African Journal of Biotechnology Vol. 6 (6), pp. 639-649, 19 March 2007
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 © 2007 Academic Journals

Review

Phylogenetics in plant biotechnology: principles, obstacles and opportunities for the resource poor

Joel W. Ochieng[1*], Anne W. T. Muigai[2] and George N. Ude[3]

[1]Sections of Genetics, College of Agriculture and Veterinary Sciences, University of Nairobi, P.O. Box 29053 Nairobi, 00625 Kenya.
[2]Institute of Biotechnology Research, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000 Nairobi, 00200 Kenya.
[3]Department of Natural Sciences, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA.

Accepted 5 February, 2007

Phylogenetic inference has become routine for most studies of genetic variation among plant taxa. However, inferring phylogenies can be confounded by both biological and computational or statistical complexities, resulting in misleading evolutionary hypotheses. This is particularly critical because the "true tree" can only truly be known in exceptional circumstances. Moreover, selecting appropriate marker(s), characters, sample sizes and the appropriate reconstruction methods offers a challenge to most evolutionary geneticists. Textbooks are generic (and sometimes outdated), and in resource poor labs, they may altogether be inaccessible. In this review, we take the worker through the low-down on reconstructing a phylogeny, review the enigmatic biological and computational problems, and examine cases where cheaper markers and extremely small sample sizes can recover a reliable phylogeny.

Key words: Phylogeny, tree incongruence, homoplasy, lineage sorting, molecular markers.

INTRODUCTION

In nearly all cases of biological evolution, it is impossible to witness speciation events. As such, a variety of methods have been designed to reconstruct these events, with the usual model being phylogenetic trees. Because it reflects the history of transmission of life's genetic information, phylogeny is a central interpretative framework for studying evolutionary processes, organization and interpretation of information on all characteristics of organisms, from structure and physiology to genomics. A reconstructed phylogeny helps guide our interpretation of the evolution of organismal characteristics, providing hypotheses about the lineages in which traits arose and under what circumstances, thus playing a vital role in studies of adaptation and evolutionary constraints (Felsenstein, 1985; Martins, 1995). Patterns of divergence of species lineages indicated by the tree inform the dynamics of speciation (and extinction), the forces that generate and reduce biodiversity (Futuyma, 1998). The evolutionary histories of genes bear the marks of the functional demands to which they have been subjected, so that phylogenetic analyses can elucidate functional relationships within living cells (Gu, 2001; Zhu et al., 2000). There is, thus, increasing use of phylogenetic analysis to make functional predictions from sequence databases of gene families (Bader et al., 2001), to predict ligands (Chambers et al., 2000), and to help in the development of vaccines and antimicrobials and herbicides (Brown and Warren, 1998).

Because phylogenies are such an important part of biological investigations, many methods exist for reconstructing them. For the most part, these methods assume that the phylogeny underlying the data is a tree. In strict biological sense, such is not always the case: for many organisms, a significant level of genetic exchange occurs between lineages, and for some groups, lineages can combine to produce new independent lineages. Such cases include meiotic and sexual recombination, horizontal gene transfer and hybrid speciation, which cannot be modelled by bifurcating trees. This review focuses on bio-

*Corresponding author. E-mail: j.ochieng.10@scu.edu.au, Tel: +61 2 6620 3961, Fax: +61 2 6622 2080.

logical and statistical problems encountered, cheaper ways of data production relevant to the resource poor, and a stepwise guide to reconstruction of phylogenies by the methods available, rather than on the technical flaws that characterise these methods.

PHYLOGENETIC TREES

Phylogenetic trees are an illustration of evolutionary relationships among a group of organisms or between collections of "things" (such as genes, proteins, and organs), which are derived from a common ancestor. A tree is composed of nodes (which represent the relationships among taxonomic units) and branches (which represent the taxa). Each node represents speciation events in evolution; terminal nodes represent the data under comparison (operational taxonomic units), while internal nodes are the inferred ancestral units (hypothetical taxonomic units). The length of the branches (branch lengths), represent the number of changes that occurred in the characters prior to the next level of separation. Hence, very similar characters (sequences, alleles, body parts) will be neighbours on the outer branches. The topology of the tree is the branching pattern.

Kinds of phylogenetic trees

There are different kinds of phylogenetic trees:

(1) A cladogram.
(2) Phenogram.
(3) Phylogram.
(4) Dendrogram.

The difference between these trees is not trivial; there are a number of published papers in which trees have been given the wrong name. A cladogram shows the relative recency of common ancestry; all the objects on it share a known common ancestor (the root taxon that branched earlier of all the other taxa but is related to them) (Figure 1A). On a cladogram, the paths from the root to the nodes correspond to evolutionary time (Figure 2A). When a cladogram has branch lengths, it is termed a phylogram (also called metric or additive trees). A dendrogram is a special kind of cladogram in which the tips of the trees are equidistant from the root Figure 2B. Phenograms are phylogenetic trees where all the objects on it are related descendants, but there is not enough information to specify the common ancestor (root). A phenogram is therefore an unrooted tree. The paths between nodes do not specify an evolutionary time. The number of tree topologies of a rooted tree is much higher than that of an unrooted tree for the same number of OTUs. Therefore the error for the latter topology is smaller than that of the rooted tree.

BOOTSTRAP AND JACKNIFE VALUES

Bootstrapping is a validation procedure (Felsenstein, 1985) in which the character columns are resample from the data matrix (with replacement) to produce bootstrap pseudoreplicates (each pseudoreplicate contains each of the original taxa, but some original characters are presented more than once and some not at all). Each pseudoreplicate is then analysed phylogenetically, with a consensus tree constructed to summarize the results of all replicates. The proportion of trees/replicates in which a grouping is recovered is presented as a measure of support for that group. However, bootstrap confidence levels apply to nodes- they are not joint confidence statements; the joint confidence drops as additional nodes are considered. Felsenstein (1985) explicitly stated that bootstrapping provides a confidence interval that contains "the phylogeny that would be estimated from repeated sampling of many characters from the underlying set of all characters", NOT the true phylogeny. Thus Felsenstein viewed bootstrap values as a measure of "repeatability" rather than measures of "accuracy". Unlike bootstrapping, jacknife resampling is a method in which either characters or taxa are resampled without replacement (Efron, 1979).

TREE RECONSTRUCTION METHODS

There are three basic types of phylogenetic reconstruction methods in common use: distance-based methods, maximum parsimony, and maximum likelihood heuristics.

