GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Co-evolution of genomic islands and their bacterial hosts revealed through phylogenetic analyses of 17 groups of homologous genomic islands

F.-B. Guo, W. Wei, X.L. Wang, H. Lin, H. Ding, J. Huang and N. Rao

School of Life Science and Technology,
University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: F.-B. Guo
E-mail: fbguo@uestc.edu.cn

Genet. Mol. Res. 11 (4): 3735-3743 (2012)
Received December 7, 2011
Accepted June 4, 2012
Published October 15, 2012
DOI http://dx.doi.org/10.4238/2012.October.15.5

ABSTRACT. Horizontal gene transfer is an important mechanism for the evolution of microbial genomes, and many horizontal gene transfer events are facilitated by genomic islands (GIs). Until now, few reports have provided evidence for the co-evolution of horizontally transferred genes and their hosts. We obtained 17 groups of homologous GIs, all of which appear in 8 or more bacterial strains of the same species or genus. Using phylogenetic analyses, we found that the topological structure of a distance tree based on the proteins of each group of homologous GIs was consistent with that based on the complete proteomes of the hosts. This result clearly indicates that GIs and their bacterial hosts have co-evolved. In addition to presenting and providing evidence for a novel concept, i.e., the co-evolution of GIs and their bacterial hosts, we also describe a new and interesting detail for the phylogenetic analysis of horizontally transferred genes: consistent phylogenetic trees can be obtained by focusing on homologous GIs despite the commonly accepted theory that the phylogenies of horizontally transferred sequences and host organisms should be inconsistent.

Key words: Co-evolution; Genomic island; Bacterial host; Horizontally transferred gene; Consistent phylogenetic trees

INTRODUCTION

Microbial genomes consist of core and accessory sequences. Core sequences have a fairly homogeneous G+C content and codon usage and encode housekeeping functions, whereas the accessory sequences differ from the rest of the genome in their G+C content and codon usage (Hacker and Kaper, 2000; Ochman et al., 2000; Vernikos and Parkhill, 2006). Accessory sequences are usually acquired via horizontal gene transfer (HGT) (Garcia-Vallvé et al., 2000, 2003; Nakamura et al., 2004). HGT is an important mechanism for the evolution of microbial genomes (Lawrence, 1999; Dobrindt et al., 2004). It contributes to the composition of the genomes, provides novel metabolic capabilities, and causes drastic changes in the ecological or pathogenic characters of bacterial species, thereby promoting microbial diversification and speciation (Lawrence, 1999; Gogarten and Townsend, 2005; Juhas et al., 2009).

Many HGT events are facilitated by genomic islands (GIs) (Hacker and Kaper, 2000). GIs are discrete DNA segments of a genome that show evidence of horizontal origins (Juhas et al., 2009). GIs appear in the genomes of most pathogenic and nonpathogenic bacteria (Dobrindt et al., 2004), and they usually encode accessory functions - such as additional metabolic activities, antibiotic resistance, symbiosis, pathogenesis, or properties involved in microbial fitness - that are not essential for bacterial growth but are advantageous under particular conditions (Hacker and Carniel, 2001; Hentschel and Hacker, 2001; Ho Sui et al., 2009). Most recently inserted GIs differ from core genomes with respect to G+C content and codon usage (Hacker and Kaper, 2000). The evolutionary advantage of GIs over small inserts is that a large number of genes can be transferred and incorporated as a whole into the recipient genome. The transfer may lead to dramatic changes in an organism and ultimately result in a quantum leap in evolution (Hentschel and Hacker, 2001).

As more and more microbial genomes have been sequenced, HGT events have been observed more frequently (Lawrence, 1999; Ochman et al., 2000; Garcia-Vallvé et al., 2000, 2003; Nakamura et al., 2004; Keeling and Palmer, 2008; Touzain et al., 2010). At the turn of this century, researchers believed that the extent of these events would cast doubt on the feasibility of constructing a "tree of life" (Pennisi, 1998; Wolf et al., 2002), because trees based on horizontally transferred genes and those based on genes with vertical inheritance are usually inconsistent (Doolittle, 1999). The concept of a universal "species" tree had not been considered appropriate until the appearance of numerous tree-constructing methods based on whole-genome comparisons (Wolf et al., 2001, 2002). The inconsistency between the phylogenies of horizontally transferred genes and an organism has been a gold standard for recognizing HGT (Keeling and Palmer, 2008; Touzain et al., 2010). However, herein we show that phylogenetic trees with consistent topological structures were obtained when homologous GIs were adopted to construct distance trees in 17 groups of bacterial genomes. Such consistency suggests co-evolution between GIs and their bacterial hosts. With the exception of the amelioration model proposed by Lawrence and Ochman (1997), our study is the first to report evidence of the co-evolution of horizontally transferred genes. Lawrence and Ochman (1997) found that the sequences of a gene cluster, spa, was ameliorated and had converged with that of the recipient enteric bacterial genome through mutation pressure. However, they did not adopt a phylogenetic analysis to show the amelioration. Herein, we present direct evidence for the co-evolution of GIs and their bacterial hosts through performing phylogenetic analysis.

