Plant phylogenomics: approaches to make phylogenetic inference more reliable

Dissertation

for the obtainment of the academic degree doctor rerum naturalium (Dr. rer. nat.)

presented to the Council of the
Faculty of Biology and Pharmacy
of the Friedrich-Schiller-University Jena

by Svetlana Nikiforova, M. sc.
born on 26 of March 1983 in Saratov, Russia

Jena, March 2013

Referees:

1. ________________________

2. ________________________

3. ________________________

Date of the public defense: ________________________

Abbreviations

AFLP – amplified fragment length polymorphism
AIC – Akaike information criterion
DLS – differential lineage sorting
DNA – deoxyribonucleic acid
cpDNA – chloroplast DNA
mtDNA – mitochondrial DNA
nuDNA – nuclear DNA
GNB – abbreviation of authors names (Goremykin, Nikiforova, Bininda-Emonds)
GTR model – general time reversible
GWD – genome wide duplication
IR – inverted repeat
ITS – internal transcribed spacer
LBA – long branch attraction
LQP – Le Quesne probability
LSC – large single copy region
MD – *Malus domestica*
ML – maximum likelihood
MP – maximum parsimony
MPI – message passing interface
MYA – million years ago
NGS – next-generation sequencing
NJ – neighbor joining
OTU – operational taxonomic unit
OV – observed variability
PCR – polymerase chain reaction
PTP – permutation tail probability
RAPD – random amplification of polymorphic DNA
SSC – small single copy region
SSR – simple sequence repeat
TBR – tree bisection-reconnection
UPGMA – unweighted pair group method with arithmetic mean

Table of contents

1 General introduction

1.1 Limitations and errors in phylogeny reconstruction and ways to overcome them

At the earliest step of plants systematics development, biologists were trying to elucidate evolutionary relationships of taxa based on the morphological characters (Cronquist, 1981; Takhtajan, 1987). Researchers, directly comparing such characters, estimated their importance without explicit analytical framework. By contrast, gene systematics is rooted in comparison of character states in biopolymers, proteins or nucleic acids, wherein homology of characters compared and their states are easier to ascertain and importance of each substitution can be determined statistically (Palmer, 1965; Zuckerkandl and Pauling, 1965). The founders of the new approach justified it based on explicity of assumptions, as compared to intuitive approach to systematics, which was used before. First molecular studies were hoped to finally resolve contradictions between different classification systems, based on the morphological characters. However, new phylogenetic approach could not fulfill these expectations from the very start.

1.1.1 Rise and fall of the marker analyses

Thought the ninetieth and at the beginning of twenty-first century, advance of Sanger sequencing methodology facilitated a large number of phylogenetic studies based on so-called phylogenetic "markers". These markers, usually relatively short gene and intragenic spacer regions were amplified from the total DNA by means of polymerase chain reaction (PCR). In 1993 in the "Annals of the Missouri Botanical Garden" a series of papers were published which were based on the analysis of chloroplast *rcb*L gene for many taxa of flowering plants (see Annals of the Missouri Botanical Garden. 1993. Volume 80). The most prominent and often cited of these studies is the one conducted by Chase with co-authors in which the gene sequences were sampled from over 500 species (Chase et al., 1993). Besides *rbc*L, other coding and non-coding regions were used for phylogenetic reasons: 26S and 18S rDNA nuclear gene (Martin and Dowd, 1991; Hamby and Zimmer, 1988; Soltis et al., 1997; Barkman et al., 2000; Zanis et al., 2002) nuclear ITS (Campbell et al., 1995; 1997; Oh and Potter, 2003; Lo et al., 2007; Feng et al., 2007), chloroplast genes (*ndh*F, *mat*K*, atp*B*, atp*A*, rpl*16*, rpo*C2*, ysf*2*)* (Zanis et al., 2002; Barkman et al., 2000; Qiu et al, 1999, 2000, 2001, 2005, 2006, 2010; Davis et al., 2004; Hoot et al., 1999; Savolainen et al., 2000; Soltis et al., 2000; Hilu et al., 2003; Graham and Olmstead, 2000; Huang et al., 2010), non-coding cpDNA spacers (*tnr*T-*trn*L*, trn*T-*trn*F*, trn*L-*trn*F *atp*B-*rbc*L*, psb*A-*trn*H*)* (Renner et al., 2000; Drabkova et al., 2004; Mes and Thart, 1994; Mes et al., 2000; Applequist and Wallace, 2002; Borsch et al., 2003), *pet*D intron (Löhne and Borsch, 2005) and mitochondrial genes (*atpA, matR, nad5, nad1, nad2, cox1,19S, rps2)* (Qiu et al., 2006; Beckert et al., 2001; Dombrovska and Qiu, 2004; Barkman et al., 2000; Zanis et al., 2002), etc.

The variety of markers led to huge discrepancies among phylogenetic trees inferred from analyses of the individual DNA regions. Striking examples of conflicting results were obtained in studies aimed at establishing phylogenetic relationships of the gymnosperms and angiosperms. Based on the short fragments analyses various authors placed either Gnetales as close relatives to all angiosperms (Crane, 1985) or conifers (Chaw et al., 1997) or Cupressophyta (Nickrent et al., 2000; Doyle, 2006) or Pinaceae (Bowe et al., 2000; Chaw et al., 2000).

In addition, the species composition of the most basal clade within angiosperms could not be agreed upon. Different researches advocated either Ceratophyllaceae (Les et al., 1991; Qiu et al., 1993; Chase et al., 1993; Endress, 1994), Chloranthaceae (Taylor and Hickey, 1992), Nymphaeales (Graham and Olmstead, 2000; Kim and Lee, 2004), *Amborella* (Soltis and Soltis, 2004; Mathews and Donoghue, 1999; Qiu et al., 1999; Stefanović, et al., 2004), or *Amborella*+Nymphaeales (Barkman et al., 2000) as the basal-most clade of angiosperms.

The incongruence in the trees obtained fored scientists to look for alternative approaches and finally has led to abandoning one-marker studies. The suitability of the markers to concrete tasks was normally not analytically tested, which has led to once-popular but mistaken conclusions, like, for example, support of the anthophyte hypothesis (Gnetales sister to angiosperms) by *rbc*L (Chase et al., 1993; Baum, 1994; Manhart, 1994). In comparison between angiosperms and gymnosperms this gene sequence is completely saturated at synonymous sites, but has almost no characters at the non-synonymous sites to meaningfully resolve the phylogeny (Goremykin et al., 1996).

First genome-scale phylogenetic studies based on data sampled from the complete chloroplast genomes (Goremykin et al., 1997, 2003a, b, 2004, 2005; Martin et al., 1998; Turmel et al., 1999; Lemieux et al., 2000; Kugita et al., 2003; Wolf et al., 2005; Bausher et al., 2006; Jansen et al., 2006; Lee et al., 2006; Ravi et al., 2006; Ruhlman et al., 2006) have demonstrated that amassing large number of characters routinely leads to far better resolution of a tree structure, compared to the one usually obtained in marker analyses.

It should be noted that first studies based on the concatenation of large number of genes were criticized by proponents of markers analysis, which advocated the importance of large number of taxa compared to large number of genes (Baldauf et al., 2000; Peterson and Eernisse, 2001; Moncalvo et al, 2002; Sanderson and Driskell, 2003; Soltis et al., 2004, Stefanovic et al., 2004). Yet subsequent development in phylogenetics has demonstrated that certain limitations inherent to marker-based analyses cannot be overcome by increased taxon sampling and render it useless in many situations. Among these limitations is the lack of characters contained in individual genes to support tree branches (Delsuc et al., 2005), stochastic error related to variation in a gene length (Jeffroy et al., 2006) and differential lineage sorting, i.e. incongruence between evolution of individual genes and the evolution of species or other genes (Maddison, 1997). The fact that even opponents of the phylogenomic approach, turned to using genome-scale data (e.g. Moore et al., 2007; Jansen et al., 2007; Soltis et al., 2011) signifies that phylogenomics was recognized by the scientific community as the way to overcome limitations of the marker era. The initial problems, related to small number of complete genomes available have been overcome with introduction of new sequencing techniques, which led to accumulation of big volume of genome-related information in public databases.

1.1.2 Using fingerprinting methods for phylogeny reconstruction on shallow taxonomic level

The issue of incongruence was perhaps even more pronounced in studies of genetic polymorphism aimed at phylogenetic reconstruction at shallow taxonomic levels. Attaining high resolution of the phylogenetic relationships within genus *Malus* has been particularly problematic (Luby, 2003, Li et al., 2012). A number of authors tried to address the issue; however, these studies did not converge on any particular outcome. NJ and UPGMA trees based on the matrix of RAPD fragment distribution from cultivated and wild apple accessions (Zhou and Li, 2000) were incongruent, and no branch received strong support (>80%). In the RAPD-based analyses of 155 cultivated and wild *Malus* accessions (Oraguzie et al., 2001) the results obtained contradicted all previous knowledge to the degree that the authors stated that "the grouping of genotypes based on the phenogram and scatter plot generally did not reflect the pedigree or provenance of the genotypes". Inadequacy of RAPD methodology for analysis of phylogenetic affinities of domesticated apple was attributed (Iketani, 1998) to the complex heterozygous structure of apple genome.

AFLP-based analyses, encompassing most of the wild apple species and a wide range of *Malus domestica* (MD) varieties (Zhang et al., 2007; Ling et al., 2009) recovered very dissimilar tree topologies, with *Malus x domestica* accessions assuming sister positions to a wide array of branches, subtending representatives of a majority of the sections within the genus. Similar outcome was obtained in SSR-based analysis encompassing 142 wild and domesticated *Malus* accessions (Hokanson et al., 2001). The authors noted: "SSR data were not useful in identifying genetic relationships among this diverse collection of accessions, with the majority of the accessions not clustering in ways concordant with taxonomic information and/or geographic origin." The same type of result was observed in recent SSR-based analysis of a broad spectrum of *Malus* accessions (Potts et al., 2012), wherein all 6 major clades within *Malus*, as defined by the authors, contained terminal branches bearing different *M. x domestica* cultivars. High dependence of the outcome of the SSR analysis within the genus on the marker set can be seen in Zhang et al. (2012). In this study, topological changes, affecting most of the tree branches were observed when two sets of the SSR markers were applied to the same *Malus* accessions. Lack of statistical support for the majority of the branches in the Neighbor-Joining analyses of another recently published SSR matrix (Micheletti et al., 2011, Fig 1b) also highlights the need to utilize the data with more robust signal structure.

Great disarray of the results obtained in genetic fingerprinting analyses precludes them from serving as a reference to elucidate phylogenetic relationships within *Malus*. Comparison of these studies tells the reader not as much about evolution of the genus, as of the need to look for better data.

1.1.3 Phylogenomics

The most robust tree-building strategy currently in use, utilizes the wealth of information available in complete genomes (Jeffroy et al., 2006). It is termed phylogenomics. The term refers to a group of diverse methods aimed at elucidating phylogeny based on various ways of assessing and summarizing evolutionary meaningful characters. Figure 1-1 (taken from Dutlih et al., 2007, with modifications) outlines the most important aspects of the four broadly used phylogenomic methods (termed in the figure as 'gene content', 'superalignment', 'superdistance' and 'supertree' approaches).

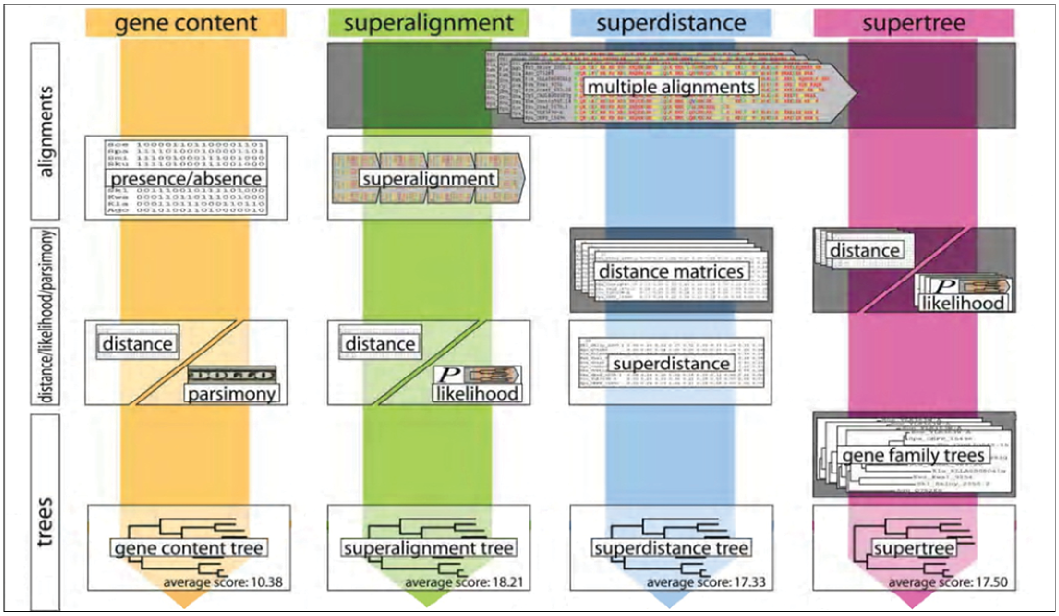


Fig. 1-1. Phylogenomic approaches (Dutlih et al., 2007).

The first step of every phylogenomic approach is making multiple alignments for all orthologous characters (Delsuc et al., 2005). The characters used in the framework of the ‘Gene content’ approach are 0 or 1 digits, representing presence (1) or absence (0) of the orthologous gene families in the genomes under analysis, with the subsequent phylogeny reconstruction employing Dollo parsimony or distance methods. It should be noted that the incongruence among trees based on the individual gene families is increasing with adding more genes (Ebersberger et al., 2007). Additionally more taxa tend increase the amount of missing data (Dutlih et al., 2007).

All other methods use nucleotides or amino acid residues as characters. The ‘superalignment’ methodology summarizes genomic information at the most basic level: alignments of individual genes/markers are concatenated in one alignment directly without any intermediate computations. Then, this alignment is analyzed. The ‘superdistance’ approach summarizes phylogenetic information at a more advanced step, by computing the average distances between each pair of OTUs in all distance matrices built from alignments of individual genes/markers (Kunin et al., 2005). The ‘supertree’ approach combines phylogenomic information already on the level of a tree topology, by creating a consensus tree, retaining majority of the branches found in individual gene/marker trees (Bininda-Emonds et al., 2002; Bininda-Emonds, 2004; Daubin et al., 2002).

Simulation studies demonstrated that ‘superalignment’ is the most consistent phylogenomic approach (Dutlih et al., 2007; Swenson et al., 2010; Kupczok et al., 2010). The ‘superalignment’ method incorporates restrictive selection of orthologous characters, minimizing the chance of stochastic errors, related to i) artificial definition of a gene family based on some arbitrarily chosen threshold value for sequence similarity, ii) “big genome attraction” phenomenon (large genomes are more likely to contain at least one member of every gene family, and thus will share more similar characters in the matrix of gene content) (Lake and Rivera, 2004), and iii) different length and information content of genes (summarizing trees from information-poor markers with trees from information-rich markers will result in the final topology strongly affected by unreliable signal from the former marker category) (Gatesy and Springer, 2004). Taking these considerations into account, the 'superalignment' method was taken as the method of choice in the framework of this thesis.

1.1.4 Even the best method is error-prone

Using the 'superalignment' strategy for phylogeny reconstruction solves a range of issues related to lack of supporting characters and differential lineage sorting that plagued marker-based studies. At the same time it poses new challenges to molecular phylogenetics, since it reveals the existence of drawbacks in methods of phylogeny reconstruction which, being masked by other problems, were difficult to discern in small data sets. The appearance of conflicting topologies of gymnosperms (Zhong et. al., 2010; Finet et al., 2010) basal angiosperms (Goremykin et al., 2005; Leebens-Mack et al., 2005; Jansen et al., 2006; Graham and Iles, 2009), and red algae (Moreira et al., 2000) strongly supported by bootstrap proportion values in genome-scale studies when different methods and substitution models are applied represents compelling evidence of such drawbacks.

It has long been noted that multiple substitutions per site, derived from an elevated rate of molecular evolution, lead to topological errors in phylogeny reconstruction, because these positions have lost historical signal (Olsen, 1987).

Saturated positions lead to appearance of the long branch attraction artifacts (LBA). LBA causes grouping of two or more unrelated, fast mutating operational taxonomic units (OTUs) on a tree as sisters (Felsenstein, 1978). Felsenstein (1978) described that this artifact only for maximum parsimony analysis. Sensitivity of maximum parsimony algorithm to LBA stems from no explicit substitution model assumed by the method, which takes all the difference in observed character states for historical signal. However, later LBA was shown to affect model-based methods in the case of model misspecification too (Gaut and Lewis, 1995; Chang, 1996; Lockhart et al., 1996; Sullivan and Swofford, 1997; Pol and Siddal, 2001). Substitution models describe the probability of transformation from one character state to another for nucleotide and amino acids alignments. Distance methods average model settings over the whole sequence length, reducing all local variations in substitution rate in genomic data to a single number, which makes them not particularly fit to account for multiple substitutions in variable sequence regions, and, hence, LBA-prone (Huelsenbeck, 1995). By contrast, the maximum likelihood method estimates probability of observing a given substitution pattern at every alignment position, given the settings of the substitution model. This makes it less susceptible to LBA, yet in the case a model does not describe the data well, this method is becoming LBA-prone (Huelsenbeck, 1995; Swofford et al., 2001; Kolaczkowski and Thornton, 2004; Ho and Jermiin, 2004; Jermiin et al., 2004; Lockhart et al., 1996). Even if the model is correct, results still can be affected by LBA due to the finite dataset (Yang, 1997).

Currently no model perfectly describes any data set. For nucleotide sequences, the most widely accepted model is general time reversible (GTR) model (Tavaré, 1986). This model has eight free parameters (five substitution rate parameters, 3 frequency parameters) plus the overall number of substitutions per time unit). Special cases of GTR can be divided onto 64 distinct models with different combinations of free parameters. Most of the currently used models (JC69 (Jukes and Cantor, 1969), K80 (Kimura, 1980) and its variation K3P (Kimura, 1981), F81 (Felsenstein, 1981), HKY85 (Hasegawa et al., 1985), T92 (Tamura, 1992), TN93 (Tamura and Nei, 1993)) are such special cases of the GTR model.

It should be noted that, within the GTR model family, rates of substitutions are assumed to be homogeneous, i.e. the same for the whole alignment, and the substitution rates between character states are assumed to be reversible, that is, the same in either direction of change (Jayaswal et al., 2005). Also, the character frequencies are assumed to be stationary, which implies that they remain uniform for all alignment regions. Much of inability to describe data well by the GTR model is due to above model restrictions, over-simplifying the real substitution process. Tests of matched-pair symmetry among the sequences (Ababneh et al., 2006) demonstrate that above assumptions of reversibility, stationarity, and homogeneity are only very rarely met (Jayaswal et al., 2011a, b).

Real data normally evolve heterotachously (Pickett et al., 2005), which means that substitution rate does not remain constant, but shifts in different lineages over time (Lopez and Philippe, 2001). They are typically characterized by compositional bias (Collins et al., 1994) which would not be possible under stationarity and reversibility of the substitution process.

Moreover, ability of any model to account for multiple substitutions per site is limited due to increased variance of model predictions for saturated substitution paths. Basically, in maximum likelihood (ML) framework, it is impossible to predict all changes among character states at internal branch tips, at saturated positions (Strimmer, 1997). Erosion of phylogenetic information over time due to multiple hits is an objective factor, and robustness of tree reconstruction generally declines with substitutional saturation (Misof et al., 2001).

Relevance of these error sources increase with increasing taxonomic range of organisms under analysis (Dutlih et al., 2007). Saturation was cited as affecting correct placement of the gymnosperm outgroup within the angiosperm ingroup (e.g. Graham and Iles, 2009) and placement of Gnetales (Zhong et al., 2010). However, within a closely related group of organisms, characterized by similar life span, as demonstrated below, on the example of apple cpDNA data, the assumption of stationarity, homogeneity and reversibility of the substitution process might not be violated for the majority of sequences, and multiple substitutions do not pose an obstacle, as they are not present in the data. Thus, amassing enough characters to resolve the topology can be expected to improve an outcome of phylogeny reconstruction in this case.

1.1.5 A range of available solutions to counter errors in phylogeny reconstruction

Development of more realistic models is perhaps the most straightforward way to counter model mis-specification and related tree-building artifacts. One of the most relevant recent developments in this field is publication of the CAT model (Lartillot and Philippe, 2004). This is Bayesian mixture model that allows the composition at different sites of an alignment to be described by a number of different frequency matrices (Lartillot and Philippe, 2004). This model better accounts for the possibility of multiple substitutions, and subsequently most effectively avoids LBA artifacts compared to the GTR family models which assume a singular character frequency matrix for the whole alignment (Lartillot et al., 2007; Phillipe et. al, 2007; Brinkman and Philippe, 2008; Talavera and Vila, 2011). However, this model does not account for heterotachy and compositional heterogeneity among species and does not guarantee finding the correct tree.

Perhaps the most widely cited remedy to avoid LBA is adding taxa to break up long branches (Hendy and Penny, 1989; Hillis, 1996, 1998; Swofford et al., 1996; Graybeal, 1998; Page and Holmes, 1998). The improvement of phylogenetic reconstruction in the case of curated/increased taxon sampling is well-documented (Philippe, 1997; Halanych, 1998; Moreira et al., 1999; Chen et al., 2001; Jenkins and Fuerst, 2001; Soltis and Soltis, 2004). In addition, to overcome LBA and improve phylogeny one might try excluding problematic (e.g. born on long branches) taxa from the analysis, however it does not help if we are interesting to determine relationships between these branches. Some authors practiced this approach (Hanelt et al., 1996; Lyons-Weiler and Hoelzer, 1997; Farias et al., 2001; Dacks et al., 2002). The problem with taxon sampling is that there are no set rules telling the analyst when taxon sampling becomes "sufficient". Not seldom the claim of imbalanced/insufficient taxon sampling was instrumentalized to criticize an unwanted outcome of phylogeny reconstruction (Lecointre et al., 1993; Rijk et al., 1995; Omland et al., 1999; Saunders and Edwards, 2000; Tuinen et al., 2000; Murphy et al., 2001a, b; Johnson, 2001). Moreover, one cannot improve taxon sampling when the relevant taxa have become extinct.

In the case when data are highly affected by saturation, exclusion of the saturated positions from analysis helps to recover the correct tree (Dutlih et al, 2007; Jeffroy et al., 2006). In coding sequences, removal of the third codon positions, which has an elevated substitution rate, and, thus has higher chance of being saturated and randomized (Swofford et al., 1996; Sullivan and Swofford, 1997), was shown to be effective to avoid LBA artifacts (Huelsenbeck and Lander, 2003). However, this approach is accompanied by loss of historical signal and has a very high cost in terms of reduction in resolution (Källersjö et al., 1999).

More direct way to deal with saturation is to identify and remove only the fastest-evolving positions. Several methods of determination and removal of the most saturated positions have been suggested.

A maximum likelihood-based method was first proposed by Ruiz-Trillo et al. (1999) and Hirt et al. (1999). The method involves sorting alignment positions according to their gamma rate assignment. As the assignment depends on the input topology, these methods are tree-dependent; LBA artifacts present on the initial tree would lead to incorrect estimation of substitution rate, and, subsequently, to incorrect sorting results (Rodriguez-Ezpeleta et al., 2007).

At the same year, Brinkmann & Philippe (1999) and Lopez (1999) presented the Slow-Fast method. This method involves sorting characters according to parsimonious tree length, computed based on each position in alignment and requires a pre-determined tree topology to estimate the number of changes for each position using MP algorithm. The advantage of this method is an option to omit problematic branches, using the sum of lengths of the remaining branches; however, even this approach requires a priori knowledge of the "true" tree.

It should be noted, that noise reduction methods are required exactly in the situations, in which the tree topology is uncertain, and, potentially wrong. Therefore, tree-dependence of the above methods severely limits their usefulness.

By contrast, Pisani developed the compatibility method, which is independent of tree topology (Pisani, 2004). This method is based on the compatibility principle (Le Quesne, 1969), which implies that two characters are compatible if they can be mapped on a tree without homoplasy. The method calculates the number of other positions with which each site is incompatible. Highly incompatible sites are identified as fast-evolving positions. According to this method, the correct tree is derived from the largest number of sites compatible to any tree, which is assumed to represent historical signal, and the most incompatible characters are considered to represent non-phylogenetic signals.

Despite the availability of several approaches to remove noisy data, the issue of character-stripping should cease has never been settled. The problem of finding the stopping criterion for character removal has been acknowledged (Pisani, 2004) as crucial for site-stripping, however, the spectrum of recommendations (Pisani, 2004) for the criterion was so broadly defined (Site removal should be stopped if further character deletion results in a significant and systematic deterioration of the results, i.e., appearance of obviously senseless clusters and/or substantial loss of resolution, support, or a decrease in the likelihood of the recovered trees. Testing presence of clustering signal in deleted characters (e.g., the PTP test and/or phylogenetic analyses to visualize their information should be conducted, etc.) that it precluded its practical use. In fact, nobody knows which information content level, or concrete stage of result deterioration should be considered “significant”, and how “substantial” the loss of support should be. Obviously, loss of support for LBA artifacts is a good and not a bad sign. Thus, the need for determination of cut-off point for sites removal is great.

1.2 Data choice – issue in plant phylogenomics

Evolution of the nuclear genome is characterized by numerous deletions and duplications and consequently by appearance of paralogous genes and gene families (Ohno, 1970). As a result, phylogenetic analyses based on “similar” nuclear genes always harbor a certain risk of being misled. In order to construct phylogenetic trees using a nuclear data set it is necessary to know the structure of investigated genomes well enough to ensure that only orthologous genes are considered. Currently there are only few completely sequenced plants genomes, which limits usefulness of nuclear-based phylogenomics analysis. Within shallow taxonomic level, nuclear genomes might exchange loci among species, leading to mixed genome structure (Harrison and Harrison, 2011). This provokes criticisms related to non-orthologous nature of nuclear phylogenetic markers sampled across the *Malus* (Harrison and Harrison, 2011).

Plant mitochondrial genomes are characterized by large and varying sizes, even within species (*Zea*) and by extensive recombination and ongoing import of nuclear sequences in non-coding mtDNA regions (i.a. *Arabidopsis* (Giege and Brennicke, 1999), *Brassica* (Handa, 2003; Chang et al., 2011), *Oryza* (Tian et al., 2006), *Triticum* (Ogihara et al., 2005), *Zea* (Clifton et al., 2004), *Vitis* (Goremykin et al., 2009a), *Malus* (Goremykin et al., 2012)), which makes most of mtDNA sequence unique, and therefore not suitable for phylogenetic analysis. At the same time, mitochondrial genes are few and very slow-evolving (Drouin et al., 2008) which limits the number of characters to resolve the tree structure. It has been speculated that mtDNA can even harbor mitochondrial genes from other, unrelated plant species (Bergthorsson et al. 2003, 2004; Adams and Palmer, 2003), which raises the question of gene orthology in this molecule. These factors limit usefulness of plant mtDNA sequence as a reference in phylogenetic studies.

Spermatophyte cpDNA encodes, on average, 90 proteins (Ravi et al., 2008), which is two to three times more compared to corresponding mtDNAs. In comparison with the other two plant genomes, chloroplast genes are characterized by intermediate substitution rates (for all plants the ratio of synonymous substitution rates for mitochondrial, chloroplast and nuclear genes is 1:3:10, respectively) (Drouin et al., 2008).

Nearly-perfect cpDNA colinearity among even unrelated angiosperm species (e.g Goremykin et al., 2003a) speaks in favor of rarity of recombination in this molecule. Also, important for phylogenetic studies on shallow taxonomic levels, generally uniparental mode of cpDNA inheritance causes exchanges between the genomes of different individuals to be rare; and even if cpDNA is biparentally inherited, chloroplast fusion is a very rare event (Gillham et al., 1991; Kuroiwa, 1991).

Thus, as all genes in chloroplast genome are expected to be strictly vertically inherited (e.g. Yamane and Kawahara, 2005; Ravi et al., 2008; Soltis et al., 2009), differential lineage sorting of various genome loci and horizontal gene transfer cannot obscure the inference of phylogeny as it might in the case of the apple nuclear DNA (Harrison and Harrison, 2011) and angiosperm mitochondrial DNA (Bergthorsson et al., 2003, 2004; Martin et al., 2005). Taking all this into account, cpDNA, in contrast to mitochondrial and nuclear genomes, appears to be well-suited as versatile tool for plant phylogeny reconstruction on both shallow (Bortiri et al., 2008; Parks et al., 2009) and deep taxonomic levels (Raubeson and Jansen, 2005; Moore et al., 2006). Thorough investigation of its reliability as universal phylogenetic marker and working out an analytical framework to expand its potential constitutes the main research goal in the framework of the thesis.

1.3 Practical taxonomic issues to be addressed

1.3.1 Phylogeny of species and cultivars of Malus

The domesticated apple, *Malus domestica* Borkh. – is one of the most important temperate fruit crops. Despite a wide number of recent studies, phylogenetic affinities of domesticated apple remain not well-resolved. In general, attaining high resolution of the phylogenetic relationships within *Malus* has been very problematic (Luby, 2003; Feng et al., 2007; Quian et al., 2006; Li et al., 2012; Micheletti et al., 2011). A number of authors tried to address the issue using genetic fingerprinting techniques; however, these studies did not converge on any particular outcome. NJ and UPGMA trees based on the matrix of RAPD fragment distribution from cultivated and wild apple accessions (Zhou and Li, 2000) were incongruent, with no strongly supported (BP index >80%) branches subtending representatives of different species. In the RAPD-based analyses of 155 cultivated and wild *Malus* accessions (Oraguzie et al., 2001) the results obtained contradicted all previous knowledge to the degree that the authors stated that "the grouping of genotypes based on the phenogram and scatter plot generally did not reflect the pedigree or provenance of the genotypes". Inadequacy of RAPD methodology for analysis of phylogenetic affinities of domesticated apple was attributed (Iketani, 1998) to complex heterozygous structure of apple genome, but more likely, it is due to general volatility of the methodology.

AFLP-based analyses, encompassing most of the wild apple species and a wide range of *Malus domestica* (MD) varieties (Zhang et al., 2007; Ling et al., 2009) recovered very dissimilar tree topologies, with MD accessions assuming sister positions to a wide array of branches, subtending representatives of a majority of the sections within the genus. Similar outcome was obtained in a SSR-based analysis encompassing 142 wild and domesticated *Malus* accessions (Hokanson et al., 2001). The authors noted: "SSR data were not useful in identifying genetic relationships among this diverse collection of accessions, with the majority of the accessions not clustering in ways concordant with taxonomic information and/or geographic origin". The same type of result was observed in recent SSR-based analysis of a broad spectrum of *Malus* accessions (Potts et al., 2012), wherein all six major clades within *Malus*, as defined by the authors, contained terminal branches bearing different MD varieties. High dependence of the outcome of the SSR analysis within the genus on the marker set can be seen in Zhang et al. (2012). In this study, topological changes, affecting most of the tree branches were observed when two sets of the SSR markers were applied to the same *Malus* accessions. Lack of statistical support for the majority of the branches in the Neighbor-Joining analyses was also evident in analysis of another recently published SSR matrix (Micheletti et al., 2011, Fig 1b).

Comparison of the genetic fingerprinting studies does not tell as much of infrageneric phylogeny, as of the need to utilize the data with more robust signal structure.

However, application of the phylogenetic methodology so far offered only a partial improvement in resolution. Application of the nuclear ribosomal ITS, commonly used in phylogenetic reconstruction of the Rosaceae, including Pyreae (Campbell et al., 1995; 1997; Oh and Potter, 2003; Lo et al., 2007) to phylogeny reconstruction within the genus yielded trees with large share of unresolved branches and generally low bootstrap support (Feng et al., 2007; Li et al., 2012). Attempts to combine ITS data with the data based on the chloroplast *mat*K (Robinson et al., 2001; Harris et al., 2002; Juniper, 2007) did not improve the situation, as *mat*K gene contains only 16 informative characters across the genus *Malus*. The chloroplast *atp*B-*rbc*L spacer, also widely used as a phylogenetic marker in Rosaceae (Wissemann and Ritz, 2005; Campbell et al., 2007; Lo and Donoghue, 2012) contains only 5 polymorphic sites in *Malus* (Savolainen et al., 1995).

This problem is not unique to *Malus*. At low taxonomic levels, recent divergence, rapid radiations, and conservative genome evolution commonly yield scarcity of characters to resolve branches (Parks et al., 2009). The need to improve character sampling for studying phylogenetic relationships within the genus has been long recognized (Forte et al., 2002), but only a few authors included many markers in their analyses, with rather modest impact on clarifying infrageneric relationships within *Malus*. Presence of only one moderately (73% BP) supported branch on the sub-tree containing representatives of *Malus* (as shown on Fig. 2 in Lo and Donoghue, 2012), and many unresolved polytomies within the genus recovered in ML analysis of data from 11 chloroplast regions plus nuclear ribosomal ITS sequences (Lo and Donoghue, 2012) does not allow to classify taxa within the genus reliably.

Recently, Velasco et al., 2010 and Micheletti et al., 2011 sequenced and analyzed the largest data set utilized so far to elucidate internal phylogeny of *Malus*. The authors maintained that cumulative evidence (Neighbor-net from *p*-distances plus ML analyses of a subset of species) from these analyses points to the common ancestry of *Malus domestica* (MD) and *Malus sieversii*, however, an alternative interpretation, assuming close affinity between MD and *M. sylvestris* were also suggested based on analyses of single markers utilized by Velasco et al. (2010) (Harrison and Harrison, 2011) and distribution of cpDNA polymorphisms (Coart et al., 2006). In the latter study, assignment of polymorphic PCR products amplified from total DNA, to chloroplast "haplotypes" was done without considering the influence of sequences of chloroplast origin residing in nuclear and mitochondrial genomes, which might have caused polymorphic amplificates (Arthofer et al., 2010). Therefore, the results obtained are difficult to interpret.

Harrison and Harrison (2011) criticized suggestion of affinity of domesticated apple to *M. sieversii* based, in part, on suspected differential lineage sorting (DLS) among the sequences amplified from various nuclear genome regions. In attempt to counter presumed negative effects of DLS, Harrison and Harrison (2011) reverted to analysis of single markers amplified by Velasco et al. (2010) and found some trees wherein MD clusters with *Malus sylvestris*. Despite absence of any statistical support for these findings, Harrison and Harrison considered conclusions of Velasco et al. (2010) and Micheletti et al. (2011) as premature and suggested that more analyses are necessary.

Because introgression of *M. sylvestris* DNA into the nuclear genome of *Malus domestica* (Harrison and Harrison, 2011) was cited as a factor obscuring the results of phylogenetic inference based on the concatenation of sequences amplified from various nuclear genome regions (Velasco et al., 2010), the results obtained should be checked based on DLS-free data.

One way to avoid DLS-derived artifacts, and at the same time to be able to apply the 'superalignment' strategy, found in simulation studies to be the most reliable methodology of all the phylogenomic methods tested (Dutlih et al., 2007; Kupczok et al., 2010), would be to use chloroplast DNA data (see Introduction, 1.2). In angiosperms, maternal plasmid transmission of cpDNA is prevalent and the original type of cpDNA inheritance, whereas biparental inheritance is more rare and a derived trait (Hu et al., 2008). The latter mode occurs in approximately 20% of the angiosperms (Zhang and Sodmergen, 2010), yet not in Rosaceae (Hu et al., 2008). Since cpDNA is generally inherited uniparentally, exchanges between the genomes of different individuals are rare, and, even if biparentally inherited, chloroplast fusion is very rare (Gillham et al., 1991; Kuroiwa, 1991). Thus, it can be assumed that cpDNA data should be unaffected by DLS.

Complete chloroplast genomes have been introduced in few systematic studies at the shallow taxonomic level in seed plants (Bortiri et al., 2008; Parks et al., 2009). The level of statistical support for the branches observed was very high (Bortiri et al., 2008; Parks et al., 2009) and using complete chloroplast genomes was suggested as a general way to improve resolution within most land plant genera (Parks et al., 2009). It remains to be seen, if such data can improve resolution within *Malus*.

1.3.2 Determination of basal-most lineage of flowering plants

One of the most thorny issues in plants systematic, termed once by Darwin "abominable mystery," (Darwin, 1871) and cited as such ever since, is the origin of angiosperms. As discussed above (chapter 1.1.1), early marker-based studies resulted in great incongruence of topologies at the base of the angiosperm tree branch.

Yet even with introduction of genome-scale data sets, contradictory findings persist in the literature concerning relationships among basal angiosperms. A basal position for *Amborella* or alternatively *Amborella* plus Nymphaeales (water lilies) is most commonly reported. Seldom have the Nymphaeales been recovered as sister to the rest of the angiosperms when *Amborella* was included into phylogenetic analyses (e.g. Graham and Olmstead, 2000; Kim and Lee, 2004; Yang et al., 2007). In fact, constraining water lilies to be sister to the rest of the angiosperms has resulted in significantly worse likelihood scores (Leebens-Mack et al., 2005). *Amborella* as the most basal angiosperm is not consistent with fossil evidence that Nymphaeales currently represents the most basal lineage amongst extant angiosperms (Friis et al., 2001).

It should be noted that in previous parsimony (MP) analyses, when Nymphaeales have been included, the placement of *Amborella* as the most basal lineage has always received strong support (Soltis et al., 1997, 1999, 2000, 2004; Leebens-Mack et al., 2005; Jansen et al., 2006; Qiu et al., 2006; Graham and Iles, 2009). At the same time, less support has been reported in these studies for the basal placement of *Amborella* within a maximum likelihood (ML) or Bayesian Inference (BI) framework

In contrast, an *Amborella* plus Nymphaeales grouping at the base of the angiosperm subtree has been recovered in some model-based analyses (Leebens-Mack et al., 2005; Jansen et al., 2006; Bausher et al., 2006; Qiu et al., 2006; Goremykin and Hellwig, 2006, 2009; Graham and Iles, 2009; Finet et al., 2010) but not others (Moore et al., 2007; Jansen et al., 2007). MP has never indicated a basal placement for *Amborella* plus Nymphaeales branch.

Since the parameter space is much larger over which MP is susceptible to long branch attraction compared to model-based methods (e.g. Felsenstein, 1978; Hendy and Penny, 1989; Philippe et al., 2005; Lockhart et al., 2006), MP trees have been favored less by some researchers (Qiu et al., 2006; Graham and Iles, 2009: p. 222). The disparity in MP and ML analyses tentatively indicates that *Amborella* constituting the basal-most branch might be an LBA artifact.

Qiu et al. (2006) have observed that mitochondrial genes with lower substitution rates than chloroplast genes support a basal placement for *Amborella* plus Nymphaeales, whereas more divergent chloroplast genes support *Amborella* as a sister group to all other angiosperms. Qiu et al. suggest that homoplasy in the faster evolving chloroplast sequences may explain these different results. In particular, saturation at third codon positions of chloroplast sequences has been inferred between some angiosperms and outgroups (Goremykin et al., 1996, 2003, Chaw et al., 2000, 2004). Model-based analyses that exclude 3rd codon positions of chloroplast genes tend to favor recovery of a basal-most clade of *Amborella* plus Nymphaeales. (Moore et al., 2007 ; Wu et al., 2007; Mardanov et al., 2008 ; Goremykin et al., 2009b). Similarly, noise reduction analyses, which attempt to identify and remove sites that might not well described by time reversible stationary substitution models, such as GTR+I+G models, have also favored a basal *Amborella* plus Nymphaeales grouping (Barkman et al., 2000; Goremykin et al. 2009b). Noise reduction, and in particular the character-sorting algorithm of Goremykin et al., (2009b), has recently been applied and found to provide an efficient means to overcome LBA artifacts in reconstructing the evolutionary relationships among placental mammals (Goremykin et al., 2010). As sorting of alignment positions based on character state variation or compatibility criteria allows the properties of sites that impact on tree building to be more easily studied (Sperling et al., 2009), character stripping might provide good analytical framework to resolve basal angiosperm diversification.