Distance methods

Distance-based methods operate by first estimating pair wise distances and then computing an edge-weighted tree using those distances. These methods are guaranteed to reconstruct the true tree if their estimates of pair wise distances are sufficiently close to the number of evolutionary events between pairs of taxa (Kim and Warnow, 1999). For many models of biomolecular sequence evolution, estimation of sufficiently accurate pair wise distances is possible (Li, 1997).

Maximum parsimony

Maximum parsimony heuristics is typical in the analysis of DNA sequences, in which the objective is the tree with the minimum number of nucleotide substitutions across the tree (Felsenstein, 2003 for a review on heuristics). These heuristics operate by hill climbing through an exponentially sized tree space. However, there are reservations to the use of Maximum parsimony as not being statistically consistent (Felsenstein, 1978); neverth-

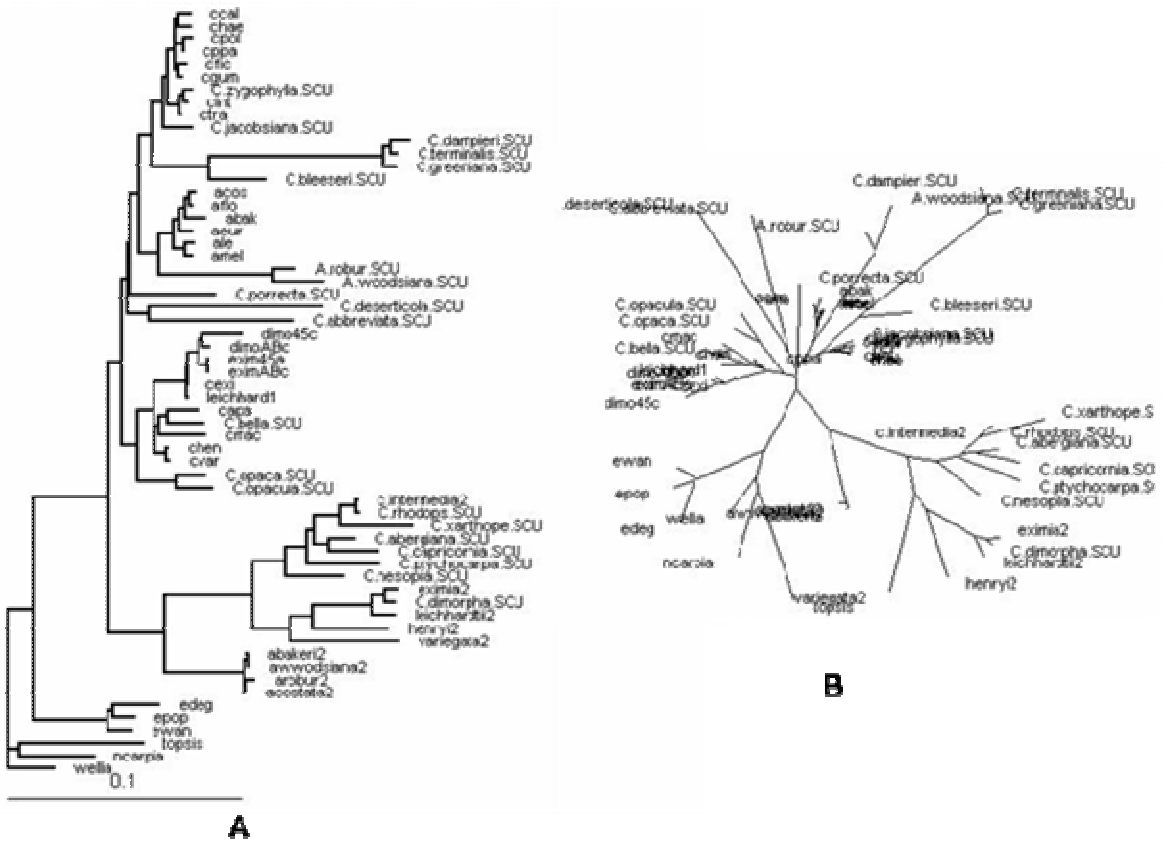

Figure 1. A. Rectangular Cladogram; B. Phenogram showing the systematic relationships among eucalypt genera *Corymbia*, *Angophora* and *Eucalyptus* (Ochieng et al., unpublished).

eless, it is a very popular approach in systematics because it is more computationally efficient with large numbers of taxa than maximum likelihood.

Maximum likelihood

Maximum Likelihood (ML) seeks the tree T and its associated parameters (such as edge-lengths, rates of evolution for each site, etc.) that maximize the probability of generating the given set of sequences. The general idea behind maximum likelihood is the estimation of a model by finding the model that maximizes the conditional probability of data/model, the likelihood of the data.

MARKERS FOR PHYLOGENETIC INFERENCE

One of the greatest impediments to genetic research in developing countries is the high cost associated with molecular analysis. These include costs associated with direct consumables as well as a lack of appropriate equipment to carry out such analysis. In contrast, institutions in developed economies have facilities for 'state of the art' methodologies such as microarray and real-time PCR analysis of gene expression, high throughput DNA sequencing, and nanotechnology for DNA quantitation. Ironically, developing countries are a reliable source of skilled manpower for universities and research centres in developed countries.

Due to inadequate research funds and the lack of equipment, developing country researchers still utilize old tools for molecular analyses such as allozymes, RAPD, RFLP and others, that were revolutionary around the third quarter the last century, but which are now mundane. To emphasize this passage of time (and technology), some journals (examples are Molecular Ecology and Heredity) no longer accept manuscripts that primarily report RAPD-based analyses. It is argued that such data have low repeatability. Although admittedly, the non-repeatability of data such as RAPD might be suggested on weak grounds; case studies have shown that even DNA sequencing is not the most corroborated and often leads to intraindividual sequence variation (Dunning et al., 1988; Ling et al., 1991). Microsatellites, commonly referred to as SSRs, are another useful marker whose utility in higher-level phylogenetic inference has been questioned.

