Inferring species trees given coalescence and reticulation

Habilitation

Dr. Michael D. Pirie

Institut für Organismische und Molekulare Evolutionsbiologie

Johannes Gutenberg-Universität, Mainz

Mainz, June 2017

Contents

Abstract

Hybridisation is an important process in plant evolution, its impact apparent at all levels from the genome duplications shared by all angiosperms through to recent speciation events. Patterns of inheritance caused by hybridisation (and other reticulate processes) appear to be incompatible with one of the fundamental tools of evolutionary biologists and ecologists seeking to understand the evolution of biological diversity: the phylogenetic tree. The current paradigm for inferring relatedness of organisms – the 'species tree' – can be summed up in one word: coalescence. Current analytical approaches serve to bias against inferring reticulate processes, even though they may be common and of direct importance both for the evolutionary process itself and for the performance of methods used to infer it. I present a brief account of current methodologies and draw on examples from both plant and virus datasets to illustrate the importance of reticulate processes in evolutionary inference. An example from danthonioid grasses shows how our inference of the direction and frequency of long distance dispersal events can be impacted by post-dispersal hybridisation; one from the northern heathers (genus *Erica*) shows the impact on interpretations of morphological evolution. In viruses, recombination can lead to pathogenic strains, and our ability to infer this process, even on human timescales, is dependent on correctly interpreting differences between gene trees. I illustrate an easily implemented approach that may be used to infer the sequence and timing of gene and genome divergences given conflict between individual gene trees, even when the processes underlying that conflict cannot be distinguished. It can in principle be applied to any group of organisms and can be extended to explicitly model both reticulation and coalescence without prior knowledge of the species tree topology.

Zusammenfassung

Hybridisierung ist in der Evolution von Pflanzen ein wichtiger Prozess. Das Ergebnis von Hybridisierung ist auf allen Ebenen erkennbar, von der allen Angiospermen gemeinsamen Genom-Vervielfältigung bis hin zu Hybridartbildung in der jüngeren Vergangenheit. Durch Hybridisierung (und andere retikulate Prozesse) verursachte Muster genetischer Ähnlichkeit erscheinen unvereinbar mit einem der wichtigsten Werkzeuge für das Verständnis der Evolution biologischer Vielfalt, dem phylogenetischen Stammbaum. Das derzeitige Paradigma in der Rekonstruktion der Verwandtschaft von Organismen - der „Species Tree“ (Artstammbaum) - kann in einem Wort zusammengefasst werden: „Coalescence“. Heutige analytische Ansätze tendieren dazu, retikulate Prozesse unterzubewerten, obwohl Retikulation häufig auftritt und sowohl für den Evolutionsprozess als auch für die Leistungsfähigkeit von Methoden für seine Rekonstruktion von großer Bedeutung ist.

Ich präsentiere einen kurzen Überblick über aktuelle Methoden für die Rekonstruktion von Evolution und verwende Pflanzen- und Virus-Datensätze, um die Bedeutung retikulater Prozesse zu illustrieren. Ein Beispiel aus den Danthonioiden Gräsern (Poaceae, Unterfamilie Danthonioideae) zeigt, wie Schlussfolgerungen über die Richtung und Häufigkeit von Fernausbreitung durch Hybridisierung nach der Ausbreitung beeinflusst werden können. Ein Beispiel aus den Heidekrautgewächsen (Gattung *Erica*) zeigt die Auswirkungen retikulater Prozesse auf die Interpretation morphologischer Evolution. In Viren kann Rekombination zu pathogenen Stämmen führen. Unsere Fähigkeit, diesen Prozess, der in vom Menschen beobachtbaren Zeiträumen abläuft, zu untersuchen, ist abhängig von der richtigen Interpretation von Unterschieden zwischen Genstammbäumen. Ich stelle einen einfach durchzuführenden Ansatz vor, mit dem Reihenfolge und Zeitrahmen der Gen- und Genomdivergenz erforscht werden können, wenn die einzelnen Genstammbäume sich widersprechen. Dieser Ansatz kann auch verwendet werden wenn sich die Prozesse, die den Widersprüchen zugrunde liegen, nicht ermitteln lassen. Der hier präsentierte Ansatz kann im Prinzip auf jede Organismengruppe angewendet werden, und er kann erweitert werden, um sowohl Retikulation als auch Coaleszenz ohne vorherige Kenntnis des Artstammbaums zu modellieren.

Species trees and gene trees

Organismal phylogenies, or "species trees", depict the sequence and timing of past speciation events that led to the species diversity we observe today. They are a fundamental tool for evolutionary biologists seeking to understand the conditions under which biological diversity originated, offering by means of e.g. ancestral state reconstruction (Harvey and Pagel, 1991; Pagel, 1999; Joy et al., 2016), or molecular dating approaches (Benton and Ayala, 2003; Magallon, 2004; Ho, 2014), insight into phenomena such as the origins of complex morphological traits or intercontinental distributions (as illustrated below). However, the species tree is also a result of evolutionary inference, a result that is influenced to a greater or lesser degree by our assumptions regarding exactly the evolutionary processes that we may wish to investigate.

Fig. 1 shows a hypothetical species tree with branches outlined in black. The tips of the branches represent extant species (X, Y and Z), whilst the lengths of the branches (the vertical axis) are proportional to time and their widths (horizontal axis) are proportional to the effective size of the populations. Reading this graph, we can follow the branches of the species tree from the common ancestors of the group at its root, up through a series of speciation events whereby ancestral populations diverged from one another (indicated with dashed lines and stars on the time axis), to the tips, observing the changes in population sizes in time and across different lineages. This tree is what we are interested in knowing and using. The data we generally have to hand, by contrast, are individual gene trees based on non-recombining DNA sequences, three of which are represented in Fig. 1 in orange, red and blue.

Given what we know about the evolutionary process, we should expect there to be a degree of discrepancy between our gene trees and the underlying species tree (Cavalli-Sforza, 1966; Kingman, 1982; Avise et al., 1987; Pamilo and Nei, 1988; Takahata, 1989; Doyle, 1992, 1997; Maddison, 1997; Wendel and Doyle, 1998; Nichols, 2001; Edwards, 2009). We generally infer gene trees from linked single-copy plastid or mitochondrial genome markers and from potentially unlinked, ideally low-copy nuclear markers. These gene trees can differ both as the result of various kinds of analytical artefacts and because of different biologically meaningful phenomena (Sanderson and Shaffer, 2002; Linder and Rieseberg, 2004).

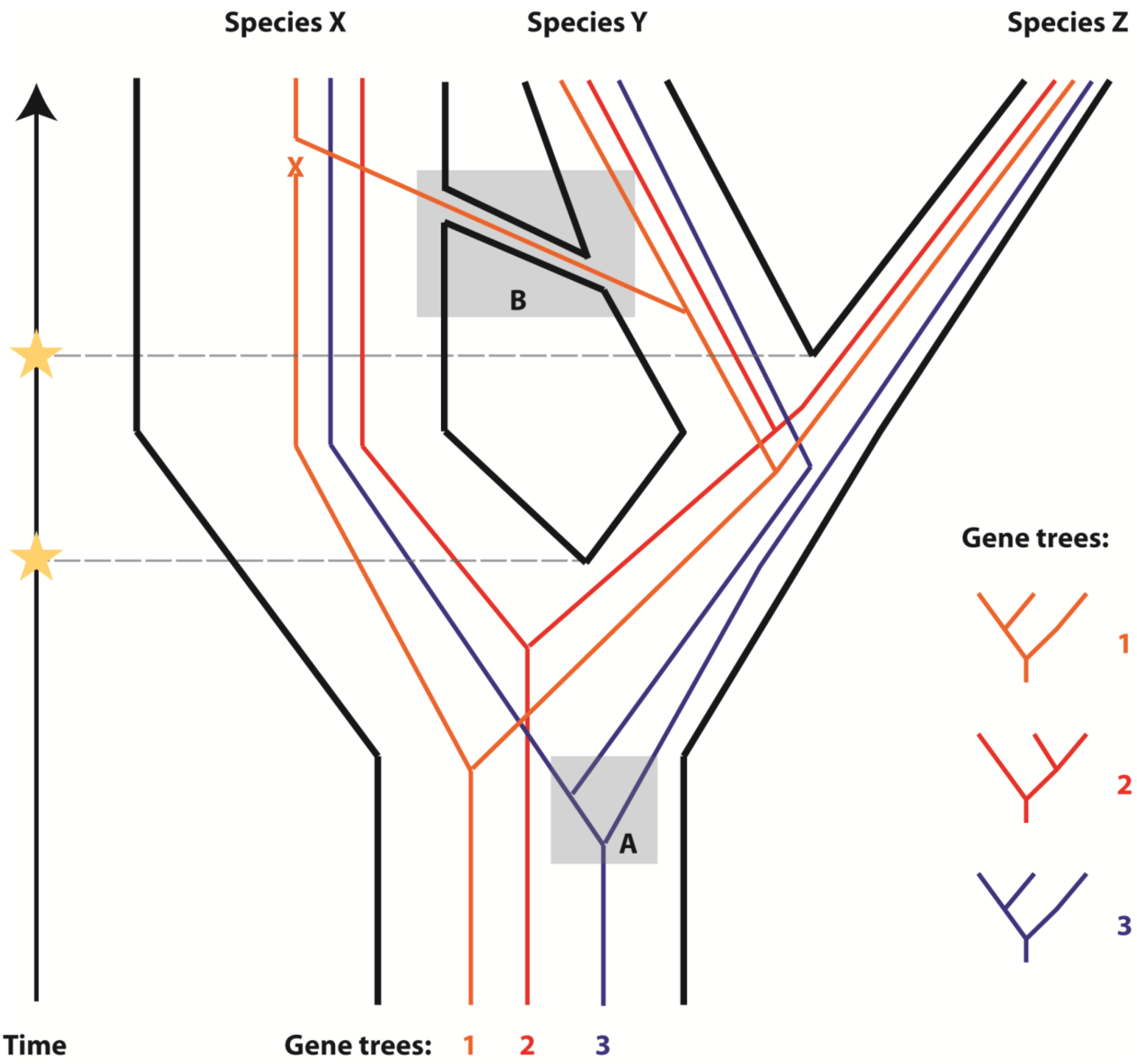


Fig. 1: A hypothetical species tree. The branches of the species tree are outlined in black and the tips of the branches represent extant species (X, Y and Z). Lengths of branches (the vertical axis) are proportional to time and their widths (horizontal axis) are proportional to the effective size of the populations. Speciation events are indicated with dashed lines and stars on the time axis. Three hypothetical individual gene trees are represented in orange, red and blue. The topology of gene tree 3 deviates from that of the species tree topology due to incomplete lineage sorting (A); that of gene tree 1 deviates due to reticulation (e.g. hybridisation) (B).

Sources of analytical artefacts include paralogy e.g. resulting from gene duplication (Page and Charleston, 1997), which is common in nuclear encoded sequences, but can also occur with putatively single copy plastid markers (Pirie et al., 2007); or from recombination within a given gene sequence (Posada and Crandall, 2002). They can also be caused by violations of model assumptions (Huelsenbeck et al., 1996) beyond that of the fundamental assumption of an underlying bifurcating tree. These may include the well-known impact of saturation and/or distant outgroups (e.g. Rosenfeld et al., 2012; Nosenko et al., 2013) as well as less well-known phenomena caused by deviations from assumptions implicit in increasingly complex likelihood based inference methods (e.g. rate distributions assumed in relaxed-clock dating approaches such as the widely used BEAST; Dornburg et al., 2012; Pirie and Doyle, 2012). These are nuisance factors in species tree inference. Other phenomena are potentially rather more informative, causing gene tree differences that are directly relevant for inference of the species tree.

In particular, we know that the stochastic process of coalescence causes the genetic diversity of populations to shift, generation by generation (Kingman, 1982; Doyle, 1992, 1997). “Coalescence” or “coalescent stochasticity” is the term for the random process whereby genotypes in populations with few interbreeding individuals (i.e. low “effective population size”; Ne) rapidly become fixed (the “genetic bottleneck” effect), whereas those with many interbreeding individuals (higher Ne) remain more heterogeneous for longer. Amongst the greater diversity of differing genotypes harboured in larger populations can be genes more typical of other closely related species. The resulting perplexing appearance of individuals of one species in the place of or amidst those of another in a given gene tree is known as “Incomplete Lineage Sorting” (ILS). The resulting conflict with other gene trees is sometimes referred to as a “lineage sorting artefact”. Implicit in the latter terms is that over time, with stochastic extinction of populations and lineages, gene tree topologies that deviate from the true order of speciation events will, if not altogether disappear, at least diminish greatly in frequency. However, such polymorphism can be maintained over multiple speciation events, for example because of selection (Saitou and Yamamoto, 1997), and under certain circumstances (described below) the majority of gene trees will in fact deviate from the species tree. Hence, due to coalescence, individual genes may not match the branching pattern of the underlying species tree (as illustrated in Fig. 1 A), and this effect may be apparent given both recent and more ancient speciation events.

Gene tree discordance can also be caused by reticulate processes, such as lateral/horizontal gene transfer (HGT) or hybridisation (illustrated in Fig. 1 B), which are also (long assumed) important phenomena in the evolutionary process (Anderson and Stebbins, 1954; Rieseberg, 1995). Genome-scale reticulate processes are prevalent across the tree of life (Dagan and Martin, 2006; Le et al., 2012; Martis et al., 2012). HGT is prevalent in prokaryotes but also important in eukaryotes (Schaack et al., 2010; Huang, 2013). The origins by endosymbiosis of eukaryotes (Sagan, 1967) was a reticulate process. Within eukaryote genomes, there is evidence for reticulate origins of a large proportion of noncoding DNA, representing mobile elements resulting from HGT (Schaack et al., 2010). At organismal level, although individual hybridisation events may be rare, up to 25% of plant and 10% of animal species are affected (Grant and Grant, 1992; Mallet, 2005; Whitney et al., 2010). The parental species tend to be phylogenetically clustered (Whitney et al., 2010), with the chances of successful hybridisation decreasing roughly exponentially with increasing genetic distance between parents (Mallet et al., 2007). Genes and traits can be obtained via hybridisation (Rieseberg et al., 2003; Rieseberg, 2009), and HGT can facilitate adaptation (Schönknecht et al., 2013). In a recent review paper, Abbott et al. (2013) discuss in depth hybridisation and its link to speciation, the diversity of processes involved, including adaptive divergence, reinforcement, and genetic/genomic factors, and the importance of their influence. The practical implications of reticulation are important for society, including the spread of antibiotic, herbicide (Ellstrand et al., 2013), and insecticide (Adler et al., 2010) resistance in pathogens and important pests.

Various fields of research (including some referred to above) assess directly the occurrence and impact of processes such as hybridisation in extant organisms. The same qualities of data are generally not available to biologists interested in evolutionary events at deeper timescales. Much of what we can learn about the historical evolutionary process must inevitably be inferred indirectly from patterns observed across extant organisms, notably including the mosaic of more or less differing phylogenetic signals encoded in their genomes, and the underlying “species tree” that these represent. A wide range of different approaches with different underlying assumptions are commonly employed to infer species trees from gene trees, which I will briefly review in the following sections.

Species trees from gene trees, 1: Concatenation, consensus and concordance

In reviewing the different general approaches to species tree inference from multiple gene trees, it seems appropriate to start with an approach that is often employed in the absence of gene tree discordance: matrix concatenation. This strategy is frequently employed in systematic studies above the species level. The data, representing multiple gene trees, is analysed as a single matrix assuming a single underlying bifurcating tree. When gene trees do not exhibit significant incongruence (i.e.

beyond that expected to result from the stochastic nature of DNA sequence evolution) and it is otherwise reasonable to assume that they track the same underlying species tree, combining the data may give an accurate hypothesis of relationships that is more precise than could be inferred from any individual gene tree alone (de Queiroz et al., 1995; Huelsenbeck et al., 1996; Kellogg et al., 1996; Leigh et al., 2008; Pirie, 2015).

However, where this assumption is violated, the negative impact on phylogenetic inference (both topological – the branching order - and branch lengths) is well known (McDade, 1992; Posada and Crandall, 2002; Mossel and Vigoda, 2005; Edwards et al., 2007; Kubatko and Degnan, 2007). As a general principle, "Total Evidence" (Kluge, 1989) also known as "simultaneous analysis" (Nixon and Carpenter, 1996) or "character congruence" (Mickevich, 1978) – matrix concatenation, irrespective of incongruence – is therefore not widely accepted for phylogenetic analyses of discrete gene trees (Bull et al., 1993; Huelsenbeck et al., 1996; Barker and Lutzoni, 2002; Hipp et al., 2004).