MATERIAL AND METHODS

Known GIs and methods for identifying homologous GIs

All of the GIs selected were known GIs that have been thoroughly studied. They have been used frequently in published researchs and constitute fairly reliable data sets. From the Pathogenicity Island Database (http://www.gem.re.kr/paidb/) (Yoon et al., 2005, 2007), we obtained 63 of these GIs. In addition, 41 known GIs were obtained through an automatic PubMed search and subsequent manual check. In total, 104 well-documented GIs were prepared to perform a thorough analysis. DNA sequences for the bacterial hosts of these GIs were downloaded from the National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Two theoretical approaches can be used to identify GIs (Langille et al., 2008, 2010). One is based on sequence composition and the other is based on comparative genomics. The latter relies on the definition of HGT rather than on its outcome and, hence, has a lower rate of false positives. To achieve reliable results, all of the GIs used in our study were identified with the comparative genomics method.

Known GIs are often reported (or identified) only in a few bacterial strains. However, closely related bacterial strains contain GIs homologous with those reported. To find all of the homologues for a known GI, we performed a homology search for its core genes within the genomes of closely related bacterial strains. Homologous GIs are those with homologues of the core genes of the known GI. Homologous genes were determined using "bi-directional best hit" criteria. Boundaries (namely, integration sites) of newly identified GIs were determined by comparing the flanking genes of all homologous GIs. Consequently, 17 known GIs were found to have homologues in 8 or more bacterial strains of the same species or genus.

Phylogenetic methods

To check the co-evolution of GIs and their hosts, phylogenetic trees showing the evolutionary relationships among each group of bacterial strains were constructed based on homologous GIs and the complete genome of the hosts, respectively. The feature frequency profile (FFP) method was used to construct the trees. The FFP method, proposed by Sims et al. (2009a), is an alignment-free method in which feature (or l-mer) frequency profiles of whole proteomes are compared. The proteomes of organisms are stored as a collection of individual protein sequences. To use the method, we first calculated the frequency of the l-mer for each proteome corresponding to a chromosome or a chromosomal fragment. A sliding window of length l was run along a protein sequence with length L from position 1 to L - l + 1. The sliding window was moved to the next protein sequence and so on until the entire proteome was scanned. During the sliding process, the occurrence number for l-mer was counted. The counts were tabulated in the vector C_l for all possible features of length l:

$$C_l = <c_{l,1}, \ldots, c_{l,K}> \quad \text{(Equation 1)}$$

where K denotes the number of all possible features and is equal to 20^l.

The raw frequency counts were then normalized to form a probability distribution vector or FFP,

$$F_l = C_l \Big/ \sum_i c_{li} \qquad \text{(Equation 2)}$$

Obviously, the FFP determines the relative abundance of each l-mer. Thus, an organism is represented as an FFP of its proteome. The Jensen-Shannon divergence with FFPs was used to calculate distance (dissimilarity) between organisms.

Once a distance matrix was calculated, constructing phylogenetic trees using the BIONJ method was straightforward (Gascuel, 1997). Currently, the FFP method has been applied to the phylogenetic study of viruses (Wu et al., 2009), mammals (Sims et al., 2009b), and prokaryotes (Jun et al., 2010; Sims and Kim, 2011). We implemented the FFP method by writing a Python code. As suggested by the authors of FFP, an optimal feature length of 8 was used in this study. We used jackknife tests to estimate the statistical confidence of the resulting trees - that is, the phylogenetic tree was constructed by excluding one sample at a time.