New insights into the issue might also bring discovery of Nymphaealean affinities of the small family *Hydatellaceae,* uniting small aquatic herbs, mostly from Southern hemisphere (Saarela et al., 2007). Previously this group was identified as highly reduced monocots by morphological evidence and included in *Centrolepidaceae* (Hamann, 1976).

1.3.3 Phylogenetic placement of Gnetales

Gnetales – is a morphologically and physiologically diverse group of gymnosperms which has a number of morphological characters (vessel elements in xylem, reticulate net venation of leafs (Gnetum), megagametophyte reduction, perianth-like structures) reminding those of the flowering plants. Phylogenetic placement of this diverse group has been hotly debated in the past.

Initially, based on the morphological properties this group was classified as nearest relatives to flowering plants (Crane, 1985). However, molecular studies based on chloroplast ITS and 18S rRNA have separated Gnetales from the flowering plants (Goremykin et al., 1996; Chaw et al., 1997) revealing their common origin with conifers. Subsequent multigene studies all rejected gnetalean affinity of the angiosperms, corroborating above cited evidence that Gnetales are related to conifers and none of extant gymnosperms is more related to angiosperms then others, as they are all monophyletic. Yet the exact placement of Gnetales could not be agreed upon. To date, some studies have placed Gnetales as sister group to conifers (the "Gnetifer" hypothesis, Chaw et al., 1997), others as sisters to Pinaceae (the “Gnepine” hypothesis, Goremykin et al, 1996; Bowe et al., 2000; Chaw et al., 2000; Finet et al., 2010; Zhong et al., 2010), and others within conifers but close to ‘Cupressophyta’ (i.e. non-Pinaceae conifers; the “Gnecup” hypothesis, Nickrent et al., 2000; Doyle, 2006). These hypotheses are summarized in Fig. 1-2.

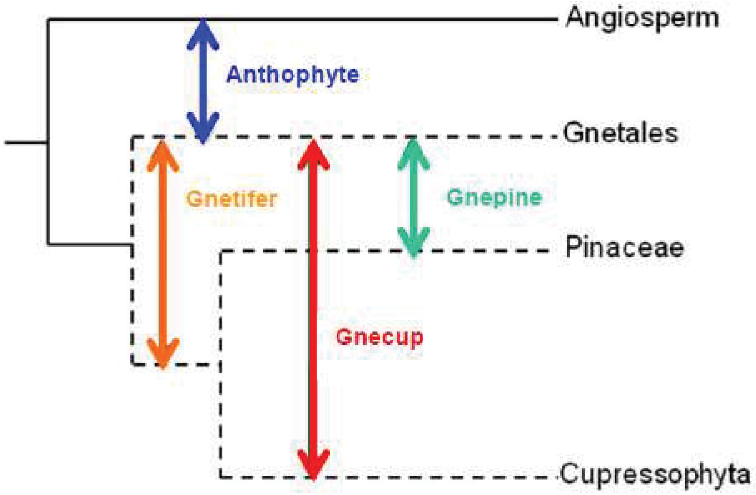


Fig. 1-2. The scheme represents four main hypotheses for relationships involving Gnetales.

A common feature of practically all recent trees including Gnetales is a particularly long branch, subtending representatives of this order from the nearest sister clade, irrespectively of the concrete composition of the latter (e.g. Chaw et al., 2000; Bowe et al., 2000; Soltis et al., 2002; Magallon and Sanderson, 2002; Burleigh and Mathews, 2004 etc.). A higher substitution rate of sequence evolution in Gnetales than in other gymnosperms (see Zhong et. al., 2010 for details) raises a possibility that volatile placement of this order within gymnosperms might be related to an LBA artifact, which changes the outcome of phylogeny reconstruction depending on taxon and gene sampling.

Besides higher rates, causes provoking LBA include model-misspecification related to certain properties of sequences not well described by stationary time reversible models (Foster, 2004; Jermiin et al., 2004; Lockhart and Steel, 2005), and poor taxon sampling due to extinction or limited availability of some taxa (Hendy and Penny, 1989).

Improving poor sampling of Cupressophyta, represented only by *Cryptomeria japonica* in recent phylogenomic studies (Zhong et al., 2010; Finet et al., 2010) and noise reduction to decrease model misspecification are obvious steps, which might shed light onto reasons behind different placements of Gnetales within gymnosperm radiation.

1.4 Aim of the thesis

Advancement of next generation sequencing techniques slashed sequencing costs and has led to unprecedented rate of accumulation of genomic data in public databases. This development is bound to have a strong impact on significance of phylogenomics, which is the most reliable way of phylogenetic tree reconstruction (Jeffroy et al., 2006). The tasks which in the past were addressed using PCR-amplified marker regions and genetic fingerprinting can now be solved at an affordable price using the wealth of characters from entire genomes.

Phylogenomics is a relatively new discipline, and is currently in the state of development. A number of studies demonstrated its outstanding potential to resolve phylogenetic relationships (Goremykin et al., 1997; 2003a; b; 2004; 2005; Martin et al., 1998; Turmel et al., 1999; Lemieux et al., 2000; Kugita et al., 2003; Wolf et al., 2005; Bausher et al., 2006; Jansen et al., 2006; Lee et al., 2006; Ravi et al., 2006; Ruhlman et al., 2006 etc.), however its applicability still remains practically unexplored for the studies on shallow taxonomic levels, (within species and genera) and is severely limited by systematic errors on deep taxonomic levels.

1. Exploring the potential of chloroplast phylogenomics to provide solutions to previously intractable phylogenetic problems constitutes the main methodological goal of this work. Approaches leading to reliable results in such difficult areas constitute a useful guideline for future studies.

2. In contrast to other plant genomes, chloroplast genome is free from horizontal gene transfer from other unrelated species and from hybridization-based sequence introgressions caused by sexual propagation. Whereas mitochondrial genomes greatly vary in size and composition even among genetic lines of one species (Tian et al., 2006; Chang et al., 2011), which does not allow establishing homology of mtDNA, except for the gene sequences, the chloroplast genome structure is substantially conserved, which allows to use its whole sequence as a phylogenetic marker. I wished to investigate whether this marker has enough resolution power to uncover origin and evolution of domesticated apple varieties, which was especially difficult to elucidate in the past due to lack of resolution and volatile tree structure.

3. Notwithstanding obvious advantages of phylogenomics, systematic errors in phylogeny reconstruction cannot be overcome by amassing large number of characters (Philippe et al., 2005, 2011; Delsuc et al., 2005; Jeffroy et al., 2006). Such errors, termed “non-phylogenetic signal” (Philippe et al., 2005) were categorized as rate, compositional, and heterotachous signals. Removing the part of the alignment which is dominated by non-phylogenetic signal (Philippe et al., 2005; Pisani, 2004) reduces the chance of error even without the use of more realistic substitution models and methods (Philippe et al., 2005). Correlation between substitutional saturation and non-phylogenetic signals was documented (Jeffroy et al., 2006; Rodriguez-Ezpeleta et al., 2007); thus, effective methodology of removal of saturated positions holds promise of liberating phylogenetic reconstruction from tree-building artifacts (Ruiz-Trillo et al., 1999 and Hirt et al., 1999; Brinkmann & Phillipe, 1999 and Lopez, 1999; Pisani, 2004). Developing relevant methodology and demonstrating its prowess in solving thorny phylogentic problems of systematic botany, such as identification of the basal-most clade of angiosperms and identifying the closest relatives of Gnetales constitutes another goal of the thesis.

2 Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line[1]

Abstract

Both the origin of domesticated apple and the overall phylogeny of the genus *Malus* are still not completely resolved. Having this as a target, we built a 134,553 position long alignment including two previously published cpDNAs and 45 *de novo* sequenced, fully co-linear chloroplast genomes from cultivated apple varieties and wild apple species. The data produced are free from compositional heterogeneity and from substitutional saturation, which can adversely affect phylogeny reconstruction. Phylogenetic analyses based on this alignment recovered a branch, having the maximum bootstrap support, subtending a large group of the cultivated apple sorts together with all analysed European wild apple (*Malus sylvestris*) accessions. One apple cultivar was embedded in a monophylum comprising wild *M. sieversii* accessions and other Asian apple species. The data demonstrate that *M. sylvestris* has contributed chloroplast genome to a substantial fraction of domesticated apple varieties, supporting the conclusion that different wild species should have contributed the organelle and nuclear genomes to domesticated apple.

Keywords: *Malus domestica*, chloroplast genome phylogeny, base compositional heterogeneity, hybridisation

[1] Original manuscript: Nikiforova S, Cavalieri D, Velasco R, Goremykin V. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. Molecular biology and evolution (accepted to journal with a revision, 25 February 2013). Own contribution: 90% (analyses, writing)

Introduction

The domesticated apple, *Malus domestica* Borkh. – is one of the most important temperate fruit crops. The origin of the crop from wild progenitors is, for several reasons, relevant both to the breeders and to taxonomists. Attaining high resolution of the phylogenetic relationships within the genus *Malus* is however, still problematic (e.g. Luby, 2003; Li et al., 2012). For example, high dependence of the outcome of the SSR analyses within the genus *Malus* on the marker set can be seen in Zhang et al. (2012): profound changes in the phylogenetic tree topology were observed when two sets of the SSR markers were applied to the same *Malus* accessions.

In general, improvement of the phylogenetic methodology so far offered partial improvements in tree resolution within the genus. Nuclear ribosomal ITS, used in phylogenetic reconstruction of the Rosaceae (Campbell et al., 1995, 1997; Oh and Potter, 2003; Lo et al., 2007), for the genus *Malus* yielded trees with many unresolved branches and low bootstrap support (Feng et al., 2007; Li et al., 2012). Attempts to combine ITS data with the chloroplast *mat*K gene sequences (Robinson et al., 2001; Harris et al., 2002; Juniper 2007) did not improve the situation, as *mat*K gene contains only 16 informative characters across the genus *Malus*. The chloroplast *atp*B-*rbc*L spacer, also widely used as a phylogenetic marker in Rosaceae (Wissemann and Ritz, 2005; Cambpell et al., 2007; Lo and Donoghue, 2012), contains only 5 polymorphic sites in *Malus* (Savolainen et al., 1995).

Recently, Velasco et al. 2010 and Micheletti et al. 2011 sequenced and analyzed the largest data set utilized so far to elucidate inter-generic phylogeny of *Malus*. The authors maintained that cumulative evidence (Neighbor-net from *p*-distances plus ML analyses of a subset of species) from these analyses points to the common ancestry of *Malus domestica* (MD) and *Malus sieversii.* However, a certain degree of affinity between MD and *M sylvestris* was also suggested based on single markers utilized by Velasco et al. (Harrison and Harrison, 2011) and distribution of cpDNA polymorphisms (Coart et al., 2006). In the latter study, assignment of polymorphic PCR products amplified from total DNA to chloroplast "haplotypes" was done without considering the influence (Arthofer et al., 2010) of sequences of the chloroplast origin residing in nuclear and mitochondrial genomes, which might have caused not genome-specific amplification.

Introgression of *M. sylvestris* DNA into the nuclear genome of *Malus domestic*a was cited (Harrison and Harrison, 2011) as a factor obscuring the results of phylogenetic inference based on the concatenation of sequences amplified from various nuclear genome regions (Velasco et al., 2010).

However, evidence of wild species introgression is of complex interpretation, considering that, while nuclear genome is inherited biparentally, chloroplast and mitochondrial genomes are maternally transmitted. Given the situation, phylogenetic relationships among closely related plant species, particularly of those of economic interest that underwent multiple cycles of conventional breeding, should be investigated independently for the different cell genomes. The target of this paper was to investigate the phylogenetic relationships of wild and domesticated apples based on chloroplast genome data.

In angiosperms, transmission of cpDNA is monoparental (Rosaceae included, Hu et al., 2008), with biparental inheritance observed only for a few crown-group lineages (Hu et al., 2008). When cpDNA is inherited uniparentally, exchanges between the genomes of different individuals are rare, and chloroplast fusion is even rarer (Gillham et al. 1991; Kuroiwa 1991). A nearly-perfect cpDNA colinearity among even unrelated angiosperm species (e.g Goremykin et al., 2003) also speaks in favour of rarity of recombination in cpDNA. Thus, introgression of sequence material from different species into the chloroplast genome molecule cannot obscure the inference of chloroplast genome-based phylogeny.

Complete chloroplast genomes have already been successively used in systematic studies at the shallow taxonomic level in seed plants (Bortiri et al. 2008; Parks et al. 2009). The level of statistical support for the branches observed was very high (Bortiri et al. 2008; Parks et al. 2009). We analyze in this paper a dataset including 46 completely or nearly completely sequenced chloroplast genomes sampled across the genus *Malus*, with emphasis on the sampling within the *domestica*-*sylvestris*-*siversii* lineage. Phylogenetic analyses of the chloroplast genome data have resulted in a tree topology characterized by a resolution previously unattained within the genus *Malus*.

Results

Overall data properties

The dot-plot of the evolutionary vs. observed distances among the OTUs based on the 134,553 position long alignment of 47 chloroplast genomes (Fig. 2-1) showed a nearly perfect linear distribution.

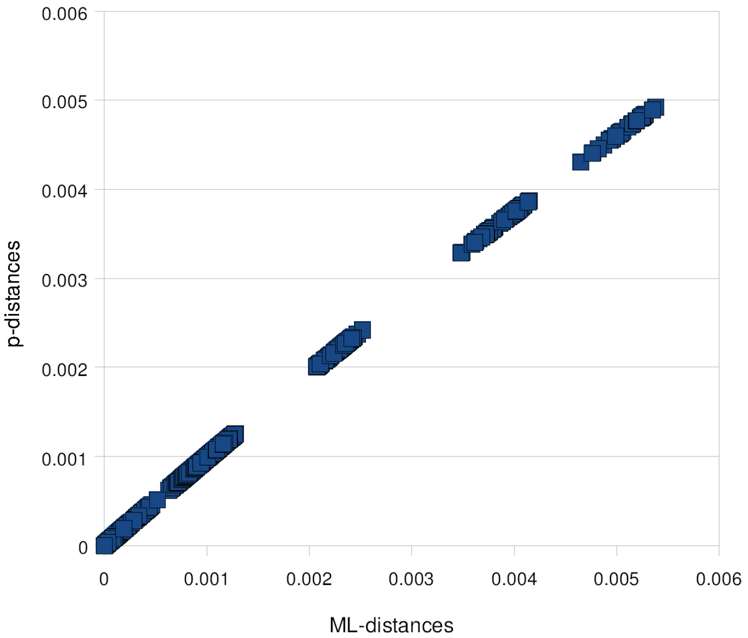


Fig. 2-1. Plot showing the distribution of the uncorrected p-distances vs. ML distances estimated using the settings of the best-fitting TVM+I+G model. The distances were calculated based on the 134,553 pos. long alignment of chloroplast genomes, including *Pyrus*.

The mean ML distance among all the OTUs, estimated using the settings of the best-fitting TVM+I+G model in PAUP*, was 0.00134, which is only marginally different from the corresponding uncorrected p-distance (0.00128). Thus, superimposed substitutions, causing on deeper taxonomic levels model-misspecification and related tree-building artifacts in phylogenectic analyses based on cpDNA data (Zhong et al. 2011; Goremykin et al., 2013) should not pose a problem in the current analysis. The 5% chi-square-test, implemented in Tree-Puzzle program was adopted to determine if the base composition in the alignment was similar to the average base composition of the whole 47 OTU alignment. All accessions, except *M. mandjurica*, passed the test. Overall results of Bowker's test of matched pairs symmetry, as implemented in SeqVis program, indicated that, out of 1081 pair-wise comparisons, only 50 (~ 4.6%) showed significant compositional heterogeneity (P-value < 0.05). Thus, the null hypothesis of evolution under stationary, reversible and homogeneous conditions could not be rejected for the majority of the sequences under analysis.

The 134,553 pos. long alignment of cpDNA sequences from *Pyrus* and 46 *Malus* species and cultivars contains 773 informative positions (in the sense of Maximum Parsimony). Of all informative positions, only three had three character states, the rest contained two character states. The data structure indicates no erosion of the historical signal in the cpDNA sequences under analysis. Good resolution of the overall tree topology (Fig. 2-2) can thus be attributed to the fact that phylogenetic signal is well-preserved in the data and is not distorted by multiple substitutions and strong compositional bias.

At the same time, unresolved clusters with zero or nearly-zero branch lengths at the crown part of the tree (Fig. 2-2) point at the resolution limit that chloroplast genome sequences have at the shallowest taxonomic range. For example, *M domestica* cv. Gala and *M. domestica* cv. Florina diverged from the common ancestor of the maternal line, Red delicious at an estimated time corresponding to 100 years ago. Based on the observation that chloroplast genomes contain no informative characters to distinguish pedigree of apple cultivars in the 6-species monophyletic cluster, including cultivars Gala and Florina (Fig. 2-2), one can conclude that cpDNA data might be of limited use for intraspecific, population-based studies of plant biodiversity.

Tree structure

Apple species *M. angustifolia* and *M. ioensis* of the *Malus* Section Chloromeles, as defined in the Germplasm Resources Information Network – GRIN, which we use as taxonomic reference, form the the most basal clade on the *Malus* subtree (Branch E on Fig. 2-2). This placement receives maximum bootstrap proportion (BP) support. Thus, among the species tested, *M. angustifolia* and *M. ioensis* can be considered the ancestral lineage of *Malus.*

Next representatives of the Section Sorbomalus, (*M. kansuensis, M. honanensis, M. prattii, M. yunnanensis*) are branching off (Branch D on Fig. 2-2, 100% BP support). Affinity of *M. fusca* to *M. kansuensis* (which are sometimes recognized within Ser. *Kansuensis* (Robinson et al. 2001)) was not confirmed in our analysis: *M. fusca* cpDNA line have a sister group relationship in all bootstrap replicas made, to the broad assemblage, uniting *M. sieversii* and related species. Further up in the tree, resemblance between tree topology and taxonomy of the genus was lost. The next strongly supported (90% BP) branch C unites species attributed to: Section Malus (*M. floribunda*, *M. prunifolia*, *M. spectabilis*, *M. xanthocarpa*, *M. zhaojiaoensis*); Section Sorbomalus (*M. sargentii*, *M. sieboldii*, *M. transitoria*); and Section Gymnomeles (*M. rockii*, *M. sikkimensis*). A further, strongly supported (100% BP) large branch uniting a number of wild species (Branch A), contains representatives of section Malus (*M. asiatica*, *M. sieversii*, *M. kirghisorum, M. orientalis*, *M. pumila,*), Section Sorbomalus (*M. mandshurica*) and Section

Gymnomeles (*M. baccata*, *M. halliana*, *M. hupehensis* and two hybrids with the chloroplasts deriving from different accessions of *M. baccata - M. adstringens and M. micromalus*). A conclusion based on these results is that the overall taxonomic subdivision of the genus *Malus* does not correspond to the phylogeny of the maternal line of the species analyses.

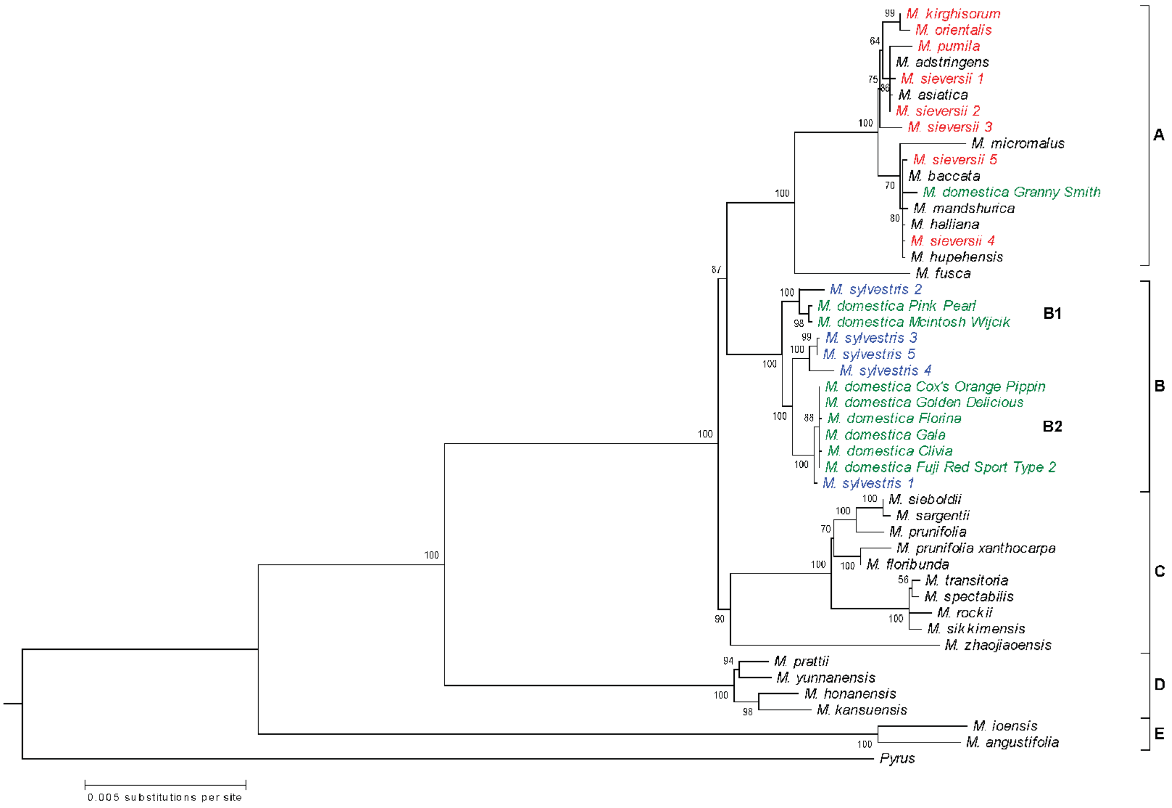


Fig. 2-2. Tree reconstructed from ML analysis using the settings of the optimal substitution model (TVM+I+G model) found by double-fitting procedure (Goremykin *et al.*, 2010) for the 134,553 pos. long alignment of chloroplast genomes. The numbers next to the tree branches represent bootstrap support values.

Accessions of *M. sieversii*, a central Asian species, whose nuclear genome was suggested to be the ancestor of domesticated apple, were scattered across branches containing other wild species. The clade supported by 80% BP, subtending *M. sieversii*, 4 and 5, included also *M. baccata, M. mandshurica, M. halliana, M. hupihensis* and *M. domestica* cv. Granny Smith, is clearly separated from a second well-supported (86% BP) branch subtending, among other OTUs, *M. sieversii* 1 and 2. These data indicate that genetic diversity of chloroplast genomes within *Malus sieversii* exceeds that between other species and might justify its splitting onto at least two species.

Eight of nine *Malus x domestica* cultivars analysed formed a branch with accessions of European crab apple, *M. sylvestris*, which was recovered in all 100 bootstrap replicas made (Branch B on Fig. 2-2). Within this large branch, *M. x domestica* chloroplasts have polyphyletic origin, evidenced by two strongly supported monophyla, comprising *M. x domestica* accessions only, each sharing a strongly supported sister group relationship with different accessions of *M.*

sylvestris. Polyphyly of *M. x domestica* maternal line is further evidenced by *M. domestica* cv. Granny Smith embedded within a strongly supported (80% BP) branch with five Asian species including *M. sieversii*.

Dating results

To estimate when separation of three cpDNA lines of M. domestica occurred, we conducted two experiments. In the first, the age of the diversification of *Malus* from *Pyrus* was assumed to be about 45 million years; in the second the age of *Malus* was constrained with the earliest possible date based on molecular dating experiments (about 20 million years, Lo and Donoghue 2012). The results of our dating experiments are presented in Figure 2-3. The separation of the cpDNA line shared by Asian species and *Malus x domestica* cv. Granny Smith (Fig. 2-2, Branch A) from the *M. x domestica/M. sylvestris* lineage (Branch B) occurred somewhere between 15.54 and 10.78 million years ago (MYA). Within the lineage including *M. x domestica* and *M. sylvestris,* the separation between the cpDNA line shared by Pink Pearl and McIntosh Wijcik from the cpDNA line of other apple cultivars occurred within the 8.11-5.46 MYA range. Separation between the wild *M. sieversii* specimen and apple cultivars forming branches B1 and B2 occurred, correspondingly, 3.45-2.36 and 1.69-1.19 million years ago.

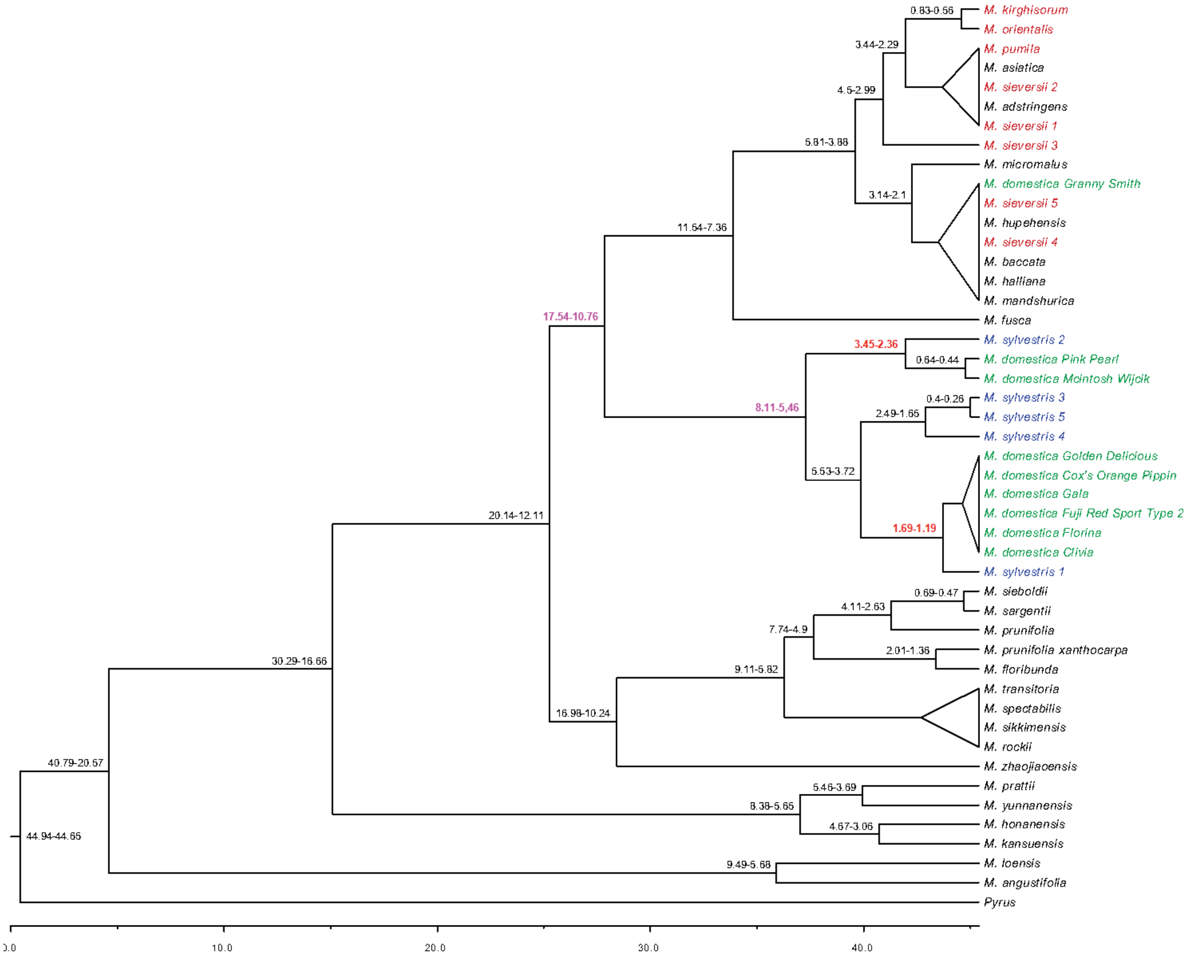


Fig. 2-3. Chronogram of *Malus* built employing Bayesian analysis as implemented in BEAST program from 134,553 pos. long alignment of chloroplast genomes. Numbers on the left side of the tree nodes denote the age of the nodes in million years. The numbers to the left of the dashes were obtained when constraining the root age to normal distribution with a mean of 45, and a standard deviation of 1. The numbers to the right of the dashes were obtained when, in addition, the age of *Malus* was constrained by a normal distribution with a mean of 20, and a standard deviation of 1. Dating for the clusters, which branching pattern could not be resolved in the ML analysis (Fig. 2-2), was considered to be unreliable and is not shown here.

Discussion

The main conclusion of this paper is that the chloroplast genome of *Malus x domestica* derives from at least two wild species, with *M. sylvestris* being the main contributor. The common origin of cpDNA of *M. sylvestris* and the majority of *M. x domestica* cultivars analysed was supported by 100% BP. The evidence provided opens a major question: apparently the nuclear and chloroplast genomes of a large part of apple cultivated varieties (Fig. 2-3) have different phylogenies.

The nuclear genome donor seems to be *M. sieversii,* as supported by the data of Velasco et al. (2010) and Micheletti et al. (2011) which i) compared 74 accessions, including 12 *M. x domestica,* 10 *M. sieversii* and 21 *M. sylvestris,* based on re-sequencing of 23 gene amplicons for a total length of 11,300 bp, with 1,507 polymorphic informative sites. The data were analyzed by a split-tree planar graph (Velasco et al., 2010) and by a maximum- likelihood method under GTR model (Micheletti et al., 2011), as suggested by Harrison and Harrison (2011); ii) by comparing the same accessions (excluding putative *M. x domestica/M. sylvestris* hybrids) and using 27 SSR markers (Micheletti et al., 2011) unrelated to the above mentioned 23 amplicons; the phylogetic tree was computed based on the "shared allele" distance index and the NJ clustering algorithm. All three phylogenies were based on nuclear genes; the separation of *M. sylvestris* from *M. sieversii* was clear and highly supported by bootstrapping. *M. x domestica* varieties clustered together with *M. sieversii.*

It is true, however, that also *M. sylvestris* has been recurrently indicated as a possible contributor to the nuclear genome of *M. x domestica* (summarized in Juniper and Mabberley (2006) and in Harrison and Harrison (2011) and Micheletti et al. (2011)), but this was, almost always, discussed considering the possibility that introgression resulted in nuclear genes private to *M. sylvestris* and not to *M. sieversii* (Micheletti et al. 2011; Harrison and Harrison 2011). Apple has been introduced to Europe by Romans and Greeks, and then from Europe it spread all over the world (Juniper and Mabberley, 2006). It was proposed to have originated either in Europe, from *M sylvestris*, a European crab apple bearing small astringent and acidulate fruits (Zohary and Hopf, 1994; Coart et al., 2006; Harrison and Harrison, 2011) or in Asia, from *M. sieversii* (Velasco et al., 2010; Micheletti et al., 2011; Cornille et al., 2012), a diverse central Asian species, characterized by a wide range of forms, colors and flavors (Way et al., 1990). Abundant reports of hybridization among domesticated apple, *M. sieversii* and *M. sylvestris* suggest polyphyletic origin of *M. x domestica* DNA loci. Cornille et al. (2012) found that 61% of the *M. x domestica* genome derives from *M. sylvestris,* which has been attributed to a recent massive introgression from the European wild apple. The introgression from *M. sylvestris* should be facilitated by self-incompatibility, long lifespan of the species and cultural practices, including selection from open-pollinated seeds. On this subject, it must be considered that in interspecific Rosaceae hybrids, the chloroplast DNA is inherited from the maternal line (Hu et al., 2008). Thus, pollination of *M. x domestica* or *M. siversii* genotypes by *M. sylvestris* would not had led to the formation of the branch B (Fig. 2-2), while the reciprocal cross remains a credible hypothesis. If the origin of the *M. x domestica* nuclear genome from *M. sieversii* is accepted, the apple varieties included in branch B would derive from hybridisation events involving *M. sylvestris* as mother, followed by back- crossing with pollen from "sweet apple" genetic lines, under a strong selection to eliminate astringency components negative for fruit taste and to increase fruit size. Such a procedure was, for example, employed in the creation of scab-resistant apple cultivars, by incorporating the *Vf* gene from *Malus floribunda 821* into *M. x domestica* (Crosby et al., 1992). Studying the pedigrees of *M. x domestica* cultivars included in branch B (Fig. 2-2), reveals that their maternal lines can be traced back to seven old "founders" (Table 2-1). The oldest founder in branch B2, Ribston Pippin, derives from seeds brought from Rouen (Normandy) to England around 1700 (Cecil, 1910). McIntosh, the oldest representative of the branch B1, was selected in Ontario, Canada, in 1792. Because, in apple, controlled breeding schemes were adopted only around 1800 (Sandlers, 2010), intentional crossing and backcrossing to wild species preceding the origin of B1 and B2 branches is unlikely. Consideration of the branching of the phylogenetic tree, in particularly, clear subdivision of clade B into sub-clades

B1 and B2 (Fig. 2-2) suggests that in *M. x domestica* the process of chloroplast genome substitution, which took place in historical time, before apple intentional breeding, occurred at least two times.

Table 2-1. Origins and pedigrees of the *M. x domestica* cultivars taken into analysis.

Variety	Date of origin
Clade B1	
Pink Pearl	
Surprise X	1944
McIntosh Wijick	
discovered in Ontario, Canada	1796
Clade B2	
Florina	
PRI 612-1 x Jonathan	1977
Delicious x PRI 14-126	
Delicious originated in Iowa, USA	1870
Fuji	
Ralls Janet x Red delicious	1939
Ralls Janet originated in Virginia, USA	late 1700s
Golden Delicious	
Grimes Golden x Golden Reinette	1891
Grimes Golden was found in West Virginia, USA	1804
Clivia	
Geheimrat Dr.Oldenburg x Cox Orane Pippin	1930
Minister von Hammerstein x Baumann's Reinette	1897
Landsberger Reinette X	1822
Cox Orange Pippin	
Ribston Pippin X	1825
Ribston Pippin originated from seeds	
brought from Rouen (France) in	1700
Gala	
Kidd's Orange Red x Golden delicious	1934
Delicious x Cox Orange Pippin	1924
Delicious originated in Iowa, USA	1870
Clade A	
Granny Smith	
Eastwood, near Sydney, Australia	1868

While the mechanisms responsible for the process of genome introgression are easily predicted, the forces which favored the end result can only be speculated upon: central roles may have played the selection for palatability and fruit size, unilateral compatibility in crosses, and even the fitness superiority of genotypes having cellular genomes deriving from different species. The testing of these hypotheses remains a subject of future studies. The finding that the *M. x domestica* variety Granny Smith has chloroplasts sharing a monophyletic origin with wild Asian accessions, *M. sieversii* included (Clade A, Fig. 2-2), indicates that the process of chloroplast genome substitution in the *M. x domestica* did not affect all cultivated apple varieties.

In the study of Velasco et al. (2010), the distribution of synonymous substitution rates (Ks) - an indication of the relative age of gene duplication based on the number of synonymous substitutions in DNA coding sequences - peaked around 0.2 for recently duplicated genes, indicating that a (recent) genome wide duplication (GWD) has shaped the genome of the domesticated apple. Dating of this GWD was based on the construction of penalized likelihood trees. Given a node of grape to rosids fixed at 115 million years ago (MYA), the GWD has been dated to between 30 and 45 MYA (Fawcett et al., 2009; other references in Velasco et al., 2010). If similar rates of protein evolution are assumed for apple and poplar, the recent apple GWD may be as old as that of poplar, about 60 to 65 MYA (Tuskan et al., 2006). Because the genetic maps of *Malus* and *Pyrus* are co-linear, this dating becomes the starting point for the radiation within the tribe Pyreae. At this time point, available molecular data indicate that the most probable ancestors of the event which generated the GWD were American Rosaceae species, corresponding to extant *Gillenia* and related taxa. In fact, the earliest fossils (48-50 MYA) of Pyreae genera are from North America (Campbell et al., 2007; Wolf and Wehr, 1988).

Our dating results, based on chloroplast DNA (see Materials and Methods), are reported in Fig. 2-3. Under 45 MYA fossil-based constraint for the common origin of *Malus* and *Pyrus*, they indicate that radiation of extant *Malus* species might have already started 40 MYA. Further diversification occurred likely in Central Asia, which is the center of origin of domesticated apple (Vavilov, 1930): between 25 and 47 different *Malus* species, including *M. x domestica,* are currently recognized there (Robinson et al., 2001), among which the asiatic *M. x asiatica, M. baccata, M. micromalus, M. orientalis, M. prunifolia* and *M. sieversii.* Around 17-11 MYA two major groups of species separated, the first one including *M. sylvestris* and *M. x domestica* (clade B, Fig. 2-2) and the other subtending the Asian wild species, including *M. sieversii,* (clade A). This subdivision corresponds to a major split among cpDNA lines of *M. x domestica*: the line (clade B) shared with the *M. sylvestris*, which later on, 8-5 MYA, divided in the B1 and B2 haplotypes; and the other line, shared among the Asian wild *Malus* species, but also present in the gene pool of *M. x domestica* variety Granny Smith (Clade A).

Comparison of the topology of Branch B (Fig. 2-2) with the geographic origin of the *M. sylvestris* reveals that the chloroplast genomes from the German *M. sylvestris* specimens (accessions 3, 4, and 5 in Fig. 2-2) separated around 5-3 MYA from those present today in cultivated apple sorts. Moreover, cpDNAs of these accessions are not related to the chloroplast genomes of cultivated apple varieties, while southern European accessions are. Six *M. x domestica* cultivars share the chloroplast genome relationship with a *M. sylvestris* specimen collected on Monte Pollino, Calabria, Italy *(Malus sylvestris* 1; Fig. 2-2). Two other cultivars build a common branch with a *M. sylvestris* accession collected in Macedonia (*Malus sylvestris* 2; Fig. 2-1). With the limitations due to the number of accessions considered in this study, it suggests that the region where *M. sylvestris* introgressed *M. x domestica* was Southern Europe.

We conclude that using *Malus* chloroplast genome data practically free from compositional heterogeneity and from substitutional saturation, we have been able to perform a highly reliable phylogeny reconstruction. Phylogenetic analyses based on this alignment demonstrate that *M. sylvestris* contributed cpDNA to a large fraction of the domesticated apple sorts, indicating that chloroplast and nuclear genomes of domesticated apple may have independent evolutionary histories. Our results provide further evidence of the mosaic origin of domesticated apple from diverse wild contributors.

Materials and methods

Sequencing

Fresh leafs of 47 wild and cultivated apple accessions, including 9 accessions of *Malus x domestica*, 5 accessions of *M. sieversii*, 5 accesions of *M. sylvestris* and 29 samples of other cultivars (please see Suppl. Materials, Table 1) were gathered from the apple tree collection maintained at the Fondazione Edmund Mach. DNA was extracted using the DNeasy Plant Mini kit (Qiagen, The Netherlands) and subsequently quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Life Technology, USA). Shotgun genomic libraries were generated via fragmentation of 0.5 µg of genomic DNA as described in 454 Life Sciences (Branford, CT, USA) protocol. Briefly, DNA was randomly sheared via nebulization and Rapid Library adaptors were blunt-ligated to fragment ends. The Multiplex Identifier (MID) Adaptors were used to distinguish reads of different specimen. Libraries were quantified via quantitative PCR using Library quantification kit – Roche 454 titanium (KAPA Biosystems, Boston, MA).

Assembly of chloroplast DNA from single reads

The chloroplast genome of *M. domestica*, cultivar Golden Delicious was previously sequenced at FEM (Velasco et al., 2010). The reads from 454 sff files were mapped onto this genome sequence, wherein a copy of the inverted repeat region was deleted, by gsMapper (454 Life Sciences, Branford, CT, USA). The selected reads were subjected to *de-novo* assembly employing gsAssembler program from the same vendor. Assemblies were transferred into the Staden package (http://sourceforge.net/projects/staden/files/) and manually edited.