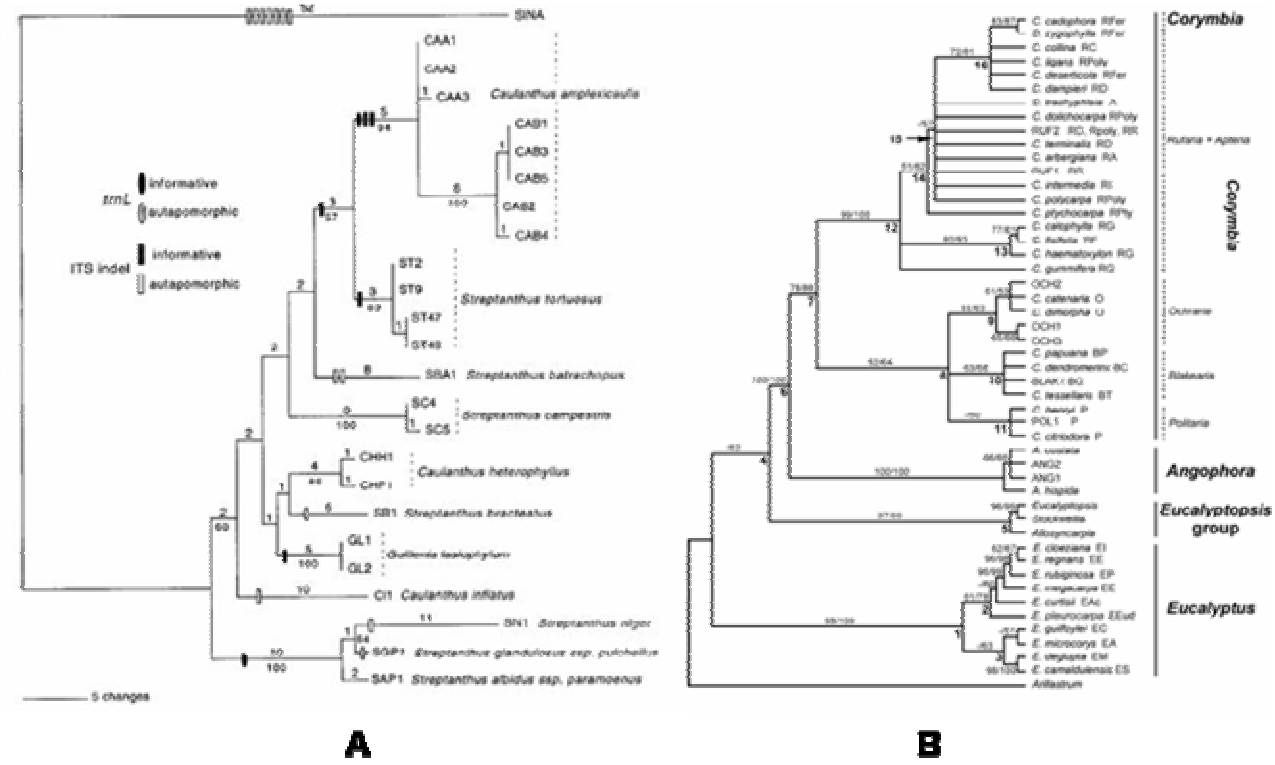

Figure 2. A- Phylogram; a cladogram with branch lengths. (Pepper and Norwood, 2001); B- Dendrogram, a cladogram with branch lengths and the tips of the trees are equidistant from the root. (Paraa et al., in press).

However, some of these 'condemned' markers are cheaper, can be genotyped without elaborate equipment, and in phylogenetic inference, might have superior signals to gene sequences.

Supporting microsatellite utility

Microsatellites, also referred to as simple sequence repeats (SSRs), are segments of DNA with tandem repeat of short sequence motifs, each generally less than 5 bp in length (Bruford and Wayne, 1993). SSRs have many advantages over DNA sequencing, including a greater representation of different genomic regions in a single dataset. Their faster evolution may lead to more informative characters. However, the utility of SSRs in reconstructing phylogenetic relationships, especially among divergent taxa, is a matter of current debate. Apart from the technical difficulty in amplifying SSRs across taxa, they are believed to possess three interrelated attributes that may limit their use in reconstructing phylogenies of divergent taxa:

1) A constraint on allele size range (Goldstein and Pollock, 1997)
2) High mutation rates.
3) Size homoplasy (Bruford and Wyne, 1993).

Another limitation is that even when it is possible to amplify SSRs in divergent taxa, the sequences may not be similar enough to permit confident assessment of orthology. These reasons partly explain why many phylogenetic studies utilizing microsatellites have been restricted to intra-specific relationships (Goldstein et al., 1999), or to the use of the SSR flanking sequence in higher order phylogenies (Streelman et al., 1998; Zhu et al., 2000).

Some notable cases exist for the use of repeat sequence variations in highly divergent taxa:

1) Richard and Thorpe (2001) used SSR size variation to analyse the phylogenetic relationships among the western canary island lizards, a group that diverged five million years ago (MYA). This divergence time corresponds to five million generations given their short generation time of one year (Richard and Thorpe, 2001).
2) Ritz et al. (2000) applied repeat size variation at SSR loci to resolve the relationships among four genera (Bos, Bison, Bubalus and Syncerus) in the tribe Bovini. To overcome issues of homoplasy, the authors used the average square $(\delta\mu)^2$ genetic distance measure (Goldstein et al., 1995). They found this measure to be robust despite fluctuations in population size, and retained linearity with increasing time. The tree topology was retained when data were reanalysed with Cavalli-Sforza and Edw-

ards' (1967) chord distance (D_C) that is, interestingly, based on the infinite allele model.

3) SSR length variation has been used in reconstructing the phylogeny of Darwin's finches (Petren et al., 1999). Although considered to be congeneric, these birds are believed to have radiated at least three MYA (Petren et al., 1999). With their short generation time of four months to one year (Zink, 2002), they have evolved for over five million generations.

Although these examples are mainly from animals, the rarity of SSR use in phylogenies of plant taxa may be due mainly to low levels of transferability (Peakall et al., 1998) and a low level of SSR conservation among many plant taxa (Whitton et al., 1997), rather than concerns relating to high mutation rates or other evolutionary considerations. This argument was recently put to test when, in a pioneer study using SSR to resolve an enigmatic plant phylogeny, Ochieng et al. (in review), using eight polymorphic SSRs, resolved a well-corroborated phylogeny for a plant group that diverged over 70 Million years ago. The authors addressed the factors that limit the use of SSRs in higher order phylogenetic inferences. Thus it would appear that when the problems of range constraints, high mutation rates and size homoplasy are addressed, SSRs may be utilised in phylogenetic studies, even among very divergent taxa, so long as SSR primers amplify across such taxa.

When SSRs can be used in higher order phylogeny

It is a widely held view that SSRs may not be useful in phylogenetic studies above the species level (Streelman et al., 1998; Zhu et al., 2000). Realistically, the problems that limit the use of SSRs in higher order phylogenetic relationships are relevant only when 'above threshold' number of generations have passed since divergence from a common ancestor. Taxonomic rank, which creates the artificial constructs 'higher order' or 'lower order' phylogeny, is irrelevant when evaluating the potential for SSRs use for a group of taxa. We believe that SSRs can reconstruct deep phylogenies under the following scenarios.