How can we assess when gene trees are incongruent? Testing for incongruence is a practical problem akin to the process of recombination detection (Martin, 2009), but simpler since the most likely breakpoints are already known (i.e. as the bounds of each independent marker). Various approaches are commonly used (Sanderson and Shaffer, 2002; Planet, 2006), including overall statistics such as the well-known "Incongruence Length Difference" (ILD) test based on parsimony tree length (Farris et al., 1994), likelihood-based approaches (Huelsenbeck and Bull, 1996; Leigh et al., 2008), distance based methods (Campbell et al., 2011), and methods based on topology and clade support (De Queiroz, 1993; Salichos et al., 2014). Considerable debate has arisen from the development of such tests and their application to empirical data (Wendel and Doyle, 1998; Barker and Lutzoni, 2002; Planet, 2006). One aspect of the debate is the reliability of particular methods, whereby the ILD test has come in for particular criticism (e.g. Barker and Lutzoni, 2002; Darlu and Lecointre, 2002) due to its high rates of Type I and Type II error (both false positive and false negative results). Despite its popularity (possibly driven by its relative ease of use with the long established software PAUP*; Swofford, 2003), this has led to a general distrust of ILD test results that either do not conform to expectations or are contradicted by other means, rendering it redundant at best (Pirie, 2015). If we regard gene tree incongruence as the result of evolutionary events that occurred in specific lineages, then the usefulness of overall metrics for congruence in general is limited: they do not indicate directly whether particular clades violate the assumption of an underlying bifurcating tree. Topology-based approaches by contrast can be used to directly identify specific examples of incongruence (MacLeod et al., 2005; Pirie, 2015). They are also easier to interpret, based on the generally understood properties of clade support values such as bootstrap support (Felsenstein, 1985; Hillis and Bull, 1993).

If gene tree conflict is identified, trees inferred from separate matrices can be summarised using consensus trees (Swofford, 1991), supertrees (Sanderson et al., 1998) or Bayesian concordance analysis (Cranston et al., 2009; Larget et al., 2010) (of which the latter can also be used to infer a species tree assuming coalescence). When the conflicting gene trees are only partially resolved, the power of this approach will be limited. Under such circumstances, particular taxa or markers might be considered to represent exceptions to otherwise consistent phylogenetic signal and excluded from the concatenated analyses to obtain a better resolved tree (Lecointre and Deleporte, 2005; Pirie et al., 2009). However, this assumption may be difficult to justify, and even if it is not obviously violated, the "conditional combination" approach (de Queiroz et al., 1995; Kellogg et al., 1996) does not take advantage of the information contained within gene tree differences, including differences in branch lengths (i.e. the relative timing of gene tree divergences).

Species trees from gene trees, 2: Coalescence-based methods

Gene tree discordance makes inferring species trees a challenge. However, it should not be regarded as some kind of barrier preventing us from inferring the 'true tree'. Rather, differing gene trees are our primary evidence for the complexity of processes underlying the species tree – exactly the processes we want to understand better. Assuming an underlying coalescent process, we might use the pattern of differences between gene trees – the degree of ILS in different parts of the tree – to infer not only the sequence and timing of speciation events, but also past effective population sizes (Degnan and Rosenberg, 2009) and even appropriate delimitation of species (Fujita et al., 2012; Harrington and Near, 2012).

This is the approach underlying a host of coalescence-based species tree inference methods. These include consensus methods such as STEAC (Liu et al., 2009), STAR (Liu et al., 2009), and MDC (Maddison, 1997; Than and Nakhleh, 2009), which take previously inferred phylogenetic trees (with or without branch length information) as input; and those implemented in popular applications *BEAST (Heled and Drummond, 2010) and BEST (Liu, 2008), that infer gene trees and corresponding species trees simultaneously from nucleotide alignments partitioned by linkage group. Further coalescence-based approaches have been developed to infer species trees directly from biallelic genetic markers, under the assumption that each is unlinked and inherited independently (e.g. SNAPP; Bryant et al., 2012).

Coalescence-based methods are not a panacea, even when the underlying assumptions are appropriate in principle. Within a so-called “anomaly zone”, where four or more lineages arose in quick succession, species trees will tend to converge on an incorrect solution (Degnan and Rosenberg, 2006). Inconsistent results can be obtained from different methods with the same data (Flórez-Rodríguez et al., 2011) for example given differing balances of mutational and coalescent variance in particular datasets (Huang et al., 2010). Phylogenetic uncertainty in individual gene trees is a potential source of error (Bayzid and Warnow, 2013; Mirarab et al., 2014). It can be incorporated directly into the results e.g. using methods that simultaneously estimate gene and species trees (Drummond and Rambaut, 2007; Boussau et al., 2013); or excluded, e.g. by means of a filtering procedure to remove genes with low signal (Salichos and Rokas, 2013); or incompletely resolved genes with apparently consistent phylogenetic signal can be identified and combined prior to analysis (statistical binning) (Bayzid and Warnow, 2013; Mirarab et al., 2014). Each of these approaches has apparent drawbacks, such as the potential for biasing the distribution of gene trees (sampling uninformative trees; failing to sample independent yet similar trees, or trees representing meaningfully low levels of divergence). Yang and Smith (2014) presented a gene jackknifing approach to examine sensitivity of results to gene inclusion, which may prove useful to assess the impact of this phenomenon in general.

The main conceptual and practical problem is that most current species tree inference methods – including all of the above – assume that gene tree differences are exclusively due to coalescent stochasticity. The exceptions (such as the method of Meng and Kubatko, 2009; and the "isolation with migration" model; Hey, 2010; as discussed below), are effectively limited to testing a limited range of reticulate scenarios. Similar to the impact of violations of assumptions given data concatenation, species trees inferred assuming coalescence are distorted by gene flow (Leaché et al., 2014; Reid et al., 2014). For example, many species tree techniques assume that the gene tree exhibiting the most recent divergence between taxon A and taxon B establishes a hard upper limit on the divergence time of those species in the species tree: gene flow subsequent to speciation would imply an erroneously recent age for the speciation event. Leaché et al. (2014) described two types of distortions to species trees caused by gene flow, in addition to impact on the topology and clade support values. “Species tree compression”, as described above, when the speciation times to appear more recent; and “species tree dilation”, when the population size is overestimated.

Some authors have argued against recombination (i.e. a special form of reticulation) being problematic for species tree inference (Lanier and Knowles, 2012), but this is of course to assume that representing a reticulate evolutionary scenario with a bifurcating tree is not a problem in itself. Irrespective of its impact on coalescence-based species tree inference, if hybridisation was an important factor in the evolution of the group, failing to represent that is clearly to throw away information.

Species trees from gene trees, 3: Network-based approaches

The dichotomy of the phylogenetic tree as a powerful descriptor of an evolutionary process that is nevertheless often distinctly non-treelike was described colourfully in a recently work by Koonin (2011). He wrote:

"… Trees are inalienable from any description of evolution, for the simple reason that replication of the genetic material is an intrinsically tree-like process"; "The reconstruction of the history of life (obviously, not the entire history, but its "skeleton") […] is not as simple as an analysis of the topology of the TOL [Tree of Life]. Instead, such a reconstruction requires charting the FOL [Forest of Life] in search of "groves" of similar trees that might reflect longterm trends of coherent (vertical) evolution of gene ensembles and "vines" of HGT."

Whereas the approaches described above are wedded to the dichotomous tree as a representation of the evolutionary process – irrespective of the assumptions made to infer that tree – various other approaches apply a different model: the network. A number of different methods have been developed for inferring networks from trees and/or molecular data (reviewed in Posada and Crandall, 2001; Linder and Rieseberg, 2004; Vriesendorp and Bakker, 2005; Huson and Bryant, 2006). These are often aimed at analyses below the species level, for which a bifurcating species tree is not relevant, and employ assumptions that may be inappropriate at higher levels. For example, the most frequently sampled haplotypes in a population level study may be ancestral, and appropriately represented as such in an unrooted haplotype network using TCS (Clement et al., 2000). However, given low molecular variation between species or higher taxa, a haplotype network may give a misleading impression of relationships, equivalent to grouping on the basis of symplesiomorphy (shared ancestral character states) rather than synapomorphy (shared derived states) (sensu Hennig, 1966).

Split decomposition (Bandelt and Dress, 1992; Huson and Bryant, 2006) is a popular method that summarises incompatibilities in data in the form of a network. The "splits" in question describe each branch of a tree or network as two lists of taxa: one on one side of the split, the other on the other. The collection of such splits implied by a given dataset or collection of trees [Konowalik et al. (2015) even used as input a matrix based on multiple gene trees coded according to the supertree "Matrix

Representation with Parsimony" approach (Ragan, 1992)] can be summarized with one or more networks using a number of different methods with differing underlying assumptions. Split decomposition is implemented in the frequently used SplitsTree (Huson, 1998; Huson and Bryant, 2006). Other network approaches involve adding reticulations to a tree (e.g. as implemented in T-REX; Boc et al., 2012) or reconciling multiple trees (e.g. median networks; Bandelt et al., 1995; or MacLeod et al., 2005; or the methods of Beiko and Hamilton, 2006). One limitation of reconciliation methods employed e.g. in Horizstory (MacLeod et al., 2005), "Efficient Evaluation of Edit Paths" (Beiko and Hamilton, 2006) and PhyloNet (Than et al., 2008) is the requirement of a guide tree – a hypothesis of the species tree with which gene trees are reconciled. Split decomposition and median networks were among a number of methods compared by Woolley et al. (2008), using a simulation approach. The results showed a perhaps surprising and alarming degree of failure of network approaches to correctly infer recombination, particularly at higher substitution rates.

No guide tree is required in order to summarise rooted hybridisation networks from separately inferred gene trees, e.g. using Dendroscope 3 (Huson and Scornavacca, 2012). Where different gene trees include different taxa, a supernetwork approach (Huson et al., 2004; Holland et al., 2008; Hassanzadeh et al., 2012) might be appropriate. However, an important limitation to network methods that summarise multiple sets of phylogenetic trees was described by Huber et al. (2015): "…even if we are given all of the subnetworks induced on all proper subsets of the leaves of some rooted phylogenetic network, we still do not have all of the information required to completely determine that network." Because a set of fully resolved conflicting gene trees could translate into different alternative networks, even with perfect gene tree data the species network could remain uncertain.

It is obvious that whilst assuming a species tree will fail to represent hybridisation, assuming a network in the face of coalescence will overestimate it (Yu et al., 2011). A natural step from inferring networks from gene trees is therefore to infer networks under models that also invoke the coalescent.

Species trees from gene trees, 4: Networks incorporating coalescence

A rather smaller number of methods have been developed to incorporate both reticulation and coalescence compared to those that consider just one of these processes. The "Filtered Supernetworks" approach (Holland et al., 2008) is similar to splits networks approaches in that it also employs a description of potential tree or network topologies based on "splits". The number of splits that can describe a single fully resolved bifurcating tree is equal to the number of taxa in the tree plus 1. Any additional splits will represent reticulations. In principle, all conflict in a dataset or between a set of trees could be represented by additional splits, but in the Filtered Supernetworks approach, the number of such additional splits is restricted. This filtering step is crucial. The authors claim that the approach can be used to distinguish incomplete lineage sorting from hybridization, that to this end the choice of supernetwork method (i.e. how to summarise multiple trees representing differing sets of taxa) was less important than the choice of filtering criteria (i.e. how to decide which splits to exclude). Count-based filtering was concluded to be the most effective. This effectively means assuming a set maximum number of splits and excluding those exceeding this number, which is to assume that we can know a priori how much conflict is worth representing as reticulation (even if we do not know to which incidences in particular this applies). This decision is not based on the factors influencing lineage sorting (i.e. effective population size, Ne; and time in numbers of generations) so it is not obvious why it should be effective.

Further approaches based on 'Minimizing deep coalescence' (Maddison, 1997) have been developed by Nakhleh, Than, Yu, and co-workers (Nakhleh et al., 2005; Than and Nakhleh, 2009; Nakhleh, 2011). Parsimony-based techniques developed in Yu et al. (2011); and Yu et al. (2013) were followed by a Maximum Likelihood (ML) application (Yu et al., 2014) that models simultaneously both coalescence and reticulation. One aspect of this method is a more effective means of finding optimal networks from the (in all but the simplest cases) immense number of potential solutions (‘tree-‘, or ‘network-space’). As with the Filtered Supernetworks approach, we can consider the minimum number of branches/splits represented by a fully resolved bifurcating tree. All potential resolutions of such a tree represent “Tree Space”, and although the number of such trees is generally far too great to be able to compare them all, it is standard practice to apply any of a number of heuristic approaches to approximate the best tree(s). Going from a fully resolved bifurcating tree, every additional branch that reticulates the tree results in an expansion of Tree Space into the even more complex “Network Space”. In a likelihood framework, it also adds a further parameter to the model. Rather than to waste computational effort searching through increasingly complex network space, Yu et al. (2014) proposed to test the likelihood of models with increasing numbers of parameters to see at which degree of reticulation the model fit no longer improves, and hence within which bounds of model complexity the heuristic search should proceed. Similar to the filtering step above, the criteria for this test are crucial, and in this case coalescence is explicitly incorporated into the model. However, there is still no objective and meaningful means to determine effective population size: in this approach, Ne is assumed to be proportional to mutation rate (which seems an over-simplification at best), and constant. A non-ultrametric tree is the result, reflecting real variation e.g. in Ne. A divergence time prior is also important for distinguishing reticulation from lineage sorting (Yu et al., 2011), a greater or lesser degree of uncertainty in which is inevitable.

A challenge for species tree inference approaches: discerning coalescent stochasticity from reticulation

Whilst each of these combined network/coalescence approaches clearly has its merits, it seems that the major limitation of each lies in the means used to distinguish between coalescent and reticulate processes. This is widely acknowledged to be a challenge (Maureira-Butler et al., 2008; Joly et al., 2009; de Villiers et al., 2013; Heled et al., 2013). In the ‘filtered supernetworks’ approach the filtering criterion is crucial; and the ML method of Yu et al. relies on an assumption linking Ne to the mutation rate. In the first case the investigator is effectively dictating the degree of reticulation to be inferred, and in the second this is derived more objectively, but arguably somewhat arbitrarily, from the data. An excellent review of the challenges involved and techniques available was presented by Anderson et al. (2012).

An alternative to tacking this difficult problem as an integral part of a given species tree inference method is to first identify reticulate patterns in the data and to exclude the corresponding taxa from species tree inference (assuming exclusively coalescence). The importance of removing hybrids from species tree inference has been further emphasised by various authors (Blanco-Pastor et al., 2012; Jackson and Austin, 2012). Where the identification of hybrids is successful, this approach will clearly avoid the above pitfalls associated with including them in species tree analyses.

Meng and Kubatko (2009) developed a model to test for hybrid speciation (using the likelihood ratio test), and to estimate the proportional contribution to the genome from each parental species. This can be used to test whether conflict in individual gene histories can be better explained by hybridization events in the presence of coalescence than by coalescence alone (Meng and Kubatko, 2009; Gerard et al., 2011), but only when the species phylogeny and location of the putative hybridization event are known. If the objective is to infer the species tree, these limitations are clearly a problem.

Another coalescence-based approach that can be used to test reticulation is the “isolation-with-Migration” model (Nielsen and Wakeley, 2001; Hey, 2010; Jackson and Austin, 2012; Melo-Ferreira et al., 2012). This multi-species coalescent model relaxes the assumption of reproductive isolation following speciation, whereby instead of a complete cessation, speciation is followed by period of gene flow. Importantly, while gene flow prior to the speciation event is dictated by the coalescent process and thus informative with regards Ne, post speciation gene flow (i.e. hybridisation) is not. However, where to draw the line – i.e. the timing of the speciation event – is in effect determined by a prior – the migration prior. Setting a value for this prior is a difficult problem, and on top of this, the model implies a bifurcating tree and can therefore only represent introgression between sister species/lineages.

Buckley et al. (2006) and Maureira-Butler et al. (2008) used coalescent simulations to identify cases of hybridisation/reticulation. However, coalescent simulations require prior knowledge of Ne through time and absolute age, which are exactly the kind of factors we might want to infer in the first place. The approach of Maureira-Butler et al. (2008) involves using the best estimate of the age of the group in question and simulating gene trees based on the inferred gene trees – representing the putative species tree – under a range of conceivable Ne. By plotting the simulated tree distances it can be seen how high Ne has to be before the simulated trees are as different one from another as the conflicting gene trees. Under this approach, coalescent stochasticity then serves as the null hypothesis: where tree differences modelled under coalescence overlap with observed differences, coalescent stochasticity is assumed to be the cause of the gene tree conflict.

Konowalik et al. (2015) developed a further approach, the “hybrid index”, aiming to test for the signal of hybridisation in individual taxa. In summary, it involves taking a network and breaking it down into its corresponding collection of three taxa “triplets” (two putative parents and one putative hybrid). The method of Yu et al. (2012) is then used with these triplets, with each taxon treated sequentially as hybrid or as parent, testing the fit to the data of a series of models that include a hybridisation parameter (γ). From this, the fit of models including a significant value for γ can be tested, and the value for γ can be summarised for all triplets including a particular focal taxon: the latter is the hybrid index. The approach is not independent of assumptions regarding age, generation time and Ne, which are represented in the model and branch lengths; and according to the authors may best be considered as a means of identifying gene tree conflict, rather than identifying its causes. In Konowalik et al. (2015), its use was followed by coalescent simulations to test whether particular instances of significant hybrid index values might nevertheless be explained by ILS alone.

The Achilles heel of such approaches is that coalescent simulations can only indicate the conditions under which gene tree differences could be explained by incomplete lineage sorting – they cannot be used to reject hybridisation. They are also limited by our potentially very vague notion of absolute ages and of population sizes through time. This means that where Ne and/or timescales are not known with confidence, and even where they are known, but populations were large and timescale short, we fall into the trap which we have termed the “coalescent stochasticity zone” (de Villiers et al., 2013). Under these circumstances, incomplete lineage sorting cannot be rejected and reticulate processes (such as hybridisation) will hence be systematically ignored. In de Villiers et al. (2013) we proposed alternative means to discern coalescent stochasticity (ILS) from reticulation that could also be used in combination with coalescent simulations, but are not dependent on them (as described below).

