Chapter 11
Bioinformatics Approaches to Deciphering Alien Gene Transfer: A Comprehensive Analysis

Rajeev K. Azad, Nitish Mishra, Firoz Ahmed, and Rakesh Kaundal

Abstract A large number of bioinformatics methods have been developed in recent years for detecting gene transfers between distantly related or unrelated organisms. These have been mainly classified as parametric and phylogenetic methods. While the former methods have been frequently invoked for detecting recent gene transfers, detection of ancient gene transfers have relied upon phylogenetic methods. Numerous evidences emerging from the applications of these methods have firmly established interspecies gene transfer as a significant force-driving prokaryotic genome evolution. The focus is now shifting to assessing the extent and impact of this mechanism in eukaryotic genome evolution. The methods developed for detecting alien genes in unicellular organisms have been adapted for identifying and cataloging instances of gene transfers in multicellular organisms. A significant interest is in cataloging gene transfers in plants which have more leaky barriers to gene transfer than highly evolved animals. We review the advances in this field with a focus on alien gene transfer in plants and the bioinformatics methods frequently used to detect such transfers.

Keywords HGT quantification • Bayesian method • Bootstrapping • Phylogenetic tree • Parametric methods • UPGMA

R.K. Azad
Departments of Biological Sciences and Mathematics,
University of North Texas, Denton, TX 76203, USA

N. Mishra • F. Ahmed
The Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA

R. Kaundal, Ph.D. (✉)
Department of Biochemistry & Molecular Biology, National Institute
for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB), Oklahoma State
University, 246 Noble Research Center, Stillwater, OK 74078, USA
e-mail: r.kaundal@okstate.edu

A. Pratap and J. Kumar (eds.), *Alien Gene Transfer in Crop Plants, Volume 1: Innovations, Methods and Risk Assessment*, DOI 10.1007/978-1-4614-8585-8_11,

11.1 Introduction

Classical genetics has traditionally focused on vertical gene transfer that has helped shape the “tree thinking” in explaining the evolution of extant or extinct organisms. Advances in genome era have brought a change in this thinking, triggered by plethora of compelling evidences emerging in support of horizontal genetic inheritance, particularly in the prokaryotic domain (Ochman et al. 2000; Koonin et al. 2001; Gogarten and Townsend 2005). Horizontal Gene Transfer (HGT), also referred to as lateral gene transfer, is the transfer of genetic material between organisms by means other than parent-to-offspring (vertical) inheritance (Syvanen and Kado 1998; Ochman et al. 2000; Koonin et al. 2001; Gogarten and Townsend 2005; Keeling and Palmer 2008). While HGT is now recognized as a potent force-driving prokaryotic genome evolution, a relatively better sampling of eukaryotic genomes now available as a consequence of DNA sequencing revolution has necessitated a reassessment of the extent and impact of HGT in eukaryotic genome evolution. Numerous instances of eukaryotic HGT events reported in recent years have further galvanized this field, bringing the spotlight on gene flow among eukaryotes (Andersson 2005; Keeling and Palmer 2008; Sanchez 2011).

The evolutionary history of plant genomes is also replete with intracellular gene transfer (IGT)—the transfer of genes between organelles within a plant cell (Keeling and Palmer 2008; Bock 2010). Single to multiple instances of HGTs involving plants have been the subject of numerous recent studies and have been reviewed by several authors. Plants have served as both recipients and donors of alien genes (see Richardson and Palmer 2007; Keeling and Palmer 2008; Bock 2010 for comprehensive reviews on HGTs in plants). However, a comprehensive treatise is lacking on the methods for detecting HGTs in plants. This review is intended to provide the plant community an overview of the methods and protocols for detecting HGTs in plants. In what follows, we briefly narrate the case studies of plant HGT as reported in recent articles and reviews (Bock 2010; Keeling and Palmer 2008) and follow this up with an elaborate description of the methodology for detecting HGTs in plants.

Plants have also integrated genomes of viruses which often act as carriers for foreign DNAs. Tobacco plants have been found to have Gemini viral DNAs in their nuclear genomes (Bejarano et al. 1996). Interestingly, evidences exist even of the transfer of viral RNA sequences into plant genomes: viral sequences likely originating from closteroviruses were found in the mitochondrial genome of grape (Goremykin et al. 2009); the host’s reverse transcriptase is likely to have transcribed viral RNAs to cDNAs thus facilitating their integration into the host genome.

Ralph Bock and colleagues recently designed a genetic screen to demonstrate plant to plant HGT (Stegemann and Bock 2009). Genetic engineering is a classic example of man-made HGT. Initially thought to be very rare, advances in genome sequencing and development, in parallel, of more sophisticated phylogenetic methods have helped elucidate numerous instances of natural plant–plant HGT. In most cases, evidences appear to support cell to cell contact as a mechanism of transfer of genetic material; this has led to hypothesize that plant parasitism and natural grafting are the major factors in plant–plant HGT (Bock 2010). Plant parasites are known to be both the recipients and donors of foreign DNAs mobilized via cell to cell contact. However, the importance of other mechanisms such as transformation (uptake of naked DNA), illegitimate pollination, and vector-mediated transfers may have been understated; this needs to be reassessed in light of new genomic data emerging from plant sequencing projects.

Perhaps due to the ability of mitochondria to fuse and recombine, mitochondrion–mitochondrion HGTs are much more prevalent than plastid-initiated transfers. Unlike chloroplasts, plant mitochondria contain active DNA uptake system (Koulintchenko et al. 2003; Logan 2006). Of the few cases of alien gene transfer involving chloroplast genomes, it has been argued that these transfers may actually be mediated by mitochondria and less likely be de novo chloroplast HGTs. A more plausible explanation for the presence of chloroplast *pvs-trnA* genic sequence in the mitochondrial genome of *Phaseolus* is the IGT of this sequence from donor's chloroplast to its mitochondrion followed by mitochondrion to mitochondrion HGT (Woloszynska et al. 2004). The transfer of whole chloroplast genome by performing grafting experiments involving *Nicotiana tabacum* (donor), *Nicotiana glauca* (recipient), and *Nicotiana benthamiana* (recipient) has been demonstrated recently (Stegemann et al. 2012). This study thus provides a strong case for natural grafting as a possible mechanism for chloroplast transfer among plant species.

