Dealing with incongruence in phylogenomic analyses

Nicolas Galtier^{1,*} and Vincent Daubin^{2,3}

¹CNRS UMR 5554, Institut des Sciences de l'Évolution, Université Montpellier 2, CC64, Place E. Bataillon, 34095 Montpellier, France

²Université de Lyon 1, 69000 Lyon, France

³Laboratoire de Biométrie et Biologie Évolutive, CNRS, UMR5558, Université Lyon 1, 69622 Villeurbanne, France

Incongruence between gene trees is the main challenge faced by phylogeneticists in the genomic era. Incongruence can occur for artefactual reasons, when we fail to recover the correct gene trees, or for biological reasons, when true gene trees are actually distinct from each other, and from the species tree. Horizontal gene transfers (HGTs) between genomes are an important process of bacterial evolution resulting in a substantial amount of phylogenetic conflicts between gene trees. We argue that the (bacterial) species tree is still a meaningful scientific concept even in the case of HGTs, and that reconstructing it is still a valid goal. We tentatively assess the amount of phylogenetic incongruence caused by HGTs in bacteria by comparing bacterial datasets to a metazoan dataset in which transfers are presumably very scarce or absent. We review existing phylogenomic methods and their ability to return to the user, both the vertical (speciation/extinction history) and horizontal (gene transfers) phylogenetic signals.

Keywords: horizontal gene transfer; phylogenomics; phylogenetic conflict; bacteria

1. INTRODUCTION

Classical molecular phylogenetics aims at reconstructing one tree from one sequence alignment (or gene), usually with the goal of inferring species evolutionary history. In the genomic era, the typical dataset is made of many genes, each available in a variable number of taxa. Using such large amounts of data to reconstruct a more reliable tree of species is tempting. Unfortunately, distinct genes can support distinct trees, challenging the traditional methodology of molecular phylogeny.

Between-genes phylogenetic incongruences, or conflicts, can occur for many reasons. First, estimated gene trees can be different from the species tree owing to tree-building errors, either stochastic (in case of insufficient sequence length) or systematic (in case of departure from the model assumptions, Jeffroy *et al.* 2006). Alternatively, conflicts can reflect biological processes, when gene trees are truly different from each other. Three major evolutionary mechanisms potentially resulting in true phylogenetic discordance between genes are known: incomplete lineage sorting; hidden paralogy; and horizontal gene transfer (HGT). Incomplete lineage sorting occurs when an ancestral species undergoes several speciation events in a short period of time. If, for a given gene, the ancestral polymorphism is not fully resolved into two monophyletic lineages when the second speciation occurs, then with some probability the gene tree will be

different from the species tree (Tajima 1983; Pamilo & Nei 1988). A recent genome-wide analysis of hominid primates indicated that, as a consequence of incomplete lineage sorting, roughly 30 per cent of our genome support the (chimpanzee, (human, gorilla)) or (human, (chimpanzee, gorilla)) branching order, i.e. topologies different from the true (gorilla, (human, chimpanzee)) species tree (Hobbolt *et al.* 2007). A very different reason why gene trees can be truly incongruent is hidden paralogy. If a dataset includes paralogous copies, then the true phylogeny will partly reflect the duplication history of the gene that is independent of species divergence history. The third mechanism is HGT. If genetic exchanges occur between species, then the phylogeny of individual genes will be influenced by the number and nature of transfers they have undergone.

The discovery, during the last decade, of a substantial amount of HGT between bacteria has modified our view of genome evolution in this group (Doolittle 1999). HGT is obviously an important, previously neglected evolutionary mechanism in prokaryotes. Remarkably, the bacterial phylogeny is still largely unresolved (Creevey *et al.* 2004), although hundreds of full genome sequences are available. These two observations led to the proposal that HGT have been so frequent during bacterial evolution that the species tree can no longer be recovered. From a theoretical point of view, several authors have also questioned the meaning of trees as a representation of the evolutionary history of a group of species in the case of HGT (Gogarten *et al.* 2002; Doolittle & Bapteste 2007).

* Author for correspondence (galtier@univ-montp2.fr).

