PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2009) **364**, 2209–2219
doi:10.1098/rstb.2009.0042

Gene and genome trees conflict at many levels

Leanne S. Haggerty[1,†], **Fergal J. Martin**[1,†], **David A. Fitzpatrick**[2] **and James O. McInerney**[1,*]

[1] *Department of Biology, The National University of Ireland, Maynooth, County Kildare, Ireland*
[2] *Conway Institute, University College Dublin, Ireland*

Horizontal gene transfer (HGT) plays a significant role in microbial evolution. It can accelerate the adaptation of an organism, it can generate new metabolic pathways and it can completely remodel an organism's genome. We examine 27 closely related genomes from the YESS group of gamma proteobacteria and a variety of four-taxon datasets from a diverse range of prokaryotes in order to explore the kinds of effects HGT has had on these organisms.

Keywords: phylogeny; prokaryotes; conflict

1. INTRODUCTION

The amount of horizontal gene transfer (HGT) that is observable in completed genomes makes it impossible to define a single unifying phylogenetic tree that can describe the evolutionary history of all prokaryotes. HGT confuses and confounds prokaryotic relationships by implying different, incongruent relationships within a set of taxa. The set of relationships derived for a particular gene is a combination of both the vertical and the horizontal history for that gene. Alongside that, it is impossible to find a species concept or definition that will please everybody. What constitutes a species barrier between prokaryotic genomes? When should two prokaryotic genomes be considered to be different species, as opposed to simply different strains of the same species? In this study, we investigate the evolutionary history between a group of closely related prokaryotic genomes using a variety of methods and filters, to determine if the genomes can be separated into distinct species in the face of HGT and high levels of sequence similarity. This will be an important aspect of future phylogenetic studies; prokaryotic genomes are now relatively inexpensive and easy to sequence. The level of sampling we have for the different prokaryotic 'species' is rapidly increasing. Moving forward, it is important to adapt our strategies for determining and classifying these relationships.

Recently, we have questioned whether or not there is a future for the Tree of Life metaphor (Dagan & Martin 2006). Many have gone further and feel that the time has long since gone when this metaphor was useful (Doolittle & Bapteste 2007). The central issue is that HGT has affected all or nearly all genes in every genome at one stage in their evolutionary history (Dagan & Martin 2007; Dagan *et al.* 2008). The most recent estimate is that in each genome an average of 81 ± 15 per cent of the genes have experienced an HGT event at some stage (Dagan & Martin 2007). In the next few years, we must precisely describe how the prokaryotic world, in particular, is structured and what exactly HGT has done.

There are two categories of HGT events: homology dependent and homology independent (though the most important factor is similarity level, not whether the sequences are homologous). Homologous recombination, according to Ochman *et al.* (2005), occurs mainly within a bacterial species, but there is very little recombination (approx. 1%) between any given species and its close relatives. However, the process of non-homologous recombination or the introduction of new genes that have no similarity to incumbent genes is mostly a process that involves organisms that we consider to be very far outside the species boundary. In referring to non-homologous recombination, we also encompass recombination events where regions with no significant similarity to anywhere in the recipient genome are carried into that genome by flanking regions that do have similarity to the recipient genome. Lawrence (2002) has put forward the theory that integration of foreign non-homologous DNA into a genome is a driver of speciation in prokaryotes, and this is a testable hypothesis.

On the question of what boundaries might exist that prevent a gene from being successfully incorporated into a recipient genome, Sorek *et al.* (2007) have indicated that gene dosage and promoter structure might be barriers. By contrast, it has been our opinion (McInerney & Pisani 2007) that the barriers to HGT, if they exist, might be very low. However, these opinions relate to the artificial scenario where barriers to HGT have been measured *in vitro.*

In this article, we have taken an exemplar densely sampled set of genomes, from the YESS group of prokaryotes, and examined a number of methods that are routinely used in order to infer phylogenies. We want to explore what might happen if trees of genomes or subsets of genomes are inferred. Additionally, we have explored the apparent rate of HGT for a diversity of organisms.

[*] Author for correspondence (james.o.mcinerney@nuim.ie).
[†] These authors have contributed equally to this work.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2009.0042 or via http://rstb.royalsocietypublishing.org.

One contribution of 11 to a Theme Issue 'The network of life: genome beginnings and evolution'.

(a) *What is a bacterial species?*

What seems indisputable is that we can identify organisms that have synapomorphies, both genetic and phenotypic. However, even though we recognize groupings, we do not have a bacterial species concept and we do not understand how these groupings (species, subspecies, even genera) form. Multi-locus sequence analysis (Gevers *et al.* 2005) has shown that there is some structure among currently defined species (Falush *et al.* 2001; Kidgell *et al.* 2002; Achtman & Wagner 2008, Buckee *et al.* 2008). However, this kind of analysis, which has been carried out extensively in thousands of isolates, has the limitation that it only examines the evolutionary history of a set of core genes. Not only does this limit the amount of information used in the analysis, core genes are not representative of the rest of the genes in a genome in terms of factors such as functional category and rate mutation. For a modern system of classification to work, it must use complete genomes and be able to accommodate HGT.

The concept of prokaryotic species is difficult to address, and there is considerable diversity of opinion on what constitutes a species among the prokaryotes. HGT might be considered to be a form of sex and, therefore, all prokaryotes might be considered to be a single species. Alternatively, we might consider a species to be an 'irreducible cluster' of organisms (Staley 2006), and this seems in many ways to be sensible. Staley has advanced the idea that we might use a genomic-phylogenetic species concept (Staley 2006). Doolittle has suggested that if a species concept is not needed, we should let it go, whereas if it can be found, it might be useful (Doolittle & Papke 2006; Papke *et al.* 2007).

At the moment, we have a polyphasic *definition* of a bacterial species. Depending on the data that are available, this polyphasic definition can involve the use of ribosomal RNA sequence identity, reciprocal DNA–DNA reassociation values, biochemical traits and so forth. In this paper, we take as an example the YESS group of γ-proteobacteria and we examine what kinds of phylogenetic signals we get when we use different parts of genomes, different genes and different analysis methods. Naturally, if analysis methods are consistent and efficient, the phylogenetic signals are congruent and the amount of data that are available is sufficient, we should inevitably get the same answer. However, we know that HGT exists, so the question is 'what kind of answer will we get when we use different datasets and different methods of analysis?'.

While much of the focus on the issue of HGT has been on the long-term evolutionary history of prokaryotes, a number of studies have examined shallower relationships. Ochman *et al.* (2005) analysed HGT at the shallower taxonomic levels and concluded that while there was relatively frequent HGT between homologous genes within species, there was a much lower amount of HGT between homologues across the species boundary. Given that new genomes are being sequenced on a daily basis, we can examine what this structure means for microbiology. In particular, this might have an important consequence for our concept of a bacterial species.

If there is a valid, biological bacterial species concept, we might in the future ask what drives speciation; so from a number of perspectives, it is interesting to explore evolution at the boundaries of recognized species.

