Molecular Ecology Resources (2009) **9** (Suppl. 1), 130–139 doi: 10.1111/j.1755-0998.2009.02652.x

BARCODING PLANTS

Are plant species inherently harder to discriminate than animal species using DNA barcoding markers?

ARON J. FAZEKAS,* PRASAD R. KESANAKURTI,* KEVIN S. BURGESS,†¶ DIANA M. PERCY,‡ SEAN W. GRAHAM,‡ SPENCER C. H. BARRETT,† STEVEN G. NEWMASTER,* MEHRDAD HAJIBABAEI§ and BRIAN C. HUSBAND*

**Department of Integrative Biology, University of Guelph, Guelph, ON, Canada N1G 2W1, †Department of Ecology and Evolutionary Biology, 25 Willcocks St., University of Toronto, Toronto, ON, Canada M5S 3B2, ‡UBC Botanical Garden and Centre for Plant Research, Faculty of Land and Food Systems, 2357 Main Mall, and Department of Botany, 6270 University Boulevard, University of British Columbia, Vancouver, BC, Canada V6T 1Z4, §Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, ON, Canada N1G 2W1*

Abstract

The ability to discriminate between species using barcoding loci has proved more difficult in plants than animals, raising the possibility that plant species boundaries are less well defined. Here, we review a selection of published barcoding data sets to compare species discrimination in plants vs. animals. Although the use of different genetic markers, analytical methods and depths of taxon sampling may complicate comparisons, our results using common metrics demonstrate that the number of species supported as monophyletic using barcoding markers is higher in animals (> 90%) than plants (~70%), even after controlling for the amount of parsimony-informative information per species. This suggests that more than a simple lack of variability limits species discrimination in plants. Both animal and plant species pairs have variable size gaps between intra- and interspecific genetic distances, but animal species tend to have larger gaps than plants, even in relatively densely sampled genera. An analysis of 12 plant genera suggests that hybridization contributes significantly to variation in genetic discontinuity in plants. Barcoding success may be improved in some plant groups by careful choice of markers and appropriate sampling; however, overall fine-scale species discrimination in plants relative to animals may be inherently more difficult because of greater levels of gene-tree paraphyly.

Keywords: barcode, genetic distance, hybridization, incomplete lineage sorting, monophyly, paraphyly

Received 17 November 2008; revision received 16 January 2009; accepted 30 January 2009

Introduction

Efforts to identify a DNA barcode for discriminating among recognized species have been more successful in animals than plants. Since the initial proposal for a standardized barcoding region (Hebert *et al*. 2003), researchers have reported variable but relatively high rates of species discrimination (> 95%) using a portion of the mitochondrial gene cytochrome *c* oxidase subunit 1 (*cox1/CO1*) for animal groups such as birds (Kerr *et al*. 2007), fishes (Ward *et al*. 2005), amphibians (Smith *et al*. 2008) and lepidopterans (Hajibabaei *et al*. 2006a).

In contrast, plant studies report a more modest ability to discriminate among closely related species. Kress & Erickson (2007) found that nine plastid DNA and nuclear ribosomal intergenic DNA regions were able to discriminate species pairs in 40.6% to 82.6% of all genera examined, with seven loci exhibiting under 70% resolution. Similarly, Fazekas *et al*. (2008) evaluated the utility of seven plastid DNA regions for their ability to discriminate 92 species in 32 genera of land plants. Differences in amplification success

Correspondence: Aron J. Fazekas, Fax: 519 767-1656; E-mail: afazekas@uoguelph.ca

¶Present address: Department of Biology, 163A LeNoir Hall, Columbus State University, Columbus, GA 31907-5645, USA.

© 2009 Blackwell Publishing Ltd

notwithstanding, individual plastid DNA regions resolved between 29% and 59% of species. Combining the more variable plastid markers provided clear benefits for species discrimination, although with diminishing returns. All combinations that were assessed using four to seven regions had only marginally different success rates (69–71%), despite an increasing amount of variation (parsimony informative characters) (Fazekas *et al*. 2008).

Certainly, neither animals nor plants are homogeneous with respect to species resolution. However, the lower rate of overall resolution observed in Fazekas *et al*. (2008) may indicate a general limit to the precision of plant species discrimination (relative to animal species) using markers from a single genetic linkage group. It also raises the question of how discrete plant species are according to plastid markers, and whether the potential for resolving species boundaries using DNA barcoding is fundamentally different than in animals.