Towards an general solution for the species tree/gene tree problem

As I have noted above, reticulation can be important for the evolutionary process and has implications for the performance of species tree inference methods: it should not simply be ignored. Ideally, we would want to apply coalescence-based methods that take advantage of the information contained within gene tree differences, including differences in branch lengths (i.e. the relative timing of gene tree divergences) to infer both the sequence and timing of speciation events. We would also want to recover and represent reticulate evolutionary processes accurately and precisely, thus avoiding distortions to species trees caused by gene flow, but without having to know a priori anything about either the species tree or to what degree it might in fact be a species network. Currently, there is no single method available with which we can achieve this.

In the next section, I am going to use examples to demonstrate a number of approaches based on standard phylogenetic inference methods that, in combination, might bring us closer to the species tree/network of particular clades, despite the challenges outlined above. It includes an easily implemented supermatrix approach that may be used to infer the sequence and timing of gene and genome divergences given conflict between individual gene trees, even within the coalescent stochasticity zone. This approach, to my knowledge first used in Pirie et al. (2008), has been employed and developed in a number of subsequent publications by various authors (e.g. Pirie et al., 2009; Pelser et al., 2010; Blanco-Pastor et al., 2012; Visser et al., 2012; Pimentel et al., 2013; Maree et al., 2015; Mugrabi de Kuppler et al., 2015; Peterson et al., 2015). It is not subject to the same ambiguities involved in summarising multiple gene trees as networks, and can be extended to explicitly model both reticulation and coalescence without prior knowledge of the species tree topology.

Case study 1: The intercontinental diversification of the danthonioid grasses

My first example comes from the grass subfamily Danthonioideae, which comprises c. 280 species in 17 genera (examples illustrated in Fig. 2) mostly distributed across the continents of the Southern Hemisphere in temperate regions (Linder et al., 2010). Danthonioideae have been the subject of phylogeny-based analyses of character evolution (Galley and Linder, 2007; Humphreys et al., 2011) and biogeographic history (Pirie et al., 2010; Linder et al., 2013), and used for empirical tests of phylogenetic inference approaches (Pirie et al., 2009; Pirie et al., 2012). The phylogeny of Danthonioideae, and clades within it, has been inferred in a series of increasingly detailed analyses using plastid and/or nuclear ribosomal DNA sequences (Barker et al., 2003; Barker et al., 2007; Galley and Linder, 2007; Pirie et al., 2008; Humphreys et al., 2010; Linder et al., 2013). The data illustrates a phenomenon that commonly recurs in phylogenetic inference in plants: a chloroplast topology that conflicts with one based on a nuclear gene(s) (Fig. 3). We were faced with two problems: First, incomplete resolution of the nuclear ITS tree, which limits how we might use it; and second, how to analyse and interpret two differing hypotheses of phylogenetic relationships.

Fig. 2: Examples of species of the grass subfamily Danthonioideae. A. *Cortaderia pilosa* (d'Urv.) Hackel. Chile, Malleco National Park, Nahuelbuta, Ceno Anai (voucher: Pirie M.D. 344; 15/12/2005; Z). B. *Cortaderia sp.* Chile, Biobio National Park, Laguna del Laja (voucher: Pirie M.D. 353; 19/12/2005; Z). C. *Rytidosperma pictum* (Nees & Meyen) Nicora. Chile Biobio National Park, Laguna del Laja (voucher: Pirie M.D. 355; 19/12/2005; Z). D. *Rytidosperma gracile* (Hook. f.) Connor & Edgar. New Zealand, Hawkes Bay Land District, Whanahuia Range, Ruahine Forest Reserve (voucher: Humphreys A.M. 142; 18/2/2006; CHR). E. *Rytidosperma pulchrum* (Zotov) Connor & Edgar. New Zealand, Hawkes Bay Land District, Whanahuia Range, Ruahine Forest Reserve (voucher: Pirie M.D. 471; 18/2/2006; CHR). F. *Rytidosperma pauciflorum* (R.Br.) Connor & Edgar. Australia, Tasmania, Walls of Jerusalem National Park, Herod's Gate (voucher: Pirie M.D. 396; 16/1/2006; Z). G. *Notochloe microdon* (Benth.) Domin. Australia, New South Wales, Blue Mountains, Lason Town, Cataract Falls (voucher: Pirie M.D. 326; 29/11/2005; CANB). H. *Chionochloa frigida* (Vickery) Conert. Australia, New South Wales, Kosciuszko National Park, Charlottes Pass (voucher: Pirie M.D. 417; 26/1/2006; MEL). I. *Merxmuellera drakensbergensis* (Schweick.) Conert. South Africa, KwaZulu Natal, Sani Pass (voucher: Pirie M.D. 494; 16/10/2006; PRE).

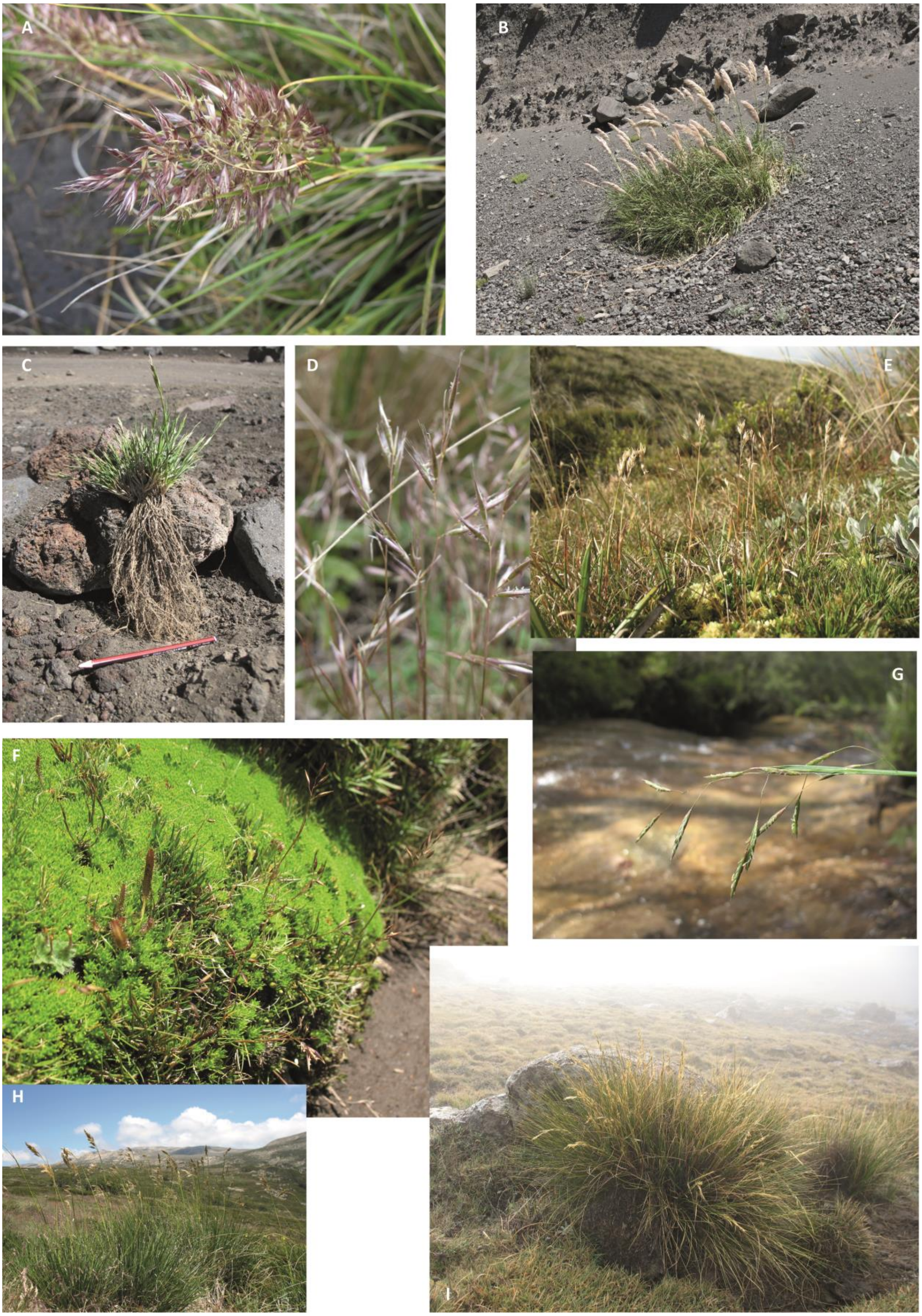
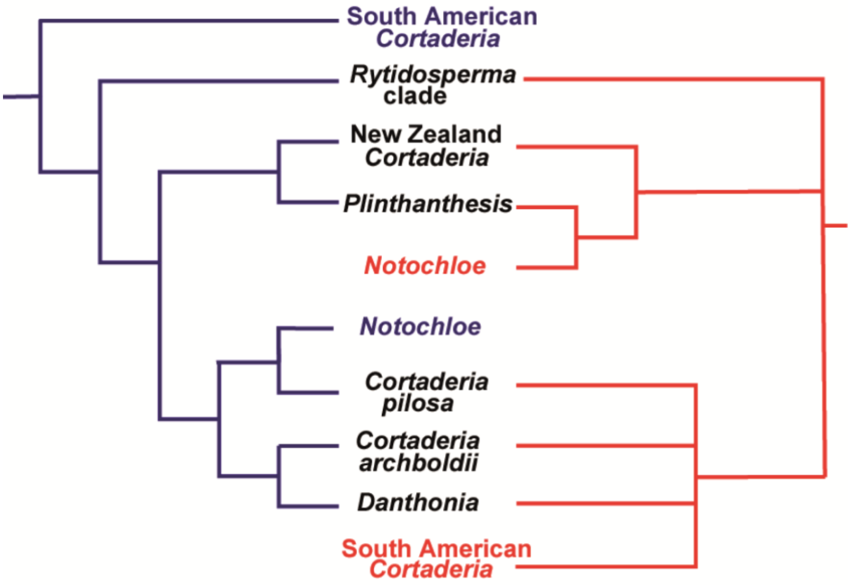
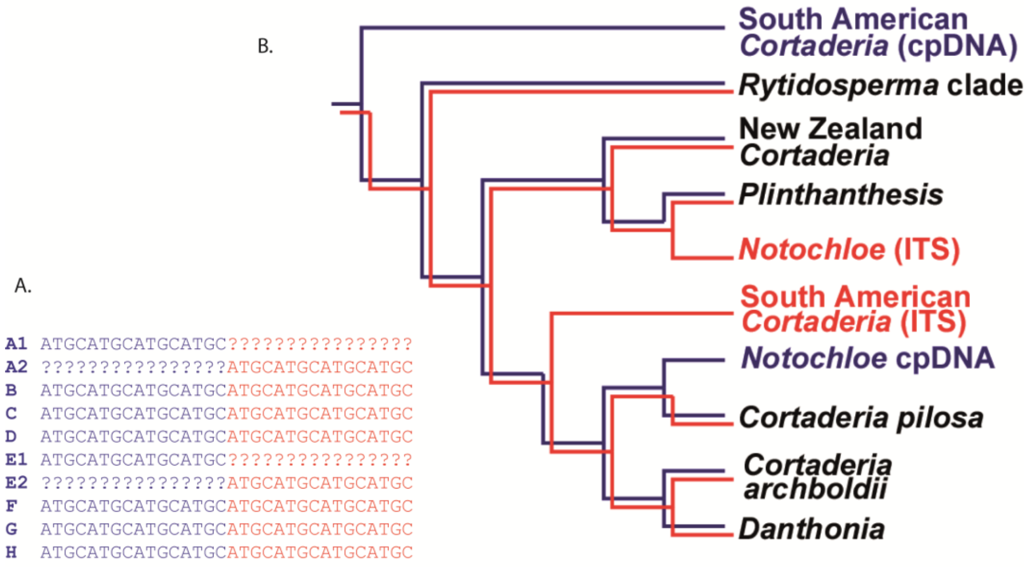

A
B
C
D
E
G
F
H
I

Fig. 3: Conflicting gene trees in Danthonioideae. Summaries of the plastid (left, in blue) and nuclear ribosomal ITS (right, in red) gene trees.



This first possible solution might be to remove the conflicting taxa and then combine the data. This results in increased support for the remaining clades which lends support to our expectation that the data partitions are tracking fundamentally the same underlying tree. It of course tells us nothing about the conflicting taxa, and nothing about the implications of the differences between the gene trees. In order to address this, the subsequent approach we took was to represent putative hybrid lineages as independent taxa in a single analysis. In practical terms, this entails making a duplicate of each of the conflicting taxa in your data matrix. For one of the duplicates you then replace one data partition with question marks, for the other, you replace the other partition with question marks. Analysis of this supermatrix (equivalent to the “compatibility matrix” that can be exported from the recombination detection package RDP3; Martin et al., 2010) results in a single so-called multi-labelled tree – that is, a tree in which some taxa and clades – the putative hybrids – are represented more than once, as can be seen in blue and red according to the chloroplast and ITS partitions respectively (Fig. 4). The same approach, with subsequent data concatenation, was used by Pelser et al. (2010); and in a coalescent framework (as expanded upon below) by Blanco-Pastor et al. (2012) and Pimentel et al. (2013). In this way we can represent a reticulate species history with a maximally resolved bifurcating tree, and use this for further inference with standard phylogenetic methods.

Fig. 4: Multilabelled "taxon duplication" phylogenetic tree of Danthonioideae. The schematic matrix (A) illustrates how sequence data representing conflicting phylogenetic signals of individual terminals is segregated into multiple 'duplicated' taxa (A1 and A2; E1 and E2). Analysed as a supermatrix, this results in a multilabelled tree (summarised in B) with the positions of duplicated taxa (South American *Cortaderia* and *Notochloe*) determined by the phylogenetic signals of individual partitions (gene trees) and those of non-duplicated taxa with congruent phylogenetic signals (and subtending nodes) determined by the combined data.

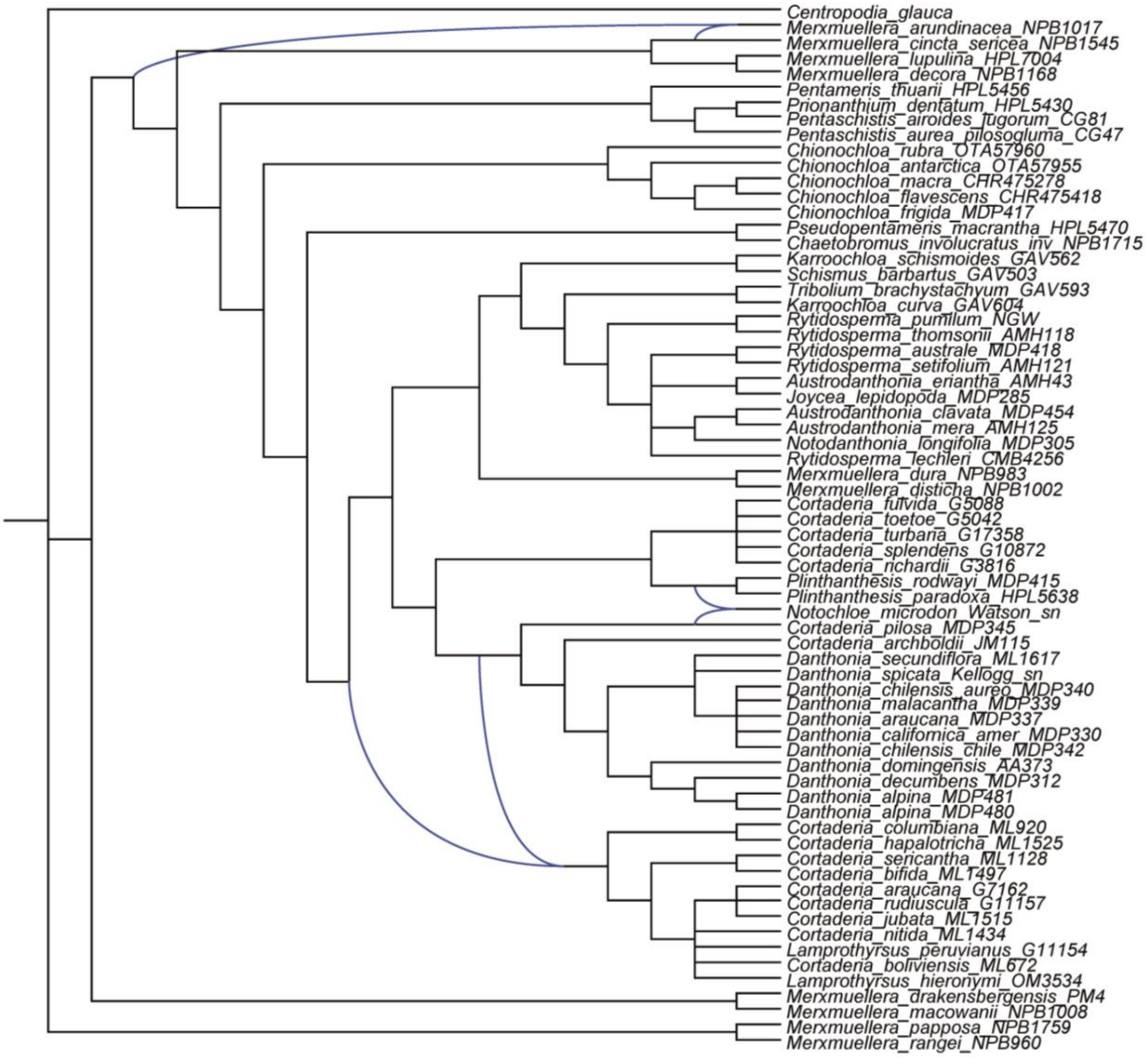


```
A1 ATGCATGCATGCATGC????????????????
A2 ????????????????ATGCATGCATGCATGC
B  ATGCATGCATGCATGCATGCATGCATGCATGC
C  ATGCATGCATGCATGCATGCATGCATGCATGC
D  ATGCATGCATGCATGCATGCATGCATGCATGC
E1 ATGCATGCATGCATGC????????????????
E2 ????????????????ATGCATGCATGCATGC
F  ATGCATGCATGCATGCATGCATGCATGCATGC
G  ATGCATGCATGCATGCATGCATGCATGCATGC
H  ATGCATGCATGCATGCATGCATGCATGCATGC
```

This approach is particularly useful for ancestral area reconstruction. It is obvious that hybrids cannot form in the physical absence of either parent, and this tree representation allows us to include this highly pertinent information directly in the analysis. This can be particularly useful when direct descendants of the parent species are not included in the analyses – through sampling limitations, or extinction for example. The only remaining evidence for particular evolutionary events may be in the conflicting gene trees of hybrid lineages. In extreme cases failing to recognise this can result in incorrect inference of ancestral areas and dispersal directions (Pirie et al., 2009).