Appropriate genetic distance measures: Homoplasy is expected under the stepwise mutation model (SMM; Kimura and Ohta, 1978), which assumes loss or gain, with equal probability, of a single repeat unit through mutation. However, the infinite allele model (IAM; Kimura and Crow, 1964) expects no homoplasy because a mutation is assumed to result in an allelic state not previously encountered in the population. Several genetic distances that make different assumptions have been developed for use with microsatellite data, however, the appropriateness of each of these distance methods will vary from case to case, depending on the model of microsatellite evolution, mutation rates, effective population size, and time since divergence. The ideal distance measure will therefore depend on the charact-eristics of the SSRs and on the phylogenetic question being addressed. If it is not clear under what model the SSRs used in a study evolves, we advocate the use of two genetic distance measures: the SMM model based average square distance $(\delta\mu)^2$; analogous to *D1* of Goldstein et al. (1995), and Nei's (1972) IAM based standard genetic distance (G_{st}). The average square distance (Goldstein et al., 1995) addresses size range constraints, thereby accounting for homoplasy. The distance retains linearity with increasing evolutionary distance, and hence is suitable for reconstructing trees that include more distantly related taxa (Goldstein et al., 1995; Pollock et al., 1998; Petren et al., 1999; Ritz et al., 2000; Richard and Thorpe, 2001). This distance has been successfully used in recovering well-corroborated phylogenetic hypotheses in a number of studies involving divergent taxa (Petren et al., 1999; Ritz et al., 2000; Richard and Thorpe, 2001; Ochieng et al., in review). On the other hand, Nei's (1972) distance is expected to become more linear while the linearity of average square distance wanes as the SSR mutations become more like the IAM model (Goldstein et al., 1995). In the Bovini study (Ritz et al., 2000) mentioned earlier, the authors used the genetic distance measure, $(\delta\mu)^2$ (Goldstein et al., 1995) to account for size homoplasy. They found the measure to be robust despite fluctuations in population size, and retained linearity with increasing time. In real data, however, both distances often recover a similar tree topology (Ritz et al., 2000). This is expected to be the case when the data comprise a minimum proportion of homoplasious alleles, as well as when SSRs evolve at a lower rate and are highly conserved, both in the repeats and in the flanking regions.

Range constraint and size homoplasy: Homoplasy may arise due to:

I. Mutations in microsatellite repeat region or flanking region that result in alleles being similar in state but not by descent.
II. A constraint to the upper (and sometimes lower) bound on the number of repeat units at a locus may exacerbate homoplasy in the repeat region, as these size limits allow only a finite number of character states.
III. Insertion and deletions in the flanking region making alleles similar in state but not by descent.

At longer time intervals, homoplasy is expected to increase, while phylogenetic signals move to obscurity as saturation is approached (Takazeki and Nei, 1996). Sometimes, the data and results do not support a likelyhood of phylogenetic signal saturation reasons, such as when:

1) The average of the means of allele size range for each species or clade considered separately across all loci is significantly less than the mean for all species combined (Ochieng et al., in review). This would suggest that saturation of phylogenetic signal through homoplasy due to range constraint is minimal because the allele size range of species or subgroups does not reach the total observed allele size range.
2) Sizes of most alleles in the dataset differ by a number divisible by their repeat unit length, implying a low likelihood of homoplasy due to mutations in the regions flanking the repeats. Insertions and deletions should be equally likely to involve odd and even numbers of bases (Ochieng et al., in review)
3) Theoretically, variation in the amount of size homoplasy is expected among SSR loci because variation in mutation rates reflects the stochasticity among loci of the coalescence process (Garza and Freimar, 1996). However, when the bootstrap support for tree topology recovered is high, this would reflect concordance among loci. Bootstrapping characters from loci with varied levels of homoplasy is expected to recover discordant phylog-enetic hypotheses, usually signified by low bootstrap values on the consensus tree.

'Below threshold' number of generations: The properties that limit SSR use in phylogenetics (mutation rates, size constraint and homoplasy) relate to the number of generations since the divergence of taxa, rather than to their classification. If SSRs correctly resolved phylogenies of lizards that have diverged for five million generations (Richard and Thorpe, 2001), then they may recover the correct phylogeny for plant genera that have diverged for up to five million generations, assuming the mutation rates are comparable. Notably, Richard and Thorpe, (2001) analysed only five SSR loci, and the results corroborated the true and confirmed organismal phylogeny. Apart from the average square distance of Goldstein et al. (1995), the authors utilized other distances such as Nei's (1972) G_{st} and allele sharing statistic (P_{SA}) for comparison. Their data contradicted the expectation that the SSR genetic distances may lose linearity after several thousands of generations, essentially due to range constraints in allele sizes (Feldman et al., 1997). As the authors noted, the fact that the essentials of a well-corroborated tree can be reconstructed from such a relatively small number of loci argue for their utility in this area. As stated earlier, SSR length variation has also been used in reconstructing the phylogeny of Darwin's finches, which are believed to have radiated at least three MYA, corresponding to over five million generations (Petren et al., 1999). Thus we argue that with limited financial resources, SSRs would be a suitable and cheaper marker compared to DNA sequencing for phylogenetic reconstruction.

How many samples can infer a phylogeny

In phylogenetic analyses using sequence data such as nuclear ribosomal ITS, chloroplast genes in plants, mitochondrial genes such as Cytochrome *b*, and the D-Loop regions in animals, a single individual is often used to represent a clade. However, when using the more polymorphic markers such as SSRs and allozymes, there is a tendency to use hundreds of samples per clade. Thus there has been considerable discussion regarding the optimal sample size in population genetic analyses, with some workers recommending large sample sizes to account for sampling variance (Nei, 1978; Ruzzante, 1998). In phylogenetic studies, however, pairwise genetic distance between individuals (or groups of individuals) rather than allele frequencies are relevant. Kalinowski (2005) recently simulated the relationship between sample size, polymorphism, and the coefficient of variation of genetic distances derived from microsatellite markers. He found that when the differentiation among the taxonomic units to be measured is large, one or two samples per group would give similar results to a large sample size.

Increasing sample size under a large F_{ST} scenario produced diminishing effect on the coefficient of variation of the genetic distance. Kalinowski's simulated data showed that the rate at which increasing sample size decreased the coefficient of variation was determined principally by the amount of differentiation between populations. This means that more individuals are necessary only when the degree of differentiation is low. In the case of interspecific and intergeneric, or the more amorphous 'higher order' phylogenies, the differentiation in question is among species rather than just between populations of the same species. Hence the between species and between genera F_{ST} values are expected to be large since the two genera have diverged for a longer evolutionary timescale. This view has recently been qualified using real data. To resolve a notoriously difficult phylogenetic question, Ochieng et al. (in review) recently used a single individual to represent a species in among-genera eucalypt phylogeny, recovering well corroborated evolutionary hypotheses.

