Horizontal gene transfer among microbial genomes: new insights from complete genome analysis

Jonathan A Eisen

The determination and analysis of complete genome sequences has led to the suggestion that horizontal gene transfer may be much more extensive than previously appreciated. Many of these studies, however, rely on evidence that could be generated by forces other than gene transfer including selection, variable evolutionary rates, and biased sampling.

Introduction

‘The view commonly entertained by naturalists is that species, when intercrossed, have been specially endowed with sterility, in order to prevent their confusion. This view certainly seems at first highly probable, for species living together could hardly have been kept distinct had they been capable of freely crossing.’

Charles Darwin, On the Origin of Species.

Analysis of complete genome sequences is providing insights into many issues concerning microbe evolution. One such area involves the transfer of genetic material between distinct evolutionary lineages — a phenomena known as horizontal, or lateral, gene transfer. (I prefer the use of ‘transfer’ instead of ‘exchange’ to refer to the process as, in most known cases, the process is unidirectional and rarely involves reciprocal exchanges of DNA. In addition, although the process can involve many genes or parts of genes, for the sake of simplicity I refer to it simply as ‘gene transfer.’) This contrasts with what is considered the ‘normal’ process of inheritance — the transmission of traits from parents to offspring, also known as vertical inheritance. The occurrence of horizontal gene transfer (HGT), which blurs the boundaries between species even more than the hybridization referred to by Darwin quoted above, has been generally accepted for many years [1]. Examples include transfers from organellar to nuclear genomes in eukaryotes and of plasmids between bacterial species. Analysis of complete genome sequences has led to suggestions that the extent of HGT is far greater than was previously appreciated [2,3,4–9]. The reliability of some of these estimates, however, has been called into question recently by many researchers [10**]. In this review, I discuss the methods that have been used to infer past occurrences of HGT from analysis of genome sequences, and then some of the recently proposed examples.

Why study horizontal gene transfer?

There are a variety of reasons for better understanding the type and extent of HGT. As HGT results in chimeric species — that is, organisms in which different portions of the genome have different histories — attempts to infer the evolutionary history of species must take such transfer into account. Understanding HGT is also valuable in studying evolutionary processes such as the origins of new functions. In addition, HGT apparently plays an active role in many biological processes including the emergence and spread of virulence and resistance to antibiotics and the long-term maintenance of organelles [11]. The inference of HGT can also be useful in predicting gene functions, an integral part of most genome-sequencing projects. For example, predicting the targetting of nuclear genes to organelles can be aided by determining if those genes have an organellar ancestry.

Steps in horizontal gene transfer

HGT is more than just the transfer of genes — it should really be considered a multi-step process (Figure 1). Understanding HGT and the design of methods to detect it requires an understanding of all of these steps. I summarize some of the important parts of each step here. First, the gene to be transferred evolves within its donor lineage (Figure 1a). Two types of evolution are of particular importance to the study of HGT: co-evolution with other genes in the genome, and incorporation of genome-wide features of the donor lineage such as codon usage, GC content, splicing signals, and promoters (referred to here as ‘genome phenotype’). At some point, the gene is then transferred to another lineage either via a vector (e.g. a virus) or by direct (e.g. mating) or indirect (e.g. competence) DNA exchange (Figure 1b). The gene must also be in or get into a format that allows long-term maintenance and replication (e.g. either as a self-replicating extrachromosomal element or by insertion into existing elements) (Figure 1c). The mechanisms of transfer and maintenance influence the patterns of HGT. For example, mismatch repair systems in some species prevent homologous recombination based insertions of distantly related genes but have no effect on insertion via non-homologous recombination [12]. A frequently overlooked component of HGT is the spread of a transferred gene in the recipient population (Figure 1c). Although this spread can occur by neutral means, it is generally thought to be driven by selective forces (e.g. selection for antibiotic resistance). For the
Methods for inferring horizontal gene transfer

A variety of methods have been developed to infer the occurrence of HGT. The utility of each method depends, in part, on the nature of all the steps in HGT that have occurred. In the following sections I discuss these methods as well as some of their strengths and weaknesses.