There has been considerable debate regarding the discreteness of plant species relative to animals. Botanists have questioned whether plant species are natural, evolutionarily independent entities and whether characteristics such as polyploidy, hybridization and apomixis preclude the application of a single species concept (Stebbins 1950; Levin 1979). In contrast, studies based on evidence from floras and monographs have concluded that plant species can usually be readily separated with minimal ambiguity (Mayr 1992; McDade 1995). More recently, Rieseberg *et al*. (2006) tested for phenotypic and reproductive discreteness of a large sample of taxonomically recognized plant species. They estimated that < 60% were phenotypically discrete and 70% corresponded to reproductively isolated groups. Significantly, plant species were no less likely to exhibit phenetic clusters than animals, and were more likely to exhibit reproductive discontinuity. A similar comparison on the nature of species boundaries in plant vs. animals using DNA sequences has not been conducted. Data from recent barcoding research offer an opportunity to explore this problem further.

The goals of this paper are to synthesize and compare results on species discrimination from recent animal and plant barcoding studies and discuss the potential causes for their differential success. As published results are based on different criteria for measuring species resolution, we first re-evaluate their success in discriminating described animal and plant species using two common criteria: (i) support for species monophyly (e.g. Hajibabaei *et al*. 2006b; Fazekas *et al*. 2008; Lahaye *et al*. 2008); (ii) differences between intra- and interspecific genetic distances (the genetic distance gap; see Hebert *et al*. 2004; Barrett & Hebert 2005). We then consider potential causes of the differences in genetic discreteness of plant and animal species that we observe. Finally, we discuss the future development of plant DNA barcodes and potential strategies that could lead to further improvement of plant species identification.

Genetic divergence among species in plants vs. animals

To investigate whether plant species are less genetically differentiated than animal species based on DNA barcodes, we downloaded several large barcoding projects (Table 1) from the Barcode of Life Data Systems (BOLD) (www.boldsystems.org/views/login.php) for comparison with our recently published barcoding data derived from major land-plant clades, as sampled in a regional flora (Fazekas *et al*. 2008). The plant plastid and the animal mitochondrion have very different levels of nucleotide variability (the animal mitochondrion exhibits ~10–30 times more nucleotide substitution; Wolfe *et al*. 1987). However, differences in nucleotide substitution rates only affect the ease of recovering sufficient differences to discriminate species. To ensure the number of parsimony-informative characters (PIC) compared are similar between genomes, we compare the resolution obtained from ~600 bp of mitochondrial *cox1* (the animal barcode) to that from ~4000 bp of plant plastid data.

To compare species resolution among projects, we re-analysed the animal data from BOLD using the same criterion as used for the plant data, that being membership in a monophyletic group, well supported by the gene tree (bootstrap value of at least 70%) (see Fazekas *et al*. 2008 for justification and limitations of this approach). We averaged bootstrap support across all species-level monophyletic groups as an overall measure of the ability to resolve species (Table 1).

The re-analysis of animal data sets indicates species resolution between 90–98% based on well-supported species monophyly (Table 1). These values agree well with estimates using other methods in the original publications; small differences may be a consequence of our exclusion of species that were represented by only a single sample, in addition to the different measures of resolution used (e.g. clustering methods, reciprocal monophyly, genetic distance). A plateau in resolution of ~70% was achieved from the plant data set using four to seven plastid regions in combination (Table 1, Fig. 1a). Individually, none of the plastid regions achieved resolution comparable to that obtained with the mitochondrial *cox1* locus in animals, despite having similar levels of parsimony informative characters per species (highest value = 46% in plants, lowest value = 90% in animals; Table 1, Fig. 1b). This difference may be in part an artifact of how densely closely related species were sampled. However, many of the plant genera were sampled relatively sparsely, suggesting that, if anything, the values we obtained for plants represent upper limits.

We further explored genetic divergence among plant and animal species by examining evidence for a 'gap' between intra- and interspecific genetic distances (Kimura 2-parameter estimate) within individual genera. A gap between the

