Molecular Phylogenetics and Evolution 64 (2012) 145–155

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Sorting through the chaff, nDNA gene trees for phylogenetic inference and hybrid identification of annual sunflowers (*Helianthus* sect. *Helianthus*)

Michael L. Moody [a,*], Loren H. Rieseberg [b,c]

[a] School of Plant Biology, University of Western Australia, 35 Sterling Highway, Crawley, WA 6009, Australia
[b] Botany Department, University of British Columbia, 3529-6270 University Blvd., Vancouver, BC, Canada V6T 1Z4
[c] Department of Biology, Indiana University, 1001 E. Third St., Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 22 May 2011
Revised 1 March 2012
Accepted 19 March 2012
Available online 5 April 2012

Keywords:
nDNA
Helianthus sect. *Helianthus*
Hybridization
Gene trees
Incongruence
Phylogeny

ABSTRACT

The annual sunflowers (*Helianthus* sect. *Helianthus*) present a formidable challenge for phylogenetic inference because of ancient hybrid speciation, recent introgression, and suspected issues with deep coalescence. Here we analyze sequence data from 11 nuclear DNA (nDNA) genes for multiple genotypes of species within the section to (1) reconstruct the phylogeny of this group, (2) explore the utility of nDNA gene trees for detecting hybrid speciation and introgression; and (3) test an empirical method of hybrid identification based on the phylogenetic congruence of nDNA gene trees from tightly linked genes. We uncovered considerable topological heterogeneity among gene trees with or without three previously identified hybrid species included in the analyses, as well as a general lack of reciprocal monophyly of species. Nonetheless, partitioned Bayesian analyses provided strong support for the reciprocal monophyly of all species except *H. annuus* (0.89 PP), the most widespread and abundant annual sunflower. Previous hypotheses of relationships among taxa were generally strongly supported (1.0 PP), except among taxa typically associated with *H. annuus*, apparently due to the paraphyly of the latter in all gene trees. While the individual nDNA gene trees provided a useful means for detecting recent hybridization, identification of ancient hybridization was problematic for all ancient hybrid species, even when linkage was considered. We discuss biological factors that affect the efficacy of phylogenetic methods for hybrid identification.

1. Introduction

The increasing availability of EST libraries and the development of next generation sequencing technology has provided a growing pool of candidate low copy nuclear DNA (nDNA) sequences for phylogenetic analyses of non-model organisms (Sang, 2002; Hughes et al., 2006). While the use of multiple nDNA genes to elucidate species level phylogeny is rapidly growing, the application of this approach to land plants has been fairly limited to date (but see, Zhu and Ge, 2005; Levin et al., 2009).

The use of nDNA genes for phylogeny does not come without potential challenges for phylogenetic inference. Recent studies have identified high levels of topological heterogeneity among nDNA gene trees in species level phylogenetic studies (Gatesy and Baker, 2005; Carstens and Knowles, 2007; Hulsey et al., 2011). This heterogeneity has been attributed to the three well characterized processes, horizontal gene transfer (including hybridization), gene duplication and deep coalescence (Maddison, 1997; Edwards, 2009). Carstens and Knowles (2007) present data from multiple nuclear loci from recently evolved species demonstrating the lack of reciprocal monophyly at any locus as well as topological incongruence likely associated with the effects of lineage sorting. They suggest incorporating explicit models of coalescence into phylogenetic analyses in multi-gene phylogeny. Deep coalescence is a particular problem for recently evolved species with a low rate of genetic drift, as found among taxa with high effective population size. While concatenation of multiple nuclear DNA regions has been general practice as a method to overcome deep coalescence effects, recent simulations have shown that concatenation alone in some circumstances will lead to well-supported, but incorrect results (Kubatko and Degnan, 2007). However, to some extent partitioned Bayesian methods incorporating specific models to incorporate differential rates of divergence among genes can correct for these effects (Edwards, 2009). Coalescent methods have also been developed to estimate species trees from gene trees and incorporate the effects of divergence time and branch lengths recovered from individual gene trees (Liu and Pearl, 2007; Edwards, 2009). Unfortunately, these approaches cannot simultaneously account for heterogeneity caused

* Corresponding author. Present address: Department of Biological Sciences, University of Texas at El Paso, 500 W. University Ave., El Paso, TX 79968, USA. Fax: +1 915 747 5808.
E-mail address: mlmoody@utep.edu (M.L. Moody).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.03.012

by hybridization and paralogy, which may also be present for any taxonomic group.

Hybridization is a significant issue for plant phylogenetics, as many plant groups have long been considered to be prone to reticulate evolution through polyploid and homoploid hybridization (Stebbins, 1950; Grant, 1981; Arnold, 1997; O'Brien et al., 2000). Estimates vary considerably, but 5–25% of plant species appear to hybridize in nature (Ellstrand et al., 1996). It is crucial to differentiate the heterogeneity among gene trees due to hybridization from that caused by deep coalescence to recover a species tree. Methods for depicting reticulate evolution in phylogenetic trees utilizing a network approach and existing data are becoming more advanced (Huson, 1998; Bryant and Moulton, 2002; Holland et al., 2008). However, reticulate evolution can become intractable with high levels of incongruence among many gene trees (McBreen and Lockhart, 2006) and combined with lineage sorting (Nakhleh, 2010; Yu et al., 2011). Strictly empirical methods still need to be explored.