Comparing phylogenetic trees of different genes

One method that has been used to determine whether HGT has occurred involves inferring evolutionary trees for many genes in many genomes. HGT should cause different genes to have different trees. It is important to recognize that phylogenetic reconstruction methods are not perfect. Many factors in addition to HGT can lead to differences in inferred trees of different genes, including poor data; misuse of phylogenetic reconstruction methods; inaccurate alignments [14]; convergence; the misidentification of orthologs versus paralogs; and different species sampling for different genes [15]. Even when used correctly, phylogenetic reconstruction methods do not always produce the correct history [16]. In addition, phylogenetic methods do not always work well in inferring exchanges among very closely related species. Therefore there are valid reasons for using other approaches to infer HGT. Nevertheless, phylogenetic reconstruction methods remain the only way to reliably infer historical events from gene sequences. In addition, all the methods used to infer HGT, phylogenetic methods are the only ones that are based on a large body of work. For example, phylogenetic methods are designed to accommodate variation in evolutionary rates and patterns within and between taxa. It is unfortunate that many studies of HGT have completely ignored the uses of phylogenetic reconstruction.

Patterns of best matches to different species

A common alternative method for inferring HGT involves using similarity search techniques to determine the ‘best match’ for each gene in a genome. HGT is frequently invoked for those genes that have best matches to supposedly distant species. The main advantages of these best-match methods are their speed and automatability — but their limitation is accuracy. For example, evolutionary rate variation can lead to two distantly related genes being detected as best matches (if the rate of evolution was very slow in their lineages) or two closely related genes not matching well (if they have evolved rapidly) [17]. In addition, there has not yet been any work that has determined how many unusual best matches one might expect by chance alone. Other problems with best-match methods include being easily misled by multidomain proteins, ignoring the possibility of gene loss, and not allowing reliable identification of orthology. Finally, this method is also significantly biased by genome size in that the number of best matches to a particular species is dependent in part on the total number of open reading frames in that species and not just evolutionary relatedness [18].

Distribution patterns of genes

Lateral gene transfer can result in the addition of novel genes in a particular species — thus, in theory, gene transfers could be identified by detecting genes with uneven distribution patterns (e.g. if a particular gene is present in gram-positive bacteria and all plants but not other bacteria or eukaryotes). Although such an approach can be effective, it is frequently misleading. For example, uneven distribution patterns can also be caused by gene loss or rapid sequence divergence (such that homologs are not detected in some lineages even when they are present). In addition, analysis of distribution patterns cannot be used to infer transfers involving homologous recombination of parts of genes.

Identifying regions of the genome with unusual compositions

As discussed above, there is significant uniformity in genome phenotype within genomes. As it takes time for a transferred gene to ameliorate to the recipient genome’s phenotype, ‘foreign’ genes in a genome can be detected by identifying genes with unusual phenotypes (e.g. nucleotide composition or codon usage) [4,13,19]. An advantage of this approach is that it only requires the genome sequence of
one species and is not prone to the problems of phylogenetic reconstruction. A disadvantage is that unusual compositions can also be caused by factors other than HGT, such as selection, mutation bias, and even the direction of transcription relative to the replication origin [20**]. The use of compositional analysis to detect HGT requires detailed examination of multiple features for any particular genome to identify which characteristics are most likely caused by HGT [21]. This method cannot detect transfers between species with similar compositions nor those that have occurred a long enough time ago such that amelioration has been completed. In some cases, unusual composition may be a reflection more of the biology of vectors (e.g. phage and plasmid replication) and not adaptation to a particular host genome.

Combined approaches
As each type of evidence commonly used to infer HGT can be explained by other forces (see Table 1), it can also be useful to combine multiple lines of evidence. A good example of a combined approach comes from studies of gene transfer from chloroplast genomes to nuclear genomes [22]. This analysis has included the use of distribution patterns to identify cases of gene loss from organelles and corresponding gain in nuclear genomes; phylogenetic reconstruction methods to prove that these nuclear genes are evolutionarily derived from organelar genomes and composition analysis to document the amelioration of these genes with the nuclear genome. The limitation of combined approaches is that each method is designed to detect transfers of different types and ages.

Examples of horizontal gene transfer proposed on the basis of genome analysis
Thermophilic bacteria and Archaea
Analysis of the complete genomes of the bacteria *Aquifex aeolicus* and *Thermotoga maritima* and comparison with other complete genomes has led to the suggestion that large numbers of genes have been transferred between thermophilic bacteria and Archaea [23,24*]. In both cases, best-match methods have revealed that a high percentage of each proteome was most similar to Archaeal genes rather than bacterial genes — ~20% for *A. aeolicus* [23] and 25% for *T. maritima* [24*]. As indicated above, however, best matches can be misleading. The high percentage of Archaeal-like genes in these genomes could be caused by any of the following: a high rate of evolution in the mesophilic bacteria, the loss of these genes from mesophilic bacteria, or convergence [10**]. The authors present additional evidence supporting the possibility of HGT of some of the Archaeal-like genes including clustering in the genome, nucleotide-composition anomalies, and evolutionary trees showing some genes branching next to Archaeal genes [23,24*,25] but the results remain inconclusive. For example, a large fraction of the Archaeal-like genes had no additional support for transfer, indicating that the best-match method may be significantly misleading [26].