© 2009 Blackwell Publishing Ltd

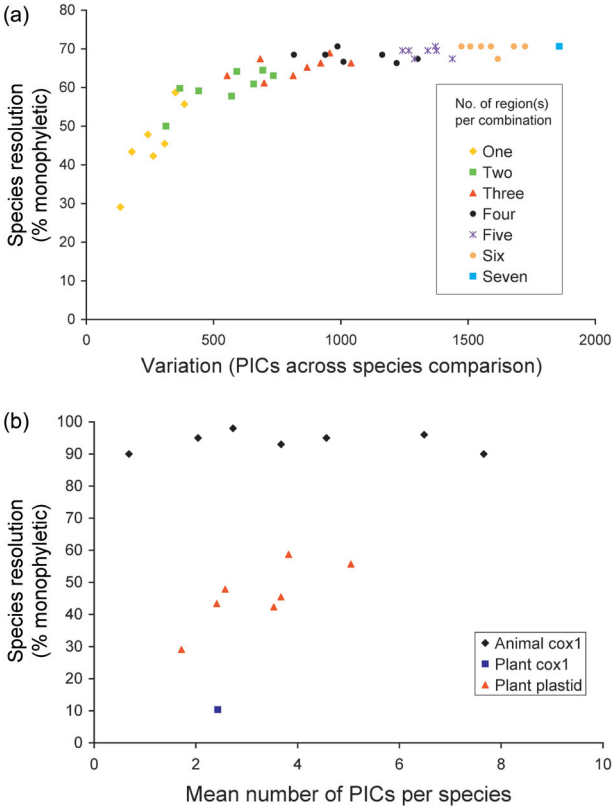

Fig. 1 Variation in barcode species resolution (percentage) as a function of sequence variation (PIC, parsimony-informative characters) for: (a) a selection of single and multilocus combinations of plant plastid regions (the same species across all loci) (from Fazekas *et al*. 2008), and (b) seven individual plant plastid regions compared to mitochondrial *cox1* for seven animal groups (see Table 1). In (a), PICs represent the sum across all genus-level comparisons. To account for differences in species number among projects in (b), variation was expressed as the mean number of PICs per species. For animal projects, the total number of PICs was divided by the number of species in the data set. For the plant data set, this was determined as the average of the number of PICs per species (calculated within genera due to alignment difficulties with non-coding regions). The mean number of PICs per species for plant mitochondrial DNA is greatly inflated by the inclusion of three species of *Plantago*, which has elevated rates of mitochondrial nucleotide substitution.

largest intraspecific distance and the smallest interspecific distance is an ideal situation for unambiguous species assignment in the taxonomic group of interest (see also Meier *et al*. 2008). We restricted the distance measure to species within genera as it is more appropriate to calculate distances between individuals at this level than at higher taxonomic ranks. Fazekas *et al*. (2008) found that the individual genera they examined corresponded to well-supported clades at current levels of taxonomic sampling. For plants, we used our own published data set (Fazekas *et al*. 2008), supplemented with data from the plant barcoding literature, specifically studies that included multiple

Table 1 Species resolution as determined by well-supported (≥ 70% bootstrap support) monophyly for seven animal DNA barcoding projects on BOLD (Birds of North America – phase II, Fishes of Australia Part I, Mosquitoes of North America, CO1 Barcoding Amphibians, Barcoding the Aphididae, Bats of Guyana, Saturniidae of the ACG 1) and one plant barcoding project (from Fazekas *et al*. 2008). We estimated plant monophyly within genera due to alignment problems in non-coding regions. We report the results for plants as the mean resolution per single locus (average of seven plastid DNA regions) or resolution based on all seven plastid DNA regions combined

Project	Reference	No. of species	No. of species with multiple accessions	No. of genera represented	Published resolution	Resolution based on monophyly (percentage)*	Average bootstrap value (percentage)**	SD of bootstrap value**
Birds of North America	Kerr *et al*. (2007)	641	553	131	94	90	97	7.63
Australian fish	Ward *et al*. (2005)	189	157	101	100	94	97	10.37
Mosquitoes	Cywinska *et al*. (2006)	52	39	4	100	98	97	6.37
Amphibians of Canada	Smith *et al*. (2008)	39	38	3	94	95	94	11.36
Aphids	Foottit *et al*. (2008)	334	109	49	96	90	96	10.81
Bats of Guyana	Clare *et al*. (2006)	87	77	22	93	96	98	6.9
Tropical silk moths	Hajibabaei *et al*. (2006a)	65	62	16	100	98	100	1.22
Plants (single genes)	Fazekas *et al*. (2008)	92	92	32	46	46	89	14.59
Plants (multiple genes combined)	Fazekas *et al*. (2008)	92	92	32	71	71	99	4.72

*Percentage of species with at least 70% support. Includes only species with multiple accessions; **based on all species resolved as monophyletic on the shortest trees.