Nuclear genes are also not immune to plant–plant HGT. HGT of a transposon, MULE (Mu-like elements), involving nuclear genomes of *Setaria* and *Oryza*, could be an example of vector-mediated nuclear HGT (Diao et al. 2006). A just published study on the evolution of C_4 photosynthesis trait in the grass lineage *Alloteropsis* implicates plant–plant nuclear HGT involving donors from the C_4 lineage that diverged from *Alloteropsis* more than 20 million years ago (Christin et al. 2012).

The horizontal acquisition of alien DNAs is not restricted to a single gene or multiple genes but may even involve fragments of genes. A few cases of horizontal intron transfer in plants have been reported: *Peperomia polybotrya*, a basal angiosperm, has integrated an intron from a fungal donor into its mitochondrial *cox1* gene (Vaughn et al. 1995); another example is a self-splicing intron likely originating from a cyanobacterium found in the *psbA* gene of the alga *Euglena myxocylindracea* (Sheveleva and Hallick 2004). Won and Renner provided an striking example of plant to plant horizontal intron transfer: the intron 2 belonging to group II introns along with its flanking exons from the mitochondrial gene *nad1*of an asteroid (angiosperm) was transferred to Gnetum (gymnosperms) 2–5 million year ago (Won and Renner 2003). An interesting case is of the *rps11* gene in the mitochondrial genome of *Sanguinaria*, an eudicot; this gene has a chimeric structure with its 3′ half acquired from a monocot (Bergthorsson et al. 2003; Richardson and Palmer 2007).

11.2 Mechanisms of HGT

Although cell to cell contact has been the most cited mechanism of gene transfer in plants, the contributions of other mechanisms including transformation and transduction might have remained underestimated. Plants can acquire alien DNAs via all the three basic mechanisms reported for gene transfer among prokaryotes (Ochman et al. 2000).

11.2.1 Transformation

Through this mechanism, a recipient cell can take in naked DNA directly from the environment. Although a common mechanism for gene transfer among bacteria, this is less common among eukaryotes. Short DNA fragments can be readily transferred using this mechanism.

11.2.2 Conjugation

Conjugation requires the physical contact of donor and recipient cells and the transfer is mediated through plasmids. This process can facilitate transfer of genetic material between distantly related organisms, and by its very nature, conjugation can move large fragments of DNAs.

11.2.3 Transduction

In transduction, the transfer of genetic material is mediated through bacteriophages which package alien DNAs from a donor cell and inject it into a recipient cell during infection. The amount of transferred DNAs is limited by the size of phage.

However, there are several barriers to HGT, which help to protect the recipient organism from deleterious effects by maintaining the integrity of the host genome (Kurland et al. 2003; Kurland 2005; Thomas and Nielsen 2005). These barriers include physiological state of donor and recipient cell, adaptability of the incoming DNA into a recipient cell, surface exclusion for the plasmid-mediated transfers, cleavage of foreign DNA by recipient's restriction system, hindrance to plasmid replication within recipient cell, successful integration into host genome, and the likelihood of acquired gene's expression within the recipient system. An understanding of these barriers will help advance the field of genetic engineering, the artificial counterpart of natural HGT, which has become an important tool to secure a desired phenotype by augmenting the physiological repertoire of an organism through gene transfer.

11.3 Quantifying HGT

The prevalence and significance of HGT has necessitated the development of novel methodologies for robust quantification of horizontal gene flow. Detection of HGT is often confounded by many factors and no single method is capable of addressing this problem. Therefore, several complementary approaches have been proposed, and a combination of disparate approaches appears to address the detection of HGT more convincingly (Azad and Lawrence 2012). The extent and impact of HGT in plants have not been realized until recently, mainly due to lack of sequenced genomes of close relatives of a species of interest, and also because of the limitation of experimental methods frequently invoked by plant biologists in cataloging gene transfer events. Post genome sequencing revolution, detection of alien genes has come to rely upon computational methods which can assess, on a genome-wide scale, the extent and consequence of HGT in plant evolution. Several computational methods have been developed to detect horizontally transferred genes, which can be categorized into two types: phylogenetic methods and parametric methods (sometimes also called composition based or surrogate methods) (Azad and Lawrence 2012). While the former methods have almost always been invoked in detecting alien genes in plants, the latter methods have not yet been seriously explored for assessing gene transfer among eukaryotes. We discuss below the principles underlying both approaches, and the different questions or hypothesis they test to infer alien genes in a given genome.

11.4 Phylogenetic Methods for Alien Gene Detection

This class of methods is focused on detecting aberrant phylogenetic patterns, that is, the gene relationships that differ significantly from the canonical organismal phylogeny (Beiko and Hamilton 2006; Poptsova 2009). Phylogenetic methods, as the name suggests, infer relationships by constructing phylogenetic trees based on complex morphological features or nucleotide sequences of genes. This is perhaps the most commonly used approach for detecting HGT in eukaryotes including plants (Keeling and Palmer 2008). HGT is primarily inferred by detecting discrepancies in the phylogenetic tree of orthologous genes when compared to species tree which represents the overall phylogenetic relationships among all considered species. The requirement of presence of homologues of a gene in *all* genomes of interest limits the applicability of phylogenetic tree-based methods; complementary phylogenetic methods that do not explicitly require building trees in order to infer alien genes have also been developed. We summarize below the frequently invoked phylogenetic approaches for alien gene detection.