(b) *A test dataset for exploring groups of genomes*

The exemplar group we have chosen to study consists of *Yersinia*, *Escherichia*, *Salmonella* and *Shigella*, sometimes termed the YESS group of γ-proteobacteria. These are facultatively anaerobic Gram-negative rod-shaped bacteria that are catalase positive and oxidase negative (Brenner 1984). The YESS group is of particular interest as many members are human pathogens. For instance, *Yersinia pestis* was the causative agent of the bubonic plague that killed an estimated 75 million worldwide during the 1300s. *Shigella* and enteroinvasive *Escherichia coli* (EIEC) are the aetiological agents of bacillary dysentery or shigellosis, of which there are an estimated 160 million cases worldwide a year, with approximately 1.1 million deaths, mainly in children under the age of five (Kotloff *et al.* 1999). *Salmonella* infection, known as salmonellosis, induces vomiting, diarrhoea, fever and abdominal cramps and can last several days. Outbreaks of YESS group-associated diseases are common (Tacket *et al.* 1985; Mahon *et al.* 1997; Lee *et al.* 2000; Varma *et al.* 2003), and consequently at the time of writing, this group of prokaryotes is the most extensively sampled in genome sequencing projects, making it useful to illustrate the points we wish to make.

The phylogenetic relationships of different *Shigella* strains have been the subject of intense debate in recent years. Joshua Lederberg famously said that Enterohaemorrhagic *E. coli* were '*Shigella* in a little cloak of *E. coli* antigens'. *Shigella* are essentially *E. coli* that have acquired a virulence plasmid (VP) (Sansonetti *et al.* 1981; Lan *et al.* 2001). There are two conflicting theories on the origin of *Shigella*. The multiple independent origin theory (Pupo *et al.* 2000) suggested that *Shigella* strains formed through multiple acquisitions of the VP. The analysis of Pupo *et al.* found three clusters of *Shigella* strains occurring within *E. coli* and concluded that *Shigella* strains, much like EIEC, do not have a single evolutionary origin. Later it was argued that there was a single origin of *Shigella* (Escobar-Paramo *et al.* 2003). The argument was based upon similarities between the phylogenies of genes on the VP with phylogenies for chromosomal genes. Because the phylogenies did not conflict significantly, Escobar-Paramo *et al.* (2003) suggested that there was a single ancestral VP that accounted for the emergence of *Shigella* and that the VP has not been horizontally transferred (as the multiple origins theory would imply). Any conflicts in the trees were said to be accounted for by transfer of fragments of the VP as opposed to the transfer of an entire VP. More recently, in 2007, Yang *et al.* (2007) revisited the two hypotheses using more robust data and found support for the multiple origin hypothesis. Like Pupo *et al.*, they found three major clusters of

Shigella. They concluded that ancestral VPs entered various strains of *E. coli* and that convergent evolution explains why we see diverse *Shigella* genomes with similar phenotypic properties.

The issue that we see with *Shigella* and *E. coli* typifies the problem that we have with microbiology in the age where we still have not routinely sequenced whole genomes for many isolates in order to assist in defining a species. Given a phylogeny based on a single gene we might make one kind of inference, whereas a phylogeny based on a different gene might result in an entirely different inference. Soon, we will routinely sequence completed genomes from multiple strains of the same species. This means that if we want to use a genome-assisted phylogenetic scheme (and indeed, we probably should), then we need to understand the signals we see in genomes.

(c) *Many methods and many data types*

Phylogenies based on housekeeping genes such as *gyrB*, *tufA* and *atpD* are often compared with those based on 16S rRNA phylogenies (Dauga 2002; Purkhold *et al.* 2003; Paradis *et al.* 2005). The goal of comparing genes is to examine linkage disequilibrium or recombination or to overcome systematic biases (Cooper & Feil 2004) that might be present in one molecule and not in another. Methodological problems that are encountered during phylogenetic analysis include artifacts related to both molecular and lineage-specific differences in evolutionary rates and mutational saturation (Doolittle 1999). These processes can sometimes be detected and if an appropriate model of sequence evolution is available, they can be overcome (Rodriguez-Ezpeleta *et al.* 2007). It has been shown that HGT can occur in genes that have been cited as unlikely candidates, including ribosomal proteins (O'Neil *et al.* 1969). One study has even shown that it is possible to replace the 16S rRNA of *E. coli* with the corresponding sequence from *Proteus vulgaris*, though there is an associated drop in growth rate of between 10 and 30 per cent (Asai *et al.* 1999).

The technique of data concatenation is often used in order to reconstruct phylogenetic relationships (Sanderson *et al.* 2003). This usually involves multiple gene sequences being concatenated and aligned as a single sequence. Using this greater number of genes is supposed to bring out the true phylogenetic relationships, the theory being that signal, even when it is weak, is cumulative, whereas homoplastic noise will be dispersive (Sanderson *et al.* 2003). However, in general data concatenation is usually based on small sets of genes. For example Ciccarelli *et al.* (Ciccarelli *et al.* 2006) used only 31 genes, or less than 1 per cent of the genes in the average genome (Dagan & Martin 2006), in their data set, to determine the relationships for 191 species. Also, data concatenation can sometimes produce misleading results. Rokas *et al.* (2003) claimed to have found the correct species tree for eight yeast genomes using data concatenation. The concatenated data, consisting of 106 nuclear genes, resulted in maximum bootstrap support for a single topology. Later, this dataset was re-analysed by Phillips *et al.* (2004) who found that by using a different method of analysis a different tree, also with maximum bootstrap support could be found, and that longer sequences (typical of those used in data concatenation) exacerbate the potential for systematic error in phylogenetic analysis. This is not the fault of concatenation *per se*; however, concatenation generally leads to long sequences, so this is an importation factor to consider when using concatenated data.

Supertree methods of inferring phylogeny address the weakness of using a tree based on a single alignment by combining data from several input trees into a single representative phylogeny (Creevey & McInerney 2005). Supertree methods offer the advantage that the leaf sets of the input trees need not match each other exactly, merely overlap. At the level of gene families, this means that it is not necessary for every organism under investigation to have a copy of every gene. Additionally, it is possible to carry out a *post hoc* analysis of agreement between input trees and supertrees in order to assess congruence (Creevey *et al.* 2004). These are key points in favour of phylogenetic supertrees. If we intend on generating concatenated alignments of multiple loci for prokaryotes, then we will have to trim the data in order to remove genes with conflicting phylogenies (Roure *et al.* 2007) and the final datamatrix could be very sparse with more gaps than filled cells.

Suitable gene families can be identified using accepted criteria for asserting homology, the phylogenetic relationships inferred from these homologues can be extracted and used to build the supertree. By using large numbers of gene families, the final supertree is based upon many more relationships between the genomes in a given data set than by simply using a small number of genes to build a phylogeny. On this basis, supertree-based studies have become increasingly popular in recent times (see Beiko *et al.* 2005; Pisani *et al.* 2007). However, there are limitations associated with supertree construction. Probably the biggest drawback is the inability of current software to handle gene families where paralogous sequences are present. This limits the number of gene families used to build the final supertree, given that paralogues are frequent, even in prokaryotic genomes. Some methods can be used to deal with this in part, such as deletion of lineage-specific duplication events, but ultimately the problem is still a serious one. Another problem with supertree methods is that the quality of the supertree is based on the quality of the input data, in this case the input trees. If the input trees themselves have low levels of support for the relationships they represent, or if they do not overlap sufficiently (Scornavacca *et al.* 2008), or if some organisms are not well represented, then the quality of the supertree will also suffer. However, unlike the issue of using single-gene families, these problems can be addressed to a certain extent by employing various methods to ensure the input trees are of sufficient quality for supertree construction, such as removing poorly aligned regions (Talavera & Castresana 2007), removing alignments with little signal or removing very short alignments. These kinds of alignment and regions of alignments are expected to confound phylogenetic inference (Talavera & Castresana 2007).

