Chapter 12

Composition-Based Methods to Identify Horizontal Gene Transfer

Diego Cortez, Luis Delaye, Antonio Lazcano, and Arturo Becerra

Abstract

The detection of horizontal gene transfer (HGT) events has become an increasingly important issue in recent years. Here we discuss a simple theoretical analysis based on the *in silico* artificial addition of known foreign genes from different prokaryotic groups into the genome of *Escherichia coli* K12 MG1655. Using this dataset as a control, we have tested the efficiency of four methodologies commonly employed to detect HGT, which are based on (a) the codon adaptation index, codon usage, and GC percentage (CAI/GC); (b) the distributional profile (DP) approach with a gene search in the closely related phylogenetic genomes; (c) the Bayesian model (BM); and (d) the first-order Markov model (MM). All methods exhibit limitations as shown here, with BM and MM giving better approximations. The MM has a better detection rate when genes from closely related organisms are evaluated. The application of the MM to detect recently transferred genes in the genomes of *E. coli* strain K12 MG1655 shows that this organism has undergone a rather significant amount of HGT, several of which have well-defined functions that appear to be involved in the direct interaction of the organisms with their environment.

Key words: Horizontal gene transfer, composition-based methods, methods to identify HGT.

1. Introduction

Horizontal gene transfer (HGT) is generally assumed to have played an important role in the innovation of genomes, especially during the early stages of biological evolution *(1–4)*. However, some authors have suggested that the role of HGT has been overestimated *(5,6)*. The current controversy lies, in part, in the inadequate or inefficient methods to identify transfer events.

Many attempts have been made to characterize the number and types of genes involved in HGT, but as shown by ongoing

Maria B. Gogarten et al. (eds.), *Horizontal Gene Transfer: Genomes in Flux, vol. 532*

DOI 10.1007/978-1-60327-853-9_12 Springerprotocols.com

debates, as of today no infallible detection methods have been developed *(7–16)*. During the past decade, different approaches have been proposed, which can be classified in two major categories: (a) the composition-based methods and (b) anomalous phylogenetic distribution. These methods can be divided into more specific groups. Although it has been suggested that phylogenetic methods are more powerful than compositional methods to detect HGT in particular situations (i.e., when the donor is closely related to the recipient genome), they tend to be time-consuming. Therefore, the development of a compositional method with an accurate detection level of horizontally transferred genes would be an effective approach in the analysis of possible HGT in large numbers of completely sequenced genomes.

Determination of the levels of HGT depends on (i) the time of their occurrence; (ii) the phylogenetic relationship between the donor and the acceptor; and (iii) the "camouflage" of the gene; it is also affected by gene loss and the rate of divergence. Although one should not abandon all hope, it is important to keep in mind that the sequences may have undergone important signal loss.

Two examples of the controversies created by different estimates of HGT are the discrepancies in shaping the topology of the tree of life *(1, 4, 17)* and the reconstruction of the Last Common Ancestor (LCA) of all extant organisms *(18)*. The lack of congruency between different universal phylogenies may result not only from the statistical issues involved in the alignment and comparison of a large number of sequences that may have diverged more than 3.5×10^9 years ago, but also from even older additional duplications *(19)* and HGT events *(1)*, both of which may be obscuring the natural relationship between the lineages. Several authors have argued that the use of genes that are less likely to be transferred are expected to yield deep phylogenies with suitable results *(20, 21)*, and they allow not only to explore the evolution of early stages but also to backtrack the characteristics of the Last Universal Common Ancestor *(22)*. Is it generally accepted that HGT was rampant during the early evolution of genomes. Faster and better methods of HGT detection could improve our understanding of the early evolution of the life significantly.

