Critical Reviews in Plant Sciences, 34:454–468, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 0735-2689 print / 1549-7836 online
DOI: 10.1080/07352689.2015.1047712

A Primer to Molecular Phylogenetic Analysis in Plants

Ayse Ozgur Uncu, Ali Tevfik Uncu, İbrahim Celık, Sami Doganlar and Anne Frary
Izmir Institute of Technology, Department of Molecular Biology & Genetics, Urla, Izmir, Turkey

Table of Contents

Reconstructing a tree of life by inferring evolutionary history is an important focus of evolutionary biology. Phylogenetic reconstructions also provide useful information for a range of scientific disciplines such as botany, zoology, phylogeography, archaeology and biological anthropology. Until the development of protein and DNA sequencing techniques in the 1960s and 1970s, phylogenetic reconstructions were based on fossil records and comparative morphological/physiological analyses. Since then, progress in molecular phylogenetics has compensated for some of the shortcomings of phenotype-based comparisons. Comparisons at the molecular level increase the accuracy of phylogenetic inference because there is no environmental influence on DNA/peptide sequences and evaluation of sequence similarity is not subjective. While the number of morphological/physiological characters that are sufficiently conserved for phylogenetic inference is limited, molecular data provide a large number of datapoints and enable comparisons from diverse taxa. Over the last 20 years, developments in molecular phylogenetics have greatly contributed to our understanding of plant evolutionary relationships. Regions in the plant nuclear and organellar genomes that are optimal for phylogenetic inference have been determined and recent advances in DNA sequencing techniques have enabled comparisons at the whole genome level. Sequences from the nuclear and organellar genomes of thousands of plant species are readily available in public databases, enabling researchers without access to molecular biology tools to investigate phylogenetic relationships by sequence comparisons using the appropriate nucleotide substitution models and tree building algorithms. In the present review, the statistical models and algorithms used to reconstruct phylogenetic trees are introduced and advances in the exploration and utilization of plant genomes for molecular phylogenetic analyses are discussed.

Keywords Bayesian methods, distance methods, maximum likelihood methods, maximum parsimony methods, molecular evolution

Address correspondence to Anne Frary, Izmir Insitute of Technology, Department of Molecular Biology & Genetics, Urla, Izmir 35430, Turkey. E-mail: annefrary@iyte.edu.tr

I. INTRODUCTION

Phylogenetic inferences are not only used for reconstructing the evolutionary history of living things on earth, but also provide evidence of the climatic and geological history of the earth. A phylogenetic tree displays taxonomic groups in a hierarchical order (Futuyma, 2005). For example in a species dendrogram reconstructed according to hierarchical clustering, the hypothesized evolutionary history of a species is displayed as a branching tree (Figure 1). In such a tree, species or taxonomic groups whose evolution from a common ancestor is relatively recent, share a common branch point (node) and they are clustered together as a monophyletic group. The first common node shared by two sister taxa represents the hypothetical common ancestor. The number of characters shared among taxa increases toward branch tips. This means that two sister taxa share a greater number of common characters than the clade they constitute shares with another clade.

While fossils seem to be the ideal material for evolutionary research, the incompleteness of the fossil record makes it impossible to reconstruct a complete tree of life. Moreover, in most cases, fossils enable observation of only morphological characters. Thus, many scientists have moved toward comparative morphological and physiological analysis (Nei and Kumar, 2000). However, phylogenetic analysis based on only morphological/physiological data has several shortcomings. One of the most important of these is the limited number of data points provided by such comparative analyses. Progress in molecular biology has led to the increased use of sequence data in phylogenetic analysis. In addition to the vast number of potential data points provided by DNA/peptide sequence analysis, it is important to note that, except for a group of viruses that store their genetic information in the form of RNA, DNA is the common genetic material of all life forms. Thus, analysis at this level enables the comparison of organisms from diverse taxa. For example, it is possible to perform a phylogenetic analysis that includes both plant and animal species by using molecular data whereas such a comparison would be impossible using morphological characters. In addition to the limited number of morphological/physiological characters and an inadequate level of their conservation among diverse taxa, another critical issue is that such characters are often subject to the variable influence of the environment. A morphological character that is manifested similarly in two taxa is not necessarily an ancestral or shared derived character and similarity may be the result of environmental influence, mimicry or convergent evolution. Thus, phylogenetic analysis based on such a character would be misleading. Moreover, evolutionary change of morphological/physiological characters is a complex process, involving multiple independent and dependent events that define a change in a single character. It should be kept in mind that the consequent phenotype is a result of the combinatory effect of the genotype and the environment, including the physiological and cellular environments that modulate the expression of the genotype. As a result of such complexity, it is difficult to extract information that correctly reflects evolutionary history. In addition, interpretation of such characters requires expertise and there is the risk of subjectivity. Due to these shortcomings and problems, trees reconstructed based on morphological comparisons often reflect inconsistent phylogenies (Nei and Kumar, 2000). However, regardless of the type of data used, it should be noted that, phylogenies are reconstructions of evolutionary events and are inferred from the available evidence. Therefore, every phylogenetic tree is indeed an estimation.

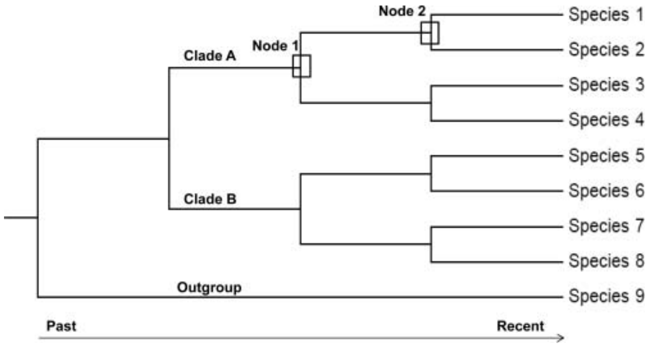

FIG. 1. A representative phylogenetic tree displaying the evolutionary relationships among nine plant species. In the tree, divergence events are represented as bifurcations at each node with relatively more recent events toward the branch tips. The topology of the displayed phylogenetic tree consists of two clades (Clades A and B) and one outgroup (Species 9). A clade is a group of taxa that are descended from a common ancestor. In Clade A, the common hypothetical ancestor shared by species 1-4 is indicated as Node 1. The outgroup (Species 9) is equally distant from Clade A and B, and indicates the root of the phylogenetic tree. A monophyletic clade consists of a common ancestral taxon with all its descendent taxa. Clade A and B are two monophyletic clades. Sister taxa are taxa that diverge from the most recent common ancestor. Species 1 and 2 are sister taxa and their common hypothetical ancestor is indicated as Node 2.

For researchers who are interested in using sequence data in their analyses, clade specific databases such as TAIR for Arabidopsis (Swarbreck *et al.*, 2008), Gramene for grasses (Liang *et al.*, 2008), SGN for Solanaceae (Bombarely et al., 2011), GDR for Rosaceae (Jung *et al.*, 2008) and LIS for legumes (Gonzales et al., 2005), and the GenBank database that hosts all publicly available sequences (Benson *et al.*, 2013) constitute freely and readily accessible data resources. In addition, comparative plant genomic databases such as GreenPhylDB (Conte *et al.*, 2008), Plaza (Proost *et al.*, 2009) and Phytozome (Goodstein *et al.*, 2012) provide both sequence data and analyses of plant genomes and genes, therefore, enable both data mining and comparative evolutionary analyses of plant genes.

II. MOLECULAR PHYLOGENETICS

A. Evolutionary Changes in Amino Acid Sequences

Substantial progress in molecular phylogenetics was made during the 1960s and 1970s due to the development of protein sequencing methodologies (Graur and Li, 2000). Work with protein sequences during these two decades led to the molecular clock (Zuckerkandl and Pauling, 1962, 1965; Margoliash, 1963) and evolution by gene duplication (Ingram, 1963; Ohno, 1970) hypotheses. Since the development of rapid and reliable DNA sequencing techniques (Sanger *et al*., 1977), nucleotide sequence data have been extensively used in phylogenetic studies. However, amino acid sequences are more conserved than nucleotide sequences and optimal alignment of the nucleotide sequences of protein-coding genes requires a reference peptide sequence to define the homologous sequence portions. Thus, amino acid sequences are still very valuable for phylogenetic analyses. Mutations that are fixed in DNA sequences by natural selection and/or genetic drift (Hartl and Clark, 1997) constitute the basis of evolution (Nei and Kumar, 2000). However, over long evolutionary periods, a large number of nucleotide substitutions accumulate, resulting in a loss of information for accurate alignment of DNA sequences. Due to the degeneracy of the genetic code (Watson *et al*., 2008), not every nucleotide substitution is manifested as an amino acid substitution in a peptide sequence. As a result, amino acid sequences display a higher degree of conservation over evolutionary time. Therefore for a more accurate approximation of the true phylogeny, amino acid sequences should be used instead of nucleotide sequences when investigating evolutionary relationships among organisms from distant taxa.

Phylogentic comparisons of proteins require the alignment of peptide sequences using multiple sequence alignment programs such as ClustalW (Thompson *et al*., 1994) and MUSCLE (Edgar, 2004). Although insertions and deletions (indels) occur in peptide sequences over the course of evolution, indels are often eliminated while calculating the extent of divergence between two sequences and only substitutions are taken into account. The simplest measure of the extent of evolutionary divergence between two peptides is the number of amino acid differences (n_d). However, because the real extent of evolutionary divergence is determined not only by the number of amino acid substitutions but also by the length of the compared sequences, a more reliable measure of divergence is not the number, but the proportion of amino acid substitutions. This proportion is referred to as p distance. A lower p distance is associated with more recent divergence of two taxonomic units. Such a relationship between divergence time and p distance suggests that, for a given peptide sequence, the number of accumulated amino acid substitutions increases with increasing time after divergence (Nei and Kumar, 2000). This observation led to the molecular clock hypothesis which states that a roughly linear relationship exists between the number of amino acid substitutions and divergence time (Zuckerkandl and Pauling, 1962; Margoliash, 1963; Zuckerkandl and Pauling, 1965). When the molecular clock hypothesis is valid and, thus, the rate of substitution for a given protein/DNA sequence among lineages is constant, p distance can be used to estimate the time of divergence from a common ancestor (Futuyma, 2005).