In the case of the danthonioid grasses, it allowed us to infer the direction and timing of intercontinental dispersals including two that led to the origins of just such hybrid lineages (Pirie et al., 2009; Pirie et al., 2012; Linder et al., 2013). The pampas grass, genus *Cortaderia*, is of hybrid origin, resulting from two separate colonisations of the New World from Africa. The first of these colonisations leaves no other evidence in present day biota – the original parent lineage appears to be extinct. The case for hybridisation in Danthonioideae is strong – intercontinental dispersal is rare, so the occasional propagule that made it over must surely have resulted in a severe genetic bottleneck – this minimal Ne reduces the likelihood that coalescent stochasticity might explain gene tree conflict. Thus we might treat this multi-labelled tree as a representation of reticulate species tree – perhaps converting it into a rooted network, if this is a more intuitive representation (Fig. 5).

Fig. 5: Rooted phylogenetic network of Danthonioideae. A rooted network inferred from a multilabelled phylogenetic tree of Danthonioideae plastid and ITS sequence data using Dendroscope.


Centropodia_glauca
Merxmuellera_arundinacea_NPB1017
Merxmuellera_cincta_sericea_NPB1545
Merxmuellera_lupulina_HPL7004
Merxmuellera_decora_NPB1168
Pentameris_thuarii_HPL5456
Prionanthium_dentatum_HPL5430
Pentaschistis_airoides_jugorum_CG81
Pentaschistis_aurea_pilosogluma_CG47
Chionochloa_rubra_OTA57960
Chionochloa_antarctica_OTA57955
Chionochloa_macra_CHR475278
Chionochloa_flavescens_CHR475418
Chionochloa_frigida_MDP417
Pseudopentameris_macrantha_HPL5470
Chaetobromus_involucratus_inv_NPB1715
Karroochloa_schismoides_GAV562
Schismus_barbartus_GAV503
Tribolium_brachystachyum_GAV593
Karroochloa_curva_GAV604
Rytidosperma_pumilum_NGW
Rytidosperma_thomsonii_AMH118
Rytidosperma_australe_MDP418
Rytidosperma_setifolium_AMH121
Austrodanthonia_eriantha_AMH43
Joycea_lepidopoda_MDP285
Austrodanthonia_clavata_MDP454
Austrodanthonia_mera_AMH125
Notodanthonia_longifolia_MDP305
Rytidosperma_lechleri_CMB4256
Merxmuellera_dura_NPB983
Merxmuellera_disticha_NPB1002
Cortaderia_fulvida_G5088
Cortaderia_toetoe_G5042
Cortaderia_turbaria_G17358
Cortaderia_splendens_G10872
Cortaderia_richardii_G3816
Plinthanthesis_rodwayi_MDP415
Plinthanthesis_paradoxa_HPL5638
Notochloe_microdon_Watson_sn
Cortaderia_pilosa_MDP345
Cortaderia_archboldii_JM115
Danthonia_secundiflora_ML1617
Danthonia_spicata_Kellogg_sn
Danthonia_chilensis_aureo_MDP340
Danthonia_malacantha_MDP339
Danthonia_araucana_MDP337
Danthonia_californica_amer_MDP330
Danthonia_chilensis_chile_MDP342
Danthonia_domingensis_AA373
Danthonia_decumbens_MDP312
Danthonia_alpina_MDP481
Danthonia_alpina_MDP480
Cortaderia_columbiana_ML920
Cortaderia_hapalotricha_ML1525
Cortaderia_sericantha_ML1128
Cortaderia_bifida_ML1497
Cortaderia_araucana_G7162
Cortaderia_rudiuscula_G11157
Cortaderia_jubata_ML1515
Cortaderia_nitida_ML1434
Lamprothyrsus_peruvianus_G11154
Cortaderia_boliviensis_ML672
Lamprothyrsus_hieronymi_OM3534
Merxmuellera_drakensbergensis_PM4
Merxmuellera_macowanii_NPB1008
Merxmuellera_papposa_NPB1759
Merxmuellera_rangei_NPB960

This was a fairly straightforward example, involving only two gene trees, and a fairly clear idea of the underlying processes. Particularly in the age of next generation sequencing, when datasets representing hundreds of independent gene trees are increasingly becoming available (Cronn et al., 2012; Harrison and Kidner, 2012), an obvious question is whether the approach works for more complex systems. My next example will illustrate this – in an entirely different organism.

Case study 2: The recent recombinant evolution of pathogenic Potato Virus Y

The 'organism' in question is a virus, 'potato virus Y' (PVY), and the example is from Visser et al. (2012). Many virus groups appear to have very recent origins and it is hypothesised that the genus *Potyvirus*, to which PVY gives its name, evolved along with human agriculture (Gibbs et al., 2008). PVY is a short single strand of RNA, coding for just 11 proteins, in a little coat-protein bag. Most strains of PVY make their way between plants on the proboscis of aphids. The aphids are infectious for less than a day – a brief window of opportunity, but enough to spread it across a potato field. The virus persists through the winter in potato tubers. Importantly, this includes the so called seed tubers that are commonly traded. Where multiple virus strains infect a host simultaneously there is the possibility for recombination during replication. This is a potentially disastrous occurrence: A number of different PVY strains are known, and whilst all reduce crop yields, the most damaging are known to be recombinant between those strains, and the recombinants have increased dramatically in occurrence in the last couple of decades. Some of these cause the so-called Potato Tuber Necrotic Ring Disease which renders crops unsaleable.

Although on the face of it PVY looks like a practical agricultural problem, in fact it represents an evolutionary phenomenon – albeit one happening on an apparently human timescale. If we want to prevent the origin and spread of pathogenic organisms we need to know how it occurs. Just as with radiations of other organisms on longer time-scales, the best data available to infer the evolutionary history of viruses is the unit of inheritance – in this case, RNA.

We could do all sorts with a phylogenetic tree of PVY – including inferring the geographical origin and history of spread of the virus and the timeframe in which it occurred. However, we could not do this with any of the phylogenetic trees previously published. Some represent a combined analysis of whole genomes assuming incorrectly that you can represent recombinant evolution with a simple bifurcating tree, and producing results that for example include a non-recombinant clade nested within a recombinant one – which clearly cannot be true – on the end of a spurious long branch (Hu et al., 2009). Other phylogenetic analyses were either limited to non-recombinant strains (e.g. Cuevas et al., 2012), or represent only parts of the genome (e.g. Galvino-Costa et al., 2012). These only tell part of the story and neither use all the data available nor directly address what is going on with the all-important pathogenic recombinant strains.

This is a somewhat different kind of data compared to my previous simple two-gene example. For genomes the size of PVY it is not a problem to sequence the whole thing. In Fig. 6 (modified from Fig. 4 of Visser et al., 2012) the horizontal bars represent different PVY genomes and the different shades represent the four major PVY strains. I have labelled the non-recombinant genomes only – the rest represent the different combinations of those genome types that we found. We arrived at this mosaic by using recombination detection software combined with phylogenetic analyses of the putatively non-recombinant regions of the genomes defined by these breakpoints. You can analyse each of these 12 non-recombinant regions separately without violating the assumption of a strictly bifurcating tree and this results in a sequence of 'gene' trees, 12 in this case compared to the two in my Danthonioideae example. These represent the changing phylogenetic signal as you move from one end of the genome to the other – so not only do we have gene trees, we know their order in the genome.

For the purpose of comparison, we summarised these 12 trees (having collapsed nodes subject to <70% bootstrap support) in a split decomposition network (Fig. 7, adapted from Visser et al., 2012, Fig. S1). The resulting highly reticulate network represents the (supported) conflicting splits, but I would argue that this is an uninformative result of uncertainty regarding the relationships of the recombinants that is unavoidable when using this network approach. The gene trees are individually poorly resolved, but improved gene tree resolution would likely increase rather than decrease the signal of conflict. Since we are working with whole genomes, no more sequence data is available in any case. We can however apply the taxon duplication approach to infer both a better resolved multi-labelled tree, and a more precise network. Although in principle the process is the same as for the two-gene example, there is an added complication: some of the recombinant strains have not just recombined once, they have recombined several times, and so we need to extend the approach to accommodate multiple homologous recombination events in particular taxa and clades.

Fig. 6: Recombination map of PVY genomes. Modified from Fig. 4 of Visser et al. (2012): http://dx.doi.org/10.1371/journal.pone.0050631.g004 . The horizontal bars represent different PVY genomes and the different shades represent the four major PVY strains. The non-recombinant genomes only are labelled (PVY^{O}: white; PVY$^{N\text{-North America}}$: dark grey; PVY$^{N\text{-Europe}}$: black; and PVY^{NONPOT}: light grey. The rest represent the different combinations of those genome types that we found, with dashed red lines representing the 11 inferred recombination breakpoints and the intervening 12 non-recombinant genome regions numbered 1-12.

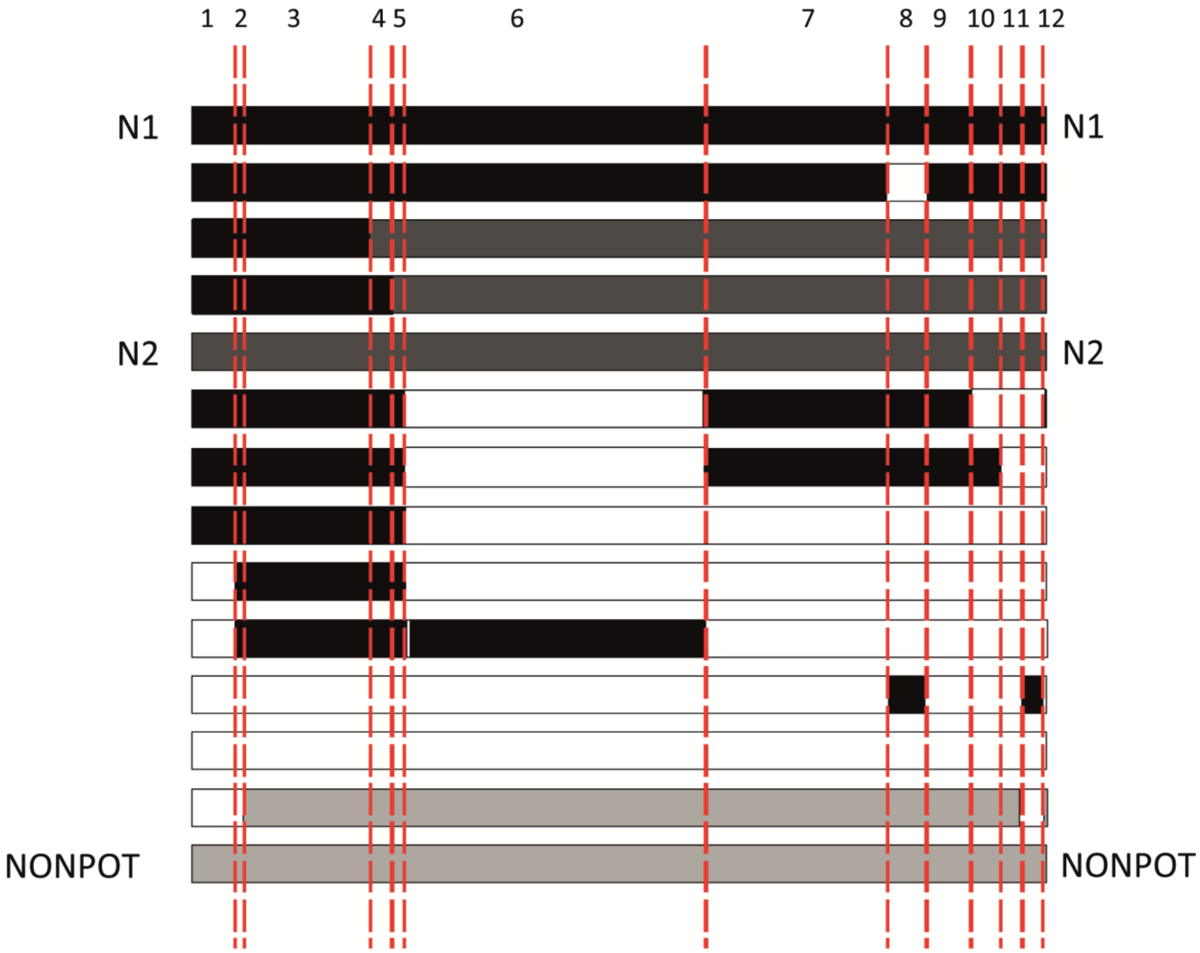
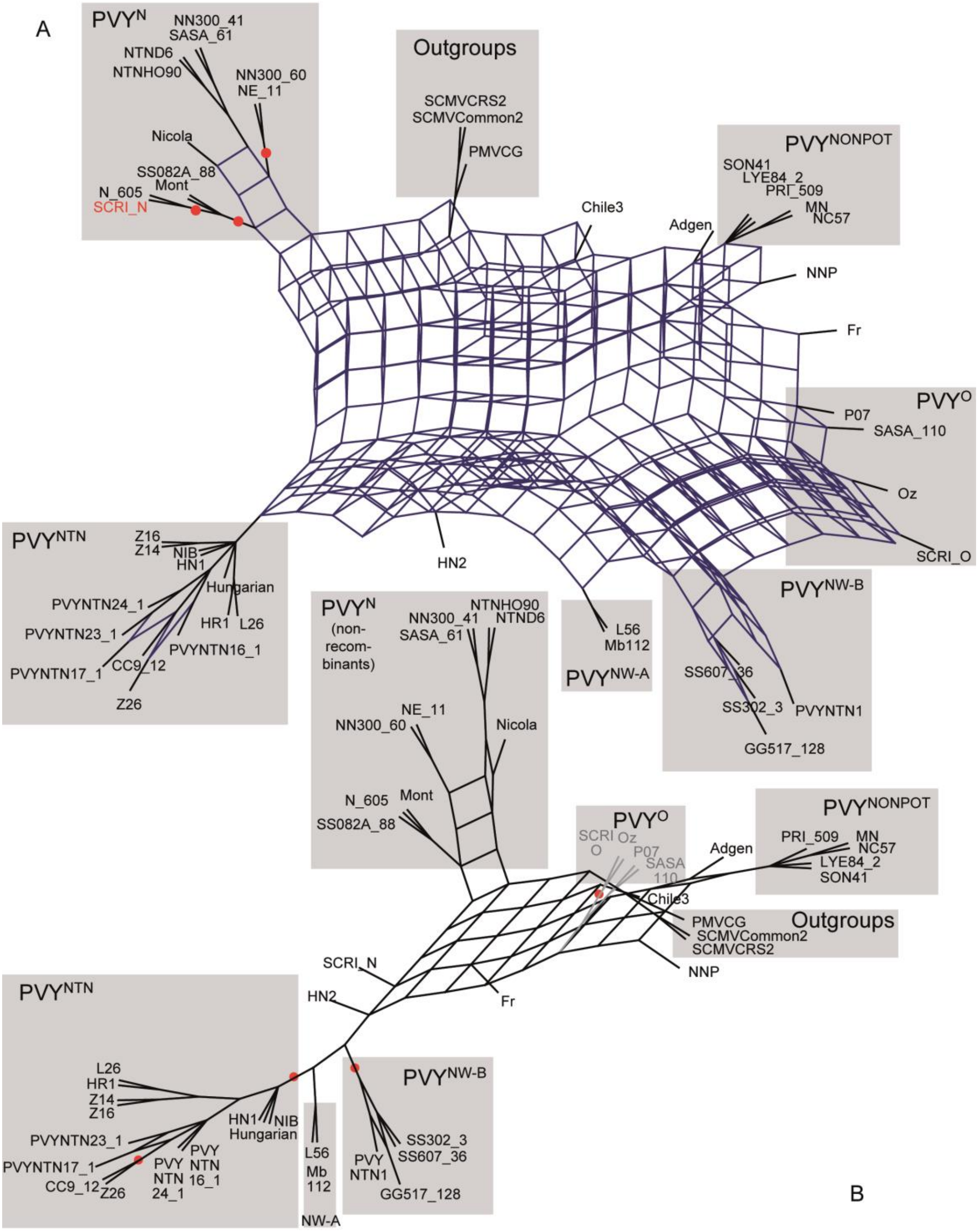


Fig. 7: Phylogenetic networks of PVY. Modified from Fig. S1 of Visser et al. (2012). Phylogenetic networks summarised using split decomposition in SplitsTree. A: Based on 12 70% BS consensus trees representing the contiguous, non-recombinant genome regions numbered and bounded by the breakpoints illustrated in Fig. 6. B: Based on a single 70% BS consensus of the multi-labelled 'genome' tree. Nodes recovered in one network but contradicted in the other are indicated with red dots on the corresponding branches. Major PVY strains and recombinant clades are indicated.