11.4.1 Phylogenetic Tree

To construct a phylogenetic tree representing relationships among organisms, highly conserved molecular sequences of DNA, RNA, or protein molecules that have evolved slowly yet engendered subtle differences to reliably compare taxa over large evolutionary distances have been used. A frequently used phylogenetic marker, initially proposed by Woese and colleagues, is the nucleotide sequence of 16S small subunit ribosomal RNA gene which has primarily been relied upon for inferring organismal phylogeny (Woese et al. 1990; Woese 1991; Olsen and Woese 1993). However, organismal phylogenies inferred from other conserved sequences differ among themselves and from ribosomal RNA phylogeny (Hilario and Gogarten 1993; Brown et al. 1994; Gogarten 1995; Nesbo et al. 2001; Poptsova 2009). This has led to developing other strategies for extracting a reliable species or organismal tree from molecular sequence data. One approach is to find a consensus from orthologous gene trees. Variants of consensus methods include strict consensus, majority-rule consensus (Day and McMorris 1992; Dong et al. 2010), Adams consensus (Adams 1972), and super tree consensus methods (Bininda-Emonds and Sanderson 2001; Eulenstein et al. 2004; Bininda-Emonds 2005; Nguyen et al. 2012; Swenson et al. 2012). This is based on the premise that a majority of genes are acquired vertically and therefore the phylogenetic signal representing vertical inheritance can be reconciled to an acceptable degree of confidence from the orthologous gene trees. Another approach to infer species tree is based on concatenation of orthologous gene alignments (Wolf et al. 2002), referred to as super matrix approach (Lapierre et al. 2012). Both super tree and super matrix methods are used frequently. A recent study used genome simulations to assess the accuracy of these methods in recovering species tree when subjected to HGT (Lapierre et al. 2012). The methods were found sensitive to the amount of HGT. The super matrix approach performed better for low amount of HGT, while the super tree approach was more accurate for moderate amount of HGT. Any prior information on the frequencies of HGT in the evolution of organisms of interest could thus help in selecting the most appropriate method. The species tree thus obtained represents the null hypothesis that there was no HGT in the history of orthologous genes. If a gene tree deviates significantly from species tree, this indicates an HGT in the history of this gene. One major advantage of this approach is that the likely scenarios of horizontal gene flow are assessed directly and the direction of gene flow determined unambiguously, thus identifying the recipient and donor organisms involved in gene transfer. Because of this attribute, phylogenetic tree methods have been often invoked to infer roadmap of gene transfers. There are five steps to phylogenetic tree construction:

11.4.1.1 Identify Orthologues of a Gene of Interest

Identification of homologues of a gene diverging following speciation events, namely, the orthologues, is the first step in phylogenetic gene tree construction.

Given a set of genes, one can use all against all BLAST similarity search (Altschul et al. 1990) to identify reciprocal best hits within the set followed by elimination of paralogous genes (homologous as a consequence of gene duplication). There are databases of orthologous genes that one can also use such as Clusters of Orthologous Groups (COGs) (Tatusov et al. 2000), OrthoMCL-DB (Chen et al. 2006), and MultiParanoid (Alexeyenko et al. 2006). NCBI's HomoloGene is a useful repository (http://www.ncbi.nlm.nih.gov/homologene) for eukaryotic orthologues and paralogues.

11.4.1.2 Perform a Multiple Sequence Alignment of Gene Orthologues

Dynamic programming methods as well as heuristic methods have been developed for multiple sequence alignment of members of gene or protein families. Programs based on progressive alignment methods such as ClustalW (Thompson et al. 2002) and MUSCLE (Edgar 2004) use a guide tree to perform multiple sequence alignment, progressively assembling most similar pair of sequences into a multiple alignment. Iterative refinement methods refine the progressive alignment by recursively aligning a sequence to the rest of the sequences in the progressive alignment. This is repeated for each sequence in the alignment or until the convergence of the alignment score (Durbin et al. 1998). Popular programs implementing iterative refinement include (Katoh et al. 2009), INTERALIGN (Pible et al. 2005) and PRALINE (Simossis and Heringa 2003, 2005). Probabilistic models, namely, the profile hidden Markov models, have been used in the consistency-based methods to achieve greater accuracy in alignment (e.g., the ProbCons program) (Do et al. 2005).

11.4.1.3 Select an Evolutionary Model of Nucleotide/Amino Acid Substitution

A simple approach to measure differences between two sequences in an alignment is to count the alignment positions where the residues (nucleotides or amino acids) differ and divide this difference by the alignment length. More sophisticated substitution models include the Jukes-Cantor model and Kimura 2- or 3-parameter model (Durbin et al. 1998).

11.4.1.4 Use One of the Tree Construction Methods

The five tree construction methods are classified as distance-based methods (UPGMA and neighbor-joining), character-based methods (maximum parsimony), and model-based methods (maximum likelihood and Bayesian) (Durbin et al. 1998; Pevsner 2003). Distance-based methods use a distance measure to perform pairwise comparison of DNA or protein sequences. This way two sequences with least nucleotide or amino acid changes observed in their alignment form the first two sister branches of the phylogenetic tree, joining at a node representing their common ancestor. This process is repeated recursively to generate other branches and ancestral nodes of the tree. In contrast, character-based methods process the information within multiple sequence alignment all at once; maximum parsimony approach accomplishes this by evaluating the likely scenarios in evolution giving rise to variations in characters (nucleotides or amino acids) at the informative sites of multiple sequence alignment. The tree postulating relationship among given taxa with minimal number of character variations or mutations is the most parsimonious explanation of relationship among taxa and is therefore considered the optimal tree given the sequence data. The maximum likelihood methods are based on the premise that the most likely tree representing the given data is the one that maximizes the likelihood of generating the observed data. Here, all possible trees with different topologies and branch lengths are explored in order to find the optimal tree representing the evolutionary history of the given sequence data. Unlike maximum parsimony which requires counting of nucleotide or amino acid substitutions, maximum likelihood associates probability to each evolutionary event and so requires specifying probabilistic evolutionary models. Bayesian methods are similar in spirit to the maximum likelihood methods, searching for most probable tree given the data; however, the optimal tree is now inferred from the posterior distribution of trees computed via Markov Chain Monte Carlo (MCMC) (Gelman and Rubin 1996) simulations. Bayesian methods add the flexibility to incorporate prior information about the model (tree parameters, etc.). The above approaches have been implemented in different software programs such as PHYLIP (distance based, maximum parsimony, maximum likelihood) (Felsenstein 1989), PAUP (maximum parsimony) (Swofford 1998), TREE-PUZZLE (maximum likelihood) (Schmidt et al. 2002; Schmidt and von Haeseler 2007), and MrBayes (Bayesian) (Huelsenbeck and Ronquist 2001).

11.4.1.5 Evaluate Trees Using Bootstrapping

Bootstrapping methods are used to assess confidence over the branching patterns of a tree topology (Efron et al. 1996; Durbin et al. 1998). Each node with bifurcating or multi-furcating branches is given a confidence score as follows. Columns from a multiple sequence alignment are selected randomly and with replacement in order to construct a random replicate of the original alignment. Confidence on a clade in a tree is obtained as the proportion of times that clade appears in the random replicates of the tree.