In cases where there is not a linear relationship between p distance and time of divergence, the molecular clock is invalid (Nei and Kumar, 2000). One reason for the deviation from a linear correlation is superimposed substitutions (multiple hits) that occur at a single locus. Multiple hits lead to a discrepancy between the actual number of amino acid substitutions and the number counted (n_d) (Nei and Kumar, 2000). To take into account multiple hits, PC distance (Poisson Correction distance), calculated using the Poisson distribution (Zuckerkandl and Pauling, 1965), can be used, and is a more accurate estimation of the number of substitutions compared to p distance.

A second reason for a nonlinear correlation between divergence time and p distance is functional constraint which causes differential conservation of individual amino acid residues in a peptide sequence. p distance is calculated with the assumption of a constant substitution rate over an entire peptide sequence (Nei and Kumar, 2000). Multiple hits and differential substitution rates are not taken into account in p distance. Substitution rate varies with the functional constraint of the sequence being considered (Kimura, 1983). Strong functional constraint prohibits or limits alteration of an amino acid for a site to a narrow range of alternative amino acids, resulting in a relatively slow rate of substitution and evolution (Blouin *et al*., 2003; Barriere *et al*., 2011). The actual, differential substitution rates for different sites in a peptide vary according to the gamma distribution (Ota and Nei, 1994). Gamma distance (d_G), is the distance corrected according to a gamma distribution with an appropriate gamma parameter. The gamma parameter is estimated based on the principle that, when the rate of substitution among sites follows a gamma distribution, the observed number of substitutions per site follows a negative binomial distribution. The gamma parameter is also called the shape parameter, as the shape of the gamma distribution depends on the value of the gamma parameter α (>0). When $\alpha \leq 1$, the distribution is exponential, and for values of α greater than 1, the gamma distribution starts to resemble normal distribution. Smaller values of α (<1) reflect greater variation in substitution rates among sites (Yang, 1996; Nei and Kumar, 2000). The difference between the values of p distance and d_G increases with an increasing number of different amino acids (i.e., increased divergence) between two peptides. Similarly, the smaller the number of amino acid differences between two peptides, the lower the discrepancy between distances calculated using these two methods. In cases where the calculated p distance is lower than 0.2, and α is higher than 0.65, the difference between p and d_G is insignificant and p distance is preferentially used, as it is based on a simpler model with fewer parameters (Nei and Kumar, 2000).

In a peptide sequence, not only the position of an amino acid, but also the amino acids that interchange, define the substitution rate of a given site. Due to similar biochemical properties, some amino acids, such as lysine and arginine which are both basic, are more likely to interchange over the course of evolution (Dayhoff, 1972). Grishin distance (d_R) (Grishin,

1995) takes into account the differential probability of interchange between amino acid pairs. Another method for estimating evolutionary relationships among peptides is an empiricial method developed by Dayhoff *et al.* (1978). This method uses an amino acid substitution matrix that displays the probability of interchange between amino acid pairs that was constructed using data from experimental work with well-conserved proteins such as hemoglobins, cytochrome c and fibrinopeptides. Using this matrix, amino acid substitutions over the course of evolution are estimated for a given peptide and Dayhoff distance (d_D) between peptides is calculated.

Calculation of Gamma, Grishin and Dayhoff distances takes into account factors that influence the evolution of peptide sequences theoretically resulting to a more accurate estimation of evolutionary relationships. However, more complex substitution models introduce additional parameters to the calculation of distance which, in turn, result in higher variances and standard errors in these estimates. Therefore, when the divergence time of the taxonomic units under study is relatively recent, it is more appropriate to use simpler substitution models for more precise calculations. For example, when p distance is below 0.2, there is no need to use more complex models, p distance itself should be used (Nei and Kumar, 2000).

B. Evolutionary Changes in DNA Sequences

Genomes are composed of different types of nucleotide sequences, such as coding regions, non-coding regions, exons, introns and repetitive elements. Evolution of these distinct sequence types follows different patterns, thus their analysis requires different statistical models (Nei and Kumar, 2000; Graur and Li, 2000).

In protein coding sequences, degeneracy of the genetic code (Watson *et al.*, 2008) results in differential substitution rates for the first, second and third codon positions. Degeneracy refers to the fact that 61 codons encode only 20 amino acids, therefore some codons, called synonymous codons, encode the same amino acid. The second nucleotide in a codon is conserved among synonymous codons and all changes at this codon position are nonsynonymous and result in an amino acid substitution. However, changes in the first and third positions of codons are more likely to be synonymous. In fact due to wobble, only 31% of substitutions at the third position alter the amino acid sequence (Nei and Kumar, 2000). Thus, a mutation in the second codon position will have a low probability of fixation because of purifying (negative) selection while a mutation in the third position will more likely get fixed over the course of evolution. In addition, depending on the strength of functional constraint, nonsynonymous substitutions in coding sequences are eliminated by purifying selection or may be fixed if the mutation introduces a selective advantage.

Different DNA sequences diverge from a common ancestral sequence as a result of nucleotide substitutions. A simple measure of the distance between two DNA sequences is the p distance. p distance is the proportion of different nucleotides between two sequences determined by counting the total number of differences and dividing by the total number of nucleotides in the sequence examined. As with amino acid sequences, however, increasing evolutionary distance will result in a concomitantly increasing discrepancy between the calculated p distance and actual evolutionary distance. Multiple hits including parallel and convergent substitutions accumulate over evolutionary time and are the cause of this discrepancy (Nei and Kumar, 2000). Therefore, mathematical models are essential for better estimation of the evolutionary distance between nucleotide sequences with greater evolutionary divergence.

The Jukes-Cantor model of nucleotide substitution (Jukes and Cantor, 1969) is among the simplest models. While this model includes the probability of multiple substitutions at the same site in distance estimation, differential substitution rates of individual nucleotides in a sequence are not taken into account. Thus, only one substitution rate is used to estimate the probability of change from one nucleotide to any other nucleotide at a particular site. In addition, transitional and transversional mutation rates are assumed to be equal in the model. Although experimental data indicate that the probabilities of transitional and transversional mutations are often not equal and that transitions are more likely to occur (Nei and Kumar, 2000), the Jukes-Cantor model is often sufficient to obtain good estimates of divergence time and for reconstruction of phylogenetic trees.

If, however, the difference between transition and tranversion rates is very high, a more appropriate model should be selected. Kimura's two parameter model (Kimura, 1980) includes estimates of both transition and transversion rates in distance calculations. Similar to Kimura's two parameter model, transition/transversion bias is included in Tamura's nucleotide substitution model (Tamura, 1992). In addition, Tamura's model extends Kimura's two parameter model by taking into account the unequal nucleotide frequencies in a DNA sequence. Many models assume that all four nucleotides occur at equal frequencies in a sequence. However, this is generally not the case and a GC content of 50% is rarely observed. For example, it is known that the GC content of coding portions of the genome is higher than that of non-coding portions (Messeguer *et al.*, 1991). In addition to the models described, there are other substitution models such as Hasegawa's (Hasegawa *et al.*, 1985) and Tamura and Nei's (Tamura and Nei, 1993), which involve more parameters and require more complicated calculations.

Equal substitution rates for all sites in a DNA sequence is an assumption of all of the models described above. However, as with peptide sequences, substitution rates may not be equal for all genomic regions. For example, substitution rates of intronic regions are often higher than that of exonic regions due to lower selection pressure on non-coding sequences. The bias in substitution rates across different sites follows a gamma distribution and can be corrected by using an experimentally determined gamma parameter (α). Taking into account substitution rate bias is a more realistic approach and gamma distance will give a more accurate measure of the extent of divergence, which is represented in the form of branch lengths in a phylogenetic tree. However, at the same time, the gamma parameter introduces additional variance to a distance estimate. Therefore, unless the number of nucleotides compared is significantly high, gamma distance does not necessarily estimate a more accurate phylogeny (Nei and Kumar, 2000).

Given the availability of so many models of nucleotide substitution, model selection is an important consideration for every phylogenetic study. Model choice should be made according to preliminary knowledge of the taxonomic units and DNA sequence under study. Simpler models that involve fewer parameters are more suitable for closely related nucleotide sequences. Greater evolutionary distances may require additional parameters, thus, more complex models. However, using a more complex model that better fits the data set is not a prerequisite to obtain the correct topology when reconstructing a phylogenetic tree. Indeed, the more complex the substitution model, the higher the variance calculated. Thus, although a correct mathematical model will result in more accurate branch length calculations, it may not be as efficient in reflecting the true evolutionary topology (Nei and Kumar, 2000).

C. Reconstruction of Phylogenetic Trees

Because evolutionary history cannot be directly observed, it must be inferred by comparative morphological/physiological or molecular analyses. Reconstructing a tree of life by resolving evolutionary and genealogical relationships among organisms has been an important focus of evolutionary biology since the late 1800s (Futuyma, 2005). In addition to showing the evolutionary relationships among taxa, phylogenetic trees are useful for understanding adaptive evolution and the evolution of multigene families. Several statistical methods are available for reconstructing phylogenetic trees based on molecular data. Such methods can be classified into three main groups: distance methods, parsimony methods and maximum likelihood methods (Nei and Kumar, 2000). Different algorithms are available for processing nucleotide or amino acid sequences to generate phylogenetic trees that reflect the estimated evolutionary relationships among taxonomic units. In a phylogenetic tree, inferred evolutionary relationships are displayed by tree topology and branch lengths. Therefore, approximation of these two parameters should be as accurate as possible. While branch length calculations are based on relatively simple statistical models, estimation of the true topology is challenging due to the large number of possible topologies. For example, when reconstructing a rooted phylogenetic tree of five taxa, the number of alternative topologies calculated according to Cavalli-Sforza and Edwards (1967) is 105, and for six taxa, it is 945. Thus, small increases in the number of taxa examined lead to drastic increases in the number of alternative topologies. When ten taxa are examined with the same method, the number of alternative topologies becomes so high that algorithms may be required to limit the search to a subset of topologies.