The high coverage of the cpDNA contigs obtained allowed to successfully assemble chloroplast genomes. As reported (Goremykin et al., 2012), the coverage values for nuclear, mitochondrial and chloroplast genome assemblies built from the total *M. domestica* DNA preparation are 15.4 X, 168 X and 847X, respectively. Thus, the majority option for consensus sequence building used, ensures correct representation of the cpDNA sequence. Among the genomes assembled, 12 contained no gaps, for the others the mean number of gaps per sequence was 4.2, and the mean estimated gap length 237 bp.

Alignment and phylogenetic analyses

Assembled sequences were aligned manually with the help of Seaview alignment editor, because sequence similarity among the cpDNA sequences was no less than 99%. *Pyrus* cpDNA sequence was downloaded from the Genbank (Acc. no NC_015996). The alignment of 47 OTUs - 134,553 aligned positions in length, available from the Dryad database (datadryad.org, Acc. no ##) - was subject to Maximum Likelihood (ML) analysis employing the PAUP* program. The search for the best-fitting model was conducted with the help of the gamma_sorter.pl script in two stages (supplementary material, Goremykin et al., 2013). In the first stage model parameters were fitted to the Neighbour-Joining tree and the best model was selected under Akaike information criterion (AIC); in the second stage – models were fitted to the ML tree built using parameters of the best model found at the first model-fitting stage, and the next best-fitting model was also selected employing AIC.

The ML tree (Fig. 2-1) was computed in PAUP* using settings of the best-fitting TVM+I+G model and the Tree Bisection-Reconnection (TBR) search option. Bootstrap support values for the tree branches, were calculated using faster MPI version of Phyml 3.0 program, which was run with the specification of the i) TVM+I+G model, ii) the BEST search option and iii) the ML tree previously obtained employing PAUP*.

Matrices of p-distances and of the ML-distances computed under specification of the optimal TVM+I+G model settings, used to produce the Figure 2-1, were generated with the help of the noiserductop.pl script embedding PAUP*, previously published (supplementary material, Goremykin et al., 2013).

Calibration for estimating divergence times within Malus

Macrofossils assigned to Pyrinae were described in middle-to late Eocene fossil floras from the north-western North America. Clarno Formation (appr. 44 Ma) of central Oregon contains well-preserved silicified fruit of *Quintacava velosida,* sharing similarity with the Maloideae (Manchester, 1994) and wood assigned to the Maloideae (Wheeler and Manchester, 2002). Leaves classified as from *Malus* or P*yrus* are part of the middle Eocene (about 45 Ma) flora of the Republic site in Washington, USA (Wehr and Hopkins, 1994). Thunder Mountain flora of central Idaho of the same geological age contains a leaf fossil described as "*Malus collardii*" (Axelrod, 1998). Pollen assigned to *Malus* or *Pyrus* has been reported from the late Eocene Florissant locality in Colorado (Leopold and Clay-Poole, 2001) estimated to be of 34.07 ± 0.10 MYA age. Fossils with similarity to *Amelanchier*, *Crataegus* and *Photinia,* as well as some relatives of *Malus* and *Sorbus,* are known from the early middle Eocene (48–50 million years ago) (Campbell et al., 2007; Wolfe and Wehr, 1988).

Previous molecular dating for Pyraeae (including *Aronia*, *Malus*, *Amelanchier*, and *Crataegus*) assumed an age of 44 million years for the group (Lo et al., 2009), as based on estimates of DeVore and Pigg (2007). We based our calibration on 45 million years old leaf *Malus* fossils (Wehr and Hopkins, 1994 and Axelrod, 1998). Because of difficulty of distinguishing fossilized leaves of *Malus* from *Pyrus*, 45 million years was assumed to be the approximate age of the common *Malus/Pyrus* lineage.

An alternative calibration corresponded to the minimum possible age for *Malus*, estimated by Lo and Donoghue (2012) as 20 million years. This calibration point provides the minimum estimate for the divergence of apple species from a common progenitor.

Dating divergence times within Malus

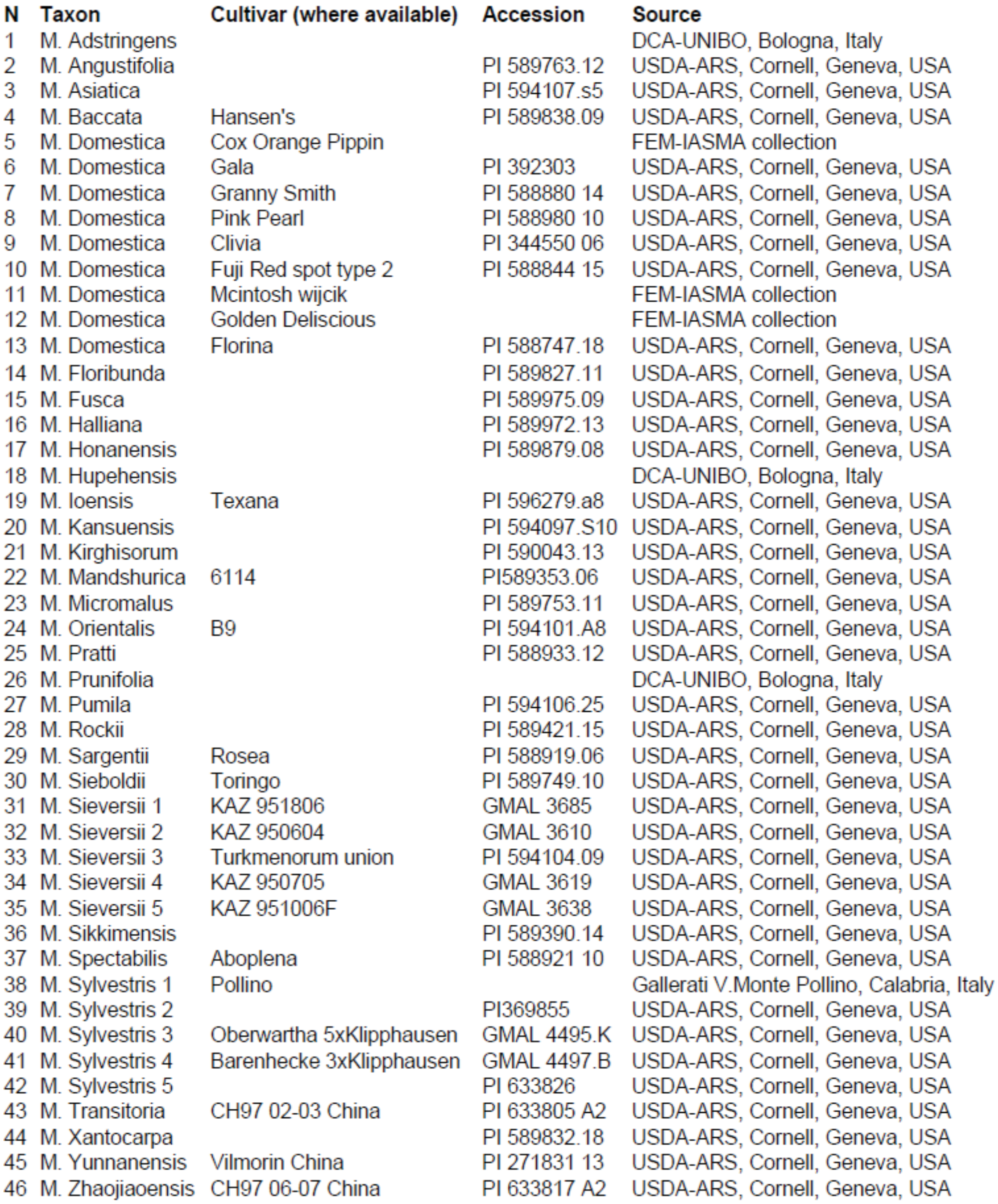
The data set used for dating was produced by manually aligning published chloroplast genomes of Rosaceae to the previously made alignment of 47 OTUs containing sequences from the genus *Malus* only. Divergence times for the major lineages were estimated using the Bayesian method as implemented in BEAST program (Drummond and Rambaut, 2007). The program was let to compute the tree topology and to optimize substitution model parameters under general definition of GTR+I+G substitution model (BEAST incorporates HKY and GTR models only). Two independent MCMC runs were performed for 10,000,000 generations, sampling every 100th generation. In both runs uncorrelated lognormal relaxed-clock model was used which allows rate variation across branches, and a Yule tree prior to model speciation. In one experiment, *Pyrus* was constrained to be the outgroup and the root age was constrained by a normal distribution with a mean of 45 MYA, and a standard deviation of 1. In the other dating experiment, *Pyrus* was constrained to be the outgroup, and the age of *Malus* was constrained by a normal distribution with a mean of 20 MYA, and a standard deviation of 1.

Acknowledgments

The authors thank Prof. Francesco Salamini for critical reading of the manuscript, several useful comments, insight and support. The authors thank Dr. Massimo Pindo for technical support of the study.

Supplementary materials

Table 1. *Malus* accessions analyzed.



N	Taxon	Cultivar (where available)	Accession	Source
1	M. Adstringens			DCA-UNIBO, Bologna, Italy
2	M. Angustifolia		PI 589763.12	USDA-ARS, Cornell, Geneva, USA
3	M. Asiatica		PI 594107.s5	USDA-ARS, Cornell, Geneva, USA
4	M. Baccata	Hansen's	PI 589838.09	USDA-ARS, Cornell, Geneva, USA
5	M. Domestica	Cox Orange Pippin		FEM-IASMA collection
6	M. Domestica	Gala	PI 392303	USDA-ARS, Cornell, Geneva, USA
7	M. Domestica	Granny Smith	PI 588880 14	USDA-ARS, Cornell, Geneva, USA
8	M. Domestica	Pink Pearl	PI 588980 10	USDA-ARS, Cornell, Geneva, USA
9	M. Domestica	Clivia	PI 344550 06	USDA-ARS, Cornell, Geneva, USA
10	M. Domestica	Fuji Red spot type 2	PI 588844 15	USDA-ARS, Cornell, Geneva, USA
11	M. Domestica	Mcintosh wijcik		FEM-IASMA collection
12	M. Domestica	Golden Deliscious		FEM-IASMA collection
13	M. Domestica	Florina	PI 588747.18	USDA-ARS, Cornell, Geneva, USA
14	M. Floribunda		PI 589827.11	USDA-ARS, Cornell, Geneva, USA
15	M. Fusca		PI 589975.09	USDA-ARS, Cornell, Geneva, USA
16	M. Halliana		PI 589972.13	USDA-ARS, Cornell, Geneva, USA
17	M. Honanensis		PI 589879.08	USDA-ARS, Cornell, Geneva, USA
18	M. Hupehensis			DCA-UNIBO, Bologna, Italy
19	M. Ioensis	Texana	PI 596279.a8	USDA-ARS, Cornell, Geneva, USA
20	M. Kansuensis		PI 594097.S10	USDA-ARS, Cornell, Geneva, USA
21	M. Kirghisorum		PI 590043.13	USDA-ARS, Cornell, Geneva, USA
22	M. Mandshurica	6114	PI589353.06	USDA-ARS, Cornell, Geneva, USA
23	M. Micromalus		PI 589753.11	USDA-ARS, Cornell, Geneva, USA
24	M. Orientalis	B9	PI 594101.A8	USDA-ARS, Cornell, Geneva, USA
25	M. Pratti		PI 588933.12	USDA-ARS, Cornell, Geneva, USA
26	M. Prunifolia			DCA-UNIBO, Bologna, Italy
27	M. Pumila		PI 594106.25	USDA-ARS, Cornell, Geneva, USA
28	M. Rockii		PI 589421.15	USDA-ARS, Cornell, Geneva, USA
29	M. Sargentii	Rosea	PI 588919.06	USDA-ARS, Cornell, Geneva, USA
30	M. Sieboldii	Toringo	PI 589749.10	USDA-ARS, Cornell, Geneva, USA
31	M. Sieversii 1	KAZ 951806	GMAL 3685	USDA-ARS, Cornell, Geneva, USA
32	M. Sieversii 2	KAZ 950604	GMAL 3610	USDA-ARS, Cornell, Geneva, USA
33	M. Sieversii 3	Turkmenorum union	PI 594104.09	USDA-ARS, Cornell, Geneva, USA
34	M. Sieversii 4	KAZ 950705	GMAL 3619	USDA-ARS, Cornell, Geneva, USA
35	M. Sieversii 5	KAZ 951006F	GMAL 3638	USDA-ARS, Cornell, Geneva, USA
36	M. Sikkimensis		PI 589390.14	USDA-ARS, Cornell, Geneva, USA
37	M. Spectabilis	Aboplena	PI 588921 10	USDA-ARS, Cornell, Geneva, USA
38	M. Sylvestris 1	Pollino		Gallerati V.Monte Pollino, Calabria, Italy
39	M. Sylvestris 2		PI369855	USDA-ARS, Cornell, Geneva, USA
40	M. Sylvestris 3	Oberwartha 5xKlipphausen	GMAL 4495.K	USDA-ARS, Cornell, Geneva, USA
41	M. Sylvestris 4	Barenhecke 3xKlipphausen	GMAL 4497.B	USDA-ARS, Cornell, Geneva, USA
42	M. Sylvestris 5		PI 633826	USDA-ARS, Cornell, Geneva, USA
43	M. Transitoria	CH97 02-03 China	PI 633805 A2	USDA-ARS, Cornell, Geneva, USA
44	M. Xantocarpa		PI 589832.18	USDA-ARS, Cornell, Geneva, USA
45	M. Yunnanensis	Vilmorin China	PI 271831 13	USDA-ARS, Cornell, Geneva, USA
46	M. Zhaojiaoensis	CH97 06-07 China	PI 633817 A2	USDA-ARS, Cornell, Geneva, USA

3 Automated removal of noisy data in phylogenomic analyses[1]

Abstract

Noisy data, especially in combination with misalignment and model misspecification can have an adverse effect on phylogeny reconstruction; however, effective methods to identify such data are few. One particularly important class of noisy data is saturated positions. To avoid potential errors related to saturation in phylogenomic analyses, we present an automated procedure involving the step-wise removal of the most variable positions in a given data set coupled with a stopping criterion derived from correlation analyses of pairwise ML distances calculated from the deleted (saturated) and the remaining (conserved) subsets of the alignment. Through a comparison with existing methods, we demonstrate both the effectiveness of our proposed procedure for identifying noisy data and the effect of the removal of such data using a well-publicized case study involving placental mammals. At the least, our procedure will identify data sets requiring greater data exploration and we recommend its use to investigate the effect on phylogenetic analyses of removing subsets of variable positions exhibiting weak or no correlation to the rest of the alignment. However, we would argue that this procedure, by identifying and removing noisy data, facilitates the construction of more accurate phylogenies by, for example, ameliorating potential long-branch attraction artifacts.

Keywords: noise reduction, saturation, long-branch attraction, model testing, noisy data, placental mammals, Rodentia

[1] Original manuscript: Goremykin VV, Nikiforova SV, Bininda-Emonds ORP (2010) Automated removal of noisy data in phylogenomic analysis. Journal of Molecular Evolution71: 319-331. Own contribution: 30% (analysis)

Introduction

Noise reduction

It has been long appreciated that multiple substitutions per site (e.g., as derived from an elevated rate of molecular evolution), can impede phylogeny reconstruction because they are indistinguishable from historical signal by phylogeny reconstruction methods (Olsen, 1987). The mere presence of noise in the data does not automatically mean that the tree is wrong, however, when the substitution model fails to compensate for the high level of saturation, the quality of topology inference decreases.

Noisy sites are one important source of long branch attraction (LBA) (Felsenstein, 1978), a tree-building artifact that causes clustering of the unrelated, fast mutating species on the tree. Though initially thought to affect only maximum parsimony, LBA was later shown to also affect model-based methods under certain conditions (e.g., model misspecification) (e.g. Pol and Siddal, 2001). Highly variable sites are especially detrimental when elucidating deeper phylogenetic relationships because they mostly add noise to the data over the generally longer time frames involved (Gribaldo and Philippe, 2002).

Several explicit methods for the identification and removal of noisy alignment positions have already been suggested (e.g., Brinkmann and Philippe, 1999; Hirt et al., 1999; Lopez et al., 1999; Ruiz-Trillo et al., 1999; Burleigh and Mathews, 2004; Pisani, 2004; Kostka et al., 2008). From the practical point of view, these methods can be divided into two categories: tree-dependent and tree-independent methods.

Tree-dependent methods, which are the more popular of the two types, were first suggested as a tool in molecular phylogenetics in 1999. The maximum likelihood (ML) -based method by Ruiz-Trillo et al. (1999) and Hirt et al. (1999) involves sorting characters according to their gamma rate assignment by the Tree-Puzzle program, (which uses a neighbor-joining (NJ) tree to assign rates to characters) and the deletion of those sites with the most variable rates. Later, Burleigh and Mathews (2004) used a similar approach, but instead used a maximum parsimony (MP) tree to assign rates to characters. These methods do not require any predefined input tree, but tree-building remains an integral part of the sorting procedure. Thus, their effectiveness depends on the outcome of the tree search through which rates are assigned to positions. All LBA-related artifacts present on the tree can be expected to contribute to the inaccurate estimation of the rates of the sites and ultimately to biasing the results of the subsequent site-stripping. Rodriguez-Espelenta et al. (2007) observed that the tree topology used in the gamma-rate based approach affected the degree to which removal of variability helped to recover the predefined benchmark clades.

Another method, the Slow-Fast approach suggested by Brinkmann and Philippe (1999) and Lopez et al. (1999), does require a pre-determined tree topology to estimate the number of changes for each position within a tree using MP (position-based tree lengths). Thereafter, the position classes with the highest the number of changes are iteratively removed from the data, creating a series of the alignment subsets with the reduced variability. This approach also allows selecting monophyletic clusters (subtrees) on the input tree, thereby escaping the influence of individual branches, which would be suspect in case LBA is present. In this case, the sum of the subtree lengths at every position is used as a proxy for the substitution rate. This taxon set partitioning is implemented in the SlowFaster program (Kostka et al., 2008) and should, in theory, represent an improvement over the input of a complete tree. Even so, the method remains topology-dependent because omitting some branches does not liberate the results from any potential incorrect order of branching within the chosen subtrees.

By contrast, the rationale underlying development of the compatibility method (Pisani, 2004; Pisani et al., 2006) was rooted in the observation that because the length of a character in a MP sense is topology-dependent, erroneous tree topologies can lead to character lengths being miscalculated and thus the misidentification of putatively fast evolving sites by the Slow-Fast method. Instead, the method of Pisani (2004) is based on the compatibility principle by Le Quesne (1969), namely, that two characters are compatible if they can be mapped on a tree without homoplasy. Practically, the method involves calculating the number of other positions with which each site in the alignment is incompatible. High incompatibility scores are used to identify and to discard fast-evolving sites. The method is based on the expectation that the correct tree will derive from the largest number of compatible characters (which are all assumed to be uniquely derived) and that noisy characters will be incompatible with this set and can be discarded. (In other words, sites that contain phylogenetic information are expected to show fewer incompatibilities than those containing little or no phylogenetic information.)

Despite the availability of several methods of character-stripping, the removal of noisy data has not yet found routine use in analyses of genome-scale alignments that contain saturated sites (e.g. Springer et al., 2001; Goremykin et al., 2009b). This is proximally so because the relative capacity of the character-stripping methods in solving real problems of systematics is still not well-investigated, and ultimately so because the practical problem of finding an objective stopping criterion for character removal remains unsolved. As perhaps first acknowledged by Pisani (2004), “All methods of character selection pose the problem of finding an optimal cut-off value under which characters should not be deleted. How to discriminate characters the deletion of which could improve phylogenetic accuracy, therefore, is key. Still, this is the most complex step of any character selection protocol.” In seeking to address this issue, Pisani (2004) listed the significant and systematic deterioration of the results (as indicated by the appearance of obviously nonsensical clusters and/or substantial loss of resolution, support, or decrease in the likelihood of the recovered trees) as the key factor to be considered when deciding to stop the character-stripping process; he also suggested testing for the presence of the clustering signal among the deleted characters. Yet no particular guidelines for decision-making (e.g. which concrete stage of the result deterioration should be considered significant? How substantial the loss of support should be?) were suggested in this or the follow-up study (Sperling et al., 2009).

Indeed implementing some of the above criteria is difficult. For instance, relying on a decrease in the support for the tree branches as a criterion requires not only some threshold specification, but also the knowledge of which branches are correct and which are not (since, obviously, decrease of support for LBA artifacts is a good sign). Likewise, at least partial knowledge of the true tree is required to define “nonsensical” clusters. Finally, a decrease in the likelihood scores computed based on the different models and different (shortened) data may be indicative of several factors, not only of the degradation of phylogenetic signal in the data.

We therefore sought to build on the previous methods by designing an automated procedure of noise removal and evaluation of results that also provides a stopping criterion for the removal of the characters. Crucial to our method is that both steps are tree-independent, thereby avoiding the potential errors noted above. We test our method using genome-scale data (which are affected by saturation and long-branch attraction artifacts) for a well-publicized test case: the hypothesis of rodent polyphyly. The issue of the basal phylogeny of placental mammals caused an intensive discussion in the recent past. It is generally held that support for the hypothesis of basal rodent polyphyly derived in part from noisy data and the hypothesis is largely discredited today. The case therefore provides an ideal proof-of-concept for our proposed method. From there we go on to compare the effectiveness of our proposed method with several of the methods introduced above.

The test case: rodent polyphyly and relationships among placental mammals

The hypothesis of rodent polyphyly, first postulated by Graur et al. (1991) on the basis of MP analyses of nuclear protein sequences is a well-publicized issue of apparent long branch attraction artefacts in phylogenetics. The tree obtained by Graur et al. suggested that the guinea pig (and, by extension, other caviomorph rodents) diverged earlier in the eutherian radiation than did the remaining rodents. This hypothesis received strong criticism at the time from both classical morphologists (e.g. Luckett and Hartenberger, 1993) and other molecular phylogeneticists, who suggested the use of either ML-based analyses (e.g. Hasegawa et al., 1992; Cao et al., 1994) or less problematic data (e.g. Allard et al., 1991). Nonetheless, numerous molecular studies supported the hypothesis of rodent polyphyly for nearly a decade, with either Caviomorpha (Li et al., 1992; Graur et al., 1992; D'Erchia et al., 1996) or Myomorpha (mouse, rats, and allies; Ma et al., 1993; Janke et al., 1997; Phillips et al., 2001; Reyes et al., 1998, 2000b) being identified as having diverged earlier than the remaining rodents, and often as one of the earliest diverging branches among placental mammals.

Analyses of nuclear genes with increased taxon and character sampling (Madsen et al., 2001; Murphy et al., 2001a; 2001b; Amrine-Madsen et al., 2003; de Jong et al., 2003), however, have laid the foundation for the currently held view on eutherian evolution (see Fig 1b for a summary; see also Springer et al., 2004) and the virtual abandonment of the rodent polyphyly hypothesis. Here, a monophyletic Rodentia that as sister to Lagomorpha (rabbits and pikas) forms Glires is favored. Glires together with Euarchonta (containing Primates among several other orders) comprise the superorder Euarchontoglires, which, in turn, is sister to the superorder Laurasiatheria (together forming Boreoeutheria). The two remaining superorders, Afrotheria and Xenarthra, are often viewed as more basal lines in eutherian evolution (e.g., Springer et al., 2004), although the placement of the root of the placental tree remains hotly debated (see Prasad et al., 2008 and references therein).

However, analyses of nuclear genes and of complete mitochondrial genomes remain in conflict, even in the face of increased taxon sampling and the use of global GTR-based ML models for the latter. To date, mtDNA-based trees congruent to ones obtained using nuclear (nu) DNA data were achieved only by means of constraining the ML search space (Lin et al., 2002a), by creating a custom substitution model (Gibson et al., 2005), by discarding some mitochondrial genes from the analysis and eliminating synonymous leucine sites (Reyes et al., 2004), or by employing a mix of models to the customary chosen data partitions (Kjer and Honeycutt, 2007). By themselves, ML analyses of mitochondrial data employing global models from the GTR family (i.e. with or without correction for invariable sites and rate heterogeneity) tend to support rodents being basal or near-basal and polyphyletic (Reyes et al., 1998, 2000a, b; Mouchaty et al., 2001; Arnason et al., 2002; Lin et al., 2002b). In investigating the utility of mtDNA data for inferring deep nodes in the mammalian radiation, Springer et al. (2001) made specific reference to saturation in mtDNA data as a factor causing incongruence between mitochondrial and nuclear-based analyses. Saturation in mtDNA data was also noted by several other authors (e.g.

Pesole et al., 1999; Phillips and Penny, 2003; da Fonseka et al., 2008).

Together, these results indicate that rodent polyphyly is likely an artefact deriving from superimposed mutations in mitochondrial genomic data, thereby providing an ideal test case for our method. If mammalian mtDNA data are indeed more prone to saturation effects than are nuDNA due to their generally higher rate of substitution (see Bininda-Emonds, 2007), then the step-wise removal of the most variable characters from an alignment of complete mitochondrial genomes should yield both a monophyletic Rodentia and a topology for placental mammals that is more consistent globally with the currently accepted view.

These data, therefore, provide an important proof of concept for our suggested methodology of character-sorting based on the observed variability. Using sparse taxon sampling characteristic of the data sets from the time of the above discussion and choosing distant outgroups subtended by long branches we re-created the LBA-affected tree (with Rodentia and Lagomorpha appearing at the base of the subtree of placental mammals). Knowledge of the "true tree" of the placentals allowed us to compare the performance of the different character-sorting methods on the basis of the recovery of benchmark clades in this test case.

Materials and Methods

Description of the test procedure

Overview

The complete test procedure we describe here is implemented in the Perl script NoiseReductor.pl. During its execution, the script minimally invokes PAUP* for UNIX (Swofford, 2002), and one of two additional Perl scripts depending on the chosen character sorting method. In addition to the tree-independent criterion of the observed variability (OV) or some other supplied measure of the substitution rate (calculated by sorter.pl), a second script (gamma_sorter.pl) can sort the characters according to their gamma rates. Both gamma_sorter.pl and NoiseReductor.pl will invoke ModelTest (Posada and Crandall, 1998) as needed to determine the optimal model of evolution. All three Perl scripts are available upon request.

The procedure consists of two main steps. In the first step, the desired measure of character variability is calculated and then the alignment is sorted accordingly. Alternatively, one can put a text file named "sortme_so" in the current directory, which contains positive numbers, one number per line, that correspond to the desired approximation of the substitution rate for the alignment positions (the first line in the file should have the proxy of the substitution rate for the first alignment position, the second line for the second position and so on). Such files, for instance, (*.plot) are automatically produced by the COMPASS program by Simon Harris (www.ncl.ac.uk/microbial_eukaryotes/downloads.html) in which compatibility-based sorting approach is implemented (Pisani, 2004; Sperling et al., 2009). If the file "sortme_so" is found in the current directory, the script sorter.pl will skip computation of OV, and will sort positions in the alignment by the measure of substitution rate provided in the file. In the second step, the most variable subsets of the alignment are removed iteratively, with two series of correlation analyses of pairwise distances between the same sets of species pairs being used to determine when character removal should cease. We now describe each step in more detail.

Character Sorting

The user can choose to sort positions in ascending order of their variability based on 1) their observed variability independent of any tree topology 2) the assignment of substitution rates provided in the input file “sortme_so” or 3) their largest contribution to a gamma rate category under the best ML model. Prior to the analysis, the user should specify a threshold value, either a certain gamma rate class or a certain value of the chosen proxy of substitution rate as needed, at which the shortening process should be stopped to prevent the removal of constant partitions or those of minimal variability.

To calculate our sorting criterion (observed variability, OV), all sequences for a given position are compared in a pair-wise fashion. Mismatches are scored as 1 and matches as 0; the mean value among all the comparisons for a given position is used as the measure of character variability in the subsequent data sorting:

OV = sum(1.. k){dij} / k.

Here k is the number of pair-wise comparisons made for a given position and d_ij is the score of character variability in each pair-wise comparison made (can be either 0 or 1). If n is the number of aligned sequences which do not have a gap at the given alignment position, than k = $(n^2-n)/2$. The OV measure is actually similar to PI (average pair-wise differences) from population genetics, which is calculated by taking all pairs of individuals in a population and computing the average number of differences between them.

Thus, observed variability (like compatibility scores) is calculated without reference to any specific tree and are free from any systematic bias in the estimation of the substitution rates for all sites in alignment that a wrong input topology might cause. At the same time, an obvious drawback of the method is that it does not incorporate any fitted substitution model, something that has been shown repeatedly to improve the accuracy of phylogenetic inference (e. g. Gadagkar and Kumar, 2005; Gaucher and Miyamoto, 2005). Indeed, sorting based on the observed variability actually would yield the same results as the sorting based on the native Jukes and Cantor model, applied to a single alignment position (because the latter is proportional to the scope of the observed sequence dissimilarity), which was previously observed to be among the worst ML models describing character distribution in non-simulated genome-scale data (Goremykin and Hellwig, 2006).

Thus, NoiseReductor.pl also includes a gamma-rates option to increase the range of situations in which the ML-based sorting methods of Ruiz-Trillo et al. (1999) and Hirt et al. (1999) are useful by employing more realistic trees to assign rates to characters. When sorting by gamma category, the settings of the best model are first determined via gamma_sorter.pl using the standard ModelTest procedure (i.e. using a NJ tree built from uncorrected distances to fit the ML model parameters) based on the Akaike information criterion. Optionally, gamma_sorter.pl can further refine the model by having PAUP* compute the ML tree using the settings of the above model, and then using this ML tree as the base tree for a second round of model fitting via ModelTest. Indeed, we suggest to not to use NJ tree for assignment of rates to characters (as is the case in the above studies), but to use it as the foundation for the second round of model fitting, which should make the results of character-sorting less LBA-prone. Using the combination of the best model (chosen either using the NJ or ML tree as the base tree for calculation) and the ML tree obtained using this model, PAUP* is then used to assign sites to up to 100 gamma-rate categories as a prelude to sorting the positions in ascending order of their variability. (We have observed that applying 100 rate categories occasionally causes PAUP* to crash because of memory allocation problems. In such cases, gamma_sorter.pl automatically reduces the number of categories by one until a result is achieved.) The user can also optionally perform the model selection procedure under topological constraints as specified in a separate tree file.

Naturally, if the optimal model does not indicate the need for gamma correction, sorting by gamma rate category should not be done. Empirically, however, we have yet to observe a single case where a correction for rate heterogeneity was not recommended by ModelTest for a large genome-scale alignment.

Finally, with regard to the OV-based sorting, it should be noted that sites which support internal splits among large groups of species will tend to have higher OV scores than sites that support splits of few versus many species. This will have the effect of reducing internal branch lengths disproportionately compared to external branch lengths. Thus if OV-based noise removal is applied for doing dating estimates, it might be advisable to use a two-step procedure where the noise-reduced alignment is used to get the topology and the complete alignment is used to then estimate the branch lengths.

Correlation analyses

Regardless of the model-fitting and sorting procedure, the sorted alignment is written to a new file (the master file, A_0) that is subjected to iterative rounds of shortening by repeatedly removing a user-specified number of positions from the most divergent end. At each shortening step (n), three data partitions are written to disk: the shortened alignment (partition A_n), all deleted positions (partition B_n) and only those positions removed in the current shortening step (partition C_n). Note that partition C_n is a subset of partition B_n, and that, after the first shortening step (n = 1), partitions B_1 and C_1 are identical.

At each step of the shortening procedure, NoiseReductor.pl then fits the optimal ML model to each of the three partitions according to the same procedure used for the initial model fitting above (i.e., using either the NJ tree or the ML tree). We have observed that execution of some of the models suggested by ModelTest for the most divergent data partitions can occasionally cause PAUP* to crash. In such cases, NoiseReductor.pl automatically modifies the model in an attempt to derive a result: first, for any rate assigned a value of 0.0000 in the substitution rate matrix (such values cause crashes with lset command) numbers from 1 to 9 are added to the fifth decimal place; second, if all rates are higher then 0, numbers from 1 to 9 are added to the fifth decimal place of the first substitution rate in the matrix (in the case of the GTR model, this would be the A-C rate); and third, if the rates are equal (and absent in the model specification in PAUP format), numbers from 1 to 9 are added to the fifth decimal position of the first base frequency given in the model specification. If all the above modifications fail to deliver a result, or the model description contains no base frequencies and no rate parameters (which is the case if the base frequencies are in equilibrium and all the rates are equal), then all the above modifications are tried with the optimal model found on the basis of hierarchical likelihood ratio tests. Finally, if this also fails, the script attempts to fit a GTR+I+G model to the data using the lscore command.

For each fitted model, a distance matrix for each data partition (A_n, B_n, and C_n) is determined using PAUP* at each shortening step. (When model fitting is done using the refined two-step procedure above, two matrices are determined, one for each of the models.) These distance matrices form the basis of a pair of correlation analyses between the same taxa in the matrix derived from partition A_n and those in the each of the matrices derived from B_n and C_n to determine when the noisiest sites have been removed from the alignment such that the results based on the remaining data (A_n) are unlikely to be misled significantly by saturation. When model-fitting is done using the refined procedure, the script first compares the matrices determined in the first round of model-fitting, as outlined above, and then it compares the matrices determined in the second round of model fitting. All correlation analyses use both Pearson product-moment correlation coefficients (r) and Spearman's rank correlation coefficients (ρ) based on the absolute distance values and their relative ranks, respectively, within each analogous matrix.

These correlation analyses form the basis of a stopping criterion indicating when the noisy sites have largely been removed from the main data partition. In the ideal case where the model and its settings describe all the data effectively, strong positive correlation will be present between the distances between the same sets of species pairs estimated from each partition because the correction for superimposed mutations works effectively for both data partitions. The correlation would also be relatively linear, and not a curve as for the plot of corrected vs. uncorrected distances.

By contrast, the distribution of distance values should become cloud-like when the model fails to describe the data adequately (e.g. at high levels of saturation when the variance in the distance estimation exceeds the distance values themselves). High levels of saturation will lead to a loss of exactness in the prediction of evolutionary distances because of the higher variances in the estimates of those distances. Thus, weak correlations in the relative ranking of distance estimates for the same species pairs between the conserved partition and in each of the variable alignment partitions B_n and C_n would indicate that the chosen model is not compensating adequately for superimposed mutations within the latter partitions. (Because weak correlations can also derive from poor model choice, optimal model settings are determined by Noise Reductor.pl via ModelTest each time prior to distance estimation, thereby minimizing the chances of severe model misspecification.) As such, the strength of the correlation should improve with the continued removal of noisy sites. In other words, the absence of strong correlation is interpreted as the failure of the fitted ML model to accommodate multiple substitutions (and the associated non-phylogenetic signals) that remain in A_n and, consequently, as an argument for the discarding and continued removal of the most variable partitions. It is only when all correlation analyses at a given shortening step start to yield strong positive values that the level of saturated data in partition A_n has been reduced to a degree where the tree derived from these data would be largely unaffected by the artifacts from noisy data and could be viewed as robust in the sense of our test.

In addition, NoiseReductor.pl also provides the means to estimate the absolute scope of saturation in the variable partitions B_n, and C_n as further information for the user. These estimates are based on either a correlation analysis (both r and ρ values) between the ML and uncorrected *p*-distances estimated from the variable partitions or the deviations between the mean values of each of the ML and *p*-distances calculated from these partitions.

Additional features

Although the Perl scripts are designed primarily to identify noisy positions in a DNA alignment, they can also be used to help facilitate standard ML analysis using PAUP*. For instance, if character sorting via gamma_sorter.pl is not invoked, the two-step model fitting procedure can be used to automate the process of determining the ML tree under the optimal model of evolution. In addition, because all model fitting and tree building analyses for all partitions can be automatically run in parallel by sorter.pl, the scripts will benefit from being run on machines with multiple CPUs or CPU cores. gamma_sorter.pl can also be used to perform a non-parametric bootstrap analysis (Felsenstein, 1985) of the original data, with the script starting a user-defined number of PAUP* processes in parallel, each computing a tree for a single bootstrap replicate. The ability of the scripts to automatically make use of multiple processors and cores as far as possible also counteracts the computationally intensive nature of the method, which derives in part from the numerous rounds of model-fitting (i.e. after each shortening step).

Assembly of the data set for the tests

The nucleotide data set was obtained from GenBank in the form of complete mitochondrial genomes for 55 mammalian species (49 eutherians, three marsupials and three monotremes). Alignment was performed initially using ClustalW and subsequently improved by eye, both using the Seaview sequence editor (Galtier et al., 1996). The final alignment length was 11549 bp after elimination of all regions of insecure alignment as determined by visual inspection.

All model-fitting analyses used the two-step model outlined above, and all correlation analyses were based on shortening steps of 200 bp. In testing the efficacy of our procedure, we focused in part on four benchmark clades: a monophyletic Rodentia, Glires, Euarchontoglires (= Glires + Primates here), and Boreoeutheria.

Proxies of the positional substitution rates were also determined used compatibility-based criteria (Pisani, 2004; Sperling, Peterson and Pisani, 2009) implemented in COMPASS program by Simon Harris (www.ncl.ac.uk/microbial_eukaryotes/downloads.html): 1) Le Quesne probability (LQP) values (Le Quesne, 1969), and 2) direct compatibility.

Results and Discussion

Failure of the optimal global ML model to identify the true tree

ML analyses of the complete mitochondrial data set did not support rodent monophyly, nor did they recover any of the remaining benchmark clades (Fig 3-1a). The tree shown was built from 9549 bp long A10 subset. Trees built from A11 and A12 subsets have the same topology. The bootstrap numbers in the trees were produced by PHYML, using ML start trees (built employing fitted models, GTR+I+G, with the help of PAUP*) and using general specification of GTR+I+G model.

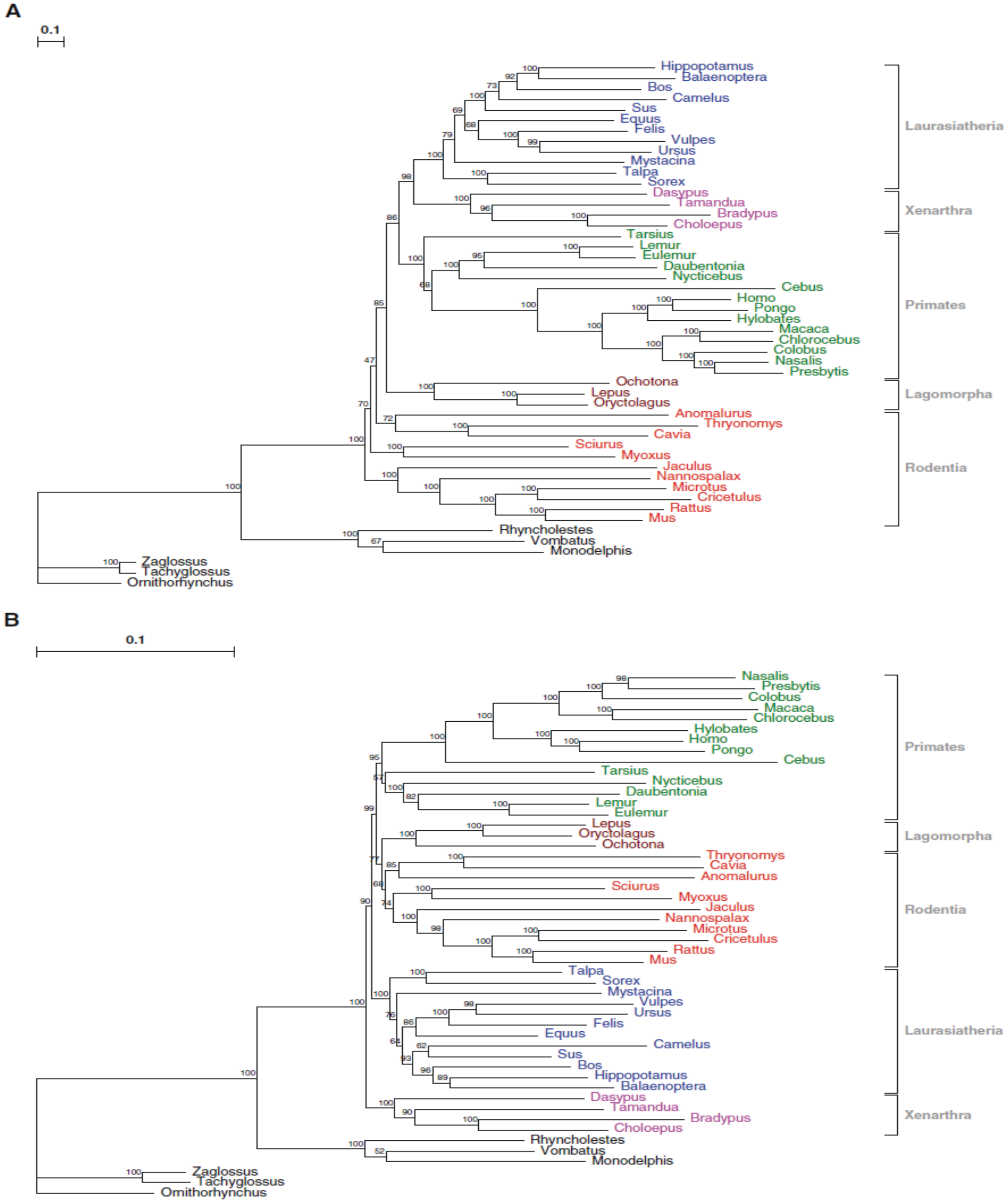


Fig. 3-1. Representative changes in tree topology as a result of the removal of the most divergent characters in multigene alignments for mammals (mtDNA). A) Topology obtained from the ML analysis of the complete mammalian mtDNA alignment (A_0). B) Topology obtained immediately after the sharp rise in r values in correlation analyses of the conserved (A_n) and each of the variable partitions (B_n and C_n) (see Fig. 3-2).