The next step in this sequence of protocols is to assess the gene tree against the background (organismal) tree. Likelihood-based methods such as Shimodaira-Hasegawa (S-H) (Shimodaira and Hasegawa 1999), Kishino-Hasegawa (K-H) (Kishino and Hasegawa 1989), and Approximately Unbiased (AU) tests (Shimodaira 2002) are frequently used for this purpose. These tests allow testing the null hypothesis that a gene tree is similar to the organismal tree; if the p-value for the likelihood statistics is less than a significance threshold (typically 0.05 or less), the null hypothesis is rejected thus inferring HGT in the evolutionary history of the gene. The other approach is to compute Robinson-Foulds (R-F) distance (Robinson and Foulds 1981) between gene tree and species tree, which is essentially the minimum number of operations required to transform a gene tree into a species tree. Assuming most genes in an organism to have been vertically inherited, a significant deviation from the mean of R-F distances between gene trees and species tree is an indicator of HGT. Similar in spirit to R-F distance is the sub-tree prune-and-graft (SPR) distance (Swofford and Olsen 1990), which is equivalent to minimum number of rearrangements required to change the topology of a gene tree to that of the species tree.

A nontrivial issue in alien gene detection is the fidelity of the phylogenetic methods. To assess the phylogenetic methods, one must have orthologous gene sets with no history of HGT (the "null" datasets for estimating the false-positive rate) and orthologous gene sets with history of HGT (for estimating false-negative rate). Since evolutionary events are often difficult to validate, alternative approaches have been developed to construct test datasets. One approach is based on absolute consensus; if none of the phylogenetic methods find a support for HGT in the history of orthologous genes, the set of such genes defines the "backbone" signifying vertical inheritance. Gene transfers could be simulated within the same dataset to construct a set of genes with one or more HGT events happening in the course of their evolution. The power of a phylogenetic method could thus be assessed on these datasets. The other approach is to simulate species evolution. Evolsimulator (Beiko and Charlebois 2007) starts with a set of genes in an ancestral genome which is evolved through speciation and other evolutionary processes sans the HGT. This gives sets of orthologous genes that have evolved vertically and therefore could be used for estimating the false-positive rate. One can also simulate HGT in the history of orthologous genes and this data could be used for estimating false-negative rate.

Keeling and Palmer (2008) elucidated six likely scenarios of gene transfer which include (1) duplicative transfer, where the recipient genome retains both the horizontally acquired and original copies of a homologous gene, (2) recent homologous replacement where a gene transfer event between extant organisms results in replacement of the recipient's gene by a homologous copy from a distantly related donor, (3) ancient homologous replacement where the homologous gene replacement involves ancestors of different lineages, (4) duplicative transfer with differential loss where the lineage-specific gene losses follow the gene transfer event, (5) sequential transfer where the same gene gets transferred more than once to different lineages, and (6) new gene transfer where a gene of recent origin in a lineage gets transferred to another lineage with no history of this gene via illegitimate recombination (Fig. 11.1). Phylogenetic methods are thus subjected to different sets of challenges arising from different scenarios of gene transfer, and the differential gene loss, in particular, has a deeper confounding effect on deciphering HGT.

In addition to lineage-specific gene loss, other confounding factors in HGT detection via tree building include biased mutation rates, improper clade selection, long branch length attraction, and segregation of paralogues (Kurland et al. 2003; Kurland 2005). Further, the phylogenetic HGT prediction is only as good as the consensus organismal tree which is hard to reconcile despite recent advances.

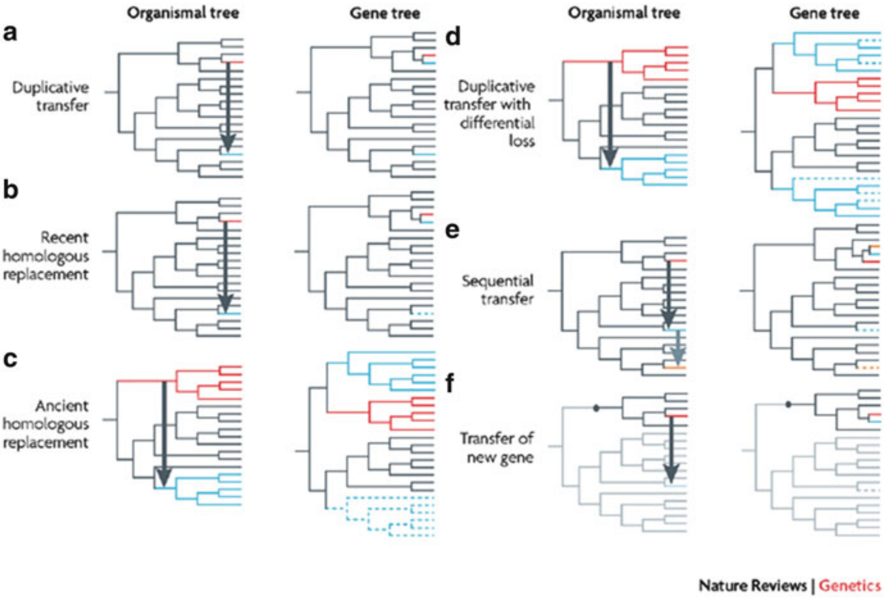
Organismal tree
Gene tree
Organismal tree
Gene tree
a
Duplicative transfer
b
Recent homologous replacement
c
Ancient homologous replacement
d
Duplicative transfer with differential loss
e
Sequential transfer
f
Transfer of new gene
Nature Reviews | Genetics

Fig. 11.1 Incongruent gene phylogenies as a consequence of different kinds of gene transfers (Keeling and Palmer 2008; Reprinted by permission from Nature Publishing Group: Nature Reviews Genetics)

Perhaps one of the biggest bottlenecks is determining the phylogeny of “orphan” genes (those lacking homologues in the database). It is plausible that many orphan genes might have arrived horizontally; however, due to the absence of their orthologues, phylogenetic methods cannot be applied to detect orphan gene transfer. Despite these shortcomings, phylogenetic methods are considered most reliable in inferring ancient gene transfer.

11.4.2 Unusual Phyletic Pattern

Phylogenetic tree methods may lead to confounding interpretation as discussed above. At the same time, comparative genomics provide alternative routes to avoid the vagaries of the tree methods. This approach examines the genomes of closely related organisms for the presence of unusual phyletic pattern (Lander et al. 2001; Gophna et al. 2006; Vernikos and Parkhill 2008; Arvey et al. 2009). If a gene is present in the genome of an organism but absent in the genomes of closely related organisms, it is likely to have been acquired horizontally. This approach is now receiving greater acceptance due to more reliable sampling of closely related genomes.