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

In this paper, we discuss the consequences of HGT and other sources of phylogenetic incongruence on phylogenomic analyses. We argue that the existence of incongruences is not sufficient to dismiss the notion of a species tree, nor to preclude its reconstruction. We review the behaviour and the principles of existing phylogenomic approaches with respect to HGT, and suggest that appropriate methods should aim at simultaneously reconstructing the species tree and HGT events. Comparing eukaryotic and bacterial datasets, we show that HGT significantly influences bacterial phylogenomic pattern, but probably does not preclude the reconstruction of the species tree in this group.

2. CAN WE STILL DEFINE THE SPECIES TREE?

If the evolution of genetic material was primarily governed by processes other than the species divergence, then the notion of a species tree would be of little, if any, empirical value. In this section, we examine the conditions under which the species tree is still a useful notion or becomes irrelevant. This is obviously strongly dependent on the causes of phylogenetic conflict between genes. Tree-building errors are not an issue here; we ask whether the species tree is or is not a meaningful notion, assuming that we know gene trees.

In case of HGT, the distribution of the genetic variation across genomes becomes independent of the species tree only if the rate of HGT is much greater than the rate of species diversification (i.e. the rate of speciation minus the rate of extinction). If this was true, then the age of the common ancestor between any two genes would be essentially independent of the common ancestor between their two host species. In this case, the tree describing species divergence would be irrelevant, and we should rather consider ecological traits—i.e. ability to exchange genes—as potential explanatory factors of genome biodiversity patterns (Gogarten *et al.* 2002).

The situation is a bit different in case of hidden paralogy; the relevance of the species tree is affected only if the rate of gene duplication/gene loss is of the same order of magnitude as the rate of species diversification. When gene duplications occur very rarely, they obviously do not influence gene trees much. At the other extreme, when the dynamics of gene duplication/gene loss is much faster than the pace of species divergence, then the paralogous members of a given gene family in a given species have a common ancestor younger than their common ancestor with gene copies from other species (Nei & Rooney 2005), so that if we sample one gene copy per genome, the gene tree will reflect the species tree. A similar pattern occurs in the case of frequent gene conversion between paralogues, or unequal crossing over between tandemly repeated sequences, leading to concerted evolution (Eickbush & Eickbush 2007). Only intermediate rates of gene duplication/gene loss, therefore, can lead to a substantial amount of phylogenetic variance across true gene trees.

As far as incomplete lineage sorting is concerned, finally, what matters is the ratio between the depth of within-species genealogies (i.e. the age of the common ancestor of conspecific individuals) and the waiting

time between two successive ‘successful’ speciation events—successful speciation events are such that the resulting two species have living descendent species represented in the sample. If this ratio was high, then the history of species divergence would be irrelevant to the structure of genomic variation: lineage sorting being random, gene trees would be essentially independent of each other. The depth of within-species genealogies depends on several population genetic parameters, including effective population size, natural selection and recombination. The average waiting time between successful speciations (i.e. the average internal branch length of the species tree) largely depends on the age of the taxonomic group under consideration and on the number of extant (sampled) species. ‘Dense’ trees will be more prone to incomplete lineage sorting than trees including few species diverged over long periods of time.

Having characterized the conditions under which the species tree is no longer a useful concept, the next question is obviously: are these conditions met in real life? A number of empirical arguments suggest that it is only rarely the case. As far as hidden paralogy is concerned, we have suggested that only rates of gene duplication/gene loss in the order of the rate of species diversification can lead to substantial discrepancy between gene trees and species trees. It would appear quite unlikely that the true biological distribution of this parameter (across genes) was concentrated within this relatively narrow window. Empirically, when facing datasets affected by duplication events, biologists typically attempt to reconcile gene trees with the species tree: they try to clarify the ortho/paralogy relationships between copies by distinguishing duplication nodes from speciation nodes in gene trees (e.g. Dufayard *et al.* 2005; Wapinski *et al.* 2007). Obviously, in practice, gene duplications have not affected the central role played by the species tree in molecular evolutionary analyses, however heterogeneous gene trees may be.