2. Methodologies Used for the Detection of HGT

Current approaches for detecting HGT can be broadly divided into three major groups: (a) methods based on codon usage such as the codon adaptation index (CAI), the GC percentage analysis, Bayesian models, and higher order Markov models, all of which attempt to identify genes with anomalous compositions *(10, 12, 13, 16, 23)*; (b) comparison of the gene content of an organism

with that of closely related species, based on the distributional profiles (DP) determined for every single gene in the genome *(15, 24)* – this method is based on the idea that if a gene present in the target genome is not found in any closely related genomes, given a variety of threshold values, the sequences are considered to have undergone HGT – and (c) phylogenetic reconstruction, which is based on phylogenetic conflict by alien-gene acquisition *(25)*.

Compositional approaches based on codon usage and GC content have been criticized since in some case the dissimilarities in base composition and codon usage between possible transferred genes and host sequences can be truly minor *(26, 27)*. Indeed, these results may be the outcome of compositional heterogeneity, which is now recognized as a characteristic of cellular genomes *(28, 29)*. For instance, when the compositional method developed by Lawrence and Ochman *(12)*, the high-order MM used by Hayes and Borodovsky *(10)*, the phylogenetically discordant approach followed by Clarke et al. *(30)*, and an anomalous phylogenetic distribution model by Ragan and Charlebois *(24)* were all applied to analyze the *Escherichia coli* K12 MG1655 genome, each method detected a very different set of possible horizontally transferred genes, and the intersections between these sets were less than expected by chance alone *(31)*. The development of a compositional method with an accurate detection level of horizontally transferred genes would be a powerful approach that could avoid the application of exhaustive processes and slow phylogenetic reconstructions that, moreover, might not lead to better results than the compositional methods.

In this chapter we describe a different approach based on the analysis of the significance of the most frequently used methodologies for the detection of HTG and describe a simple theoretical approach that uses the *in silico* inclusion of known foreign genes from different prokaryotes into a chosen model genome *(8)*.

2.1. The Accuracy of the Methodologies Based on Composition

The methodologies discussed here have been extensively employed during the past few years. However, they have usually been used without proper experimental controls that could allow an estimate of their accuracy. A simple theoretical approach was recently developed by Cortez et al. *(8)*, which consists in the *in silico* artificial inclusion of known foreign genes from different mayor prokaryotic groups into a given genome. Using this methodology we have tested the efficiency of four approaches, which are based on (a) the codon adaptation index, codon usage, and GC percentage *(12)*; (b) a distributional profile for every single gene in a genome *(15)*; (c) a Bayesian model *(16)*; and (d) a first-order Markov model we developed *(8)*. Genes were selected randomly from 30 different organisms in order to test these methodologies. The foreign genes were then randomly inserted *in silico* into the genome of *E. coli* K12 MG1655. The genes

detected by each method were compared *(8)*. Our results show that the MM appears to be the most reliable approach to identify horizontally transferred genes, especially when they come from closely related species. The MM was then employed to analyze the foreign sequences acquired by *E. coli* K12 MG1655. Enterobacteria have a significant percentage of foreign genes, many of which have defined functions that might be involved in the direct interaction of the organism and its environment.

3. "Who Is Who" in Composition-Based Detection?

The average detection levels of artificially introduced foreign genes are shown in **Fig. 12.1**. The BM and the MM consistently detected the foreign, introduced genes. However, the MM appears to be a much better approach when genes from closely related species are studied.

The CAI/GC method appears to be a less effective approach in the detection of HGT. However, it has better detection levels when the artificially introduced foreign genes came from phylogenetically distant species, or when they belonged to