The purpose of this study is to demonstrate the difficulty associated with using genome data to construct a phylogeny of the YESS group using many of the methods listed in the previous paragraphs, namely single-gene phylogenies, data concatenation and supertree analysis. We are doing this in order to test whether these organisms can be robustly classified, whether there is a meaningful phylogenetic tree that is agreed upon by a considerable amount of the data and whether there is general agreement across all methods and all data. We have specifically chosen to look at shallow-level relationships both inside and outside the species boundaries, as they are currently understood.

2. MATERIAL AND METHODS

(a) *Genome sequences*

The GOLD database (http://www.genomesonline.org/) was used to obtain the genome for 27 completed YESS group genomes. This included eight *Yersinia*, eight *Escherichia*, five *Salmonella* and six *Shigella* genomes. A full list of the individual genomes can be found in table 1 in the electronic supplementary material.

16S rRNA tree: 187 16S rRNA sequences from the 27 YESS group genomes were downloaded from GenBank. These sequences were aligned using ClustalW v. 1.83 (Thompson *et al.* 2002). A phylogenetic tree was constructed using model selection and maximum likelihood as described in the Multiphyl (Keane *et al.* 2007) documentation.

(b) *Identification of single-gene families*

Gene families were identified using the Random-BLAST method as described in Fitzpatrick *et al.* (2006). A total of 8736 gene families were recovered. The set of gene families was then filtered to remove families with fewer than four sequences, which is the smallest number of sequences required to build a non-trivial phylogenetic tree. This left 4693 gene families. Out of these families, 3109 were found to be single-gene families, with at most one representative sequence from each of the 27 genomes.

(c) *Multiple sequence alignment of remaining single-gene families*

The corresponding amino acid sequences of the 3109 single-gene families were used as input to ClustalW v. 1.83 (Thompson *et al.* 2002) for multiple sequence alignment. A total of 3109 alignments were produced. Each of the 3109 alignments was input into Gblocks (Talavera & Castresana 2007) to remove poorly aligned regions. A shell script (available on request) was created to remove badly aligned regions in a more relaxed manner than the default Gblocks settings. We set the minimal length of a block to 8 amino acid positions, and the maximum number of allowed contiguous non-conserved amino acid position to 15. Gapped sites were not systematically removed; rather they were treated as any other site in the alignment. Perl scripts were written to remove alignments that had fewer than 150 residues following analysis by Gblocks. This left a total of 1960 alignments.

The remaining alignments were converted to nexus format, and a PAUP* (Maddison *et al.* 1997) block for carrying out a permutation-tail-probability (PTP) test was added to each nexus file. The nexus files were then executed in PAUP* and a PTP test was carried out on each alignment. The resulting p-values gave a measure of confidence in the strength of the signal within the alignment. Only alignments passing the PTP test, i.e. those with a p-score of ≤ 0.01 were retained. A total of 1408 alignments were found to pass the PTP test. Nucleotide sequence alignments were then constructed based on these amino acid alignments.

(d) *Construction of phylogenetic trees*

Maximum likelihood phylogenetic trees for the 1408 alignments were constructed using MultiPhyl (Keane *et al.* 2007), with the model selection option set to *yes*. This resulted in 100 bootstrapped trees for each alignment. Each set of 100 bootstrap replicates was then summarized as a majority-rule consensus tree using Consense (Felsenstein 1993). The default settings were changed so that only nodes receiving 70 per cent support or greater were shown to be resolved on the resultant output tree. This produced 1408 consensus trees, one tree for each of the 1408 alignments. These trees were used for the supertree analysis.

(e) *Supertree construction*

Clann (Creevey & McInerney 2005) was used for supertree construction. A variety of different supertrees were constructed using the *dfit* optimization function. All other settings were left on their default values. Bootstrap resampling (100 replicates) of the input data was carried out and supertrees generated using these pseudoreplicates were summarized using a majority-rule consensus method.

(f) *Input tree-to-supertree distances*

The Treecompare software (available from authors) was used to measure the level of incongruence between the input trees and the dfit supertree. A score was generated for each of the 1408 input trees in terms of dissimilarity to an appropriately pruned supertree. This score was based on the Robinson–Foulds distance metric (Robinson & Foulds 1981). An Excel spreadsheet containing the input trees and their score against the dfit supertree can be found in the electronic supplementary material.

(g) *Minimum-evolution tree*

The nucleotide data for the 1408 single-gene families was aligned by translating the individual sequences into their corresponding amino acid sequences, aligning the proteins using ClustalW v. 1.83 and putting the gap characters into the nucleotide sequences according to where they were found in the amino acid sequences. These data were then analysed using

PAUP (Wilgenbusch & Swofford 2003) using the GTR distance matrix method with the optimality criterion set to minimum evolution.

(h) *Four-taxon trees*

We obtained genome sequences for six groups of prokaryotes for which there were multiple within-species genomes available. We chose four taxa in each case and using homology-based detection, we identified putative orthologues by choosing only gene families where there was one copy of the gene in every genome. We then constructed alignments for these putative orthologues, removed badly aligned regions using Gblocks and carried out a PTP test. This identified all alignments where there was some kind of evolutionary signal that was stronger than might be expected from random sequences. We labelled these alignments with the epithet 'signal'. We then used maximum likelihood to infer which of the three possible topologies was the best fit to the data. We carried out a Shimodaira–Hasegawa (SH) test (Shimodaira & Hasegawa 2001) in order to identify those datasets where the best-fitting topology was significantly better than the other two topologies. We labelled these datasets 'strict signal'. Lastly, we used bootstrap resampling of all alignments in order to find out what proportion of the alignments supported each of the three topologies. We partitioned the data into sets of alignments based on whether the internal branch on the four-taxon tree received at least 70 per cent bootstrap support or 100 per cent bootstrap support. We used such a variety of filtering regimes in order to explore the sensitivity of the results to differences in data treatment.

(i) *Ribosomal RNA sequence analysis*

All 16S rRNA sequences were aligned using ClustalW v. 1.83. The alignment was inspected by eye and ambiguously aligned regions were removed. The alignment is available as electronic supplementary material. Using standard methods for finding the optimal model of nucleotide substitution (Keane *et al.* 2006), we used the HKY+I+G model for all subsequent phylogenetic analyses. Confidence in phylogenetic hypotheses was assessed using bootstrap resampling and results are presented following 100 bootstrap replicates.

(j) *Housekeeping gene analysis*

The three housekeeping genes, *atpD*, *gyrB* and *trpB*, were retrieved from each genome using BLAST. The sequences were aligned using ClustalW v. 1.83. Upon inspection of the alignments, no further changes were felt necessary because the sequences were strongly conserved and the alignments seemed sensible (alignments available on request).

3. RESULTS

We have explored a small number of genomes to see the levels of conflict and congruence that different phylogeny reconstruction approaches will uncover. We demonstrate that there is a high level of heterogeneity in the data and the results we might see for one species do not necessarily generalize to all species. For a selection of *Enterobacteriacece*, we have analysed the entire complement of 16S rRNA genes, three commonly analysed housekeeping genes, a concatenated alignment of all 1408 genes where we could unambiguously assign orthology and a supertree of the same 1408 genes. In addition, we have analysed a number of four-taxon phylogenetic trees from strains within the same species and for species within the same genus in order to examine if the results generalize outside of our group of Enterobacteriacece.