1. Distance methods

Distance methods rely on simpler calculations and algorithms compared to parsimony, maximum likelihood and Bayesian methods. Distance methods include UPGMA (Unweighted Pair Group Method with Arithmetic Averages), Least Squares, Minimum Evolution and Neighbor Joining.

UPGMA (Sokal and Michener, 1958; Sneath and Sokal, 1973) is considered the simplest of distance methods. This method assumes a constant rate of substitution among lineages. In the UPGMA method, evolutionary relationships are determined according to a substitution model and a distance matrix is generated for clustering. A distance matrix can be based on biochemical, morphological or DNA/protein sequence data and presents pairwise distances for each pair of taxa under study. In UPGMA, clustering starts with the two taxa with the smallest distance and these taxa are treated in the subsequent step as a composite taxon. The next step is the recalculation of distance values to generate another distance matrix to determine a new pair of taxa with the smallest pairwise distance. The process continues until all taxa are clustered. Because of its simplistic nature, UPGMA may result in an incorrect topology when the molecular clock hypothesis is not valid for the lineages examined (Nei and Kumar, 2000). For this reason, UPGMA is not recommended for phylogenetic analysis but can be used for other types of clustering analyses.

While the UPGMA method involves the assumption of a constant molecular clock, the Least Squares method allows unequal rates of substitution for different lineages. Thus, the Least Squares method is more appropriate when the rate of evolution is not constant among lineages. In this method, the residual sum of squares is calculated for every possible topology, and the topology that gives the smallest value is selected. The Fitch-Margoliash (Fitch and Margoliash, 1967) or the least squares method (Rzhetsky and Nei, 1992, 1993) is used to calculate branch lengths. The Least Squares method ensures the selection of the topology with branch lengths comparable to the actual ones.

The Minimum Evolution method relies on a mathematical proof (Rhetsky and Nei, 1993) which states that, when sequence estimates do not deviate from real evolutionary distances, regardless of the number of sequences compared, the sum of branch lengths (S) is smallest for the topology that reflects the true phylogeny. S is calculated for all possible topologies to find the one with the smallest S. When the number of taxa for analysis is high, topology selection with such an algorithm is time consuming. Thus, the method can be used in combination with the Neighbor Joining method to reduce analysis time. In the combined approach, a Neighbor Joining tree is initially generated and S is calculated for the possible trees that are similar to the Neighbor Joining tree. The algorithm searches for a topology with a smaller S value compared to the Neighbor Joining tree, and this tree is selected as the provisional Minimum Evolution tree. The search for a smaller S continues by analyzing topologies that are similar to the provisional tree. The process proceeds until no other alternative topology with a smaller S value is left (Nei and Kumar, 2000).

The Neighbor Joining method relies on the minimum evolution principle and is a simplified version of this method (Saitou and Nei, 1987). In order to shorten the time for data processing, the Neighbor Joining algorithm does not analyze every possible topology. Instead, the minimum evolution principle is used while clustering each taxon. The algorithm first generates a star-like tree where all lineages branch from a single point. Then, every possible pair of taxa is tested as neighbors to find the two taxa with the smallest S. These two taxa are then treated as a composite taxon to be tested as neighbors with the remaining taxa, in search of the smallest S. The process is complete when all taxa are clustered.

Distance methods are often criticized for being too simplistic. However, they are far faster than more sophisticated and potentially more accurate methods such as Maximum Likelihood and Bayesian inference. This simplicity becomes an advantage when dealing with computationally challenging, massive datasets that require speed in data processing (Pardi and Gascuel, 2012). The most frequently used distance method, Neighbor Joining, was proven to be consistent in producing accurate phylogenies with exact distances or distances with very small errors. However, when dealing with distant taxa, distance methods may yield incorrect topologies as errors of distance estimates increase exponentially with increasing sequence divergence (Bruno et al., 2000). Nevertheless, to date, Neighbor Joining is the most cited method in phylogenetics (Pardi and Gascuel, 2012) and phylogenetic software packages are still being extended with new tools for distance based evolutionary analysis (Popescu et al., 2012).

2. *Maximum Parsimony methods*

The Maximum Parsimony method was initially developed for morphological characters and was used with amino acid sequence data for the first time by Eck and Dayhoff (1966). Maximum Parsimony algorithms for use with nucleotide sequence data were developed by Fitch (1971) and Hartigan (1973). These methods search for the topology that involves the smallest number of evolutionary steps (nucleotide or amino acid substitutions) and rely on fewer assumptions compared to distance and likelihood methods. Compared to other methods, Maximum Parsimony methods are more likely to produce topologies closer to the true phylogeny for smaller evolutionary distances. However, Maximum Parsimony methods reconstruct phylogenies under the assumption of a molecular clock, leading to topological errors when the substitution rate is not constant among lineages. In addition, the probability of obtaining an incorrect topology increases with an increasing number of parallel and reverse substitutions and, therefore, an increasing number of sites that are identical by state but not by descent, due to convergent evolution. The method calculates the total number of substitutions in a nucleotide or amino acid sequence and selects the topology that involves the smallest number of changes by neglecting the probability of parallel and backward substitutions. Thus, the Maximum Parsimony tree is the tree for which the tree length (L) is at a minimum. Maximum Parsimony methods are divided into two types: weighted and unweighted. In the unweighted method, the possibility of nucleotide/amino acid substitution is assumed to be constant in all directions. For example transitional and transversional substitutions are treated equally by the algorithm. Conversely, the weighted method assigns different weights for different types of substitutions. Therefore, the weighted method is more appropriate when the rate of transitional substitutions is not equal to that of transversional substitutions.

Several tree searching methods are used for selection of the Maximum Parsimony tree. One of these is the Exhaustive Search method which calculates L for every possible topology. Since topology searching requires considerable processing time, the method is not appropriate for analyses that involve a large number of taxa. In contrast to Exhaustive Search, the Specific Tree Search method calculates L only for potentially correct topologies. For this method, preliminary knowledge of the evolutionary relationships among analyzed taxa should be available to eliminate incorrect topologies. The Branch and Bound tree search method (Hendy and Penny, 1982; Kumar *et al.*, 1993) starts with the construction of a core tree that involves three taxa. The tree search algorithm continues with the addition of the remaining taxa to the core tree according to an order determined by a maximum of the minimum algorithm. An upperbound of tree length (L_u) is set for the elimination of topologies for which L exceeds L_u. The length of the shortest tree is set as L_u at each cycle of the algorithm to ensure that the topology with the shortest length is selected as the most parsimonious tree at the end of the process. While the method is effective for finding the true Maximum Parsimony tree, it is time consuming when analyzing data sets of more than 20 taxa (Nei and Kumar, 2000). Branch and Bound-like algorithms, which are faster than the Branch and Bound method, are available (Kumar *et al.*, 1993). However, since they examine fewer topologies, they may not be as effective as the original Branch and Bound algorithm in finding the Maximum Parsimony tree.

Heuristic tree search is a relatively fast method which examines only a subset of possible topologies. This method involves the construction of a temporary provisional tree of shortest length using a stepwise addition algorithm, and then the application of Branch Swapping algorithms to find the

Maximum Parsimony tree (Swofford and Begle, 1993). The most frequently used Branch Swapping Algorithms are Nearest Neighbor Interchange (NNI), Subtree Pruning Regrafting (SPR) and Tree Bisection Reconnection (TBR). NNI examines all trees that are different from the provisional tree by a topological distance of two. SPR separates the provisional tree into two: the pruned tree and the residual tree. The pruned tree is regrafted onto each branch of the residual tree to produce alternative topologies, in search of the Maximum Parsimony tree. Similar to SPR, the TBR algorithm separates the provisional tree into two subtrees. Subtrees are reconnected to produce alternative topologies. Among the three algorithms, TBR is the most frequently used, since it examines a larger number of alternative topologies (Nei and Kumar, 2000). However, the probability of finding the true topology decreases with an increasing number of taxa even with the TBR algorithm (Maddison, 1991). Application of multiple rounds of stepwise addition and TBR algorithms increases the chance of finding the true Maximum Parsimony tree.

Maximum Parsimony methods tend to underestimate the actual branch lengths. Therefore, they are often used for estimating only topology with Least Squares or Maximum Likelihood methods used for branch length calculations. When multiple topologies are selected as equally parsimonious, which is often the case, a composite tree, referred to as consensus tree, is required to represent all of the alternative topologies. While there are several different types of consensus trees (Swofford and Begle, 1993), the most frequently used are strict consensus trees, majority rule consensus trees and bootstrap consensus trees. Strict consensus trees display discrepancies among the branching patterns of equally parsimonious trees by multifurcations. In contrast, the branching patterns of majority rule consensus trees are only displayed as bifurcations, following the general assumption of evolutionary divergence as a bifurcating process. In such consensus trees, only the branching patterns that appear over a certain frequency are represented. When constructing bootstrap consensus trees, the reliability of the topology is tested using the bootstrap method. In the bootstrap test, a set of nucleotide sites is randomly resampled from the dataset with replacement, resulting in a new dataset with an equal number of nucleotides as the original dataset. This new, randomly resampled dataset is used to construct a new tree and the process of resampling and reconstruction is repeated multiple times (most commonly 1000s of times). The reliability of the branching pattern of the consensus tree is evaluated by calculating the percentage of times that a given branching pattern appears in replicate bootstrap trees. The bootstrap method is used not only to construct Maximum Parsimony consensus trees but also to test the reliability of phylogenetic trees reconstructed with all of the described methods.