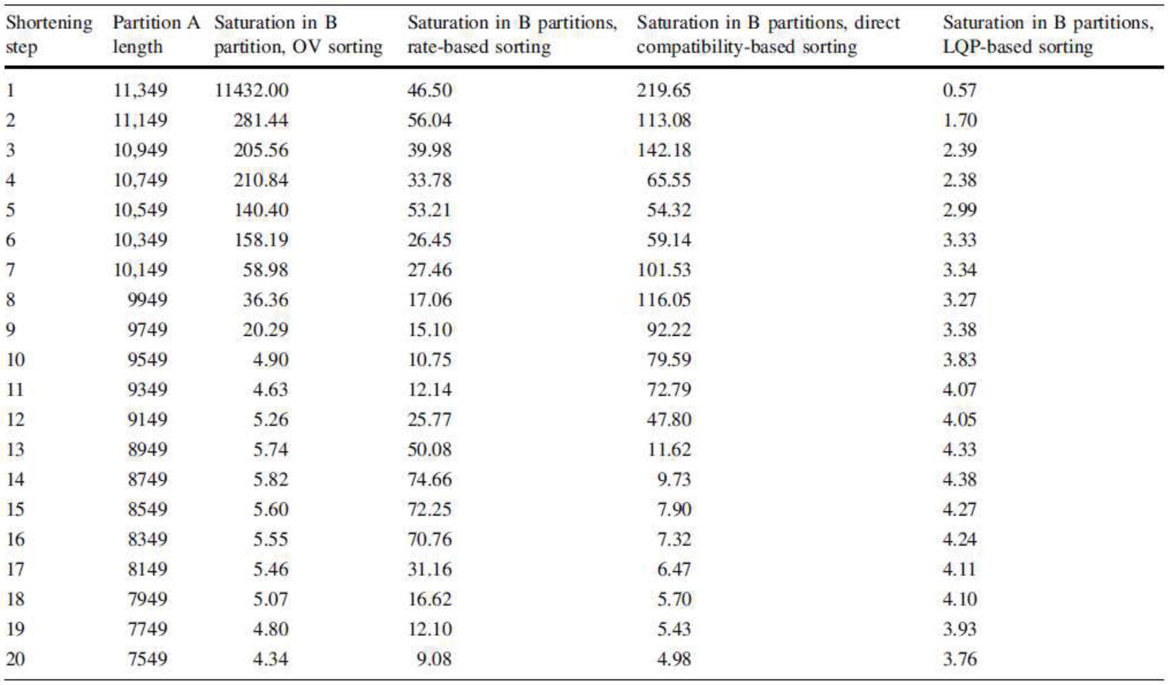
Instead, rodent polyphyly was indicated, with lagomorphs and caviomorph and murid rodents forming successively more distant sister groups to the remaining placentals. Bootstrap support for the relevant nodes was comparatively weak (85%, 47%, and 70%, respectively). In addition, the widely accepted monophyly of the benchmark clade Boreoeutheria was contradicted with Xenarthra being strongly supported as sister to Laurasiatheria (bootstrap support = 98%). These general results were true even with the use of the two-step model fitting procedure, which should provide a better choice of the substitution model. We therefore applied our procedure to identify noisy positions and to assess if their removal led to the recovery of our four benchmark clades. By doing so, we also tested the performance of our procedure with regard to 1) efficiency of character sorting using OV vs. assignment to gamma rate categories, Slow-Fast-based sorting and compatibility-based sorting and 2) determination of a stopping criterion for the removal of noisy data.

Efficiency of character sorting

The best character-sorting approach should yield the highest concentration of saturated positions at the most divergent end of the sorted alignment. Estimates of the scope of saturation based on the deviation of the mean values of the corrected ML distances from those of the uncorrected *p*-distances (see above) in the variable partitions B_n revealed that OV-based sorting outperformed gamma-based sorting and compatibility-based sorting in all analyses (Table 3-1). (These conclusions could not be verified with the Slow-Faster program because it does not create variable data partitions).

OV-based sorting eliminated the basal rodent paraphyly artifact and recovered Euarchontoglires at the seventh shortening step (1400 positions deleted). This was much faster than all other methods tested. Neither the gamma-rate based sorting nor the two compatibility-based methods tested recovered a single benchmark clade at this or the next four subsequent shortening steps. The Slow-Faster program also did not recover a single benchmark clade within this shortening range (up to 2200 deleted positions) even when the input tree (unrooted ML tree shown in Fig. 3-1b) was correct and Primates, Glires, Laurasiatheria, Xenarthra and the outgroop (the branch subtending monotremes plus marsupials on the above tree) were specified as input taxon partitions.

Table 3-1. Relative prowess of character sorting based on observed variability, gamma-rate categories, direct compatibility scores (as implemented in the program COMPASS) and Le Quesne probability (as implemented in the program COMPASS) in concentrating saturated positions towards the most variable end of the sorted alignment. Saturation in the variable data partitions was estimated using deviations of mean uncorrected distances from the mean evolutionary distances among all terminal taxa calculated using optimally-fitted ML models.



Shortening step	Partition A length	Saturation in B partition, OV sorting	Saturation in B partitions, rate-based sorting	Saturation in B partitions, direct compatibility-based sorting	Saturation in B partitions, LQP-based sorting
1	11,349	11432.00	46.50	219.65	0.57
2	11,149	281.44	56.04	113.08	1.70
3	10,949	205.56	39.98	142.18	2.39
4	10,749	210.84	33.78	65.55	2.38
5	10,549	140.40	53.21	54.32	2.99
6	10,349	158.19	26.45	59.14	3.33
7	10,149	58.98	27.46	101.53	3.34
8	9949	36.36	17.06	116.05	3.27
9	9749	20.29	15.10	92.22	3.38
10	9549	4.90	10.75	79.59	3.83
11	9349	4.63	12.14	72.79	4.07
12	9149	5.26	25.77	47.80	4.05
13	8949	5.74	50.08	11.62	4.33
14	8749	5.82	74.66	9.73	4.38
15	8549	5.60	72.25	7.90	4.27
16	8349	5.55	70.76	7.32	4.24
17	8149	5.46	31.16	6.47	4.11
18	7949	5.07	16.62	5.70	4.10
19	7749	4.80	12.10	5.43	3.93
20	7549	4.34	9.08	4.98	3.76

Similarly, sorting and deletion based on OV also recovered all four benchmark clades faster than all other methods tested (9 shortening steps, 1800 positions removed). Gamma-based sorting required 13 shortening steps (2600 positions deleted) to recover the four clades as did COMPASS with the direct compatibility option turned on. When LQP values were used instead, an additional three steps were required (3200 positions removed). The Slow-Faster program with the correct tree divided onto five input subtrees as described above yielded a tree with these clades when 2629 positions were deleted.

It should be mentioned that when we used the (wrong) tree shown in Fig. 3-1a, as input for the Slow-Faster program, and specified i) the branch subtending monotremes plus marsupials plus Glires and ii) the branch subtending Xenarthra plus Laurasiatheria as input subtrees, a tree containing all four benchmark clades was not recovered at any shortening step. By contrast, gamma-rate based sorting with the same wrong input tree (for model specification) still managed to recover the benchmark clades and at about the same point as did COMPASS and Slow-Faster with the correctly specified input branches (see above). These results demonstrate the difficulty in appraising the validity of the Slow-Faster method in situations where the true tree is not known (which unfortunately, is exactly the situation when the method is needed) because the absence of topological changes during the shortening process could be misinterpreted as indicating a robust initial topology. Also, despite the valid arguments regarding importance of topology-independence testing by Pisani (2004), LQP based noise reduction as implemented in COMPASS faces the same problem.

Determination of a stopping criterion

Successive removal of the noisiest data from the mammalian data set (see above) yielded a tree that was increasingly congruent with currently accepted views on eutherian evolution: Xenarthra as sister to the remaining placentals and the monophyly of Rodentia, Glires, and Euarchontoglires (Fig. 3-1b). However, the recovery of such benchmark clades as a stopping criterion for character removal is generally unsuitable for two reasons. First, it can often be applied in only a few cases (e.g. when the path of evolution is definitely known from the fossil record). Second, benchmark clades, when available, are often chosen on the basis of their overall robustness, suggesting that their recovery might be more resistant to the negative effects of noisy data.

Thus, we examined if the correlation analyses of pairwise distances might provide an alternative solution to this problem of determining a stopping criterion for character removal. Specifically, we evaluated two approaches: a comparison of the corrected and uncorrected distances calculated from the variable partitions C_n and B_n (referred to hereafter as analysis 1), and a comparison of evolutionary distances calculated from the conserved A_n partition to those from the variable data partitions C_n and B_n (analysis 2). In so doing, we focused solely on OV-based sorting here based on its superior performance in i) recovering the benchmarks clades and ii) concentrating saturated positions towards the most variable end of A_0 compared to all other methods tested (see Table 1 and discussion above).

Intervals exist in both analyses 1 and 2 where r values rise sharply (ends marked with yellow lines in Fig. 3-2). These “break intervals” are short, never exceeding two shortening steps, and are observed at the same or immediately adjacent shortening step in analyses 1 and 2 (step 9 or 10, when 2000 bps ((C_n), results not shown) or 2200 bps ((B_n), Fig. 3-2) have been removed, respectively). Inspection of the variable alignment partitions (B_n and C_n) of removed positions from before the break intervals revealed that they contained positions that are absolutely dissimilar among certain OTUs, with pairwise ML distances being unreliable and unrealistically high (from 10 to ~20000 substitutions per site). At such high values, small changes in base composition can cause disproportionally large differences in ML distance values (results not shown). Examination of all distance matrices produced during the shortening process revealed that the break intervals occur after all such outliers are finally removed, in part because of the sensitivity of Pearson correlation coefficients to outliers. Thus, r values estimated in comparisons involving the most divergent alignment partitions behave erratically on the left side of the graphs and show greater similarity to the Spearman’s rank correlation coefficients on the right side after the removal of the outliers.

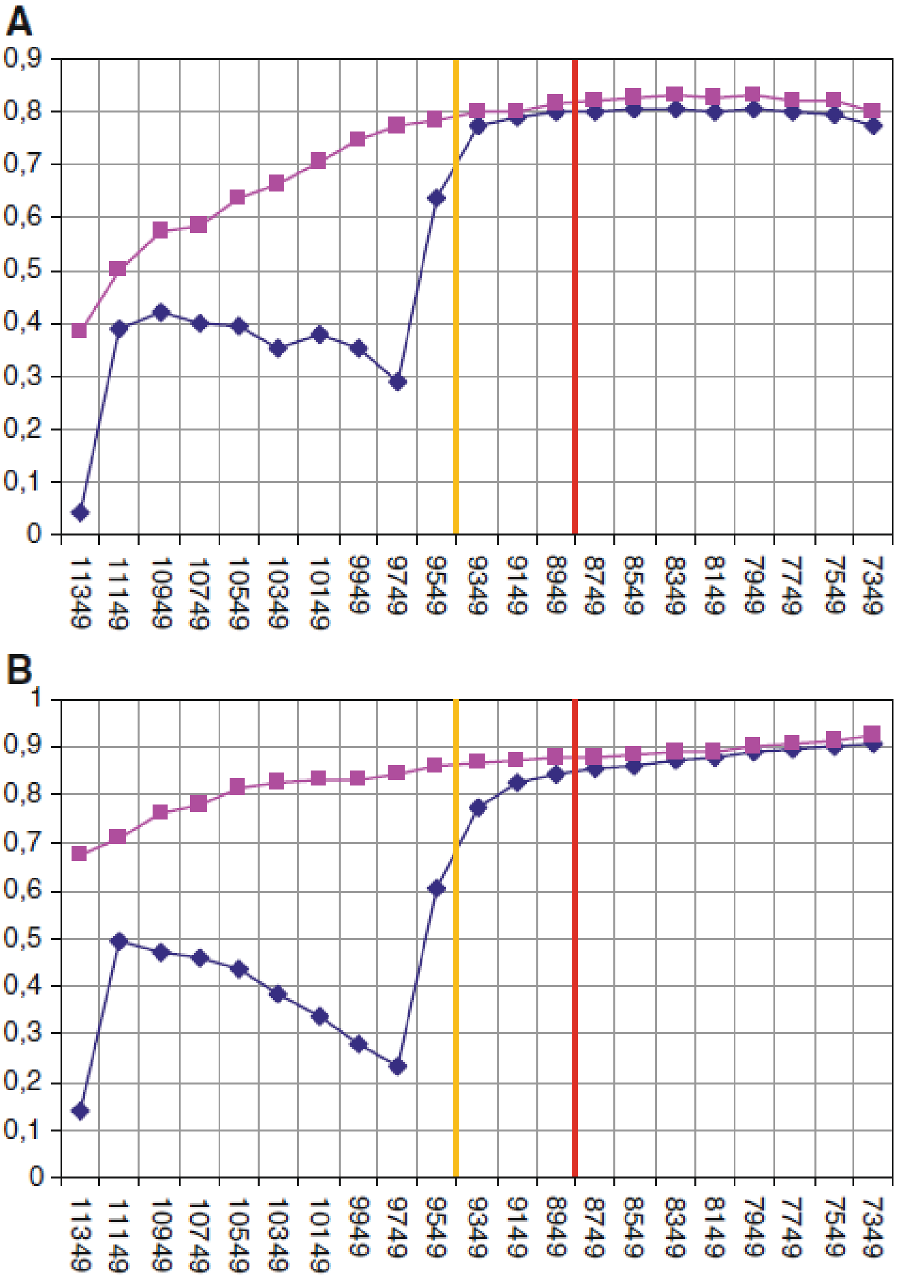


Fig. 3-2. Results of the correlation analyses involving pairwise comparisons between the conserved (A_n) and variable partitions (B_n). Pink and blue dotted lines on the graphs represent ρ and r values, respectively. Digits on the x-axis depict shortening steps (i.e. lengths of the remaining A_n partitions). Orange lines on the graphs denote the ends of the break intervals (see main text), whereas red lines denote points beyond which removal of variable positions caused the collapse of some tree branches. Graph A presents the results of the analysis 2 (see main text) and graph B presents the results of the analysis 1.

Empirically, all mtDNA-based trees calculated from the conserved alignment partition after the end of the break intervals but before the shortening step leading to the point of information loss (defined by the start of branch collapse leading to unresolved clusters on the tree; marked by red lines in Fig. 3-2) recover all benchmark clades, including a monophyletic

Rodentia. Indeed, all benchmark clades were recovered also before the break intervals (see above), indicating that the data set appeared to contain more residual noise than just the substitutional saturation associated with the rodents and surrounding clades. Examination of the first tree obtained after the break interval at 10th shortening step (Fig. 3-1b) reveals higher bootstrap support for all nodes along the backbone of the tree (all now 100%) than were obtained using the entire alignment. Support for Glires (bootstrap = 77%) and Rodentia (bootstrap = 68%) is admittedly relatively low, but still generally higher than for the conflicting nodes in Fig. 3-1a. Virtually all the trees obtained after the break interval but before the point of information loss are congruent to the topology shown in Fig. 3-1b. The exception is the last tree obtained at the 13th shortening step, where Tarsius was placed as the sister group to the remaining primates and lagomorphs were sister to the primates. This tree, however, might indicate the first signs of information loss in the matrix.

In general, our correlation-based stopping criterion only operates effectively when the sorting method is good and efficiently concentrates noisy data to the right-hand side of the sorted alignment. For example, in applying LQP-based sorting, which was the slowest in recovering the four benchmark clades (see above), we observed a steady growth of the correlation values without any breaking intervals (result not shown). This was because the sorting algorithm failed to efficiently concentrate saturated positions in the variable partitions, such that ML distance matrices based on these partitions did not include any outliers. This failure, in turn, would have lead to the erroneous conclusion that the tree built from A_0 partition could have been accepted in the sense of our correlation test based on the absence of a breaking interval in variability reduction process.

By contrast, OV-based character sorting does appear able to effectively identify and concentrate the most highly variable sites from genome-scale alignments, showing good performance compared to other sorting methods (see Table 3-1). Thus, we can expect the appearance of some extremely high distances in the variable matrices, with the consequence that failure of the optimal ML model to correct for superimposed substitutions in divergent data partitions will become apparent in the correlation analyses as break intervals (Fig. 3-2). We therefore suggest that the removal of variable positions from large-scale alignments should be continued at least until the very end of sharp rise in r values in any series (A or B) of correlation analyses, as these most variable alignment subsets contain the most misleading positions in the data. By contrast, lack of any break intervals during the shortening process with the OV method would indicate that the tree obtained from the original unsorted alignment data is not strongly biased by noise, and can be accepted in the sense of our test. This does not mean that the tree is correct; it means that the tree structure is not likely to be affected by one factor which might pose obstacles to correct phylogeny reconstruction - superimposed mutations. Other factors, related, for example, to the compositional heterogeneity, to the assumptions about time-reversibility of the substitution process implemented in the currently available substitution models, to the regional variation in rates of substitutions among different taxa, to concerted gene evolution, etc. might cause errors too and, thus, additional data exploration is always advisable.

Conclusions

To date, recommendations in the literature to improve the accuracy of phylogenetic analyses have focused largely on increased taxon and character sampling (including the reduction of missing data) and the application of increasingly more complex ML models. Although such steps are undoubtedly beneficial, they still might fail to counter the negative impact of non-phylogenetic signals, which, as Jeffroy et al. (2006) and Rodriguez-Espelenta et al. (2007) have pointed out, are associated with saturated positions. Our study highlights in part the potential impact of saturated positions in phylogenomic analyses. This is not a trivial issue given the growing use of unfiltered genomic-scale data to infer phylogenetic relationships. Phylogenetic analyses of angiosperms represent a prime example of this trend (e.g. Stefanović et al., 2004; Leebens-Mack et al., 2005; Qiu et al., 2005; Moore et al., 2007; Jansen et al., 2007). In fact, Stefanović et al. (2004) and Leebens-Mack et al. (2005) explicitly advocate the retention of all characters in their studies.

Building on previously suggested methods of noise-reduction, and on the novel method of nose reduction based on the observed sequence variability, we present here an automated procedure of noise removal and provide guidelines for its optimal use. At a minimum, the method provides means of data exploration for the end user with respect to the potential impact of noisy, saturated positions on the inferred phylogenetic hypotheses. Rate variability among nucleotide positions in the alignment can be estimated using any published measure. However, our results indicate that our suggested proxy of the substitution rate, observed variability, appear to be more effective at identifying the most saturated positions compared to other noise-reduction methods tested.

Importantly, our results point to the existence of an apparently objective stopping criterion, indicating the point at which the noisiest and most misleading positions have been removed from the alignment. The stopping criterion is designed to estimate the ability of the fitted ML model to compensate for multiple substitutions per site in light of the observation that the mere presence of saturation does not automatically impact negatively on likelihood-based phylogenetic analysis because ML can account for non-observed, superimposed substitutions. Instead, the potential for negative effects depends on whether or not the model applied can accommodate the given level of saturation. This can be observed by comparing the pairwise ML distances calculated from two partitions of the same alignment, one containing conserved characters and the other the variable characters. Absence of correlation between ML distances based on conserved and variable partitions indicates that correction for superimposed substitution does not work well in the variable partition. Strong positive correlation indicates that correction works well and character removal process may be stopped. Such analysis of evolutionary distances in alignment partitions have been used to measure and compare the degree of correlation between synonymous and non-synonymous substitutions in coding sequences (e.g. Goremykin et al., 1996, Jabbari et al., 2003). However, we are not aware of its application to guide the identification and potential removal of saturated positions.

This is not to say, however, that all noisy data are removed at the point where the stopping criterion is applied. Visual inspection of the scatterplots of the evolutionary distances vs. non-corrected distances in B_n partitions after application of the stopping criterion in our two examples revealed that the distribution of the dots is still curve-like, which is indicative of at least some degree of saturation in the data. But because variability in DNA sequence data is a continuum (from hyper variable to invariant sites), there is no method to define noisy sites per se. Indeed, our criterion might be held to be too liberal in that some noisy sites are retained. However, it does seem to point to a natural dichotomy in the data, where r values increase sharply indicating improved performance of ML model in compensating for multiple substitutions per site. Whether character removal beyond this point is justified cannot be answered on the basis of our tests. As variability in the data will often be increasingly reduced with subsequent character removal, correlation values will continue to increase to the point where the drawbacks from the loss of phylogenetic information outweigh any additional gain from the removal of saturated positions. Because the correlation analysis we suggest here provides estimates only for the removal of noise and not for the retention of phylogenetic signal, it can only be used to identify and eliminate that part of the alignment causing ML models to yield extremely unreliable estimates of the substitution process.

We believe that the scripts we present here can help to identify better supported hypotheses of phylogenetic relationships of species, and therefore, can enrich the means available to assess the reliability of phylogenetic analyses. As our analyses of the mammalian mtDNA data set showed, our method can arguably improve the net result of phylogeny reconstruction, especially for deeper nodes, where the inference of phylogeny is especially obscured by multiple substitutions and the resulting long-branch attraction. In many such cases, including among the basal angiosperms, the usual suggestion of breaking of long branches cannot be applied (e.g. for isolated taxa with extinct relatives, taxa that are not infrequently found at the base of the tree). Instead, coupled with more traditional approaches of adding more characters and species to the alignment, the assessment and potential removal of saturated positions should be a necessary procedure in helping to improve building of phylogenetic trees.

Acknowledgments:

O.B.E. was supported in part by the Deutsche Forschungsgemeinschaft (BI 825 / 3-2).

4 The evolutionary root of flowering plants[1]

Abstract

Correct rooting of the angiosperm radiation is both challenging and necessary for understanding the origins and evolution of physiological and phenotypic traits in flowering plants. The problem is known to be difficult due to the large genetic distance separating flowering plants from other seed plants and the sparse taxon sampling among basal angiosperms. Here we provide further evidence for concern over substitution model misspecification in analyses of chloroplast DNA sequences. We show that support for *Amborella* as the sole representative of the most basal angiosperm lineage is founded on sequence site patterns poorly described by time reversible substitution models. Improving the fit between sequence data and substitution model identifies *Trithuria*, Nymphaeales and *Amborella* as surviving relatives of the most basal lineage of flowering plants. This finding indicates that aquatic and herbaceous species dominate the earliest extant lineage of flowering plants.

Keywords: *Trithuria inconspicua*, chloroplast genome, angiosperm origins, heterotachy, base compositional heterogeneity, data model fit

[1] Original manuscript: Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, Delange P, Martin W, Woetzel S, Atherton RA, Mcleanachan T, Lockhart JP (2013) The evolutionary root of flowering plants. Systematic Biology 62: 50–61. Own contribution: 10% (analyses)

Introduction

While there is increasing consensus about many relationships among major lineages of flowering plants (Soltis et al., 2011) and convergence towards more similar dates for the origin of angiosperms (Sun et al., 2011; Jiao et al., 2011), determining the root of the angiosperm phylogeny has been more problematic. This difficulty is not unique to the study of angiosperms; reconstructing basal relationships in species radiations is known to be hard (Shavit et al., 2007; Graham and Iles, 2009). Not only can the shape of the true underlying phylogeny make it difficult to accurately reconstruct gene trees (Whitfield and Lockhart, 2007), even correct gene trees can be incongruent with the underlying species phylogeny (Degnan and Rosenberg, 2009).

In phylogenetic studies of chloroplast DNA (cpDNA), nuclear DNA (nuDNA), and mitochondrial DNA (mtDNA), *Amborella* has often been recovered as the sole survivor of the first lineage to diverge from that leading to all the other extant flowering plants (Soltis and Soltis, 2004; Soltis et al., 2011; Mathews and Donoghue, 1999; Saarela et al., 2007; Qiu et al., 1999; Graham and Iles, 2009; Stefanović et al., 2004; Leebens-Mack et al., 2005; Jansen et al., 2007). However, a closer relationship between *Amborella* and aquatic angiosperm species has been reported in analyses of mitochondrial and nuclear DNA (Soltis et al., 2011; Qiu et al., 2010; Jiao et al., 2011) as well as in model-based analyses of chloroplast genes that typically exclude or reduce the impact of third codon positions (Barkman et al., 2000; Wu et al., 2007). Opinion has been divided over how to treat third codon positions in cpDNA. While inclusion of these sites might improve phylogenetic resolution between some taxa (Stefanović et al., 2004; Leebens-Mack et al., 2005; Zanis et al., 2002), they also exhibit evidence of a decayed historical signal (due to multiple substitutions at the same site) between some taxa (Goremykin et al., 2003a; Chaw et al., 2004). Analyses of short independent nuclear markers have not provided improved phylogenetic resolution, suggesting instead alternative relationships among basal angiosperms (e.g., Mathews and Donoghue, 1999; Soltis et al., 2011; Jiao et al., 2011). This finding is perhaps not unexpected given the short internal branches typically reconstructed for angiosperm phylogenies (e.g., see Martin et al., 2005).

We have previously suggested that a poor fit between commonly used phylogenetic models and sequence data contributes to uncertainty concerning relationships among early diverging lineages of flowering plants (Martin et al., 2005; Lockhart and Penny, 2005). Here we provide further evidence for this hypothesis in a study of the substitution properties of concatenated chloroplast genome sequences, and in particular of sites in the alignment that are most varied. These sites, often called "fast sites", show the greatest character state variation as well as evidence of multiple substitutions. Numerous methods for sorting, identifying and removing fast sites have been suggested, and the impact of removal of the fastest evolving sites on phylogenetic reconstruction is well known (e.g., Brinkmann and Philippe, 1999; Hirt et al., 1999; Lopez et al., 1999; Ruiz-Trillo et al., 1999; Hansmann and Martin, 2000; Burleigh and Mathews, 2004; Pisani, 2004). Less well appreciated is the observation that the sorting of sites based on character state variation or compatibility criteria allows the properties of sites that impact on tree building to be more easily studied (Sperling et al., 2009). We have examined the compositional heterogeneity of fast sites and the fit of concatenated chloroplast sequences to the GTR + I + Γ substitution model commonly used in angiosperm phylogeny studies. We address the problem of identifying which of the fast sites to exclude from the phylogenetic data by applying the GNB criterion (named after the inventors: Goremykin et al., 2010) to the concatenated alignment after the sites in this alignment had been reordered according to their observed variability (OV: see Materials and Methods). This criterion has been suggested as suitable for identifying sites most affected by multiple substitutions in a multiple sequence alignment. Here, we examine the properties of the fast sites identified under the GNB criterion and the contribution of these sites to topological distortion in phylogenetic trees reconstructed for angiosperm and conifer sequences. To obtain optimal phylogenetic estimates, we employed the CAT + covarion model, which was consistently identified in our cross validation analyses as the best-fitting model to our original data and to data partitions generated in the "noise reduction" protocol of Goremykin et al. (2010). This substitution model better accommodates a restricted substitution profile across sites and describes spatial heterogeneity of substitutions in terms of simple covarion models (Ane et al., 2005).

To improve taxon sampling at the base of the angiosperm radiation, we also sequenced the chloroplast genome of *Trithuria inconspicua,* a species from a genus of minute aquatic herbs, which recently has been found to be closely related to Nymphaeales (Saarela et al., 2007). Our findings highlight the importance of the fit between model and data when evaluating relationships among basal angiosperms.

Materials and methods

Sequencing of the chloroplast genome of Trithuria inconspicua

Trithuria inconspicua was collected from the Kai Iwi Lakes (Lakes Waikare and Taharoa), Northland, North Island, New Zealand, and sent by courier to Massey University, Palmerston North. Voucher specimens have been deposited at the Auckland War Memorial Museum Herbarium AK (see AK 308938, AK 320388). Enriched cpDNA was sequenced on an Illumina GAII platform as described in Atherton et al. (2010). Contigs were assembled using Velvet version 0.7.60 (Zerbino and Birney, 2008) and odd kmer values ranging from 25 to 61. Because the copy number of cpDNA was higher than that for the nuDNA (though not a higher absolute amount), coverage cutoffs of 10, 20, 40 and 80 were applied during the assembly of contigs. Staden 2.0.0b7 (http://staden.sourceforge.net/) was used to join the contigs generated by Velvet. Nine gaps remained after the assembly; eight gaps were closed by designing primers to flanking regions and sequencing the missing parts using standard ABI3730 sequencing protocols (Massey Genome Service http://genome.massey.ac.nz/).

Taxon selection and multiple sequence alignment

Protein-coding sequences of 61 genes common to 31 chloroplast genomes from angiosperms and gymnosperms were downloaded from GenBank. NAD dehydrogenase genes were not included in analyses as these are absent from the cpDNAs of gnetophytes and conifers (Wakasugi et al., 1994; Braukmann et al., 2009). In our taxon sampling, we included representatives of all available basal angiosperm lineages but not all crown group angiosperm species for which chloroplast genomes have been determined. This taxon selection retained species most important for inferring relationships among basal angiosperms and reduced computation time for model-fitting and tree-building analyses on a 16-core Linux server.

Eudicots were represented by six basal species. We excluded grasses, known to be subtended by a very long branch in previous analyses (Goremykin et al., 2005), keeping all other monocots.

As concern over alignment procedures remains an important practical consideration for phylogenomic analyses (Philippe et al., 2011), multiple sequence alignments were generated using two alignment protocols in the present study. The first protocol, used as a basis for figures shown in the manuscript, uses the same principles described in Goremykin et al. (2004). This alignment protocol provides a rapid and reliable method of aligning similar gene sequences and for producing data sets comprising first and second codon positions and all three codon positions. With this approach, gene sequences were sorted into 61 Fasta files, each containing orthologues. For each file, first and second codon positions were aligned using the program MUSCLE (Edgar, 2004). Alignments for sequences that included all three codon positions were also generated by the same script. The resulting 122 alignment files were each manually edited, such that regions of low similarity between the ingroup and outgroup sequences were discarded. Individual gene alignment files were concatenated using Geneious v5.5.4. (Drummond et al., 2010) to produce i) a gapped alignment of 40553 positions in length, provided as a supplementary material (File S1) and ii) an alignment of first and the second codon positions 25246 positions in length (supplementary File S2). An OV sorted (see below) version of the 40553 pos. long alignment has been provided as a supplementary material file S3.

A MUSCLE alignment of translated nucleotide sequences from 56 individual Fasta files was also generated and used to confirm results of phylogenetic analyses obtained using the first alignment protocol. This second alignment approach used the same principles as previously implemented for obtaining conservative alignments between anciently diverged sequences (Lockhart et al., 1996). With this method, we imported each Fasta file into MEGA 5.0 (Tamura et al., 2011), translated the sequences, and then aligned them with MUSCLE (default options). We concatenated these aligned files using Geneious v5.5.4. (Drummond et al., 2010) and then imported the concatenated file into Se-Al. v2.0a11. (Rambaut, 2002). Site patterns adjacent to indels were then removed if they did not contain amino acids with similar physical/chemical properties as specified in Se-Al. Finally, the columns with gaps were removed and the sequences back-translated. This alignment protocol produced a much shorter concatenated alignment than did the first method (31674 ungapped positions). This alignment has been provided as a supplementary material (File S4).

OV sorting and "noise reduction"

Site patterns in our concatenated alignments were reordered according to their OV scores and data partitions identified for tree building using the GNB criterion (Goremykin et al., 2010). Previously, this approach was found effective in the recovery of benchmark clades of mammalian phylogeny, and more effective than other methods in identifying fast-evolving sites that cause long branch attraction (LBA) artifacts (Goremykin et al., 2010).

OV sorting involves calculating a sum-of-pairs mismatch score for each site in the full alignment (including positions with gaps) and then ordering the sites according to the OV scores (Goremykin et al., 2010). This produces an alignment with the most conserved (least varied) site patterns at one end, and the least conserved (most varied) positions at the other end. We refer to this alignment as the OV alignment. The OV alignment was generated using the script Sorter.pl. This script also splits the OV alignment into several bipartitions of sites. Each bipartition contains an "A" partition, which includes site patterns from the conserved end of the alignment, and a “B” partition, which includes site patterns from the least conserved end of the OV alignment. In the present study, the bipatrition of sites into partitions A and B occurred at position *i* x 250 (where *i* = 1,2,3, …) upstream from the most varied end of the OV alignment. The incremental increase in interval length of 250 sites for the B partition is an arbitrary size previously found suitable for monitoring change in the properties of the ordered sites at the most-varied end of the OV alignment. Once the bipartitions are formed, the script Sorter.pl calls ModelTest (Posada and Crandall, 1998) to identify an optimal time-reversible substitution model for each of the A and B partitions using a two-step procedure (for further details, see: Goremykin et al., 2010). The script then calls PAUP* (Swofford, 2002; Unix v. 4.0b10) to calculate a matrix of maximum likelihood (ML) distances for the A and B partitions. A matrix of *p*-distances (number of sites with observed differences/total number of sites) is also calculated for each B partition.

Sorter.pl also calculates the average of the ML-distances minus the average of the *p*-distances and reports this mean deviation of the ML- and *p*-distances for the B partitions, and Pearson correlation coefficient values (*r*) between these estimates (Goremykin et al., 2010). Dissimilarity between relative ranking of ML- and *p*-distances calculated from the B partitions occurs if distance estimates are not similar between taxa. Stochastic error associated with the short sequence length of the initial B partitions will cause such dissimilarity, as will substitution model violations and saturation with multiple substitutions. By monitoring the *r* values as the length of the B partition is increased, it is possible to identify a point of transition with respect to the similarity of the distances compared. As the relative ranking of absolute distance values within two groups of distance estimates (*p*- and ML- distances) becomes similar, there is a dramatic rise in the value of *r*.

As well as comparing the ML- and *p*-distances for B partitions, the script Sorter.pl also compares optimal ML-distances for the A and B partitions. Deviation is again measured in terms of *r*. As with the ML- and *p*-distance comparison for the B partition, a dramatic rise in *r* occurs when the distances become proportional, and their ranking becomes similar. The comparison identifies the relative length of the A and B partitions, at which point the evolutionary properties of the B partition become similar to those of the A partition.

Goremykin et al. (2010) have suggested that the site stripping process should cease when there is a dramatic increase in the value of *r* in both correlation analyses. At this point, positions added from the conserved A partition to the variable B partition clearly start to mask the non-phylogenetic signal associated with the most-varied positions in the B partitions. Here we also report that the topological distortion induced by the presence of B partition sites is also greatly reduced at this point. Model misspecification contributed by compositional heterogeneity, as we also show, still persists beyond this point. However, this has little impact on the relative ranking of distances in B partitions. Thus, further character removal is not justified on the basis of the GNB criterion.

As demonstrated in Zhong et al. (2011), the GNB criterion also identifies and provides a basis for removing sites from a concatenated alignment that have a poor fit to phylogenetic model assumptions. While this criterion does not remove all model-violating sites from the data, it has been shown to remove sites that significantly impact on phylogenetic estimates, and thus sites that have significant effect in misleading tree building. In particular, it appears very useful for reducing LBA artifacts in phylogenetic reconstruction. This was demonstrated in reanalysis of mitochondrial DNA sequences, which previously and consistently had yielded a rodent polyphyly artifact (Goremykin et al., 2010) and also in recent analyses of chloroplast sequences from Gnetales and other seed plants (Zhong et al., 2011).

To study the relationship between changes in *r* and branch length support in reconstructed trees, splits can be calculated for individual A and B partitions. We calculated NeighborNet (NNET: Bryant and Moulton, 2004) splits from the optimal ML-distances obtained for each B partition generated during the noise reduction protocol. These were calculated using SplitsTree 4.0 (Huson and Bryant, 2006). Of particular interest are the splits that separate outgroup and ingroup taxa as these are relevant for the question of rooting the angiosperm radiation. In the present study, we plotted the relative size of the splits separating (a) angiosperms from gymnosperms and (b) Gnetales from other species. Such a "heterotachy plot", as it was referred to in Zhong et al. 2011, allows visualization of the relationship between B-partition distances and any topological distortion (Lockhart et al., 1996; Bruno and Halpern, 1999) of reconstructed trees due to including the most-varied sites of the OV alignment when tree building.

Base composition heterogeneity

Base compositional heterogeneity (Jermiin et al., 2004) was examined over the most-varied end of the OV alignment. To investigate this, intervals of sites with the same length (360 jacknife resampled ungapped positions; 3 replicates for each interval) were sampled from non-overlapping locations at the most-varied end of the OV alignment (between 0-500 sites, 500-1000 sites, 1000-1500 sites, …, 9500-10,000 sites). We examined each of these sets of sites using Bowker's matched-pair symmetry test (Ababneh et al., 2006), as implemented in Seq-Vis (Ho et al., 2006). We used Seqboot from the PHYLIP v3.69 (Felsenstein, 2004) package for jackknife resampling of sites (sampling without replacement) and SeqVis v1.5 (Ho et al., 2006) for the symmetry test. The smallest interval from which sites were resampled was the first interval: 0-500 sites (these 500 gapped positions contained 380 ungapped positions).

Goodness of fit analyses

We used MISFITS (Nguyen et al., 2011) and Tree-Puzzle-5.2 (Schmidt et al., 2002) to identify those site patterns in the OV alignment whose observed frequencies were unexpected under a GTR + I + Γ substitution model. This model was identified as the best-fitting model to the OV alignment among all models that assumed a single matrix of base frequencies. This was also the case for the increasingly short A partitions according to a double-fitting procedure that employed an AIC criterion (described in Goremykin et al., 2010). The fit of the GTR + I + Γ model to chloroplast data sets is also of significant interest as this model has been commonly used in phylogenetic analyses of basal angiosperms (e.g. Saarela et al., 2007; Graham and Iles, 2009; Stefanović et al., 2004; Leebens-Mack et al., 2005; Soltis et al., 2011; Jiao et al., 2011; Qiu et al., 2010; Barkman et al., 2000; Wu et al., 2007; Zanis et al., 2002). While there are computational issues with co-estimation of the I + Γ parameter values (e.g. see discussion in Yang, 2006), this model has been found to have higher reconstruction accuracy than GTR + Γ models in more biologically realistic simulations (Gruenheit et al., 2008). The impact that deletion of sites from the most-varied end of the OV alignment had on the fit of this substitution model was also studied at different shortening steps. Log-likelihood scores for the evolutionary model obtained for the increasingly short A partitions were also compared with the log-likelihood scores for equal length partitions that were jackknife resampled from the complete OV alignment. We used Seqboot for jackknife resampling and PhyML 3.0 (Guindon et al., 2010) for calculating log-likelihood scores.