Apart from SSRs, proteins have been used in phylogenetic reconstruction. Demastes and Remsen (1994) analysed allozyme variation to reconstruct the phylogeny of eight bird genera in the family Cardinalinae, using a single individual to represent each genus in the family. Their tree topologies supported phylogenetic analyses of morphological characters. As the authors noted, in a phylogenetic context the priority switches from more samples to more phylogenetic characters (Demastes and Remsen, 1994, and references therein). We are aware that allozymes are less polymorphic compared to SSRs, however, Kalinowski's (2005) simulation addresses this difference in variability and its implications. Nonetheless, to be sure the small sample sizes are taken fully into account, a surrogate analysis should be conducted for

samples pooled into the main groupings. The finding that such low sample numbers can infer a 'correct' phylogeny is welcome for research groups with limited funds available for field collections and molecular analyses.

PHYLOGENETIC TREE INCONGRUENCE

Phylogenies inferred from independent data partitions may differ from one another in topology despite the fact that they are drawn from the same set of organisms. We do not of know cases in which phylogenetic incongruence has been attributed to non-repeatability of data, even in cases where AFLP or RAPD markers were used. Cases of low statistical support or contradictory results in molecular studies are commonly attributed to insufficient information from short sequences (Cummings et al., 1995; Renner and Chanderbali, 2000), poor taxon samp-ling (Steane et al., 1999, but see Steane et al., 2002), strong rate variation across nucleotide sites or taxa (Kuhner and Felsenstein, 1994; Takezaki and Gojobori, 1999), homoplasy and hybridization (McCracken and Sorenson, 2005), the use of inappropriate phylogenetic models (Cunningham, 1997), paralogy (Ochieng et al., in review) or a combination of these factors. Hence incongruence may arise from statistical or biological causes. Common statistical (or computational) causes, such as inadequate or non-judicious sampling, inappropriate tree reconstructtion methods, and inadequate phylogenetic characters can be addressed by expanded or judicious sampling, addition of phylogenetic characters or by modifying analysis and tree construction models (Udovicic et al., 1995; Steane et al., 1999, 2002; Udovicic and Ladiges, 2000). However, incongruence that has its origin in genealogical discordance is not easily resolved. Notorious biological causes of tree incongruence include hybridization (Dumolin-Lapegue et al., 1997; McKinnon et al., 1999; Avise, 2000), homoplasy (McCracken and Sorenson, 2005), paralogy (Ochieng et al., in review) and lineage sorting (Avise et al., 1990; Maddison, 1997; Avise, 2000; Lu, 2001; Takahashi et al., 2001).

Paralogy and lineage sorting

Sequences that have evolved from a single most recent common ancestor (MRCA) at the root of a clade are said to be orthologous. In contrast, DNA sequences that are the result of gene duplication events are paralogs. Gene trees that are reconstructed from orthologs will be identical to the species trees, while those reconstructed from paralogs may be incongruent because the tree topology reflects both gene duplication as well as speciation events. An example of genes known to exist in multiple copies in plant genomes is the nuclear ribosomal internally transcribed spacer (nrITS), implicated in RNA maturation (Baldwin et al., 1995). The nrITS is generally located in one to several arrays, with generally low intragenomic rDNA diversity due to concerted evolution within and between ribosomal loci (Dover, 1982; Arnheim et al., 1983; Baldwin et al., 1995). However, there are a number of scenarios in which evolution may not be concerted for some paralogous sets:

1) When concerted evolution is slower than speciation, then a single genome will contain divergent paralogues.
2) When paralogous rDNA are present in non homologous chromosomes; this may also preclude complete sequence homogenization between the paralogous sets (Ochieng et al., in review).

Such paralogs would be divergent, and may also form separate clades in a phylogenetic framework. Paralogs may be the result of hybridization, polyploidization, or stochastic gene duplications (whole genome, large scale chromosomal, or single gene duplications). However, the evolutionary fate of duplicated genes is largely unknown. Duplicated genes could either evolve novel functions, or become pseudogenes (i.e., non-functional paralog) through loss-of-function mutations. Several lines of evidence can show that a paralog is a pseudogene: comparatively lower GC content suggesting lower structural stability, deamination-like mutations at potential methylation sites, lack of conserved helices and hairpins, and conspicuously lower thermodynamic stability in secondary structures.

A number of population genetic processes can cause orthologs to be randomly or systematically lost in some species: Genetic drift and population bottlenecks (random), and natural selection (systematic). Thus, when a species lacks a particular ortholog, it is possible to use a paralog and be unable to detect the mistake. This can cause paraphyly or polyphyly, and erroneous evolutionary interpretations if orthology of alleles were assumed. However, recognition of divergent paralogues and pseudogenes will provide workers with more out group opportunities. Indeed, the first report of pseudo-genes recovering a more corroborated phylogeny than ITS functional paralog is currently under review (Ochieng et al., in review). The freedom from functional constraint would appear to make pseudogenes more suited for phylogenetic work than the functional genes.

The random loss of orthologs is termed "lineage sorting" (Nei, 1987). During speciation or successive rounds of speciation, ancestral polymorphisms may be stochastically (by chance) sorted in descendants. This incomeplete lineage sorting has caused phylogenetic difficulties in a number of studies (Pamilo and Nei, 1988; Takahata, 1989; Lu, 2001; Takahashi et al., 2001; McCracken and Sorenson, 2005).

Homoplasy

Homoplasy is the similarity in characters but no common ancestry; similarity due to convergent, parallel or 'loss' evolution. In phylogenetic reconstruction, of interest are

characters that are shared due to common ancestry (homologs). Homoplasy would tend to increase with time since speciation, and with increasing rates of mutation. Though homoplastic characters can obscure synapomorphies (shared derived characters), it is easily detected in cases of very recent speciation, and where retained ancestral polymorphism is still extant.