However, this approach is also not free from caveats. The unusual phyletic pattern may be a consequence of lineage-specific gene loss than a gene gain. Further, the gene of interest may actually be a paralogue that has diverged following the duplication event and therefore does not appear to reside in the closely related genomes. Incomplete genome or the loss of original gene copy complicates the verification of this hypothesis. Another caveat is that the gene displaying unusual pattern might be evolving rapidly, due to selective pressure resulting in unusually high substitution rate. This could occur either in the gene of interest or in the orthologues of this gene in the related genomes. Frequent gene or genome rearrangement and the requirement of multiple strains of closely related species potentially limit the applicability of this approach to lineages or clades with good sampling of completely sequenced genomes. Further, the arbitrary choice of phylogenetic distance to define close or distant relationship renders this approach susceptible to incorrect interpretations.

11.4.3 Similar Genes in Distant Lineages

Pairwise sequence similarity methods such as BLAST are used to find genes with unusually high degree of similarity in otherwise distant lineages (Aravind et al. 1998; Nelson et al. 1999; Lander et al. 2001; Armbrust et al. 2004). Many interdomain gene transfers were reported through this approach. For example, if a gene in a plant appears more similar to bacterial genes than plant genes, this presents an evidence of transfer of this gene from a bacterial genome to a plant genome. Note that such transfers are easier to detect due to large evolutionary distance between donor and recipient organisms, and thus much stronger phylogenetic signal to resolve in order to infer gene transfer events. Although relatively rare, interdomain transfers have contributed significantly to shaping the evolution of extant organisms. Such transfer events have been documented as important players in evolutionary and ecological processes such as host-parasite interaction. However, these methods are also not immune to the vagaries of the comparative approaches. Perhaps, they are highly vulnerable to misinterpretation, and a well-known example comes from the human genome project which reported hundreds of bacterial genes in human genome (Lander et al. 2001) but was strongly refuted by subsequent studies (Salzberg et al. 2001; Stanhope et al. 2001). Therefore, one needs to carefully weigh the caveats including high conservation of the gene of interest coupled with the likely scenario of its differential loss in certain lineages, and also the feasibility of convergent evolution contributing this non-concordance before inferring HGT using this approach. Despite its inherent limitations, this approach is still used frequently but in conjunction with other approaches to add confidence over predictions (Richards et al. 2009).

11.4.4 Phylogenetic Methods and Their Inferences: Most Comprehensive yet Most Confounding

The scope and advantages of phylogenetic methods arise from their inherent ability to construct a roadmap of gene transfer, identifying both the recipient and donor organisms as well as the paths of gene flow. However, the just discussed other likely scenarios that may also explain the observed pattern could be sometimes overwhelming and it is often nontrivial to assign probabilities to each of the alternative scenarios let alone unambiguously rule out the rest in favor of one. By the very nature of their design, the success of these methods solely depends on the breadth and depth of sequence database. Given the sheer complexity and sophistications involved in tree making, quantifying horizontal gene flow at genome scale could be difficult. Although alternative approaches have been developed to make this task computationally less intensive, for example, by prioritizing genes that are more likely to have evolved via HGT, but this issue still remains at the core of phylogenetic limitations (Beiko and Hamilton 2006; Beiko and Ragan 2008, 2009).

In order to catalog plant–fungi HGT, Richards et al. (2009) recently proposed a pipeline that excluded a large proportion of plant genes from further downstream phylogenetic tree analysis; only those plant genes showing the greatest similarity to fungal genes (excluding other plant genes) were selected for phylogenetic tree analysis. Although HGT quantification becomes much faster and applicable genome-wide, this approach is biased towards detecting recent gene transfers from distinct lineages. Another attempt to address this problem culminated in the development of Efficient Evaluation of Edit Paths (EEEP) method (Beiko and Hamilton 2006), however, the computer memory still remained a limiting factor, and further it is hard to resolve the equally parsimonious edit paths and the direction of gene transfer.

Consensus-based methods have been developed to infer an organismal tree, however, since the HGT prediction methods are highly sensitive to heuristically derived organismal tree, any error in extracting the consensus phylogenetic signal would have a profound negative effect on the reliability of inference on all genes being tested for HGT hypothesis.

11.5 Parametric Methods for Alien Gene Detection

This class of methods is based on the premise that an alien gene having evolved in a different (donor) genomic context appears compositionally distinct in the recipient genome context, and could therefore be identified by measuring the compositional disparities against the recipient genome background. Note that ancient transfers are difficult to detect using parametric methods as these alien genes, constrained by recipient's mutation-selection pressure, may have their composition ameliorated to that of the recipient genome (Lawrence and Ochman 1997). However, since most acquired genes are lost over the course of evolution, the repertoire of alien genes in a genome is replete with recently acquired genes. And, therefore, parametric methods have often been invoked to assess the scale and impact of recent gene transfers, particularly, among the microbes (Lawrence and Ochman 1998; Ochman et al. 2000). These methods have sparingly been used for detecting HGT in plants, partly because most remarkable developments in parametric alien gene detection have happened only recently (Arvey et al. 2009; Azad and Lawrence 2011, 2012). The earlier parametric methods used simpler discrimination criteria such as G+C compositional bias to identify alien genes (Lawrence and Ochman 1998); more recent parametric methods have much greater sophistication and have shown consistently high performance in detecting bacterial gene transfer (Vernikos and Parkhill 2006; Azad and Lawrence 2007, 2011; Arvey et al. 2009; Azad and Li 2013). Many of these recent methods hold the promise to robustly quantify the horizontal gene flow among eukaryotes. Since these methods are computationally less intensive and amenable to genome scale analysis, their adaptation for detecting gene transfers in plants will significantly advance our understanding of plant evolution via HGT.

11.5.1 Bottom Up Parametric Methods

These methods perform gene-by-gene analysis to classify each gene as either native or alien (see, for example, Lawrence and Ochman 1998; Garcia-Vallve et al. 2000). Alternatively, without gene information, one can move a fixed size window along a genome sequence and assess the compositional character of the region within the window (Karlin 1998). The bottom up methods can be further categorized as clustering and non-clustering methods.

