PLOS | COMPUTATIONAL BIOLOGY

TOPIC PAGE

Inferring Horizontal Gene Transfer

Matt Ravenhall[1☯], Nives Škunca[2,3☯], Florent Lassalle[1], Christophe Dessimoz[1,4]*

1 University College London, London, United Kingdom, **2** ETH Zurich, Zurich, Switzerland, **3** Swiss Institute of Bioinformatics, Zurich, Switzerland, **4** European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

☯ These authors contributed equally to this work.
* c.dessimoz@ucl.ac.uk

Abstract

Horizontal or Lateral Gene Transfer (HGT or LGT) is the transmission of portions of genomic DNA between organisms through a process decoupled from vertical inheritance. In the presence of HGT events, different fragments of the genome are the result of different evolutionary histories. This can therefore complicate the investigations of evolutionary relatedness of lineages and species. Also, as HGT can bring into genomes radically different genotypes from distant lineages, or even new genes bearing new functions, it is a major source of phenotypic innovation and a mechanism of niche adaptation. For example, of particular relevance to human health is the lateral transfer of antibiotic resistance and pathogenicity determinants, leading to the emergence of pathogenic lineages [1]. Computational identification of HGT events relies upon the investigation of sequence composition or evolutionary history of genes. Sequence composition-based ("parametric") methods search for deviations from the genomic average, whereas evolutionary history-based ("phylogenetic") approaches identify genes whose evolutionary history significantly differs from that of the host species. The evaluation and benchmarking of HGT inference methods typically rely upon simulated genomes, for which the true history is known. On real data, different methods tend to infer different HGT events, and as a result it can be difficult to ascertain all but simple and clear-cut HGT events.

This is a 'Topic Page' article for PLOS Computational Biology.

Citation: Ravenhall M, Škunca N, Lassalle F, Dessimoz C (2015) Inferring Horizontal Gene Transfer. PLoS Comput Biol 11(5): e1004095. doi:10.1371/journal.pcbi.1004095

Editor: Shoshana Wodak, University of Toronto, Canada

Published: May 28, 2015

Funding: FL was supported by ERC grant BIG_IDEA 260801. CD was supported in part by SNSF advanced researcher fellowship #136461. The funders had no role in the preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Horizontal gene transfer (HGT) was first observed in 1928, in Frederick Griffith's experiment. Showing that virulence was able to pass from virulent to nonvirulent strains of *Streptococcus pneumoniae*, Griffith demonstrated that genetic information can be horizontally transferred between bacteria via a mechanism known as transformation [2]. Similar observations in the 1940s [3] and 1950s [4] showed evidence that conjugation and transduction are additional mechanisms of horizontal gene transfer [5].

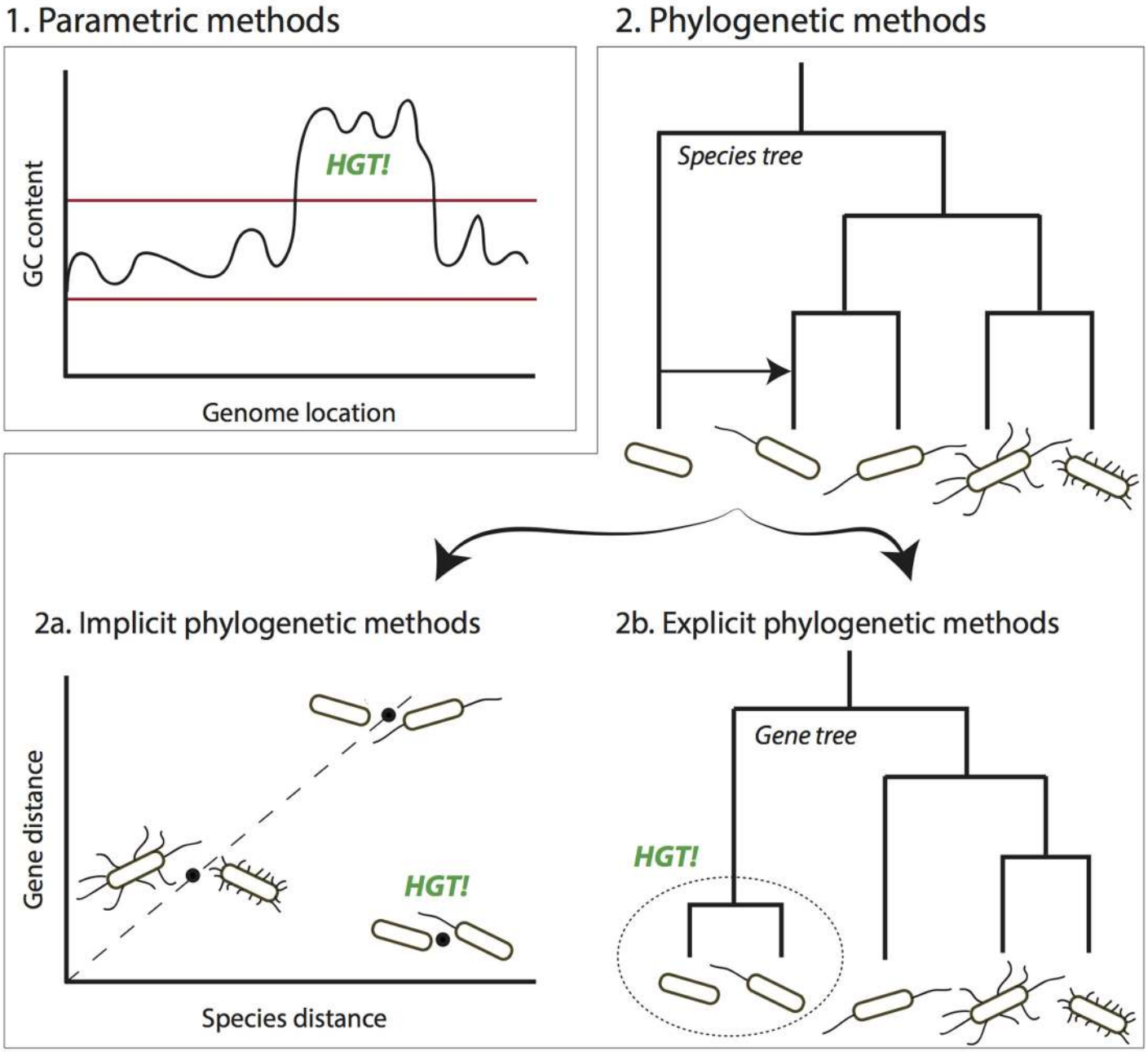

Fig 1. Conceptual overview of HGT inference methods. (1) Parametric methods infer HGT by computing a statistic, here GC content, for a sliding window and comparing it to the typical range over the entire genome, here indicated between the two red horizontal lines. Regions with atypical values are inferred as having been horizontally transferred. (2) Phylogenetic approaches rely on the differences between genes and species tree evolution that result from HGT. Explicit phylogenetic methods reconstruct gene trees and infer the HGT events likely to have resulted into that particular gene tree. Implicit phylogenetic methods bypass gene tree reconstruction, e.g., by looking at discrepancies between pairwise distances between genes and their corresponding species.

doi:10.1371/journal.pcbi.1004095.g001

To infer HGT events, which may not necessarily result in phenotypic changes, most contemporary methods are based on analyses of genomic sequence data. These methods can be broadly separated into two groups: parametric and phylogenetic methods (Fig 1). Parametric methods search for sections of a genome that significantly differ from the genomic average,

such as guanine-cytosine (GC) content or codon usage [6]. Phylogenetic methods examine evolutionary histories of genes involved and identify conflicting phylogenies. Phylogenetic methods can be further divided into those that reconstruct and compare phylogenetic trees explicitly and those that use surrogate measures in place of the phylogenetic trees [7].