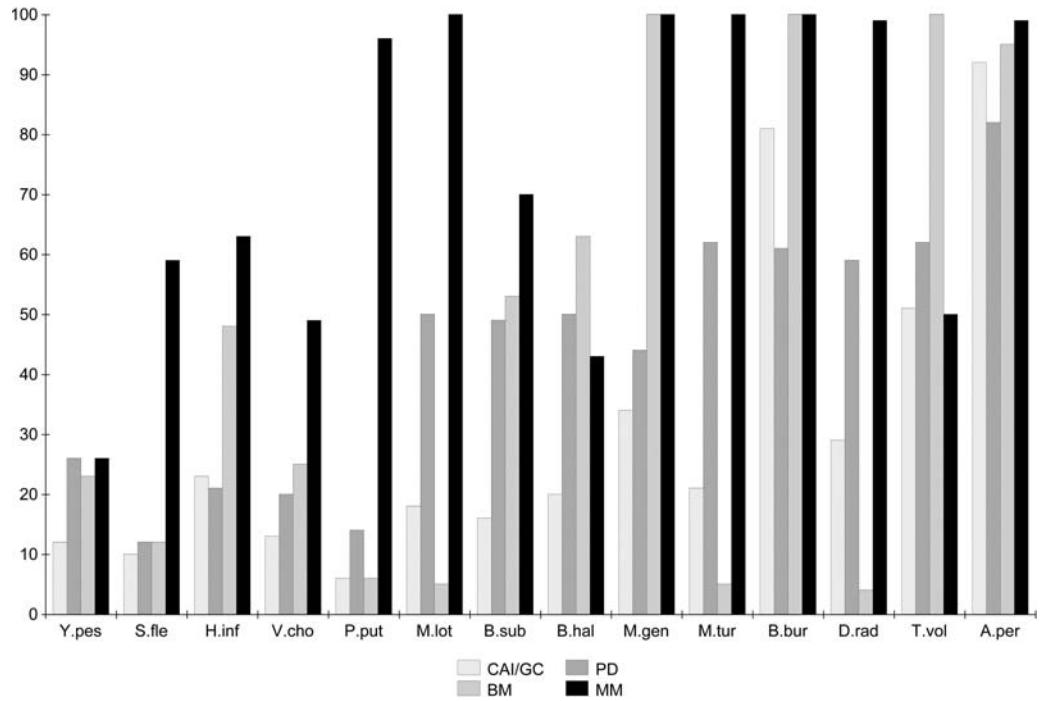

Fig. 12.1. Detection levels of 100 artificially introduced foreign genes from different organisms by the four methods discussed here: CAI/GC, PD, BM, and MM. Where: Y.pes, *Y. pestis*; S.fle, *S. flexneri*; H.inf; *H. influenzae*; V.cho, *V. cholerae*; P.put, *P. putida*; M.lot, *M. loti*; B.sub, *B. subtilis*; B.hal, *B. halodurans*; M.gen, *M. genitalium*; M.tur, *M. tuberculosis*; B.bur, *B. burgdorferi*; D.rad, *D. radiodurans*; T.vol, *T. volcanium*; A.per, *A. pernix*.

Table 12.1
Functional characteristics of the four methodologies. Average detection levels of genes from close and distant prokaryotic species based on the model's performance in the experiment of *in silico* introduction of foreign genes. Probabilities of having false negatives and false positives are shown. The faculty to discriminate between gene loss and HT is shown too (for more information *see (8)*)

Model	Average detection levels of genes from distant species	Average detection levels of genes from Proteobacteria species	Probability of having false positives	Probability of having false negatives	Faculty to discriminate between gene loss and HT	Number of possible HTG in *E. coli* K12
MM ($p < 0.01$)	High	Medium	Low	Low	YES	422
BM ($p < 0.01$)	High	Low	Low	Medium	NO	443
CAI/GC ($p < 0.33$)	Medium	Low	Medium	High	NO	324
PD (e^{-6})	Medium	Low	Medium	High	NO	545

genomes whose average %GC content was much lower than that of the *E. coli* K12 MG1655 (50.8% GC) genome. The DP exhibited average detection levels of approximately 50% of the introduced foreign genes. These methods (CAI/GC and DP) are efficient only when the artificially introduced foreign genes came from phylogenetically distant species. In this case, the results are phylogenetically coherent: the more related the species that the genes are taken from are to the analyzed genome, the lower the HGT detection level was. However, the DP method failed to go beyond the 80% threshold of detection, with the exception of those genes coming from *Aeropyrum pernix*, which is comparable to the CAI results observed in almost all cases.