The primary disadvantage of parsimony is that parameters that alter sequence evolution cannot be incorporated into the method. Therefore, superimposed and parallel substitutions lead to unreliable topologies, evidenced by a tendency of the method to group long branches together, a concept known as long-branch attraction. However, selection of substitution models that are too simplistic to explain a dataset also leads to long-branch attraction with distance, likelihood and Bayesian methods (Yang and Rannala, 2012).

3. Maximum Likelihood methods

An algorithm based on the Maximum Likelihood principle for reconstructing phylogenies using nucleotide sequence data was developed by Felsenstein (1981). Later, the algorithm was modified for amino acid sequence data by Kishino *et al.* (1990). The method aims at finding a tree which maximizes the probability of observing the data for a specific substitution model (Nei and Kumar 2000; Bromham, 2008; Hall, 2008). Phylogeny is inferred based on likelihood values and the topology with the highest likelihood is selected (Nei and Kumar, 2000). The Maximum Likelihood function calculates branch lengths by considering every possible nucleotide/amino acid for each interior node (hypothetical ancestor) to maximize the likelihood for each observed site. The probability of a topology is the sum of the probabilities calculated for each site. For easier computational handling, the probability (likelihood) is expressed as a log likelihood (Hall, 2008). Given the data and the model of substitution, the topology with the highest probability (log likelihood) is selected as the Maximum Likelihood tree. At any point of the tree search, the tree that is kept by the algorithm is the one with the highest likelihood among the trees that were examined (Bromham, 2008). Since all sites in a nucleotide/amino acid sequence are considered and analyzed, and the number of alternative topologies to examine increases drastically with increasing number of taxa, tree search is computationally demanding and long processing times are required. For example, 2,027,025 alternative trees should be examined in case of a sample set of only ten taxa. Therefore, heuristic search methods such as NNI and TBR can be applied. Because the Maximum Likelihood method maximizes the probability of an observed nucleotide/amino acid based on a substitution model, the accuracy of the method is strongly dependent on the selection of the most appropriate model.

An important shortcoming of the Maximum Likelihood method is that it does not involve a parameter for tree topology. In fact, the function does not actually estimate a topology by maximizing the likelihood (Nei, 1987; Yang *et al.*, 1995; Nei, 1996). Instead, a Maximum Likelihood tree is selected under the assumption that the topology with the highest Maximum Likelihood value is most likely to reflect the true phylogeny. It is important to note that the probability of choosing a Maximum Likelihood tree with incorrect topology is high when the rate of substitution varies significantly among lineages. In addition, the tree search may fail to find the best tree by getting stuck in a local optimum of likelihoods (Bromham, 2008).

4. Bayesian methods

Bayesian inference of phylogeny was first introduced in the 1990s (Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Mau *et al*., 1999). Bayesian inference is similar to the Maximum Likelihood method in that both methods examine and calculate the likelihood of possible trees. However, unlike Maximum Likelihood, Bayesian methods sample trees from the tree space and do not calculate the likelihood of all possible branch lengths per tree (Bromham, 2008). In addition, while Maximum Likelihood maximizes the probability of observing the data given the tree and the substitution model, Bayesian analysis maximizes the probability of the tree, given the data and the model (Hall, 2008). Bayesian inference is based on posterior (conditional) probabilities, which means that, prior information (prior probability) about the data is used to estimate the probabilities of the examined trees. Markov chain Monte Carlo (MCMC) methods are used for the approximation of the posterior probabilities of trees (Huelsenbeck and Ronquist, 2001). MCMC starts with a random tree (or a tree specified by the user) and proceeds by generations that involve modifications (moving a branch and/or changing a branch length) and posterior probability ratio calculations to decide on accepting or rejecting a tree. The chain should eventually converge on a stable likelihood value, where accepting or rejecting a tree becomes a random choice, implying that the best (equally likely) Bayesian trees are established (Bromham, 2008; Hall, 2008). An advantage of Bayesian inference is that, MCMC spends more time on the best trees, as the probability that the chain moves from a tree with a high posterior probability is not so likely (Bromham, 2008).

While the ability to incorporate prior knowledge into a tree search algorithm is considered an advantage, it also stands as a drawback. Posterior probability which is used in Bayesian statistics implies accepting a hypothesis without taking the data into account. The fact that prior information affects the outcome of the analysis makes it essential to avoid using inaccurate priors. In phylogenetic applications of Bayesian inference, lack of prior knowledge or lack of certainty about this knowledge is solved by using uninformative priors that do not affect the outcome of the analysis. Similar to Maximum Likelihood, MCMC can get stuck in a local optimum of the tree space (Bromham, 2008), thereby failing to find the best tree.

D. A Case Study Using the *rbcL* Gene

To illustrate the similarities and differences among the methods described herein, nucleotide sequence data from the plastid *rbcL* gene were analyzed in eight angiosperm species and one gymnosperm. The angiosperm taxa included four monocot species in the Poaceae: *Avena sativa* (oat), *Hordeum vulgare* (barley), *Oryza sativa* (rice) and *Triticum aestivum* (bread wheat) and four dicot species in the Solanaceae: *Nicotiana tabacum* (tobacco), *Solanum lycopersicum* (tomato), *Solanum melongena* (eggplant) and *Solanum tuberosum* (potato). The gymnosperm taxon was *Pinus ponderosa* (Ponderosa pine) and was used as an outgroup. The sequences were retrieved from the GenBank database. Unless otherwise metioned, the MEGA computer program version 6.0 (Tamura *et al*., 2013) was used for all analyses. The sequences were aligned with ClustalW method and all positions containing gaps and missing data were eliminated. Overall mean distance among the sequences was 0.102, indicating that the data were suitable for reconstruction of phylogenetic trees using the methods described in this review.

The data were first analyzed with three distance methods: UPGMA, Minimum Evolution and Neighbor Joining with p-distance used for each. The UPGMA reconstruction (Figure 2A) clustered the monocots and dicots separately, as expected. However, within the solanaceous species, *S. lycopersicum* was most closely related to *S. melongena*. This is contrary to what is known about Solanum evolutionary relationships. As mentioned previously, UPGMA is a very simplistic algorithm and not a reliable phylogenetic method. Therefore, it is not surprising that it failed to reconstruct the correct topology with the *rbcL* sequence data. The Minimum Evolution (Figure 2B) and Neighbor Joining (not shown) trees had identical topologies which agreed with angiosperm evolution. These two trees had identical branch lengths and correctly placed the *Pinus* species as outgroup.

Maximum Parsimony was performed with the SPR search option. This method returned three equally parsimonious trees with the outgroup (Ponderosa pine) in the monocot cluster. When the 70% consensus tree was rooted on the outgroup, the monocot species showed the expected relationships; however, the method failed to resolve the solanacous species (Figure 2C). This highlights a shortcoming of the Maximum Parsimony method which uses only parsimony informative sites and therefore, utilizes a smaller dataset than the other methods. In this case study, the number of resulting nucleotides for analyses after gap and missing data removal was 524, whereas the number of parsimony informative sites (sites with at least two types of nucleotides that are represented at least twice) was 80. The inability of the method to separate the Solanum species may also be due to the fact that the *rbcL* gene is highly conserved and has not undergone sufficient divergence to be useful in Maximum Parsimony analysis.

Maximum Likelihood analysis was performed with the NNI tree search option. Once the root was specified, this method resulted in the same topology as Minimum Evolution and Neighbor Joining (Figure 2B). Bayesian analysis was performed with the MrBayes plugin of the Geneious computer program version 8.1. (Kearse *et al*., 2012) by pre-defining the root and applying the MCMC settings of 500,000 chain length, 100 subsampling frequency and 100,000 burn-in length. The most probable tree reconstructed by this method was identical to the Minimum Evolution, Neighbor Joining and Maximum Likelihood trees (Figure 2B). Thus, four of the six methods

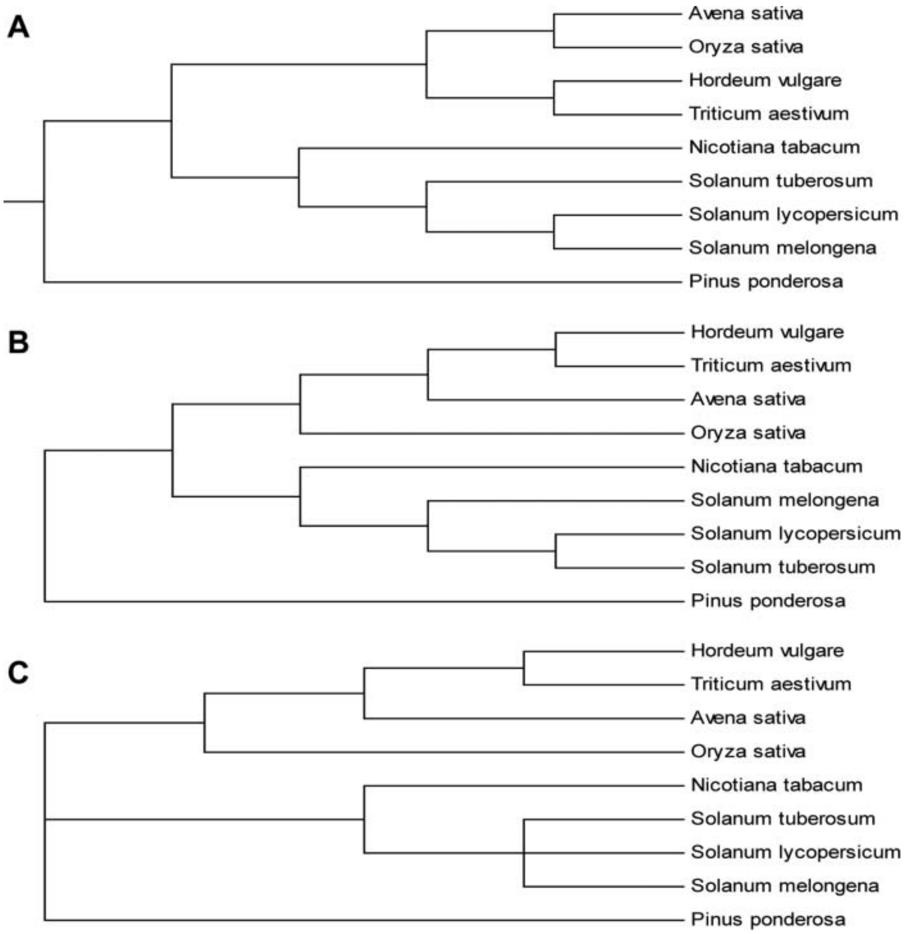

FİG. 2. Alternative topologies obtained by analyzing a dataset of nine plant *rbcL* nucleotide sequences with different tree reconstruction methods. A, Phylogenetic tree reconstructed with UPGMA method; B, Phylogenetic tree reconstructed with Minimum Evolution/Neighbor Joining/Maximum Likelihood/Bayesian methods; C, Maximum Parsimony majority rule consensus tree.