Incomplete lineage sorting apparently poses a more difficult theoretical problem. According to basic coalescence theory, the within-species genealogical depth is proportional to effective population size. This parameter is probably extremely variable across species; the effective population size of a bacterial species must be many orders of magnitude higher than that of, say, a vertebrate. Assuming a more or less constant rate of species diversification per time unit, should not we expect an overwhelming impact of incomplete lineage sorting in bacteria, knowing that it can affect vertebrate phylogenomic patterns (Hobbolt *et al.* 2007)? There are several reasons why this rationale is too simplistic. First, the within-species genealogical depth as defined by the coalescence theory is measured in units of generations, not in absolute time. Generation time is (most probably) negatively correlated with population size, what reduces the between-species variance of genealogical depth when appropriately measured in years. Second, the effective population size is not the only population genetic parameter controlling within-species genealogical depth. Natural selection, and especially positive selection, also matters. Selective sweeps, i.e. the rapid fixation of advantageous mutations leading to a sudden

drop of the genealogical depth, are more frequent in larger populations, simply because the population rate of advantageous mutation is higher. Gillespie (2000) showed that this effect, which he called genetic draft, essentially compensates for the decreased genetic drift in large populations. This is probably the reason why the observed level of within-species genetic diversity does not reach extremely high values, even in extremely abundant species.

Besides these theoretical considerations, the most obvious and perhaps the strongest arguments indicating that we can still define the species tree in most cases are empirical. If incomplete lineage sorting was a major problem, many of the morphologically or ecologically defined species should be found paraphyletic or polyphyletic, something we observe only rarely. Molecular data have largely corroborated existing species delineation, with exceptions. Most importantly, there is no evidence, as far as we know, for a higher frequency of non-monophyletic species in groups with larger expected population sizes. Similarly, if HGTs were frequent enough to make species tree a useless concept, we would expect no or very little correlation between gene trees. This is obviously not the case, even in bacteria, as discussed by Daubin *et al.* (2003) and Susko *et al.* (2006) (and see below). Phylogenetic agreement is much more common than disagreement, indicating that HGT is not prevalent enough to erase the vertical signal.

It should be noted, finally, that these considerations vary strongly depending on the genes we are considering. Some genes are highly prone to incomplete lineage sorting (typically owing to overdominant selection, e.g. immunity genes, Klein *et al.* (2007), and self-incompatibility genes, Castric & Vekemans 2004), duplications (e.g. olfactory receptors, Niimura & Nei 2006) or HGT (e.g. *mutS*, Denamur *et al.* 2000). Their history is little, or not at all, influenced by the species tree. Obviously, the existence of such genes is not sufficient to imply that the species tree is generally useless. Quantifying the proportion of genes strongly, weakly or not affected by non-vertical evolutionary processes is one of the major challenges of current comparative genomics, especially in bacteria (Lawrence & Hendrickson 2005).

3. SHOULD WE TRY TO RECOVER THE SPECIES TREE?

Recently, several authors have questioned the usefulness of the species tree in prokaryote evolutionary genomics, based on the observation that only a tiny fraction (1% or less) of reconstructed gene trees are congruent with the reconstructed species tree (Dagan & Martin 2006; Baptiste *et al.* 2008). This, we think, is not a correct argument. In our view, the species tree could still be a useful concept even if incongruent with every gene tree, as we now discuss in more detail.

Owing to incomplete lineage sorting, roughly 30 per cent of human genes do not support the (gorilla, (human, chimpanzee)) topology (Hobbolt *et al.* 2007). This problem must concern other recently diverged triplets of primate species. If we think of a dataset made of 10 such triplets (plus other species), then only 3 per cent (0.7^{10}) of true gene trees will be identical to the

species tree. As we add more species, and more triplets affected by incomplete lineage sorting, this percentage decreases. Does this mean that the primate species tree is useless? Obviously not. The scientific value of the species tree cannot be dependent on (decrease with) the amount of available data. Fundamentally, the primate species tree is useful to evolutionary biologists because it traces back speciation and extinction events, two evolutionary processes worth studying. Empirically, it provides a framework necessary for any comparative analysis of biological data. These properties, exhibited and highlighted by cladists decades ago are valid, whether or not lineage sorting was incomplete in some recently diverged groups of species. Recent approaches explicitly modelling the process of gene coalescence have been developed to reconstruct a species tree based on multiple, incongruent gene phylogenies (Edwards *et al.* 2007; Liu & Pearl 2007).

We see no reason why the same arguments would not apply to HGT in bacterial phylogeny. Assuming even a very low rate of HGT, the percentage of gene trees identical to the species tree will mechanically decline towards zero, as we add more species in the dataset, for combinatorial reasons—this tells nothing about the usefulness of the species tree. Our view is that we need the species tree for bacteria as much as we need it for primates. In particular, knowing the bacterial species tree would be of invaluable interest for studying the process of HGT, this important and fascinating aspect of prokaryote evolution. By reconciling gene trees with the (supposedly known) species tree, we could annotate gene gains and losses, and learn about the preferential routes of genetic exchange between species, families or phyla. We note that, similarly, the statistical support for any one tree declines as species are added even in the absence of HGT or incomplete lineage sorting simply due to the combinatorial increase in the number of possible topologies.