Of all the methods analyzed, the MM appears to be the best methodology for a proper detection of HGT, and it appears to be the most reliable strategy to detect transfers from closely related species, which are believed to be the most frequent ones. The BM and the MM consistently detected the foreign, introduced genes. However, the MM appears to be more accurate when genes from closely related species are analyzed.

The functional properties of all models are summarized in **Table 12.1**. The average detection levels were obtained from the performance of the four methodologies in the experiment of the *in silico* introduction of foreign genes. The probabilities of having false negatives and positives were also analyzed.

4. What Kind of Genes Are Detected by Each Method (HGT and Function)?

A detailed analysis of the foreign genes found by the different methodologies shows that most of them belong to the unassigned functional category. However, a significant percentage of these putative horizontally transferred genes has well-defined functions or belongs to an assigned functional category. Few genes detected by the MM method were informational genes. In pathogenic strains, the sets of genes that have undergone HGT included a number of sequences, which are directly related to pathogenesis. Important aspects that could help to understand the HGT dynamics are: (a) most of the HGT detected belong to the unassigned functional category; these genes are mainly hypothetical insertion elements and phage-related (as shown in **Table 12.2**, most of these genes are also pseudogenes), (b) HGT belonging to all major functional categories also can be found, including a few informational genes such as a ribosomal protein, a DNA polymerase II ε subunit, etc. (nevertheless, these informational genes also appear to be pseudogenes); and (c) HGT that belongs to other functional categories (for instance, amino acid metabolism, carbohydrate metabolism, membrane transport,

Table 12.2
Functional categories of the transferred genes and the pseudogenes detected using the MM in the *E. coli* K12 MG1655 genome

Functional categories	*HGT*	*Pseudogenes*
Amino acid metabolism	16	3
Biodegradation of xenobiotics	5	3
Biosynthesis of secondary metabolites	1	0
Carbohydrate metabolism	27	6
Cell motility	0	0
Energy metabolism	4	1
Hypothetical	90	265
Insertion elements	2	16
Lipid metabolism	5	1
Membrane transport	43	8
Metabolism of cofactors and vitamins	2	1
Metabolism of complex carbohydrates	0	4
Metabolism of complex lipids	0	0
Metabolism of other amino acids	0	1
Nucleotide metabolism	4	0
Phage-related	2	4
Replication and repair	0	1
Signal transduction	1	2
Sorting and degradation	0	4
Transcription	3	3
Transcriptional regulator	10	4
Translation	0	0

nucleotide metabolism, metabolism of cofactors and vitamins, energy metabolism, which are less likely to be pseudogenes).

5. Discussion

The four methodological approaches compared in this work have been widely used by different authors during the past few years *(10, 12, 13, 15, 16)*. However, they have generally been employed

without proper experimental controls that could allow an accurate assessment. A control methodology based on the detection of prokaryotic genes and phage genes from two distinct genomic pools (one including all the prokaryotic genes from complete sequenced genomes and the other containing several phage genes) has been recently suggested *(23)*.

The introduction of a simple theoretical control, such as the artificial addition of prokaryotic foreign genes into a genome as discussed here, demonstrates that the accuracy level of the different methods may be low, with the exception of the BM and the MM methods that consistently detected the foreign, introduced genes and are accurate when genes from closely related species are under evaluation.

5.1. Different Methods Give Different Results: Is a Biological Explanation Feasible?

The proposal that massive amounts of genetic material can be promiscuously exchanged between prokaryotes raises the issue whether all the genes from a genome are equally subject to HGT *(2)*. If this is the case, foreign DNA would eventually replace all the vertically inherited genes in a few millions years *(12)*. This has led to the proposal that the history of life should not be represented as a tree but rather as a complex network *(1)*. Considerable efforts have been undertaken to discuss this hypothesis *(32–34)*. Recently, it was shown that all orthologs with non-rRNA-type phylogenies encode for unassigned proteins *(25)*. This would imply that the most essential genes from a genome, i.e., those with high selective pressure and which can be part of large complexes, such as informational genes *(35,36)*, seem to be less likely to undergo HGT. The functional analysis of HGT confirms that most of the transferred genes encode for unassigned proteins.