used in this comparison gave the expected evolutionary relationships among the studied taxa. Because every method may not reconstruct the same tree, it is common practice to analyze a given dataset with several methods. Discrepancies among topologies also highlight the usefulness of having independent information about the relationships among the taxa under study.

III. PLANT SEQUENCES FOR PHYLOGENETIC ANALYSES

Over the last 20 years, developments in molecular phylogenetics have led to substantial progress in understanding evolutionary relationships among plants (Wang *et al.*, 2014). Advances in DNA sequencing techniques have made it feasible to obtain whole genome or transcriptome sequences, enabling fuller comparisons of the potential of plant genomes for phylogenetic analysis (Zimmer and Wen, 2013). A genomic region should meet certain criteria to be used for reconstructing phylogenies. A target region should be standardized to enable comparisons among a diverse range of taxonomic groups. It should provide sufficient phylogenetic information and, at the same time, flanking, conserved sequences should be present to allow the design of universal primers that robustly amplify the target from diverse taxonomic units (Taberlet *et al.*, 2007). The target should preferably be from a single copy region to avoid problems due to paralogy (Chase *et al.*, 2005). While working with herbarium samples or fossil remains, high copy number sequences (e.g. sequences from the organellar genomes) have the advantage of improved target amplification from degraded DNA samples (Taberlet *et al.*, 2007).

A. Plastid Sequences

Despite differences in opinion regarding sequence choice, to date, single copy regions of the plastid genome are the most extensively used targets for reconstructing plant phylogenies. The chloroplast genome has many desirable attributes for such analyses including its relatively small size, conserved gene content and order, and high copy number in green plant cells (Chase *et al*., 2007; Davis *et al*., 2014). The small size and conserved gene order of the genome have enabled sequencing of 360 plant plastid genomes (Ruhfel *et al*., 2014). The high copy number of the plastid genome in plant cells enables easy recovery of DNA of sufficient quality and quantity for PCR and sequencing. Conserved gene content and order allow the design of standardized assays to amplify and sequence homologs of the target from a diverse range of taxa. When selecting a target for phylogenetic analysis, it is crucial that the sequence bears a sufficient number of phylogenetically informative sites for comparison without losing the ability to perform an accurate alignment. The search for chloroplast sequences that are optimal for phylogenetic inference started in the 1980s. In early work, the *rbcL* gene, which encodes the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) was used (Taberlet *et al.* 1991). However, due to a very high level of sequence conservation, comparisons based on *rbcL* did not always successfully resolve relationships among closely related taxa. For example, in the Triticeae, *rbcL* could not resolve relationships among the genera *Hordeum*, *Triticum* and *Aegilops* (Doebley *et al*., 1990; Gaut *et al*., 1992).

In the early 1990s, researchers focused on single copy, noncoding regions of the chloroplast genome, mainly introns and intergenic spacers in order to identify sequences with higher rates of evolution (Taberlet *et al*., 1991; Gielly and Taberlet, 1994). Since then, the phylogenetic potential of other chloroplast sequences continues to be explored in order to establish standardized protocols for analyses. For example, Shaw *et al.* (2007) compared single copy chloroplast sequences within Solanaceae (*Atropa* vs. *Nicotiana*), Fabaceae (*Lotus* vs. *Medicago*) and Poaceae (*Saccharum* vs. *Oryza*) in order to determine the most variable regions. Among the 13 variable regions detected, nine non-coding regions (*rpl32-trnL*$^{\text{(UAG)}}$, *trnQ*$^{\text{(UUG)}}$*-5'rps16*, *3'trnV*$^{\text{(UAC)}}$*-ndhC*, *ndhF-rpl32*, *psbD-trnT*$^{\text{(GGU)}}$, *psbJ-petA*, *30rps16–50trnK*$^{\text{(UUU)}}$, *atpI-atpH*, and *petL-psbE*) were selected as the most informative sequences for angiosperm phylogenetic inference at low taxonomic levels. The authors suggested that this set of nine markers be tested to determine the region/regions that meet the requirements to resolve the phylogeny of any given taxa.

Chase *et al.* (2007) proposed two alternative sets of sequences as a standard set for plant phylogenetic analysis. The first set includes the intron maturase gene *matK* and two plastid RNA polymerase genes, *rpoCl* and *rpoB*; the second set includes the *rpoCl* and *matK* genes, and the *psbA-trnH* intergenic spacer. The performance of the plastid genes *matK, rpoC1*, *rpoB* and *rbcL*, and the intergenic spacers *trnH–psbA*, *atpF–atpH* and *psbK–psbI* were assessed for universality, sequence quality and species discrimination (CBOL Plant Working Group, 2009). When *matK* and *rbcL* genes were used in combination, 72% of 550 species were discriminated and 100% were successfully assigned to their co-generic groups. As a result of this work, the authors suggested that the two genes in combination have potential for use as a universal DNA barcode for land plants. In similar work, Dong *et al.* (2012) examined whole chloroplast sequences of 12 genera (*Acorus*, *Aethionema*, *Calycanthus*, *Chimonanthus*, *Eucalyptus*, *Gossypium*, *Nicotiana*, *Oenothera*, *Oryza*, *Paeonia*, *Populus* and *Solanum*) and selected variable regions that were present in at least three of the genera. Among the 23 variable loci, four were coding regions, two were introns and 17 were intergenic spacers. *ycfI*, a coding region of unknown function, was the most variable region, however, the degree of sequence variability prevented the design of universal primers. As evidenced by the relative numbers of variable sequence types (17 of the 23 variable regions are intergenic spacers), intergenic regions of the chloroplast genome seem to harbor many promising loci for phylogenetic analysis.

Advances in next generation sequencing techniques have resulted in the accumulation of whole plastid genome sequences which are readily accessible via the GenBank database (Moore *et al*., 2007; Cronn *et al*., 2008; Parks *et al*., 2009; Cronn *et al*., 2012; Straub *et al*., 2012; Huang *et al*., 2014; Ruhfel *et al*., 2014). By taking advantage of complete plastid genome sequences, it is now feasible to compare many regions of the plastid genome in phylogenetic analysis. Following such a strategy, Jansen *et al.* (2006) sequenced the *Vitis vinifera* plastid genome and retrieved plastid genome sequences of 27 angiosperms from the GenBank database. Their analysis using 61 protein coding genes identified Vitaceae as the earliest diverging lineage of rosids. Moore *et al.* (2010) studied the origin and evolutionary relationships among the major lineages of the Pentapetalae clade by comparative analysis of 83 plastid genes of 86 seed plant species. Nikiforova *et al.* (2013) compared the chloroplast genomes of cultivated apple cultivars (*Malus domestica*) and wild Malus species. The results of their work provided valuable insight into the history of apple domestication. Ruhfel *et al.* (2014) assembled and compared protein coding sequences of 78 plastid genes from 360 species including angiosperms; gymnosperms; monilophytes; lycophytes; liverworts; hornworts; mosses; and paraphyletic, streptophytic and chlorophytic algae, in order to resolve evolutionary relationships of green plants. While trees reconstructed by different substitution models and sampling strategies were consistent in most nodes, the inconsistent portions of the tree across analyses highlighted the requirement for additional molecular data, such as data from nuclear targets, to resolve the divergence of certain lineages.