Substitution Model Selection for A partitions

The optimal substitution model was determined for the A partition data sets using cross-validation as implemented in PhyloBayes 3.2e (Lartillot and Philippe, 2004). To determine the length of time needed for convergence of posterior probabilities, we initially ran PhyloBayes on a 16-core Linux server for at least 2 weeks with alignments of the first and second codon positions, and of all three codon positions, choosing between 6 substitution models for each input file: the "classical" GTR + Γ, GTR + Γ + covarion, GTR + Γ + covext, GTR + Γ + CAT, GTR + Γ + CAT + covarion, and GTR + Γ + CAT + covext (Lartillot and Philippe, 2004). Here, "CAT" refers to the site-heterogeneous mixture model of Lartillot and Philippe (2004), "covarion" to the covarion model of Tuffley and Steel (1998), and "covext" to a variant of the Tuffley and Steel model that allows for variation in rate-heterogeneity across sites. We assumed a 4-category discrete Γ distribution in modeling rate-heterogeneity across sites. From these initial 12 runs we determined that 200 cycles were sufficient for convergence on our Linux server. Since cross-validation is multi-staged and computationally demanding, we wrote a script Cross.pl, which initiates parallel multiple PhyloBayes and cross-validation runs. This script first invokes PhyloBayes, lets it run for 1000 cycles under the abovementioned models, and builds consensus trees discarding the first 200 cycles as burn-in. Then the script invokes the PhyloBayes program cvrep to randomly sample 10 learning and 10 test data partitions from each alignment, so that each learning data partition has 90% of the input alignment length and each test partition has 10% of the input alignment length. The script then calls PhyloBayes and performs Markov chain Monte Carlo (MCMC) sampling for 200 cycles in parallel for the learning sets created by the PhyloBayes program cvrep. Subsequently, the script initiates the Phylo Bayes program readcv in parallel for all data replicates and computes a cross-validation score (i.e. calculates the likelihood under the test set, averaged over the posterior distribution of the learning set) discarding a burn-in of 50 sampling points and taking every point thereafter. Finally, the script invokes the PhyloBayes program sumcv to compute summary statistics. Using the Akaike information criterion (AIC) for the double-fitting procedure, the GTR + I + Γ model was selected as the best-fitting model among those with one matrix of base frequencies for the OV alignment and its next 20 shortened subsets.

Tree building

Phylogenetic reconstructions were performed using PhyloBayes and the PAUP*-embedded scripts in Sorter.pl (Goremykin et al., 2010). RAxML (Stamatakis et al., 2005) was also used to reanalyze a recently published data set of chloroplast, mitochondrial, and nuclear genes (Soltis et al., 2011).

Availability of scripts and of Trithuria chloroplast genome sequence

Scripts not already publically available and used in this study have been provided as supplementary material. The sequence for the *Trithuria inconspicua* chloroplast genome determined in in this study has been deposited with EMBL (Accession no. HE963749).

Results

Alignments

Two alignments were obtained using different approaches in the present study. Despite differences in their lengths, both methods produced very similar alignments. This can be visualised by comparing split networks that display the NNET split systems (*p*-distances) for each alignment (Supplementary file S5). Similar analytical results were obtained for both alignments. The figures shown in subsequent sections were based on the alignment method of Goremykin et al. (2003a).

GNB analyses

A significant improvement in *r* occurred after eight steps: 2000 sites (Fig. 4-1a); that is, once the 2000 most-varied sites were included in the B partition, *p*-distances and ML-distances for the B partition had become highly correlated. Similarly, at this shortening step ML-distances for A and B partitions also became highly correlated (Fig. 4-1b), indicating similar evolutionary distances for both partitions, and suggesting a point had been reached at which further removal of sites from the A partition was no longer justified. Most significantly, the distance between outgroup and ingroup taxa reduced dramatically by the 8th sampling step. This was visualized in Figure 4-2, which shows the relative length of outgroup splits in the NNET split system for the taxon set. The extreme branch length separating the outgroup and ingroup sequences is a property of the 2000 sites at the most-varied end of the OV alignment.

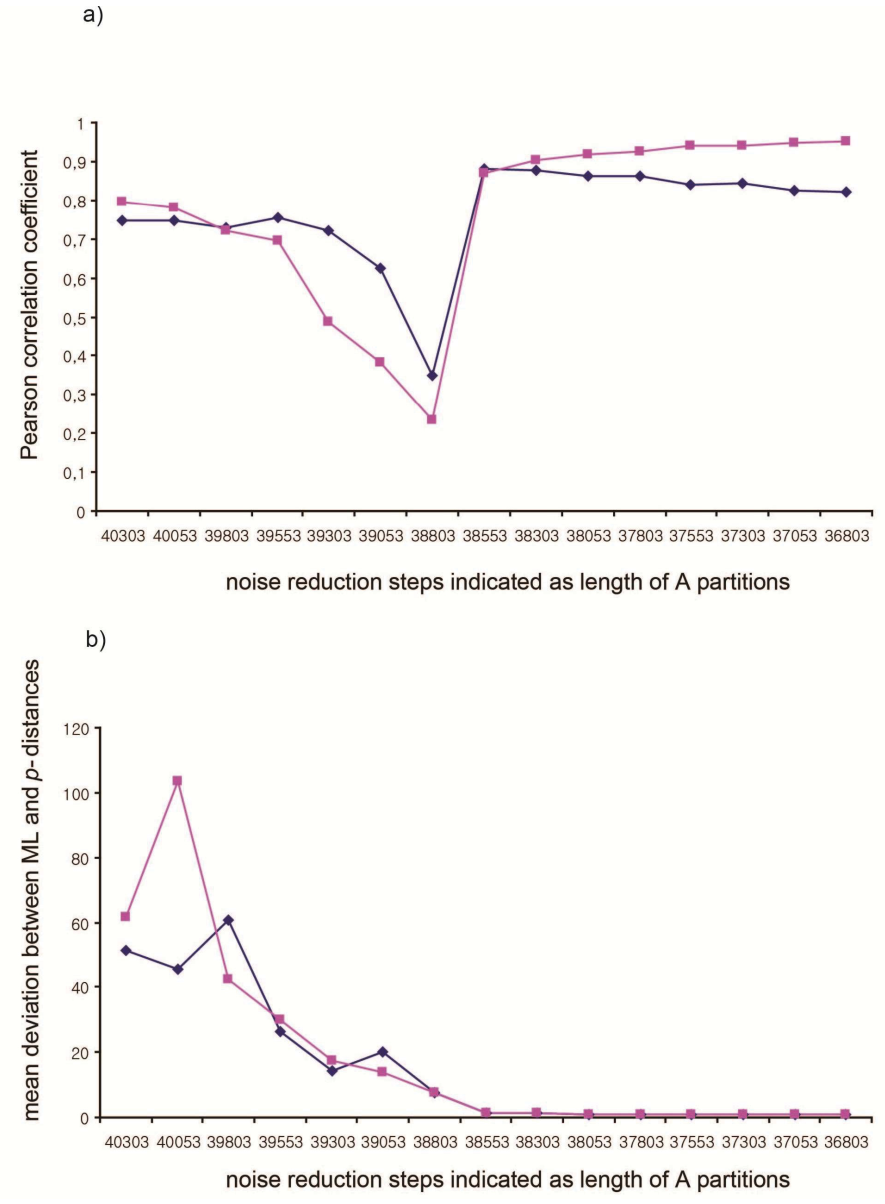


Fig. 4-1. a) Plot showing results of the correlation analyses. The dotted blue line indicates Pearson correlation coefficient values (*r*) obtained for pair-wise comparisons of ML-distances calculated from the A and B partitions whose combined length was 40553 gapped positions in the OV alignment. The dotted pink line indicates *r* values obtained for pair-wise comparisons of *p*-distances and ML-distances calculated for B partitions, discarded at each shortening step. At the 8th shortening step, when the A partition is 38553 gapped positions in length, it passes both correlation tests (Goremykin et al., 2010). b) Plot showing mean deviation between ML- and *p*-distances calculated for B partitions at each shortening step. In calculating ML-distances, the best-fitting ML model for each partition length was first determined under an AIC criterion using ModelTest (Posada and Crandall, 1998). The pink line indicates results from analyses using a Neighbor-Joining tree to fit ML model parameters. The blue line indicates results obtained when an ML tree is used to fit substitution model parameters. This ML tree was computed using settings of the best-fitted model determined by the standard ModelTest procedure employing AIC.

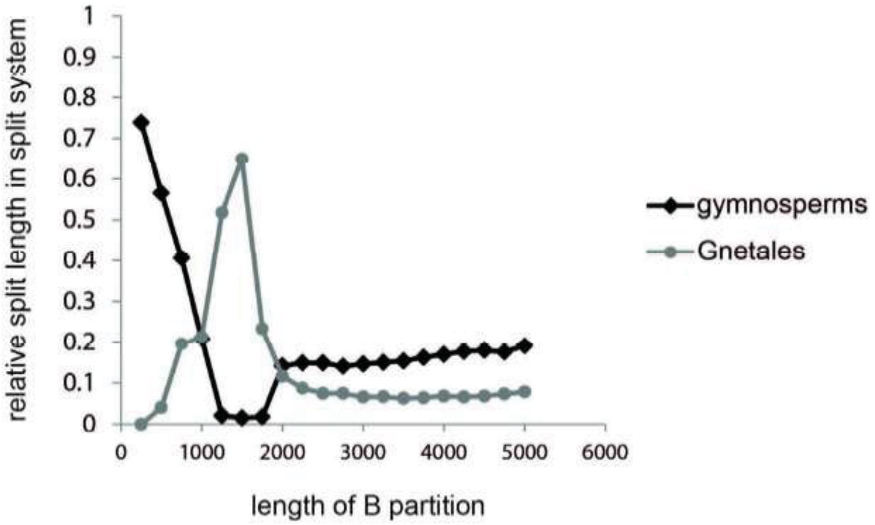


Fig. 4-2. Plot showing the relative size of NNET splits in split system separating i) angiosperms from gymnosperm and ii) Gnetales from other taxa. The NNET splits were calculated from the optimal distances estimated for each B partition formed at the most-varied end of the OV alignment.

Compositional Heterogeneity

It has been previously observed that compositional heterogeneity and the rate of substitutions of sites are tightly correlated (Rodriguez-Ezpeleta et al., 2007). Our analyses provide some support for this observation. Figure 4-3 indicates that compositional heterogeneity is a feature of the most-varied end of the OV alignment. In particular, it indicates the number of pairs failing a matched-pairs test of symmetry at $p < 0.00005$ when these are calculated on identical length partitions (360 sites each) sampled within 500 bp non-overlapping gaped intervals at the most-varied end of the OV alignment. The plot suggests that heterogeneity in composition is most significant over the first 3000-3500 most-varied positions of this alignment. This heterogeneity is most significant between angiosperm and outgroup sequences and among outgroups sequences (values for individual pairs not shown). It extends past the stopping point identified by the GNB method. Hence, while compositional heterogeneity is likely to contribute to the extreme branch length difference between ingroup and outgroup sequences, it does not appear to explain the extreme branch length differences over the first 2000 most-varied positions in the OV alignment.

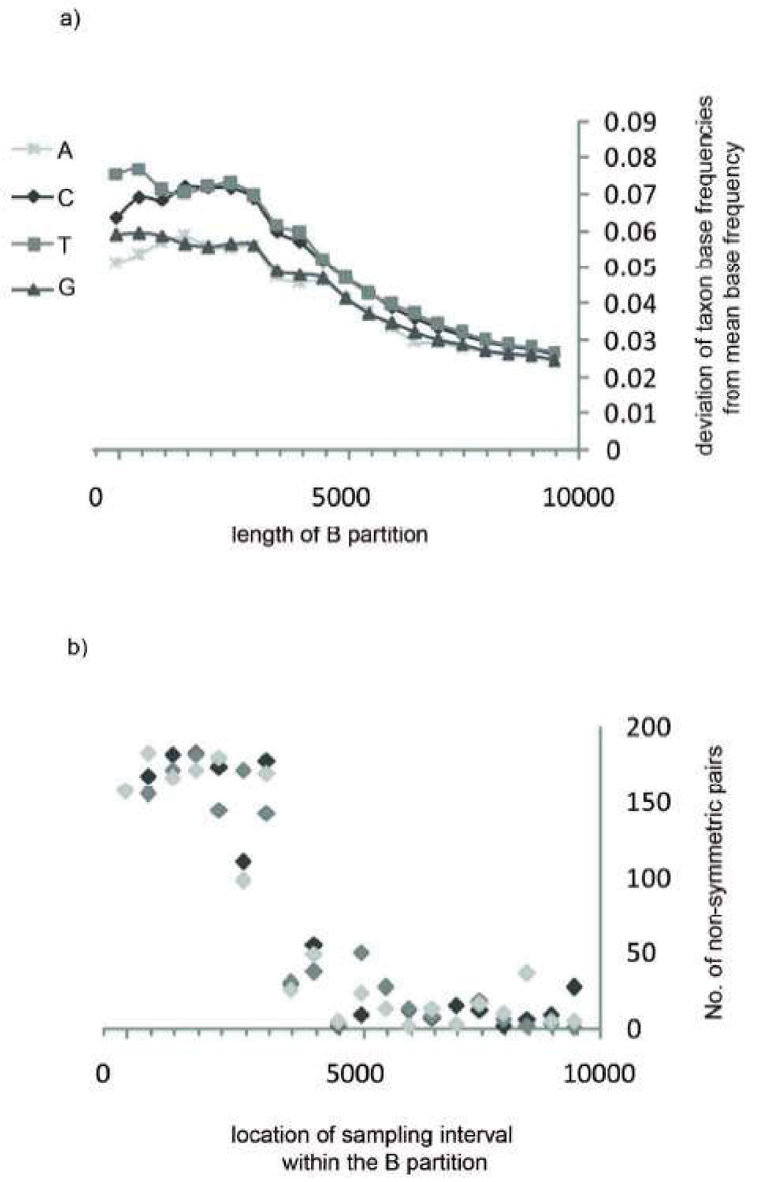
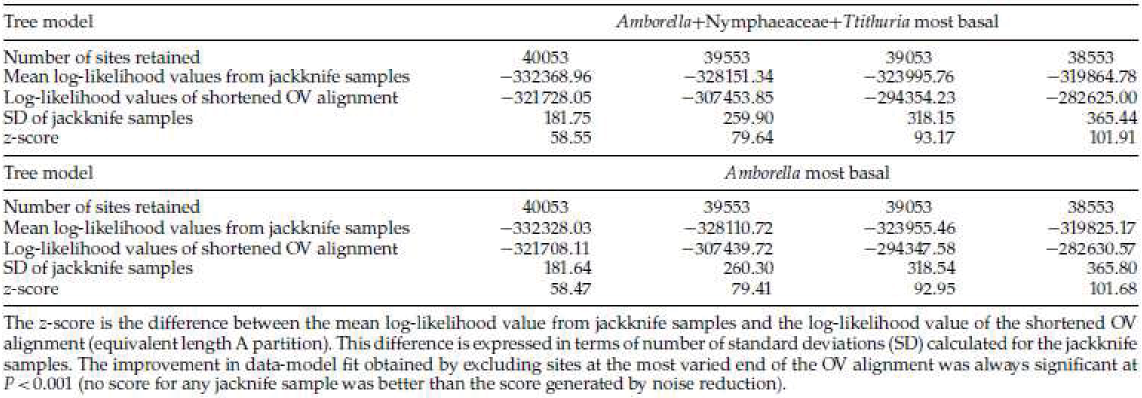


Fig. 4-3. The number of pairwise distances (645 comparisons) failing a matched-pairs test of symmetry at p<0.00005 was determined for equal length, non-overlapping intervals at the most-varied end of the OV alignment. For these estimates, we analyzed only ungapped sites (360 positions: 3 replicates per estimate) randomly sampled without replacement from 500 bp non-overlapping gapped partitions at the most-varied end of the OV alignment ("C" partitions in Goremykin et al. 2010).

Fit of data to a GTR + I + Γ substitution model

The effect of removal of the most-varied sites on the fit of the aligned data to a GTR + I + Γ substitution model was investigated. Table 4-1 reports log-likelihood scores for two tree models (*Amborella* most basal; *Amborella* + *Trithuria* + *Nymphaeae* most basal) on A partitions generated by the script Sorter.pl. These scores were compared against the log-likelihood scores for data sets identical in length to the shortened A partitions that were jackknife resampled from the OV alignment. They were always significantly better than the scores for the randomly resampled data, indicating that the sites removed by OV noise reduction significantly contribute to the poor fit between the evolutionary models and the aligned sequence data.

Table 4-1. Data-model fit after removal of 500, 1000, 1500 and 2000 sites.



Tree model	*Amborella*+Nymphaeaceae+*Ttithuria* most basal			
Number of sites retained	40053	39553	39053	38553
Mean log-likelihood values from jackknife samples	−332368.96	−328151.34	−323995.76	−319864.78
Log-likelihood values of shortened OV alignment	−321728.05	−307453.85	−294354.23	−282625.00
SD of jackknife samples	181.75	259.90	318.15	365.44
z-score	58.55	79.64	93.17	101.91
Tree model	*Amborella* most basal			
Number of sites retained	40053	39553	39053	38553
Mean log-likelihood values from jackknife samples	−332328.03	−328110.72	−323955.46	−319825.17
Log-likelihood values of shortened OV alignment	−321708.11	−307439.72	−294347.58	−282630.57
SD of jackknife samples	181.64	260.30	318.54	365.80
z-score	58.47	79.41	92.95	101.68

The z-score is the difference between the mean log-likelihood value from jackknife samples and the log-likelihood value of the shortened OV alignment (equivalent length A partition). This difference is expressed in terms of number of standard deviations (SD) calculated for the jackknife samples. The improvement in data-model fit obtained by excluding sites at the most varied end of the OV alignment was always significant at $P < 0.001$ (no score for any jacknife sample was better than the score generated by noise reduction).

Assuming the same evolution models as examined in Table 4-1, MISFITS and Tree-Puzzle were used to the identify site patterns whose relative frequencies are over- and under-represented in the OV alignment. Figure 4-4 plots the position of unexpected site patterns in the ungapped OV alignment. The height of each bar in the histogram indicates the number of consecutive sites at which unexpected site patterns occur. The most-varied end of the OV alignment is identified as containing many site patterns that contribute to the poor fit of the GTR + I + Γ substitution model.

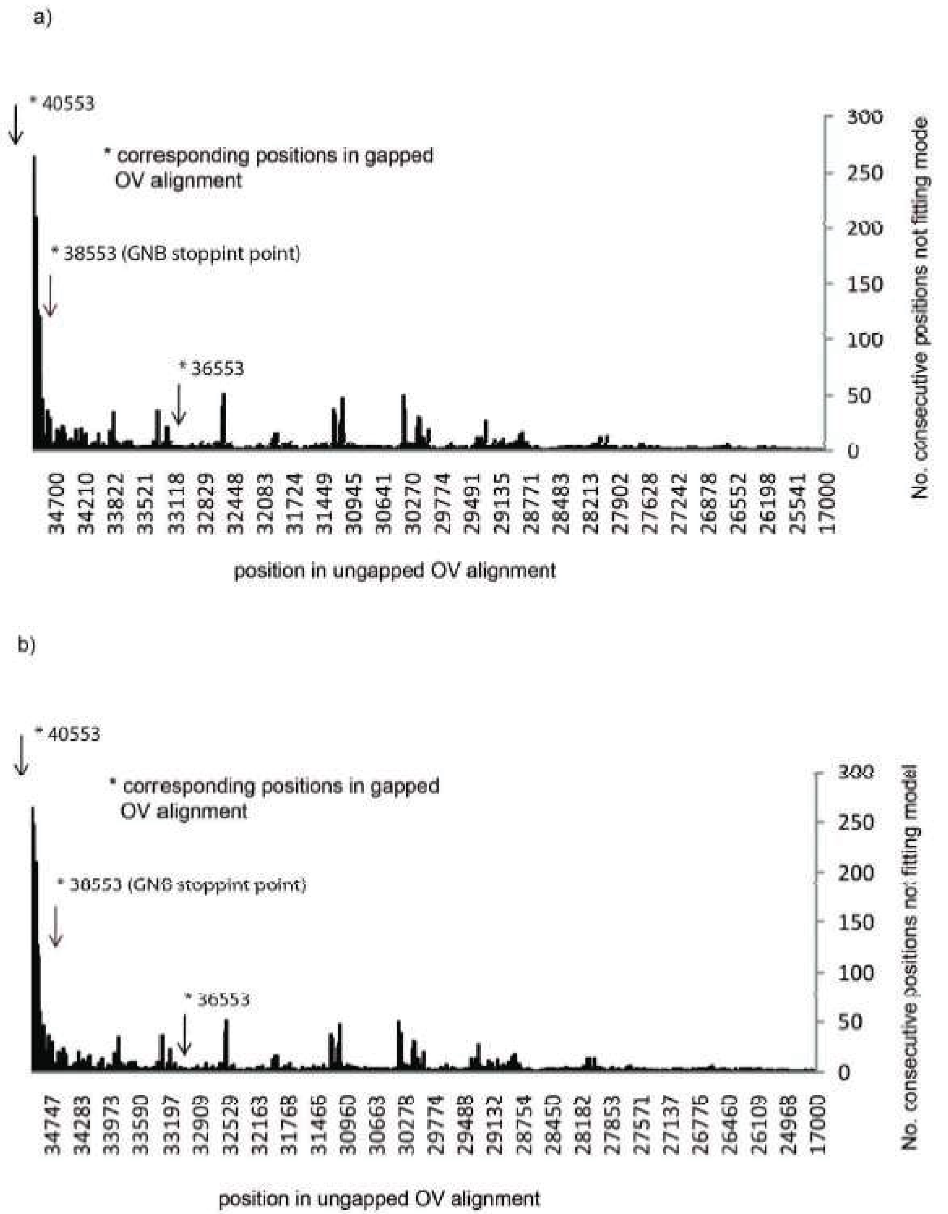


Fig. 4-4. a) Histogram showing positions of sites in the OV alignment that contain site patterns unexpected under a GTR + I + Γ substitution model and *Amborella+Trithuria+Nymphaeales* hypothesis. b) Histogram showing positions of sites in the OV alignment that contain site patterns unexpected under a GTR + I + Γ substitution model and *Amborella* most basal hypothesis. A feature of both graphs is that relatively few sites fit either model at the most-varied end of the OV alignment. Both ungapped positions and gapped positions (*) have been indicated on the figure.

Tree building

Phylogenetic trees were built from the OV alignment for the different length A partitions generated by the Sorter.pl script. This was done both for a CAT + GTR + Γ + covext model and for a GTR + I + Γ model. The former was found under cross-validation to be optimal for i) the full length OV alignment, ii) the alignment of the first and the second codon positions and iii) the alignment of the most conserved 38553 positions in the OV alignment. The optimal tree reconstructed with a CAT + GTR + Γ + covext model on the A partition at the GNB stopping point is shown in Figure 4-5.

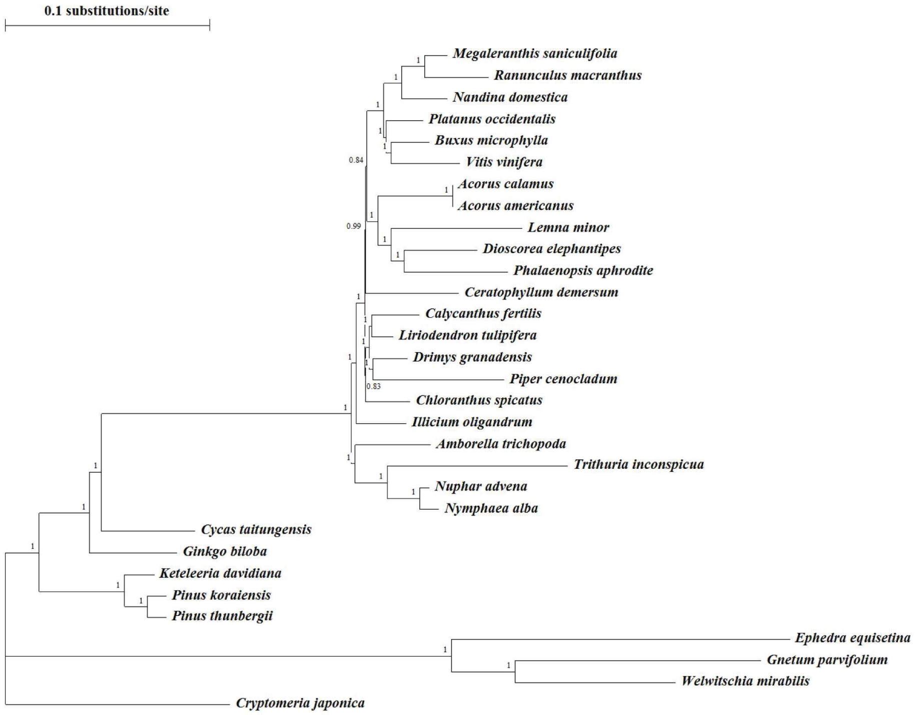


Fig. 4-5. Tree reconstructed from Bayesian analysis and best-fitting substitution model (CAT + GTR + Γ + covext model) for the conserved A partition (38553 sites) identified by the GNB criterion.

This tree indicates the same relationships among basal angiosperms as does the GTR + I + Γ tree reconstructed on the A partition at the GNB stopping point. Both reconstructions identify a lineage comprising *Amborella* + *Trithuria* + Nymphaeales as most basal in the angiosperm radiation. Figure 4-6a indicates relationships inferred when a CAT + GTR + Γ + covext model is used to analyze the full length (40553 sites) concatenated data set. With this data set, *Amborella* is inferred to be the most basal lineage in the radiation of angiosperms. Trees built from the alignment of first and second codon positions using the best-fitting CAT + GTR + Γ + covext model (Fig. 4-6b) show *Amborella* + *Trithuria* + Nymphaeales as the most basal lineage. Substitution models rejected in cross-validation supported the tree with the most basal branch subtending *Trithuria* + Nymphaeales (e.g. GTR + Γ model, Fig. 4-6c) based on the first and second position data set.

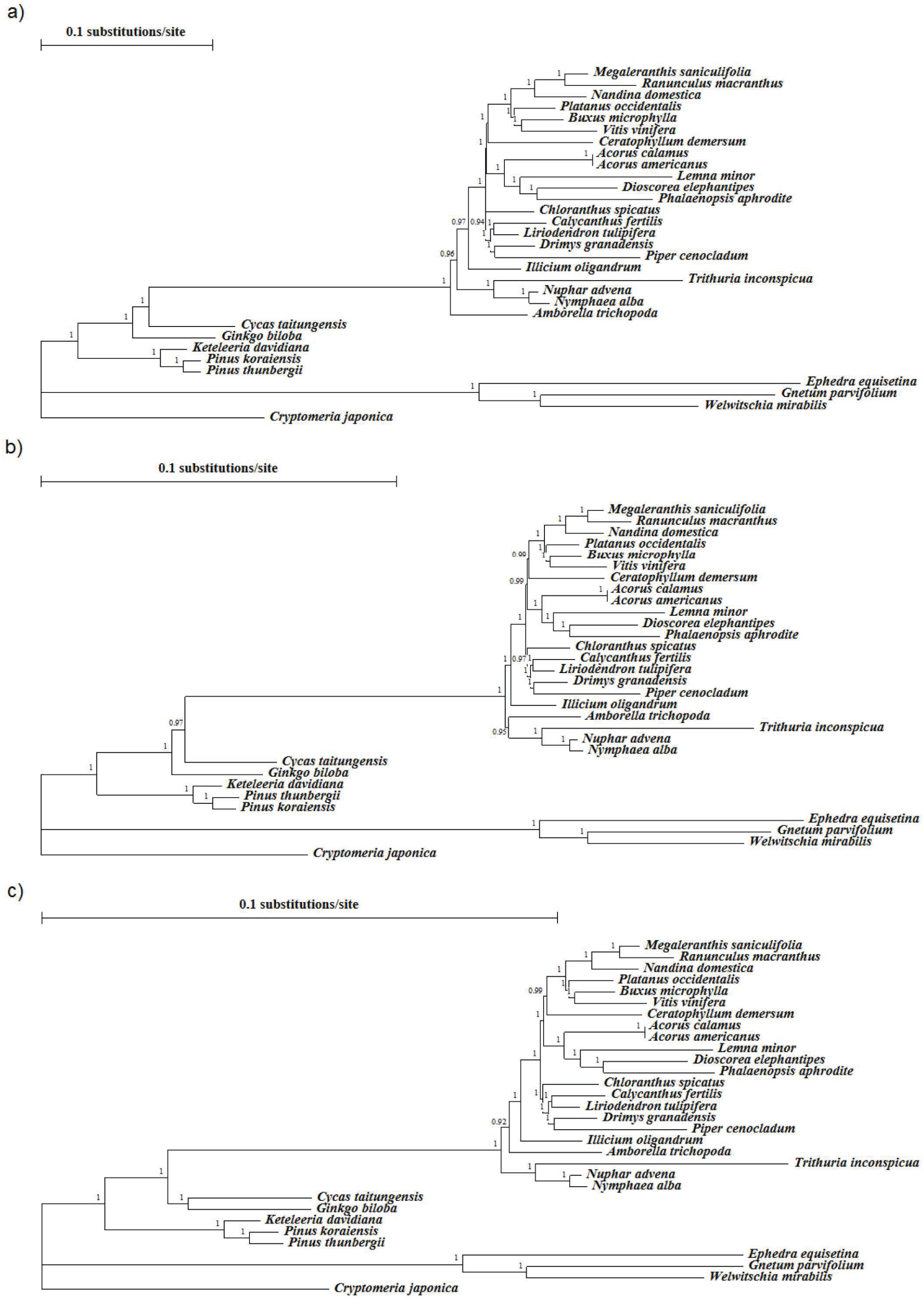


Fig. 4-6. a) Tree reconstructed from Bayesian analysis and best-fitting substitution model (CAT + GTR + Γ + covext model) for the full length (40553 sites) concatenated data set. b) Tree built from the alignment of the first and the second codon positions employing best-fitting CAT + GTR + Γ + covext model. c) Tree built from the alignment of the first and the second codon positions employing the GTR + Γ model.

The support for relationships among basal angiosperms under a CAT + GTR + Γ + covext covarion model was also investigated after each shortening step of 250 positions in the alignment of all codon positions. The results are shown in Figure 4-7. These indicate that i) support for *Amborella* joining with the outgroup occurs only when the most-varied positions of the alignment are included, ii) the grouping of *Amborella* + *Trithuria* + Nymphaeaeles is strongly favored as the most basal lineage after removal of 1750 sites and remains supported until 2500 sites are removed and iii) a basal grouping of *Amborella* + *Trithuria* + Nymphaeales + *Illicium* becomes favored after removal of 2750 sites. Note that under the CAT + GTR + Γ + covext model, support for *Amborella* + *Trithuria* + Nymphaeales as a most basal clade is realized prior to the GNB stopping point, which might indicate a better fit of this substitution model to the data.

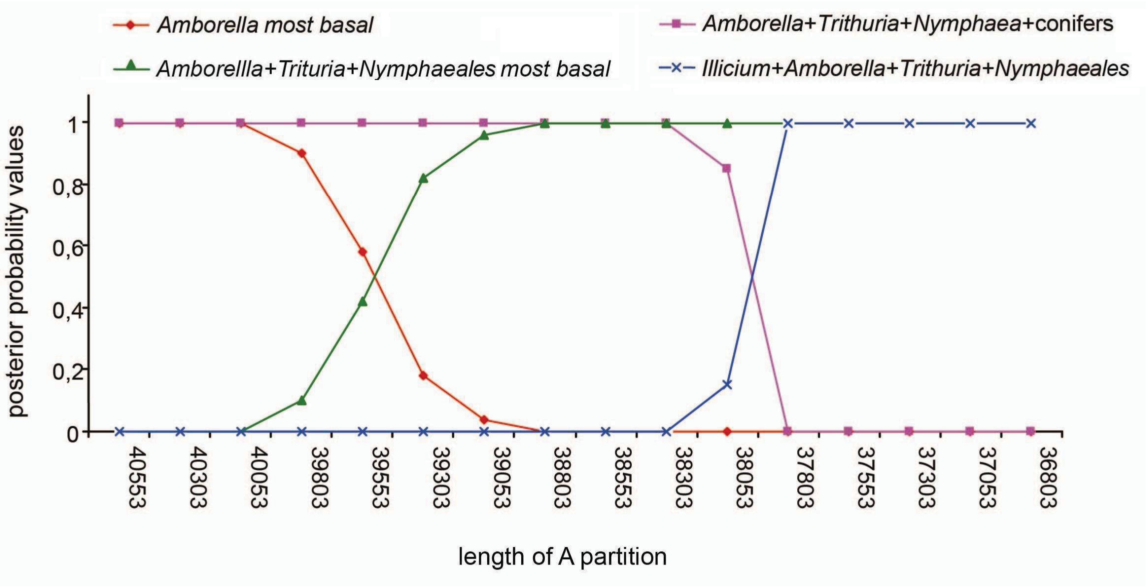


Fig. 4-7. Posterior probability support for alternative hypotheses of relationship as sites are removed from the most-varied end of the OV alignment computed under the best-fitting substitution model (CAT + GTR + Γ + covext). Similar inferences were obtained with taxon subsets that excluded the most compositionally heterogeneous sequences.

Discussion

Our findings reported here, and those in recent analyses of other seed plants (Zhong et al., 2011), re-emphasize the importance of considering the fit of time-reversible models to the fast-evolving sites in sequence alignments (Sullivan et al., 1995). We show that site sorting can facilitate studies of the substitution properties of concatenated gene alignments and help to identify site patterns relevant to substitution model misspecification and potential tree-building artifacts. The sites providing most support for the *Amborella* most basal hypothesis are characterized by poor fit between model and data and by evolutionary properties that induce extreme topological distortion in reconstructed trees. The GNB stopping criterion removes many of these sites (38% of the removed sites did not fit an *Amborella* basal + GTR + I + Γ model;

39% of the removed sites did not fit an *Amborella* + *Trithuria* + Nymphaeales basal + GTR + I + Γ model).

In the present study, when sites causing topological distortion were removed, reconstruction under the optimal CAT model and GTR + I + Γ model favors a tree indicating *Amborella* + *Trithuria* + *Nymphaea* as the most basal hypothesis. While compositional heterogeneity will contribute to topological distortion when time-reversible Markov models are used in analysis of the data, our heterotachy and matched-pairs test of symmetry plots suggest that compositional heterogeneity is alone insufficient to explain the different topologies obtained during tree building with different A partitions. In general, the impact of compositional heterogeneity needs to be evaluated in the context of the extent of divergence between sequences exhibiting this heterogeneity (Jermiin et al., 2004) and the spatial pattern of sites free to vary in the sequences (Lockhart et al., 2006).

We propose that our analyses and observations provide a basis for understanding the discrepancy among recent findings from phylogenetic analyses of cpDNA and mtDNA concerning the rooting of the angiosperm phylogeny. Our reconstructed phylogeny (Fig. 4-5a) obtained after exclusion of a large number of model-violating sites is consistent with that recently obtained in analyses of nuclear EST amino acid sequences that also implemented a CAT model. In this case, while *Trithuria* was not available for study, *Amborella* and *Nuphar* were inferred to be sister taxa (Finet et al., 2010). Our reconstruction is also congruent with recent analyses of four slowly evolving mitochondrial genes (Qiu et al., 2010).

Our phylogenetic reconstruction differs from that obtained in a recent and well-sampled ML-based phylogenetic analyses for 17 concatenated nuclear, mitochondrial, and chloroplast genes (Soltis et al., 2011). This study reported *Amborella* as most basal. Reanalysing these data with a GTR + I + Γ model and RaxML, we were unable to confirm this finding. Rather, we inferred a phylogenetic tree wherein a clade comprising *Amborella*, *Trithuria* and Nymphaeales receives 94% non-parametric bootstrap support (results not shown). Whether this result indicates a shortcoming of the heuristic search with RaxML or a more accurate reconstruction of angiosperm phylogeny from this joint data matrix requires further investigation.

We conclude that analyses of available sequence data do not support the earliest angiosperms being woody and terrestrial. Evidence from phylogenetic analyses of concatenated chloroplast genes appears equally consistent with some of the earliest species being herbaceous and aquatic. Further tests of this hypothesis are needed. We suggest that our analytical protocol provides a valuable approach, and one that is potentially useful for other questions currently being investigated with phylogenomic datasets.

Acknowledgments

We thank the Allan Wilson Centre and New Zealand Royal Society (James Cook Fellowship scheme and Marsden Fund) and for financial support. We thank Dr. Lars Jermiin and the reviewers for their constructive and helpful suggestions.

Supplementary material available on DRYAD:
http://datadryad.org/review?wfID=8426&token=f63fb0ef-009a-492c-9cc6-6c613515e618
Provisional DOI:10.5061/dryad.vs49s

5 Systematic Error in Seed Plant Phylogenomics[1]

Abstract

Resolving the closest relatives of Gnetales has been an enigmatic problem in seed plant phylogeny. The problem is known to be difficult because of the extent of divergence between this diverse group of gymnosperms and their closest phylogenetic relatives. Here, we investigate the evolutionary properties of conifer chloroplast DNA sequences. To improve taxon sampling of Cupressophyta (non-Pinaceae conifers), we report sequences from three new chloroplast (cp) genomes of Southern Hemisphere conifers. We have applied a site pattern sorting criterion to study compositional heterogeneity, heterotachy, and the fit of conifer chloroplast genome sequences to a general time reversible + G substitution model. We show that non-time reversible properties of aligned sequence positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction of these seed plants. When 2,250 of the most varied sites in our concatenated alignment are excluded, phylogenetic analyses favor a close evolutionary relationship between the Gnetales and Pinaceae—the Gnepine hypothesis. Our analytical protocol provides a useful approach for evaluating the robustness of phylogenomic inferences. Our findings highlight the importance of goodness of fit between substitution model and data for understanding seed plant phylogeny.

Keywords: compositional heterogeneity, heterotachy, Gnetales, systematic error.

[1] Original manuscript: Zhong B, Deush O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ (2011) Systematic errors in seed plant phylogenomics. Biology and Evolution 3: 1340-1348. Own contribution: 10% (analyses)

Introduction

Gnetales are a morphologically and ecologically diverse group of Gymnosperms, united as a monophyletic group based on special features of their cytology. Initially, they were thought to be the nearest relatives of flowering plants (angiosperms) based on the morphological similarities (the ‘‘Anthophyte’’ hypothesis) (Crane, 1985). However, all recent molecular work has separated Gnetales away from the angiosperms and instead placed them with or within conifers. Some analyses have placed them as sister group to conifers (the ‘‘Gnetifer’’ hypothesis, Chaw et al., 1997), others close to Pinaceae (the ‘‘Gnepine’’ hypothesis, Bowe et al., 2000; Chaw et al., 2000; Finet et al., 2010; Zhong et al., 2010), and others within conifers but close to Cupressophyta (non-Pinaceae conifers; the ‘‘Gnecup’’ hypothesis, Nickrent et al., 2000; Doyle, 2006). These alternative hypotheses are illustrated in figure 5-1A.

It has been reported that Gnetales have a faster substitution rate of sequence evolution than other gymnosperms, which could potentially cause a ‘‘long-branch attraction’’ (LBA) artifact in phylogenetic reconstruction (Zhong et al., 2010). The effects of LBA are well understood, even though the significance of contributing causes is often difficult to determine. These can include faster substitution rates in nonadjacent phylogenetic lineages (Felsenstein, 1978), poor taxon sampling due to extinction or limited availability of some taxa (Hendy and Penny, 1989), and properties of sequences not well described by stationary time reversible models. The latter include base compositional heterogeneity (Foster, 2004; Jermiin et al., 2004) and lineage-specific changes in evolutionary constraint that can alter the proportion of variable sites in homologs (Lockhart and Steel, 2005).