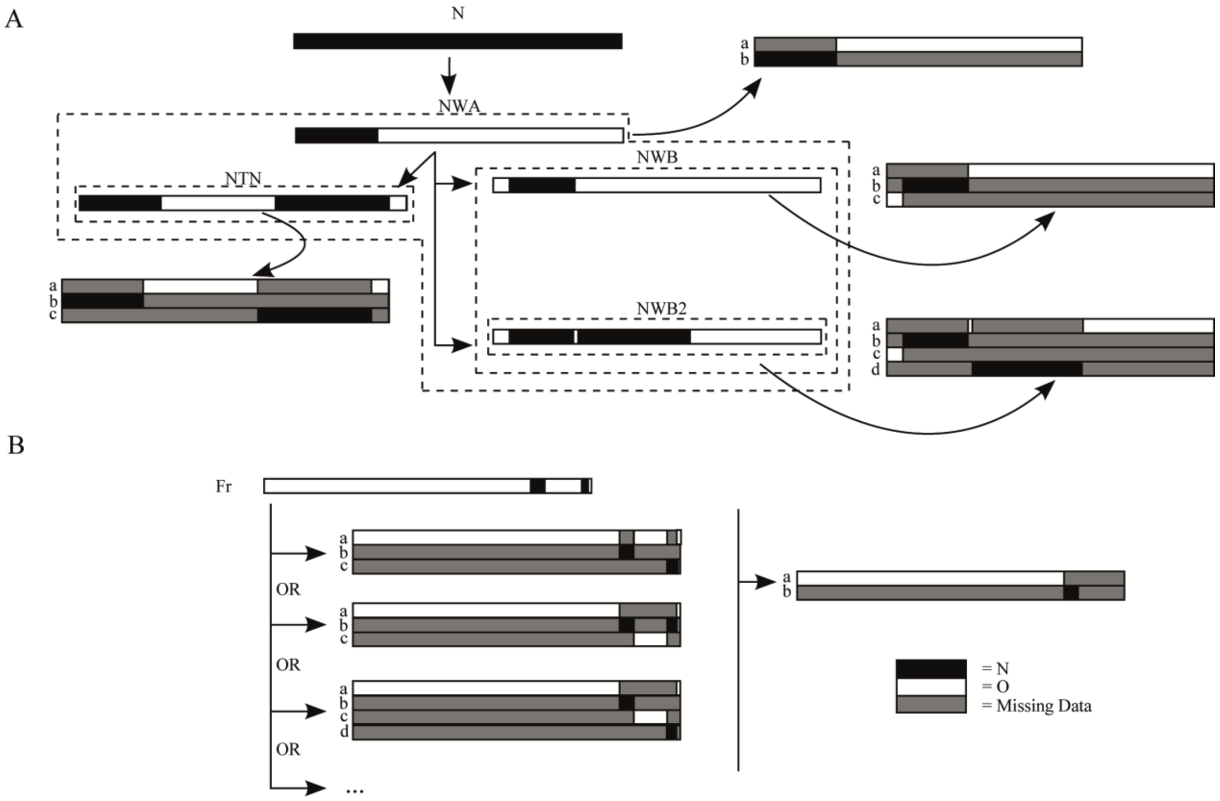


To do this first we infer the logical sequence of such events. The evidence for this is both shared breakpoints and monophyly of taxa sharing those breakpoints in the separate 'gene' trees, as illustrated in Fig. 8 (modified from Fig. 2 of Visser et al., 2012). Recombination patterns shared by a number of genomes suggest a single ancestral recombination event. We have two further recombination patterns that share a break point with the single recombinant. Assuming that these are double recombinants, we would expect them to be nested within single recombinant clades, as indicated by the dashed box. Similarly, triple recombinants should be nested within double recombinant clades and in the absence of further nested recombinants we would expect each shared derived recombination pattern to represent monophyletic groups of genomes. Having confirmed this by referring back to the gene trees we can then build a supermatrix by subdividing the recombinant genomes into their constituent phylogenetic signals. For single recombinants, the situation is similar to the chloroplast-nuclear conflict we have seen in the plant example. We create two duplicated taxa in the matrix. For double recombinants, we need to add one further duplicate taxon and for triple recombinants we need to add one more; four in total (Fig. 8).

We end up with a data matrix resembling a Swiss cheese, with single taxa for the non-recombinants and multiple taxa with varying proportions of missing data for the recombinants. In analysing this supermatrix we can infer the phylogeny of all parts of all the genomes simultaneously. Missing data per se is not problematic for phylogenetic inference (Wiens, 2003; Wiens et al., 2005; Jiang et al., 2014). However, the duplicate taxa of course share no characters in the sequence matrix and if the phylogenetic signal of the matrix as a whole is not sufficiently strong to place both (as in the case of recombination events involving closely related parents), this will cause a break-down in phylogenetic resolution (Visser et al., 2012).

The ideal solution would be to sample more non-recombinants that might place the recombinant clades with greater precision. For various reasons (including extinction) this may not be possible. What we did was to impose monophyly constraints on the derived recombinants as evidenced by the unique patterns of breakpoints shared by multiple genomes. This resulted in improved resolution and support both between and within recombinant clades. The resulting shortest trees were the same length as those found without the constraint, indicating that we were not in this way imposing a suboptimal solution. Further justification for constraining monophyly of derived recombinant/reticulate clades might be argued e.g. in cases of increasing chromosome numbers in phylogenetically derived clades that might be assumed to have resulted in instant reproductive barriers (Barber et al., 2007; although reproductive isolation is not necessarily complete in such cases; e.g. Pinheiro and Cozzolino, 2013). Alternatively, shared breakpoints could be coded as additional binary characters, as in the common approach to indel coding (Simmons and Ochoterena, 2000), which could be analysed with the sequence data instead of constraining the topology.

Fig. 8: The taxon duplication approach and multiple recombinants. Modified from Fig. S1 of Visser et al. (2012): http://dx.doi.org/10.1371/journal.pone.0050631.g002. Recombinant genomes encode a mosaic of differing phylogenetic relationships. Depending on the pattern and sequence of recombination events, separate regions of a given genome may share a common history of inheritance whilst those immediately adjacent are more distantly related. Under the 'taxon duplication' approach, these distinct 'phylogenetic signals' are segregated into separate taxa in the data matrix. Precisely which genome regions should be combined and which should be analysed independently can be inferred from the logical sequence of recombination events. A: In the case of PVY^{NW} and PVY^{NTN}, single, double and triple recombinants are apparent from shared derived recombination patterns (indicated here by black and white bars) and confirmed by exclusive ancestry (monophyly; indicated here by dotted boxes) of the pertinent genome regions. These are treated as two, three and four taxa respectively, as indicated, with the rest of the alignment re-coded as missing data. B: In the case of isolate Fr, lower recombinants are not known and phylogenetic signal is not sufficiently strong to discern congruence from conflict across the genome, thus the data could logically be combined in a number of different ways. In this case, the shorter of the non-contiguous regions are excluded from further analyses.



It is hard to compare the 12 original gene trees to the one genome tree, but in Fig. 7 beneath the original network (Fig. 7 a) is a second network (Fig. 7 b) computed in the same way but just from the single multi-labelled genome tree. From the marked reduction in conflict represented in Fig. 7 b compared to Fig. 7 a, it would seem that a significant proportion of the previous network structure (in blue) was due to uncertainty in the placement of the recombinants. Huber et al. (2015) showed that a collection of fully resolved differing gene trees could be represented by a number of different networks. The corresponding fully resolved multilabelled tree, by contrast, could be represented by just one network.

The multi-labelled tree might also be summarised as a rooted network, as in the Danthonioideae example in Fig. 5. However, the multi-labelled tree is a useful representation: we can use it to infer simultaneously the multiple phylogenetic relationships of the recombinants and the timing of the recombination events, as represented by the stem and crown group ages of recombinant clades (Visser et al., 2012). What we could learn from this was that the origins of the major strains of PVY appears to be subsequent to the introduction of potatoes to Europe, recombination in general occurred in the last century, and the multiple recombination events leading to the most pathogenic strain, PVY^{NTN} occurred within the last 50 years. As in our biogeographic analyses of Danthonioideae (Pirie et al., 2009; Pirie et al., 2012; Linder et al., 2013), the multilabelled tree could also be used to infer how the disease spread (although in this case denser taxon sampling representing isolates from across the known distribution would be required).

We have subsequently used the general approach with another economically important virus, grapevine leafroll-associated virus 3 (Maree et al., 2015), and there is no particular reason why it should not work for more closely studied recombinant systems such as influenza or HIV.

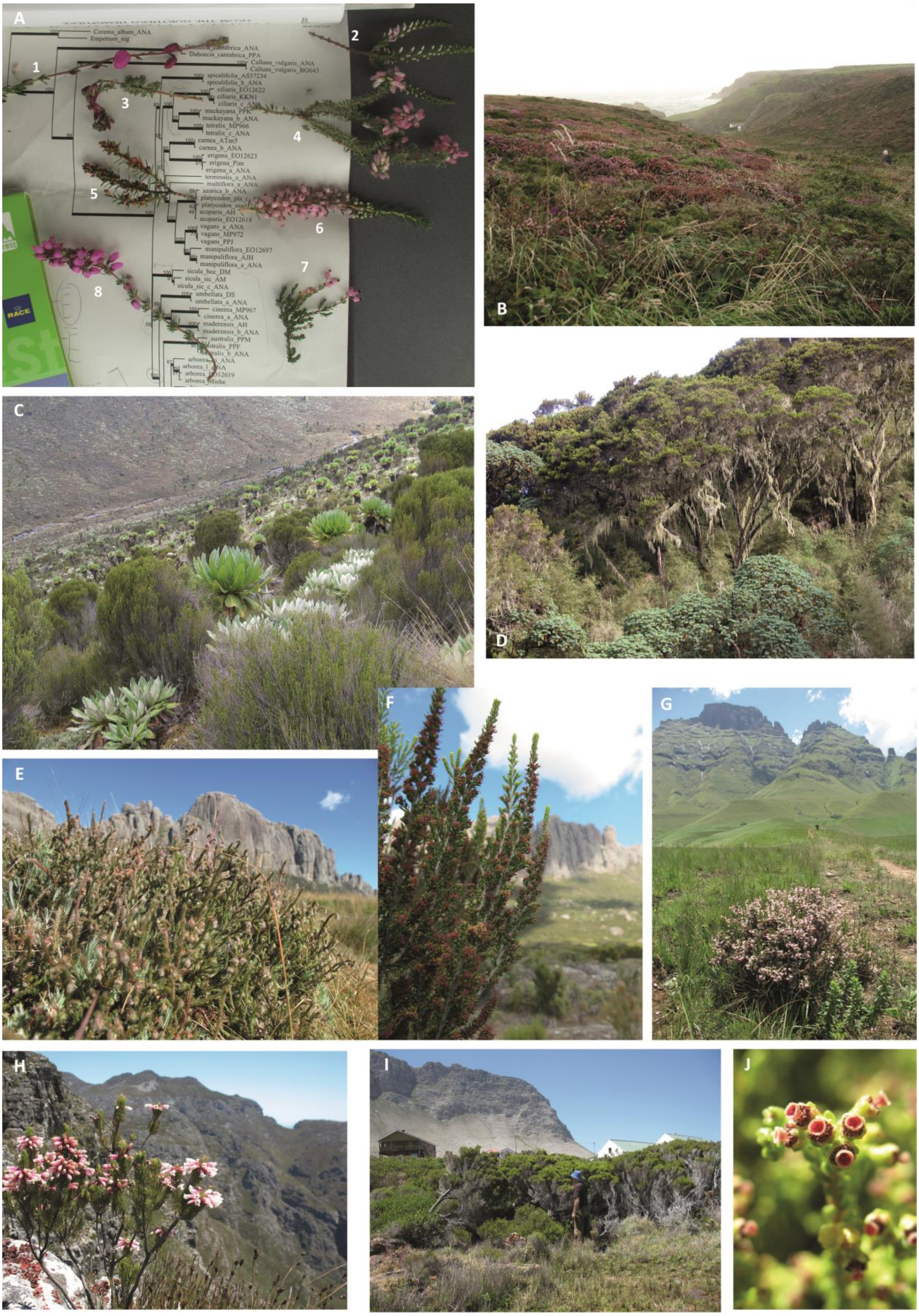
Case study 3: The hybrid origin of a tree heather, *Erica lusitanica*

I have shown how you can use the taxon duplication approach to infer reticulate phylogenies, representing hybridisation between genera of Danthonioideae and recombination between strains of PVY. The resulting multi-labelled tree could in principle represent either reticulation or incomplete lineage sorting: there is no assumption of the underlying process involved. However, it is likely that at least some degree of gene tree conflict in any multi-locus dataset will be the result of coalescent stochasticity, and under such circumstances we might wish to use the power of species tree inference methods better to infer the sequence and timing of speciation events and associated population level parameters.

My next case study represents an attempt to do exactly that. It comes from our ongoing work (Pirie et al., 2011; Pirie, 2012; Van der Niet et al., 2014; Mugrabi de Kuppler et al., 2015) on the phylogeny and evolution of *Erica* L. (Figs. 9 and 10), one of the largest genera of flowering plants (Frodin, 2004), comprising 830-840 species (Oliver and Oliver, 2003; Oliver and Forshaw, 2012). Most species of *Erica* are endemic to the Cape Floristic Region (CFR) of South Africa (Fig. 9 H-J; Fig. 10), but species of *Erica* are also found in the high mountains across Africa (Fig. 9 C-D), Madagascar (Fig. 9 E-F) and the South African Drakensberg (Fig. 9 G), and both *Erica* and closely related *Calluna* Salisb. and *Daboecia* D.Don (Ericaceae; Ericeae) are typical elements of open landscapes of the western Palearctic: Europe and surrounding areas (Fig. 9 A-B), in both the Temperate and Mediterranean biomes. A recent study of ours focussed on the relationships of the latter Northern Hemisphere species (Mugrabi de Kuppler et al., 2015).

We inferred nuclear and plastid gene trees from multiple accessions of all 21 northern *Erica* species and, although these gave largely congruent results, we discovered well supported conflict in the position of one species: *Erica lusitanica*. Having confirmed that this phenomenon was consistent across different samples and insensitive to phylogenetic methods, the next question was whether it represented incomplete lineage sorting or reticulation. We applied the standard approach, simulating gene trees based on estimated ages and a range of different generation times and Ne, and failed to reject coalescent stochasticity. Given our inevitable uncertainty in estimating past values for Ne, and the potential in *Erica* for species to occur in large populations distributed across wide areas (Nelson, 2012), this was not so surprising.