(a) *16s rRNA gene tree*

We used every16S rRNA gene from the 17 genomes. This came to a total of 187 genes in our alignment. The broad topology of the 16S rRNA tree, as outlined in figure 1, is in line with expectations. *Yersinia* and *Salmonella* both form monophyletic groups while *Shigella* groups within *Escherichia*. However, many of the other features of the tree are unusual. Firstly, in general, the 16S rRNA genes within each genome do not form monophyletic groups with one another, with only two such species-specific clades found on the tree. *Shigella* is non-monophyletic, with multiple *Shigella* groupings within the *Escherichia* clade. The simplest interpretation of the data is that homogenization of ribosomal RNA genes is not sufficiently rapid that each genome has its own unique kind of 16S gene. This means that a genome-of-origin cannot be assigned based on the sequence of the 16S rRNA gene. The alternative explanation is that 16S rRNA genes are being exchanged between strains by some recombination mechanism. What we do see, in general, is that for this collection of genomes there are three kinds of 16S rRNA—a *Yersinia*-type of rRNA, a *Salmonella*-type of rRNA and an *Escherichia*/*Shigella*-type of rRNA.

(b) *Concatenated atpD, gyrB and trpB tree*

Figure 2*a–c* shows the trees for the three housekeeping genes *atpD*, *gyrB* and *trpB*. Once again, in all trees we find a monophyletic grouping of *Yersinia*, a monophyletic grouping of *Salmonella* and the *Shigella* sequences are mixed with the *E. coli* sequences. A closer analysis of these gene trees reveals some common features. Assuming a rooting in the centre of the circle trees shown in figure 2, *Yersinia enterocolitica* is the deepest branch in each tree, followed by *Yersinia pseudotuberculosis. Yersinia pestis* Microtus and *Yersinia pestis* Mediaevalis group together in the gyrB tree and tryB tree. The relationships for the *Salmonella* genomes show a similar level of conflict. *Salmonella enterica* Typhi Ty2 and *S. enterica sv* Typhi CT18 group together on all the trees. The other three *Salmonella* strains are found located in different positions in each tree. In the *gyrB* and *trpB* trees, *E. coli* MG, *E. coli* W, *E. coli* 0157 and *E. coli* Sakai form a group outside the subclade formed by the remaining four *E. coli* and six *Shigella* strains. The *atpD* tree is different, with the 0157/Sakai group outside the *Shigella*, while *E. coli* 06 K15 moves from outside *Shigella* to a grouping with *Shigella sonnei*.

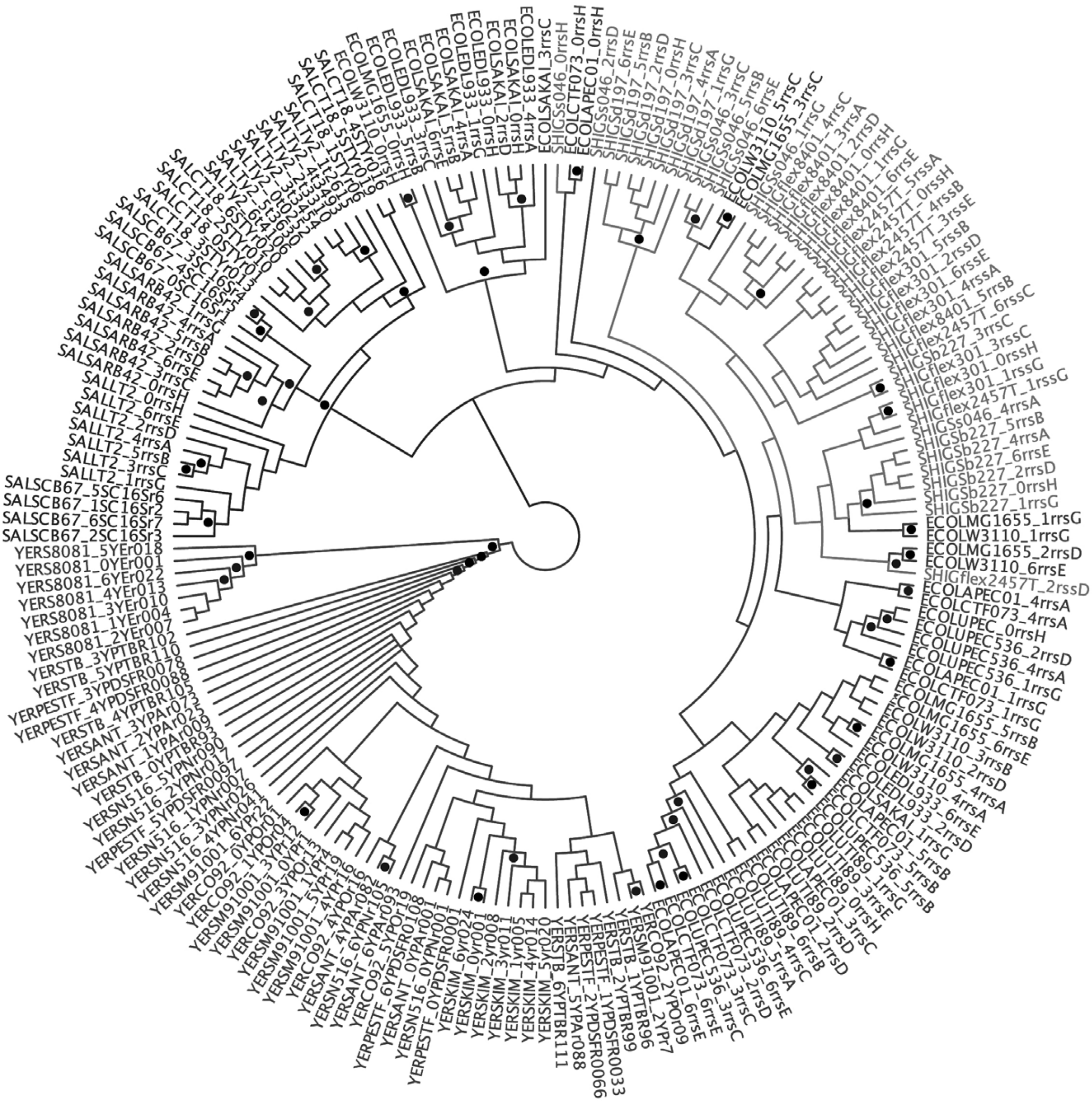

Figure 1. Phylogenetic tree of 187 16S rRNA sequences. Grey nodes denote more than 50 per cent bootstrap support, and black nodes denote more than 70 per cent bootstrap support.

Using the CONSEL software (Shimodaria & Hasegawa 2001), we carried out a number of analyses of the significance of the difference between the trees generated from the three housekeeping genes. For each alignment, we took the maximum likelihood tree and we tested whether its topology was within the confidence set of trees for the other two alignments. The results are presented in the electronic supplementary material, and for each alignment, the two trees that were not derived using that alignment were rejected by 23 out of 24 tests. The single exception among the 24 tests was where the SH test did not consider the topology of the *gyrB* tree to be outside the confidence set of trees for the *trpB* alignment and therefore did not reject that topology ($p = 0.132$). Notably, all other tests of the significance of difference for this alignment and tree combination rejected the topology of the *gyrB* tree.

The tree from the concatenated alignment is shown in figure 2*d* and it has elements of the relationships found in each of the individual gene trees though it ultimately conflicts with all of them as a result. This is not unexpected given the conflict between the gene trees themselves. All things considered, we feel that little confidence can be invested in the relationships on the tree based on concatenated data.