B. Mitochondrial Sequences

In animal phylogenetics, the mitochondrial *coxI* gene, encoding cytochrome oxidase subunit I, is a widely accepted standard target (Chase *et al*., 2005; Hollingsworth *et al*., 2011). However, the mitochondrial genome seems to be not as well-suited for plant phylogenetic analyses. Due to a high frequency of rearrangements, gene order and content of the plant mitochondrial genome is poorly conserved across plants (Duff and Nickrent, 1999; Knoop, 2004; Knoop *et al.*, 2011; Grewe *et al.*, 2014). In addition, horizontal gene transfer among plant mitochondrial genomes is more common compared to plastid and nuclear genomes (Bergthorsson, 2003; Sanchez-Puerta et al., 2008; Sanchez-Puerta *et al.*, 2011; Xi *et al.*, 2013). Horizontal transfer of the *coxI* group I intron among angiosperms, accompanied by coconversion of the flanking exons, is a well-established example of this phenomenon in the plant mitochondrial genome (Sanchez-Puerta *et al.*, 2008; Sanchez-Puerta *et al.*, 2011). Moreover, the rate of sequence evolution in plant mitochondrial genome is significantly low (Wolfe *et al.*, 1987; Drouin *et al.*, 2008; Galtier, 2011; Davis *et al.*, 2014). Calculations based on synonymous substitution rates showed that sequence evolution of the mitochondrial genome is three times slower than that of the chloroplast genome and ten times slower than that of the nuclear genome (Drouin *et al.*, 2008). Due to the above-listed attributes, the mitochondrial genome is less favored compared to the plastid and nuclear genomes for plant phylogenetic analyses. In addition to the problems related to low sequence divergence and high rates of rearrangements and horizontal gene transfer, an important consideration while working with mitochondrial sequences should be the high frequency of RNA editing sites in the mitochondrial genes. While editing also occurs in nuclear and plastid genomes, it is more prominent in the mitochondrial genome (Knoop, 2011). For example, 200 to 500 cytidine-to-uridine RNA editing sites exist in the angiosperm mitochondrial genome vs 30 to 50 such sites in its chloroplastic counterpart (Oldenkott *et al.*, 2014). In addition, RNA editing in mitochondria is not limited to protein coding sequences, but is also pronounced for tRNAs, introns, and 5' and 3' untranslated sequences (Malek et al., 1996; Grewe *et al.*, 2014). Hence, while working with mitochondrial sequences, comparisons at the cDNA or peptide level are likely to produce more accurate results. Nevertheless, mitochondrial genes including cytochrome oxidase subunits (Hiesel *et al.*, 1994; Malek *et al.*, 1996; Parkinson *et al.*, 1999; Sanchez-Puerta *et al.*, 2008; Sanchez-Puerta *et al.*, 2011; Zeng *et al.*, 2010; Liao *et al.*, 2013; Sha *et al.*, 2014), NADH dehydrogenase subunits (Sanjur *et al.*, 2002), *atpA* (alpha subunit of mitochondrial ATP synthase) (Barkman *et al.*, 2000; Seberg *et al.*, 2012), *matR* (intron encoded maturase R) (Barkman *et al.*, 2000), small subunit (19S) ribosomal DNA (Duff and Nickrent, 1999) and *rps* genes that encode ribosomal proteins (Bergthorsson, 2003), have been used for resolving plant phylogenies. More recently, researchers are reporting the use of large sets of plant mitochondrial genes in evolutionary studies (Xi *et al.*, 2013; Grewe *et al.*, 2014; Liu et al., 2014).

C. Nuclear Sequences

When employing nuclear genes in phylogenetic analyses, it is necessary to have knowledge about hybridization, introgression and polyploidization events (Duarte *et al.*, 2011). Moreover, complex patterns of orthology and paralogy, resulting from high rates of gene duplication and deletion should be taken into consideration. Nuclear ribosomal DNA, especially the internal transcribed spacer (ITS) region of the 18S–5.8S–26S nuclear ribosomal cistron, is a widely used target in evolutionary studies (Poczai and Hyvonen, 2010). However, using ribosomal DNA (rDNA) for plant phylogenetic inference has certain drawbacks and its appropriateness for phylogenetic analysis is questionable. For example, while using rDNA sequences such as ITS, distinguishing orthologs from paralogs is problematic. Ribosomal genes are found as tandem arrays in the nuclear genome and a typical plant genome harbors thousands of such arrays. For example, the Arabidopsis genome has 1400 ribosomal RNA coding genes located as arrays on different chromosomes (Poczai and Hyvonen, 2010). Multiple arrays of rDNA are introduced into plant genomes during evolution by hybridization, polyploidization, gene/chromosome segmental duplication and recombination events (Alvarez and Wendel, 2003). While concerted evolution acts to homogenize multiple ribosomal gene copies among and within the arrays (Buckler *et al.*, 1997), sequence divergence always occurs and orthology is not fully maintained. Therefore, rDNA sequences isolated for comparison are a mixture of paralogs and orthologs which makes their use in phylogenetic inference error-prone (Alvarez and Wendel, 2003; Poczai and Hyvonen, 2010), as accurate phylogenetic reconstructions require comparison of orthologs, not paralogs. The use of rDNA in phylogenetics is also problematic due to the technical aspects of amplifying multigene families, including, possible isolation of different sequence variants from the same sample depending on amplification conditions and the problem of obtaining clean, reliable sequences from a mixture of divergent copies (Hollingsworth *et al.*, 2011).

In order to circumvent the problem of amplifying paralogs or non-functional pseudogenes while targeting nuclear genes from multigene families, single or low copy nuclear genes have been identified for use in evolutionary studies (Zimmer and Wen, 2013). *Alcohol dehydrogenase* (Fukuda *et al.* 2005), *β-amylase* (Rajapakse *et al.* 2004), *Chalcone synthase* (Inda *et al.*, 2010), *Cycloidea* (Marten Rodriguez *et al.*, 2010), *Granule-bound starch synthase I* (Mason Gamer, 2008), *Chloroplast-expressed glutamine synthetase* (Clarkson *et al.*, 2010), *LEAFY* (Kim *et al.*, 2010), and *DNA-directed RNA polymerase II subunit B* (Sun *et al.*, 2010) are examples of low copy nuclear genes that have been identified as potentially useful targets for plant phylogenetic studies.

Thanks to the reduced costs, improved speed and massive data output of DNA sequencing, complete or almost complete sequences of nuclear genomes and transcriptomes are accumulating in databases. Thus, it is becoming possible to investigate near-entire genomes of organisms to find single copy nuclear targets or to compare thousands of loci at a time for phylogenetic inference (Zimmer and Wen, 2013). In addition to enabling the comparison of a large number of loci, retrieving sequences from public databases greatly reduces the cost of such analyses. Duarte *et al.* (2011) performed a comparative analysis of the complete genome sequences of *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera* and *Oryza sativa* in order to detect shared, single copy sequences. Their analysis identified 959 potentially useful, single copy genes. EST (Expressed Sequence Tag) sequences from 69 plant species were retrieved from public databases for 13 of the identified loci. The phylogeny reconstructed using these single copy, nuclear genes in combination, was largely concordant with studies performed using single (Hilu *et al.*, 2003) or multiple (Jansen *et al.*, 2007) plastid genes and work based on the combined use of plastid and nuclear ribosomal DNA targets (Soltis *et al.*, 2000). Using publicly available EST sequences, Burleigh *et al.* (2011) constructed a total of 18,896 gene trees with a variable number of taxa (a minimum of three taxa) represented by each tree. By employing a gene tree parsimony approach that utilized the topology data of the 18,896 trees, a consensus phylogeny was reconstructed that represented all of the 136 species sampled in the gene trees. While the indirect use of sequence alignment data yielded results that were consistent with studies based on direct sequence alignments (Soltis *et al.*, 2000; Hilu *et al.*, 2003; Jansen *et al.*, 2007), the authors indicated that, data from at least 1000 genes were required to obtain sufficient statistical (bootstrap) support for their analysis. In another recent study, the complete genome sequences of *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Vitis vinifera* and *Physcomitrella patens*, along with EST sequences from 145 other plant species were used to determine a set of orthologous sequences (22,833 orthologs) for phylogenetic analysis (Lee *et al.*, 2011). A functional phylogenomics approach was employed to identify candidate genes associated with plant diversification and adaptation. This study produced interesting results such as the significant representation of RNA interference mechanism genes in the pool of candidates for angiosperm and gymnosperm divergence. Moreover, genes associated with salt/drought tolerance and oxygen radical detoxification were also found to be relevant in plant diversification.

IV. CONCLUSIONS

While comparative morphological and physiological analyses have long been used for phylogenetic inference, advances in molecular biology enable comparison at the molecular level. The use of molecular data for phylogenetic inference compensates for the various shortcomings of morphology based approaches. There are several statistical models designed to explain changes in protein and DNA sequences. Correct model choice and the appropriate phylogeny reconstruction method are expected to lead to the most accurate phylogenetic reconstructions. However, it is important to remember that selection of the most appropriate model and reconstruction method should be guided by preliminary knowledge of the evolutionary relationships among taxa based on morphology and physiology. Thus, combined approaches that initially employ morphological and physiological comparisons prior to molecular phylogenetic analysis are more likely to produce the most accurate phylogenies. Molecular data have been extensively used in plant phylogenetics over the last two decades and plant molecular geneticists have identified several regions in nuclear and plastid genomes that enable reconstruction of consistent phylogenies across different studies. As a result of the advances in high throughput DNA sequencing technologies, comparative analysis of entire genomes has become feasible and it is rational to anticipate that genome wide comparisons will become a routine in plant phylogenetics.

ACKNOWLEDGMENT

The authors are grateful to TUBITAK for financial support (project no. 109O979) and to Amy Frary for critical comments on the manuscript.

REFERENCES

Alvarez, I., and Wendel, J. F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol.* **29**: 417–434.

Barkman, T. J., Chenery, G., McNeal, J. R., Lyons-Weiler, J., Ellisens, W. J., Moore, G., Wolfe, A. D., and dePamphilis, C. W. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *PNAS* **97**: 13166–13171.

Barriere, A., Gordon, K. L., and Ruvinsky, I. 2011. Distinct functional constraints partition sequence conservation in a *cis*-regulatory element. *PLoS Genet.* **7**: e1002095. doi:10.1371/journal.pgen.1002095.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. 2013. GenBank. *Nucleic Acids Res.* **41**: D36–D42.

Bergthorsson, U., Adams, K. L., Thomason, B., and Palmer, J. D. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424**: 197–201.

Blouin, C., Boucher, Y., and Roger, A. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucleic Acids Res.* **31**: 790–797.

Bombarely, A., Menda, N., Tecle, I. Y., Buels, R. M., Strickler, S., Fischer-York, T., Pujar, A., Leto, J., Gosselin, J., and Mueller, L. A. 2011. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.* **39**: D1149–D1155.

Bromham, L. 2008. *Reading the story in DNA*. Oxford University Press Inc., New York.