The main feature of parametric methods is that they only rely on the genome under study to infer HGT events that may have occurred on its lineage. It has been a considerable advantage at the early times of the sequencing era, when few closely related genomes were available for comparative methods. However, because they rely on the uniformity of the host's signature to infer HGT events, not accounting for the host's intragenomic variability will result in overpredictions—flagging native segments as possible HGT events [8]. Similarly, the transferred segments need to exhibit the donor's signature and to be significantly different from the recipient's [6]. Furthermore, genomic segments of foreign origin are subject to the same mutational processes as the rest of the host genome, and so the difference between the two tends to vanish over time, a process referred to as amelioration [9]. This limits the ability of parametric methods to detect ancient HGTs.

Phylogenetic methods benefit from the recent availability of many sequenced genomes. Indeed, as for all comparative methods, phylogenetic methods can integrate information from multiple genomes and in particular integrate them using a model of evolution. This lends them the ability to better characterize the HGT events they infer—notably by designating the donor species and time of the transfer. However, models have limits and need to be used cautiously. For instance, the conflicting phylogenies can be the result of events not accounted for by the model, such as unrecognized paralogy due to duplication followed by gene losses. Also, many approaches rely on a reference species tree that is supposed to be known, when in many instances it can be difficult to obtain a reliable species tree. Finally, the computational costs of reconstructing many gene and species trees can be prohibitively expensive. Phylogenetic methods tend to be applied to genes or protein sequences as basic evolutionary units, which limits their ability to detect HGT in regions outside or across gene boundaries.

Because of their complementary approaches—and often nonoverlapping sets of HGT candidates—combining predictions from parametric and phylogenetic methods can yield a more comprehensive set of HGT candidate genes. Indeed, combining different parametric methods has been reported to significantly improve the quality of predictions [10,11]. Moreover, in the absence of a comprehensive set of true horizontally transferred genes, discrepancies between different methods [12,13] might be resolved through combining parametric and phylogenetic methods. However, combining inferences from multiple methods also entails a risk of an increased false positive rate [14].

Parametric Methods

Parametric methods to infer HGT use characteristics of the genome sequence specific to particular species or clades, also called genomic signatures. If a fragment of the genome strongly deviates from the genomic signature, this is a sign of a potential horizontal transfer. For example, because bacterial GC content falls within a wide range (see Fig 2), GC content of a genome segment is a simple genomic signature. Commonly used genomic signatures include nucleotide composition [15], oligonucleotide frequencies [16], or structural features of the genome [17].

To detect HGT using parametric methods, the host's genomic signature needs to be clearly recognizable. However, the host's genome is not always uniform with respect to the genome signature; for example, GC content of the third codon position is lower close to the replication terminus [18], and GC content tends to be higher in highly-expressed genes [19]. Not accounting for such intragenomic variability in the host can result in overpredictions, flagging native

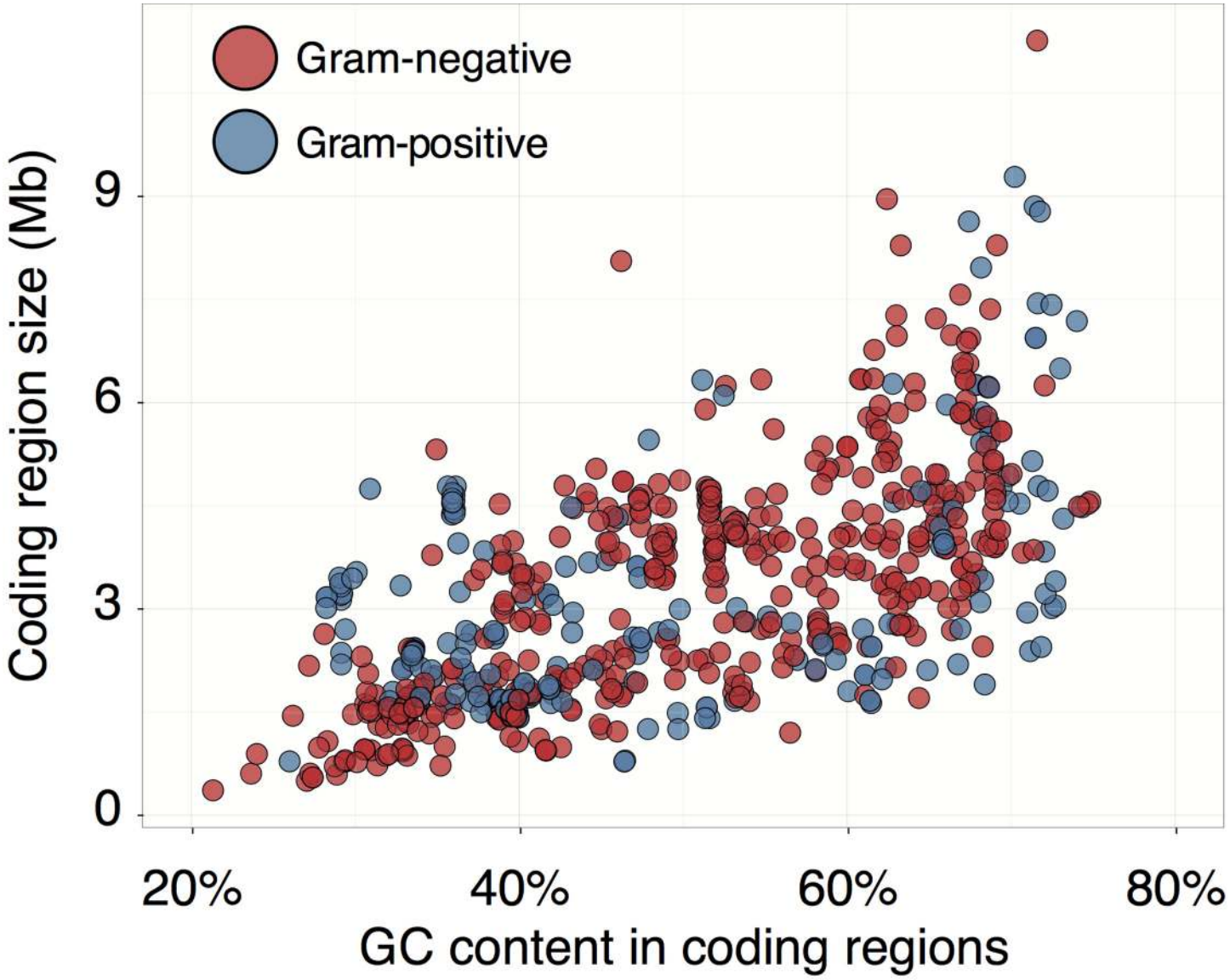

Fig 2. Average GC content of coding regions compared to the genome size for selected bacteria. There is considerable variation in average GC content across species, which makes it relevant as a genomic signature.

doi:10.1371/journal.pcbi.1004095.g002

segments as HGT candidates [8]. Larger sliding windows can account for this variability at the cost of a reduced ability to detect smaller HGT regions [12].