(c) *Supertree of 1408 single-gene families*

Figure 3 shows a supertree constructed from 1408 single-gene families derived from nucleotide alignments. The supertree recovered using these shows strong support across the majority of the tree. Some low support values exist in the *Yersinia* clade, but in general the tree has strongly supported relationships,

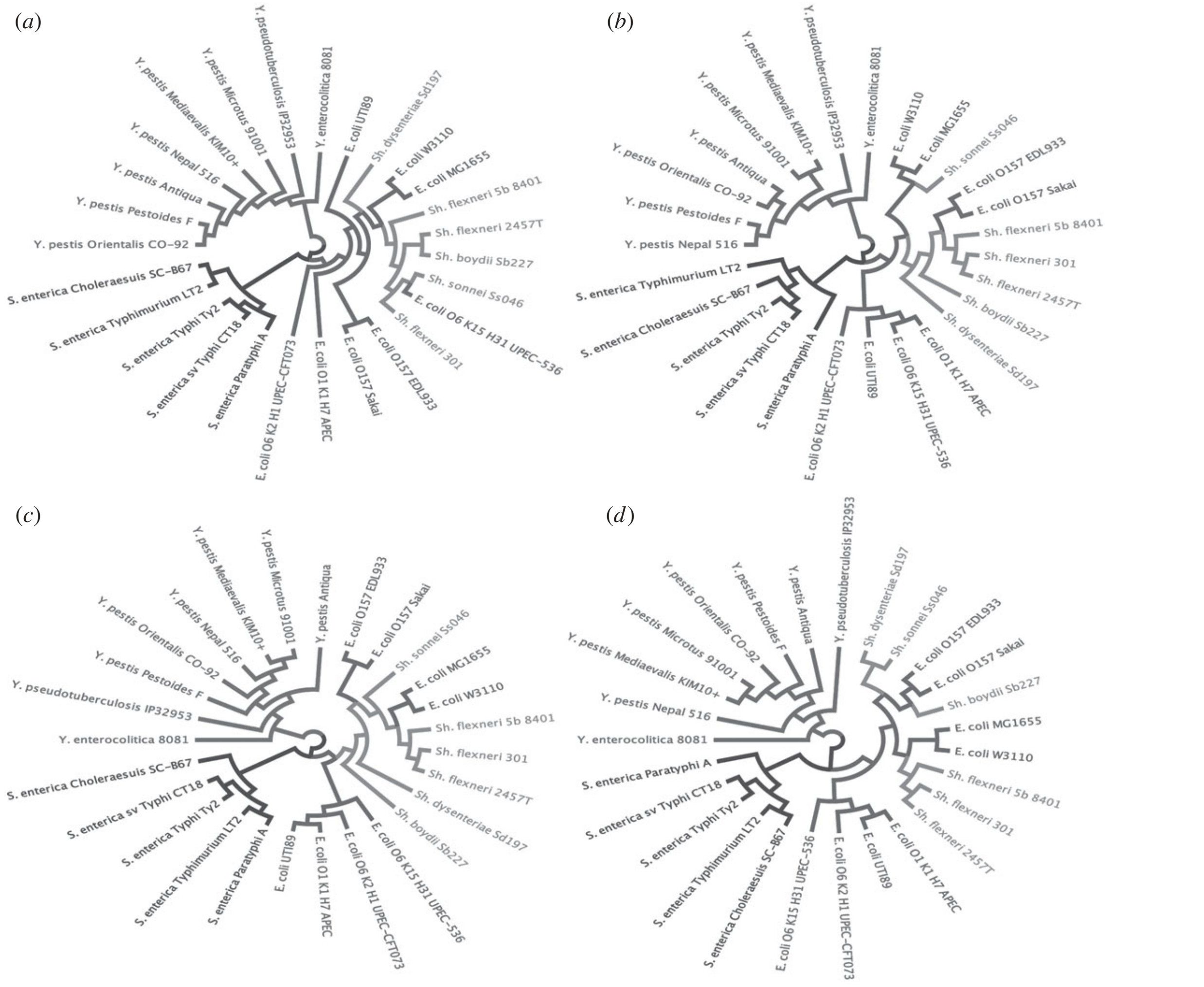

Figure 2. Phylogenetic trees for (*a*) *atpD* (*b*) *gyrB*, (*c*) *trpB*. (*d*) Phylogenetic tree based on concatenated gene sequences for *atpD*, *gyrB* and *trpB*.

probably indicative of the greater amount of signal in the nucleotide sequence data.

(d) *Minimum-evolution tree of concatenated data*

Figure 4 displays the tree recovered using minimum-evolution criterion for a concatenated alignment of the same 1408 single-gene families used for the construction of the supertree. Minimum evolution was used instead of maximum likelihood because of the length of the alignment (1 537 155 bases). The concatenated data tree shows strong support for the majority of the nodes in the tree. Weak support is present only towards the base of the *Yersinia* clade and at the node separating the *Salmonella* clade from the *Escherichia*/*Shigella* clade.

(e) *Tree-to-supertree distances for 1408 source trees*

One of the most interesting questions is whether or not the various phylogenetic trees we use as input to generate the supertree are similar in topology to an appropriately pruned supertree. Tree-to-tree distances from the 1408 ML input trees to the supertree were calculated using the Robinson–Foulds distances (Robinson & Foulds 1981) as implemented in the Treedist program in the PHYLIP package. This calculates the number of elementary operations required to convert one phylogenetic tree into another. Therefore, a distance of two indicates that two nearest-neighbour branch swaps are required to convert one tree into another. The average input tree-to-supertree distance was 1.1733 (median 1.168, range 0.181–3.458), with no trees receiving a score of zero, meaning no conflict between the topologies of these trees and the supertree. The number of leaves on these 11 non-conflicting trees varied from four to seven sequences. This is understandable, given that fewer leaves on the tree provide fewer opportunities for conflict with the topology of the supertree. Input trees based upon families with larger numbers of sequences were, in general, responsible for much of the conflict observed, though it should be noted that since 587 of the 1408 families were universally distributed across the 27 genomes, some of this is simply a reflection of the abundance of widely distributed genes in the data. In terms of phylogenetic conflict, of the families with the largest tree-to-tree distances compared with the supertree, many were found to be ribosomal or ribosome-associated proteins. Because of the high level of similarity in the sequences of these genes,

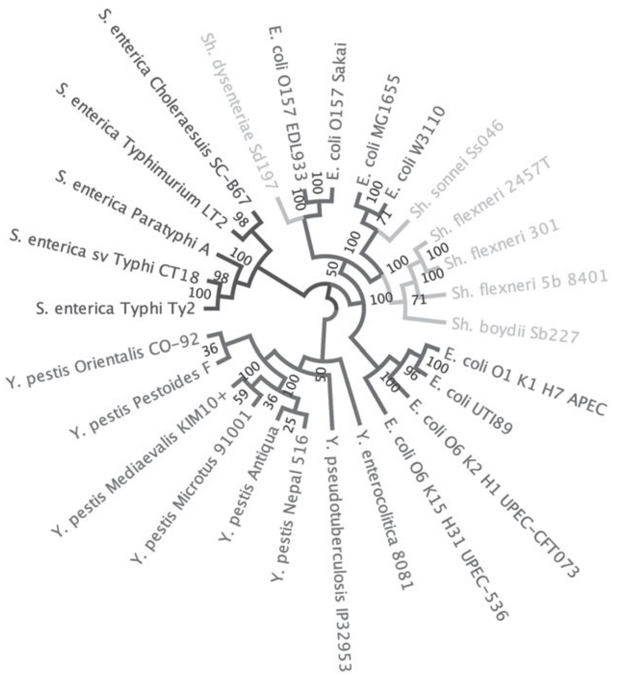

Figure 3. Supertree of 1408 single-gene families using nucleotide data.