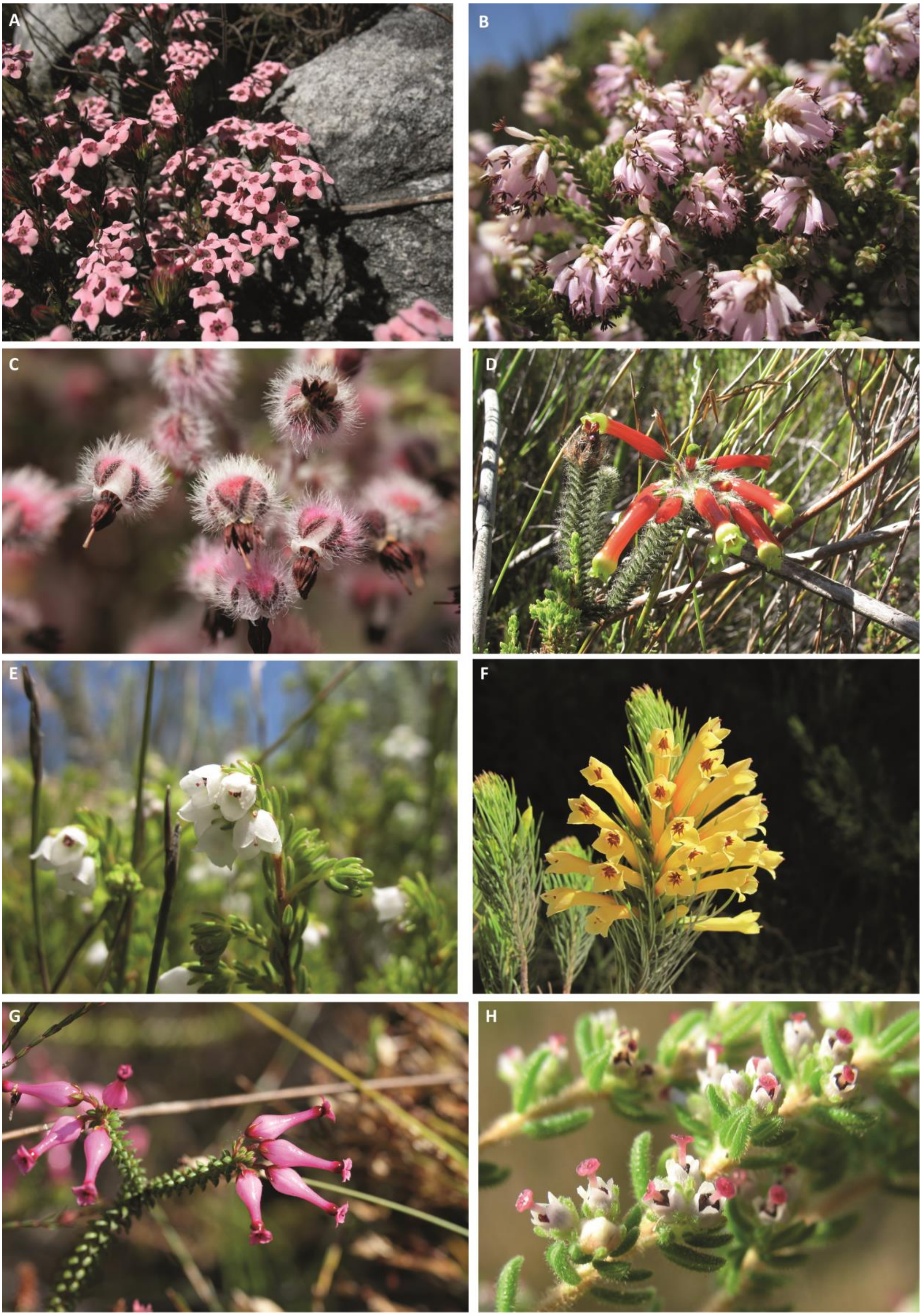
Fig. 9: *Ericeae* (Ericaceae) in Europe, across the African continent and in Madagascar. A: Eight species of Ericeae in Gallicia (Spain) displayed on a phylogenetic tree based on DNA sequence data (*Daboecia cantabrica* [1], *Calluna vulgaris* [2], *Erica ciliaris* [3], *E. mackayana* [4], *E. scoparia* [5], *E. vagans* [6], *E. umbellata* [7], and *E. cinerea* [8]). B: *C. vulgaris*, *E. cinerea* and *E. vagans* in sympatry on the Lizard, Cornwall, U.K.. C: Afroalpine landscape on Mt. Kenya with *Erica* and giant senecio. D: Tree heathers in the Ericoid zone, Abedares, Kenya. E: *Erica madagascariensis* (voucher: Oliver, EGH, 12658) and F: *E. sp.* cf. *goudotiana* (voucher Oliver, EGH, 12659), Andringitra National Park, Madagascar. G: *E. straussiana* (voucher: Pirie, MD, 638), South Africa, KwaZulu-Natal, Monk's Cowl. H: *E. viscaria ssp. longifolia* (voucher: Pirie, MD, 554), South Africa, Western Cape, Jonkershoek near Stellenbosch. I and J: *E. tristis* (voucher: Pirie, MD, 743), South Africa, Western Cape, Betty's Bay. I: Taxonomic expert E.G.H. (Ted) Oliver in the foreground; mountains of the Kogelberg Biosphere Reserve in the Background. J: Detail of wind-pollinated flowers. Photos: MDP; except C and D: Berit Gehrke.


A
Corema_album_ANA
Empetrum_nig
Daboecia_cantabrica_ANA
Daboecia_cantabrica_PPA
Calluna_vulgaris_ANA
Calluna_vulgaris_BG643
spiculifolia_AS57234
spiculifolia_b_ANA
ciliaris_EO12622
ciliaris_KKN1
ciliaris_c_ANA
mackayana_PPK
mackayana_b_ANA
tetralix_MP966
tetralix_c_ANA
carnea_ATsn5
carnea_b_ANA
erigena_EO12623
erigena_Pim
erigena_a_ANA
terminalis_a_ANA
multiflora_a_ANA
azorica_b_ANA
platycodon_pla_c
platycodon_ma
scoparia_AH
scoparia_EO12618
vagans_a_ANA
vagans_MP972
vagans_PPJ
manipuliflora_EO12697
manipuliflora_AJH
manipuliflora_a_ANA
sicula_boc_DM
sicula_sic_AM
sicula_sic_c_ANA
umbellata_DS
umbellata_a_ANA
cinerea_MP967
cinerea_a_ANA
maderensis_AH
maderensis_b_ANA
australis_PPM
australis_PPF
australis_b_ANA
arborea_b_ANA
arborea_I_ANA
arborea_EO12619
arborea_Miehe
1
2
3
4
5
6
7
8
RACE
B
C
D
E
F
G
H
I
J

Nevertheless, failure to reject incomplete lineage sorting does not in itself represent a rejection of reticulation, and two factors in particular suggested that a hybridisation scenario might still be worth considering: First, the lack of additional gene tree conflict. Although the standard simulations approach gave plausible values for past effective population sizes that might cause this particular gene tree conflict, at these values gene tree conflict would be expected to be rather more extensive than we observed. Second, a number of characteristics are shared between *E. lusitanica* and another species, *E. arborea*, which according to our nuclear marker might represent the sister species. The plastid phylogeny would imply that these characteristics – including tree-like habit, presence of lignotubers, similar floral morphology, and a distributional overlap – had originated independently. The alternative interpretation: that the plastid had been inherited from a more distantly related, morphologically dissimilar species following hybridisation (a 'chloroplast capture' like scenario); might explain the morphological similarity more parsimoniously. It also has implications for the biogeographic scenario of the group as a whole, specifically the ancestral area of *E. arborea*, which is widespread, also occurring in Tropical East Africa.

We therefore applied species tree analyses (using *BEAST) separately under both assumptions. *E. lusitanica* was either treated as a single taxon or as two separate taxa – represented by nuclear and plastid data respectively. The data partitions were then treated as independent gene trees with (the remaining) gene tree differences interpreted as resulting from coalescent stochasticity and used to infer a standard or multi-labelled species tree. The results were somewhat different, with the reticulation plus coalescence scenario supporting the sister-group relationship of *E. arborea* and *E. lusitanica* (for the nuclear partition) but the coalescence scenario not, implying two contrasting scenarios for character evolution and geographic range shifts.

Fig. 10: The floral morphological diversity of *Erica* (Ericaceae): examples from South Africa's Cape Floristic Region. A: *Erica fastigiata*, Jonkershoek (voucher: Pirie, MD, 555). B: *E. labialis*, Pringle Bay. C: *E. bruniades*, Cape of Good Hope Nature Reserve. D: *E. massonii*, Kogelberg Biosphere Reserve. E: *E. subdivaricata* (voucher: Pirie, MD, 739), Pringle Bay. F: *E. abietina ssp. perfoliosa* (voucher: Pirie, MD, 1071), Jonkershoek. G: *E. retorta*, Babylon's Tower. H: *Erica hispidula*, Limietberg Reserve, Bain's Kloof (voucher: Pirie, MD, 801). Photos: MDP.


A
B
C
D
E
F
G
H

The taxon duplication approach was used in species tree analyses previously by Blanco-Pastor et al. (2012), in an analysis of *Linaria* (Scrophulariaceae) and by Pimentel et al. (2013), in an analysis of *Anthoxanthum* (Poaceae). They first used coalescent simulations to identify hybrids then represented those hybrids as multiple taxa. Where gene tree conflict fell within the range that might be expected under coalescence – under a range of potential Ne – single taxa were maintained. It is a logical extension of our concatenation approach in an otherwise standard coalescence framework, and useful if you are willing to assume that your identification of hybrids is accurate. I have already suggested that this may often not be the case if you rely exclusively on coalescent simulations.

Case study 4: Cape primroses in the coalescent stochasticity zone

The concerns about coalescent simulations and the 'coalescent stochasticity zone' are presented in de Villiers et al. (2013), an empirical study of South African *Streptocarpus* (Gesneriaceae). *Streptocarpus* comprises 178 species in Africa and Madagascar (Nishii et al., 2015). Species formerly classified under *Saintpaulia* are known as African violets (although they are not violets), others as Cape primroses (although they are not primroses). *Streptocarpus* are widely cultivated and crossed with ease. In analyses of ITS data, Möller and Cronk (2001a, 2001b) identified a 'Cape primrose clade' (CPC) of species native to Southern Africa, which was the subject of more detailed analyses by De Villiers et al. (2013). The gene trees that we inferred based on plastid and ITS sequences show extensive incongruence (de Villiers et al., 2013). The question is whether this could be the result of hybridisation in the wild. Age estimates for the genus are based on inferred evolutionary rates in other groups, since there is no direct fossil evidence, and range from 25 to 50 MYA (Möller and Cronk, 2001b). Since we also know little or nothing about likely population sizes through time, coalescent simulations are unlikely to ever reject incomplete lineage sorting.

The approach we took to identify likely hybrids given all this uncertainty included use of the Genealogical Sorting Index (GSI; Cummings et al., 2008) and an assessment of patterns of morphological inheritance.

For the former, as also applied by Konowalik et al. (2015), we assessed whether individuals of a species were significantly associated according to in this case two independent gene trees. The cases in which the association was significant according to one gene tree but not according to another were of particular interest: such a result could be compared to that expected given the differing Ne of markers from plastid and nuclear genomes. The smaller Ne single copy plastid would be expected to coalesce faster, and where it nevertheless shows lack of species association this might be an indication of reticulation.

For the latter, we reasoned that a character state that originated once (at least, within a particular clade) according to one gene tree in contrast to multiple independent origins implied by another gene tree was a pointer towards a reticulate scenario. The two characters that we compared were vegetative habit and floral morphology. A number of apparent independent shifts in habitat suggested candidates for chloroplast capture, whilst the notable absence of such conflict with the floral morphology character suggested that the pollination syndromes it represented were indeed acting as a reproductive barrier, preventing hybridisation between lineages with different morphologies. The limitations of the approach are that it will fail to identify hybrids when they are not associated with an apparent shift in morphology and that it will be less effective with characteristics that tend to show homoplasy even without reticulation. In combination with e.g. a simulation approach, however, it promises to increase the chances of identifying hybrids and hence being able to handle them appropriately in subsequent species tree inference.

Conclusions and future prospects

Both the available data and the methods for species tree inference have in recent years undergone a revolution. With high throughput “next generation” sequencing we can obtain greater volumes of data at an ever decreasing cost (Harrison and Kidner, 2012). For the purposes of accurate species tree inference, we should be using these methods to obtain data that will allow us to infer numerous, meaningfully-resolved independent gene trees. At the moment, the most efficient means of achieving this is probably through approaches, such as hybrid enrichment (Cronn et al., 2012; Lemmon et al., 2012), used to target longer and more variable contiguous sequences known to be low copy in clades of interest. With this data, we can apply a range of methods that can implement potentially highly complex models incorporating partitioned substitution and relaxed clock model parameters and coalescence and even reticulate processes. The theoretical challenge that must still adequately be met is to reliably distinguish these processes.

Methods that incorporate both reticulate and coalescent processes simultaneously using genome scale data are a tempting prospect (Szöllősi et al., 2015). However, currently available approaches are either inappropriate for analyses of clades of multiple closely related species (i.e. the most challenging phylogenetic problems such as in the case of rapid radiations) (Joly et al., 2009; Meng and Kubatko, 2009) or distinguish the processes arbitrarily (Holland et al., 2008; Yu et al., 2014). It may be that in such cases multiple step approaches will be necessary, first to identify potentially hybridising sister species, then to test what the underlying processes might be, and finally to infer the species tree/network based on those assumptions. For the first step, the taxon duplication supermatrix approach (Pirie et al., 2008; Pirie et al., 2009) is well suited for inferring hypotheses of relationships in the form of a multi-labelled "genome tree". The results are independent of assumptions regarding the processes underlying gene tree conflict (Visser et al., 2012) and can show the relationships of potential hybrids precisely, i.e. avoiding uncertainty resulting both from failing to combine congruent data and from the differing ways in which multiple trees can be summarised as networks.

Testing potential reticulation may involve coalescence-based simulations (Maureira-Butler et al., 2008) or methods such as migration with isolation (Joly et al., 2009), or use of other statistical tests (de Villiers et al., 2013; Konowalik et al., 2015) or evidence (such as patterns of morphological inheritance or geographic distribution; de Villiers et al., 2013; Mugrabi de Kuppler et al., 2015). Ideally, a combination of multiple approaches under a range of assumptions should be used and the results compared. Thereafter, the taxon duplication approach could also be applied in coalescence-based analyses (Blanco-Pastor et al., 2012; Pimentel et al., 2013; Mugrabi de Kuppler et al., 2015) to infer a more or less reticulate species trees. The results may not be conclusive, and different methods, particularly for identifying reticulation, may lead to different hypotheses of relationships. These hypotheses can nevertheless be used to ask evolutionary questions whilst taking into account more of the complexity of the past evolutionary process and the potential sensitivity of our inferences to the unavoidable uncertainty involved in inferring it.

Acknowledgements

Lars Chatrou and Paul Maas (Wageningen University) supervised my PhD in Utrecht. Various other supervisors and collaborators have supported my work and contributed to the ideas presented here (for which I however take responsibility) since then. In particular, Peter Linder and Aelys Humphreys at the University of Zurich; Dirk Bellstedt and Ted Oliver at the University of Stellenbosch; Joachim Kadereit at JGU Mainz; and my wife Berit Gehrke throughout.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C.A., Buggs, R., Butlin, R.K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S.H., Hermansen, J.S., Hewitt, G., Hudson, A.G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J., Martinez-Rodriguez, P., Möst, M., Mullen, S., Nichols, R., Nolte, A.W., Parisod, C., Pfennig, K., Rice, A.M., Ritchie, M.G., Seifert, B., Smadja, C.M., Stelkens, R., Szymura, J.M., Väinölä, R., Wolf, J.B.W., & Zinner, D. 2013. Hybridization and speciation. *J Evol Biol* 26:229-246. http://dx.doi.org/10.1111/j.1420-9101.2012.02599.x

Adler, P.H., Cheke, R.A., & Post, R.J. 2010. Evolution, epidemiology, and population genetics of black flies (Diptera: Simuliidae). *Infection, Genetics and Evolution* 10:846-865. http://dx.doi.org/10.1016/j.meegid.2010.07.003

Anderson, C.N., Liu, L., Pearl, D., & Edwards, S.V. 2012. Tangled trees: the challenge of inferring species trees from coalescent and noncoalescent genes. Pp. 3-28 in: Anisimova, M., (ed), *Evolutionary Genomics*. Humana Press. p 3-28.

Anderson, E., & Stebbins, G.L., Jr. 1954. Hybridization as an Evolutionary Stimulus. *Evolution* 8:378-388. http://dx.doi.org/10.2307/2405784

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., & Saunders, N.C. 1987. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics* 18:489-522. http://dx.doi.org/10.2307/2097141

Bandelt, H.-J., & Dress, A.W.M. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Mol Phylog Evol* 1:242-252. http://dx.doi.org/10.1016/1055-7903(92)90021-8

Bandelt, H.J., Forster, P., Sykes, B.C., & Richards, M.B. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-753.

Barber, J.C., Finch, C.C., Fransisco-Ortega, J., Santos- Guerra, A., & Jansen, R.K. 2007. Hybridisation in Macaronesian Sideritis (Lamiaceae): evidence from incongruence of multiple independent nuclear and chloroplast sequence datasets. *Taxon* 56:74-88.

Barker, F.K., & Lutzoni, F.M. 2002. The Utility of the Incongruence Length Difference Test. *Sys. Biol.* 51:625-637. http://dx.doi.org/10.1080/10635150290102302

Barker, N.P., Galley, C., Verboom, G.A., Mafa, P., Gilbert, M., & Linder, H.P. 2007. The phylogeny of the austral grass subfamily Danthonioideae: Evidence from multiple data sets. *Plant Sys. Evol.* 264:135-156. http://dx.doi.org/10.1007/s00606-006-0479-9

Barker, N.P., Linder, H.P., Morton, C.M., & Lyle, M. 2003. The paraphyly of *Cortaderia* (Danthonioideae; Poaceae): evidence from morphology and chloroplast and nuclear DNA sequence data. *Annals of the Missouri Botanical Garden* 90:1-24.