Another alignment-free tree construction method, the composition vector (CV) (Qi et al., 2004), was also used to ensure the reliability of our results. Currently, the CV method has been applied to the phylogenetic study of viruses (Gao et al., 2003), chloroplasts (Chu et al., 2004), and prokaryotes (Qi et al., 2004). The CVTree server (Xu and Hao, 2009) that implements the method is available online (http://tlife.fudan.edu.cn/cvtree/), and we used its default settings. The trees generated using the 2 methods were displayed using Tree Explorer in Molecular Evolutionary Genetics Analysis 4 (Tamura et al., 2007).

RESULTS AND DISCUSSION

Consistent phylogenetic trees based on homologous GIs and complete proteomes of the hosts of 17 groups of bacterial strains

We selected known GIs that have homologues in 8 or more bacterial strains of the same species or genus. A total of 17 known GIs met this criterion. Details of the 17 groups of homologous GIs are listed in Table S1. We constructed phylogenetic trees for each group and compared them with trees constructed based on the complete genomes of the hosts. Carrying out phylogenetic analysis requires an approximate tree-constructing method. The method should construct trees based on the complete sequences of GIs or their hosts because investigating the evolutionary relationship between the overall GIs and the whole hosts is necessary. The FFP (Sims et al., 2009a) and CV methods (Qi et al., 2004) satisfy these requirements and are based on complete proteomes of organisms or adequately large collections of protein sequences.

The FFP method was used to construct the phylogenetic trees based on each group of homologous GIs and their hosts, respectively. Then, 2 types of trees were thoroughly compared. In the species *Staphylococcus aureus* and its GI vSaα, for example, 15 strains are found to have vSaα homologous GIs. After comparison, a sub-tree of 11 strains based on GIs was found to have a topological structure consistent with that based on the complete proteomes of the hosts. The 4 exceptional strains were MRSA252, Newman, RF122, and ED98. The 2 types

of phylogenetic trees were inconsistent when these exceptional strains were merged into them.

The GI tree and the host tree for the 11 strains of *S. aureus* are shown in Figure 1. They have the same topological structure, which means that the 2 trees reflect the same evolutionary relationship for the 11 strains. In other words, the GI tree also reflects the phylogenetic relationship of the hosts. However, these homologous GIs have the same source or donor - that is, they contained the same sequences at the time of insertion. After a period of time, they evolved into different sequences, and interestingly, the resulting protein sequences of the GIs contain information from the host. Therefore, we reasonably concluded that GIs vSaα and their host genomes co-evolved.