Just as importantly, horizontally transferred segments need to exhibit the donor's genomic signature. This might not be the case for ancient transfers where transferred sequences are subjected to the same mutational processes as the rest of the host genome, potentially causing their distinct signatures to "ameliorate" [9] and become undetectable through parametric methods. For example, *Bdellovibrio bacteriovorus*, a predatory δ-Proteobacterium, has homogeneous GC content, and it might be concluded that its genome is resistant to HGT [20]. However, subsequent analysis using phylogenetic methods identified a number of ancient HGT events in the genome of *B. bacteriovorus* [21]. Similarly, if the inserted segment was previously ameliorated

to the host's genome, as is the case for prophage insertions [22], parametric methods might miss predicting these HGT events. Also, the donor's composition must significantly differ from the recipient's to be identified as abnormal, a condition that might be missed in the case of short- to medium-distance HGT, which are the most prevalent. Furthermore, it has been reported that recently acquired genes tend to be more AT-rich than the recipient's average [15], which indicates that differences in GC content signature may result from unknown postacquisition mutational processes rather than from the donor's genome.

Nucleotide composition

Bacterial GC content falls within a wide range, with *Candidatus Zinderia insecticola* having a GC content of 13.5% [23] and *Anaeromyxobacter dehalogenans* having a GC content of 75% [24] (see Fig 2). Even within a closely related group of α-Proteobacteria, values range from approximately 30% to 65% [25]. These differences can be exploited when detecting HGT events as a significantly different GC content for a genome segment can be an indication of foreign origin [15] (see Fig 2).

Oligonucleotide spectrum

The oligonucleotide spectrum (or k-mer frequencies) measures the frequency of all possible nucleotide sequences of a particular length in the genome. It tends to vary less within genomes than between genomes and therefore can also be used as a genomic signature [26]. A deviation from this signature suggests that a genomic segment might have arrived through horizontal transfer.

The oligonucleotide spectrum owes much of its discriminatory power to the number of possible oligonucleotides; if n is the size of the vocabulary and w is oligonucleotide size, the number of possible distinct oligonucleotides is n^w; for example, there are $4^5 = 1{,}024$ possible pentanucleotides. Some methods can capture the signal recorded in motifs of variable size [27], thus capturing both rare and discriminative motifs along with frequent but more common ones.

Codon usage bias, a measure related to codon frequencies, was one of the first detection methods used in methodical assessments of HGT [16]. This approach requires a host genome which contains a bias towards certain synonymous codons (different codons which code for the same amino acid), which is clearly distinct from the bias found within the donor genome. The simplest oligonucleotide used as a genomic signature is the dinucleotide; for example, the third nucleotide in a codon and the first nucleotide in the following codon represent the dinucleotide least restricted by amino acid preference and codon usage [28].

It is important to optimise the size of the sliding window in which to count the oligonucleotide frequency; a larger sliding window will better buffer variability in the host genome at the cost of being worse at detecting smaller HGT regions [29]. A good compromise has been reported using tetranucleotide frequencies in a sliding window of 5 kb with a step of 0.5 kb [30].

A convenient method of modelling oligonucleotide genomic signatures is to use Markov chains. The transition probability matrix can be derived for endogenous versus acquired genes [31], from which Bayesian posterior probabilities for particular stretches of DNA can be obtained [32].

Structural features

Just as the nucleotide composition of a DNA molecule can be represented by a sequence of letters, its structural features can be encoded in a numerical sequence. The structural features include interaction energies between neighbouring base pairs [33], the angle of twist that makes

two bases of a pair noncoplanar [34], or DNA deformability induced by the proteins shaping the chromatin [35].

The autocorrelation analysis of some of these numerical sequences show characteristic periodicities in complete genomes [36]. In fact, after detecting archaea-like regions in the thermophilic bacteria *Thermotoga maritima* [37], periodicity spectra of these regions were compared to the periodicity spectra of the homologous regions in the archaea *Pyrococcus horikoshii* [17]. The revealed similarities in the periodicity were strong supporting evidence for a case of massive HGT between the bacteria and the archaea kingdoms [17].

Genomic context

The existence of genomic islands, short (typically 10–200 kb long) regions of a genome which have been acquired horizontally, lends support to the ability to identify non-native genes by their location in a genome [38]. For example, a gene of ambiguous origin which forms part of a non-native operon could be considered to be non-native. Alternatively, flanking repeat sequences or the presence of nearby integrases or transposases can indicate a non-native region [39]. A machine-learning approach combining oligonucleotide frequency scans with context information was reported to be effective at identifying genomic islands [40]. In another study, the context was used as a secondary indicator, after removal of genes which are strongly thought to be native or non-native through the use of other parametric methods [10].

Phylogenetic Methods

The use of phylogenetic analysis in the detection of HGT was advanced by the availability of many newly sequenced genomes. Phylogenetic methods detect inconsistencies in gene and species evolutionary history in two ways: explicitly, by reconstructing the gene tree and reconciling it with the reference species tree, or implicitly, by examining aspects that correlate with the evolutionary history of the genes in question, e.g., patterns of presence and absence across species, or unexpectedly short or distant pairwise evolutionary distances.

Explicit phylogenetic methods

The aim of explicit phylogenetic methods is to compare gene trees with their associated species trees. While weakly-supported differences between gene and species trees can be due to inference uncertainty, statistically significant differences can be suggestive of HGT events (see Fig 1A). For example, if two genes from different species share the most recent ancestral connecting node in the gene tree, but the respective species are spaced apart in the species tree, an HGT event can be invoked. Such an approach can produce more detailed results than parametric approaches because the involved species, time, and direction of transfer can potentially be identified.

As discussed in more details below, phylogenetic methods range from simple methods merely identifying discordance between gene and species trees to mechanistic models inferring probable sequences of HGT events. An intermediate strategy entails deconstructing the gene tree into smaller parts until each matches the species tree (genome spectral approaches).

Explicit phylogenetic methods rely upon the accuracy of the input rooted gene and species trees, yet these can be challenging to build [41]. Even when there is no doubt in the input trees, the conflicting phylogenies can be the result of evolutionary processes other than HGT, such as duplications and losses, causing these methods to erroneously infer HGT events when paralogy is the correct explanation. Similarly, in the presence of incomplete lineage sorting, explicit phylogeny methods can erroneously infer HGT events [42]. That is why some explicit model-

based methods test multiple evolutionary scenarios involving different kinds of events and compare their fit to the data, given parsimonious or probabilistic criteria.