From a statistical point of view, rejecting the species tree because of the existence of conflicts between gene trees means denying calculation of the mean of a distribution because its variance is non-zero, which appears too extreme a policy. Note that calculating the mean is not sufficient, and this rule applies to phylogenomics as well: we need to analyse the variance, i.e. patterns of conflict between genes. By chance, the mean and variance in phylogenomics have distinct biological interpretations: speciations/extinctions versus HGT and other non-vertical processes. The two deserve to be studied and understood, and this can only be done jointly.

It is true, however, that the predictive value of the mean of a distribution decreases with its variance. The mean of a multimodal distribution, for instance, is not very informative, much less than the modes and their respective weights. In phylogenomics, multimodal distributions of true gene trees can occur in case of genome chimerism. The eukaryote nuclear genome, for example, is made of genes of ancient (proto)eukaryotic origin, and genes of organellar origin, transferred to the nucleus posterior to the endosymbiotic events. In such a situation, building a single 'average' (consensus) species tree is nonsense. Rather, we should try to reconstruct the (small number of) distinct trees

reflecting the distinct histories of the subsets of genes, as in [Pisani *et al.* \(2007\)](#). If the distribution of gene trees is ‘unimodal’, i.e. more or less randomly dispersed around the species tree (as in [Baptiste *et al.* 2008](#)), then taking the average (or consensus) makes sense, even if a single true gene tree is not identical to the true species tree.

4. CAN WE STILL RECOVER THE SPECIES TREE?

Assuming that recovering the species tree is a valid scientific goal, the next question is whether we can reasonably hope to reconstruct it, for a given taxonomic group, knowing that a proportion of genes are affected by non-vertical evolutionary processes, and that tree-building methods can fail—two independent sources of incongruence between gene trees. In principle, even a high variance between gene trees can be overcome by increasing the number of analysed genes provided that heterogeneity is random, as discussed previously. [Galtier \(2007\)](#) showed that a supertree method (which essentially returns the average estimated gene tree) recovers the true species tree with strong accuracy from phylogenomic data simulated under a model incorporating HGT, even when the amount of HGT is such that two random gene trees share only 50 per cent of their internal branches, on average. Under such conditions, no true gene tree is identical to the species tree; this does not preclude reconstruction of the correct tree.

This result, however, was obtained under a model in which true and reconstructed gene trees are randomly distributed around the species tree, i.e. the optimal situation. The challenge is bigger when departure from the species tree is non-random, i.e. correlated across gene trees. This can occur for biological or methodological reasons. Biologically, preferential HGT between specific phylogenetic lineages can lead to such correlated patterns, when true gene trees tend to share branching orders contradicting the species tree. An extreme case is genome chimerism, as discussed previously. As far as bacteria are concerned, neither [Beiko *et al.* \(2005\)](#) nor [Ge *et al.* \(2005\)](#) detected strong evidence for preferential genetic exchanges between specific phyla: according to these studies, bacterial gene trees are more or less randomly distributed around the species tree (but see [Matte-Tailleux *et al.* 2002](#); [Jain *et al.* 2003](#)). We note that there is a convergence between practical and conceptual aspects here; the task of reconstructing ‘the’ species tree is more and more difficult, and more and more pointless, as the distribution of true gene trees becomes multimodal.

The second reason why estimated gene trees can depart the species tree in a correlated way, and thus mislead inferences, is methodological bias. The maximum-likelihood (ML; and Bayesian) methods have solved the long-branch attraction problem in theory ([Felsenstein 1981](#)), but not in practice: when the actual evolutionary process violates the underlying model of sequence evolution, consistency is no longer ensured, and fast-evolving lineages typically branch as sister groups ([Lartillot & Philippe 2008](#)). Recent studies of metazoan phylogeny ([Philippe *et al.* 2005](#)), or of microsporidian origins ([Thomarat *et al.* 2004](#)), have

revealed the existence and strong phylogenomic impact of such genome-wide fast evolving lineages. If the bias affects a majority of genes, then the reconstructed species tree will be wrong, and this will not be solved by adding new genes. This problem is conceptually very different from the case of correlated HGT: here the correlation between gene trees is of artefactual, not biological, origin. It does not affect the scientific relevance of the species tree.