Tests for homoplasy employs different measures of character quality, including successive approximations (Farris, 1969), compatibility analysis (Meacham and Estabrook, 1985), and the optimization method of Goloboff (1993). Goloboff's criterion searches for trees that imply the highest weights for all characters. Character quality is assessed by a function that relates the fit of a character on a tree to its homoplasy: $k+1/(s+k+1-m)$, where k is a constant describing the concavity of the fit/homoplasy relationship, s is the number of steps required for a character to fit a particular tree, and m is the minimum possible number of steps for that character on any tree. A posteriori weights is assigned to each character based on its maximum consistency index (CI- Kluge and Farris 1969) and homoplasy index over all trees and are used in subsequent parsimony searches. Heuristic searches can be implemented in PAUP using random taxon addition and TBR branch swapping option. This procedure is repeated until character weights (and tree topologies) are stable for two consecutive iterations. Another method developed by Ochieng J.W. (unpublished) to test for homoplasy involves sequential character exclusion to create polytomies, followed by reinstatement in a stepwise manner and observing the changes in tree topology. This test tries to determine if there is a statistically significant excess of homoplasies compared to that expected by mutation in the absence of recombination. This test relies on the observation that, as the time depth of a soft polytomy increases, homoplasy will tend to obscure relationships and increase the number of characters needed to resolve a given gene tree, regardless of whether that gene tree matches the species tree or not. For soft polytomies that occurred farther in the past, homoplasy may overwrite phylogenetic signal, such that each gene tree effectively becomes a hard polytomy with internal branch lengths that do not differ significantly from zero in a statistical framework.

Insufficient phylogenetic characters

Operationally, insufficient data in a phylogeny is represented by polytomies, in which three or more descendant lineages diverge from a single node. "Soft" polytomies may be due to insufficient data and often can be resolved by adding more or different kinds of characters (Maddison, 1989; DeSalle et al., 1994). However, previous studies have shown that even massive datasets may fail to get the true tree (Kuhner and Felsenstein, 1994, Takezaki and Gojobori, 1999; Qiu et al., 1999) when other biological factors compound the analysis. In contrast, a "hard" polytomy represents the simultaneous origin of three or more gene lineages from a common ancestor and has no bifurcating resolution. Instances in which intervals between successive branching events are too short to accumulate informative variation also are effectively hard polytomies (Hoelzer and Melnick, 1994). Hard polytomies can be identified by internal branch lengths that do not differ significantly from zero (Walsh et al., 1999).

Recombination, hybridization and reticulate evolution

Recombination is the reciprocal exchange of genetic material between two homologous chromosomes during meiosis (Griffiths et al., 2005). Its frequency varies between loci, is influenced by chromosomal location (regions near centromeres and telomeres show little recombination) and sequence structure, and has been found to occur within single genes (Zhang and Hewitt, 2003). Intragenic recombination generates alleles that are chimeric between parental alleles (Small et al., 2004) and therefore when it has occurred, the evolutionary history of a set of sequences forms a group of contradictory phylogenetic trees rather than a single tree (Zhang and Hewitt, 2003). When using nuclear genes for phylogeny reconstruction, avoiding regions that have been influenced by recombination may be difficult, because a positive correlation exists between recombination rate and the level of sequence polymorphism, and polymorphism is desired in such analyses (Zhang and Hewitt, 2003). Therefore, genomic regions with low recombination rates may not have enough sequence variation for phylogenetic analysis. The likely strategy for overcoming this problem is to analyse recombination events in the dataset and incorporate them into the models of evolution. In general, chloroplast genome does not recombine; it is a conserved maternally inherited organelle, which would maintain its integrity throughout hybridization, except in a few plant systems.

Several methods have been developed for detecting the presence of recombination, identifying the parental and recombinant individuals and approximating the positions of recombination breakpoints (reviewed in Posada and Crandall, 2001). These generally involve four different detection strategies: distance methods that look for inversions of distance patterns among the sequences; phylogenetic methods that compare the branching patterns of adjacent sequences; compatibility methods that partition phylogenetic incongruence site by site; and substitution distribution methods that look for a significant clustering of nucleotide substitutions or an expected statistical distribution (Posada and Crandall, 2001). Detection of recombination by more than one method should always be attempted before conclusions are drawn about the presence of recombination.

Tree reconstruction models and rate heterogeneity

Evolutionary rate heterogeneity among sites or lineages can cause phylogenetic problems. Simulation studies have shown that correct phylogenetic reconstruction can be hampered by heterogeneity in molecular evolutionary rates among lineages or sites, but that except in extreme cases, it should be possible to reconstruct the correct tree by selection of appropriate methods, models, and parameters (Felsenstein, 1988; Li, 1997; Siddall, 1998). Simulations using matrices with up to ninefold substitution rate differences among taxa indicate that ML may be robust against unequal rate effects (Li, 1997, p. 135). Neighbor joining too is considered robust against unequal rates, as long as distances are estimated accurately (Felsenstein, 1988; Li, 1997). However, the tenet that some methods are less robust in recovering the correct phylogenies may not always hold: Russo et al. (1996) evaluated the different construction methods in recovering the true tree of a known phylogeny, and concluded that the different trees obtained by these methods were congruent. Doolittle (1999) concluded that molecular phylogeneticists may fail to find the 'true tree', not because their methods are inadequate or because we have chosen the wrong genes, "but because the history of life cannot properly be represented as a "tree".

LONG BRANCH ATTRACTION

Long Branch Attraction (LBA; Felsenstein, 1978; Carmean and Crespi, 1995; Andersson and Swofford, 2004) is a phenomenon (most pronounced in maximum parsimony) when rapidly evolving lineages are inferred to be closely related, regardless of their true evolutionary relationships. The problem arises when the DNA of two (or more) lineages evolve rapidly. There are only four possible nucleotides and when DNA substitution rates are high, the probability that two lineages will convergently evolve the same nucleotide at the same site increases. When this happens, one can erroneously interpret this similarity as a synapomorphy (that is, evolving once in the common ancestor of the two lineages). Methods suggested to avoid LBA-artifacts include excluding long-branch taxa, excluding faster evolving third codon positions, using inference methods less sensitive to LBA such as maximum likelihood, adding taxa that are related to those with the long branches to break up the long branches, and sampling more characters especially of another kind, (Felsenstein, 2003; Bergsten, 2005).

Monophyletic and paraphyletic taxa

There has been ongoing discussion regarding the concept of paraphyletic taxa (Brummit, 1996; Funk, 2001; Nelson et al., 2003; Brummit, 2003). In this paper, phylogenetic definition of the concept is adopted: a monophyletic group is a group consisting of members descended from a single most recent common ancestor, whereas "paraphyletic" means a "clade" that includes within it, a non-descendant of their most recent common ancestor. Cladistics has alternative definition of paraphyletic group as a group whose members are descended from a common ancestor, but which does not include all of the known or considered descendants of that common ancestor. Brummit (2002) gives an example of such definition as: one, which includes a single common ancestor, but not all its descendants. Implicit is that a paraphyletic taxon is a monophyletic taxon in which a member, other than the most recent common ancestor, is excluded. Each of these definitions may be right in the context of the corresponding disciplines, which differ in principle and aim. Since cladistics is an older 'lineage', a phylogenetic definition of paraphyly might sound like walking north in a southbound train. However, the definition is well considered, and is shared by phylogenetic doyens such as Joe Felsenstein (Felsenstein, personal communication to first author).