However, detailed analysis shows that some of these genes have well-defined functions or belong to well-defined assigned functional categories. These genes might be encoding for the proteins involved in the direct interaction of the organism with its environment, such as membrane proteins, cytosolic enzymes, and pathogenesis-related proteins. This possibility is supported by the recognition that high levels of HGT events are recognizable in prokaryotic populations living in environments polluted with xenobiotics *(37)*. Furthermore, some badly preserved, informational genes were also detected (see results). This may imply that this sort of genes can be transferred but, for these enterobacterial species, they have not been selected in the host genome, and thus, they have experienced a sequence decay by accumulating non-sense mutations.

The results analyzed here demonstrate that it is possible to detect HGT through the compositional approaches (BM and MM) assembled with Markov chains and using Monte-Carlo simulations for statistical purposes. It is particularly interesting to observe the complementary nature of these two approaches; this

suggests that the ideal detection model could be shaped using the combination of several of these approaches.

The MM exhibited its best detection levels when the transferred genes belong to distantly related organisms (**Fig. 12.1**). This implies that the detected horizontally transferred genes in *E. coli* have originated from very distant organisms, and still reflect their previous genome context. Thus it could be concluded that the barriers that limit gene exchange among prokaryotes are ecological-environmental rather than species-dependent *(36)*. As shown in **Table 12.1**, there are a high number of independent events of lateral acquisition of genes. This is consistent with the previous discussion by Ochman et al. *(2)*, Blattner *(38)*, and Perna et al. *(39)*.

During the past two years, new detection methods have been developed based on (a) the composition-based methods and (b) anomalous phylogenetic distribution (*see* **Chapters 11 and 13**). Some of these methods have a good degree of accuracy and with proper controls may yield reliable estimates. These methods include those proposed by (i) Azad and Lawrence *(40)*, which is based on a entropic clustering method and genome position; (ii) Choi and Kim *(6)*, which estimate the global extent of HGT by statistical method and analyzing protein domain families; and (iii) Linz et al. *(41)*, which propose the use of a likelihood framework to measure HGT.

In conclusion, the proper detection of HGT is affected by several, frequently occurring processes that dim the evolutionary history of the organisms and their genes. No method will be infallible, but the different approaches developed during the past few years suggest that, if we use suitable controls and consider the processes that affect the search, it is possible to generate reliable results, at least enabling us to discuss the most important hypotheses on the evolution of genomes. Furthermore, the determination of overall HGT rates and their impact on genome evolution requires that the calculation explicitly considers the rates of false positives and false negatives, and that *in silico* transfer provides an elegant approach to estimate these parameters.

Acknowledgments

This work was supported in part by CONACYT-Mexico (Project 50520-Q) to A.L. and (Project 52226) to A.B.

References

1. Doolittle, W. F. (1999) Phylogenetic classification and the universal tree. *Science* **284**, 2124–29.
2. Ochman, H., Lawrence, J. G., Groisman, E. A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304.