Organellar to nuclear transfers
The understanding of organellar to nuclear gene transfers has been greatly increased in the past few years with the completion of more genome sequences, both organelles and eukaryotes [22,27,28]. For example, tracing gene loss from chloroplast genomes [29] is now being supplemented with analysis of nuclear genome of plants, which will allow the identification of genes with chloroplast origins [30*]. Additional complete chloroplast genomes are also helping in this analysis [31–33].

Transfers involving pathogen genomes
A variety of studies have suggested the occurrence of extensive HGT in the history of many of the pathogen species for which complete genome sequences are now available. For example, analysis of nucleotide composition and codon usage has identified many regions of possible HGT in *Escherichia coli* (up to 18% of the genome) [4]. It will be interesting to repeat this type of analysis with pathogenic strains of *E. coli* and to try to identify the source of

### Table 1
Methods of detecting horizontal gene transfer.

<table>
<thead>
<tr>
<th>Observation</th>
<th>For what HGTs does it not occur?</th>
<th>Other possible causes</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unusual distribution patterns.</td>
<td>If recipient already has gene.</td>
<td>Sampling bias, gene loss, rapid divergence and ancestral polymorphisms.</td>
<td>Only works for HGT of novel genes.</td>
</tr>
<tr>
<td>Unusual nucleotide composition</td>
<td>If donor/recipient similar. If transfer was a long time ago.</td>
<td>Selection and mutation bias. Replication direction.</td>
<td>Only need a single genome. Rapid.</td>
</tr>
<tr>
<td>(e.g. codon usage, GC content).</td>
<td></td>
<td>Selection. Rate variation. Gene loss.</td>
<td>Fast but error prone.</td>
</tr>
<tr>
<td>High hit to distant species.</td>
<td>If transfer is among close relatives.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incongruent phylogenetic trees.</td>
<td>If transfer is among close relatives.</td>
<td>Bad trees. Missed paralogs.</td>
<td>Accurate but slow.</td>
</tr>
<tr>
<td>Clustering of any of the above</td>
<td>If genes were not transferred together. If gene order has been rearranged.</td>
<td>Selection.</td>
<td></td>
</tr>
<tr>
<td>in genome.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

---
this foreign DNA. Genome analysis suggests that a large percentage of the Xylella fastidiosa genome (~7%) appears to be derived from λ-like phage [34]. The finding that only a modest proportion of the genes in the Campylobacter jejuni genome had matches to genes from other e-Proteobacteria, led to the suggestion that this species may have acquired many genes by HGT [35]. In contrast, genome analysis indicates that HGT appears to have played little role in shaping Chlamydial genomes: most of the differences within and between C. pneumoniae and C. trachomatis strains involve chromosomal rearrangements and sequence divergence [36,37].

Genome analysis has also lent support to one of the more interesting possible cases of HGT — that between the Haemophilus influenzae and Neisseria meningitidis lineages. This would be interesting because although these species are only moderately closely related, they occupy similar niches in the human ecosystem. The possibility of HGT between these two species was suggested initially on the basis of individual gene analysis [38]. Analysis of the complete genome of N. meningitidis (best matches and distribution of uptake signal sequences) supports this possibility [18,39].

Transfers involving whole genetic elements (small chromosomes, megaplasmids, plasmids)
Analysis of complete genomes is revealing that many of the smaller genetic elements in microbes may have separate origins from the main chromosomes. For example, the linear plasmids of Borrelia appear to be prone to HGT among strains or species [40]. Other proposed cases of discordant evolution of smaller genetic elements versus chromosomes include the Xylella fastidiosa plasmid [34], the megaplasmids and plasmid of D. radiodurans [41] and the smaller chromosome of V. cholerae [42]. The source of these smaller elements in each species remains unclear. For D. radiodurans and V. cholerae, any transfer probably occurred long ago because the smaller genetic elements have undergone extensive exchange with the main chromosome in each species. Interestingly, phylogenetic analysis suggests that the megaplasmids of Rhizobia are only transferred among closely related strains or species and not across long distances [43].