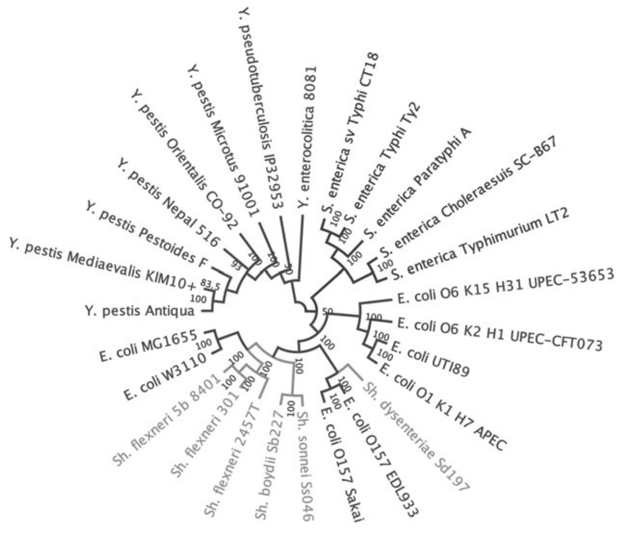

Figure 4. Minimum-evolution tree built from an alignment of 1408 single-gene families.

there is little statistical support for the input trees. While the groups themselves may be well defined, the lack of resolution of the internal relationships within a group leads to conflict with the supertree.

We examined a number of four-taxon datasets from a diverse range of prokaryotes. A four-taxon dataset can return only three possible unrooted bifurcating trees. Therefore, we can carry out some useful comparisons. We can ask if one of the trees is preferred (supported by the majority of the data) all of the time, or if two of the trees are preferred and one is not, or if none of the trees is particularly better supported than any of the others. We expect that some of the trees have no significant support for any given topology, either because superimposed substitutions have overwritten the phylogenetic signal or because there are too few mutations to make a statistically robust decision on which tree(s) is/are preferred. However, we might also expect to see some trees strongly supported, and by analysing these particular trees, we might gain some insight into the nature of HGT in the organisms under study.

Our results in table 1 illustrate that there is considerable variation in the frequency of HGT in our datasets. The rate of HGT is generally decreasing from the taxa on the left of table 1 compared with the taxa on the right. Our filters include the elimination of four-sequence alignments that do not pass a standard PTP test (the ‘signal’ datasets) and a further test of whether the ‘winning tree’ seems to be significantly better fit to the data than the two alternatives (the ‘strict signal’ datasets). We further screened the resulting phylogenetic trees by analysing those where the single internal branch received a bootstrap support value of 70 per cent or whether it got a score of 100 per cent (meaning that none of the alternative hypotheses was seen during bootstrapping). These filters were implemented because we do not wish to make any decisions on whether there are conflicting phylogenetic signals unless there is strong support for incompatible hypotheses of relationships.

4. DISCUSSION

We have analysed a set of taxa that are known to be closely related and where there is some confusion over whether there is a total of three or four valid taxa within the group. This test dataset is emblematic of the issues that crop up in employing genome-scale data to answer questions concerning the evolutionary history of prokaryotes.

Three groups were consistently recovered from our analysis, irrespective of the method chosen to infer phylogenetic relationships. These were the *Yersinia* group, the *Salmonella* group and the *Escherichia*/*Shigella* group. We did not find that there was a single origin of *Shigella* (Escobar-Paramo *et al.* 2003), rather, we found multiple origins, in accordance with the findings of Pupo *et al.* (2000). We did not find that the three groups were a single homogenous entity; we did find partitions. There were clear boundaries, and none of our analysis methods broke these boundaries. This sets up the possibility of describing bacterial species according to a bacterial genomic species concept. If we found that using genome data, there were no clear distinctions between the groups, then there would be no need for a bacterial species concept because the idea of a species would be outdated.

One of the weak features of this kind of analysis is the sampling issue. Having so few genomes to work with means that we are unlikely to have probed the boundaries of the species lines as they are depicted in the figures in this paper (we would naturally place *E. coli* and *Shigella* spp. into the same genome–species group). We will know in the future if these discrete boundaries hold steady or if they break down.

When we look within each of the groups, the story is clearly somewhat different. There were very few recurrent themes across different analyses and different

Table 1. Genes that are said to have signal have passed a PTP test, while those that are considered to be strict signal have passed both a PTP and SH test, respectively.

dataset	*Neisseria*		*Escherichia*		*Streptococcus*		*Staphylococcus*		*Chlamydia*		*Buchnera*	
genes in dataset	9089		18 912		7642		12 071		4178		7029	
single-gene families	1184		909		678		697		631		281	
signal	390		132		64		70		2		208	
support[a] (%)	70	100	70	100	70	100	70	100	70	100	70	100
topology 1 (%)	30	32	91	91	32	19	0	0	n/a	n/a	0	0
topology 2 (%)	32	33	4	3	33	25	96	98	n/a	n/a	0	0
topology 3 (%)	37	36	5	6	35	56	4	2	n/a	n/a	100	100
total[b]	388	162	132	61	366	34	70	53	0	0	208	208
strict signal	348		27		25		8		0		174	
support[c] (%)	70	100	70	100	70	100	70	100	70	100	70	100
topology 1 (%)	31	32	85	67	48	50	0	0	n/a	n/a	0	0
topology 2 (%)	33	32	7.5	22	28	25	100	100	n/a	n/a	0	0
topology 3 (%)	36	36	7.5	11	24	25	0	0	n/a	n/a	100	100
total[d]	310	115	27	9	25	8	8	8	0	0	174	80

[a] Bootstrap support of genes that have signal.
[b] The number of genes that were used in signal.
[c] Bootstrap support of genes that have strict signal.
[d] The number of genes that were used in strict signal bootstrap analysis.

datasets. Unlike Ochman *et al.* (2005), we did not analyse only those gene families that were found in all genomes; we analysed all gene families, even those with a patchy distribution.

The 16S phylogeny produced a result that might be considered contrary to expectations. Homogenization of all 16S sequences within a genome was not complete. Depending on the copy of the 16S gene that might be used, the resulting phylogeny can be different. Although we speculate that this is because homogenization is not fast enough that all copies of the sequence are the same in each genome, it is just as parsimonious to hypothesize that there has been recombination between strains and this is the reason for the absence of within-genome monophyly. The three major groups are recovered on this 16S rRNA tree, and this suggests that either homogenization is rapid enough to avoid the intermingling of sequences across the three major groups or sequence divergence has been sufficient that homologous recombination is much less frequent across the genome–species divide.