5. HOW STRONGLY IS BACTERIAL PHYLOGENOMICS AFFECTED BY HGT?

HGTs are an important source of conflicts between gene trees in bacteria. However, it is not the only one; hidden paralogy and reconstruction errors must also contribute to the phylogenetic variance between gene trees. Quantifying the amount of phylogenetic conflict caused by HGT versus other factors would appear worthwhile for a correct interpretation of bacterial phylogenomic data: how frequently, and under which conditions, should we invoke HGT?

To investigate this issue, we analysed the level of phylogenetic conflict between genes in several phylogenomic datasets. The first dataset was made of 10 fully sequenced metazoan species, and 284 genes. Each gene alignment included 100 amino acid positions or more. The other four datasets were made of 10 bacterial species each, sampled in four distinct phyla and 184–227 genes (100 amino acid positions or more). Bacterial datasets were chosen to grossly match the reference metazoan dataset with respect to average gene tree length (sum of branch lengths) and average gene tree diameter (the longest pathway between any two leaves). HGT is expected to be very rare or absent in animals; the metazoan dataset will therefore be considered as reflecting the basic level of conflict to be expected in the absence of HGT. The excedentary amount of conflict detected in bacterial datasets, when compared with the metazoan dataset, will be taken as a measure of the impact of HGT in bacterial phylogenomics.

Sequences were retrieved from the HOGENOM database (<http://pbil.univ-lyon1.fr/databases/hogenom.php>). We selected genes present in one copy per species, or for which duplications appeared to have occurred on terminal branches—in such cases, we selected a single representative copy per species. Sequences were aligned using CLUSTAL W ([Thompson *et al.* 1994](#)). Ambiguously aligned sites were removed with G-blocks ([Castresana 2000](#)), with the -b5 option (gaps conserved). Gene trees were reconstructed with PhyML ([Guindon & Gascuel 2003](#), LG matrix, all parameters estimated). Trees and alignments are available at ftp://pbil.univ-lyon1.fr/pub/datasets/DAUBIN/Galtier_Daubin08/.

Table 1 provides the main results of this analysis. The first six columns summarize the characteristics of the analysed datasets, and confirm that they are pretty comparable in terms of phylogenetic depth. To make the comparison even fairer in terms of sequence length, we also analysed a reduced metazoan dataset made of the 200 shortest genes of the main metazoan dataset (table 1, line 2). The last four columns of table 1 report various measures of phylogenetic conflict between gene

Table 1. Measures of phylogenetic conflict between genes in real and simulated datasets.

dataset	sp. ^a	genes ^b	seq lg. ^c	tree lg. ^d	diam. ^e	RF ^f	ML ^g	SPR ^h	%SPR=0 ⁱ
metazoa	10	284	507.5	3.47	1.88	1.76	1.72	0.010	98.6
metazoa_short	10	200	343.8	3.55	1.89	2.09	2.09	0.010	99.0
γ -proteobacteria	10	199	338.2	3.38	2.01	4.07	5.78	0.009	99.1
bacilli	10	227	299.8	2.67	1.25	5.26	8.74	0.089	92.1
α -proteobacteria	10	184	306.9	4.41	1.90	8.01	14.65	0.145	87.0
actinobacteria	10	187	314.5	3.92	1.71	10.33	26.83	0.289	77.6
sim random	10	200	300	3.96	2.20	12.44	134.01	—	—

^a Number of species.^b Number of genes.^c Average alignment length.^d Average gene tree length (sum of branch lengths).^e Average gene tree diameter (longest leave-to-leave pathway).^f Average Robinson–Foulds distance between gene trees.^g Average maximum-likelihood reciprocal rejection.^h Average number of necessary SPR per gene tree.ⁱ Percentage of gene trees showing no significant conflict with the reference.

trees. RF is the mean Robinson–Foulds topological distance between gene trees (Robinson & Foulds 1981), averaged over every pair of genes. This standard measure does not account for the statistical support of internal branches, so that unresolved trees will typically be considered as conflicting. To correct this problem, we introduce a measure of conflict based on the ML criterion (column ML in table 1). Let T_1 be the ML tree for gene 1 and L_1 the associated maximal log likelihood. Similarly define T_2 and L_2 for gene 2. Now calculate L'_1 , the log likelihood of tree T_2 for gene 1, and L'_2 , the log likelihood of T_1 for gene 2 (reoptimizing branch lengths). We define the ML conflict between gene 1 and gene 2 as:

$$\text{ML_conflict}(1, 2) = \min(L_1 - L'_1, L_2 - L'_2).$$

With this definition, only genes significantly rejecting the tree supported by each other are considered conflicting. This variable was computed for every pair of genes and the average was taken.