To improve taxonomic sampling of the Cupressophyta, we determined sequences for 52 genes from the chloroplast DNA (cpDNA) genomes of *Halocarpus kirkii*, *Podocarpus totara*, and *Agathis australis* using Illumina GAII sequencing. In phylogenetic analyses of concatenated seed plant chloroplast genome sequences, we demonstrate that sites exhibiting greatest character state variation are not well described by a time reversible substitution model. We show that this data property significantly impacts on the reconstruction accuracy of seed plant phylogeny.

Materials and Methods

Sample Collection and DNA Sequences Tissue for Cupressophyta (*H. kirkii, P. totara, and A. australis*) was obtained with permission from the living collection at Massey University, Palmerston North. Chloroplasts were isolated and enriched DNA sequenced using the protocols described in Atherton et al. (2010). Short reads were filtered for the longest contiguous subsequences below 0.05 error probability using DynamicTrim (Cox et al., 2010). Filtered reads were assembled with Velvet (Zerbino and Birney, 2008) and a k-mer range from 23 to 63. Contigs were further assembled using the Geneious assembler (Drummond et al., 2011). Initial annotations for protein coding genes were carried out using DOGMA (Wyman et al., 2004). Annotations were manually refined by comparison with genes of more closely related species.

We retrieved 13 cp genomes from the NCBI database, including the three genera of Gnetales, one Cupressophyta conifer (*Cryptomeria japonica*), three representatives of Pinaceae conifers (*Pinus thunbergii, Pinus koraiensis, and Keteleeria davidiana*), and three species from the *Cycads/Ginkgo* group, with three angiosperms representing the outgroup. GenBank accession numbers for gene sequences used and determined in the present study are listed in supplementary table S1 (Supplementary Material online). Fifty-two protein-coding genes were first aligned as proteins using MUSCLE (Edgar, 2004). Gaps were excluded from these alignments so that only blocks of ungapped residues bounded by similar or identical amino acids were used in phylogenetic analyses. Se-Al v2.0all (Rambaut, 2002) was used to edit the underlying DNA sequences into the amino acid alignments. These alignments were then concatenated using Geneious v5.4 (Drummond et al., 2011). This approach produced an alignment of 33289 ungapped positions (not divisible by three as some gaps occur in Genbank sequences).

Sorting Sites Based on Character State Variation

The positions in our concatenated alignments were sorted based on their character state variation. As we demonstrate, this facilitated the study of systematic error in these data. Several methods have been suggested for ordering sites (e.g., discussed in Hansmann and Martin, 2000; Goremykin et al., 2010). We used the method of observed variability (OV) sorting as described in Goremykin et al. (2010), which previously has been found to be efficient in concentrating saturated positions toward the most varied end of the sorted alignment. The alignment was ordered from the most highly varied sites to the most conserved sites, and a series of alignments was generated by successively shortening the OV-sorted alignment in steps of 250 sites. For each shortening step, two data partitions were obtained: 1) the shortened alignment containing the most conserved sites (partition ‘‘A’’) and 2) an alignment containing the more varied sites (partition ‘‘B’’). After model fitting for each partition data, the maximum likelihood (ML) distance and uncorrected p distance were calculated using PAUP* (Swofford, 2002). Two Pearson correlation analyses of pairwise distances were conducted at each shortening step: 1) correlation of the ML and uncorrected p distances for partition B and 2) correlation of the ML distances for partition A and B. The stopping point for site removal was determined as the point at which the two correlations showed a significant improvement (Goremykin et al., 2010).

Data Model Fit

We used MISFITS (Nguyen et al., 2011) to determine the occurrence of site patterns in our sorted alignment that were unexpected under a general time reversible (GTR) + G model using three alternative Gnetales phylogenetic trees incorporated as part of the evolutionary model. That is, given a GTR + G substitution model and weighted tree, the expected pattern likelihood vector was computed. For each entry in the vector, a simultaneous α=95% Gold confidence region was calculated. Sequence positions in the alignment indicating unexpected patterns were recorded. We also successively shortened our alignment by 250 positions and compared the log-likelihood scores for our OV sorted alignment (partition A) to log-likelihood scores for identical length partitions jackknife resampled from the complete 33289 position alignment. PhyML 3.0 (Guindon et al., 2010) was used for log-likelihood calculations. Seqboot, implemented in the Phylip3.6 package (Felsenstein, 2004), was used for jackknife resampling. Z-scores were calculated by subtracting the log-likelihood score on the original data from the mean log-likelihood score for the pseudoreplicate data sets and dividing by the standard deviation (SD) of mean scores.

Compositional Heterogeneity

MEGA5.0 (Tamura et al., 2011) was used to calculate the average nucleotide composition of 1) all codon sites, 1st + 2nd codon sites, and 3rd codon sites, and 2) intervals of increasing length (250 bp) beginning from the most varied end of the OV-sorted alignment. The SD of mean nucleotide frequencies was plotted to visualize compositional heterogeneity among taxa.

Phylogenetic Analyses

ML trees were built assuming a GTR + G model implemented in PhyML 3.0 (Guindon et al., 2010). The relative length of branches and extent of heterotachy (lineagespecific differences in evolutionary rate) in these trees was visualized using SplitsTree 4.0 (Huson and Bryant, 2006).

Results

Effect of Improved Taxon Sampling

In ML analyses of all codon positions and 1st + 2nd sites, inclusion of the newly determined sequences from three Cupressophyta genomes halved the length of the internal branch subtending Gnetales and Cupressophyta when compared with phylogenetic reconstructions made without these taxa. Inclusion of sequences from these additional genomes did not change the topology. In the trees with additional taxa, the Gnecup hypothesis (Fig. 5-1B) was strongly supported (96% and 97% bootstrap support for all positions and 1st + 2nd sites, respectively). However as we show below, support for this hypothesis was also strongly dependent on the inclusion of sites in the data that showed a poor fit to the GTR + G substitution model.

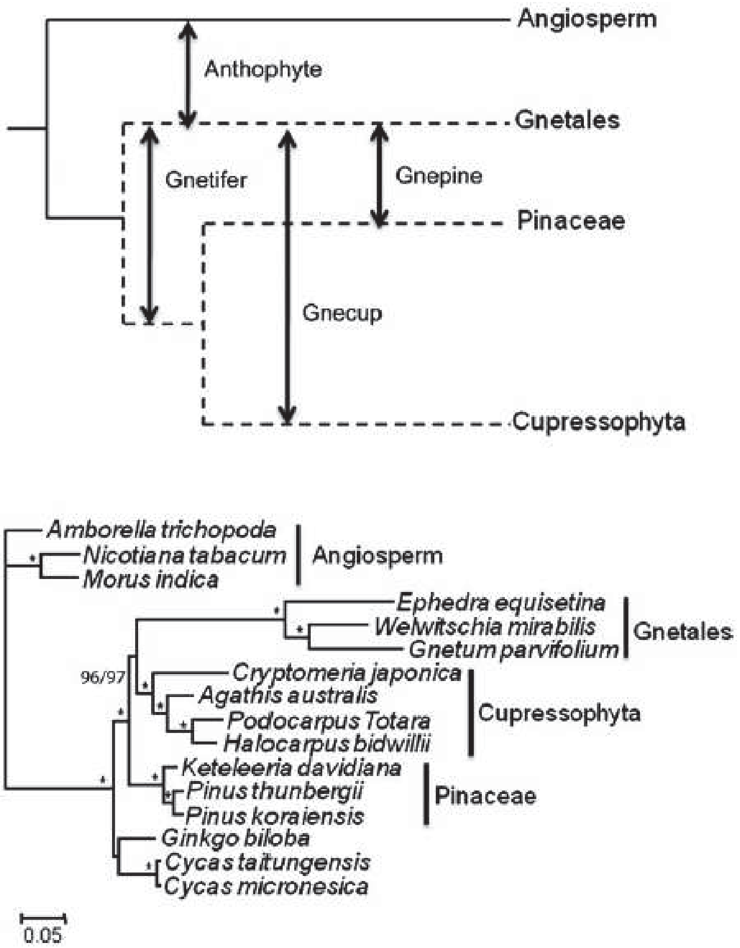


Fig. 5-1. (A) Four major hypotheses for phylogenetic relationships involving Gnetales. (B) Optimal PhyML tree (GTR + G substitution model) reconstructed from all codon positions. The same topology is obtained using 1st + 2nd position sites. Bootstrap support for Gnecup hypothesis is 96% for all sites and 97% for 1st þ 2nd position sites.

The impact of site removal

We used the OV sorting criterion of Goremykin et al. (2010) to rank site patterns from most varied to least varied. Blocks of columns in steps of 250 sites were then removed sequentially. This produced a series of shortened alignments. ML trees under a GTR + G model were reconstructed for each partition, and the bootstrap support for alternative hypotheses was measured for each partition. This analysis was made for all sites, 1st + 2nd codon position sites, and 3rd codon position sites. Figure 5-2A (all sites) shows that the Gnecup hypothesis was favored only while the 2000 most varied positions were included in the analysis. After these sites were removed, the Gnepine hypothesis became favored until 3,250 sites were removed. After this point, alternative hypotheses were unresolved.With 1st and 2nd codon position data alone, the Gnepine hypothesis was favored after removal of 750 sites and before removal of 1,250 sites (Fig. 5-2B). With 3rd codon position data, the Anthophyte hypothesis was initially weakly supported, but this support decreased as sites were removed (Fig.5-2C).

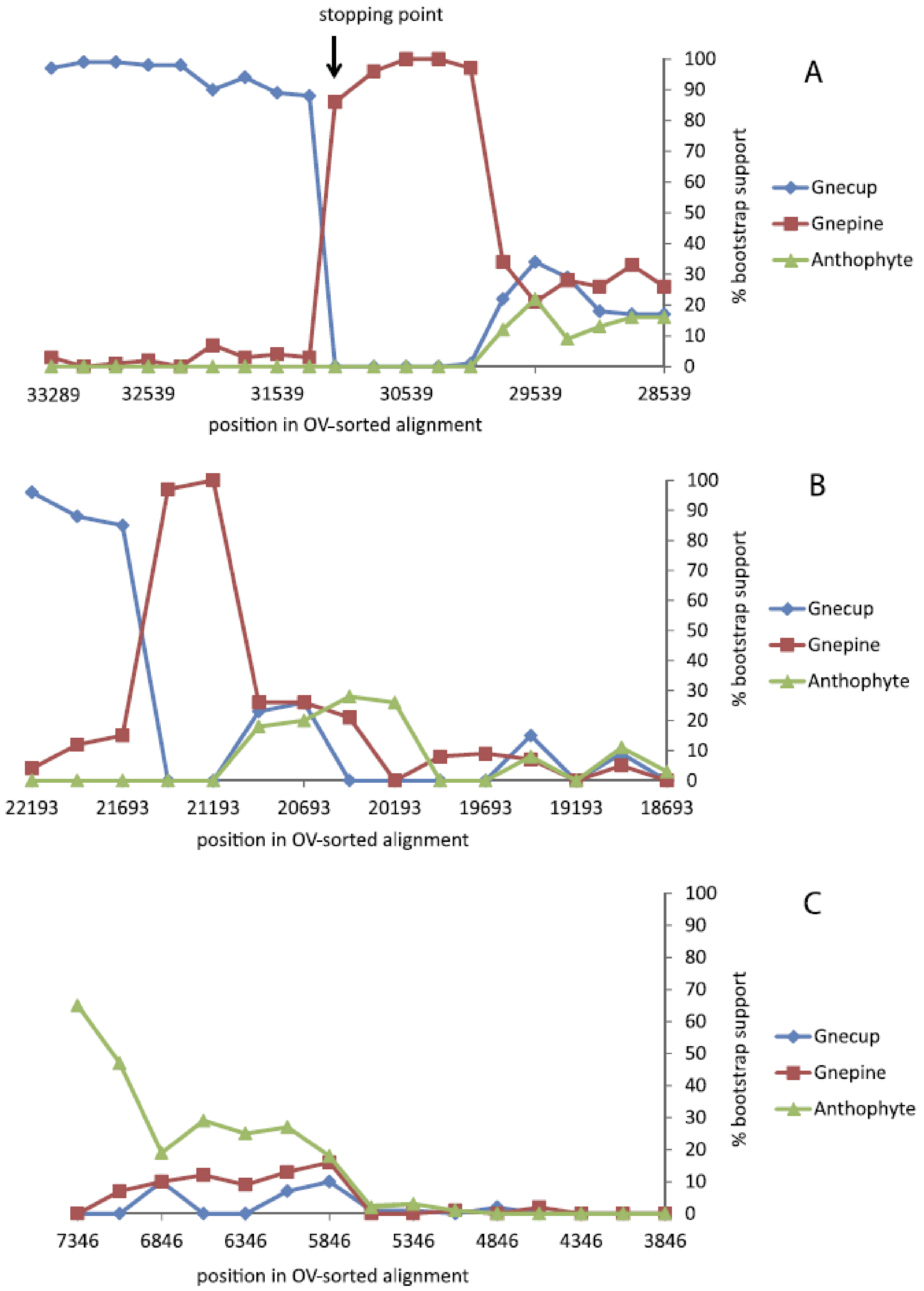


Fig. 5-2. Bootstrap support in optimal PhyML trees for three alternative relationships as intervals of 250 bases were successively removed from the OV-sorted alignment. (A) all sites, (B) 1st þ 2nd codon positions, and (C) 3rd codon positions.

Data Model Fit

To help understand the impact of site removal, we investigated the fit of site patterns to three alternative evolutionary models (Gnecup, Gnepine, and Gnetifer trees) that assumed an optimal GTR + G substitution model. Using MISFITS (Nguyen et al. 2011), we computed the overrepresented and underrepresented site patterns in the OV-sorted data. For the Gnepine hypothesis, we observed that 46%of the sites not fitting the evolutionary model occurred within the 2250 most varied positions (i.e., in 7% of the total alignment length; 15% of all variable sites). About 3.1% (691/22193) of the 1st þ 2nd position sites and 15.2% (1687/11096) of the 3rd position sites do not fit the Gnepine tree. A similar poor fit was also obtained for tree topologies that supported the Gnetifer and Gnecup hypotheses (Fig. 5-3), suggesting that in the most varied positions of the OV-sorted alignment, misspecification was a general property of the GTR + G substitution model and not specific to any one hypothesis of evolutionary relationship. To further evaluate the impact of the most varied positions on data model fit with our three tree models, we also compared the log-likelihood scores for the sequentially shorted (partition A) data sets, with scores for identical length data sets comprised of jackknife resampled site patterns taken from the original 33289 position alignment. The results from this analysis corroborated those obtained with MISFITS in identifying an extremely poor data model fit for sites at the most varied end of the OV-sorted alignment (supplementary Fig. S1, Supplementary Material online).

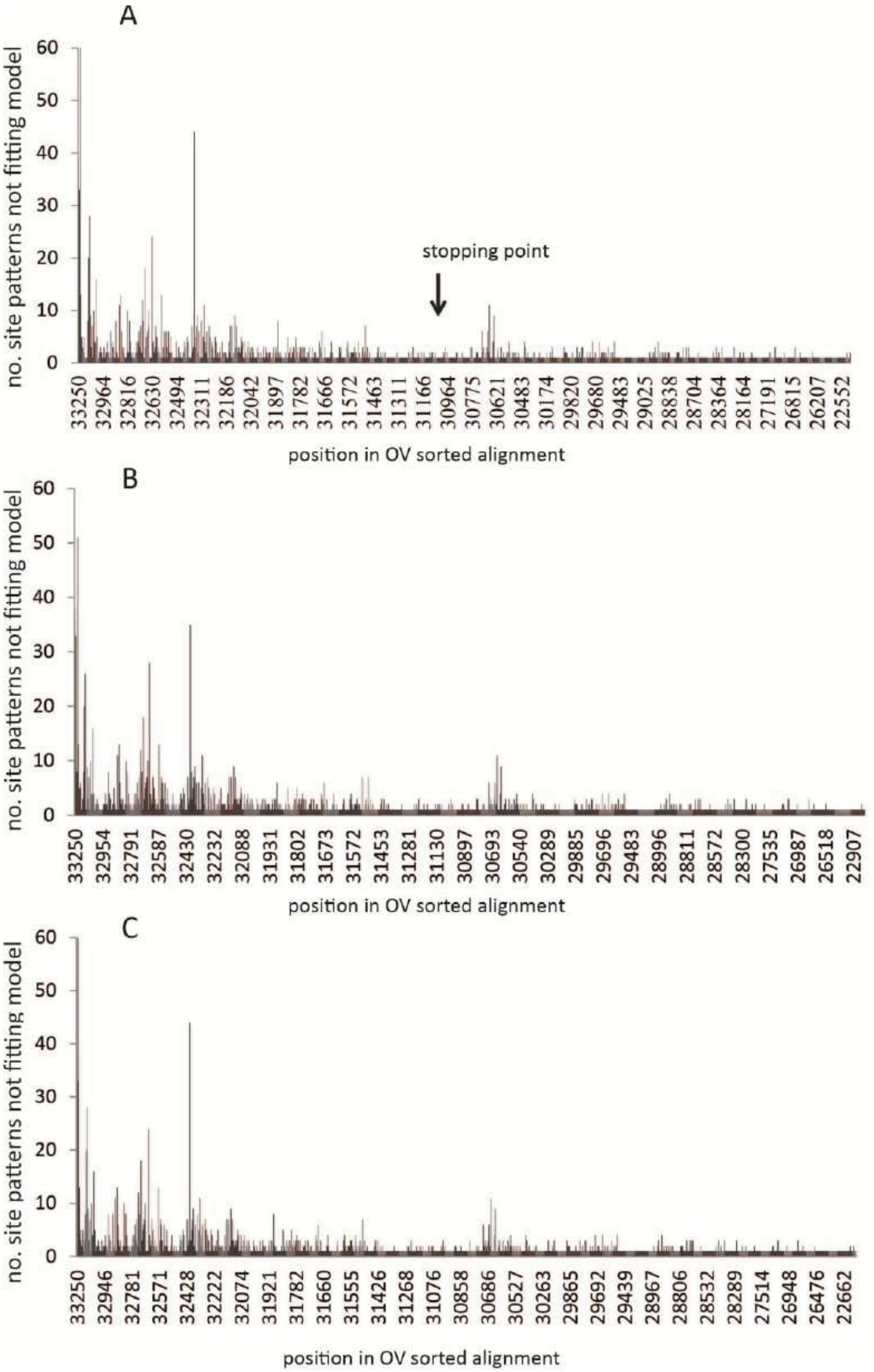


Fig. 5-3. Histogram indicating consecutive misfitting site patterns under the (A) GTR + G + Gnepine, (B) GTR + G + Gnetifer, and (C) GTR + G + Gnecup evolutionary model. The height of each histogram indicates the number of unexpected site patterns.

Compositional Heterogeneity

Figure 5-4 shows the SD of individual base frequencies from mean (stationary) estimates for intervals increasing in length by 250 bases sampled from the most varied end of the OVsorted alignment. While the average nucleotide compositional frequencies of all sites, 1st þ 2nd sites, and 3rd sites are relatively homogeneous (Results not shown), the most varied OV-sorted sites in the alignment exhibit significant compositional heterogeneity. This decreases incrementally toward the more conserved positions of the OV-sorted alignment.

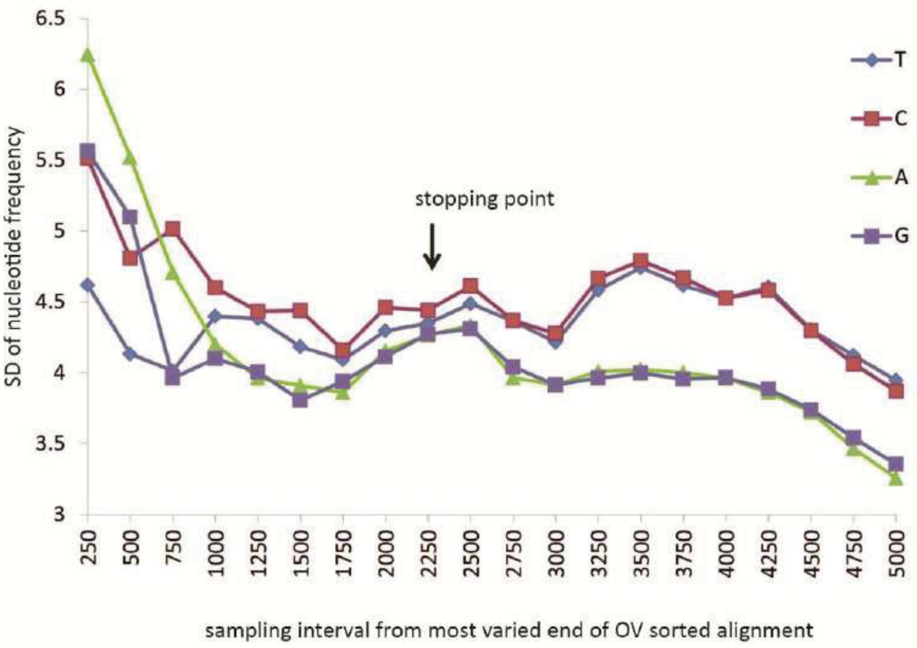


Fig. 5-4. Plot indicating nucleotide compositional heterogeneity within intervals sampled from the most varied end of the OV-sorted alignment. Subsequent intervals increased in length by 250 bases per interval.

Heterotachy

Optimal PhyML trees (GTR + G substitution model) were reconstructed for sampling intervals that increased in length by 250 bases from the most varied end of the OV sorted alignment. The relative length of the Gnetales internal branch separating Gnetales from other species in the 16 taxon data set for each sampling interval is shown in Figure 5-5 (A). The relative length of the branches subtending the Cupressophyta, Pinaceae and angiosperms in the 13 taxon data set are shown in Figure 5-5 (B). A striking feature of the 16 taxon trees is that the branch leading to the Gnetales lineage is disproportionately much longer than branches subtending other seed plant lineages (more than 70x longer over the first 1750 bases, and between 10x-5x between 2000 and 2500 bases) at the most varied end of the OV sorted alignment (Fig. 5-5). This extreme branch length difference is a feature of both the 1st+2nd codon position and 3rd codon position data (not shown).

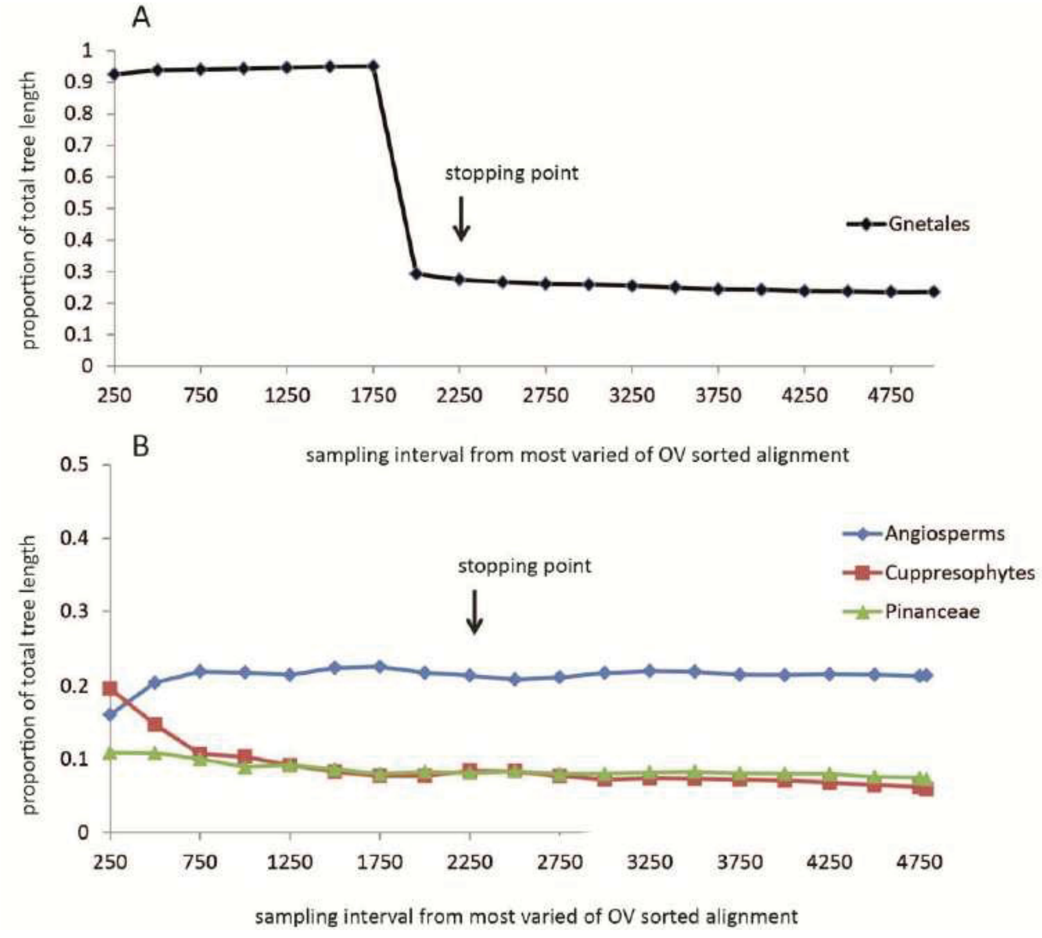


Fig. 5-5. Relative length of internal branch leading to (A) Gnetales in a 16 taxon data set, (B) Non-Pinaceae, Pinaceae, and Angiosperms in a 13 taxon data set (this second data set excluded Gnetales). The branch lengths are shown as a proportion of total tree length. Optimal PhyML trees were reconstructed for the same sampling intervals as used in Figure 5-4.

Removal of most varied sites from the alignment

We used the stopping criterion of Goremykin et al. (2010) to make an assessment of the number of most varied sites that should be excluded prior to tree building. This criterion considers the alignment partitions created by the sequential shortening steps described previously and compares (i) ML distances for the conserved (A) and the variable (B) bipartition and (ii) *p*-distances and ML distances for the B partition. The authors have suggested that the removal of variable positions should be continued at least until the very end of the sharp rise in Pearson correlation values in either analysis. The stopping criterion identifies the point where the substitution properties of most varied sites (partition B) becomes more similar to those of the more conserved sites in the alignment (partition A), and where corrected and uncorrected distances for the variable B partition begin to show a strong positive correlation. As such it provides a means to objectively decide a cut off point for excluding from tree building sites that exhibit site saturation and or model misspecification. Figure 5-6 indicates change in the correlation coefficient (*r*) and similarity of distances estimates as sites are removed. A sharp rise in (*r*) occurs when 2000 sites have been removed and it ceases with removal of 2250 sites in the correlation of *p*-distances and ML distances estimated from B partitions. Reference to Figure 5-5 shows that this is accompanied by reduction of heterotachy associated with the Gnetales lineage. It also marks the transition zone for bootstrap support of the Gnecup and Gnepine hypotheses. The Gnepine hypothesis is strongly favored after removal of 2250 sites (position 31039). It continues to be favored until 3,250 sites are removed when the PhyML trees become unresolved.

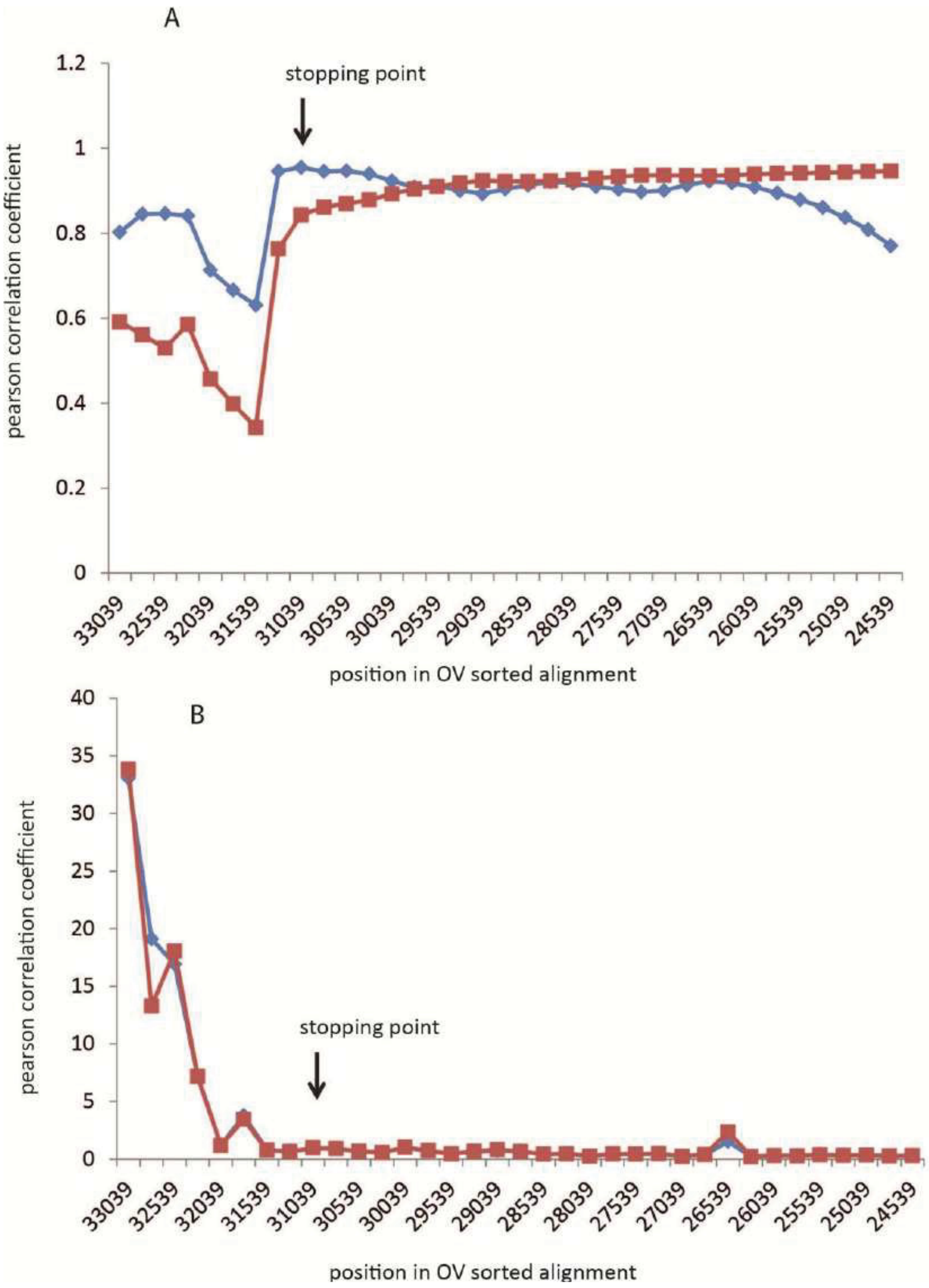


Fig. 5-6. (A) Pearson correlation analyses. The blue dotted line indicates the Pearson correlation coefficients (*r*) of ML distances for (the more conserved) partition "A" and (less conserved) partition "B". The red dotted line represents *r* value of uncorrected *p*-distances and ML distances for partition "B". The *r* values begin to increase sharply at the 8th shortening step (31,289 position remained). (B) Mean deviations of ML distances from *p*-distances for "B" partitions. The red dotted line shows deviations between *p*-distances and ML distances calculated using the best-fitting ML model as determined by ModelTest (Posada and Crandall, 1998) using the AIC criterion. (The NJ tree was used to estimate ML model parameters). The blue dotted line indicates the deviation between *p*-distances and ML distances calculated as above, but using an ML tree to fit model parameters.

Discussion

Most phylogenetic methods assume that DNA sequences have evolved under stationary, reversible, and homogeneous conditions. Violation of this model assumption is well known to lead to inaccurate tree reconstruction (e.g. Lanave et al., 1984; Lockhart et al., 1994; Foster, 2004; Jermiin et al., 2004; Delsuc et al., 2005; Lockhart and Steel, 2005). Our Misfit analyses indicate a poor fit between the most varied nucleotide sites in the Gnetales chloroplast concatenated data set and a GTR+G model - one of the more general models of substitution currently used in phylogenetic reconstruction. Although more complex mixture models exist (e.g. such as the CAT model, Lartillot and Philippe, 2004), like GTR+G they also assume a stationary distribution of base frequencies and have the expectation for a constant proportion of variable sites in all sequences.

Deviation from compositional homogeneity occurs in the most varied regions of the OV sorted alignment. However, this heterogeneity extends past the OV sorting stopping point and shows no obvious relationship to it. Thus, compositional homogeneity appears an insufficient explanation for the significant increase in value of the Pearson statistic after removal of 2000 sites and an insufficient explanation for the extent of poor model fit observed in the most varied part of the OV sorted alignment.

More important for explaining the sharp rise in the Pearson statistic is the extent of substitution rate difference inferred for the Gnetales lineage across the sampling intervals at the most varied end of the OV sorted alignment. This property of the aligned data causes high variance in ML distance estimation between Gnetales and other species when estimates are made from B partitions. This property of the sorted data explains much of the Pearson coefficient behavior in the correlation analyses. By the final shortening step, at 2250 bases, the relative length of the internal branch separating Gnetales shows approx. 60x reduction in length. This reduction is accompanied by a rapid change in the bootstrap support for the Gnepine hypothesis.

The extreme branch length differences between Gnetales and other lineages for sites at the most varied end of the OV sorted alignment suggests an issue with alignment of some amino acid positions, despite a conservative approach being used in generating the sequence alignments in the present study. To investigate this further we also aligned seed plant DNA sequences using the approach of Goremykin et al. (2010) and excluded regions of low sequence similarity (analyses not shown). Working with these alignments we also obtained very similar results and conclusions regarding heterotachy, compositional heterogeneity, misfit analyses and bootstrap support. Thus, we conclude that heterotachy is a strong feature of the data, and is not a feature of a specific alignment method.

Very recently, a similar study has been undertaken to that reported here. Wu et al. (in press GBE) have determined chloroplast genome sequences for five Cupressophytes and a cycad. They also studied the phylogenetic placement of Gnetales with respect to other seed plants. Our general conclusions are similar to theirs – phylogenetic reconstruction of Gnetales in seed plant phylogeny is misled by non-time reversible properties of chloroplast sequence evolution. From their sampling of taxa, Wu et al. obtain stronger evidence than we do for lineage specific change in the Cupressophyta that parallels Gnetales. Our studies also differ in that these authors did not evaluate the relative contribution of compositional heterogeneity and heterotachy in causing problems for tree building. Our analyses suggest that heterotachy is a more significant cause of systematic error in the seed plant sequences analysed. As we have discussed below, our analyses also suggest that removal of sites rather than individual genes provides a better strategy for dealing with this problem.

Wu et al. divided chloroplast sequences into L (low heterotachy) and H (high heterotachy) genes and provide evidence that only phylogenetic inference from genes in the L dataset is reliable. The H data set contains genes involved in translation including the rpo genes, which previously have been shown to exhibit non conservative substitutions, indels and increased proportions of variable sites in green algae (Lockhart et al., 2006). Our analyses indicate that while heterotachy is most pronounced in genes of the H data set, a significant level of heterotachy also occurs in the L data set for conifers that we have studied (not shown). There is also a significant amount of useful phylogenetic information in the H genes, as indicated from our results that favour the Gnepine hypothesis. This conclusion is based on an analysis of 31,039 sites, while that of Wu et al. is based on 21945 DNA positions (7315 amino acids in the L data set). In general we suspect that it will be more phylogenetically informative to remove model violating sites rather than genes prior to phylogenetic analyses.

Wu et al. suggest that the example of Gnetales follows the classic LBA scenario of Felsenstein (1978), wherein there is LBA between Gnetales and Cupressophyta. However, it is important to note that whilst similar, the LBA scenario for seed plants is likely to differ from this. The properties of seed plant sequences better fit the LBA scenario described by Lockhart and Steel (2005) in which proportions of variable sites change in a lineage specific fashion, and where parallel changes occur (Zhong et al., 2010) because of similar proportions and convergent patterns of variable sites (modelled in Gruenheit et al., 2008). The significance of the difference in scenarios is important because current methods of tree building do not model lineage specific change the proportion of variable sites in homologues (Lockhart and Steel, 2005; Lockhart et al., 2006; Gruenheit et al., 2008; Shavit, 2008). Although, it is possible to model changes in proportions of variable sites using branch length mixtures, these can be complex under some scenarios and thus problematic to identify (Gruneheit et al., 2008; Matsen and Steel, 2007; Lartillot et al., 2009). Further to this, Wu et al. observe that a mixture branch lengths model was unsuccessful in alleviating LBA with the H data set.

Conclusions

Observations of a poor fit between fast evolving sites and time reversible models such as the GTR + G model of sequence evolution are not novel (e.g. Sullivan et al., 1995; Goremykin et al., 2004). However, the significance of having a poor fit becomes much more obvious in analysis of concatenated sequences. In the present study, systematic error arising from lineage specific differences in evolutionary constraint dominates phylogenetic signal and misleads phylogenetic reconstruction. When systematic error contributing to most of the model misfit is removed prior to tree building, our analyses favor the Gnepine hypothesis for seed plant phylogeny (Chaw et al., 2000; Bowe et al., 2000; Zhong et al., 2010; Finet et al., 2010; Soltis et al., 2011; Wu et al., 2011).

We studied site removal in the context of substitution model misspecification and the stopping criterion of Goremykin et al. (2010). In respect of this, our study provides more insight into the performance of this method. Our results indicate that use of the stopping criterion also removes sites that provide a poor fit to tree building assumptions. While this criterion does not remove all model violating sites from data, it removes sites that significantly impact on phylogenetic estimates, and thus sites most important for misleading tree building. Thus it provides a useful tool to guide phylogenomic analyses.

Wu et al. note that improved taxon sampling was insufficient to overcome LBA between Curessophytes and Gnetales. We also obtained this result. However, we wish to be more positive about the contribution that improving taxon sampling of conifers will make to phylogenetic reconstruction of seed plant phylogeny. In our study, addition of sequences from three Cupressophytes reduced the length of the internal branch leading to Gnetales and Cupressophytes two fold, even if it was not sufficient to change the topology. Together with international efforts currently underway to sequence and analyze conifer genomes, we believe that analytical approaches such as those used here will be essential for evaluating and mitigating the impact of systematic error in large scale phylogenomic data sets for seed plants.

Acknowledgements

This work was financially supported by the Allan Wilson Centre, Massey University, the New Zealand Royal Society and the D.A.A.D. We thank Jennifer Tate, the anonymous reviewers and Associate Editor for helpful comments.

Supplementary table S1, figure S1, and data matrix concatenated gapped alignment are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals. org/).

6 Final discussion

6.1 Use of cpDNA for phylogeny reconstruction at shallow taxonomic level

Elucidating phylogenetic relationships within *Malus* has been particularly thorny (chapter 2), and no well-resolved topology for the genus has ever been published. This problem is not unique to the genus, as infrageneric relationships are routinely investigated with the set of methods, such as genetic fingerprinting and marker-based phylogeny reconstruction, which, as it was discussed above (chapter 1.1.1 & 1.1.2) have serious flaws. Thus, demonstrating the methodological feasibility of accumulating genome-scale chloroplast data using the NGS approach and its advantage in terms of tree resolution should provide a guideline for the future studies aimed at clarifying phylogenetic relationships on the shallow taxonomic levels.

Apple (*Malus × domestica*) has been cultivated more than 4,000 years (Zohary and Hopf, 1994, Luby, 2003). It has been introduced to Europe by Romans and Greeks, and then from Europe it had spread all over the world (Juniper and Mabberley, 2006). It was assumed to be originated either in Europe, from *M. sylvestris*, a European crab apple bearing small astringent and acidulate fruits (Zohary and Hopf, 1994; Coart et al., 2006; Harrison and Harrison, 2011) or in Asia, from *M. sieversii* (Velasco et al., 2010; Micheletti et al., 2011; Cornille et al., 2012), a diverse central Asian species, characterized by a wide range of forms, colors and flavors (Way et al., 1990), which is capable of producing sweet fruits.