Tests of topologies. To detect sets of genes that fit poorly to the reference tree, one can use statistical tests of topology, such as the Kishino-Hasegawa (KH) [43], Shimodaira-Hasegawa (SH) [44], and Approximately Unbiased (AU) [45] tests. These tests assess the likelihood of the gene sequence alignment when the reference topology is given as the null hypothesis.

The rejection of the reference topology is an indication that the evolutionary history for that gene family is inconsistent with the reference tree. When these inconsistencies cannot be explained using a small number of nonhorizontal events, such as gene loss and duplication, an HGT event is inferred.

One such analysis checked for HGT in groups of homologs of the γ-Proteobacterial lineage [46]. Six reference trees were reconstructed using either the highly conserved small subunit ribosomal RNA sequences, a consensus of the available gene trees or concatenated alignments of orthologs. The failure to reject the six evaluated topologies, and the rejection of seven alternative topologies, was interpreted as evidence for a small number of HGT events in the selected groups.

Tests of topology identify differences in tree topology taking into account the uncertainty in tree inference, but they make no attempt at inferring how the differences came about. To infer the specifics of particular events, genome spectral or subtree pruning and regraft methods are required.

Genome spectral approaches. In order to identify the location of HGT events, genome spectral approaches decompose a gene tree into substructures (such as bipartitions or quartets) and identify those that are consistent or inconsistent with the species tree.

Removing one edge from a reference tree produces two unconnected subtrees, each containing a disjoint set of nodes—a bipartition. If a bipartition is present in both the gene and the species trees, it is compatible; otherwise, it is conflicting. These conflicts can indicate an HGT event or may be the result of uncertainty in gene tree inference. To reduce uncertainty, bipartition analyses typically focus on strongly supported bipartitions such as those associated with branches with bootstrap values or posterior probabilities above certain thresholds. Any gene family found to have one or several conflicting, but strongly supported, bipartitions is considered as an HGT candidate [47,48].

Alternatively, trees can be decomposed into quartets. Quartets are trees consisting of four leaves. In bifurcating (fully resolved) trees, each internal branch induces a quartet whose leaves are either subtrees of the original tree or actual leaves of the original tree. If the topology of a quartet extracted from the reference species tree is embedded in the gene tree, the quartet is compatible with the gene tree. Conversely, incompatible strongly supported quartets indicate potential HGT events [49]. Quartet mapping methods are much more computationally efficient and naturally handle heterogeneous representation of taxa among gene families, making them a good basis for developing large-scale scans for HGT, looking for highways of gene sharing in databases of hundreds of complete genomes [50,51].

Subtree pruning and regrafting. A mechanistic way of modelling an HGT event on the reference tree is to first cut an internal branch—i.e., prune the tree—and then regraft it onto another edge, an operation referred to as subtree pruning and regrafting (SPR) [52]. If the gene tree was topologically consistent with the original reference tree, the editing results in an inconsistency. Similarly, when the original gene tree is inconsistent with the reference tree, it is possible to obtain a consistent topology by a series of one or more prune and regraft operations applied to the reference tree. By interpreting the edit path of pruning and regrafting, HGT candidate nodes can be flagged and the host and donor genomes inferred [48,53]. To avoid reporting false positive HGT events due to uncertain gene tree topologies, the optimal "path" of SPR

operations can be chosen among multiple possible combinations by considering the branch support in the gene tree. Weakly supported gene tree edges can be ignored a priori [54], or the support can be used to compute an optimality criterion [55,56].

Because conversion of one tree to another by a minimum number of SPR operations is NP-Hard [57], solving the problem becomes considerably more difficult as more nodes are considered. The computational challenge lies in finding the optimal edit path, i.e., the one that requires the fewest steps [58,59], and different strategies are used in solving the problem. For example, the HorizStory algorithm reduces the problem by first eliminating the consistent nodes [60]; recursive pruning and regrafting reconciles the reference tree with the gene tree and optimal edits are interpreted as HGT events. The SPR methods included in the supertree reconstruction package SPRSupertrees substantially decrease the time of the search for the optimal set of SPR operations by considering multiple localised subproblems in large trees through a clustering approach [61].

Model-based reconciliation methods. Reconciliation of gene and species trees entails mapping evolutionary events onto gene trees in a way that makes them concordant with the species tree, given a mechanistic model. Different reconciliation models exist, differing in the types of event they consider to explain the incongruences between gene and species tree topologies. Early methods exclusively modelled horizontal transfers (T) [52,55]. More recent ones also account for duplication (D), loss (L), incomplete lineage sorting (ILS), or homologous recombination (HR) events. The difficulty is that by allowing for multiple types of events, the number of possible reconciliations increases rapidly. For instance, conflicting gene tree topologies might be explained in terms of a single HGT event or multiple duplication and loss events. Both alternatives can be considered plausible reconciliation depending on the frequency of these respective events along the species tree.

Reconciliation methods can rely on a parsimonious or a probabilistic framework to infer the most likely scenario(s), where the relative cost and probability of D, T, and L events can be fixed a priori or estimated from the data [62]. The space of DTL reconciliations and their parsimony costs—which can be extremely vast for large multicopy gene family trees—can be efficiently explored through dynamic programming algorithms [63–65]. In some programs, the gene tree topology can be refined where it was uncertain to fit a better evolutionary scenario as well as the initial sequence alignment [63,66,67]. More refined models account for the biased frequency of HGT between closely related lineages [68], reflecting the loss of efficiency of HR with phylogenetic distance [69], for ILS [70], or for the fact that the actual donor of most HGT belong to extinct or unsampled lineages [71]. Further extensions of DTL models are being developed towards an integrated description of the genome evolution processes. In particular, some of them consider horizontal transfer at multiple scales—modelling independent evolution of gene fragments [72] or recognising coevolution of several genes (e.g., due to cotransfer) within and across genomes [73].

Implicit phylogenetic methods

In contrast to explicit phylogenetic methods, which compare the agreement between gene and species trees, implicit phylogenetic methods compare evolutionary distances or sequence similarity. Here, an unexpectedly short or long distance from a given reference compared to the average can be suggestive of an HGT event (see Fig 1). Because tree construction is not required, implicit approaches tend to be simpler and faster than explicit methods.

However, implicit methods can be limited by disparities between the underlying correct phylogeny and the evolutionary distances considered. For instance, the most similar sequence as obtained by the highest-scoring BLAST hit is not always the evolutionarily closest one [74].

Top sequence match in a distant species. A simple way of identifying HGT events is by looking for high-scoring sequence matches in distantly related species. For example, an analysis of the top BLAST hits of protein sequences in the bacteria *Thermotoga maritima* revealed that most hits were in archaea rather than closely-related bacteria, suggesting extensive HGT between the two [37]; these predictions were later supported by an analysis of the structural features of the DNA molecule [17].