The RF and ML measures do not explicitly account for the nature of the topological differences between gene trees. A single HGT event can result in large RF (and ML) differences if it occurs between distantly related taxa. The subtree pruning regrafting (SPR) distance accounts for this aspect; it corresponds to the minimal number of SPR rearrangements necessary to reconcile two trees. This number was calculated using the EEEP program (Beiko & Hamilton 2006). This method accounts for the support of internal branches: only SPR moves reconciling statistically significant conflicts are counted. We considered as statistically significant every node supported by a probability over 95 per cent, as defined by the min (SH, χ^2) test implemented in PHYLML (Shimodaira & Hasegawa 2001). EEEP was used to calculate the SPR distance between each gene tree and the consensus tree, reconstructed using the matrix representation parsimony method (Ragan 1992). The per-gene average distance is given. The last column of table 1 (SPR=0%), finally, gives the percentage of gene trees showing no significant discordance with the consensus tree (SPR distance equal to zero).

All four measures of average conflict largely agree. Not surprisingly, phylogenetic conflict is the lowest in

the metazoan dataset, for which HGTs are presumably very rare. The four bacterial datasets show quite distinct patterns. Conflict is low in the γ -proteobacteria dataset (even lower than in the metazoan dataset as far as SPR and SPR=0% measures are concerned), confirming earlier reports by Daubin *et al.* (2003) and Lerat *et al.* (2003). Phylogenetic conflicts are much more common in actinobacteria, in which two average genes support trees distinct by over 25 log-likelihood units. The bacilli and α -proteobacteria datasets show intermediate levels of conflicts.

These analyses therefore suggest that the rate of HGT varies substantially between bacterial phyla. The optimistic conclusions of Daubin *et al.* (2003) and Lerat *et al.* (2003), based on γ -proteobacteria, are apparently not general to all bacterial groups. We note, however, that even in the most highly self-conflicting datasets, more than 75 per cent of the genes do not significantly reject the consensus tree—although we agree that these percentages would decrease if we added more species, as previously discussed in §3. Overall, agreement seems to be the rule and disagreement the exception. To illustrate this point, we simulated sequence datasets comparable in size to the real ones, using a random tree for every gene. This simulation mimics what data would look like if the HGT rate was high enough to erase the signal of speciations and extinctions. The average phylogenetic conflict in this simulated dataset is much higher than in any of the real datasets, indicating that vertical inheritance is a major component of the phylogenetic distribution across genes (the EEEP program could not be used for this dataset owing to excessively long running time). Our conclusions would be that HGTs significantly impact bacterial phylogenomics, even at a relatively recent time scale, but do not obviously preclude the reconstruction of species trees in this group.

6. HOW SHOULD WE ANALYSE AND REPRESENT PHYLOGENOMIC DATASETS KNOWING THERE ARE CONFLICTS?

Two viewpoints are prominent in current phylogenomic literature. The first one means focusing on conflicts, and denying the building of a species tree

as soon as significant disagreement is detected (Baptiste *et al.* 2005; Comas *et al.* 2006; Susko *et al.* 2006). The second view means averaging the phylogenetic signal across genes without considering the variance (basic supertree and supermatrix approaches). These two views are probably unnecessarily extreme. Both the vertical and horizontal components of the phylogenetic variance among genes are worth analysing, since they reflect distinct biological and historical processes, as discussed above.

