (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 407, 508–513

3 Nelson, K.E. *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329

4 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

5 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304

6 Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 201, 187–191

7 Lawrence, J.G. and Hartl, D.L. (1992) Inference of horizontal genetic transfer: an approach using the bootstrap. *Genetics* 131, 753–760

8 Stanhope, M.J. *et al.* (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411, 940–944

9 Salzberg, S.L. *et al.* (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292, 1903–1906

10 Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2653–2657

11 Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44, 383–397

12 Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9413–9417

13 Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* 1, 598–610

14 Mrazek, J. and Karlin, S. (1999) Detecting alien genes in bacterial genomes. *Ann. New York Acad. Sci.* 870, 314–329

15 Médigue, C. *et al.* (1991) Evidence of horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* 222, 851–856

16 Moszer, I. *et al.* (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.* 2, 524–528

17 Hayes, W.S. and Borodovsky, M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.* 8, 1154–1171

18 Koski, L.B. *et al.* (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* 18, 404–412

19 Wang, B. (2001) Limitations of compositional approach to identifying horizontally transferred genes. *J. Mol. Evol.* 53, 244–250

20 Guindon, S. and Perrière, G. (2001) Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol. Biol. Evol.* 18, 1838–1840

21 Ochman, H. and Jones, I.B. (2000) Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J.* 19, 6637–6643

Jeffrey G. Lawrence*
Pittsburgh Bacteriophage Institute and
Dept of Biological Sciences, University of
Pittsburgh, Pittsburgh, PA 15260, USA.
*e-mail: jlawrenc@pitt.edu

Howard Ochman
Dept of Ecology and Evolution,
University of Arizona, Tucson,
AZ 85721, USA.

Reconciling the many faces of lateral gene transfer

Response from Ragan

Until a few years ago, most biologists assumed – or would have, had the question arisen – that genome evolution is basically gene evolution writ large. Genomes were viewed as physically and functionally linked sets of genes that travel coherently together, through time, along tree-like bifurcating lineages. As evidence accumulated that some genes have instead been transmitted across lineages, methods were introduced to search more systematically for them. Four of these methods [1] confirm that lateral gene transfer (LGT) has indeed been quantitatively important in shaping the genome of *Escherichia coli* MG1655; but most pairs of these methods identify a common set of suspect genes less often than expected by chance under a simple model [1,2]. How are we to interpret this?

To some extent, the different methods test different null hypotheses, for example, events of different antiquities [1]. If this is the problem, better alignment of null hypotheses should increase agreement among tests. Lawrence and Ochman now show that comparing their compositional test, which most efficiently identifies recent transfers, with a test based on the presence or absence of 'positional homologs' in genomes thought to have diverged within this same time frame – for example, other enteric bacteria, not all bacteria [3] or all organisms (G.D.P. Clarke *et al.*, unpublished), as considered earlier [1] – dramatically improves agreement. Some 83% of compositionally atypical genes in *E. coli* MG1655 are distributed among enteric bacteria in a way suggestive of LGT. The remaining 17% are, on average, shorter than those in the intersection, suggesting a purely stochastic explanation for many false negatives (and, by extrapolation, also false positives). Distributionally suspect and compositionally anomalous genes are often contiguous. Thus, genes were transferred into the *E. coli* genome as large blocks, not singly, and compositionally typical genes within these blocks are secondarily so.

This view of genome evolution has four consequences:

First, LGT might better be represented as the union, not the intersection, among tests addressing compatible null hypotheses. This would shift the burden of proof, with genes in suspect regions assumed to be of lateral origin unless proven otherwise.

Second, it points the way towards more-complex models of genome evolution, in which the parameters include the number and size distribution of blocks, their spatial distribution along the genome and the rate and patchiness at which regions within blocks are ameliorated. Are these parameters genome- or taxon-specific? What ranges of intersections and gene spacings arise from simulations under such a model?

Third, it could explain why putatively older events are more widely spaced [1]: the intervening genes have been transferred there more recently, perhaps in single events.

Finally, as the two tests deployed by Lawrence and Ochman address only relatively recent transfers, much more than 24.5% of the *E. coli* MG1655 genome is likely to be of lateral origin. Far from suggesting a basic flaw with LGT, 'disagreement' among tests seems to be pointing the way towards a more comprehensive hypothesis of genome evolution.

References

1 Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 201, 187–191

2 Koski, L.B. *et al.* (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* 18, 404–412

3 Ragan, M.A. and Charlebois, R.L. Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *Int. J. Syst. Evol. Microbiol.* (in press)

Mark A. Ragan
The Institute for Molecular Bioscience,
The University of Queensland, Brisbane,
Qld 4072, Australia.
e-mail: m.ragan@imb.uq.edu.au

http://tim.trends.com 0966-842X/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0966-842X(01)02283-1