Is there a core recalcitrant to transfer?
Despite the apparent occurrence of extensive gene transfers in the history of microbes, it appears that there is a ‘core’ to each evolutionary lineage that is recalcitrant to transfer. Recent evidence for this comes from the construction of ‘whole-genome trees’ that are based on analysis of the presence and absence of particular homologs or orthologs in different complete genomes [44]. Although these are not really trees, they do represent a measure of the average of the evolutionary patterns including gene loss and transfer as well as phylogeny — of all genes in a genome. That these average trees are nearly identical to the trees of rRNA genes suggests that those HGT events that occur are either constrained by phylogenetic relationships indicated by the rRNA tree or only occur for a moderate portion of the genome. Initial studies indicate that both appear to be true. For example, phylogenetic analysis suggests that the more universal genes can be exchanged within [45] but not between major microbial lineages [23]. The fact that these rRNA lineages share many genes [45,46] supports the core concept, as does evidence that only particular classes of genes are prone to transfer (see below).

What types of genes are transferred?
Analysis of complete genome sequences is helping determine what types of genes are most prone to transfer. Two recent studies [45,47] using phylogenetic analysis of many genes, conclude that genes which interact with many other genes (informational genes in [47]) are less prone to gene transfer than those with fewer interactions (operational genes in [47]). This was referred to as the ‘complexity hypothesis’ [47]. It is still not clear if complexity of interactions is the key factor here. For example, it is possible that sequence conservation of genes between species also influences the likelihood of transfer [25], which is probably what allows even rRNA genes to be transferred [48,49]. In addition, because informational genes are more likely to be required for survival, they are probably less likely to be lost from a species. Therefore, there will be little opportunity to replace them with a homologous gene. In addition, some genes may even facilitate their own transfer. For example, genes could avoid the need for amelioration by carrying the features needed for their own expression. Possible examples include self-regulatory transcription factors (which could bind to their own promoter), amino-acyl tRNA synthetases (which have been shown to be prone to transfer [50,51] and could help avoid codon usage problems), splicing factors, recombinases, and DNA polymerases.

Conclusions and future directions
The analyses described here, as well as many others not reported for space reasons, are beginning to reveal a great deal about HGT and its influence of gene and genome evolution. Some of the more interesting proposals include the following: that HGT may be more important than mutation in evolving new functions [4], that the tree of life may be unresolvable because of the extent of HGT [3,4], and that what we consider phylogeny may be defined in a large part by HGT [2]. However interesting these proposals are, I recommend caution before jumping on the HGT bandwagon. In most cases, the proposals for HGT involve only one type of evidence, such as nucleotide composition bias or unusual distribution patterns. As each of these can be caused by factors other than HGT, it will be important to see if they are supported by other lines of evidence. Furthermore, in most of the proposed cases, there is little information regarding what the likely donor lineage was or when in the course of evolution the transfer likely occurred. Even when donors and recipients have been proposed, there is rarely supporting
evidence regarding the absence of the genes from relatives of the recipient lineage that diverged prior to the transfer. For a complete picture of any proposed case of HGT it will be important to have information regarding the vectors, what, if any, selective forces were involved in the transfer, and what was the extent of amelioration.

With all these caveats aside, it seems likely that HGT has been frequent in the course of microbial evolution. It remains to be seen just how frequent HGTs have been and what the rules that govern the whole process are. As we learn more about the mechanisms and rules of HGT, we will be better equipped to model the process and develop more rigorous tests of its occurrence. Only then will we be able to determine the role HGT has played in shaping microbial evolution and biology.

Acknowledgements
I thank all the people at TIGR for support, assistance, and helpful discussions.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:
• of special interest
** of outstanding interest
Discussion of how the concept of classification of species may need to be revised in light of extensive horizontal gene transfer.
A review of HGT in microorganisms and the difficulties that extensive gene transfer presents to attempts to study the evolution of species. Outlines the need to search for principles governing genetic exchange.
Excellent commentary on the pitfalls of using sequence similarity methods to infer HGT.
Horizontal gene transfer among microbial genomes

Eisen 611


A detailed comparison of four closely related genomes including two previously published genomes. Shows how comparisons of closely related species can be used to identify hypervariable loci. Most remarkably, they find extensive polymorphisms in the library used for sequencing showing that the mutation rate in this species, especially at microsatellite loci, is very high.


One of the first attempts to build a ‘whole genome tree’. The authors show that a tree based on gene presence/absence is very similar to the rRNA tree for the same species, suggesting that there is some average phylogeny of a genome that is accurately reflected in the rRNA of life.