During recent years, various approaches have been proposed to deal with phylogenomic incongruences in a more balanced way, and return both the vertical and horizontal signals to the user. Phylogenetic networks (Huson & Bryant 2006) are a natural option; conflicting internal edges are kept and represented, when classical supertree methods will only return the majoritary one. Networks, however, are ambiguous in that they do not distinguish, at least in their basic version, edges (nodes) corresponding to species divergence from edges (nodes) corresponding to HGT (but see Kunin *et al.* 2005). Their biological interpretation is therefore not always straightforward. Scornavacca *et al.* (in press) have recently introduced a supertree-like method that aims at identifying and resolving conflicts before combining trees. Significant incongruences are annotated and returned to the user, together with the estimated species tree. Leigh *et al.* (2008) proposed a method in which conflicts are measured in the likelihood framework. Genes are concatenated into supergenes and analysed jointly, only if congruent with each other. The method therefore returns several trees, one per subset of congruent genes. Very recently, N. Galtier & J. Dutheil (2008, personal communication) have developed an approach similar in spirit to that of Scornavacca *et al.* (in press)—resolving conflicts before merging datasets—but technically related to that of Leigh *et al.* (2008): conflicts are measured in the ML framework, as introduced in §5. The output is a tree in which ambiguously located taxa appear several times, indicating the various alternative positions supported by substantial subsets of genes. Suchard (2005), finally, proposed the statistically most satisfying method, modelling HGT and fitting the model to the data taken as a whole. This promising approach is computationally highly demanding, since it involves integrating the likelihood over all possible tree topologies for every gene. Suchard (2005) performed the integration numerically using Markov chain Monte Carlo in the Bayesian framework, for a small example dataset. Arvestad *et al.* (2004) also introduced a Bayesian approach to the gene tree/species tree problem, this time focusing on gene duplications.

The main issue all these methods have to address is to define the threshold of conflict above which two genes will be considered significantly incongruent. Scornavacca *et al.* (in press), and Dutheil and Galtier let the user define arbitrary bootstrap or ML thresholds, while Suchard (2005) and Leigh *et al.* (2008) use more rigorous statistical procedures. One important problem comes from the fact that incongruence can occur for many reasons whose biological interpretations are very different. Integrated methods

such as that of Suchard (2005), for instance, will probably overestimate the rate of HGT, because this parameter recapitulates every possible source of conflict between genes, including systematic tree-building errors due to departures from the model assumptions. So paradoxically enough, one of the major methodological challenges in phylogenomics might belong to the classical one gene—one tree area: we need appropriate models of sequence evolution to correctly estimate gene trees and their associated statistical supports and appropriately interpret phylogenetic conflicts between gene trees and species trees.

REFERENCES

- Arvestad, L., Berglund, A. C., Lagergren, J. & Sennblad, B. 2004 Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proc. Eighth Annu. Int. Conf. of RECOMB*, pp. 326–335. New York, NY: ACM.
- Baptiste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. L. & Doolittle, W. F. 2005 Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.* **5**, e33. (doi:10.1186/1471-2148-5-33)
- Baptiste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J. & Doolittle, W. F. 2008 Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryote phylogeny. *Mol. Biol. Evol.* **25**, 83–91. (doi:10.1093/molbev/msm229)
- Beiko, R. G. & Hamilton, R. 2006 Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.* **6**, e15. (doi:10.1186/1471-2148-6-15)
- Beiko, R. G., Harlow, T. J. & Ragan, M. A. 2005 Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 14 332–14 337. (doi:10.1073/pnas.0504068102)
- Castresana, J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552.
- Castric, V. & Vekemans, X. 2004 Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. *Mol. Ecol.* **13**, 2873–2889. (doi:10.1111/j.1365-294X.2004.02267.x)
- Comas, I., Moya, A., Azad, R. K., Lawrence, J. G. & Gonzalez-Candelas, F. 2006 The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. *Mol. Biol. Evol.* **23**, 2049–2057. (doi:10.1093/molbev/msl075)
- 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)
- Daubin, V., Moran, N. A. & Ochman, H. 2003 Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832. (doi:10.1126/science.1086568)
- Denamur, E. *et al.* 2000 Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**, 711–721. (doi:10.1016/S0092-8674(00)00175-6)
- Doolittle, W. F. 1999 Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129. (doi:10.1126/science.284.5423.2124)
- Doolittle, W. F. & Baptiste, E. 2007 Pattern pluralism and the tree of life. *Proc. Natl Acad. Sci. USA* **104**, 2043–2049. (doi:10.1073/pnas.0610699104)
- Dufayard, J. F., Duret, L., Penel, S., Gouy, M., Rechenmann, F. & Perrière, G. 2005 Tree pattern matching in

- phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21**, 2596–2603. (doi:10.1093/bioinformatics/bti325)
- Edwards, S. V., Liu, L. & Pearl, D. K. 2007 High-resolution species trees without concatenation. *Proc. Natl Acad. Sci. USA* **104**, 5936–5941. (doi:10.1073/pnas.0607004104)
- Eickbush, T. H. & Eickbush, D. G. 2007 Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**, 177–185. (doi:10.1534/genetics.107.071399)
- Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)
- Galtier, N. 2007 A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst. Biol.* **56**, 633–642. (doi:10.1080/10635150701546231)
- Ge, F., Wang, L. S. & Kim, J. 2005 The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol.* **3**, e316. (doi:10.1371/journal.pbio.0030316)
- Gillespie, J. H. 2000 Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155**, 909–919.
- Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. 2002 Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238.
- Guindon, S. & Gascuel, O. 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704. (doi:10.1080/10635150390235520)
- Hobbolt, A., Christensen, O. F., Mailund, T. & Schierup, M. 2007 Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* **3**, e7. (doi:10.1371/journal.pgen.0030007)
- Huson, D. H. & Bryant, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267. (doi:10.1093/molbev/msj030)
- Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. 2003 Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**, 1598–1602. (doi:10.1093/molbev/msg154)
- Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. 2006 Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231. (doi:10.1016/j.tig.2006.02.003)
- Klein, J., Sato, A. & Nikolaidis, N. 2007 MHC, TSP and the origin of species: from immunogenetics to evolutionary genetics. *Annu. Rev. Genet.* **41**, 281–304. (doi:10.1146/annurev.genet.41.110306.130137)
- Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. 2005 The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* **15**, 954–959. (doi:10.1101/gr.3666505)
- Lartillot, N. & Philippe, H. 2008 Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Phil. Trans. R. Soc. B* **363**, 1463–1472. (doi:10.1098/rstb.2007.2236)
- Lawrence, J. G. & Hendrickson, H. 2005 Genome evolution in bacteria: order beneath chaos. *Curr. Opin. Microbiol.* **8**, 572–578. (doi:10.1016/j.mib.2005.08.005)
- Leigh, J. W., Susko, E., Baumgartner, M. & Roger, A. J. 2008 Testing congruence in phylogenomic analysis. *Syst. Biol.* **57**, 104–115. (doi:10.1080/10635150801910436)
- Lerat, E., Daubin, V. & Moran, N. A. 2003 From gene trees to organosomal phylogeny in prokaryotes: the case of gamma-proteobacteria. *PLoS Biol.* **1**, e19. (doi:10.1371/journal.pbio.0000019)
- Liu, L. & Pearl, D. K. 2007 Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* **56**, 504–514. (doi:10.1080/10635150701429982)
- Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. 2002 Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**, 631–639.
- Nei, M. & Rooney, A. P. 2005 Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152. (doi:10.1146/annurev.genet.39.073003.112240)
- Niimura, Y. & Nei, M. 2006 Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *J. Hum. Genet.* **51**, 505–517. (doi:10.1007/s10038-006-0391-8)
- Pamilo, P. & Nei, M. 1988 Relationship between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583.
- Philippe, H., Lartillot, N. & Brinkmann, H. 2005 Multigene analysis of bilaterian anomalies corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* **22**, 1246–1253. (doi:10.1093/molbev/msi111)
- Pisani, D., Cotton, J. A. & McInerney, J. O. 2007 Supertrees disentangle the chimaerial origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760. (doi:10.1093/molbev/msm095)
- Ragan, M. A. 1992 Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1**, 53–58. (doi:10.1016/1055-7903(92)90035-F)
- Robinson, D. R. & Foulds, L. R. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
- Scornavacca, C., Berry, V., Douzery, E. J. & Ranwez, V. In press. PHYSIC IST: cleaning source trees to infer more informative supertrees. *BMC Bioinformatics*.
- 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)
- Suchard, M. A. 2005 Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics* **170**, 419–431. (doi:10.1534/genetics.103.025692)
- Susko, E., Leigh, J., Doolittle, W. F. & Baptiste, E. 2006 Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol. Biol. Evol.* **23**, 1019–1030. (doi:10.1093/molbev/msj113)
- Tajima, F. 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.
- Thomarat, F., Vivarès, C. P. & Gouy, M. 2004 Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. *J. Mol. Evol.* **59**, 780–791. (doi:10.1007/s00239-004-2673-0)
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680. (doi:10.1093/nar/22.22.4673)
- Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. 2007 Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61. (doi:10.1038/nature06107)