Bruno, W. J., Socci, N. D., and Halpern, A. L. 2000. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17**: 189–197.

Buckler, E. S., Ippolito, A., and Holtsford, T. P. 1997. The evolution of ribosomal DNA: divergent paralogues and phylogenetic implications. *Genetics* **145**: 821–832.

Burleigh, J. G., Bansal, M. S., Eulenstein, O., Hartmann, S., Wehe, A., and Vision, T. J. 2011. Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees. *Syst. Biol.* **60**: 117–125.

Cavalli-Sforza, L. L., and Edwards, W. F. 1967. Phylogenetic analysis Models and Estimation Procedures. *Am. J. Hum. Genet.* **19**: 233–257.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *PNAS* **106**: 12794–12797.

Chase, M. W., Cowan, S., Hollingsworth, P. M., van den Berg, C., Madrinan, S., Petersen, G., Seberg, O., Jorgsensen, T., Cameron, K. M., Carine, M., Niklas, P., Hedderson, T. A. J., Conrad, F., Salazar, G. A., Richardson, J. E., Hollingsworth, M. L., Barraclough, T. G., Kelly, L., and Wilkinson, M. 2007. A proposal for a standardized protocol to barcode all land plants. *Taxon* **56**: 295–299.

Chase, M. W., Salamin, N., Wilkinson, M., Dunwell, J. M., Kesanakurthi, R. P., Haidar, N., and Savolainen, V. 2005. Land plants and DNA barcodes: short-term and long-term goals. *Philos. T. Roy. Soc. B.* **360**: 1889–1895.

Clarkson, J. J., Kelly, L. J., Leitch, A. R., Knapp, S., and Chase, M. W. 2010. Nuclear glutamine synthetase evolution in *Nicotiana*: phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Mol. Phylogenet. Evol.* **55**: 99–112.

Conte, M. G., Gaillard, S., Lanau, N., Rouard, M., and Perin, C. 2008. GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res.* **36**: D991–D998.

Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., and Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing by synthesis technology. *Nucleic Acids Res.* **36**: 1–11.

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., and Udall, J. 2012. Targeted enrichment strategies for next-generation plant biology. *Am. J. Bot.* **99**: 291–311.

Davis, C. C., Xi, Z., and Mathews, S. 2014. Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *BMC Biol.* **12**: 11.

Dayhoff, M. O. 1972. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Springs, MD.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. 1978. A model of evolutionary change in proteins. **In**: *Atlas of Protein Sequence and Structure*. pp. 345–352. Dayhoff, M. O., Eds., National Biomedical Research Foundation, Silver Springs, MD.

Doebley, J., Durbin, M. L., Golenberg, E. M., Clegg, M. T., and Ma, D. P. 1990. Evolutionary analysis of the large subunit of carboxylase (*rbcL*) nucleotide sequence among the grasses (Graminae). *Evolution* **44**: 1097–1108.

Dong, W., Liu, J., Yu, J., Wang, L., and Zhou, S. 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* **7**: e35071.

Drouin, G., Daoud, H., and Xia, J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* **49**: 827–831.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, J. C., Leebens-Mack, J., and de Pamphilis, C. W. 2011. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**: 61.

Duff, R. J., and Nickrent, D. L., 1999. Phylogenetic relationships of land plants using mitochondrial small-subunit rDNA sequences. *Am. J. Bot.* **86**: 372–386.

Eck, R. V., and Dayhoff, M. O. 1966. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Springs, MD.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**: 51792–1797.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol. Evol.* **17**: 368–376.

Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20**: 406–416.

Fitch, W. M., and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155**: 279–284.

Fukuda, T., Yokoyama, J., Nakamura, T., Song, I. J., Ito, T., Ochiai, T., Kanno, A., Kameya, T., and Maki, M. 2005. Molecular phylogeny and evolution of alcohol dehydrogenase (*Adh*) genes in legumes. *BMC Plant Biol.* **5**: 6.

Futuyma, J. D. 2005. *Evolution*. Sinauer Associates Inc., Sunderland, MA.

Galtier, N. 2011. The intriguing evolutionary dynamics of plant mitochondrial DNA. *BMC Biol.* **9**: 61.

Gaut, B. S., Muse, S. V., Clark, W. D., and Clegg, M. T. 1992. Relative rates of nucleotide substitutions at the *rbcL* locus of monocotyledonous plants. *J. Mol. Evol.* **35**: 292–303.

Gielly, L., and Taberlet, P. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbcL* sequences. *Mol. Biol. Evol.* **11**: 769–777.

Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W. D., and Waugh, M. E. 2005. The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.* **33**: D660–D665.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**: D1178–1186.

Graur, D., and Li, W. H. 2000. *Fundamentals of Molecular Evolution.* 2nd ed., Sinauer Associates Inc., Sunderland, MA.

Grewe, F., Edger, P. P., Keren, I., Sultan, L., Pires, J. C., Ostersetzer-Biran, O., and Mower, J. P. 2014. Comparative analysis of 11 Brassicales mitochondrial genomes and the mitochondrial transcriptome of *Brassica oleracea*. *Mitochondrion* **19**: 135–143.

Grishin, N. V. 1995. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* **41**: 675–679.

Hall, B. G. 2008. *Phylogenetic trees made easy*. 3rd ed., Sinauer Associates Inc., Sunderland, MA.

Hartigan, J. A. 1973. Minimum evolution fits to a given tree. *Biometrics* **29**: 53–65.

Hartl, D. L., and Clark, A. G. 1997. *Principles of Population Genetics*. Sinauer Associates Inc., Sunderland, MA.

Hasegawa, M., Lida, Y., Yano, T., Takaiwa, F., and Iwabuchi, M. 1985. Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. *J. Mol. Evol.* **22**: 32–38.

Hendy, M. D., and Penny, D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* **59**: 277–290.

Hiesel, R., von Haeseler, A., and Brennicke, A. 1994. Plant mitochondrial nucleic acid sequences as a tool for phylogenetic analysis. *PNAS* **91**: 634–638.

Hilu, K. W., Borsch, T., Muller, K., Soltis, D. E., Soltis, P. S., Savolainen, V., Chase, M. W., Powell, M., Alice, L. A., Evans, R., Sauquet, H., Neinhus, C., Slotta, T. A. B., Rohwer, J. G., Campbell, C. S., and Chatrou, L. W. 2003. Angiosperm phylogeny based on *matK* sequence information. *Am. J. Bot.* **90**: 1758–1776.

Hollingsworth, P. M., Graham, S. W., and Little, D. P. 2011. Choosing and using a plant DNA barcode. *PLoS ONE* **6**: e19254.

Huang, H., Shi, C., Liu, Y., Mao, S. Y., and Gao, L. Z. 2014. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol. Biol.* **14**: 151.

Huelsenbeck, J. P., and Ronquist, F. 2011. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Inda, L. A., Pimentel, M., and Chase, M. W. 2010. Chalcone synthase variation and phylogenetic relationships in *Dactylorhiza* (Orchidaceae). *Bot. J. Linn. Soc.* **163**: 155–165.

Ingram, V. M. 1963. *The Hemoglobins in Genetics and Evolution*. Columbia University Press, New York.

Jansen, R. K., Kaittanis, C., Saski, C., Lee, S. B., Tomkins, J., Alverson, A. J., and Daniell, H. 2006. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol. Biol.* **6**: 32.

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., dePamphilis, C. W., Leebens-Mack, J., Muller, K. F., Guisinger-Bellian, M., Haberle, R. C., Hansen, A. K., Chumley, T. W., Lee, S. B., Peery, R., McNeal, J. R., Kuehl, J. V., and Boore, J. L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *PNAS* **104**: 19369–19374.

Jukes, T. H., and Cantor, C. R. 1969. Evolution of protein molecules. In: *Mammalian Protein Metabolism.* pp. 21–132. Munro, H. N., Eds., Academic Press, New York.

Jung, S., Staton, M., Lee, T., Blenda, A., Svancara, R., Abbott, A., and Main, D. 2008. GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.* **36**: D1034–D1040.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., and Drummond, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Kim, C., Shin, H., Chang, Y. T., and Choi, H. K. 2010. Speciation pathway of *Isoetes* (Isoetaceae) in East Asia inferred from molecular phylogenetic relationships. *Am. J. Bot.* **97**: 958–969.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111–120.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Kishino, H., Miyata, T., and Hasegawa, M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31**: 151–160.

Knoop, V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr. Genet.* **46**: 123–139.

Knoop, V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol. Life Sci.* **68**: 567–586.

Knoop, V., Volkmar, U., Hecht, J., and Grewe, F. 2011. Mitochondrial genome evolution in the plant lineage. **In:** *Plant Mitochondria.* pp. 3–29. Kempken, F., Eds., Springer, New York.

Kumar, S., Tamura, K., and Nei, M. 1993. *Manual for MEGA: Molecular Evolutionary Genetics Analysis software.* Pennsylvania State University, University Park, PA.

Lee, E. K., Cibrian-Jaramillo, A., Kolokotronis, S. O., Katari, M. S., Stamatakis, A., Ott, M., Chiu, J. C., Little, D. P., Wm. Stevenson, D., McCombie, R., Martienssen, R. A., Coruzzi, G., and DeSalle, R. 2011. A functional phylogenomic view of the seed plants. *PLoS Genet*. **7**: e1002411.

Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E. S., Casstevens, T., Hurwitz, B., McCouch, S., Ni, J., Pujar, A., Ravenscroft, D., Ren, L., Spooner, W., Tecle, I., Thomason, J., Tung, C. W., Wei, X., Yap, I., Youens-Clark, K., Ware, D., and Stein, L. 2008. Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.* **36**: D947–D953.