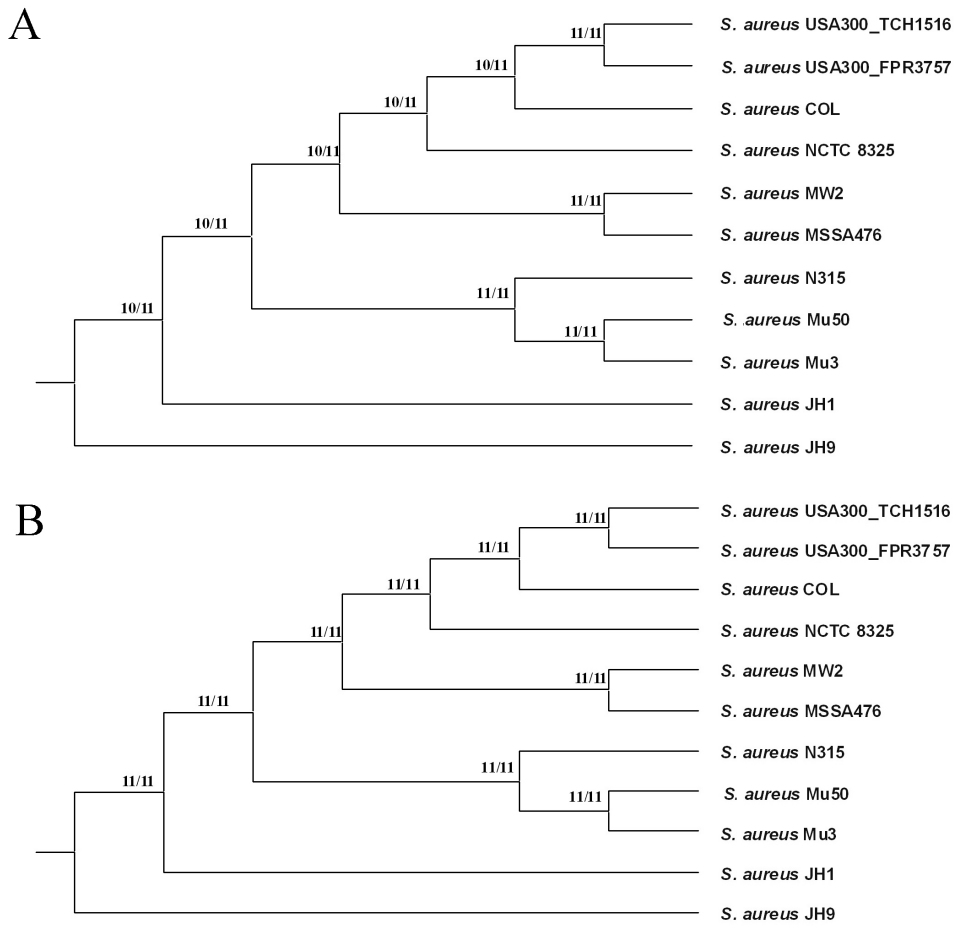

Figure 1. Trees illustrating the phylogenetic relationship among 12 *Staphylococcus aureus* strains. The two trees are displayed as rectangular cladogram by using Tree Explorer with MEGA4. **A.** The evolutionary tree is constructed based on proteins contained in vSaα and its homologous GIs. **B.** The tree is constructed based on the complete proteomes of the hosts. Marked fraction shows the appearance frequency of each branch in the jackknife test. Branch lengths are not to scale so that the clade and tree topology can be clearly displayed. As can be seen, topological structures of the two trees are consistent.

Phylogenetic trees of the other 16 groups of homologous GIs and their hosts were similarly compared and are provided in Figure S1. The number of strains with homologous GIs and the maximum number of the strains with consistent sub-trees are listed in Table 1. All 17 groups had 7 or more strains with consistent topological structures of phylogenetic trees. Therefore, the percentages of strains with consistent trees are all equal to or larger than 50%, which indicates that the 17 groups of homologous GIs and their hosts co-evolved. Therefore, the co-evolution between GIs and their bacterial hosts may be a universal phenomenon.

Table 1. Maximum number of strains that have consistent phylogenetic trees constructed by the feature frequency profile method for each of the 17 groups of samples.

GI	Host	Number of strains with consistent trees in each random groups										Average number of strains with consistent trees in random groups	Number of GIs with consistent trees[a]	Number of hosts with homologous GIs
FPI	*Francisella*	8	7	5	8	8	8	8	7	8	8	8	8 (89%)	9
HPI	*Yersinia*	7	7	8	7	7	8	7	7	7	7	7	8 (89%)	9
LEE	*E. coli*	9	9	8	9	7	8	8	8	8	9	8	7 (78%)	9
PPI-1	*S. pneumoniae*	7	7	7	6	6	7	6	7	7	7	7	7 (64%)	11
SPI-1	*S. enterica*	9	10	11	11	9	10	10	12	9	9	10	10 (63%)	16
SPI-2		10	10	9	9	11	10	11	9	10	9	10	10 (63%)	16
SPI-3		10	9	8	8	10	8	11	8	7	8	9	8 (53%)	15
SPI-4		10	8	11	8	12	9	8	9	10	10	10	13 (81%)	16
SPI-5		9	9	8	10	11	10	9	9	8	8	9	8 (50%)	16
SPI-6		7	9	7	8	7	8	8	9	8	7	8	9 (82%)	11
SPI-9		7	6	9	9	11	7	10	8	10	8	9	9 (56%)	16
SPI-11		12	6	8	9	10	8	10	9	8	7	9	10 (67%)	15
SPI-12		10	8	8	10	9	9	8	11	10	8	9	8 (50%)	16
SPI-16		9	9	8	10	10	8	10	9	8	9	9	9 (60%)	15
vSaα	*S. aureus*	11	11	12	11	13	10	12	11	11	12	11	11 (73%)	15
vSaβ		11	10	11	12	12	10	13	10	10	11	11	10 (67%)	15
vSaγ		10	11	11	11	11	12	12	10	11	10	11	11 (73%)	15

[a]The percentages in parentheses are obtained by dividing the number of genomic islands (GIs) with consistent trees by the total number of strain hosts with the homologous GIs.

Phylogenetic trees constructed using the CV method were also compared, and the results are listed in Table 2. The co-evolutionary relationship between the homologous GIs and their hosts was illustrated in this analysis.