Yet the evidence presented so far in defense of both hypotheses of domesticated apple origin is not conclusive (see Introduction 1.3.1). Single marker-based phylogenetic analyses and genetic fingerprinting approaches failed to provide good resolution within the genus, and multigene data sets did not help to shed more light onto the issue so far. Absence of statistical support for the tree branches (Harrison and Harrison, 2011; Velasco et al., 2010, Fig 5), low bootstrap support for any branch subtending *Malus domestica* (MD) accessions plus non-MD accessions obtained in ML re-analysis of the character-richest data set built so far to elucidate evolutionary history of the genus (Micheletti et al., 2011, Fig 1a), absence of crucial species in other analysis of other concatenated nuclear markers (Forte et al., 2002; Lo and Donoghue, 2012) suggests that gathering further evidence on the issue is necessary.

Taking this into account, my goal was to provide a solid phylogenomic foundation for understanding of apple domestication. As reports of hybridization among domesticated apple, *M. sieversii* and *M. sylvestris* are abundant (Cornille et al., 2012; Coart et al., 2006; Harris et al., 2002; Robinson et al., 2001), I chose to sample a number of accessions representing these species which would allow identifying the main contributor to the maternal line of *M. x domestica* by majority rule.

Results of phylogeny reconstruction based on alignment of chloroplast genomes indicate that *M. x domestica* cultivars have diverse cytoplasmic lines of inheritance. The backbone of the tree is well-resolved. *Malus x domestica* cv. Granny Smith has chloroplasts sharing monophyletic origin with the Asian accessions (*M. sieversii, M. baccata, M. mandshurica, M. halliana and M. hupihensis*). However, the majority of the varieties analyzed have cpDNA derived from *M. sylvestris*. Comparison of the topology of the corresponding subtree (Branch B in Fig 2-2) with the geographic origin of the *M. sylvestris* specimens reveals that the German *M. sylvestris* accessions (shown as 3, 4, and 5 in Fig. 2-2) are unrelated to cultivated apple sorts, while southern European accessions are. Six *M. x domestica* cultivars share the sister group relationship, supported by maximum bootstrap support value, with *M. sylvestris* specimens, collected on Monte Pollino, Calabria, Italy (*Malus sylvestris* 1 in Fig. 2-2). Two other cultivars build a common branch, supported by maximum BP value, with *M. sylvestris* accession collected in Macedonia (PI 369855, *M. sylvestris* 2 on Fig. 2-1). This finding is in good agreement with the scenario according to which systematic apple cultivation had started and had being maintained throughout long period of time in southern Europe.

The dating results (highlighted in pink on Fig 2-3) indicate that three major chloroplast lineages of *M. x domestica* suggested by the tree structure have diverged millions years ago, long before the origin of modern man. Even divergence among chloroplast lines of apple cultivars of *M. sylvestris* ancestry and their closest wild *M. sylvestris* progenitors took place more than 1 million years ago, long before historical times.

Cornille et al. (2012) also have registered close genetic affinity between the genomes of *M. x domestica* and *M. sylvestris*, exceeding that between *M. x domestica* and *M. sieversii*. Given admission model parameters accepted, this was explained by recent massive introgression of genetic material from *M. sylvestris*. The reasons facilitating introgression from *M. sylvestris*, into the *M. domestica* line, were assumed to be self-incompatibility, a long lifespan, and cultural practice including selection from open-pollinated seeds.

It should be noted that these factors do not explain the tree topology obtained here. Because chloroplast DNA is inherited by maternal line in Rosaceae (Hu et al., 2008), any pollination by *M. sylvestris* would not had led to formation of the branch B (Fig. 2-2). Given difficulty of distinguishing introgression from incomplete lineage sorting/parentage (Velasco et al., 2010; Harrison and Harrison, 2011, Micheletti et al., 2011) from *M. sylvestris*, which can explain the branch B, the validity of conclusions drawn by Cornille et al. (2012) should be re-examined. All the more so, because the shallow time scope for separation events of *M. sylvestris* and *M. x domestica* (17,700 years ago) and of *M. sieversii* and *M. sylvestris* (83,250 years ago) estimated based on arbitrarily chosen admixture model settings (Cornille et al., 2012), appears improbable given much older dating obtained based on present phylogenetic analysis of the complete chloroplast genomes, even when the constraints for the age of *Malus* were set at the unrealistically early range of values.

Given the observed tree topology, acceptance of *M. sieversii* as the main ancestor of *M. × domestica*, advocated by many authors (e.g. Janick et al., 1996; Robinson et al., 2001; Harris et al., 2002; Velasco et al., 2010), would imply a breeding scheme involving *M. sylvestris* as mother with back-crossing using pollen belonging to the “sweet apple” genetic line(s) to eliminate astringent and sour components in fruit taste. Such a scheme was, for example, employed in creation of scab-resistant apple cultivars, by incorporating the *Vf* gene from another crab apple, *Malus floribunda* 821 into *M. x domestica* (Crosby et al., 1992) and subsequent extensive back-crossing with *M. x domestica* varieties.

Taking into account that the known pedigree of cultivars analyzed (Table 2-1) does not include such breeding scheme, this should have happened before the origin of the founder varieties (Table 2-1). The most parsimonious explanation would include only two independent pollinations of *M. sylvestris* leading to branches B1 and B1 (Fig. 2-2). Studying the pedigree of *M. x domestica* cultivars forming branch B (Fig. 2-2) reveals that their maternal lines can be traced back to seven old “founder” sorts (Table 1). The origin of the oldest of the cultivars forming branch B2, Ribston Pippin, can be traced back to the seed brought from Rouen (Normandy) to England in 1707 (Cecil, 1910). McIntosh, the sort with the oldest pedigree from branch B1 was discovered in Ontario, Canada in 1796. Taking into account that apple breeding continued haphazardly before introduction of the controlled breeding schemes at the start of 1800s (Sandlers, 2010), appearance of any elaborated breeding scheme preceding the origin of B1 and B2 branches is unlikely. Even until the latter half of the twentieth century most of the world's apple cultivars were chance seedlings selected by fruit growers (Janick et al., 1996).

However, the same breeding outcome might have been facilitated by massive inclusion of local species into cultivation of *M. domestica* **(**Hokanson et al., 2001; Robinson et al., 2001**).** In fact, planting apple trees from forest to gardens using root suckers was a widespread practice in central Asia (Ponomarenko, 1983**):** the benefits of planting best apple individuals close to human dwellings were apparent enough that people from various places might have adopted this practice. Subsequent uncontrolled pollination among genetically heterogeneous apple cultivars, substantial proportion of which has had maternal *M. sylvestris* pedigree, would have produced results we obtained.

Following this scenario, large divergence times observed between chloroplast lineages of the cultivated sorts reflect evolution of paths of their diverse progenitors, which have developed under selectional pressure in natural habitats to produce fruits attractive to animals spreading seeds. Formation of the desirable fruit characters, which gave incentive to domesticate best apple individuals, has apparently gone on for a long time in natural conditions. Accumulation of sugars, attracting birds and animals was facilitated by intense insolation levels in souther Europe and Central Asia.

Our analysis has also indicated that different *M. sieversii* specimen harbor chloroplast DNA lines, showing more close relationships with other Asian species than with each other. An observation that allozyme diversity within this species is significantly greater than that found in four widely distributed North American wild apples (Lamboy, 1996; Dickson et al., 1991) also suggests heterogeneous nature of lineages united in *M. sieversii*. As *M. sieversii* appears to be an assemblage of unrelated lines and not a monophylum, its status as species should be revised. Lack of congruency between the tree structure and the taxonomic assignment of species to various sections within *Malus* points to the need to revise the latter. Future comprehensive phylogenomic analysis, including all the wild species is thus warranted.

This analysis demonstrates that using chloroplast DNA for phylogeny reconstruction within *Malus* has led to the level of resolution unattained with all previously used approaches. This study thus should constitute a useful framework for future phylogenetic analyses on shallow taxonomic levels.

6.2 Use of cpDNA at deep taxonomic level

In contrast to intrageneric analyses, on high taxonomic levels, with increased saturation among the sequences, current methods of phylogeny reconstruction are becoming error-prone. Appearance of conflicting tree topologies of basal angiosperms and gymnosperms in the literature (see Introduction) represent a compelling evidence of such errors.

6.2.1 Noise reduction

When constructing phylogenetic trees by maximum likelihood method, the algorithm tries to predict the whole substitution path from one sequence to another one, based on every position in alignment. Consequently, likelihood computation for a given internal branch/position is very dependent on correct ancestral character state identification on the tips of the internal branch. In highly saturated regions (and, as Fig. 3-2 shows, positions with 30 subst./site and higher are abundant in the cpDNA alignment of seed plants) correct identification of ancestral character states becomes impossible, as there is not phylogenetic signal left in the data. In that case, removal of the most saturated positions at which the fit between model and the data is poor makes sense.

Currently there are several methods, which sort characters based on some proxy of substitution rate (Ruiz-Trillo et al., 1999 and Hirt et al., 1999; Brinkman and Philippe, 1999; Pisani, 2004). Ideally, sorting results of these methods should not depend on the input tree, since they are needed exactly when the tree structure is unknown and is potentially affected by LBA. A general requirement to these methods is that they should effectively concentrate noisy positions towards one alignment end and ensure the quickest removal of LBA artifacts during character-stripping.

In the case study, wherein we investigated the known LBA phenomenon, our method (Goremykin, Nikiforova, Bininda-Emonds, 2010) performed far better than sorting based on the gamma rate assignment (Ruiz-Trillo et al., 1999 and Hirt et al., 1999) or based on parsimonious tree length (Brinkman and Philippe 1999) or other sorting methods based on the various implementation of compatibility principle (Pisani, 2004, COMPASS program by Simon Harris (www.ncl.ac.uk/microbial_eukaryotes/downloads.html) in i) recovering benchmark clades (LBA removal), and ii) in achieving highest difference between observed and ML distances for the most variable partitions (which shows effective noise concentration).

The problematic of finding the stopping criterion (Pisani, 2004) was addressed in Goremykin, Nikiforova, Bininda-Emonds (2010) for the first time (in history of phylogenetics). The stopping criterion based on two tests - i) correlation between observed distances and evolutionary distances in variable partitions, and ii) correlation between ML distances in the conserved partitions and ML distance in variable partitions helps to visualize the most variable part of the sorted alignment wherein signal structure as determined by ML model does not resemble both the observed signal structure as estimated using p-distances, and the signal structure estimated by ML in rest of the data. Within the framework of the criterion, the lack of resemblance is attributed to errors in correcting for multiple substitutions by the ML model for the most variable positions.

Practically, we observed that removal of variability in the scope approved by our stopping criterion was sufficient to liberate the tree topology from all four known LBA artifacts and from most of the noise present in the alignment.

It is important to stress that the method removes just one potential source of systematical error. Its application does not guarantee the absence of systematical errors related, e.g. to heterotachy or heterogeneity in a residual alignment. Heterotachy and heterogeneity in alignment were shown to be related to saturation (Rodriguez-Ezpeleta et al., 2007), yet this relationship is not always direct (Zhong et al., 2011; Goremykin et al., 2012). Thus, in parallel to the correlation analysis resulting in the stopping criterion, further analytical tests (analysis of model misspecification, analysis of presence or absence of heterotachous branches and of compositional heterogeneity) are desirable. These will help to make decision when the character-stripping should cease most judicious. Examples of such analytical tests have been published (Zhong et al., 2011; Goremykin et al., 2012). The developed approach provides a useful tool to guide phylogenomic analyses of nucleotide data, which is in suspect of LBA.

6.2.2 Evolution of basal angiosperms

Understanding the phylogenetic relationships among basal angiosperms is a long standing problem of much interest. The difficulty for phylogenetic reconstruction is exacerbated by the shape of the underlying tree topology and model misspecification (LBA; Hendy and Penny, 1989; Lockhart et al., 1996; Shavit et al., 2007). While concatenating large number of genes improves information content, it does not deal with problems of LBA and model misspecification (Rodriguez-Ezpeleta et al., 2007). The fit between substitution model and the alignment can be improved by either removing sequence positions that are difficult to model (e.g. Rodriguez-Ezpeleta et al., 2007; Goremykin et al., 2010) and/or by applying more biologically realistic models such as with site-heterogeneous mixture (CAT) models (Lartillot et al., 2007; Rodriguez-Ezpeleta et al., 2007; Pagel and Meade, 2005; Gruenheit et al., 2008). While both parsimony and model based methods can be susceptible to LBA, it has now been repeatedly observed that MP has produced a number of trees with an unexpected rooting of angiosperms, including *Ceratophyllum* basal-most (Les, 1988; Les et al., 1991; Chase et al., 1993), monocot/s basal-most (Goremykin et al., 2003b, Chang et al., 2006, Ravi et al., 2006) and eudicots basal-most (Graham and Iles, 2009). Therefore, its continued use as an optimality criterion for inference of basal angiosperm topology is difficult to justify.

This is not a trivial note, because in the last decade a number of studies based on, mostly, the same cpDNA-encoded genes appeared, where sister group relationship of *Amborella* and all other angiosperms was inferred or corroborated using MP (Leebens-Mack et al., 2005; Jansen et al., 2006; Qiu et al., 2006; Graham and Iles, 2009). For a long time the trend was that results of this kind received strong support in MP analyses, but lower support in model-based analyses within ML or Bayesian frameworks (Leebens-Mack et al., 2005; Jansen et al., 2006; Qiu et al., 2006; Graham and Iles, 2009), whereas topology where *Amborella* plus Nymphaeales branch was a sister to the remaining angiosperms never supported by MP, but was recovered in a number of model-based analyses (Leebens-Mack et al., 2005; Jansen et al., 2006; Bausher et al., 2006; Qiu et al., 2006; Goremykin and Hellwig, 2006, 2009; Graham and Iles, 2009). Since it is long known that MP is more susceptible to long branch attraction compared to the model-based methods (Felsenstein, 1978), this observation was taken as a reason to doubt the former group of results (Qiu et al., 2006). Graham and Iles (2009), observing attachment of the outgroup to the terminal branch with MP (previously suggested to be the consequence of randomized signal between in and outgroup (Graham et al., 2002)) also expressed doubts regarding correct MP tree inference for the basal angiosperms.

As our findings (Goremykin et al., 2013) suggest, the use of time reversible substitution models, such as GTR also need to be applied with caution. The support for the basal-most placement of *Amborella* in ML/Bayesian analyses is founded on the most variable sites, their evolution being poorly described by currently available time-reversible models (Goremykin et al., 2013), including even the CAT model, which is more robust to LBA. Removal of the sites causing model mis-specification reveals that the clade uniting surviving relatives of the most basal lineage of flowering plants is fairly species-large, and includes *Trithuria* (12 species), Nymphaeaceae (> 50 species), and *Amborella trichopoda*.

So do we have a definitive answer for relationships among basal angiosperms? It is most probably so for the available taxa that have been sequenced. Definitely, the basal-most position of Amborella is an artifact related to model misspecification at the most saturated sites. The fact that this LBA artifact has been published so often highlights the importance of the general issue which is inherent to current methods of phylogeny reconstruction, namely, inference of the correct placement of a distant outgroup within the ingroup radiation. Removing sites where the patterns of substitution deviate significantly from model assumptions is one way to obtain more realistic tree topology with currently available substitution models in such cases.

Given the impact of sampling on reconstructed trees, it remains to be seen if the basal-most branch of angiosperms, including Nymphaeales, *Trithuria* and *Amborella*, supported in this work, will be expanded by other taxa with broader number of species under analysis

It should be noted in this respect, that the taxon-wise most representative data set assembled so far (Soltis et al., 2011), including 640, mostly angiosperm, species consistently yielded the basal-most branch (*Amborell*a(*Trituria*(Nymphaeales s.s.))), supported by high (92-94%) bootstrap proportion values in our ML re-analyses employing GTR+I+G model.

This statistically well-supported result stands in sharp contrast to the conclusion of the study, namely that "Amborellaceae, Nymphaeales, and Austrobaileyales are successive sisters to the remaining angiosperms." Without careful analytical explanation, why results of the ML analysis with the globally optimal model, broadly applied by the authors in previous studies (Goremykin and Hellwig, 2006, 2009; Moore et al., 2007; Qui et al., 2010; Finet et al., 2010) were not reflected in the conclusions by Soltis et. al. (2011), acceptance of the above-cited basal-most angiosperm topology cannot be justified based on the data (Soltis et al., 2011).

Our result corroborates indications stemming from the fossil evidence, that the basal-most lineage of flowering plants is dominated by aquatic forms (Friis et al., 2001, 2003; Sun et al., 2002).

6.2.3 Evolution of seed plants

Phylogenetic placement of an aberrant group of gymnosperms, Gnetales, is another "classical" thorny and much-debated issue in systematic botany. The methodological issues complicating phylogenetic inference in this case are, in fact, similar to those affecting placement of non-angiosperm outgroup within the angiosperm radiation. Gnetales, characterized by elevated substitution rate in cpDNA (Zhong et al., 2010), are born on a long branch in all recent multi-gene analyses of chloroplast data (e.g. Zhong et al., 2010; Finet et al., 2010; Soltis et al., 2011; Wu et al., 2011), and time reversible models such as the GTR + G model of sequence evolution do not describe evolution of the fast evolving sites well (e.g. Sullivan et al., 1995; Goremykin et al., 2004; Chaw et al., 2000, 2004; Zhong et al., 2010). The reason for this lies in the nature of currently available substitution models, which assume stationarity, reversibility and homogeneity of molecular evolution. Violation of these assumptions, usual for fast-evolving sequences, leads to incorrect inference of the tree topology (e.g. Lanave et al., 1984; Lockhart et al., 1994; Foster, 2004; Jermiin et al., 2004; Delsuc et al., 2005; Lockhart and Steel, 2005).

Thus, it is not surprising that MISFITS analysis (Fig. 5-4) indicated extremely poor fit between the optimal GTR+G model and the data at the most divergent end of the sorted alignment. Compositional heterogeneity was also most pronounced at the most-variable end of the sorted alignment. Yet it represents an insufficient explanation for the poor data-model fit and topological change observed, as it extends a way past the cut-off point and the zone of relevant topological change (Fig. 5-3). The reason for the model mis-specification, indicated by correlation and MISFITS analyses lies, most probably, in an extremely elevated substitution rate in the Gnetales lineage, as compared to other lineages, observed for the most divergent partitions of the sorted alignment of the concatenated chloroplast genes (Fig 5-2). This is evident form the observation that on the PhyMLtrees built employing optimal GTR+G model the branch subtending Gnetales was much longer than branches subtending other seed plant taxa (more than 500x longer over the first 1750 bases, and between 5x-10x between 2000 and 2500 bases) at the most varied end of the OV sorted alignment (Fig. 5-5). By the last shortening step approved by the stopping criterion, the relative length of the branch subtending Gnetales as determined on the variable B-partition reduced 60 times (!) in length as compared to its initial size. The reduction in branch length was accompanied by sharp rise in bootstrap support values for the common branch uniting Gnetales with Pinaceae. As current methods of phylogeny reconstruction do not model heterotachous evolution, i.e. lineage-specific changes of the proportion of variable sequence positions in alignment (Lockhart and Steel, 2005; Lockhart et al., 2006; Gruenheit et al., 2008; Shavit et al., 2008), they cannot describe the substitution process at the most variable positions of the sorted data set, excluded before the final character-stripping step approved by the stopping criterion.

Recently, Wu et al. (2011) performed a similar study aimed at pinpointing placement of Gnetales within the gymnosperms. The study was based on partitioning of amino acid sequences encoded by the chloroplast genes onto the “low heterotachy” and “high heterotachy” categories and comparison of the trees built from each protein category. The authors arrived at the same conclusion – namely, that phylogenetic reconstruction of Gnetales in seed plant phylogeny is misled by non-time reversible properties of chloroplast sequence evolution and advocated the sister group relationship between Gnetales and Pinaceae.

7 Summary

Correct phylogenetic inference is especially hard on very shallow and deep taxonomic levels. On intraspecific/intrageneric levels, widely used genetic fingerprinting techniques (AFLP, SSR) tend to yield volatile results, partially because bands of equal size on the gel should not necessarily be orthologous, and absence of a band does not necessarily mean absence of a character, as it might be caused by experimental setup (see Kumar et al., 2009). At the same time, at this taxonomic range commonly used phylogenetic markers such as nuclear ITS regions and chloroplast spacers/genes are only of limited help due to lack of informative characters to resolve tree structure, end even medium-size data sets suffer from lack of resolution.

This work demonstrates the benefit of introducing cpDNA-based phylogenomic methodology to shallow level taxonomic research. Well-resolved tree structure allowed for the first time to reveal phylogenetic affinities of the maternal line of domesticated apple sorts. Three distinct chloroplast DNA lineages present in apple cultivars diverged before the origin of agriculture. One line, represented by *M. x domestica* cv. Granny Smith diverged from Asian apple species, and the other two, comprising the majority of apple cultivars analyzed share sister group relationships to different, unrelated accessions of *M. sylvestris* from southern and south eastern Europe.

The most straightforward explanation for this finding is that apple cultivation has started independently in various regions. The advantages of planting apples close to human dwellings were apparent enough, and people from various places started to grow apple trees taken from locations near about. Lack of the domestication bottleneck and clonal population structure in domesticated apples (Cornille et al., 2012) corroborate their polyphyletic origin. An alternative explanation would be substitution of cpDNA in *M. x domestica* by hybridisation, through pollination of *M. sylvestris* by pollen of domesticated apple sorts.

In contrast to intrageneric/intraspecific studies, on deeper taxonomic levels phylogenomic approach is widely used today. Yet its usefulness is hampered by systematic errors inherent to current methods of phylogeny reconstruction, which assume stationarity, homogeneity and reversibility of the substitution process. These non-phylogenetic signals, violating these assumptions, stem mainly from the subset of the most fast-evolving positions in the data (Rodriguez-Ezpeleta et al., 2007), and can provoke errors in reconstruction of phylogenetic trees.

This work provides evidence for further serious concerns regarding poor data-model fit at the fast-evolving, saturated positions in alignments of the chloroplast genes of spermatophytes. At the same time, it provides an analytical framework, which can help to identify and remove such positions, thereby improving the net outcome of phylogeny reconstruction. A novel character-sorting approach based on the criterion (OV), in contrast to gamma rate assignment, parsimonous tree length and compatibility of characters is more realistic proxy of the substitution rate (Goremykin, Nikiforova, Bininda-Emonds, 2010). As such it facilitates studying impact of most variable sites onto the tree topology. Accompanied by the tests estimating model-mis-specification, however, it achieves perhaps the most important goal in phylogenetic analysis, namely, provides a basis for understanding the reasons causing conflicting placements of divergent taxa. This procedure strongly expands the range of usability of chloroplast DNA onto deeper taxonomic levels, and uncovers the potential of cpDNA to serve as universal marker of plant evolution.

Practically, it provided solid evidence that the placement of *Amborella trichopoda* at the base of the angiosperm subtree as an LBA artifact. An extremely well-publicized nature of this erroneous branch, even taken alone, emphasizes importance of introducing exploratory phylogenetic analysis, which incorporates testing applicability of methodological assumptions to the data structure, to the deep-level phylogenetic studies in botany. Finding evidence of model-mis-specification affecting placement of Gnetales among non-angiosperm seed plant lineages provides yet another indication that LBA artifacts in the plant part of the Tree of Life might be abundant. Improving data-model fit has allowed to identify sister group relationship between Gnetales and Pinaceae and the composition of the basal-most angiosperm branch, comprising Hydatellaceae, Nymphaeales s.s. and *Amborella*.

The procedures worked out in the course of this study represent a framework to guide phylogenetic research in the future.

8 Zusammenfassung

Korrekte phylogenetische Rekonstruktion ist besonders auf sehr niedrigen und sehr hohen taxonomischen Ebenen schwer. Die auf intraspezifischen/intragenerischen Ebenen häufig genutzte Technik des genetischen Fingerabdrucks (AFLP, SSR) neigt dazu, zu unsicheren Ergebnissen zu führen, teils, weil Banden gleicher Größe auf dem Gel nicht notwendigerweise ortholog sind; teils, weil die Abwesenheit einer Bande nicht unbedingt auch die Abwesenheit eines Merkmals bedeutet, da erstere auch durch experimentelle Fehler (Kumar et al., 2009) verursacht werden kann. Gleichzeitig bieten phylogenetische Marker, die in diesem taxonomischen Bereich häufig verwendet werden, wie z.B. nukleäre ITS-Regionen und chloroplastidäre Spacer/Gene nur bedingt Hilfe und zwar aufgrund des Mangels an Information für die Baumstrukturauflösung; selbst mittelgroße, aus aneinandergehängten Markern bestehende Datensätze leiden oft unter unzureichender Auflösung.

Diese Arbeit demonstriert den Nutzen der Einführung der cpDNA-basierten phylogenomischen Methodologie in phylogenetische Studien auf niedrigem taxonomischen Ebenen. Eine gut aufgelöste Baumstruktur bietet zum ersten Mal die Möglichkeit, die phylogenetische Geschichte der mütterlichen Linie von domestizierten Apfelsorten aufzuklären. Drei verschiedene in Apfelsorten vorhandene cpDNA-Linien haben sich lange vor der Entstehung der Landwirtschaft getrennt. Eine Linie, vertreten in *Malus x domestica* cv. Granny Smith stammte von asiatischen Apfelsorten ab, und die anderen beiden, die im den Großteil der analysierten Apfelsorten vorhanden sind darstellen, bildeten Schwestergruppen zu Chloroplastengenomen vom *M. sylvestris* aus Süd- und Südosteuropa.

Die einfachste Erklärung für dieses Ergebnis ist, dass der Apfelanbau unabhängig voneinander in verschiedenen Regionen begonnen hat. Die von Menschen erkannten Vorteile der Anpflanzung von Äpfeln in der Nähe menschlicher Siedlungen waren vermutlich ausreichend, sie an verschiedenen Orten zu veranlassen, Apfelbäume, die aus der Nähe stammten, anzubauen. Die Abwesenheit eines Domestizierungsengpasses und die klonale Populationsstruktur bei domestizierten Äpfeln (Cornille et al., 2012) unterstützen die Annahme einer polyphyletischen Herkunft.

Im Gegensatz zu intragenetischen /intraspezifischen Studien ist der phylogenomische Ansatz auf hohen taxonomischen Ebenen heute weit verbreitet. Allerdings wird seine Nützlichkeit durch systematische Fehler in den aktuellen Methoden der Phylogenierekonstruktion begrenzt. Diese setzen uniforme Basenzusammensetzung sowie Homogenität und Reversibilität des Substitionsverlaufes voraus. Die sogenannten nicht-phylogenetischen Signale, die diese Annahmen verletzen, sind besonders in sich schnell verändernden Positionen in den Daten ausgeprägt (Rodriguez-Ezpeleta et al., 2007) und können zu Fehlern in der Phylogenierekonstruktion führen.

Diese Arbeit liefert Evidenz für weitere ernsthafte Bedenken in Bezug auf unzureichende Anpassung der Substitutionsmodelle an die sich schnell verändernden, mutationsgesättigten Positionen in den kodierenden Sequenzen aus der cpDNA der Spermatophyten. Gleichzeitig bietet sie einen analytischen Rahmen, der hilft, solche Positionen zu identifizieren und aus der weiteren Analyse auszuschließen, wodurch das Ergebnis der phylogenetischen Rekonstruktion verbessert werden kann. Ein neuartiger Ansatz zur Sortierung der Positionen im Alignment, basierend auf dem OV-Kriterium (positios-speziefische *p*-Distanz), stellt, im Gegensatz zur Bestimmung der positions-speziefischen parsimonischen Baumlänge, Gammaratenzuordnung und Kompatibilitätsmethoden eine realistischere Annäherung an die Substitutionsrate dar (Goremykin, Nikiforova, Bininda-Emonds, 2010) Als solcher erleichtert er die Untersuchung der Auswirkungen der gesättigten Positionen auf die Baumtopologie. Begleitet von Tests, die die Anpassung des Substitutionsmodells an die Daten einschätzen, ermöglicht er vielleicht, das wichtigste Ziel der phylogenetischen Analyse zu erreichen, nämlich, eine Grundlage für das Verständnis der Ursachen widersprüchlicher Baumtopologien zu schaffen. Dieses Verfahren erweitert die Grenzen der Nutzbarkeit der cpDNA-Daten für die Phylogenierekonstruktion auf hohen taxonomischen Ebenen und zeigt das Potenzial der cpDNA als unverseller Marker für Studien der Pflanzenevolution.

Praktisch wurden konkrete Hinweise dafür vorgelegt, dass die Platzierung von *Amborella trichopoda* an der Basis des Stammbaums der Angiospermen ein LBA-Artefakt ist.

Die extreme Verbreitung dieses Artefaktes in der rezenten Literatur betont schon für sich allein betrachtet die Wichtigkeit, in die Phylogeniestudien auf hohen taxonomischen Ebenen in der Botanik explorative phylogenetische Analysen einzuführen, welche das Testen der Anwendbarkeit der methodischen Annahmen für die Datensätze enthalten. Ein Nachweis dafür, dass unzureichende Modellanpassung auch die Platzierung der Gnetales innerhalb der Gymnospermenradiation beeinflusst, bietet noch einen weiteren Hinweis darauf, dass LBA-Artefakte in dem in dem heute weitgehend akzeptierten Stammbaum der Pflanzen reichlich vorhanden sein dürften.

Die Verbesserung der Modellanpassung an den Datensatz mittels Entfernen des Datenanteils, der nicht-phylogenetischen Signale beinhaltet, hat es erlaubt, die Schwestergrupppenbeziehung zwischen Gnetales und Pinaceae und die Zusammensetzung des basalsten Angiospermenastes zu identifizieren. Die Verfahren, die im Rahmen dieser Studie ausgearbeitet wurden, bieten einen methodologischen Rahmen für zukünftige phylogenetische Forschung.

9 References

Ababneh F, Jermiin LS, Ma CS, Robinson J (2006) Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics. **22**: 1225-1231

Adams K and Palmer J (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phyl Evol **29**: 380-395

Allard MW, Miyamoto MM, Honeycutt RL (1991) Test for rodent polyphyly. Nature **353**: 610-611

Amrine-Madsen H, Koepfli KP, Wayne RK, Springer MS (2003) A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol Phylogenet Evol **28**: 225-240

Ane C, Burleigh JG, McMahon MM, Sanderson MJ (2005) Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol **22**: 914-924

Applequist WL, Wallace RS (2002) Deletions in the plastid *trn*T-*trn*L intergenic spacer define clades within Cactaceae subfamily Cactoideae. Plant Syst Evol **231**: 153-162

Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X, Janke A (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. Proc Natl Acad Sci USA **99**: 8151-8156

Arthofer W, Schueler S, Steiner FM, Schlick-Steiner BC (2010) Chloroplast DNA-based studies in molecular ecology may be compromised by nuclear-encoded plastid sequence. Mol Ecol **19**: 3853-3856

Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods **6**: 22

Axelrod DI (1998) The Eocene Thunder Mountain Flora of central Idaho. University of California Publications Geological Sciences **142**: 1-61

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**: 972–977

Barkman TJ, Chenery G, McNeal JR, Lyons-Weiler J, Ellisens WJ, Moore G , Wolfe AD, dePamphilis CW (2000) Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc Natl Acad Sci USA **97**: 13166-13171

Baum D (1994) *rbc*L and seed-plant phylogeny. Trends Ecol Evol **9**: 39-41

Bausher MG, Singh ND, Lee SB, Jansen RK, Daniell H (2006) The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biol **6**: Article Number: 21 DOI: 10.1186/1471-2229-6-21

Beckert S, Muhle H, Pruchner D, Knoop V (2001) The mitochondrial *nad2* gene as a novel marker locus for phylogenetic analysis of early land plants: A comparative analysis in mosses. Mol Phyl Evol. **18**: 117-126

Bergthorsson U, Adams KL, Thomason B, Palmer JD (2003) Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature **424**: 197-201

Bergthorsson U, Richardson AO, Young GJ, Goertzen LR, Palmer JD (2004) Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc Natl Acad Sci USA **101**: 17747-17752

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol **19**: 315-322

Bininda-Emonds ORP (2007) Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evol Bioinf **3**: 59-85

Bininda-Emonds ORP, Gittleman JL, Steel MA (2002) The (Super)tree of life: Procedures, problems, and prospects. Annu Rev Ecol Syst **33**: 265-289

Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W (2003) Noncoding plastid *trn*T-*trn*F sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol **16**: 558-576

Bortiri E, Coleman-Derr D, Lazo GR, Anderson OD, Gu YG (2008) The complete chloroplast genome sequence of *Brachypodium distachyon*: sequence comparison and phylogenetic analysis of eight grass plastomes. BMC research notes. **1**: 61 DOI: 10.1186/1756-0500-1-61

Bowe LM, Coat G, dePamphilis CW (2000) Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Acad Sci **97**: 4092-4097

Braukmann TW, Kuzmina M, Stefanović S (2009) Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr Genet **55**: 323-337

Brinkmann H, Philippe H (1999) Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol **16**: 817-825

Brinkmann H, Philippe H (2008) Animal phylogeny and large-scale sequencing: progress and pitfalls. J Syst Evol **46**: 274-286

Bruno WJ, Halpern AL (1999) Topological bias and inconsistency of maximum likelihood using wrong models. Mol Biol Evol **16**: 564-566

Bryant D, Moulton V (2004) Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol **21**: 255-265

Burleigh JC, Mathews S (2004) Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am J Bot **91**: 1599-1613

Campbell CS, Baldwin BG, Donoghue MJ, Wojciechowski MF (1995) A phylogeny of the genera of Maloideae (Rosaceae): Evidence from Internal Transcribed Spacers of nuclear ribosomal DNA sequences and congruence with morphology. Amer. J. Bot **82**: 903-918

Campbell CS, Evans RC, Morgan DR, Dickinson TA, Arsenault MP (2007) Phylogeny of subtribe Pyrinae (formerly the Maloideae, Rosaceae): Limited resolution of a complex evolutionary history. Pl Syst Evol **266**: 119-145

Campbell CS, Wojciechowski MF, Baldwin BG, Alice LA, Donoghue MJ (1997) Persistent nuclear ribosomal DNA sequence polymorphism in the Amelanchier agamic complex (Rosaceae) Mol Biol Evol.**14**: 81-90

Cao Y, Adachi J, Yano T, Hasegawa M (1994) Phylogenetic place of guinea pigs: No support of the rodent-polyphyly hypothesis from Maximum-likelihood analyses of multiple protein sequences. Mol Biol Evol **11**: 593-604

Cecil E (1910) A history of gardening in England. John Murray, London

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu, SM, Chang CC et al. (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol **23**: 279-291

Chang JT (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134: 189-215

Chang S, Yang T, Du T, Huang Y, Chen J, Yan J, He J, Guan R (2011) Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in *Brassica*. BMC Genomics **12**: 497

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu, YL et al. (1993) Phylogenetics of seed plants – an analysis of nucleotide-sequences from the plastid gene *rbc*L. Ann Missouri Bot Gard **80**: 528-580

Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol **58**: 424-441

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD (2000) Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci **97**: 4086–4091

Chaw SM, Zharkikh A, Sung HM, Lau TC, Li WH (1997) Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol Biol Evol **14**: 56-68

Chen, JM, Cutler C, Jacques C, Boeuf G, Denamur E, Lecointre G, Mercier B, Cramb G, Ferec C (2001) A combined analysis of the cystic fibrosis transmembrane conductance regulator: implications for structure and disease models. Mol Biol Evol **18**: 1771-1788

Clifton SW, Minx P, Fauron CM, Gibson M, Allen JO, Sun H, Thompson M, Barbazuk WB, Kanuganti S, Tayloe C, et al. (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol **136**: 3486-3503

Coart E, Glabeke V, Loose MD, Larsen AS, Roland-Ruiz I (2006) Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). Mol Ecol **15**: 2171-2182

Collins TM, Wimberger PH, Naylor GJP (1994) Compositional bias, character-state bias and reconstruction using parsimony Syst Biol **43**: 482-496

Cornille A, Gladieux P, Smulders MJM, Roldán-Ruiz I, Laurens F, Le Cam B, Nersesyan A, Clavel, J, Olonova M, Feugey L, Gabrelyan I, Zhang X-G, Tenaillon MI, Giraud T (2012) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genet **8**: e1002703. DOI:10.1371/journal.pgen.1002703

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics **11**: 485

Crane PR (1985) Phylogenetic analysis of seed plants and the origin of angiosperms. Ann Missouri Bot Gard **72**: 716-793

Cronquist A (1981) An integrated system of classification of flowering plants. Columbia University press, New York, New York, USA

Crosby JA, Janick J, Pecknold PC, Korban SS, O'Connor PA, Ries SM, Goffreda J, Voordeckers A (1992) Breeding apples for scab resistance: 1945-1990. Fruit Var J **46**: 145-166

D'Erchia AM, Gissi C, Pesole G, Saccone C, Arnason U (1996) The guinea-pig is not a rodent. Nature **381**: 597-600

da Fonseca RR, Johnson, WE, O'Brien SJ, Ramos, MJ, Antunes A (2008) The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics **9**: 119

Dacks JB, Marinets A, Doolittle WF, Cavalier-Smith T, Logsdon JM (2002) Analyses of RNA polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. Mol Biol Evol **19**: 830-840

Darwin C (1871) On the origin of species by means of natural selection. Appleton, New York, New York, USA

Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. Genome Res 12: 1080-1090

Davis JI, Stevenson DW, Petersen G, Seberg O, Campbell LM, Freudenstein JV, Goldman DH, Hardy CR, Michelangeli FA, Simmons MP, Specht CD, Vergara-Silva F, Gandolfo M (2004) A Phylogeny of the monocots, as inferred from *rbc*L and atpA sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. Syst Bot **29**: 467-510

de Jong WW, van Dijk MAM, Poux C, Kappé G, van Rheede T, Madsen O (2003) Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. Mol Phylogenet Evol **28**: 328-340

Degnan JH, Rosenburg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol **24**: 332-340

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature reviews genetics **6**: 361-375

DeVore ML, Pigg KB (2007) A brief review of the fossil history of the family Rosaceae with a focus on the Eocene Okanogan Highlands of eastern Washington State, USA, and British Columbia, Canada. Plant Syst Evol **266**: 45-57

Dickson EE, Kresovich S, Weeden NF (1991) Isozymes in North American *Malus* (Rosaceae): hybridization and species differentiation. Syst Bot **16**: 363-375

Dombrovska O, Qiu YL (2004) Distribution of introns in the mitochondrial gene *nad1* in land plants: phylogenetic and molecular evolutionary implications Mol Phyol Evol **32**: 246-263

Doyle JA (2006) Seed ferns and the origin of angiosperms. J Torrey Bot Soc **133**: 169–209

Drabkova L, Kirschner J, Vlcek C, Paces, V (2004) *Trn*L-*trn*F intergenic spacer and trnL intron define major clades within *Luzula* and *Juncus* (Juncaceae): Importance of structural mutations. J Mol Evol **59**: 1-10

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

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2011) Geneious v5.4, Available from http://www.geneious.com/

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2010) Geneious v5.1, Available from http://www.geneious.com

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**: 214

Dutlih BE, Noort V, van der Heijden RTJM, Boekhout T, Snel B and Huynen MA (2007) Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics **23**: 815-824

Ebersberger I, Galgoczy P, Taudien S, Taenzer S, Platzer M, von Haeseler A (2007) Mapping human genetic ancestry Mol Biol Evol **24**: 2266-2276

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**: 113

Endress PK (1994) Evolutionary aspects of the floral structure in *Ceratophyllum.* Plant Syst Evol **8**: 175-183

Farias IP, Orti G, Sampaio I, Schneider H, Meyer A (2001) The cytochrome b gene as a phylogenetic marker: the limits of resolution for analyzing relationships among cichlid fishes. J Mol Evol **53**: 89-103