However, this method is limited to detecting relatively recent HGT events. Indeed, if the HGT occurred in the common ancestor of two or more species included in the database, the closest hit will reside within that clade, and therefore the HGT will not be detected by the method. Thus, the threshold of the minimum number of foreign top BLAST hits to observe to decide a gene was transferred is highly dependent on the taxonomic coverage of sequence databases. Therefore, experimental settings may need to be defined in an ad-hoc way [75].

Discrepancy between gene and species distances. The molecular clock hypothesis posits that homologous genes evolve at an approximately constant rate across different species [76]. If one only considers homologous genes related through speciation events (referred to as "orthologous" genes), their underlying tree should by definition correspond to the species tree. Therefore, assuming a molecular clock, the evolutionary distance between orthologous genes should be approximately proportional to the evolutionary distances between their respective species. If a putative group of orthologs contains xenologs (pairs of genes related through an HGT), the proportionality of evolutionary distances may only hold among the orthologs, not the xenologs [77].

Simple approaches compare the distribution of similarity scores of particular sequences and their orthologous counterparts in other species; HGT are inferred from outliers [78,79]. The more sophisticated DLIGHT (Distance Likelihood-based Inference of Genes Horizontally Transferred) method considers simultaneously the effect of HGT on all sequences within groups of putative orthologs [7]: if a likelihood-ratio test of the HGT hypothesis versus a hypothesis of no HGT is significant, a putative HGT event is inferred. In addition, the method allows inference of potential donor and recipient species and provides an estimation of the time since the HGT event.

Phylogenetic profiles. A group of orthologous or homologous genes can be analysed in terms of the presence or absence of group members in the reference genomes; such patterns are called phylogenetic profiles [80]. To find HGT events, phylogenetic profiles are scanned for an unusual distribution of genes. Isolated occurrence of a gene, i.e., absence of a homolog in other members of a group of closely related species is an indication that the examined gene might have arrived via an HGT event. For example, the three facultatively symbiotic *Frankia* spp. strains are of strikingly different sizes: 5.43 Mbp, 7.50 Mbp, and 9.04 Mbp, depending on their range of hosts [81]. Marked portions of strain-specific genes were found to have no significant hit in the reference database and were possibly acquired by HGT transfers from other bacteria. Similarly, three phenotypically diverse *Escherichia coli* strains (uropathogenic, enterohemorrhagic, and benign) shared about 40% of the total combined gene pool, with the other 60% being strain-specific genes and, consequently, HGT candidates [82]. Further evidence for these genes resulting from HGT was their strikingly different codon usage patterns from the core genes and a lack of gene order conservation (order conservation is typical of vertically-evolved genes) [82]. The presence and absence of homologs (or their effective count) can thus be used by programs to reconstruct the most likely evolutionary scenario along the species tree. Just as with reconciliation methods, this can be achieved through parsimonious [83] or probabilistic estimation of the number of gain and loss events [84,85]. Models can be complexified by adding processes, like the truncation of genes [86], but also by modelling the heterogeneity of rates of gain and loss across lineages [87] and/or gene families [85,88].

Clusters of polymorphic sites. Genes are commonly regarded as the basic units transferred through an HGT event. However, it is also possible for HGT to occur within genes. For example, it has been shown that horizontal transfer between closely related species results in more exchange of ORF fragments [89,90], a type a transfer called gene conversion, mediated by homologous recombination. The analysis of a group of four *E. coli* and two *Shigella flexneri* strains revealed that the sequence stretches common to all six strains contain polymorphic sites, consequences of homologous recombination [91]. Clusters of excess of polymorphic sites can thus be used to detect tracks of DNA recombined with a distant relative [92]. This method of detection is, however, restricted to the sites in common with all analysed sequences, limiting the analysis to a group of closely related organisms.

Evaluation

The existence of the numerous and varied methods to infer HGT raises the question of how to validate individual inferences and of how to compare the different methods.

A main problem is that, as with other types of phylogenetic inferences, the actual evolutionary history cannot be established with certainty. As a result, it is difficult to obtain a representative test set of HGT events. Furthermore, HGT inference methods vary considerably in the information they consider and often identify inconsistent groups of HGT candidates [6,93]; it is not clear to what extent taking the intersection, the union, or some other combination of the individual methods affects the false positive and false negative rates [14].

Parametric and phylogenetic methods draw on different sources of information; it is therefore difficult to make general statements about their relative performance. Conceptual arguments can, however, be invoked. While parametric methods are limited to the analysis of single genomes or pairs of genomes, phylogenetic methods provide a natural framework to take advantage of the information contained in multiple genomes. In many cases, segments of genomes inferred as HGT based on their anomalous composition can also be recognised as such on the basis of phylogenetic analyses or through their mere absence in genomes of related organisms. In addition, phylogenetic methods rely on explicit models of sequence evolution, which provide a well-understood framework for parameter inference, hypothesis testing, and model selection. This is reflected in the literature, which tends to favour phylogenetic methods as the standard of proof for HGT [94–97]. The use of phylogenetic methods thus appears to be the preferred standard, especially given that the increase in computational power coupled with algorithmic improvements has made them more tractable [61,71], and that the ever denser sampling of genomes lends more power to these tests.

Considering phylogenetic methods, several approaches to validating individual HGT inferences and benchmarking methods have been adopted, typically relying on various forms of simulation. Because the truth is known in simulation, the number of false positives and the number of false negatives are straightforward to compute. However, simulating data does not trivially resolve the problem, because the true extent of HGT in nature remains largely unknown, and specifying rates of HGT in the simulated model is always hazardous. Nonetheless, studies involving the comparison of several phylogenetic methods in a simulation framework could provide quantitative assessment of their respective performances and thus help the biologist in choosing objectively proper tools [56].

Standard tools to simulate sequence evolution along trees such as INDELible [98] or PhyloSim [99] can be adapted to simulate HGT. HGT events cause the relevant gene trees to conflict with the species tree. Such HGT events can be simulated through subtree pruning and regrafting rearrangements of the species tree [54]. However, it is important to simulate data that are realistic enough to be representative of the challenge provided by real datasets, and simulation

under complex models are thus preferable. A model was developed to simulate gene trees with heterogeneous substitution processes in addition to the occurrence of transfer and accounting for the fact that transfer can come from now extinct donor lineages [100]. Alternatively, the genome evolution simulator Artificial Life Simulator (ALF) [101] directly generates gene families subject to HGT by accounting for a whole range of evolutionary forces at the base level but in the context of a complete genome. Given simulated sequences which have HGT, analysis of those sequences using the methods of interest and comparison of their results with the known truth permits study of their performance. Similarly, testing the methods on sequences known not to have HGT enables the study of false positive rates.

Simulation of HGT events can also be performed by manipulating the biological sequences themselves. Artificial chimeric genomes can be obtained by inserting known foreign genes into random positions of a host genome [12,102–104]. The donor sequences are inserted into the host unchanged or can be further evolved by simulation [7], e.g., using the tools described above.

One important caveat to simulation as a way to assess different methods is that simulation is based on strong simplifying assumptions that may favour particular methods [105].