Table 2. Maximum number of strains that have consistent phylogenetic trees constructed by the CVTree method for each of the 17 groups of samples.

GI	FPI	HPI	LEE	PPI-1	SPI-1	SPI-2	SPI-3	SPI-4	SPI-5
Number of GIs with consistent trees[a]	7 (78%)	7 (78%)	6 (67%)	8 (73%)	10 (63%)	10 (63%)	7 (47%)	12 (75%)	12 (75%)
GI	SPI-6	SPI-9	SPI-11	SPI-12	SPI-16	vSaα	vSaβ	vSaγ	
Number of GIs with consistent trees[a]	8 (73%)	9 (56%)	11 (73%)	8 (50%)	9 (60%)	12 (80%)	10 (67%)	11 (73%)	

[a]The percentages in parentheses are obtained by dividing the number of genomic islands (GIs) with consistent trees by the total number of strain hosts with the homologous GIs.

Explanation of the consistency in the trees of most strains

Homologous GIs have the same source or donor and, hence, contained the same sequences at the time of insertion. Over time, they evolved into various sequences, and the dis-

crepancies can be used to resemble the phylogenetic relationship of their hosts. The dynamic evolution process of the GIs is called co-evolution. Lawrence and Ochman (1997) have analyzed the post-introgression evolution process of horizontally transferred genes through a process called amelioration. According to this process, horizontally transferred genes reflect the base composition of the donor genome at the time of insertion, but over time, these sequences may ameliorate to reflect the DNA composition of the new genome. This change occurs because the introgressed genes are subject to the same mutational processes that affect all genes in the recipient genome. Lawrence and Ochman (1997) proposed a model to describe the amelioration process by assuming that the rate of amelioration could be expressed as a function of the rates of evolutionary change or substitution rates. Using this model, they simulated the process of amelioration for a gene cluster, spa, of enteric bacteria. Over a long period of such evolution, the composition character of the donor fades as the character of the acceptor grows.

Our study illustrated the current status of GIs, which are a larger collection of horizontally transferred genes, using a phylogenetic method. Although the phylogenetic history of the donor in itself could not be erased from the homologous GIs, the current status (particularly composition information) of the GIs could be reliably used to infer the phylogenetic relationships among their new hosts. Therefore, phylogenetic trees consistent with those of the hosts are obtained when composition-based methods - namely FFP and CV - are used to construct the trees. We do not use the term amelioration because evolutionary patterns other than directional evolution or amelioration may exist during the post-introgression period. For example, GIs and their hosts may evolve concurrently owing to new genome-scale evolution pressures.

A commonly accepted expectation is that phylogenetic trees based on horizontally transferred genes will differ significantly from those based on important and conserved genes or complete genomes (Doolittle, 1999; Pennisi, 1998; Wolf et al., 2002; Keeling and Palmer, 2008). This difference occurs because horizontally transferred genes and conserved genes of a certain species are from different sources. However, our study showed that phylogenetic trees consistent with those of hosts can be obtained when homologous GIs are used to construct the tree. Therefore, we provide a new and interesting detail for the phylogenetic analysis of horizontally transferred genes.

Interpretation of the appearance of outliers for consistent trees

Although the 2 kinds of sub-trees (GI tree and host tree) for most of the bacterial strains under study had consistent topological structures, some outlier strains were present among each group of samples. To understand the appearance of the outliers, we constructed another kind of phylogenetic tree using the FFP method. However, instead of using proteins contained in homologous GIs, we chose random proteins. For a known GI, the same number of proteins was randomly chosen from the proteome of the host. Then, homologous proteins were extracted in the other hosts in the group. We choose only proteins conserved in specific outgroups and hosts. Therefore, the chosen proteins were likely to be native proteins and, hence, representative of the phylogenetic relationship of their hosts to the greatest extent. The maximum numbers of bacterial strains with consistent sub-trees among the whole proteomes and randomly chosen proteins are also listed in Table 1. To avoid sampling bias, we repeated the selection procedure 10 times and the maximum consistent numbers for all 10 comparisons are listed in Table 1.