3. Jain, R., Rivera, M. C., Moore, J. E., Lake, J. A. (2003) Horizontal gene transfer accelerates genomes innovation and evolution. *Mol Biol Evol* **20**, 1598–1602.
4. Brown, J. R. (2003) Ancient horizontal gene transfer. *Nat Rev Genet* **4**, 121–32.
5. Galtier, N. (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol* **56**, 633–42.
6. Choi, I. G., Kim, S. H. (2007) Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A* **104**, 4489–94.
7. Snel, B., Bork, P., Huynen, M. A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**, 17–25.
8. Cortez, D. Q., Lazcano, A., Becerra A. (2005) Comparative analysis of methodologies for the detection of horizontally transferred genes: A reassessment of first-order Markov models. *In silico Biology* **5**, 581–92.
9. Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R., Koonin, E. V. (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet* **14**, 442–4.
10. Hayes, W. S., Borodovsky, M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res* **8**, 1154–1171.
11. Kyrpides, N. C., Olsen, G. J. (1999) Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry? *Trends Genet* **15**, 298–9.
12. Lawrence, J. G., Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 383–97.
13. Garcia-Vallve, S., Romeu, A., Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**, 1719–25.
14. Hooper, S. D., Berg, O. G. (2002) Gene import or deletion: a study of the different genes in Escherichia coli strains K12 and O157:H7. *J Mol Evol* **55**, 734–44.
15. Daubin, V., Lerat, E., Perriere, G. (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol* **4**, R57.
16. Nakamura, Y., Itoh, T., Matsuda, H., Gojobori., T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature* **36**, 760–6.
17. Briones, C., Manrubia, S. C., Lázaro, E., Lazcano, A., Amils, R. (2005) Reconstructing evolutionary relationships from functional data: a consistent classification of organisms based on translation inhibition response. *Mol Phyl Evol* **34**, 371–81.
18. Delaye, L., Becerra, A., Lazcano, A (2005) The last common ancestor: what's in a name? *Orig Life Evol Biosph* **35**, 537–54.
19. Forterre, P. (1993) The great virus comeback – from an evolutionary perspective. *Res Microbiol* **154**, 223–5.
20. Philippe, H., Douady, C. J. (2003) Horizontal gene transfer and phylogetics. *Curr Opin Microbiol* **6**, 498–505.
21. Kurland, C. G., Canback, B., Berg, O. G. (2003) Horizontal gene transfer: A critical view. *Proc Natl Acad Sci U S A* **95**, 9413–7.
22. Becerra, A., Delaye, L., Islas, S., Lazcano, A. (2007) The very early stages of biological evolution related to the nature of the last common ancestor of the three major cell domains. *Annu Rev Ecol Evol Syst* **38** (in press).
23. Tsirigos, A., Rigoutsos, I. (2005). A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res* **33**, 922–33.
24. Ragan, M. A., Charlebois, R. C. (2002) Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *Int J Syst Evol Microbiol* **52**, 777–87.
25. Daubin, V., Moran, N. A., Ochman, H. (2003) Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–32.
26. Koski, L. B., Morton, R. A., Golding, G. B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18**, 404–12.
27. Wang, B. (2001) Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol* **53**, 244–50.
28. Guindon, S., Perriere, G. (2001) Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol Biol Evol* **18**, 1838–40.
29. Daubin, V., Perriere, G. (2003) G + C3 structuring along the genome: a common feature in prokaryotes. *Mol Biol Evol* **20**, 471–83.
30. Clarke, G. D., Beiko, R. G., Ragan, M. A., Charlebois, R. L. (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* **184**, 2072–80.
31. Ragan, M. A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* **201**, 187–91.
32. de la Cruz, F., Davies, J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* **8**, 128–33.
33. Gogarten, J. P., Doolittle, W. F., Lawrence, J. G. (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226–38.

34. Yang, S., Doolittle, R. F., Bourne, P. E. (2005) Phylogeny determined by protein domain content. *Proc Natl Acad Sci* **102**, 373–8.
35. Rivera, M. C., Jain, R., Moore, J. E., Lake, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci* **95**, 6239–44.
36. Jain, R., Rivera, M. C., Lake, J. A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci* **96**, 3801–6.
37. Top, E. M., Springael, D. (2003) The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr Opin Biotechnol* **14**, 262–9.
38. Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., et al. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–74.
39. Perna, N. T., Plunkett, G. 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–33.
40. Azad, R. K., Lawrence, J. G. (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res* **35**, 4629–39.
41. Linz, S., Radtke, A., von Haeseler, A. (2007) A likelihood framework to measure horizontal gene transfer. *Mol Biol Evol* **24**, 1312–9.