Liao, J. Q., Ross, L., Fan, X., Sha, L. N., Kang, H. Y., Zhang, H. Q., Wang, Y., Liu, J., Wang, X. L., Yu, X. F., Yang, R. W., Ding, C. B., Zhang, L., and Zhou, Y. H. 2013. Phylogeny and maternal donors of the tetraploid species with St genome (Poaceae: Triticeae) inferred from *Cox*II and ITS sequences. *Biochem. Syst. Ecol.* **50**: 277–285.

Liu, Y., Cox, C. J., Wang, W., and Goffinet, B. 2014. Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Syst. Biol.* **63**: 862–878.

Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Biol.* **40**: 315–328.

Malek, O., Lattig, K., Hiesel, R., Brennicke, A. and Knoop, V. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. *The EMBO Journal* **15**: 1403–1411.

Margoliash, E. 1963. Primary structure and evolution cytochrome c. *PNAS* **50**: 672–679.

Marten-Rodriguez, S., Fenster, C. B., Agnarsson, I., Skog, L. E., and Zimmer, E. A. 2010. Evolutionary breakdown of pollination specialization in a Caribbean plant radiation. *New Phytol.* **188**: 403–417.

Mason Gamer, R. J. 2008. Allohexaploidy, introgression, and the complex phylogenetic history of *Elymus repens* (Poaceae). *Mol. Phylogenet. Evol.* **47**: 598–611.

Mau, B., and Newton, M. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* **6**: 122–131.

Mau, B., Newton, M., and Larget, B. 1999. Bayesian phylogenetic inference via Markov chain Monte carlo methods. *Biometrics* **55**: 1–12.

Messeguer, R., Ganal, M. W., Steffens, J. C., and Tanksley, S. D. 1991. Characterization of the level, target sites and inheritance of cytosine methylation in tomato nuclear DNA. *Plant Mol. Biol.* **16**: 753–770.

Moore, M. J., Bell, C. D., Soltis, P. S., and Soltis, D.E. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *PNAS* **104**: 19363–19368.

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., and Soltis, D. E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *PNAS* **107**: 4623–4628.

Nei, M. 1987. *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* **30**: 371–403.

Nei, M., and Kumar, S. 2000. *Molecular Evolution and Phylogenetics.* Oxford University Press Inc., New York.

Nikiforova, S. V., Cavalieri, D., Velasco, R., and Goremykin, V. 2013. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Mol. Biol. Evol.* **30**: 1751–1760.

Ohno, S. 1970. *Evolution by Gene Duplication.* Springer-Verlag, Berlin.

Oldenkott, B., Yamaguchi, K., Tsukinoki, S. T., Knie, N., and Knoop, V. 2014. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. *RNA* doi: 10.1261/rna.045575.114.

Ota, T., and Nei, M. 1994. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* **38**: 542–643.

Pardi, F., and Gascuel, O. 2012. Combinatorics of distance-based tree inference. *PNAS* **109**: 16443–16448.

Parkinson, C. L., Adams, K. L., and Palmer, J. D. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr. Biol.* **9**: 1485–1488.

Parks, M., Cronn, R., and Liston, A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* **7**: 84.

Poczai, P., and, Hyvonen, J. 2010. Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. *Mol. Biol. Rep.* **37**: 1897–1912.

Popescu, A. A., Huber, K. T., and Paradis, E. 2012. Ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**: 1536–1537.

Proost, S., van Bel, M., Sterck, L., Billiau, K., van Parys, T., van de Peer, Y., and Vandepoele, K. 2009. PLAZA: A comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell* **21**: 3718–3731.

Qiu, Y. L., Lee, J., Bernasconi-Quadroni, F., Soltis, D. E., Soltis, P. S., Zanis, M., Zimmer, E. A., Chen, Z., Savolainen, V., and Chase, M. W. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404–407.

Rajapakse, S., Nilmalgoda, S., Molnar, M., Ballard, R., Austin, D., and Bohac, J. 2004. Phylogenetic relationships of the sweet potato in *Ipomoea* series *Batatas* (Convolvulaceae) based on nuclear beta-amylase gene sequences. *Mol. Phylogenet. Evol.* **30**: 623–632.

Rannala, B., and Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**: 304–311.

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., and Burleigh, J. G. 2014. From algae to angiosperms–inferring the phylogeny of green plants (*Viridiplantae*) from 360 plastid genomes. *BMC Evol. Biol.* **14**: 23.

Rzhetsky, A., and Nei, M. 1992. A simple method for estimating and testing minimum evolution trees. *Mol. Biol. Evol.* **9**: 945–967.

Rzhetsky, A., and Nei, M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* **10**: 1073–1095.

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

Sanchez-Puerta, M. V., Abbona, C. C., Zhuo, S., Tepe, E. J., Bohs, L., Olmstead, R. G., and Palmer, J. D. 2011. Multiple recent horizontal transfers of the *cox1* intron in Solanaceae and extended co-conversion of flanking exons. *BMC Evol. Biol.* **11**: 277.

Sanchez-Puerta, M. V., Cho, Y., Mower, J. P., Alverson, A. J., and Palmer, J. D. 2008. Frequent, phylogenetically local horizontal transfer of the *cox1* group I intron in flowering plant mitochondria. *Mol. Biol. Evol.* **25**: 1762–1777.

Sanger, F., Nicklen, S., and Coulson, A. R. 1977. DNA Sequencing with chain-terminating inhibitors. *PNAS* **74**: 5463–5467.

Sanjur, O. I., Piperno, D. R., Andres, T. C., and Wessel-Beaver, L. 2002. Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: Implications for a crop plant evolution and areas of origin. *PNAS* **99**: 535–540.

Seberg, O., Petersen, G., Davis, J. I., Pires, J. C., Stevenson, D. W., Chase, M. W., Fay, M. F., Devey, D. S., Jorgensen, T., Sytsma, K. J., and Pillon, Y. 2012. Phylogeny of the Asparagales based on three plastid and two mitochondrial genes. *Am. J. Bot.* **99**:875–889.

Sha, L. N., Fan, X., Zhang, H. Q., Kang, H. Y., Wang, Y., Wang, X. L., Zhang, L., Ding, C. B., Yang, R. W., and Zhou, Y. H. 2014. Phylogenetic relationships in *Leymus* (Triticeae; Poaceae): Evidence from chloroplast *trn-H-psbA* and mitochondrial *coxII* intron sequences. *J. Syst. Evol.* **52**: 722–734.

Shaw, J., Lickey, E. B., Schilling, E. E., and Small, R. L. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *Am. J. Bot.* **94**: 275–288.

Sneath, P. H. A., and Sokal, R. R. 1973. *Numerical Taxonomy*. Freeman Publishers, San Francisco, CA.

Sokal, R. R., and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28**: 1409–1438.

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C., Zanis, M., Savolainen, V., Hahn, W. H., Hoot, S. B., Fay, M. F., Axtell, M., Swensen, S. M., Prince, L. M., Kress, W. J., Nixon, K. C., and Farris, J. S. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* **133**: 381–461.

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* **99**: 349–364.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**: D1009–D1014.

Swofford, D. L. and Begle, D. P. 1993. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1. User's Manual.* Illinois Natural History Survey, Champaign, IL.

Sun, G., Pourkheirandish, M., and Komatsuda, T. 2010. Molecular evolution and phylogeny of the *rpb2* gene in the genus *Hordeum*. *Ann. Bot.* **103**: 975–983.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C., and Willerslev, E. 2007. Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **35**: e14.

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. 1991. Universal primers for Amplification of three non-coding regions of chloroplast DNA. *Plant. Mol. Biol.* **17**: 1105–1109.

Tamura, K. 1992. The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. *Mol. Biol. Evol.* **9**: 814–825.

Tamura, K., and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**: 512–526.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6. 0. *Mol. Biol. Evol.* **30**: 2725–2729.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Res*. **22**: 4673–4680.

Wang, W., Li, H. L., and Chen, Z. D. 2014. Analysis of plastid and nuclear DNA data in plant phylogenetics-evaluation and improvement. *Sci. China Life Sci*. **57**: 280–286.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., and Oosick, R. 2008. *Molecular Biology of the Gene*. 6th ed., Pearson/Benjamin Cummings, San Francisco, CA.

Wolfe, K. H., Li, W. H., and Sharp, P. M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *PNAS* **84**: 9054–9058.

Xi, Z., Wang, Y., Bradley, R. K., Sugumaran, M., Marx, C. J., Rest, J. S., and Davis, C. C. 2013. Massive mitochondrial gene transfer in a parasitic flowering plant clade. *PLoS Genet*. **9**: e1003265.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *TREE* **11**: 367–372.

Yang, Z., Goldman, N., and Friday, A. E. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistic estimation problem. *Syst. Biol.* **44**: 384–399.

Yang, Z., and Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte carlo method. *Mol. Biol. Evol.* **14**: 717–724.

Yang, Z., and Rannala, B. 2012. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**: 303–314.

Zeng, J., Fan, X., Zhang, L., Wang, X., Zhang, H., Kang, H., and Zhou, Y. 2010. Molecular phylogeny and maternal progenitor implication in the genus *Kengyilia* (Triticeae: Poaceae): Evidence from COXII intron sequences. *Biochem. Syst. Ecol.* **38**: 202–209.

Zimmer, E. A., and Wen, J. 2013. Reprint of: Using nuclear gene data for plant phylogenetics: Progress and prospects. *Molecular Phylogenet. Evol.* **66**: 539–550.

Zuckerkandl, E., and Pauling, L. B. 1962. Molecular disease, evolution, and genic heterogeneity. **In**: *Horizons in Biochemistry*. pp. 189–225. Kasha, M. and Pullman, B., Eds., Academic Press, New York.

Zuckerkandl, E., and Pauling, L. B. 1965. Evolutionary divergence and convergence in proteins. **In**: *Evolving Genes and Proteins.* pp. 97–166. Bryson, V., and Vogel, H. J., Eds., Academic Press, New York.