Supporting Information

S1 Text. Version history of the text file.
(XML)

S2 Text. Peer reviews and response to reviews. Human-readable versions of the reviews and authors' responses are available as comments on this article.
(XML)

Acknowledgments

The authors thank Daniel A. Dalquen, Nick Goldman, Kevin Gori, Jelena Repar, Hannes Röst, Greg Slodkowicz, Fran Supek, and Stefan Zoller for helpful comments and suggestions, as well as Tom Williams and Robert Beiko for constructive peer reviews. We also thank Daniel Mietchen for his engaged and valuable editorial work. This article started as an assignment for the graduate course "Reviews in Computational Biology" at ETH Zurich. The version history of the text file and the peer reviews (and response to reviews) are available as supporting information in S1 and S2 Texts.

References

1. Hiramatsu K, Cui L, Kuroda M, Ito T (2001) The emergence and evolution of methicillin-resistant Staphylococcus aureus. Trends Microbiol. 9:486–93 PMID: 11597450
2. Griffith F (1928) The Significance of Pneumococcal Types. J Hyg (Lond) 27:113–59 PMID: 20474956
3. Tatum E L, Lederberg J (1947) Gene Recombination in the Bacterium Escherichia coli. J. Bacteriol. 53:673–84
4. ZINDER N D, LEDERBERG J (1952) Genetic exchange in Salmonella. J. Bacteriol. 64:679–99
5. Jones D, Sneath P H (1970) Genetic transfer and bacterial taxonomy. Bacteriol Rev 34:40–81 PMID: 4909647
6. Jeffrey G Lawrence, Ochman Howard (2002) Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10:1–4
7. Christophe Dessimoz, Margadant Daniel, and Gaston H Gonnet. 2008. DLIGHT—Lateral Gene Transfer Detection Using Pairwise Evolutionary Distances in a Statistical Framework. Springer, 4955:315–330. doi: 10.1136/bmj.h2068 PMID: 25908437
8. Guindon S, Perrière G (2001) Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. Mol. Biol. Evol. 18:1838–40

9. Lawrence J G, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44:383–97 PMID: 9089078
10. Azad Rajeev K, Lawrence Jeffrey G (2011) Towards more robust methods of alien gene detection. Nucleic Acids Res. 39:e56 doi: 10.1093/nar/gkr059 PMID: 21297116
11. Xiong Dapeng, Xiao Fen, Liu Li, Hu Kai, Tan Yanping, He Shunmin, Gao Xieping (2012) Towards a better detection of horizontally transferred genes by combining unusual properties effectively. PLoS ONE 7:e43126 doi: 10.1371/journal.pone.0043126 PMID: 22905214
12. Becq Jennifer, Churlaud Cécile, Deschavanne Patrick (2010) A benchmark of parametric methods for horizontal transfers detection. PLoS ONE 5:e9989 doi: 10.1371/journal.pone.0009989 PMID: 20376325
13. Poptsova Maria (2009) Testing phylogenetic methods to identify horizontal gene transfer. Methods Mol. Biol. 532:227–40 doi: 10.1007/978-1-60327-853-9_13 PMID: 19271188
14. Poptsova Maria S, Gogarten J Peter (2007) The power of phylogenetic approaches to detect horizontally transferred genes. BMC Evol. Biol. 7:45 PMID: 17376230
15. Daubin Vincent, Lerat Emmanuelle, Perrière Guy (2003) The source of laterally transferred genes in bacterial genomes. Genome Biol. 4:R57
16. Lawrence J G, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. U.S.A. 95:9413–7 PMID: 9689094
17. Worning P, Jensen L J, Nelson K E, Brunak S, Ussery D W (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. Nucleic Acids Res. 28:706–9 PMID: 10637321
18. Deschavanne P, Filipski J (1995) Correlation of GC content with replication timing and repair mechanisms in weakly expressed E.coli genes. Nucleic Acids Res. 23:1350–3 PMID: 7753625
19. Wuitschick J D, Karrer K M (1999) Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila. Jk. Eukaryot. Microbiol. 46:239–47 PMID: 10377985
20. Rendulic Snjezana, Jagtap Pratik, Rosinus Andrea, Eppinger Mark, Baar Claudia, Lanz Christa, Keller Heike, Lambert Carey, Katy J Evans Alexander Goesmann, Meyer Folker, Sockett R Elizabeth, Schuster Stephan C (2004) A predator unmasked: life cycle of Bdellovibrio bacteriovorus from a genomic perspective. Science 303:689–92 PMID: 14752164
21. Gophna Uri, Charlebois Robert L, Doolittle W Ford (2006) Ancient lateral gene transfer in the evolution of Bdellovibrio bacteriovorus. Trends Microbiol. 14:64–9
22. Vernikos Georgios S, Thomson Nicholas R, Parkhill Julian (2007) Genetic flux over time in the Salmonella lineage.Genome Biol. 8:R100
23. McCutcheon John P, Moran Nancy A (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol 2:708–18 doi: 10.1093/gbe/evq055 PMID: 20829280
24. Liu Zhandong, Venkatesh Santosh S, Maley Carlo C (2008) Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. BMC Genomics 9:509 doi: 10.1186/1471-2164-9-509 PMID: 18973670
25. Bentley Stephen D, Parkhill Julian (2004) Comparative genomic structure of prokaryotes. Annu. Rev. Genet. 38:771–92 PMID: 15568993
26. Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11:283–90 PMID: 7482779
27. Vernikos Georgios S, Parkhill Julian (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22:2196–203 PMID: 16837528
28. Hooper Sean D, Berg Otto G (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. J. Mol. Evol. 54:365–75
29. Deschavanne P J, 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. 16:1391–9 PMID: 10563018
30. Dufraigne Christine, Fertil Bernard, Lespinats Sylvain, Giron Alain, Deschavanne Patrick (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 33:e6 PMID: 15653627
31. Cortez Diego, Forterre Patrick, Gribaldo Simonetta (2009) A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol. 10:R65

32. Nakamura Yoji, Itoh Takeshi, Matsuda Hideo, Gojobori Takashi (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat. Genet. 36:760–6 PMID: 15208628

33. Ornstein R L, Rein R (1978) An optimized potential function for the calculation of nucleic acid interaction energies I. base stacking. Biopolymers 17:2341–60 PMID: 24624489

34. el Hassan M A, Calladine C R (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. J. Mol. Biol. 259:95–103

35. Olson W K, Gorin A A, Lu X J, Hock L M, Zhurkin V B (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc. Natl. Acad. Sci. U.S.A. 95:11163–8 PMID: 9736707

36. Herzel H, Weiss O, Trifonov E N (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15:187–93 PMID: 10222405

37. Nelson K E, Clayton R A, Gill S R, Gwinn M L, Dodson R J, Haft D H, Hickey E K, Peterson J D, Nelson W C, Ketchum K A, McDonald L, Utterback T R, Malek J A, Linher K D, Garrett M M, Stewart A M, Cotton M D, Pratt M S, Phillips C A, Richardson D, Heidelberg J, Sutton G G, Fleischmann R D, Eisen J A, O White, Salzberg S L, Smith H O, Venter J C, Fraser C M (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323–9 PMID: 10360571