CONCLUSION

It is clear from this review that majority of the problems encountered in phylogenetic inference arise from biological, genealogical or statistical causes, rather than from the non repeatability of 'old' molecular tools. We have shown that a greater majority of these problems are typical to sequence data, which are today generated by state of the art methods. Microsatellite, whose usefulness in higher order phylogeny has been unjustifiably criticised, is today perhaps the cheapest and most efficient in capturing molecular variation among taxa. We continue to support their use in phylogenetic inference as long as the number of generations since divergence so permits. While accepting that many samples per clade are necessary in population genetic studies to account for sampling variance, we support the view that in phylogenetics, the priority shifts from the number of taxa to the number of characters. The utility of fewer samples per clade and cheaper but phylogenetically robust markers such as AFLP and microsatellites is an affordable option that can be explored, albeit with caution.

ACKNOWLEDGEMENT

The authors thank Professor Joseph Felsenstein, University of Washington, for sharing certain aspects of phylogenetics. This work was written while the first author was on research leave funded by the Federal Government of Australia.

REFERENCES

Andersson FE, Swofford DL (2004). Should we be worried about long-branch attraction in real data sets? Investigations using metazoan

18S rDNA. Mol. Phylogenet. Evol. 33:440-451
Arnheim N (1983). Concerted evolution of multigene families. In: Nei M, Koehn R (eds) Evolution of Genes and Proteins. Sinauer, Sunderland, MA, pp 38-61
Avise JC (2000). The history and formation of species. Harvard University Press, Cambridge, Massachusets
Avise JC, Ankney CD, Nelson WS (1990). Mitochondrial gene trees and the evolutionary relationship of mallard and black ducks. Evolution 44:1109-1119
Bader DA, Moret BME, Vawter L (2001). Industrial applications of high-performance computing for phylogeny reconstruction. Paper presented at Commercial Applications for High-Performance Computing (SPIE01)
Baldwin BG, Sanderson MJ, Porter MJ, Wojciechowski MF, Campbell CS, Donoghue MJ (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Annals of Misouri Botanic Gardens 82:247-277
Bergsten J (2005). A review of long-branch attraction. Cladistics 21:163-193
Brown JR, Warren PV (1998). Antibiotic discovery: Is it in the genes? Drug Discovery Today 3:564-566
Bruford MW, Wyne RK (1993). Microsatellites and their application to population genetic studies. Current Opinion in Genetics & Development 3:939-943.
Brummitt RK (2002). How to chop up a tree. Taxon 51:31-41
Brummitt RK (2003). Further dogged defence of paraphyletic taxa. Taxon 52:803-804
Carmean D, Crespi BJ (1995). Do long branches attract flies? Nature 373:666-666
Cavalli-Sforza LL, Edwards WF (1967). Phylogenetic analysis: models and estimation procedures. Evolution 21:550-570
Chambers JK, McDonald LE (2000). A G protein-coupled receptor for UDP-glucose. J. Biol. Chem. 275:10767-10771
Cummings MP, Otto SP, Wakeley L (1995). Sampling properties of DNA sequence data in phylogenetic analyses. Mol. Biol. Evol. 12:814-822
Cunningham C (1997). Is incongruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. Syst. Biol. 46:464-478
Demastes JW, Remsen Jr. JC (1994). The genus Caryonthraustes (*Cardinalinae*) is not monophyletic. Wilson Bull. 106:733-738
DeSalle R, Absher R, Amato G (1994). Speciation and phylogenetic resolution. Trends Ecol. Evol. 9:297-298
Doolittle WF (1999). Phylogenetic classifcation and the universal tree. Science 284:2124-2129
Dover G (1982). Molecular drive: a cohesive mode of species evolution. Nature 299:111–117
Dumolin-Lapegue S, Demesure B, Fineschi S, Le Corre V, Petit RJ (1997). Phylogeographic structure of white oaks throughout the European continent. Genetics 146:1475-1487
Dunning AM, Talmud P, Humphries SE (1988). Errors in the Polymerase Chain Reaction. Nucleic Acids Research 16:10393.
Efron B (1979). Bootstrap methods: Another look at the jackknife. Ann. Stat. 7:1-26
Farris JS (1969). A successive approximations approach to character weighting. Syst. Zool. 18:374-385
Feldman MW, Bergman A, Pollock DD, Goldstein DB (1997). Micro-satellite genetic distances with range constraints: Analytical descrip-tion and problems of estimation. Genetics 29:207-216
Felsenstein J (1978). Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401-410
Felsenstein J (1985). Phylogenies and the comparative method. Am Naturalist 125:1-15
Felsenstein J (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791
Felsenstein J (1988). Phylogenies from molecular sequences: Inference and reliability. Ann. Rev. Genet. 22:521-565
Felsenstein J (2003). Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, MA
Futuyma DJ (1998). Evolutionary Biology. Sinauer Assocociates, Sun- derland, Massachusets
Garza JC, Freimer NB (1996). Homoplasy for size at microsatellite loci in humans and chimpanzees. Genome Res.6:211-217
Goldstein DB, Pollock DD (1997). Launching Microsatellites: A Review of Mutation Processes and Methods of Phylogenetic Inference. Heredity 88:335-342.
Goldstein DB, Roemer G, Smith D, Reich DE, Bergman A, Wayne R (1999). The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics 151:797-801
Goldstein DB, Ruiz LA, Cavalli-Sforza LL, Feldman MW (1995). An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471
Goloboff PA (1993). Estimating character weights during tree search. Cladistics 9:83-91
Griffiths AJF, Wessler SR, Lewontin RC, Gelbart WM, Susuki DT, Miller JH (2005). An introduction to genetic analysis. W.H. Freeman, New York
Gu X (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. 18:453-464
Hoelzer GA, Melnick DJ (1994). Patterns of speciation and limits to phylogenetic resolution. Trends in Ecology and Evolution 9:104-107
Kalinowski ST (2005). Do polymorphic loci require large sample sizes to estimate genetic distances? Heredity 94:33-36
Kim J, Warnow T (1999). Tutorial on phylogenetic tree estimation. 1999. Paper presented at Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB99)
Kimura M, Crow JF (1964). The number of alleles that can be maintained in a finite population. Genetics 49:725-738
Kimura M, Ohta T (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. USA. 75:2868-2872
Kluge AG, Farris JS (1969). Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1-32
Kuhner MK, Felsenstein J (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459-468
Kuhner MK, Felsenstein J (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 4:406-425
Li W-H (1997). Molecular Evolution. Sinauer Associates, Sunderland, MA
Lu Y (2001). Roles of lineage sorting and phylogenetic relationship in the genetic diversity at the self-incompatibility locus of Solanaceae. Heredity 86:195-205
Maddison WP (1989). Reconstructing character evolution on polytomous cladograms. Cladistics 5:365-377
Maddison WP (1997). Gene trees in species trees. Syst. Biol. 46:523-536
Martins EP (1995). Phylogenies and comparative data, a microevolutionary perspective. Philosophical Transac. Roy. Soc. Lond. 349:85-91.
McCracken KG, Sorenson. MD (2005). Is Homoplasy or Lineage Sorting the Source of Incongruent mtDNA and Nuclear Gene Trees in the Stiff-Tailed Ducks (Nomonyx-Oxyura)? Syst. Biol. 54:35-55.
McKinnon GE, Steane DA, Potts BM, Vaillancourt RE (1999). Incongruence between chloroplast and species phylogenies in *Eucalyptus* subgenus *Monocalyptus* (Myrtaceae). Am. J. Bot. 86:1038-1046
Meacham CA, Estabrook GF (1985). Compatibility methods in systematics. Ann. Rev. Ecol. Syst. 16:431-446
Nei M (1972). Genetic distance between populations. Am. Naturalist 106:283-292
Nei M (1978). Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. Genetics 89:583-590
Nei M (1987). Mol. Evol. Genet. Columbia University Press, New York
Nelson G, Murphy DJ, Ladiges PY (2003). Brummitt on paraphyly: a response. Taxonomy 52:295-298
Pamilo P, nei M (1988). Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568-583
Parra-O C, Bayly M, Udovicic F, Ladiges P (2006). ETS sequences support the monophyly of the eucalypt genus *Corymbia* (Myrtaceae). Taxon: In press

Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998). Cross-Species Amplification of Soybean (*Glycine max*) Simple Sequence Repeats (SSRs) Within the Genus and Other Legume Genera: Implications for the Transferability of SSRs in Plants. Mol. Biol. Evol. 15:1275-1267

Pepper AE, Norwood LE (2001). Evolution of Caulanthus amplexicaulis var. Barbarae (Brassicaceae), a rare serpentine endemic plant: A molecular phylogenetic perspective 1. Am. J. Bot. 88:1479-1489

Petren KB, Grant R, Grant PR (1999). A phylogeny of Darwin's finches based on microsatellite DNA length variation. Proc. Roy. Soc. Lond. B 266:321-329

Posada D, Crandall KA (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA. 98:13757-13762

Qiu Y-LJL, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis MJ, Zimmer EA, Chen Z, Savolainen V, Chase MW (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404-407

Renner SS, Chanderbali S (2000). What is the relationship among Hernandiaceae, Lauraceae, and Monimiaceae, and why is this question so difficult to answer? Int. J. Plant Sci 161:s109-s119

Richard M, Thorpe RS (2001). Can microsatellites be used to infer phylogenies? Evidence from population affinities of the western Canary Island lizard (Gallotia galloti). Mol. Phylogenet. Evol. 20:351-360

Ritz LR, Glowatzki-Mullis ML, MacHugh DE, Gaillard C (2000). Phylogenetic analysis of the tribe Bovini using microsatellites. Anim. Genet. 31:178-185

Russo CAM, Takezaki N, Nei M (1996). Efficiencies of Different Genes and Different Tree-building Methods in Recovering a Known Vertebrate Phylogeny. Mol. Biol, Evol. 13:525-536

Ruzzante DE (1998). A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. canadian Journal of Fish Aquatic Science 55:1-14

Saitou N, Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425

Siddall ME (1998). Success of parsimony in the four-taxon case: longbranch repulsion by likelihood in the Farris zone. Cladistics 14:209-220

Small RL, Cronn RC, Wendel JF (2004). Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Bot. 17:145-170

Steane DA, McKinnon GE, Vaillancourt RE, Potts BM (1999). ITS sequence data resolves higher level relationships among the eucalypts. Mol. Phylogenet. Evol. 12:215-223

Steane DA, Nicolle D, McKinnon GE, Vaillancourt RE, Potts BM (2002).

Higher-level relationships among the eucalypts are rerolved by ITS-sequence data. Aust. Syst. Bot. 15:49-62

Streelman JT, Zardoya R, Meyer A, Karl SA (1998). Multi-locus phylogeny of cichlid fishes (Pisces: Perciformes): evolutionary comparison of microsatellite and single-copy nuclear loci. Mol. Biol. Evol. 15:798-808

Takahashi K, Terai Y, Nishida M, Okada N (2001). Phylogenetic Relationships and Ancient Incomplete Lineage Sorting Among Cichlid Fishes in Lake Tanganyika as Revealed by Analysis of the Insertion of Retroposons. Mol. Biol. Evol. 18:2057-2066

Takezaki N, Gojobori T (1999). Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences.

Takahata N (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957-966Mol. Biol. Evol. 16:590-601

Takezaki N, Nei M (1996). Genetic Distances and Reconstruction of Phylogenetic Trees From Microsatellite DNA. Genetics 144:389-399

Udovicic F, Ladiges PY (2000). Informativeness of nuclear and chloroplast DNA regions and the phylogeny of the eucalypts and related genera (Myrtaceae). Kew Bull. 55:633-645

Udovicic F, McFadden GI, Ladiges PY (1995). Phylogeny of Eucalyptus and Angophora based on 5S rDNA spacer sequence data. Mol. Phylogenet. Evol. 4:247-256.

Walsh HE, Kidd MG, Moum T, Friesen VL (1999). Polytomies and the power of phylogenetic inference. Evolution 53:932-937

Whitton J, Rieseberg LH, Ungerer MC (1997). Microsatellite loci are not conserved across the Asteraceae. Mol. Biol. Evol. 14:204-209

Zhang D, Hewitt GM (2003). Nuclear DNA analyses in genetic studies of populations: Practice, problems and prospects. Mol. Ecol. 12:563-584

Zhu Y, Queller DC, Strassmann JE (2000). A phylogenetic perspective on sequence evolution in microsatellite loci. J. Mol. Evol. 50:324-338

Zink RM (2002). A new perspective on the evolutionary history of Darwin's finches. The Auk. 119:864- 871