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Nat Acad Sci USA **106**: 5737-5742

Felsenstein J (1978) Cases in which parsimony or compatibility methods can be positively misleading. Syst Zool **27**: 401-410

Felsenstein J (1978) Cases in which parsimony or compatibility methods can be positively misleading. Syst Zool **27**: 401-410

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**: 368-376

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

Felsenstein J (2004) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Felsenstein J (2004) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Feng T-T, Zhou Z-Q, Tang J-M, Cheng M-H, Zhou S-L (2007) ITS sequence variation supports the hybrid origin of *Malus toringoides* Hughes. Can J Bot **85**: 659-666

Finet C, Timme RE, Delwiche CF, Marleta F (2010) Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Cur Biol **20**: 2217-2222

Forte AV, Ignatov AN, Ponamorenko VV, Dorokhov DB, Savelyev NI (2002) Phylogeny of the *Malus* (apple tree) species inferred from the morphological traits and molecular DNA analysis. Russ J Genet **38**: 1150-1160

Foster PG (2004) Modeling compositional heterogeneity. Syst Biol **53**: 485-495

Friis EM, Doyle JA, Endress PK, Leng Q (2003) *Archaefructus* – Angiosperm precursor or specialized early angiosperm? Trends in Plant Sciences **8**: 369-373

Friis EM, Pedersen KR, Crane PR (2001) Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous Nature **410**: 357-360

Gadagkar SR, Kumar S (2005) Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol **22**: 2139-2141

Galtier N, Gouy M, Gauthier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci **12**: 543-548

Gatesy J and Springer MS (2004) A critique of matrix representation with parsimony supertree. In phylogenetic Supertrees: Combining information to reveal the Tree of Life. Editing by Bininda-Emonds ORP. Dordrecht, The Netherlands: Kluwer Academic. 369-388 pp

Gaucher EA, Miyamoto MM (2005) A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol Phylogenet Evol **37**: 928-931

Gaut BS, Lewis PO (1995) Success of maximum-likelihood phylogeny inference in the 4-taxon case. Mol Biol Evol **12**: 152-162

Gibson AV, Gowri-Shankar P, Higgs G, Rattray MA (2005) Comprehensive analysis of mammalian mitochondrial genome base composition and improved methods. Mol Biol Evol **22**: 251-264

Giege P and Brennicke A (1999) RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci USA **96**: 15324-15329

Gillham NW, Boynton JE, Harris EH (1991) Transmission of plastid genes. Cell Cult Somatic Cell Genet Plants **7A**: 55-92

Goremykin V, Holland B, Hirsch-Ernst K, Hellwig F (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol **22**: 1813-1822

Goremykin V, Bobrova V, Pahnke J, Troitsky A, Antonov A, Martin W (1996) Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbc*L data do not support Gnetalean affinities of angiosperms. Mol Biol Evol **13**: 383-396

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003b) The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis* - structural and phylogenetic analyses. Plant Syst Evol **242**: 119-135

Goremykin VV, HansmannS, Martin WF (1997) Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. Pl Sys Evol **206**: 337-351

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: Whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol **21**: 1445-1454

Goremykin VV, Lockhart PJ, Viola R, Velasco R (2012) The mitochondrial genome of *Malus domestica* and the import-driven hypothesis of mitochondrial genome expansion in seed plants. Plant J **71**: 615-626

Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, Delange P, Martin W, Woetzel S, Atherton RA, Mcleanachan T, Lockhart JP (2013) The evolutionary root of flowering plants. Syst Biol **62**: 50-61

Goremykin VV, Nikiforova SV, Bininda-Emonds ORP (2010) Automated removal of noisy data on phylogenetic analyses. J Mol Evol **71**: 319-331

Goremykin VV, Hellwig F, Viola R (2009b) Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J Mol Evol 68(3): 197-204. DOI: 10.1007/s00239-009-9206-9

Goremykin VV, Hellwig, FH (2006) A new test of phylogenetic model fitness addresses the issue of the basal angiosperm phylogeny. GENE **381**: 81-91

Goremykin VV, Salamini F, Velasco R, Viola R (2009a) Mitochondrial DNA of *Vitis vinifera* and the Issue of Rampant Horizontal Gene Transfer. Mol Biol Evol **26**: 99-110

GoremykinVV, Hirsch-Ernst KI, Wo S, Hellwig FH (2003a) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal Angiosperm. Mol Biol Evol **20**:1499-1505

Graham SW and Iles WJD (2009) Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. Am J Bot **96**: 216-227

Graham SW, Olmstead RG (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. American J Bot **87**: 1712-1730

Graham SW, Olmstead RG, Barrett SCH (2002) Rooting phylogenetic trees with distant outgroups: A case study from the commelinoid monocots. Mol Biol Evol **19**: 1769-1781

Graur D, Hide WA, Zharkikh A, Li W-H (1992) The biochemical phylogeny of guinea-pigs and gundis, and the paraphyly of the order Rodentia. Comp Biochem Phys **101**: 495-498

Graur D, Hide WA, Li WH (1991) Is the guinea-pig a rodent? Nature **351**: 649-652

Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol **47**: 9-17

Gribaldo S, Philippe H (2002) Ancient phylogenetic relationships. Theor Popul Biol **61**: 391-408

Gruenheit N, Lockhart PJ, Steel M, Martin W (2008) Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol **25**: 1512-1520

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59**: 307-321

Halanych KM (1998) Lagomorphs misplaced by more characters and fewer taxa. Syst Biol **47**: 138-146

Hamann U (1976) Hydatellaceae – a new family of Monocotyledoneae. New Zeal J Bot. **14**: 193-6

Hamby RK and Zimmer EA (1988) Ribosomal RNA sequences for inferring phylogeny wirthin the grass family (Poaceae). Plant Syst Evol **160**: 29-37

Handa H (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. Nucleic Acids Res **31**: 5907-5916

Hanelt B, VanSchyndel D, Adema CM, Lewis LA, Loker ES (1996) The phylogenetic position of *Rhopalura ophiocomae* (Orthonectida) based on 18S ribosomal DNA sequence analysis. Mol Biol Evol **13**: 1187-1191

Hansmann S, Martin WT (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol **50**: 1655-1663

Harris SA, Robinson JP, Juniper BE (2002) Genetic clues to the origin of the apple Trends in genetics **18**: 426-430

Harrison N and Harrison RJ (2011) On the evolutionary history of domesticated apple. Nature Genetics **43**: 1043-1044

Hasegawa M, Cao Y, Adachi J, Yano T (1992) Rodent polyphyly? Nature **355**: 595-595

Hasegawa M, Kishino H, Yano T (1985) Dating of human-ape splitting by a molecular clock of mitochondrial DNA J Mol Evol **22**: 160-174

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool **38**: 297-309

Hillis DM (1996) Inferring complex phylogenies. Nature **383**: 130-131

Hillis DM (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol **47**: 3-8

Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Rohwer JG, Campbell CS, Chatrou LW (2003) Angiosperm phylogeny based on *mat*K sequence information. Amer J Bot **90**: 1758-1776

Hirt RP, Logsdon JM Jr, Healy B, Dorey MW, Doolittle WF, Embley TM (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci USA **96**: 580-585

Ho JWK, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, Jermiin LS (2006) SeqVis: Visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics **22**: 2162-2163

Ho SYW, Jermiin LS (2004) Tracing the decay of the historical signal in biological sequence data. Syst Biol **53**: 623-637

Hokanson SC, Lamboy WF, Szewc-McFadden AK, McFerson JR (2001) Microsatellite (SSR) variation in a collection of *Malus* (apple) species and hybrids. Euphytica **118**: 281-294

Hoot SB, Magallon S, Crane PR (1999) Phylogeny of basal eudicots based on three molecular data sets: *atp*B, *rbc*L, and 18S nuclear ribosomal DNA sequences. Ann Mo Bot Gard **86**: 1-32

Hu Y, Zhang Q, Rao G, Sodmergen (2008) Occurrence of plastids in the sperm cells of Caprifoliaceae: biparental plastid inheritance in angiosperms is unilaterally derived from maternal inheritance. Plant Cell Physiol **49**: 958-968

Huang JL, Sun GL, Zhang DM (2010) Molecular evolution and phylogeny of the angiosperm *ycf*2 gene. J Syst Evol **48**: 240-248

Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation systematic biology. Syst Biol **44**: 17-48

Huelsenbeck, JP, Lander KM (2003) Frequent inconsistency of parsimony under a simple model of cladogenesis. Syst Biol **52**: 641-648

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol **23**: 254-267

Iketani H (1998) Classification of fruit trees - What is the problem? What is important? J Japan Soc Hort Sci **67**: 1193-1196

Jabbari K, Rayko E, Bernardi G (2003) The major shifts of human duplicated genes. Gene **317**: 203-208

Janick J, Cummings JN, Brown SK, Hemmat M (1996) Apples. In: Janick J. and Moore JN, eds. Fruit Breed, Volume I: Tree and Tropical Fruits. John Wiley & Sons, Inc

Janke A, Xu X, Arnason U (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. Proc Natl Acad Sci USA **94**: 1276-1281

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Claude W, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA **104**: 19369-19374

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genomesequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids BMC Evol Biol **6**: Article Number: 32 DOI: 10.1186/1471-2148-6-32

Jayaswal V, Jermiin LS, Poladian L, Robinson J (2011a) Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Sist Biol **60**: 74-86

Jayaswal V, Ababneh F, Jermiin LS, Robinson J (2011b) Reducing model complexity of the general Markov model of evolution. Mol Biol Evol **28**: 3045-3059

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? TRENDS in Genetics **22**: 225-231

Jenkins C, Fuerst JA (2001) Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. J Mol Evol **52**: 405-418

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol **53**: 638-643

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali S, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473**: 97-100

Johnson KP (2001) Taxon sampling and the phylogenetic position of Passeriformes: evidence from 916 avian cytochrome b sequences. Syst Biol **50**: 128-136

Jukes TH and Cantor CR (1969) Evolution of Protein Molecules. New York: Academic Press. pp. 21-132

Juniper BE (2007) The mysterious origin of the sweet apple - On its way to a grocery counter near you, this delicious fruit traversed continents and mastered coevolution. American Scientist **95**: 44-51

Juniper BE, Mabberley DJ (2006) The story of the apple. Imber Press Inc. 240 p

Källersjö M, Albert VA, Farris JS (1999) Homoplasy increases phylogenetic structure. Cladistic **15**: 91-93

Kim KJ and Lee HL (2004) Complete chloroplast genome sequence from korea ginseng (Panax schinseng Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res **11**: 247–261

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 (1981) Estimation of evolutionary distances between homologous nucleotide-sequences. Proc Nat Acad Sci USA **78**: 454-458

Kjer KM, Honeycutt RL (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evol Biol **7**: 8

Kolaczkowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous Nature **431**: 980-984

Kostka M, Uzlikova M, Cepicka I, Flegr J (2008) SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of *Blastocystis.* BMC Bioinf **9**: 34

Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K (2003) The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. Nucleic Acid research. **31**: 716-721

Kumar P, Gupta VK, Misra AK, Modi DR, Pandey BK (2009) Potential of molecular markers in plant biotechnology. Plant Omics **2**: 141-162

Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA (2005) Measuring genome conservation across taxa: divided strains and united kingdoms. Nucleic Acids Research **33**: 616-621

Kupczok A, Schmidt HA, Haeseler A (2010) Accurancy of phylogeny reconstruction methods combinino overlapping gene data sets. Algorithms Mol Biol **5**: Article 37. DOI: 10.1186/1748-7188-5-37

Kuroiwa T (1991) The replication, differentiation and inheritance of plastids with emphasis on the concept of organelle nuclei. International review of cytology – A survey of cell biology **128:** 1-62

Lake JA and Rivera MC (2004) Deriving the Genomic Tree of Life in the Presence of Horizontal Gene Transfer: Conditioned Reconstruction. Mol Biol Evol **21**: 681–690. 2004

Lamboy WF, Yu J, Forsline PL, Weeden NF (1996) Partitioning of allozyme diversity in wild populations of *Malus sieversii* L and implications for germplasm collection. J Am Soc Hortic Sci **121**: 982-987

Lanave C, Preparata G, Saccone C, Serio G (1984) J Mol Evol **20**: 86-93

Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol **7**: Supplement: 1 Article Number: S4. DOI: 10.1186/1471-2148-7-S1-S4

Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics **25**: 2286-2288

Lartillot N, Philippe H (2004) A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**: 1095-1109

Le Quesne WJ (1969) A method of selection of characters innumerical taxonomy. Syst Zool **18**: 201-205

Lecointre G, Philippe HL, Le Guyader H (1993) Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol **2**: 205-224

Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006) The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. BMC Genomics **7**: Article Number: 61 DOI: 10.1186/1471-2164-7-61

Leebens-Mack J, Raubeson, LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphilis CW (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone Mol Biol Evol **22**: 1948-1963

Lemieux C, Otis C, Turmel M (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature **403**: 649-652

Leopold EB and Clay-Poole S (2001) Florissant leaf and pollen floras of Colorado compared: Climatic considerations. In Fossil Flora and Stratigraphy of the Florissant Formation Colorado. (Evanoff E, Gregory-Wodzicki KM and Johnson KR (eds.)). pp. 17-69. Denver Museum of Nature & Science Series **4, 1**: 17-69

Les DH (1988) The origin and affinities of the Ceratophyllaceae. Taxon **37**: 326-345

Les DH, Garvin DK, Wimpee CF (1991) Molecular evolutionary history of ancient aquatic angiosperms. Proc Natl Acad Sci USA **88**: 10119-10123

Li Q-Y, Guo W, Liao W-B, Macklin JA, Li J-H (2012) Generic limits of Pyrinae: Insights from nuclear ribosomal DNA sequences. Botanical Studies **53**: 151-164

Li W-H, Hide WA, Zharkikh A, Ma D-P, Graur D (1992) The molecular taxonomy and evolution of the Guinea Pig. J Hered **83**: 174-181

Lin Y, Waddell P, Penny D (2002a) Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires. Gene **294**: 119-129

Lin YH, McLenachan PA, Gore AR, Phillips MJ, Ota R, Hendy MD, Penny D (2002b) Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol **19**: 2060-2070

Ling G, Shiliang Z, Zuoshuang Z, Xiang S, Ying C, Donglin Z, Huairui S (2009) Relationships of species, hybrid species and cultivars in genus *Malus* revealed by AFLP markers. Scientia Silvae Sinicae **45**: 33-40

Lo EYY, Donoghue MJ (2012) Expanded phylogenetic and dating analyses of the apples and their relatives (Pyreae, Rosaceae). Mol Phylogen Evol **63**: 230-243

Lo EYY, Stefanović S, Christensen KI, Dickinson TA (2009) Evidence for genetic association between East Asian and Western North American *Crataegus* L. (Rosaceae) and rapid divergence of the Eastern North American lineages based on multiple DNA sequences. Mol Phylogenet Evol **51**: 157-168

Lo EYY, Stefanović S, Dickinson TA (2007) Molecular reappraisal of relationships between *Crataegus* and *Mespilus* (Rosaceae, Pyreae) - two genera or one? Syst Bot **32**: 596-616

Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T (2006) Heterotachy and tree building: A case study with plastids and eubacteria. Mol Biol Evol **23**: 40-45

Lockhart P, Steel M (2005) A tale of two processes. Syst Biol **54**: 948-951

Lockhart P, Steel M, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution Mol Biol Evol **11**: 605-612

Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA **93**: 1930-1934

Lockhart PJ, Penny D (2005) The place of *Amborella* within the radiation of angiosperms. Trends Plant Sci. **10**: 201-202

Löhne C, Borsch T (2005) Molecular evolution and phylogenetic utility of the *pet*D group II intron: A case study in basal angiosperms. Mol Biol Evol. **22**: 317-332

Lopez P, Forterre P, Philippe H (1999) The root of the tree of life in the light of the covarion model. J Mol Evol **49**: 496-508

Lopez P, Philippe H (2001) Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. CR Acad Sci Paris, Sciences de la vie / Life Sciences **324**: 201-208

Luby JJ (2003) Taxonomic classification and brief history. Editors: Ferree DC, Warrington IJ. Apples: botany, production and uses. p: 1-14

Luckett WP, Hartenberger J-L (1993) Monophyly or polyphyly of the order Rodentia: possible conflict between morphological and molecular interpretations. J Mamm Evol **1**: 127-147

Lyons-Weiler J, Hoelzer GA (1997) Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. Mol Phyl Evol **8**: 375-384

Ma D-P, Zharkikh A, Graur D, VandeBerg JL, Li WH (1993) Structure and evolution of opposum, guinea pig, and porcupine cytochrome b genes. J Mol Evol **36**: 327-334

Maddison WP (1997) Gene trees in species trees. Syst Biol **46**: 523-536

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature **409**: 610-614

Magallon S, Sanderson MJ (2002) Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. Am J Bot **89**: 1991-2006

Manchester SR (1994) Fruits and seeds of the middle Eocene nut neds flora, Clarno formation, north central Oregon. Palaeontographica Americana **58**: 1-205

Manhart JR (1994) Phylogenetic analysis of green plant *rbc*L sequences. Mol Phylogenet Evol **3**: 114-27

Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV, Skyabin KG (2008) Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: Structural organization and phylogenetic relationships to other angiosperms. J Mol Evol **66**: 555-564

Martin PG, Dowd JM (1991) A comparison of 18S ribosomal-RNA and rubisco large subunit sequences for studying angiosperm phylogeny. J Mol Evol **33**: 274-28

Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V (2005) Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Pl Sci **10**: 203-209

MartinW, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature **393**: 162-165

Mathews S, Donoghue MJ (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science **286**: 947-950

Matsen FA, Steel MA (2007) Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst Biol **56**: 767-775

Mes THM, Thart H (1994) Sedum-surculosum and S-Jaccardianum (Crassulaceae) share a unique 70-bp deletion in the chloroplast DNA-*trn*L (UAA)-*trn*F (GAA) intergenic spacer. Plant Syst Evol **193**: 213-221

Mes THM, Kuperus P, Kirschner J, Stepanek J, Oosterveld P, Storchova H, den Nijs JCM (2000) Hairpins involving both inverted and direct repeats are associated with homoplasious indels in non-coding chloroplast DNA of Taraxacum (Lactuceae : Asteraceae). Genome **43**: 634-641

Micheletti D, Troggio M, Salamini F, Viola R, Velasco R, Salvi S (2011) On the evolutionary history of domesticated apple. Nature Genetics **43**: 1044-1045

Misof B, Rickert AM, Buckley TR, Fleck G, Sauer KP (2001) Phylogenetic signal and its decay in mitochondrial SSU and LSU rRNA gene fragments of *Anisoptera*. Mol Evol Biol **18**: 27-37

Moncalvo JM, Vilgalys R, Redhead SA, Johnson JE, James TY, Catherine, Aime M, Hofstetter V, Verduin SJ, Larsson et al. (2002) One hundred and seventeen clades of euagarics. Mol Phylogenet Evol **23**:357–400

Moore MJ, Bell CD, Soltis PS and Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA **104**: 19363-19368

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM and Soltis DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol **6**: 17

Moreira D, Le Guyader H, Philippe H (1999) Unusually high evolutionary rate of the elongation factor lar genes from the ciliophora and its impact on the phylogeny of eukaryotes. Mol Biol Evol **16**: 234-245

Moreira D, Le Guyader H, Philippe H (2000) The origin of red algae and the evolution of chloroplasts. Nature **405**: 69-72

Mouchaty SK, Catzeflis F, Janke A, Arnason U (2001) Molecular evidence of an african phiomorpha-south america caviomorpha clade and support for hystricognathi based on the complete mitochondrial genome of the cane rat (*Thryonomys swinderianus*). Mol Phylogenet Evol **18**:127-135

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ (2001a) Molecular phylogenetics and the origins of placental mammals Nature **409**: 614-618

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, et al. (2001b) Resolution of the early placental mammal radiation using Bayesian inference. Science **294**: 2348-2351

Nguyen MAT, Klaere S, von Haeseler A (2011) MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. Mol Biol Evol **28**: 143-152

Nickrent DL, Parkinson CL, Palmer JD, DuV RJ (2000) Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol Biol Evol **17**: 1885-1895

Ogihara Y, Yamazaki Y, Murai K, Kanno A, Terachi T, Shiina T, Miyashita N, Nasuda S, Nakamura C, Mori N et al. (2005) Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res **33**: 6235-6250

Oh SH, Potter D (2003) Phylogenetic utility of the second intron of LEAFY in *Neillia* and *Stephanandra* (Rosaceae) and implications for the origin of *Stephanandra*. Mol Phylogenet Evol **29**: 203-215

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York. 160 pp

Olsen G (1987) Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol **52**:825-37

Omland KE, Lanyon SM, Fritz SJ (1999) A molecular phylogeny of the New World orioles (*Icterus*): the importance of dense taxon sampling. Mol Phylogenet Evol **12**: 224–239

Oraguzie NC, Gardiner SE, Basset HCM, Stefanati M, Ball RD, Bus VGM, White AG (2001) Genetic diversity and relationships in *Malus* sp. ermplasm Collections as determined by randomly amplified polymorphic DNA. J Amer Soc Hort Sci **126**: 318-328

Page RDM, Holmes EC (1998) Molecular evolution: A phylogenetic approach. Blackwell Science, Oxford

Pagel M and Meade A (2005) Mixture models in phylogenetic inference. In O. Gascuel (ed) Mathematics of Evolution and Phylogeny. Oxford: Oxford University Press 121-142 pp

Palmer AR (1965) Biomere — a new kind of biostratigraphic unit. J Paleontol **39**: 149-153

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology. **7**: 84 DOI:10.1186/1741-7007-7-84

Pesole G, Gissi C, de Chirico A, Saccone C (1999) Nucleotide Substitution Rate of Mammalian Mitochondrial Genomes. J Mol Evol **48**: 427-434

Peterson, KJ and Eernisse DJ (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol Dev **3**:170–205

Philippe H (1997) Rodent monophyly: pitfalls of molecular phylogenies J Mol Evol **45**: 712-715

Philippe H, Brinkmann H, Lavrov DV, Timothy D, Littlewood J, Manue M, Wörheide G, Baurain D (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol **9**: e1000602. DOI:10.1371/journal.pbio.1000602

Philippe H, Brinkmann H, Martinez P, Riutort M, Baguña J (2007) Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS ONE **2**: 717

Philippe H, Zhou1 Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol **5**:50 DOI:10.1186/1471-2148-5-50

Phillips MJ, Lin Y-H, Harrison GL, Penny D (2001) Complete mitochondrial sequences for two marsupials, a bandicoot and a brushtail possum. Proc R Soc Lond Ser B **268**: 533-1538

Phillips MJ, Penny D (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol **28**: 171-185

Pickett KM, Tolman GL, Wheeler WC, Wenzel JW (2005) Parsimony overcomes statistical inconsistency with the addition of more data from the same gene. Cladistics **21**: 438-445

Pisani D (2004) Identifying and removing fast evolving sites using compatibility analysis: an example from the arthropoda. Syst Biol **53**: 978-989

Pisani D, Mohun MM, Harris S, McIterney JO, Wilkinson M (2006) Molecular evidence for dim-light vision in the last common ancestor of the vertebrates. Curr Biol **16**: 318-319

Pol D and Siddall ME (2001) Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics **17**: 266-281

Ponomarenko VV (1983) History of the origin and evolution of the apple *Malus*. Trudy po prikladnoi botanike, genetike i selektsii **76**: 10-18 (in Russian, English abstract)

Posada D, Crandall KA (1998) ModelTest: testing the model of DNA substitution. Bioinformatics **14**: 817-818

Potts SM, Han YP, Khan MA, Kushad M, Rayburn AL, Korban SS (2012) Genetic diversity and characterization of a core collection of *Malus* Germplasm using simple sequence repeats (SSRs). Plant Mol Biol Rep **30**: 827-837

Prasad AB, Allard MW, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol **25**: 1795-1808

Qiu YL, Chase MW, Les DH and Parks CR (1993) Molecular phylogenetics of the Magnoliidae : cladistic analyses of nucleotide sequences of the plastid gene *rbc*L. Ann Mo Bot Gard **80**: 587-606

Qiu YL, Lee J, Whitlock BA, Bernasconi-Quadroni F, Dombrovska O (2001) Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? Mol Biol Evol **18**: 1745-1753

Qiu YL, Dombrovska O, Lee J, Li LB, Whitlock BA, Bernasconi-Quadroni F, Rest JS, Davis CC, Borsch T, Hilu KW, Renner SS, Soltis DE, Soltis PS, Zanis MJ, Cannone JJ, Gutell RR, Powell M, Savolainen V, Chatrou LW Chase, MW (2005) Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes Inter. J Plant Sci **166**: 815-842

Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW (2000) Phylogeny of basal angiosperms: Analyses of five genes from three genomes. Inter J Plant Sci **161**: Issue: 6 Supplement: S Pages: S3-S27

Qiu YL, Lee JH, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen ZD, Savolainen V, Chase MW (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature **402**: 404-407

Qiu YL, Li LB, Hendry TA, Li RQ, Taylor DW, Issa MJ, Ronen AJ , Vekaria ML, White AM (2006) Reconstructing the basal angiosperm phylogeny: evaluating information content of mitochondrial genes. Taxon **55**: 837-856

Qiu YL, Li LB, Wang B, Xue JY, Hendry TA, Li RQ, Brown JW, Liu Y, Hudson GT, Chen ZD (2010) Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J Syst Evol. **48**: 391-425

Quian GZ, Liu LF, Tang GG (2006) A new section in *Malus* (Rosaceae) from China. Ann Bot Fennici **43**: 68-73

Rambaut A (2002) Se-Al. Sequence Alignment Editor v2.0a11. http://evolve.zoo.ox.ac.uk

Raubeson LA and Jansen RK (2005) Chloroplast genomes of plants. Pages 45-68. In RJ Henry, editor. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. CABI Publishing, Cambridge, MA

Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. Pl Sys Evol 271: 101-122

Ravi V, Jitendra P. Khurana, Akhilesh K, Khurana T, Khurana P (2006) The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis. Tree genetics and genomes **3**: 49-59

Renner SS, Foreman DB, Murray D (2000) Timing transantarctic disjunctions in the Atherospermataceae (Laurales): Evidence from coding and noncoding chloroplast sequences Syst Biol **49**: 579-591

Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, Saccone C (2004) Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. Mol Biol Evol **21**: 397-403

Reyes A, Gissi C, Pesole G, Catzeflis FM, Saccone C (2000b) Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. Mol Biol Evol **17**: 979-83

Reyes A, Pesole G, Saccone C (1998) Complete mitochondrial DNA sequence of the fat dormouse, Glis glis: further evidence of rodent paraphyly. Mol Biol Evol **15**: 499-505

Reyes A, Pesole G, Saccone C (2000a) Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. Gene **259**: 177-187

Rijk P, Van de Peer Y, Van den Broeck I, Wachter R (1995) Evolution according to large ribosomal subunit RNA. J Mol Evol **41**: 366-375

Robinson JP, Juniper BE, Harris SA (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivatedapple, *Malus domestica* Borkh. Plant Syst Evol **226**: 35-58

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

Ruhlman T, Lee SB, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006) Complete plastid genome sequence of *Daucus carota*: Implications for biotechnology and phylogeny of angiosperms. BMC Genomics **7**: Article Number: 222 DOI: 10.1186/1471-2164-7-222

Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J (1999) Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science **283**: 1919-1923

Saarela JM, Rai HS, Doyle JA, Endress PK, Mathews S, Marchant AD, Briggs BG, Graham SW (2007) Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature **446**: 312-315

Sanderson MJ, Driskell AC (2003) The challenge of constructing large phylogenetic trees. Trends Pl Sci **8**: 374-379

Sandlers R (2010) The apple book. Frances Lincoln Limited, London

Saunders MA, Edwards SV (2000) Dynamics and phylogenetic implications of MtDNA control region sequences in New World Jays (Aves: Corvidae). J Mol Evol **51**: 97-109

Savolainen V, Corbaz R, Moncousin C, Spichiger R, Manen JF (1995) Chloroplast DNA variation and parentage analysis in 55 apples. Theor Appl Genet **90**: 1138-1141

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, De Bruijn AY, Sullivan S, Qiu YL (2000) Phylogenetics of flowering plants based on combined analysis of plastid *atp*B and *rbc*L gene sequences. Syst Biol **49**: 306-362

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**: 502-504

Shavit L, Penny D, Hendy MD, Holland BR (2007) The Problem of Rooting Rapid Radiations Mol Biol Evol **24**: 2400-2411

Shavit GL, Penny D, Hendy MD, Holland BR (2008) LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol **8**: 317

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature **402**: 402-404

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanović S, Rice DW, Palmer JD, Soltis PS (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Pl Sci **9**: 477-483

Soltis DE, Moore MJ, Burleigh G, Soltis PS (2009) Molecular markers and concepts of plant evolutionary relationships: progress, promise, and future prospects. Critical reviews in Plant Sciences **28**: 1-15

Soltis DE, Soltis PS (2004) *Amborella* not a 'basal angiosperm'? Not so fast. Am J Bot **91**: 1199-1199

Soltis DE, Soltis PS, Zanis MJ (2002) Phylogeny of seed plants based on evidence from eight genes. Am J Bot **89**: 1670-1681

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS (2011) Angiosperm phylogeny: 17 genes, 640 taxa. American J Bot. **98**: 704-730

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris JS (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbc*L, and *atp*B sequences. Bot J Linnean Soc **133**: 381-461

Soltis DE, Soltis PS, Nickrent DL, Johnson LA, Hahn WJ, Hoot SB, Sweere JA, Kuzoff RK, Kron KA, Chase MW et al. (1997) Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann Missouri Bot Gar **84**: 1-49

Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of eumetazoa. Mol Biol Evol **26**: 2261-2274

Springer MS, Debry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ (2001) Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny phylogeny reconstruction. Mol Biol Evol **18**: 132-143

Springer MS, Stanhope MJ, Madsen O, de Jong WW (2004) Molecules consolidate the placental mammal tree. Trends Ecol Evol. **19**: 430-438

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21**: 456-463

Stefanović S, Rice DW, Palmer JD 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol Biol **360**: 1975-1980

Strimmer SK (1997) Maximum likelihood methods in molecular phylogenetics. Thesis. pp. 56. Munchen

Sullivan J and Swofford DL (1997) Are Guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J Mamm Evol **4**: 77-86

Sullivan J, Holsinger KE, Simon C (1995) Among-site rate variation and hylogenetic analysis of 12S rRNA data in sigmodontine rodents. Mol Biol Evol **12**: 988-1001

Sun G, Dilcher DL, Wang H, Chen Z (2011) A eudicot from the Early Cretaceous of China. Nature **471**: 625-628

Sun G, Ji Q, Dilcher DL, Zheng S, Nixon KC, Wang X (2002) Archaefructaceae, a New Basal Angiosperm Family. Science **296**: 899-904

Swenson MS, Barbancon F, Warnow T, Linder CR. (2010) A simulation study comparing supertree and combined analysis methods using SMIDGen. Algorithms Mol Biol **5** Article Number: 8. DOI: 10.1186/1748-7188-5-8

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (Eds.), Phylogenetic Inference. Sinauer Associates, Sunderland, MA, USA 407-514

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol **50**: 525-539

Takhtajan A (1967) Система и филогения цветкорых растений (Systema et Phylogenia Magnoliophytorum). Nauka. Moscow

Talavera G, Vila R (2011) What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. BMC Evol biol **11**: 315

Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Mol Biol Evol **9**: 678-687

Tamura K, 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, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28**: 2731-2739

Tavaré S (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences (American Mathematical Society) **17**: 57-86

Taylor DW, Hickey LJ (1992) Phylogenetic evidence for the herbaceous origin of angiosperms. Plant Syst Evol **180**: 137-156

Tian X, Zheng J, Hu S, Yu J (2006) The rice mitochondrial genomes and their variations Plant Physiol **140**: 401-410

Tuffley C, Steel MA (1998) Modelling the covarion hypothesis of nucleotide substitution. Math BioSci **147**: 63-91

Tuinen M, Sibley CG, Hedges SB (2000) The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. Mol Biol Evol **17**: 451-457

Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci USA **96**: 10248-10253

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**: 1596-1604

Vavilov NI (1930) Wild progenitors of the fruit trees of Turkestan and the Caucasus and the problem of the origin of fruit trees. In Proceedings of the 9th International Horticultural Congress. pp. 271-286. London

Velasco R, **Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D et al.** (2010) The genome of the domesticated apple (*Malus domestica* Borkh.). Nature Genetics **42**: 833-839

Wakasugi T, Tsudzukit J, Itot S, Nakashimat K, Tsudzuki T (1994) Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci USA **91**: 9794-9798

Way RD, Aldwinckle HS, Lamb RC, Rejman A, Sansavini S, Shen T, Watkins R, Westwood MN, Yoshida Y (1990) Apples (*Malus*). In Genetic Resources of Temperate Fruit and Nuts (Moore JN and Ballington R eds), pp. 1-62

Wehr WC, Hopkins, DO (1994) The Eocene orchards and gardens of Republic, Washington. Washington Geology **22**: 27-35

Wheeler EA, Manchester SR (2002) Woods of the Eocene Nut Beds Flora, Clarno Formation, Oregon, USA. In IAWA Journal Supplement 3. International Association of Wood Anatomists, National Herbarium Nederland, The Netherlands

Whitfield JB, Lockhart PJ (2007) Deciphering ancient rapid radiations. Trends Ecol Evol **22**: 258-265

Wissemann V, Ritz CM (2005) The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and *atp*B-*rbc*L intergenic spacer (IGS) versus conventional taxonomy. Bot J Linn Soc **147**: 275-290

Wolf PG, Karol KG, Mandoli DF, Kuehl J, Arumuganathan K, Mishler BD, Kelch DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene **350**: 117-128

Wolfe JA, Wehr W (1988) Rosaceous *Chamaebatiaria*-like foliage from the paleogene of western North America. Aliso **12**: 177-200

Wu CS, Wang YN, Liu SM, Chaw SM (2007) Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol **24**: 1366-1379

Wu CS, Wang YN, Hsu CY, Lin CP, Chaw SM (2011) Loss of Different Inverted Repeat Copies from the Chloroplast Genomes of Pinaceae and Cupressophytes and Influence of Heterotachy on the Evaluation of Gymnosperm Phylogeny. Genome Biol Evol **3**: 1284-1295

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20**: 3252-3255

Yamane K, Kawahara T (2005) Intra- and interspecific phylogenetic relationships among diploid *Triticum-Aegilops* species (Poaceae) based on base-pair substitutions, indels, and microsatellites in chloroplast noncoding sequences Am J Bot **92**: 1887-1898

Yang X, Tuskan GA, Tschaplinski TJ, Cheng Z-M (2007) Third-codon transversion rate-based *Nymphaea* basal angiosperm phylogeny – concordance with developmental evidence. Nature precedings DOI:10.1038/npre.2007.320.1

Yang Z (2006) Computational Molecular Evolution. Oxford University Press

Yang ZH (1997) How often do wrong models produce better phylogenies? Mol Biol Evol **24**: 35-37

Zanis MJ, Soltis DE, Soltis PS, Mathews S, Donoghue MJ (2002) The root of the angiosperms revisited. Proc Natl Acad Sci USA **99**: 6848-6853

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**: 821-829

Zhang N, Shen H-X, Gao X-H, Yao Y-C, Wang Y, Feng Y-Q (2007) Phylogenetic relationships between ornamental and wild species of *Malus* in China. Acta Horticulturae Sinica. **34**: 1227-1234

Zhang Q, Li J, Zhao Y, Korban SS, Han Y (2012) Evaluation of genetic diversity in chinese wild apple species along with apple cultivars using SSR markers. Plant Mol Biol Rep **30**: 539-546

Zhang Q, Sodmergen (2010) Why does biparental plastid inheritance revive in angiosperms? J Plant Res **123**: 201-206

Zhong B, Deusch O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ (2011) Systematic error in seed plant phylogenomics. Genome Biol Evol **3**: 1340-1348

Zhong B, Deusch O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ (2011) Systematic error in seed plant phylogenomics. Genome Biol Evol **3**: 1340-1348

Zhong BJ, Yonezawa T, Zhong Y, Hasegawa M (2010) The position of Gnetales among seed plants: Overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol **27**: 2855-2863

Zhou ZQ, Li YN (2000) The RAPD evidence for the phylogenetic relationship of the closely related species of cultivated apple. Genet Resour Crop Ev **47**: 353-357

Zohary D, Hopf M (1994) Domestication of plants in the Old World. Clarendon Press, Oxford

Zuckerkandl E and Pauling L (1965) Evolutionary divergence and convergence in proteins. In Evolving genes and proteins. Ed. Bryson V, Vogel HJ. New York: Acad. Press.

10 Eigenständigkeitserklärung

Ich erkläre, entsprechend § 5 Abs. 3 der Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena,

- Dass mir die geltende Promotionsordnung der Fakultät bekannt ist;
- Dass ich die Dissertazion selbst angefertigt habe, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeit ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe;
- Dass ich Personen, die mich bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben, angegeben habe und das Dritte weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
- Dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;
- Dass ich nicht die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Svetlana Nikiforova

Svetlana Nikiforova

Jena, 10. März 2013

11 List of publications and presentations

Publications:

1. Goremykin VV, Nikiforova SV, Bininda-Emonds ORP (**2010**) Automated removal of noisy data on phylogenetic analyses. Journal of Molecular Evolution **71**: 319-331

2. Zhong B, Deusch O, Goremykin VV, Panny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ (**2011**) Systematic error in seed plant phylogenomics. Genome Biology and Evolution **3**: 1340-1348

3. Caputi L, Malnoy M, Goremykin V, Nikiforova SV, Martens S (**2012**) A genome-wide phylogenetic reconstruction of family during the adaptation of plants to life on land. Plant Journal **69**: 1030-1042

4. Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, Delange P, Martin W, Woetzel S, Atherton RA, McLenachan T, Lockhart PJ (**2013**) The evolutionary root of flowering plants. **62**: 50-61

5. Nikiforova SV, Cavalieri D, Velasco R, Goremykin VV (**2013**) Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. Molecular biology and evolution (under revision).

Presentations:

Svetlana Nikiforova. Automated Removal of Noisy Data in Phylogenomic Analyses // Abst. Book of the Otto Warburg International Summer School and Research Symposium 2011 on Evolutionary Genomics. – Berlin, Germany, 14-22 September, **2011**, p 66.

12 Curriculum Vitae

Personal details

Name: Svetlana Nikiforova
Date of Birth: 26 of March 1983
Nationality: Russian
Address: Evolutionary Biology research group, Computational Biology, Research and Innovation Center – Foundation Edmund Mach (Edmund Mach 1, 38010 - S. Michele all'Adige (TN), Italy)
e-mail: svetlana.nikiforova@fmach.it
nikiforovasvetlana83@gmail.com

Education and work:

2001 – 2005	M.Sc. in Biology at Saratov State University, Saratov, Russia Diploma with honors
2005 – 2009	junior scientific co-worker at the Institute of Biochemistry and Physiology of Plants and Microorganisms, Russian Academy of Sciences, Saratov, Russia
2009 – present	PhD student at Foundation Edmund Mach (San Michele All'Adige, Italy) and Friedrich Schiller University (Jena, Germany)

13 Acknowledgments

First of all, I would like to thank my tutor and doctoral adviser Vadim Goremykin. He taught and supported me during all these four years. I could always bother him with my questions and call for his help in lab and bioinformatics works. He was always a great company to talk about work and life over a beer.

Together with Vadim, Professor Frank Hellwig made me a scientist. He guided me in my research and introduced me in scientific community in Jena University. He was always strict and fair with me and sometimes very patient.

I would like to thank to all my colleagues in San Michele and especially my boss and head of department of Computational biology, Duccio Cavallieri, for help and support, too.

Last, but not least, I would like to hug my parents and sister, who always had confidence in me and supported me in many respects. Without them, my dreams could never come true. Thank you very much!