38. Langille Morgan G I, Hsiao William W L, Brinkman Fiona S L (2010) Detecting genomic islands using bioinformatics approaches. Nat. Rev. Microbiol. 8:373–82 doi: 10.1038/nrmicro2350 PMID: 20395967

39. Hacker J, Blum-Oehler G, Mühldorfer I, Tschäpe H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol. Microbiol. 23:1089–97 PMID: 9106201

40. Vernikos Georgios S, Parkhill Julian (2008) Resolving the structural features of genomic islands: a machine learning approach. Genome Res. 18:331–42 PMID: 18071028

41. Altenhoff Adrian M, Dessimoz Christophe (2012) Inferring orthology and paralogy. Methods Mol. Biol. 855:259–79 doi: 10.1007/978-1-61779-582-4_9 PMID: 22407712

42. Than Cuong, Ruths Derek, Innan Hideki, Nakhleh Luay (2007) Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. J. Comput. Biol. 14:517–35

43. Goldman N, Anderson J P, Rodrigo A G (2000) Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–70 PMID: 12116432

44. Shimodaira H, and Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16 (8): 1114.

45. Shimodaira Hidetoshi (2002) An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508 PMID: 12079646

46. Lerat Emmanuelle, Daubin Vincent, Moran Nancy A (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol. 1:E19 PMID: 12975657

47. Zhaxybayeva Olga, Hamel Lutz, Raymond Jason, Gogarten J Peter (2004) Visualization of the phylogenetic content of five genomes using dekapentagonal maps. Genome Biol. 5:R20

48. Beiko Robert G, Harlow Timothy J, Ragan Mark A (2005) Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 102:14332–7 PMID: 16176988

49. Olga Zhaxybayeva, Gogarten J Peter, Charlebois Robert L, Doolittle W Ford, Papke R Thane (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16:1099–108 PMID: 16899658

50. Bansal Mukul S, Guy Banay, Gogarten J Peter, Ron Shamir (2011) Detecting highways of horizontal gene transfer. J. Comput. Biol. 18:1087–114

51. Bansal Mukul S, Guy Banay, Harlow Timothy J, Gogarten J Peter, Shamir Ron (2013) Systematic inference of highways of horizontal gene transfer in prokaryotes. Bioinformatics 29:571–9 doi: 10.1093/bioinformatics/btt021 PMID: 23335015

52. Hallett MT, Lagergren J. RECOMB 2001. Montreal: ACM; 2001. Efficient Algorithms for Lateral Gene Transfer Problems; pp. 149–156.

53. Baroni Mihaela, Stefan Grünewald, Moulton Vincent, Semple Charles (2005) Bounding the number of hybridisation events for a consistent evolutionary history. J Math Biol 51:171–82

54. Beiko Robert G, Nicholas Hamilton (2006) Phylogenetic identification of lateral genetic transfer events. BMC Evol. Biol. 6:15 PMID: 16472400

55. Nakhleh L, Ruths DA, Wang L: RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer. COCOON, August 16–29, 2005; Kunming 2005.

56. Abby Sophie S, Tannier Eric, Gouy Manolo, Daubin Vincent (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics 11:324 doi: 10.1186/1471-2105-11-324 PMID: 20550700

57. Hickey Glenn, Dehne Frank, Rau-Chaplin Andrew, Blouin Christian (2008) SPR distance computation for unrooted trees. Evol. Bioinform. Online 4:17–27 PMID: 19204804

58. Hein, Jotun, Tao Jiang, Lusheng Wang, and Kaizhong Zhang. 1995. On the Complexity of Comparing Evolutionary Trees.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.861.

59. Allen Benjamin L., and Steel Mike. 2001. Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. Annals of Combinatorics 5 (1) (June): 1–15.

60. MacLeod Dave, Charlebois Robert L, Doolittle Ford, Bapteste Eric (2005) Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evol. Biol. 5:27 PMID: 15819979

61. Whidden Christopher, Zeh Norbert, Beiko Robert G (2014) Supertrees Based on the Subtree Prune-and-Regraft Distance. Syst. Biol. 63:566–81 doi: 10.1093/sysbio/syu023 PMID: 24695589

62. Doyon Jean-Philippe, Hamel Sylvie, Chauve Cedric (2011) An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. IEEE/ACM Trans Comput Biol Bioinform 9:26–39

63. David Lawrence A, Alm Eric J (2011) Rapid evolutionary innovation during an Archaean genetic expansion.Nature 469:93–6 doi: 10.1038/nature09649 PMID: 21170026

64. Doyon Jean-Philippe, Hamel Sylvie, Chauve Cedric (2011) An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. IEEE/ACM Trans Comput Biol Bioinform 9:26–39

65. Szöllosi Gergely J, Boussau Bastien, Abby Sophie S, Tannier Eric, Daubin Vincent (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl. Acad. Sci. U.S.A. 109:17513–8 doi: 10.1073/pnas.1202997109 PMID: 23043116

66. Nguyen Thi Hau, Ranwez Vincent, Pointet Stéphanie, Chifolleau Anne-Muriel Arigon, Doyon Jean-Philippe, Berry Vincent (2013) Reconciliation and local gene tree rearrangement can be of mutual profit. Algorithms Mol Biol 8:12 doi: 10.1186/1748-7188-8-12 PMID: 23566548

67. Szöllosi Gergely J, Tannier Eric, Lartillot Nicolas, Daubin Vincent (2013) Lateral gene transfer from the dead. Syst. Biol. 62:386–97 doi: 10.1093/sysbio/syt003 PMID: 23355531

68. Bansal Mukul S, Alm Eric J, Kellis Manolis (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28:i283–91 doi: 10.1093/bioinformatics/bts225 PMID: 22689773

69. Majewski J, Zawadzki P, Pickerill P, Cohan F M, Dowson C G (2000) Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. J. Bacteriol. 182:1016–23 PMID: 10648528

70. Sjöstrand Joel, Tofigh Ali, Daubin Vincent, Arvestad Lars, Sennblad Bengt, Lagergren Jens (2014) A Bayesian method for analyzing lateral gene transfer. Syst. Biol. 63:409–20 doi: 10.1093/sysbio/syu007 PMID: 24562812

71. Szöllősi Gergely J, Rosikiewicz Wojciech, Boussau Bastien, Tannier Eric, Daubin Vincent (2013) Efficient exploration of the space of reconciled gene trees. Syst. Biol. 62:901–12 doi: 10.1093/sysbio/syt054 PMID: 23925510

72. Haggerty Leanne S, Jachiet Pierre-Alain, Hanage William P, Fitzpatrick David A, Lopez Philippe, O'Connell Mary J, Pisani Davide, Wilkinson Mark, Bapteste Eric, McInerney James O (2014) A pluralistic account of homology: adapting the models to the data. Mol. Biol. Evol. 31:501–16 doi: 10.1093/molbev/mst228 PMID: 24273322