11.5.1.1 Gene Clustering Methods

The fundamental principle underlying gene clustering methods is that the genes that have evolved under similar evolutionary constraints appear similar to each other and thus could be grouped together and discriminated against other groups having similar genes. Since majority of the genes in a genome are ancestral or native genes, the largest cluster of genes correspond to the genome backbone and all other smaller clusters harbor similar genes that are likely arising from different donor sources. A popular approach to group similar genes is to first randomly assort given genes into *k* number of clusters, and then compute the cluster center (represents the mean of the sequence properties, e.g., nucleotide frequencies, in a cluster) of each cluster, followed by reassignment of genes to the clusters with closest cluster center. This process is repeated until convergence, that is, further reassignment will result in the same cluster configuration. Variants of *k*-means clustering procedure were used for grouping genes with similar compositional pattern in earlier studies (Médigue et al. 1991; Hayes and Borodovsky 1998). One serious limitation of this approach is that one has to specify a priori the number of clusters (value of k) which is often unknown for the given data. For identifying alien genes, a naïvely chosen value of k (e.g., $k=2$) may result in high misclassification errors (Azad and Lawrence 2005). To address this problem, Azad and Lawrence (2007) developed a gene clustering method that identifies the number of clusters inherent to genome heterogeneity in a hypothesis testing framework. Beginning with single gene clusters, a hierarchical agglomerative clustering procedure allows to group recursively two most similar gene clusters. This recursion is halted when the difference between gene clusters in any cluster pair becomes significantly large. The largest cluster is identified as native and the remaining smaller clusters as alien. While this procedure reduced the misclassification errors significantly in comparison to other methods, combining it with biological information such as gene context information for reassigning the compositionally ambiguous genes further reduced the misclassification errors (Azad and Lawrence 2007).

11.5.1.2 Non-clustering Methods

Since a large majority of genes in a typical genome are ancestral, the genome composition (average over all genes) is often taken to represent the composition of ancestral genes. One can thus infer alien genes by assessing the compositional atypicality of a gene against the genome background. Most parametric methods are based on this premise yet they test different hypothesis and thus often lead to non-convergent predictions (Ragan 2001; Lawrence and Ochman 2002). The most simple, and perhaps most used, among these methods, is to measure the discrepancies in nucleotide composition of a gene vis-à-vis the whole genome. Lawrence and Ochman (1998) proposed that if the G+C composition at first and third codon position of a gene deviates significantly from the respective means for all genes, the gene in question is likely an alien gene. Karlin (1998) went a step further, suggesting that the dinucleotide compositional bias is a stronger indicator of atypicality, perhaps inspired by the dinucleotide compositional differences he observed in pairwise comparison of genomes of different species, which led him to propose that dinucleotide composition represents genomic signature, and thus could be exploited to detect alien genes which exemplify genomic signatures of donor organisms and so appear distinct from recipient organism's genomic signature. More recent studies suggest that higher order k-mers carry greater discriminative power and thus can potentially improve alien detection (Tsirigos and Rigoutsos 2005a). Design-Island (Chatterjee et al. 2008) and a chaos game representation-based method (Deschavanne et al. 1999; Dufraigne et al. 2005) were developed for exploiting the power of tetra-nucleotide compositional bias in alien gene detection. Advantages of higher order k-mers include the utilization of codon usage information lying within trimers or longer oligomers ($k>3$), and better predictive abilities encoded within nucleotide ordering patterns arising as a consequence of differential evolutionary forces acting upon genomes of different organisms. Nakamura et al. (2004) used hexamer frequency as a discriminant criterion in a Bayesian formalism, Horizontal Transfer Index, to catalog alien genes in bacterial genomes. Another Bayesian approach, the naïve Bayesian classifier, also used oligomer frequencies to compute the a posteriori probability of a genomic segment to be originating from one of the possible donor sources (Sandberg et al. 2001). However, there is a caveat to the usage of higher order *k*-mers: longer oligomers carry greater predictive ability only if there is a good sampling (recurrence) of longer oligomers in the data. For example, a hexamer, which does not occur frequently enough in the data, cannot be used to predict the nucleotide that just succeeds this hexamer in a DNA sequence. This issue could be circumvented to an extent by using a variable length *k*-mer model, also called interpolated Markov model (Salzberg et al. 1998; Azad and Borodovsky 2004), which was implemented in the IVOM *a.k.a.* Alien Hunter program (Vernikos and Parkhill 2006). Another critical aspect of this class of methods is the choice of measure or model framework for assessing the compositional difference between DNA sequences of interest. Arvey et al. (2009) have shown that an entropy-based measure outperforms a covariance-based measure (Tsirigos and Rigoutsos 2005a) even when the former uses just the nucleotide composition while the latter uses its "optimal" octanucleotide composition as the discriminate criterion. The octanucleotide compositional bias was also exploited in a Support Vector Machine framework (Tsirigos and Rigoutsos 2005b), a frequently invoked supervised learning procedure used successfully in solving a range of biological problems, e.g., disease forecasting (Kaundal et al. 2006), subcellular localization prediction (Kaundal and Raghava 2009; Kaundal et al. 2010). Though the octanucleotide composition approach was outperformed by other methods that used dinucleotide composition in a model selection framework or codon usage in a hypothesis testing framework (Azad and Lawrence 2007). Note that where the gene information is available, one can use codon usage information to exploit the atypical codon usage biases of alien genes. This was implemented in methods by Karlin (1998), and Azad and Lawrence (2007).

11.5.2 Top Down Parametric Methods

While bottom up parametric methods robustly classify the strongly typical and atypical genes, *all* bottom up methods have difficulty in classifying compositionally ambiguous genes. Given that genes often arrive *en masse*, with tens to hundreds acquired in a single transfer event, misclassification of compositionally ambiguous genes in these alien gene islands (also, genomic islands) will lead to overestimation of gene transfer events. Consequently, it will lead to a fragmented structure of otherwise large genome islands. To address this problem, Azad and Lawrence (2011) have recently suggested the use of gene context and operon structural information embedded within the genome of an organism to classify compositionally ambiguous genes in a multiple threshold model framework. However, a robust identification of large acquired regions with dozens of alien genes requires a different approach that separate in this particular case can not just simultaneously analyze multiple genes within an acquired region but be able to do so without regard to gene information and thus predict island boundaries more precisely, which can even lie in non-genic regions. Arvey et al. (2009) have shown that this can be realized in a top down framework. They used a recursive segmentation procedure to divide a given genome sequence recursively into compositionally homogeneous regions within a hypothesis framework. If a homogeneous segment thus obtained was found sufficiently atypical vis-à-vis the genome composition, it was labeled alien. As a consequence, all genes—whether strongly, moderately, or weakly atypical—harbored by this segment, were labeled alien. This class of methods, having demonstrated their power in delineating genomic islands in bacterial genomes, holds a great promise in deciphering large acquired regions in eukaryotic genomes, including genomic islands in plant genomes, where often the gene annotation is incomplete or unavailable.