Bayzid, M.S., & Warnow, T. 2013. Naive binning improves phylogenomic analyses. *Bioinformatics* 29:2277-2284. http://dx.doi.org/10.1093/bioinformatics/btt394

Beiko, R., & Hamilton, N. 2006. Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology* 6:15. http://dx.doi.org/10.1186/1471-2148-6-15

Benton, M.J., & Ayala, F.J. 2003. Dating the tree of life. *Science* 300:1698-1700. http://dx.doi.org/10.1126/science.1077795

Blanco-Pastor, J.L., Vargas, P., & Pfeil, B.E. 2012. Coalescent simulations reveal hybridization and incomplete lineage sorting in Mediterranean *Linaria*. *PLoS ONE* 7:e39089. http://dx.doi.org/10.1371/journal.pone.0039089

Boc, A., Diallo, A.B., & Makarenkov, V. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research* 40:W573-W579. http://dx.doi.org/10.1093/nar/gks485

Boussau, B., Szöllősi, G.J., Duret, L., Gouy, M., Tannier, E., & Daubin, V. 2013. Genome-scale coestimation of species and gene trees. *Genome Research* 23:323-330. http://dx.doi.org/10.1101/gr.141978.112

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., & RoyChoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* 29:1917-1932. http://dx.doi.org/10.1093/molbev/mss086

Buckley, T.R., Cordeiro, M., Marshall, D.C., & Simon, C. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (Maoricicada Dugdale). *Sys. Biol.* 55:411-425.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L., & Waddell, P.J. 1993. Partitioning and combining data in phylogenetic analysis. *Sys. Biol.* 42:384-397.

Campbell, V., Legendre, P., & Lapointe, F.-J. 2011. The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. *BMC Evolutionary Biology* 11:64. http://dx.doi.org/10.1186/1471-2148-11-64

Cavalli-Sforza, L.L. 1966. Population Structure and Human Evolution. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 164:362-379. http://dx.doi.org/10.2307/75457

Clement, M., Posada, D., & Crandall, K.A. 2000. TCS: a computer program to estimate gene genealogies. *Mol Ecol* 9:1657-1659.

Cranston, K.A., Hurwitz, B., Ware, D., Stein, L., & Wing, R.A. 2009. Species Trees from Highly Incongruent Gene Trees in Rice. *Sys. Biol.* 58:489-500. http://dx.doi.org/10.1093/sysbio/syp054

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., & Udall, J. 2012. Targeted enrichment strategies for next-generation plant biology. *Am. J. Bot.* 99:291-311. http://dx.doi.org/10.3732/ajb.1100356

Cuevas, J.M., Delaunay, A., Visser, J.C., Bellstedt, D.U., Jacquot, E., & Elena, S.F. 2012. Phylogeography and molecular evolution of Potato virus Y. *PLoS ONE*. http://dx.doi.org/10.1371/journal.pone.0037853

Cummings, M.P., Neel, M.C., & Shaw, K.L. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* 62:2411-2422. http://dx.doi.org/10.1111/j.1558-5646.2008.00442.x

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

Darlu, P., & Lecointre, G. 2002. When Does the Incongruence Length Difference Test Fail? *Mol Biol Evol* 19:432-437.

De Queiroz, A. 1993. For consensus (sometimes). *Sys. Biol.* 42:368-372.

de Queiroz, A., Donoghue, M.J., & Kim, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* 26:657-681. http://dx.doi.org/10.1146/annurev.es.26.110195.003301

de Villiers, M.J., Pirie, M.D., Hughes, M., Möller, M., Edwards, T., & Bellstedt, D.U. 2013. An approach to identify putative hybrids in the 'coalescent stochasticity zone', as exemplified in the African plant genus *Streptocarpus* (Gesneriaceae). *New Phytologist* 198:284-300. http://dx.doi.org/10.1111/nph.12133

Degnan, J.H., & Rosenberg, N.A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. *PLoS Genet* 2:e68. http://dx.doi.org/10.1371/journal.pgen.0020068

Degnan, J.H., & Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *TREE* 24:332-340. http://dx.doi.org/10.1016/j.tree.2009.01.009

Dornburg, A., Brandley, M.C., McGowen, M.R., & Near, T.J. 2012. Relaxed Clocks and Inferences of Heterogeneous Patterns of Nucleotide Substitution and Divergence Time Estimates across Whales and Dolphins (Mammalia: Cetacea). *Mol Biol Evol* 29:721-736. http://dx.doi.org/10.1093/molbev/msr228

Doyle, J.J. 1992. Gene Trees and Species Trees: Molecular Systematics as One-Character Taxonomy. *Sys. Bot.* 17:144-163. http://dx.doi.org/10.2307/2419070

Doyle, J.J. 1997. Trees within Trees: Genes and Species, Molecules and Morphology. *Sys. Biol.* 46:537-553. http://dx.doi.org/10.2307/2413695

Drummond, A.J., & Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7:214. http://dx.doi.org/10.1186/1471-2148-7-214

Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1-19. http://dx.doi.org/10.1111/j.1558-5646.2008.00549.x

Edwards, S.V., Liu, L., & Pearl, D.K. 2007. High-resolution species trees without concatenation. *PNAS* 104:5936-5941. http://dx.doi.org/10.1073/pnas.0607004104

Ellstrand, N.C., Meirmans, P., Rong, J., Bartsch, D., Ghosh, A., de Jong, T.J., Haccou, P., Lu, B.-R., Snow, A.A., Neal Stewart, C., Strasburg, J.L., van Tienderen, P.H., Vrieling, K., & Hooftman, D. 2013. Introgression of Crop Alleles into Wild or Weedy Populations. *Annu Rev Ecol Evol Sys* 44:325-345. http://dx.doi.org/10.1146/annurev-ecolsys-110512-135840

Farris, J.S., Kallersjo, M., Kluge, A.G., & Bult, C. 1994. Testing Significance of Incongruence. *Cladistics* 10:315-319. http://dx.doi.org/10.1111/j.1096-0031.1994.tb00181.x

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.

Flórez-Rodríguez, A., Carling, M.D., & Cadena, C.D. 2011. Reconstructing the phylogeny of "Buarremon" brush-finches and near relatives (Aves, Emberizidae) from individual gene trees. *Mol Phylog Evol* 58:297-303. http://dx.doi.org/10.1016/j.ympev.2010.11.012

Frodin, D.G. 2004. History and concepts of big plant genera. *Taxon* 53:753-776. http://dx.doi.org/10.2307/4135449

Fujita, M.K., Leaché, A.D., Burbrink, F.T., McGuire, J.A., & Moritz, C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* 27:480-488. http://dx.doi.org/10.1016/j.tree.2012.04.012

Galley, C., & Linder, H.P. 2007. The phylogeny of the Pentaschistis clade (Danthonioideae, Poaceae) based on chloroplast DNA, and the evolution and loss of complex characters. *Evolution* 61:864-884. http://dx.doi.org/doi:10.1111/j.1558-5646.2007.00067.x

Galvino-Costa, S.B.F., Dos Reis Figueira, A., Camargos, V.V., Geraldino, P.S., Hu, X.J., Nikolaeva, O.V., Kerlan, C., & Karasev, A.V. 2012. A novel type of Potato virus Y recombinant genome, determined for the genetic strain PVYE. *Plant Pathology* 61:388-398. http://dx.doi.org/10.1111/j.1365-3059.2011.02495.x

Gerard, D., Gibbs, H.L., & Kubatko, L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evolutionary Biology* 11:291. http://dx.doi.org/10.1186/1471-2148-11-291

Gibbs, A.J., Ohshima, K., Phillips, M.J., & Gibbs, M.J. 2008. The Prehistory of Potyviruses: Their Initial Radiation Was during the Dawn of Agriculture. *PLoS ONE* 3:e2523. http://dx.doi.org/10.1371/journal.pone.0002523

Grant, P.R., & Grant, B.R. 1992. Hybridization of Bird Species. *Science* 256:193-197. http://dx.doi.org/10.1126/science.256.5054.193

Harrington, R.C., & Near, T.J. 2012. Phylogenetic and Coalescent Strategies of Species Delimitation in Snubnose Darters (Percidae: Etheostoma). *Sys. Biol.* 61:63-79. http://dx.doi.org/10.1093/sysbio/syr077

Harrison, N., & Kidner, C.A. 2012. Next-generation sequencing and systematics: What can a billion base pairs of DNA sequence data do for you? *Taxon* 60:1552-1566.

Harvey, P.H., & Pagel, M. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford.

Hassanzadeh, R., Eslahchi, C., & Sung, W.K. 2012. Constructing phylogenetic supernetworks based on simulated annealing. *Mol Phylogenet Evol* 63:738-744. http://dx.doi.org/10.1016/j.ympev.2012.02.009

Heled, J., Bryant, D., & Drummond, A. 2013. Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evolutionary Biology* 13:44. http://dx.doi.org/10.1186/1471-2148-13-44

Heled, J., & Drummond, A.J. 2010. Bayesian Inference of Species Trees from Multilocus Data. *Mol Biol Evol* 27:570-580. http://dx.doi.org/10.1093/molbev/msp274

Hennig, W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana.

Hey, J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol* 27:905-920. http://dx.doi.org/10.1093/molbev/msp296

Hillis, D.M., & Bull, J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Sys. Biol.* 42:182-192. http://dx.doi.org/10.2307/2992540

Hipp, A.L., Hall, J.C., & Sytsma, K.J. 2004. Congruence Versus Phylogenetic Accuracy: Revisiting the Incongruence Length Difference Test. *Sys. Biol.* 53:81.

Ho, S.Y. 2014. The changing face of the molecular evolutionary clock. *Trends Ecol Evol* 29:496-503. http://dx.doi.org/10.1016/j.tree.2014.07.004

Holland, B.R., Benthin, S., Lockhart, P.J., Moulton, V., & Huber, K.T. 2008. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol Biol* 8:202. http://dx.doi.org/10.1186/1471-2148-8-202

Hu, X., Meacham, T., Ewing, L., Gray, S.M., & Karasev, A.V. 2009. A novel recombinant strain of Potato virus Y suggests a new viral genetic determinant of vein necrosis in tobacco. *Virus Research* 143:68-76. http://dx.doi.org/10.1016/j.virusres.2009.03.008

Huang, H., He, Q., Kubatko, L.S., & Knowles, L.L. 2010. Sources of Error Inherent in Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing among Different Methods. *Sys. Biol.* 59:573-583. http://dx.doi.org/10.1093/sysbio/syq047

Huang, J. 2013. Horizontal gene transfer in eukaryotes: The weak-link model. *BioEssays* 35:868-875. http://dx.doi.org/10.1002/bies.201300007

Huber, K.T., Van Iersel, L., Moulton, V., & Wu, T. 2015. How much information is needed to infer reticulate evolutionary histories? *Syst Biol* 64:102-111. http://dx.doi.org/10.1093/sysbio/syu076

Huelsenbeck, J.P., & Bull, J.J. 1996. A likelihood-ratio test to detect conflicting phylogenetic signal. *Sys. Biol.* 45:92-98. http://dx.doi.org/10.2307/2413514

Huelsenbeck, J.P., Bull, J.J., & Cunningham, C.W. 1996. Combining data in phylogenetic analysis. *TREE* 11:152-158.

Humphreys, A.M., Antonelli, A., Pirie, M.D., & Linder, H.P. 2011. Ecology and evolution of the diaspore "burial syndrome". *Evolution* 65:1163-1180. http://dx.doi.org/10.1111/j.1558-5646.2010.01184.x

Humphreys, A.M., Pirie, M.D., & Linder, H.P. 2010. A plastid tree can bring order to the chaotic generic taxonomy of *Rytidosperma* Steud. s.l. (Poaceae). *Mol Phylog Evol* 55:911-928. http://dx.doi.org/10.1016/j.ympev.2009.12.010

Huson, D.H. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14:68-73. http://dx.doi.org/10.1093/bioinformatics/14.1.68

Huson, D.H., & Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies *Mol Biol Evol* 23:254-267. http://dx.doi.org/10.1093/molbev/msj030

Huson, D.H., Dezulian, T., Klopper, T., & Steel, M.A. 2004. Phylogenetic super-networks from partial trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1:151.

Huson, D.H., & Scornavacca, C. 2012. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Sys. Biol.* 61:1061-1067. http://dx.doi.org/10.1093/sysbio/sys062

Jackson, N.D., & Austin, C.C. 2012. Inferring the evolutionary history of divergence despite gene flow in a lizard species, *Scincella lateralis* (Scincidae), composed of cryptic lineages. *Biological Journal of the Linnean Society* 107:192-209. http://dx.doi.org/10.1111/j.1095-8312.2012.01929.x

Jiang, W., Chen, S.Y., Wang, H., Li, D.Z., & Wiens, J.J. 2014. Should genes with missing data be excluded from phylogenetic analyses? *Mol Phylogenet Evol* 80:308-318. http://dx.doi.org/10.1016/j.ympev.2014.08.006

Joly, S., McLenachan, Patricia A., & Lockhart, Peter J. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist* 174:E54-E70. http://dx.doi.org/10.1086/600082

Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T., & Poon, A.F.Y. 2016. Ancestral Reconstruction. *PLoS Comput Biol* 12:e1004763. http://dx.doi.org/10.1371/journal.pcbi.1004763

Kellogg, E.A., Appels, R., & Mason-Gamer, R.J. 1996. When genes tell different stories: the diploid genera of Triticeae (Gramineae). *Sys. Bot.* 21:321-347. http://dx.doi.org/10.2307/2419662

Kingman, J.F.C. 1982. On the Genealogy of Large Populations. *Journal of Applied Probability* 19:27-43. http://dx.doi.org/10.2307/3213548

Kluge, A.G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology* 38:7-25. http://dx.doi.org/10.2307/2992432

Konowalik, K., Wagner, F., Tomasello, S., Vogt, R., & Oberprieler, C. 2015. Detecting reticulate relationships among diploid Leucanthemum Mill. (Compositae, Anthemideae) taxa using multilocus species tree reconstruction methods and AFLP fingerprinting. *Mol Phylogenet Evol*. http://dx.doi.org/10.1016/j.ympev.2015.06.003

Koonin, E.V. 2011. The logic of chance: the nature and origin of biological evolution, 1st ed. FT Press Science, Upper Saddle River, New Jersey.

Kubatko, L.S., & Degnan, J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Sys. Biol.* 56:17 - 24. http://dx.doi.org/10.1080/10635150601146041

Lanier, H.C., & Knowles, L.L. 2012. Is Recombination a Problem for Species-Tree Analyses? *Sys. Biol.* 61:691-701. http://dx.doi.org/10.1093/sysbio/syr128

Larget, B.R., Kotha, S.K., Dewey, C.N., & Ané, C. 2010. BUCKy: Gene Tree / Species Tree Reconciliation with Bayesian Concordance Analysis. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/btq539

Le, P., Ramulu, H., Guijarro, L., Paganini, J., Gouret, P., Chabrol, O., Raoult, D., & Pontarotti, P. 2012. An automated approach for the identification of horizontal gene transfers from complete genomes reveals the rhizome of Rickettsiales. *BMC Evolutionary Biology* 12:243. http://dx.doi.org/10.1186/1471-2148-12-243

Leaché, A.D., Harris, R.B., Rannala, B., & Yang, Z. 2014. The Influence of Gene Flow on Species Tree Estimation: A Simulation Study. *Sys. Biol.* 63:17-30. http://dx.doi.org/10.1093/sysbio/syt049

Lecointre, G., & Deleporte, P. 2005. Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta* 34:101-117. http://dx.doi.org/10.1111/j.1463-6409.2005.00168.x

Leigh, J.W., Susko, E., Baumgartner, M., & Roger, A.J. 2008. Testing congruence in phylogenomic analysis. *Syst Biol* 57:104-115. http://dx.doi.org/10.1080/10635150801910436

Lemmon, A.R., Emme, S.A., & Lemmon, E.M. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Sys. Biol.* 61:727-744. http://dx.doi.org/10.1093/sysbio/sys049

Linder, C.R., & Rieseberg, L.H. 2004. Reconstructing patterns of reticulate evolution in plants. *Am. J. Bot.* 91:1700-1708. http://dx.doi.org/10.3732/ajb.91.10.1700

Linder, H.P., Antonelli, A., Humphreys, A.M., Pirie, M.D., & Wüest, R.O. 2013. What determines biogeographical ranges? Historical wanderings and ecological constraints in the danthonioid grasses. *Journal of Biogeography* 40:821-834. http://dx.doi.org/10.1111/jbi.12070

Linder, H.P., Baeza, M., Barker, N.P., Galley, C., Humphreys, A.M., Lloyd, K.M., Orlovich, D.A., Pirie, M.D., Simon, B.K., Walsh, N., & Verboom, G.A. 2010. A generic classification of the Danthonioideae (Poaceae). *Annals of the Missouri Botanical Garden* 97:306-364. http://dx.doi.org/10.3417/2009006

Liu, L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24:2542-2543. http://dx.doi.org/10.1093/bioinformatics/btn484

Liu, L., Yu, L., Pearl, D.K., & Edwards, S.V. 2009. Estimating Species Phylogenies Using Coalescence Times among Sequences. *Sys. Biol.* 58:468-477. http://dx.doi.org/10.1093/sysbio/syp031

MacLeod, D., Charlebois, R., Doolittle, F., & Bapteste, E. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evolutionary Biology* 5:27-27. http://dx.doi.org/10.1186/1471-2148-5-27

Maddison, W.P. 1997. Gene Trees in Species Trees. *Sys. Biol.* 46:523-536. http://dx.doi.org/10.1093/sysbio/46.3.523