When examining the results of concatenating the sequences of *atpD*, *gyrB* and *trpB*, the same three major groups are recovered in each tree, but the internal relationships differ significantly (as judged by a number of tests using CONSEL (Shimodaira & Hasegawa 2001)) from tree to tree. In fact there is little to no agreement over the internal relationships of the groups. By concatenating the data and reconstructing a representative phylogeny, we can produce a result that is a mixture of the information contained in three conflicting topologies, but it is not clear what this tree means and in fact, we would suggest that it is meaningless. This kind of approach has been used previously to assess congruence with 16S rRNA phylogenies in a large number of *Streptomyces*, and it has been reported that the results were 'obviously superior to the 16S rRNA gene tree in both resolution power and topological stability' (Guo *et al.* 2008). The tree that we recover from this concatenated alignment has low bootstrap support, and we feel this reflects the fact that the individual trees have conflicting histories. As an approach to understanding the evolutionary history of the YESS group, this method seems to be ambiguous.

Both the 1408 gene nucleotide-based supertree and the minimum-evolution tree of the concatenated nucleotide data fare much better with regard to support for the hypotheses that they display. The trees agree completely in terms of the relationships for the *Salmonella* clade and only minor differences exist in the *Escherichia*/*Shigella* clade with the position of *S. sonnei* and the relationships between the three *Shigella flexneri* strains changing between the two trees. It should be noted that even though the differences are minor, they receive strong support in both trees. The major area of difference between the trees is in the *Yersinia* clade. The supertree shows weak support for some of the internal relationships while the minimum-evolution tree shows strong support for all the relationships bar the split between *Y. enterocolitica* and the rest of the clade. Do these trees have more meaning than the trees from the 16S rRNA gene or the housekeeping genes? This is a difficult question to answer. Are they simply revealing the central tendency in these large datasets? Possibly. Do these trees have real meaning in terms of revealing the evolutionary history of the groups?

Furthermore, there is no obvious reason to choose one tree over another. What does this tell us about the YESS group? Surely this is not a uniquely difficult group to analyse, yet after a thorough examination of the data, apart from concluding that there are three, not four major genome–species, we are left with as many questions as when we started. It has been previously argued that a tree-like phylogeny may exist only at the tips for prokaryotes and that the deeper branches may remain a mystery (Creevey *et al.*

2004). Here we find that at the tips it may be impossible to derive a reliable phylogeny.

In an effort to illustrate further that it is difficult to derive definite conclusions, we carried out some four-taxon analyses of genes in a variety of genera. We used at least three genomes from a single species and sometimes all four genomes were from the same species. As can be seen from table 1, the evolutionary histories of the genomes of the four *Neisseria* in our analyses are completely randomized. All three topologies are equally well supported. For the *Escherichia* dataset, there is somewhat less randomization, with topology 1 tending to be supported most often. For *Streptococcus*, each topology is supported by an appreciable amount of data, whereas for *Staphylococcus* there is clear support for one particular topology. Not enough chlamydial genes passed through our filters, so we cannot say anything about this taxon using these data. In *Buchnera*, where there is a vertical pattern of inheritance of the bacterium within an aphid host, there is no evidence for any HGT, with all gene trees supporting the same topology. This pattern of genome stasis has already been reported previously (Tamas *et al.* 2002).

This analysis of these genera illustrates that whatever we might say for *E. coli*, it cannot be guaranteed that we can say the same about other organisms. Despite the fact that we now have more than 1800 sequenced prokaryotic genomes, sampling is still patchy, usually being driven by medical or economic factors and, therefore, is perhaps not representative of most species. We may have phylogenetic bias in the collection of organisms we are analysing. If in some cases we have sequenced closely related strains and in others we have sampled more distant strains, this can have an effect on our estimates of recombination rate and population structure.

Assessing deep-level phylogenetic relationships is fraught with difficulties related to HGT and erosion of phylogenetic signal; however, assessing shallow relationships is no less difficult.

The authors acknowledge the financial support of the Irish Research Council for Science, Engineering and Technology (to FJM), the Science Foundation Ireland UREKA programme (to LSH) and the Science Foundation Ireland Research Frontiers Programme (to JOM).

REFERENCES

Achtman, M. & Wagner, M. 2008 Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6**, 431–440. (doi:10.1038/nrmicro1872)

Asai, T., Zaporojets, D., Squires, C. & Squires, C. L. 1999 An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc. Natl Acad. Sci. USA* **96**, 1971–1976. (doi:10.1073/pnas.96.5.1971)

Beiko, R. G., Harlow, T. J. & Ragan, M. A. 2005 Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 14332–14337. (doi:10.1073/pnas.0504068102)

Brenner, D. J. 1984 Enterobacteriaceae. In *Bergey's manual of systematic bacteriology*, vol. 1 (eds N. R. Krieg & J. G. Holt), pp. 408–420. Baltimore, MD: Williams and Wilkins.

Buckee, C. O., Jolley, K. A., Recker, M., Penman, B., Kriz, P., Gupta, S. & Maiden, M. C. 2008 Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. *Proc. Natl Acad. Sci. USA* **105**, 15082–15087. (doi:10.1073/pnas.0712019105)

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. 2006 Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287. (doi:10.1126/science.1123061)

Cooper, J. E. & Feil, E. J. 2004 Multilocus sequence typing—what is resolved? *Trends Microbiol.* **12**, 373–377. (doi:10.1016/j.tim.2004.06.003)

Creevey, C. J. & McInerney, J. O. 2005 Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**, 390–392. (doi:10.1093/bioinformatics/bti020)

Creevey, C. J., Fitzpatrick, D. A., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., Travers, S. A., Wilkinson, M. & McInerney, J. O. 2004 Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc. Biol. Sci.* **271**, 2551–2558. (doi:10.1098/rspb.2004.2864)

Dagan, T. & Martin, W. 2006 The tree of one percent. *Genome Biol.* **7**, 118. (doi:10.1186/gb-2006-7-10-118)

Dagan, T. & Martin, W. 2007 Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl Acad. Sci. USA* **104**, 870–875. (doi:10.1073/pnas.0606318104)

Dagan, T., Artzy-Randrup, Y. & Martin, W. 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl Acad. Sci. USA* **105**, 10 039–10 044. (doi:10.1073/pnas.0800679105)

Dauga, C. 2002 Evolution of the gyrB gene and the molecular phylogeny of *Enterobacteriaceae*: a model molecule for molecular systematic studies. *Int. J. Syst. Evol. Microbiol.* **52**, 531–547.

Doolittle, W. F. 1999 Lateral genomics. *Trends Cell Biol.* **9**, M5–M8. (doi:10.1016/S0962-8924(99)01664-5)

Doolittle, W. F. & Bapteste, E. 2007 Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 2043–2049. (doi:10.1073/pnas.0610699104)

Doolittle, W. F. & Papke, R. T. 2006 Genomics and the bacterial species problem. *Genome Biol.* **7**, 116. (doi:10.1186/gb-2006-7-9-116)

Escobar-Paramo, P., Giudicelli, C., Parsot, C. & Denamur, E. 2003 The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *J. Mol. Evol.* **57**, 140–148. (doi:10.1007/s00239-003-2460-3)

Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M. & Suerbaum, S. 2001 Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc. Natl Acad. Sci. USA* **98**, 15056–15061. (doi:10.1073/pnas.251396098)

Felsenstein, J. 1993 PHYLIP 3.6 edn. Distributed by Author.

Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. 2006 Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol. Biol. Evol.* **23**, 74–85. (doi:10.1093/molbev/msj009)

Gevers, D. *et al.* 2005 Opinion: re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3**, 733–739. (doi:10.1038/nrmicro1236)

Guo, Y., Zheng, W., Rong, X. & Huang, Y. 2008 A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *Int. J. Syst. Evol. Microbiol.* **58**, 149–159. (doi:10.1099/ijs.0.65224-0)

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McLnerney, J. O. 2006 Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29. (doi:10.1186/1471-2148-6-29)

Keane, T. M., Naughton, T. J. & McInerney, J. O. 2007 MultiPhyl: a high-throughput phylogenomics webserver using distributed computing. *Nucleic Acids Res.* **35**, W33–W37. (doi:10.1093/nar/gkm359)

Kidgell, C., Reichard, U., Wain, J., Linz, B., Torpdahl, M., Dougan, G. & Achtman, M. 2002 *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect. Genet. Evol.* **2**, 39–45. (doi:10.1016/S1567-1348(02)00089-8)

Kotloff, K. L., Winickoff, J. P., Ivanoff, B., Clemens, J. D., Swerdlow, D. L., Sansonetti, P. J., Adak, G. K. & Levine, M. M. 1999 Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ.* **77**, 651–666.

Lan, R., Lumb, B., Ryan, D. & Reeves, P. R. 2001 Molecular evolution of large virulence plasmid in Shigella clones and enteroinvasive *Escherichia coli*. *Infect. Immun.* **69**, 6303–6309. (doi:10.1128/IAI.69.10.6303-6309.2001)

Lawrence, J. G. 2002 Gene transfer in bacteria: speciation without species? *Theor. Popul. Biol.* **61**, 449–460. (doi:10.1006/tpbi.2002.1587)

Lee, T. M., Chang, L. L., Chang, C. Y., Wang, J. C., Pan, T. M., Wang, T. K. & Chang, S. F. 2000 Molecular analysis of *Shigella sonnei* isolated from three well-documented outbreaks in school children. *J. Med. Microbiol.* **49**, 355–360.

Maddison, D. R., Swofford, D. L. & Maddison, W. P. 1997 NEXUS: an extensible file format for systematic information. *Syst. Biol.* **46**, 590–621. (doi:10.2307/2413497)

Mahon, B. E. *et al.* 1997 An international outbreak of Salmonella infections caused by alfalfa sprouts grown from contaminated seeds. *J. Infect. Dis.* **175**, 876–882. (doi:10.1086/513985)

McInerney, J. O. & Pisani, D. 2007 Genetics. Paradigm for life. *Science* **318**, 1390–1391. (doi:10.1126/science.1151657)

McInerney, J. O., Cotton, J. A. & Pisani, D. 2008 The prokaryotic tree of life: past, present . . . and future? *Trends Ecol. Evol.* **23**, 276–281. (doi:10.1016/j.tree.2008.01.008)

Ochman, H., Lerat, E. & Daubin, V. 2005 Examining bacterial species under the specter of gene transfer and exchange. *Proc. Natl Acad. Sci. USA* **102**(Suppl. 1), 6595–6599.

O'Neil, D. M., Baron, L. S. & Sypherd, P. S. 1969 Chromosomal location of ribosomal protein cistrons determined by intergeneric bacterial mating. *J. Bacteriol.* **99**, 242–247.

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D. & Doolittle, W. F. 2007 Searching for species in haloarchaea. *Proc. Natl Acad. Sci. USA* **104**, 14 092–14 097. (doi:10.1073/pnas.0706358104)

Paradis, S. *et al.* 2005 Phylogeny of the *Enterobacteriaceae* based on genes encoding elongation factor Tu and F-ATPase beta-subunit. *Int. J. Syst. Evol. Microbiol.* **55**, 2013–2025. (doi:10.1099/ijs.0.63539-0)

Phillips, M. J., Delsuc, F. & Penny, D. 2004 Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**, 1455–1458. (doi:10.1093/molbev/msh137)

Pisani, D., Cotton, J. A. & McInerney, J. O. 2007 Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760. (doi:10.1093/molbev/msm095)

Pupo, G. M., Lan, R. & Reeves, P. R. 2000 Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl Acad. Sci. USA* **97**, 10 567–10 572. (doi:10.1073/pnas.180094797)

Purkhold, U., Wagner, M., Timmermann, G., Pommerening-Roser, A. & Koops, H. P. 2003 16S rRNA and amoA-based phylogeny of 12 novel betaproteobacterial ammonia-oxidizing isolates: extension of the dataset and proposal of a new lineage within the nitrosomonads. *Int. J. Syst. Evol. Microbiol.* **53**, 1485–1494. (doi:10.1099/ijs.0.02638-0)

Robinson, D. & Foulds, L. 1981 Comparison of phylogenetic trees. *Biosciences* **53**, 131–147.

Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F. & Philippe, H. 2007 Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* **56**, 389–399. (doi:10.1080/10635150701397643)

Rokas, A., Williams, B. L., King, N. & Carroll, S. B. 2003 Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804. (doi:10.1038/nature02053)

Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. 2007 SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* **7**(Suppl. 1), S2.

Sanderson, M. J., Driskell, A. C., Ree, R. H., Eulenstein, O. & Langley, S. 2003 Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol. Biol. Evol.* **20**, 1036–1042. (doi:10.1093/molbev/msg115)

Sansonetti, P. J., Kopecko, D. J. & Formal, S. B. 1981 *Shigella sonnei* plasmids: evidence that a large plasmid is necessary for virulence. *Infect. Immun.* **34**, 75–83.

Scornavacca, C., Berry, V., Lefort, V., Douzery, E. J. & Ranwez, V. 2008 PhySIC_IST: cleaning source trees to infer more informative supertrees. *BMC Bioinform.* **9**, 413. (doi:10.1186/1471-2105-9-413)

Shimodaira, H. & Hasegawa, M. 2001 CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247. (doi:10.1093/bioinformatics/17.12.1246)

Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P. & Rubin, E. M. 2007 Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449–1452. (doi:10.1126/science.1147112)

Staley, J. T. 2006 The bacterial species dilemma and the genomic–phylogenetic species concept. *Phil. Trans. R. Soc. B* **361**, 1899–1909. (doi:10.1098/rstb.2006.1914)

Tacket, C. O., Ballard, J., Harris, N., Allard, J., Nolan, C., Quan, T. & Cohen, M. L. 1985 An outbreak of *Yersinia enterocolitica* infections caused by contaminated tofu (soybean curd). *Am. J. Epidemiol.* **121**, 705–711.

Talavera, G. & Castresana, J. 2007 Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577. (doi:10.1080/10635150701472164)

Tamas, I., Klasson, L., Canback, B., Naslund, A. K., Eriksson, A. S., Wernegreen, J. J., Sandstrom, J. P., Moran, N. A. & Andersson, S. G. 2002 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379. (doi:10.1126/science.1071278)

Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2002 Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinform.*, Chapter 2: Unit 2.3.

Varma, J. K. *et al.* 2003 An outbreak of *Escherichia coli* O157 infection following exposure to a contaminated building. *JAMA* **290**, 2709–2712. (doi:10.1001/jama.290.20.2709)

Wilgenbusch, J. C. & Swofford, D. 2003 Inferring evolutionary trees with PAUP*. *Curr. Protoc. Bioinform.*, Chapter 6, Unit 6.4. Hoboken, NJ: John Wiley & Son.

Yang, J., Nie, H., Chen, L., Zhang, X., Yang, F., Xu, X., Zhu, Y., Yu, J. & Jin, Q. 2007 Revisiting the molecular evolutionary history of Shigella spp. *J. Mol. Evol.* **64**, 71–79. (doi:10.1007/s00239-006-0052-8)