11.5.3 Parametric Methods and Their Inferences

Parametric methods are becoming increasingly popular because of their simplicity, genome-wide applicability, interpretability, and ease in their implementation. One of the biggest advantages of this class of methods is that these methods do not require multiple related (or sometimes, unrelated) genomes to infer alien genes. The sole input is the genome of an organism (either the whole genome sequence or the sequences of all genes). Alien genes are identified without regard to the presence or absence of their homologues in the genomes of other organisms. However, these methods often generate non-convergent results, which is perhaps because of their testing different hypotheses for being alien (Lawrence and Ochman 2002). Azad and Lawrence (2005) have argued that this is rather a strength than a weakness, for this offers an opportunity to combine the complementary strengths of different parametric methods. To buttress this claim, they combined the predictions from two methods, one using dinucleotide composition and the other using codon usage bias as discriminant criterion, and showed that a simple union of predictions at conservative thresholds significantly minimizes both Type I and Type II errors of misclassification (Azad and Lawrence 2005). Though both, bottom up and top down parametric methods, were designed for different purposes, integration of the two disparate methods will augment the power in delineating and characterizing the compositionally aberrant regions (Arvey et al. 2009; Azad and Lawrence 2012).

Like phylogenetic methods, which suffer from the vagaries related to consensus phylogenetic signals, the performance of bottom up parametric methods is also a function of consensus signal. Often the whole genome composition is assumed to represent the "native" parametric signal; however, this assumption would be severely violated for genomes that have undergone rampant gene transfers. In contrast, bottom up hierarchical clustering methods do not suffer from this limitation. Other caveats include the failure to detect HGT among phylogenetically similar organisms (for example, transfer between *E. coli* and *S. enterica*) and false predictions of otherwise differentially evolving native genes.

11.6 Conclusions

A survey of the recent developments in quantifying HGT in plants highlights the importance of gene transfer in plant genome evolution. It was not long ago that HGT was perceived extremely rare in higher eukaryotes (unlike microbes that swap genetic material frequently among themselves), but this long-held perception has now come into question due to emergence of numerous evidences supporting HGT in eukaryotes, and particularly bolstered by a plethora of plant HGTs reported in recent years. This has infused renewed interest and enthusiasm in the field. There are bottlenecks that must be addressed; this includes the tendency to filter bacterial DNAs if any during eukaryotic genome assembly, and more importantly, conflicting predictions generated by different methods. Integrative approaches to reconcile conflicting signals have remained elusive despite forceful arguments put forward in support of this (Arvey et al. 2009). Parametric methods have come a long way, and with the inclusion of more sophisticated, top down methods in parametric repertoire (Arvey et al. 2009), time is just ripe to exploit the power of these methods, which has, rather surprisingly, been overlooked for alien gene detection in plants. Future strategies should focus on integration of phylogenetic and parametric methods for robustly cataloging both ancient and recent gene transfers in the evolutionary history of plants.

Acknowledgement This work is supported by faculty start-up funds to R.K. from NIMFFAB, Department of Biochemistry & Molecular Biology, Oklahoma State University and similar start-up funds from the University of North Texas to R.K.A. The authors thank the anonymous referees for critically reviewing and help in improving the book chapter.

References

Adams EN III (1972) Consensus techniques and the comparison of taxonomic trees. Syst Biol 21:390–397

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9–e15

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andersson JO (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62:1182–1197

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 14:442–444

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH et al (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86

Arvey AJ, Azad RK, Raval A, Lawrence JG (2009) Detection of genomic islands via segmental genome heterogeneity. Nucleic Acids Res 37:5255–5266

Azad RK, Borodovsky M (2004) Effects of choice of DNA sequence model structure on gene identification accuracy. Bioinformatics 20:993–1005

Azad RK, Lawrence JG (2005) Use of artificial genomes in assessing methods for atypical gene detection. PLoS Comput Biol 1:e56

Azad RK, Lawrence JG (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. Nucleic Acids Res 35:4629–4639

Azad RK, Lawrence JG (2011) Towards more robust methods of alien gene detection. Nucleic Acids Res 39:e56

Azad RK, Lawrence JG (2012) Detecting laterally transferred genes. Methods Mol Biol 855:281–308

Azad RK, Li J (2013) Interpreting genomic data via entropic dissection. Nucleic Acids Res 41:e23

Beiko RG, Charlebois RL (2007) A simulation test bed for hypotheses of genome evolution. Bioinformatics 23:825–831

Beiko RG, Hamilton N (2006) Phylogenetic identification of lateral genetic transfer events. BMC Evol Biol 6:15

Beiko RG, Ragan MA (2008) Detecting lateral genetic transfer : a phylogenetic approach. Methods Mol Biol 452:457–469

Beiko RG, Ragan MA (2009) Untangling hybrid phylogenetic signals: horizontal gene transfer and artifacts of phylogenetic reconstruction. Methods Mol Biol 532:241–256

Bejarano ER, Khashoggi A, Witty M, Lichtenstein C (1996) Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proc Natl Acad Sci U S A 93:759–764

Bergthorsson U, Adams KL, Thomason B, Palmer JD (2003) Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197–201

Bininda-Emonds OR (2005) Supertree construction in the genomic age. Methods Enzymol 395:745–757

Bininda-Emonds OR, Sanderson MJ (2001) Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst Biol 50:565–579

Bock R (2010) The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci 15:11–22

Brown JR, Masuchi Y, Robb FT, Doolittle WF (1994) Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J Mol Evol 38:566–576

Chatterjee R, Chaudhuri K, Chaudhuri P (2008) On detection and assessment of statistical significance of Genomic Islands. BMC Genomics 9:150

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–D368

Christin PA, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA et al (2012) Adaptive evolution of C(4) photosynthesis through recurrent lateral gene transfer. Curr Biol 22:445–449

Day WH, McMorris FR (1992) Consensus sequences based on plurality rule. Bull Math Biol 54:1057–1068

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol 16:1391–1399

Diao X, Freeling M, Lisch D (2006) Horizontal transfer of a plant transposon. PLoS Biol 4:e5

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340

Dong J, Fernandez-Baca D, McMorris FR, Powers RC (2010) Majority-rule (+) consensus trees. Math Biosci 228:10–15

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res 33:e6

Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK, p 350

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A 93:13429–13434

Eulenstein O, Chen D, Burleigh JG, Fernandez-Baca D, Sanderson MJ (2004) Performance of flip supertree construction with a heuristic algorithm. Syst Biol 53:299–308

Felsenstein J (1989) PHYLIP: Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166