Taxon-specific molecular markers and linkage disequilibrium (i.e., associations among taxon-specific markers; Szymura and Barton, 1991; Bridle and Butlin, 2002) or recovery of reciprocal alleles (Sang and Zhang, 1999; Moody and Les, 2002) have been commonly employed as an empirical approach to identify contemporaneous hybridization events, but the footprints of hybridization rapidly diminish over time due to mutation and recombination. To identify more ancient hybridization events researchers have relied on heterogeneity among gene trees. This approach is not entirely satisfactory, because other evolutionary processes (e.g., deep coalescence) known to cause heterogeneity among gene trees cannot be dismissed. Also, reticulate evolution will often be missed in phylogenetic studies using few genes because the genes analyzed may be from the same side of the hybridization event, one of the parental taxa may now be extinct, or there may be insufficient resolution among hybrid and parental taxa.

Rieseberg et al. (1996, 2003) showed that the nuclear genomes of three hybrid sunflower species originating >100 thousand years ago contain segments of tightly linked markers deriving from alternate parental species and concluded that hybrid speciation is the most plausible explanation for this result. Linkage mapping of experimental hybrid crosses and subsequent modeling have supported this, showing that a proportion of tightly linked blocks of genes (<0.5 cM) will be retained from each side of a hybridization event as seen in *Helianthus* (Buerkle et al., 2000; Buerkle and Rieseberg, 2008). Therefore, gene trees from tightly linked block of genes should reflect a phylogenetic history with only one of the proposed hybrid parents whereas other blocks of genes might reflect the other side of the cross. Incongruence due to lineage sorting or unequal rates of evolution should not predict such a relationship between topological similarity and linkage (Linder and Rieseberg, 2004). With the development of an EST library for *Helianthus annuus* a number of nDNA markers have become available along with well-resolved linkage maps based on combined RFLP, RAPD, AFLP, and SNP data (Lai et al., 2005; Buerkle and Rieseberg, 2008; Heesacker et al., 2009), providing empirical comparative data to test this approach for hybrid identification.

The annual sunflowers (*Helianthus* sect. *Helianthus*; 12 spp.) have become a model group for the study of homoploid hybridization, patterns of genetic diversity and selection in plants (Rieseberg et al., 2003; Burke et al., 2004; Buerkle and Rieseberg, 2008; Kane et al., 2009). This group includes three well-corroborated homoploid hybrid species (*H. anomalus*, *H. deserticola* and *H. paradoxus*) with relatively small effective population size and strong ecological specificity (Schwarzbach and Rieseberg, 2002; Welch and Rieseberg, 2002; Gross et al., 2003; Strasburg et al., 2011). The putative parental species, *H. annuus* and *H. petiolaris* are the most widespread and abundant species, representing the deepest divergence in the annual sunflower clade (Schilling et al., 1998; Timme et al., 2007) and also have the largest effective population sizes (Strasburg and Rieseberg, 2008). Other species in the group are regionally isolated with comparatively small effective population size (Strasburg et al., 2011). However, most have populations that are regionally sympatric with *H. annuus* and/or *H. petiolaris* and some level of interspecific gene flow has been identified or is suspected in regions of sympatry (Strasburg et al., 2009, 2011; Scascitelli et al., 2010). Also, extensive interspecific gene flow has been demonstrated between *H. petiolaris* and *H. annuus*, which have widely overlapping geographic distributions, even though strong reproductive barriers are present (Yatabe et al., 2007). This may best be exemplified by Kane et al. (2009) who showed higher genetic similarity between *H. annuus* and *H. petiolaris* at a subset of sampled genes than found between *H. argophyllus* and *H. annuus*, which diverged approximately 700,000 years after the *H. annuus*/*H. petiolaris* split (Strasburg and Rieseberg, 2008).

The most comprehensive phylogenetic studies of the annual sunflowers have utilized nrDNA (Rieseberg, 1991; Schilling et al., 1998; Timme et al., 2007). The most recent of these (Timme et al., 2007), based on sequence from the external transcribed spacer (ETS) nrDNA, was the first to show high resolution with strong nodal support for relationships among species of the annual sunflower clade, when the hybrid taxa were removed (but included only a single accession of *H. annuus*). These results were consistent with earlier results from ITS (Schilling et al., 1998) or restriction site (Rieseberg, 1991) data, although both had lower resolution and comparatively weak nodal support. Similar to the earlier studies, Timme et al. (2007) report confounding results when the homoploid hybrid species were included in the phylogenetic analyses, as well as conflicting data due to presumed paralogs and recombinants. The addition of multiple nDNA sequence single copy markers can help elucidate and/or provide support for our current understanding of phylogenetic relationships among the annual sunflowers and may be a particularly important asset in phylogenetic interpretation when combined with our understanding of population level dynamics and hybrid history.

Here we utilize low copy nDNA sequence data from tightly linked blocks of genes derived from the *Helianthus annuus* expressed sequence tags (EST) library for the annual sunflower clade (*Helianthus* sect. *Helianthus*) to recover individual gene trees to assess the potential effects of interspecific gene flow and deep coalescence. We also analyze concatenated data using a partitioned Bayesian methodology to recover phylogenetic hypotheses of "species trees" and interpret these results, given our improved understanding of species level dynamics. We also use the results from individual gene trees to test the hypothesis that blocks of closely linked genes will provide congruent phylogenetic results for alternate sides of the hybridization events thought to have given rise to *H. anomalus*, *H. deserticola* and *H. paradoxus*.

2. Materials and methods

2.1. Sampling strategy

Our sampling strategy was guided by two objectives: (1) reconstruction of a species level phylogeny of the annual sunflowers and (2) determination of the efficacy of using linked genes and congruent versus incongruent gene tree topology for hybrid recognition. Thus we sampled three or four accessions of each of the putative hybrid taxa and parental species, as well as one to three accessions of each the seven remaining species in the clade (29 accessions total; Table 1). DNAs were extracted from fresh field collected leaf material or grown from field collected USDA seed stock in the Indiana University greenhouse. The monophyly of the annual sunflowers has been well-supported by all previous phylogenetic analyses and based on these, as well as PCR amplification success, the