Similar to the outcome in the GI tree, the maximum consistent number of "randomly chosen protein trees" is also lower than the number of strains with homologous GIs. For example, in *S. aureus* and its GI vSaα, the maximum consistent number between the GI tree and the whole proteome tree is 11. Comparatively, the maximum consistent number between the randomly chosen protein tree and the whole proteome tree is also 11. Therefore, GI vSaα may resemble the phylogenetic relationship to a similar extent, with the native fragment having the same number of proteins. We could not demand that the GIs be more effective than the native proteins to resemble phylogenetic relationship, so it is easy to understand why a few outliers appear in Figure 1. As shown in Table 1, the maximum consistent numbers between the 2 kinds of trees (GI trees and randomly chosen protein trees) are also similar in the other 16 groups sampled. Therefore, the reason for the appearance of outliers in the 2 kinds of trees (GI tree and the randomly chosen protein tree) may be the same: the number of contained proteins is so small that it could not act as an appropriate representative of the hosts.

ACKNOWLEDGMENTS

We thank Dr. Huixiong Zhang for carefully reading the manuscript. Research supported by the National Natural Science Foundation of China (#31071109), the special fund from the China Postdoctoral Science Foundation (#201104687), and the Program for New Century Excellent Talents in University (#NCET-11-0059).

Supplementary material

REFERENCES

Chu KH, Qi J, Yu Z-G and Anh V (2004). Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol.* 21: 200-206.

Dobrindt U, Hochhut B, Hentschel U and Hacker J (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* 2: 414-424.

Doolittle WF (1999). Phylogenetic classification and the universal tree. *Science* 284: 2124-2129.

Gao L, Qi J, Wei H, Sun Y, et al. (2003). Molecular phylogeny of coronaviruses including human SARS-CoV. *Chin. Sci. Bull.* 48: 1170-1174.

Garcia-Vallvé S, Romeu A and Palau J (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* 10: 1719-1725.

Garcia-Vallvé S, Guzman E, Montero MA and Romeu A (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.* 31: 187-189.

Gascuel O (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14: 685-695.

Gogarten JP and Townsend JP (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3: 679-687.

Hacker J and Kaper JB (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* 54: 641-679.

Hacker J and Carniel E (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* 2: 376-381.

Hentschel U and Hacker J (2001). Pathogenicity islands: the tip of the iceberg. *Microbes Infect.* 3: 545-548.

Ho Sui SJ, Fedynak A, Hsiao WW, Langille MG, et al. (2009). The association of virulence factors with genomic islands. *PLoS One* 4: e8094.

Juhas M, van der Meer JR, Gaillard M, Harding RM, et al. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* 33: 376-393.

Jun SR, Sims GE, Wu GA and Kim SH (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. U. S. A.* 107: 133-138.

Keeling PJ and Palmer JD (2008). Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9: 605-618.
Langille MG, Hsiao WW and Brinkman FS (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 9: 329.
Langille MG, Hsiao WW and Brinkman FS (2010). Detecting genomic islands using bioinformatics approaches. *Nat. Rev. Microbiol.* 8: 373-382.
Lawrence JG (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol.* 2: 519-523.
Lawrence JG and Ochman H (1997). Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44: 383-397.
Nakamura Y, Itoh T, Matsuda H and Gojobori T (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* 36: 760-766.
Ochman H, Lawrence JG and Groisman EA (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405: 299-304.
Pennisi E (1998). Genome data shake tree of life. *Science* 280: 672-674.
Qi J, Wang B and Hao BI (2004). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.* 58: 1-11.
Sims GE and Kim SH (2011). Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U. S. A.* 108: 8329-8334.
Sims GE, Jun SR, Wu GA and Kim SH (2009a). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U. S. A.* 106: 2677-2682.
Sims GE, Jun SR, Wu GA and Kim SH (2009b). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc. Natl. Acad. Sci. U. S. A.* 106: 17077-17082.
Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.
Touzain F, Denamur E, Medigue C, Barbe V, et al. (2010). Small variable segments constitute a major type of diversity of bacterial genomes at the species level. *Genome Biol.* 11: R45.
Vernikos GS and Parkhill J (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* 22: 2196-2203.
Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, et al. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* 1: 8.
Wolf YI, Rogozin IB, Grishin NV and Koonin EV (2002). Genome trees and the tree of life. *Trends Genet.* 18: 472-479.
Wu GA, Jun SR, Sims GE and Kim SH (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc. Natl. Acad. Sci. U. S. A.* 106: 12826-12831.
Xu Z and Hao B (2009). CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res.* 37: W174-W178.
Yoon SH, Hur CG, Kang HY, Kim YH, et al. (2005). A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics* 6: 184.
Yoon SH, Park YK, Lee S, Choi D, et al. (2007). Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res.* 35: D395-D400.