Magallon, S. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences* 165:S7-S21. http://dx.doi.org/10.1086/383336

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* 20:229-237. http://dx.doi.org/10.1016/j.tree.2005.02.010

Mallet, J., Beltran, M., Neukirchen, W., & Linares, M. 2007. Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology* 7:28. http://dx.doi.org/10.1186/1471-2148-7-28

Maree, H.J., Pirie, M.D., Oosthuizen, K., Bester, R., Rees, D.J.G., & Burger, J.T. 2015. Phylogenomic Analysis Reveals Deep Divergence and Recombination in an Economically Important Grapevine Virus. *PLoS ONE* 10:e0126819. http://dx.doi.org/10.1371/journal.pone.0126819

Martin, D.P. 2009. Recombination Detection and Analysis Using RDP3. Pp. 185-205 in: Posada, D., (ed), *Bioinformatics for DNA Sequence Analysis*. Humana Press. p 185-205.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., & Lefeuvre, P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462-2463. http://dx.doi.org/10.1093/bioinformatics/btq467

Martis, M.M., Klemme, S., Banaei-Moghaddam, A.M., Blattner, F.R., Macas, J., Schmutzer, T., Scholz, U., Gundlach, H., Wicker, T., Simkova, H., Novak, P., Neumann, P., Kubalakova, M., Bauer, E., Haseneyer, G., Fuchs, J., Dolezel, J., Stein, N., Mayer, K.F., & Houben, A. 2012. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci U S A* 109:13343-13346. http://dx.doi.org/10.1073/pnas.1204237109

Maureira-Butler, I.J., Pfeil, B.E., Muangprom, A., Osborn, T.C., & Doyle, J.J. 2008. The Reticulate History of *Medicago* (Fabaceae). *Sys. Biol.* 57:466-482. http://dx.doi.org/10.1080/10635150802172168

McDade, L.A. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* 46:1329. http://dx.doi.org/10.2307/2409940

Melo-Ferreira, J., Boursot, P., Carneiro, M., Esteves, P.J., Farelo, L., & Alves, P.C. 2012. Recurrent introgression of mitochondrial DNA among hares (Lepus spp.) revealed by species-tree inference and coalescent simulations. *Syst Biol* 61:367-381. http://dx.doi.org/10.1093/sysbio/syr114

Meng, C., & Kubatko, L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology* 75:35-45. http://dx.doi.org/10.1016/j.tpb.2008.10.004

Mickevich, M.F. 1978. Taxonomic Congruence. *Sys. Biol.* 27:143-158. http://dx.doi.org/10.2307/2412969

Mirarab, S., Bayzid, M.S., Boussau, B., & Warnow, T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346. http://dx.doi.org/10.1126/science.1250463

Möller, M., & Cronk, Q.C.B. 2001a. Evolution of morphological novelty: a phylogenetic analysis of growth patterns in *Streptocarpus* (Gesneriaceae). *Evolution* 55:918-929. http://dx.doi.org/10.1111/j.0014-3820.2001.tb00609.x

Möller, M., & Cronk, Q.C.B. 2001b. Phylogenetic studies in *Streptocarpus* (Gesneriaceae): reconstruction of biogeographic history and distribution patterns. *Systematics and Geography of Plants* 71:545-555. http://dx.doi.org/10.2307/3668699

Mossel, E., & Vigoda, E. 2005. Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees. *Science* 309:2207-2209. http://dx.doi.org/10.1126/science.1115493

Mugrabi de Kuppler, A.L., Fagúndez, J., Bellstedt, D.U., Oliver, E.G.H., Léon, J., & Pirie, M.D. 2015. Testing reticulate versus coalescent origins of *Erica lusitanica* using a species phylogeny of the northern heathers (Ericeae, Ericaceae). *Mol Phylog Evol* 88:121-131. http://dx.doi.org/10.1016/j.ympev.2015.04.005

Nakhleh, L. 2011. Evolutionary phylogenetic networks: models and issues. Pp. 125-158 in, *Problem solving handbook in computational biology and bioinformatics*. Springer. p 125-158.

Nakhleh, L., Warnow, T., Linder, C.R., & John, K.S. 2005. Reconstructing Reticulate Evolution in Species—Theory and Practice. *Journal of Computational Biology* 12:796-811. http://dx.doi.org/10.1089/cmb.2005.12.796

Nelson, E.C. 2012. Hardy heathers from the Northern Hemisphere. Royal Botanic Gardens, Kew, Richmond, United Kingdom.

Nichols, R. 2001. Gene trees and species trees are not the same. *TREE* 16:358-364. http://dx.doi.org/10.1016/S0169-5347(01)02203-0

Nielsen, R., & Wakeley, J. 2001. Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach. *Genetics* 158:885-896.

Nishii, K., Hughes, M., Briggs, M., Haston, E., Christie, F., DeVilliers, M.J., Hanekom, T., Roos, W.G., Bellstedt, D.U., & Möller, M. 2015. *Streptocarpus* redefined to include all Afro-Malagasy Gesneriaceae: Molecular phylogenies prove congruent with geographical distribution and basic chromosome numbers and uncover remarkable morphological homoplasies. *Taxon* 64:1243-1274. http://dx.doi.org/10.12705/646.8

Nixon, K.C., & Carpenter, J.M. 1996. On simultaneous analysis. *Cladistics* 12:221-241. http://dx.doi.org/10.1111/j.1096-0031.1996.tb00010.x

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Muller, W.E.G., Nickel, M., Schierwater, B., Vacelet, J., Wiens, M., & Worheide, G. 2013. Deep metazoan phylogeny:

When different genes tell different stories. *Mol Phylog Evol* 67:223-233. http://dx.doi.org/10.1016/j.ympev.2013.01.010

Oliver, E.G.H., & Forshaw, N. 2012. Genus *Erica* An Identification Aid Version 3.00. *Contributions from the Bolus Herbarium* 22.

Oliver, E.G.H., & Oliver, I.M. 2003. Ericaceae. Pp. 424-451 in: Germishuizen, G., & Meyer, N.L., (eds), *Plants of southern Africa: an annotated checklist. Strelitzia 19*. South African National Biodiversity Institute, Pretoria. p 424-451.

Page, R.D.M., & Charleston, M.A. 1997. From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. *Mol Phylog Evol* 7:231-240. http://dx.doi.org/10.1006/mpev.1996.0390

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877-884. http://dx.doi.org/10.1038/44766

Pamilo, P., & Nei, M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol* 5:568-583.

Pelser, P.B., Kennedy, A.H., Tepe, E.J., Shidler, J.B., Nordenstam, B., Kadereit, J.W., & Watson, L.E. 2010. Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. *Am. J. Bot.* 97:856-873. http://dx.doi.org/10.3732/ajb.0900287

Peterson, P.M., Romaschenko, K., & Arrieta, Y.H. 2015. Phylogeny and subgeneric classification of Bouteloua with a new species, B. herrera-arrietae (Poaceae: Chloridoideae: Cynodonteae: Boutelouinae). *Journal of Systematics and Evolution* 53:351-366. http://dx.doi.org/10.1111/jse.12159

Pimentel, M., Sahuquillo, E., Torrecilla, Z., Popp, M., Catalán, P., & Brochmann, C. 2013. Hybridization and long-distance colonization at different time scales: towards resolution of long-term controversies in the sweet vernal grasses (*Anthoxanthum*). *Ann Bot* 112:1015-1030. http://dx.doi.org/10.1093/aob/mct170

Pinheiro, F., & Cozzolino, S. 2013. Epidendrum (Orchidaceae) as a model system for ecological and evolutionary studies in the Neotropics. *Taxon* 62:77-88.

Pirie, M.D. 2012. What can heathers tell us about the origins of biological diversity? *Heathers* 9:15-23.

Pirie, M.D. 2015. Phylogenies from concatenated data: Is the end nigh? *Taxon* 64:421-423. http://dx.doi.org/10.12705/643.1

Pirie, M.D., & Doyle, J.A. 2012. Dating clades with fossils and molecules: the case of Annonaceae. *Botan J Linn Soc* 169:84-116. http://dx.doi.org/10.1111/j.1095-8339.2012.01234.x

Pirie, M.D., Humphreys, A.M., Antonelli, A., Galley, C., & Linder, H.P. 2012. Model uncertainty in ancestral area reconstruction: a parsimonious solution? *Taxon* 61:652-664.

Pirie, M.D., Humphreys, A.M., Barker, N.P., & Linder, H.P. 2009. Reticulation, data combination, and inferring evolutionary history: an example from Danthonioideae (Poaceae). *Sys. Biol.* 58:612-628. http://dx.doi.org/10.1093/sysbio/syp068

Pirie, M.D., Humphreys, A.M., Galley, C., Barker, N.P., Verboom, G.A., Orlovich, D., Draffin, S.J., Lloyd, K., Baeza, C.M., Negritto, M., Ruiz, E., Cota Sanchez, J.H., Reimer, E., & Linder, H.P. 2008. A novel supermatrix approach improves resolution of phylogenetic relationships in a comprehensive sample of danthonioid grasses. *Mol Phylog Evol* 48:1106-1119. http://dx.doi.org/10.1016/j.ympev.2008.05.030

Pirie, M.D., Lloyd, K.M., Lee, W.G., & Linder, H.P. 2010. Diversification of *Chionochloa* (Poaceae) and biogeographic history of the New Zealand Southern Alps. *Journal of Biogeography* 37:379–392. http://dx.doi.org/10.1111/j.1365-2699.2009.02205.x

Pirie, M.D., Oliver, E.G.H., & Bellstedt, D.U. 2011. A densely sampled ITS phylogeny of the Cape flagship genus *Erica* L. suggests numerous shifts in floral macro-morphology. *Mol Phylog Evol* 61:593-601. http://dx.doi.org/10.1016/j.ympev.2011.06.007

Pirie, M.D., Vargas, M.P.B., Botermans, M., Bakker, F.T., & Chatrou, L.W. 2007. Ancient paralogy in the cpDNA *trnL-F* region in Annonaceae: implications for plant molecular systematics. *Am. J. Bot.* 94:1003-1016. http://dx.doi.org/10.3732/ajb.94.6.1003

Planet, P.J. 2006. Tree disagreement: Measuring and testing incongruence in phylogenies. *Journal of Biomedical Informatics* 39:86-102. http://dx.doi.org/10.1016/j.jbi.2005.08.008

Posada, D., & Crandall, A.K. 2002. The Effect of Recombination on the Accuracy of Phylogeny Estimation. *J Mol Evol* 54:396-402. http://dx.doi.org/10.1007/s00239-001-0034-9

Posada, D., & Crandall, K.A. 2001. Intraspecific gene genealogies: trees grafting into networks. *TREE* 16:37-45.

Ragan, M.A. 1992. Phylogenetic inference based on matrix representation of trees. *Mol Phylog Evol* 1:53-58. http://dx.doi.org/10.1016/1055-7903(92)90035-F

Reid, N.M., Hird, S.M., Brown, J.M., Pelletier, T.A., McVay, J.D., Satler, J.D., & Carstens, B.C. 2014. Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data. *Sys. Biol.* 63:322-333. http://dx.doi.org/10.1093/sysbio/syt057

Rieseberg, L.H. 1995. The Role of Hybridization in Evolution: Old Wine in New Skins. *Am. J. Bot.* 82:944-953. http://dx.doi.org/10.2307/2445981

Rieseberg, L.H. 2009. Evolution: Replacing Genes and Traits through Hybridization. *Current Biology* 19:R119-R122. http://dx.doi.org/10.1016/j.cub.2008.12.016

Rieseberg, L.H., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J.L., Schwarzbach, A.E., Donovan, L.A., & Lexer, C. 2003. Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science* 301:1211-1216. http://dx.doi.org/10.1126/science.1086949

Rosenfeld, J.A., Payne, A., & DeSalle, R. 2012. Random roots and lineage sorting. *Mol Phylogenet Evol* 64:12-20. http://dx.doi.org/10.1016/j.ympev.2012.02.029

Sagan, L. 1967. On the origin of mitosing cells. *Journal of Theoretical Biology* 14:225-IN226. http://dx.doi.org/10.1016/0022-5193(67)90079-3

Saitou, N., & Yamamoto, F. 1997. Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol* 14:399-411.

Salichos, L., & Rokas, A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327-331. http://dx.doi.org/10.1038/nature12130

Salichos, L., Stamatakis, A., & Rokas, A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* 31:1261-1271. http://dx.doi.org/10.1093/molbev/msu061

Sanderson, M.J., Purvis, A., & Henze, C. 1998. Phylogenetic supertrees: assembling the trees of life. *TREE* 13:105-109. http://dx.doi.org/10.1016/S0169-5347(97)01242-1

Sanderson, M.J., & Shaffer, H.B. 2002. Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics* 33:49-72. http://dx.doi.org/10.1146/annurev.ecolsys.33.010802.150509

Schaack, S., Gilbert, C., & Feschotte, C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution* 25:537-546. http://dx.doi.org/10.1016/j.tree.2010.06.001

Schönknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., BrÃ¤utigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., Carr, K., Wilkerson, C., Rensing, S.A., Gagneul, D., Dickenson, N.E., Oesterhelt, C., Lercher, M.J., & Weber, A.P.M. 2013. Gene Transfer from Bacteria and Archaea Facilitated Evolution of an Extremophilic Eukaryote. *Science* 339:1207-1210. http://dx.doi.org/10.1126/science.1231707

Simmons, M.P., & Ochoterena, H. 2000. Gaps as characters in sequence-based phylogenetic analysis. *Sys. Biol.* 49:369-381.

Swofford, D.L. 1991. When are phylogeny estimates from molecular and morphological data incongruent. Pp. 295-333 in: Miyamoto, M.M., & Cracraft, J., (eds), *Phylogenetic analysis of DNA sequences*. Oxford University Press. p 295-333.

Swofford, D.L. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), version 4. In. Sinauer Associates, Sunderland, Mass.

Szöllősi, G.J., Tannier, E., Daubin, V., & Boussau, B. 2015. The Inference of Gene Trees with Species Trees. *Sys. Biol.* 64:e42-e62. http://dx.doi.org/10.1093/sysbio/syu048

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957-966.

Than, C., & Nakhleh, L. 2009. Species Tree Inference by Minimizing Deep Coalescences. *PLoS Comput Biol* 5:e1000501. http://dx.doi.org/10.1371/journal.pcbi.1000501

Than, C., Ruths, D., & Nakhleh, L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322. http://dx.doi.org/10.1186/1471-2105-9-322

Van der Niet, T., Pirie, M.D., Shuttleworth, A., Johnson, S.D., & Midgley, J.J. 2014. Do pollinator distributions underlie the evolution of pollination ecotypes in the Cape shrub *Erica plukenetii*? *Ann Bot* 113:301-316. http://dx.doi.org/10.1093/aob/mct193

Visser, J.C., Bellstedt, D.U., & Pirie, M.D. 2012. The recent recombinant evolution of a major crop pathogen, *Potato Virus Y*. *PLoS ONE* 7:e50631. http://dx.doi.org/10.1371/journal.pone.0050631

Vriesendorp, B., & Bakker, F.T. 2005. Reconstructing patterns of reticulate evolution in angiosperms: what can we do? *Taxon* 54:593-604.

Wendel, J.F., & Doyle, J.J. 1998. Phylogenetic Incongruence: Window into Genome History and Molecular Evolution. Pp. 265-296 in: Soltis, D.E., Soltis, P.S., & Doyle, J.J., (eds), *Molecular Systematics of Plants II: DNA Sequencing*. Springer US, Boston, MA. p 265-296.

Whitney, K.D., Ahern, J.R., Campbell, L.G., Albert, L.P., & King, M.S. 2010. Patterns of hybridization in plants. *Perspectives in Plant Ecology, Evolution and Systematics* 12:175-182. http://dx.doi.org/10.1016/j.ppees.2010.02.002

Wiens, J.J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Sys. Biol.* 52:528-538.

Wiens, J.J., Fetzner, J.W., Parkinson, C.L., & Reeder, T.W. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. *Sys. Biol.* 54:778-807.

Woolley, S.M., Posada, D., & Crandall, K.A. 2008. A Comparison of Phylogenetic Network Methods Using Computer Simulation. *PLoS ONE* 3. http://dx.doi.org/10.1371/journal.pone.0001913

Yang, Y., & Smith, S.A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Mol Biol Evol* 31:3081-3092. http://dx.doi.org/10.1093/molbev/msu245

Yu, Y., Barnett, R.M., & Nakhleh, L. 2013. Parsimonious Inference of Hybridization in the Presence of Incomplete Lineage Sorting. *Sys. Biol.* 62:738-751. http://dx.doi.org/10.1093/sysbio/syt037

Yu, Y., Degnan, J.H., & Nakhleh, L. 2012. The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection. *PLoS Genetics* 8:e1002660. http://dx.doi.org/10.1371/journal.pgen.1002660

Yu, Y., Dong, J., Liu, K.J., & Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences* 111:16448-16453. http://dx.doi.org/10.1073/pnas.1407950111

Yu, Y., Than, C., Degnan, J.H., & Nakhleh, L. 2011. Coalescent Histories on Phylogenetic Networks and Detection of Hybridization Despite Incomplete Lineage Sorting. *Sys. Biol.* 60:138-149. http://dx.doi.org/10.1093/sysbio/syq084