73. Szöllősi Gergely J, Tannier Eric, Daubin Vincent, Boussau Bastien (2014) The Inference of Gene Trees with Species Trees. Syst. Biol. 64: e42–e62. doi: 10.1093/sysbio/syu048 PMID: 25070970

74. Koski L B, Golding G B (2001) The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52:540–2 PMID: 11443357

75. Wisniewski-Dyé Florence, Borziak Kirill, Khalsa-Moyers Gurusahai, Alexandre Gladys, Sukharnikov Leonid O, Wuichet Kristin, Hurst Gregory B, McDonald W Hayes, Robertson Jon S, Barbe Valérie, Calteau Alexandra, Rouy Zoé, Mangenot Sophie, Prigent-Combaret Claire, Normand Philippe, Boyer Mickaël, Siguier Patricia, Dessaux Yves, Elmerich Claudine, Condemine Guy, Krishnen Ganisan, Kennedy Ivan, Paterson Andrew H, González Victor, Mavingui Patrick, Zhulin Igor B (2011) Azospirillum genomes reveal transition of bacteria from aquatic to terrestrial environments. PLoS Genet. 7: e1002430

76. Zuckerkandl E. and Pauling L.B. 1965. Evolutionary divergence and convergence in proteins. In Bryson V. and Vogel H.J. (editors). Evolving Genes and Proteins. Academic Press, New York. pp. 97–166.

77. Novichkov Pavel S, Omelchenko Marina V, Gelfand Mikhail S, Mironov Andrei A, Wolf Yuri I, Koonin Eugene V (2004)Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J. Bacteriol. 186:6575–85 PMID: 15375139

78. Lawrence J G, Hartl D L (1992) Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics 131:753–60

79. Clarke G D Paul, Beiko Robert G, Ragan Mark A, Charlebois Robert 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 PMID: 11914337

80. Pellegrini M, Marcotte E M, Thompson M J, Eisenberg D, Yeates T O (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U.S.A. 96:4285–8 PMID: 10200254

81. Normand Philippe, Lapierre Pascal, Tisa Louis S, Gogarten Johann Peter, Alloisio Nicole, Bagnarol Emilie, Bassi Carla A, Berry Alison M, Bickhart Derek M, Choisne Nathalie, Couloux Arnaud, Cournoyer Benoit, Cruveiller Stephane, Daubin Vincent, Demange Nadia, Francino Maria Pilar, Goltsman Eugene, Huang Ying, Kopp Olga R, Labarre Laurent, Lapidus Alla, Lavire Celine, Marechal Joelle, Martinez Michele, Mastronunzio Juliana E, Mullin Beth C, Niemann James, Pujic Pierre, Rawnsley Tania, Rouy Zoe, Schenowitz Chantal, Sellstedt Anita, Tavares Fernando, Tomkins Jeffrey P, Vallenet David, Valverde Claudio, Wall Luis G, Wang Ying, Medigue Claudine, Benson David R (2007) Genome characteristics of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography. Genome Res. 17:7–15

82. Welch R A, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles E L, Liou S-R, Boutin A, Hackett J, Stroud D, Mayhew G F, Rose D J, Zhou S, Schwartz D C, Perna N T, Mobley H L T, Donnenberg M S, Blattner F R (2002)Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 99:17020–4 PMID: 12471157

83. Cs rös M. 2008. Ancestral Reconstruction by Asymmetric Wagner Parsimony over Continuous Characters and Squared Parsimony over Distributions. In Comparative Genomics (eds. Nelson C.E. and Vialette S.), Lecture Notes in Computer Science, pp. 72–86, Springer Berlin Heidelberg http://link.springer.com/chapter/10.1007/978-3-540-87989-3_6

84. Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877–84 PMID: 10553904

85. Miklós Csurös, István Miklós (2009) Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol. Biol. Evol. 26:2087–95 doi: 10.1093/molbev/msp123 PMID: 19570746

86. Hao Weilong, Golding G Brian (2010) Inferring bacterial genome flux while considering truncated genes. Genetics 186:411–26

87. Hao Weilong, Golding G Brian (2006) The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16:636–43 PMID: 16651664

88. Hao Weilong, Golding G Brian (2008) Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics 9:235

89. Ochman H, Lawrence J G, Groisman E A (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304 PMID: 10830951

90. Papke R Thane, Koenig Jeremy E, Rodríguez-Valera Francísco, Doolittle W Ford (2004) Frequent recombination in a saltern population of Halorubrum. Science 306:1928–9

91. Mau Bob, Glasner Jeremy D, Darling Aaron E, Perna Nicole T (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. Genome Biol. 7:R44

92. Didelot Xavier, Falush Daniel (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–66 PMID: 17151252

93. Ragan M A (2001) On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. 201:187–91 PMID: 11470360

94. Ragan Mark A, Harlow Timothy J, Beiko Robert G (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends Microbiol. 14:4–8

95. Kechris Katherina J, Lin Jason C, Bickel Peter J, Glazer Alexander N (2006) Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. Proc. Natl. Acad. Sci. U.S.A. 103:9584–9 PMID: 16769896

96. Moran Nancy A, Jarvik Tyler (2010) Lateral transfer of genes from fungi underlies carotenoid production in aphids.Science 328:624–7 doi: 10.1126/science.1187113 PMID: 20431015

97. Danchin Etienne G J, Rosso Marie-Noëlle, Vieira Paulo, de Almeida-Engler Janice, Coutinho Pedro M, Henrissat Bernard, Abad Pierre (2010) Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. U.S.A. 107:17651–6 doi: 10.1073/pnas.1008486107 PMID: 20876108

98. Fletcher William, Yang Ziheng (2009) INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–88

99. Sipos Botond, Massingham Tim, Jordan Gregory E, Goldman Nick (2011) PhyloSim—Monte Carlo simulation of sequence evolution in the R statistical computing environment. BMC Bioinformatics 12:104

100. Galtier Nicolas (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. Syst. Biol. 56:633–42 PMID: 17661231

101. Dalquen Daniel A, Anisimova Maria, Gonnet Gaston H, Dessimoz Christophe (2012) ALF—a simulation framework for genome evolution. Mol. Biol. Evol. 29:1115–23 doi: 10.1093/molbev/msr268 PMID: 22160766

102. Cortez Diego Q, Lazcano Antonio, Becerra Arturo (2005) Comparative analysis of methodologies for the detection of horizontally transferred genes: a reassessment of first-order Markov models. In Silico Biol. (Gedrukt) 5:581–92 PMID: 16610135

103. Tsirigos Aristotelis, Rigoutsos Isidore (2005) A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res. 33:922–33 PMID: 15716310

104. Azad Rajeev K, Lawrence Jeffrey G (2005) Use of artificial genomes in assessing methods for atypical gene detection. PLoS Comput. Biol. 1:e56 PMID: 16292353

105. Iantorno Stefano, Gori Kevin, Goldman Nick, Gil Manuel, Dessimoz Christophe (2014) Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. Methods Mol. Biol. 1079:59–73 doi: 10.1007/978-1-62703-646-7_4 PMID: 24170395