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

Gelman A, Rubin DB (1996) Markov chain Monte Carlo methods in biostatistics. Stat Methods Med Res 5:339–355

Gogarten JP (1995) The early evolution of cellular life. Trends Ecol Evol 10:147–151

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

Gophna U, Charlebois RL, Doolittle WF (2006) Ancient lateral gene transfer in the evolution of Bdellovibrio bacteriovorus. Trends Microbiol 14:64–69

Goremykin VV, Salamini F, Velasco R, Viola R (2009) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. Mol Biol Evol 26:99–110

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

Hilario E, Gogarten JP (1993) Horizontal transfer of ATPase genes–the tree of life becomes a net of life. Biosystems 31:111–119

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

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

Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537:39–64

Kaundal R, Raghava GPS (2009) RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. Proteomics 9:2324–2342

Kaundal R, Kapoor AS, Raghava GPS (2006) Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC Bioinforma 7:485

Kaundal R, Saini R, Zhao PX (2010) Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis. Plant Physiol 154:36–54

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29:170–179

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742

Koulintchenko M, Konstantinov Y, Dietrich A (2003) Plant mitochondria actively import DNA via the permeability transition pore complex. EMBO J 22:1245–1254

Kurland CG (2005) What tangled web: barriers to rampant horizontal gene transfer. Bioessays 27:741–747

Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A 100:9658–9662

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lapierre P, Lasek-Nesselquist E, Gogarten JP (2012) The impact of HGT on phylogenomic reconstruction methods. Brief Bioinform (in press)

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

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

Lawrence JG, Ochman H (2002) Reconciling the many faces of gene transfer. Trends Microbiol 10:1–4

Logan DC (2006) Plant mitochondrial dynamics. Biochim Biophys Acta 1763:430–441

Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence of horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222:851–856

Nakamura A, Schmitt M, Schmitt N, Simon HU (2005) Inner Product Spaces for Bayesian Networks. J Machine Learning Res 6:1383–1403

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH et al (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329

Nesbo CL, Boucher Y, Doolittle WF (2001) Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol 53:340–350

Nguyen N, Mirarab S, Warnow T (2012) MRL and SuperFine+MRL: new supertree methods. Algorithms Mol Biol 7:3

Ochman H, Lawrence JG, Groisman E (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Olsen GJ, Woese CR (1993) Ribosomal RNA: a key to phylogeny. FASEB J 7:113–123

Pevsner J (2003) Bioinformatics and functional genomics. John Wiley & Sons, Hoboken, New Jerssey

Pible O, Imbert G, Pellequer JL (2005) INTERALIGN: interactive alignment editor for distantly related protein sequences. Bioinformatics 21:3166–3167

Poptsova M (2009) Testing phylogenetic methods to identify horizontal gene transfer. Methods Mol Biol 532:227–240

Ragan MA (2001) On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett 201:187–191

Richards TA, Soanes DM, Foster PG, Leonard G, Thornton CR, Talbot NJ (2009) Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell 21:1897–1911

Richardson AO, Palmer JD (2007) Horizontal gene transfer in plants. J Exp Bot 58:1–9

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26:544–548

Salzberg SL, White O, Peterson J, Eisen JA (2001) Microbial genes in the human genome: lateral transfer or gene loss? Science 292:1903–1906

Sanchez C (2011) Horizontal gene transfer: eukaryotes under a new light. Nat Rev Microbiol 9:228

Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res 11:1404–1409

Schmidt HA, von Haeseler A (2007) Maximum-likelihood analysis using TREE-PUZZLE. Curr Protoc Bioinformatics. Chapter 6:Unit 6 6

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Sheveleva EV, Hallick RB (2004) Recent horizontal intron transfer to a chloroplast genome. Nucleic Acids Res 32:803–810

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Syst Biol 51:492–508

Shimodaira H, Hasegawa M (1999) Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol Biol Evol 16:1114–1116

Simossis VA, Heringa J (2003) The PRALINE online server: optimising progressive multiple alignment on the web. Comput Biol Chem 27:511–519

Simossis VA, Heringa J (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 33:W289–W294

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411:940–944

Stegemann S, Bock R (2009) Exchange of genetic material between cells in plant tissue grafts. Science 324:649–651

Stegemann S, Keuthe M, Greiner S, Bock R (2012) Horizontal transfer of chloroplast genomes between plant species. Proc Natl Acad Sci U S A 109:2434–2438

Suzuki K, Yamashita I, Tanaka N (2002) Tobacco plants were transformed by Agrobacterium rhizogenes infection during their evolution. Plant J 32:775–787

Swenson MS, Suri R, Linder CR, Warnow T (2012) SuperFine: fast and accurate supertree estimation. Syst Biol 61:214–227

Swofford D (1998) PAUP* 4.0 Beta version, phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Inc., Sunderland, Massachusetts

Swofford D, Olsen GJO (1990) Phylogenetic reconstruction. In: Hillis DM, Moritz C (eds) Molecular systematics. Sinauer Associates, Inc., Sunderland, Massachusetts, pp 411–501

Syvanen M, Kado CI (1998) Horizontal Gene Transfer. Chapman & Hall, Lodon

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36

Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics. Chapter 2:Unit 2 3

Tsirigos A, Rigoutsos I (2005a) A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res 33:922–933

Tsirigos A, Rigoutsos I (2005b) A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. Nucleic Acids Res 33:3699–3707

Vaughn JC, Mason MT, Sper-Whitis GL, Kuhlman P, Palmer JD (1995) Fungal origin by horizontal transfer of a plant mitochondrial group I intron in the chimeric CoxI gene of Peperomia. J Mol Evol 41:563–572

Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22:2196–2203

Vernikos GS, Parkhill J (2008) Resolving the structural features of genomic islands: a machine learning approach. Genome Res 18:331–342

Woese CR (1991) The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria. In: Selander RK, Clark AG, Whittam TS (eds) Evolution at the Molecular Level. Sinauer Associates Inc., Sunderland, MA, pp 1–24

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci U S A 87:4576–4579

Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. Trends Genet 18:472–479

Woloszynska M, Bocer T, Mackiewicz P, Janska H (2004) A fragment of chloroplast DNA was transferred horizontally, probably from non-eudicots, to mitochondrial genome of Phaseolus. Plant Mol Biol 56:811–820

Won H, Renner SS (2003) Horizontal gene transfer from flowering plants to Gnetum. Proc Natl Acad Sci U S A 100:10824–10829