© 2009 Blackwell Publishing Ltd

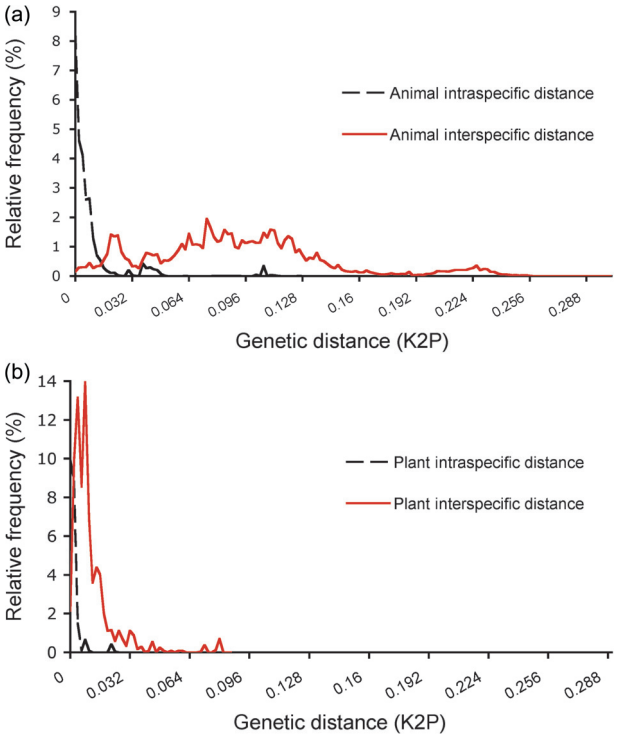

Fig. 2 Distribution of intraspecific (black broken line) and interspecific (red solid line) pairwise Kimura 2-parameter genetic distances for (a) pooled data from 326 animal genera across seven projects in BOLD (see Table 1) and (b) pooled data from 49 plant genera derived from three plant barcoding publications (Fazekas *et al*. 2008; Lahaye *et al*. 2008; Newmaster *et al*. 2008) (see Appendix S1).

samples per species and at least two species per genus (Lahaye *et al*. 2008; Newmaster *et al*. 2008, see Appendix S1, Supporting information). Collectively, these data sets include floristic sampling in temperate (Fazekas *et al*. 2008) and tropical (Lahaye *et al*. 2008) locations, as well as taxonomically focused sampling (Lahaye *et al*. 2008; Newmaster *et al*. 2008) (N = 49 genera). For comparison, we pooled the intra- and interspecific (only within-genus) distances generated from all animal data sets represented in Table 1 (N = 326 genera). The resulting histograms illustrate a continuum of genetic distances, with some degree of overlap between intra- and interspecific distances in both animals (Fig. 2a) and plants (Fig. 2b). However, two important differences are evident between the plant and animal data sets. The values of interspecific distance are generally much greater in animals than in plants, and the degree of overlap between intra- and interspecific distance is far less. Species pairs that exhibit unusually high values of intraspecific distance may reflect undetected cryptic species, which could artificially reduce the discontinuity.

A closer examination of genetic distance measures on a genus-by-genus basis reveals a more complex pattern. Maximum and minimum levels of intra- and interspecific distance vary among genera in both plants and animals, as does the magnitude of any discontinuities in genetic distance (Fig. 3). For 12 plant genera (with at least three species each) from Fazekas *et al*. (2008), we find variation in gap size, from complete overlap (e.g. *Solidago*, *Symphyotrichum*) to small but distinct gaps in intra- and interspecific distances (e.g. *Polygonum*, *Viburnum*) (see Fig. 3 for examples). It is important to note that these patterns are based on a relatively limited sampling of each genus. More complete taxonomic sampling may reveal more overlap in intra- and interspecific distance, further reducing the ability to discriminate species in these situations. Variation in the pattern of genetic distances is also observed in some animal groups (Fig. 3) but the differences between intra- and interspecific distances are typically much larger.

Whether species resolution is determined using support for monophyly or other approaches, species identification using DNA barcodes is expected to fail when species are paraphyletic according to gene trees, i.e. if some haplotypes of a species are more closely related to haplotypes of another species than to conspecifics. Our review of barcoding studies published to date suggests that such 'paraphyly' (broadly defined, since gene-tree paraphyly may have multiple sources, see below) may be more common among plants than animals, which is an idea with some support. For example, Lynch (1989) concluded that ~21% of animal species arise through mechanisms such as sympatric or peripheral isolation and therefore may include non-monophyletic species, at least initially (see Olmstead 1995). Using the same argument, Rieseberg & Brouillet (1994) suggested that paraphyly is likely to be very common in plants. Indeed, using recent surveys of phylogenetic data from two plant families, Crisp & Chandler (1996) reported that paraphyly ranged from 20% to 50% among species in eight genera. It has been estimated that 23% of 2319 species are paraphyletic in mtDNA phylogenies of animals (Funk & Omland 2003). Values varied among groups but were particularly low in mammals (17%) and birds (16.7%), two groups commonly represented in barcoding studies. These values are in contrast to the high species resolution reported in most animal barcoding studies, perhaps because most barcoding studies have been geographically focused and taxonomically incomplete. Although additional confirmation is required, as the methods used to estimate paraphyly are frequently coarse, the available evidence is consistent with the hypothesis that paraphyly is more widespread in plants than animals.

Sources of paraphyly and modest species discrimination in plants

Using DNA sequences as barcodes to discriminate between species (e.g. Hebert *et al*. 2003) rests in part on the assumption that species are monophyletic with respect to barcode

© 2009 Blackwell Publishing Ltd

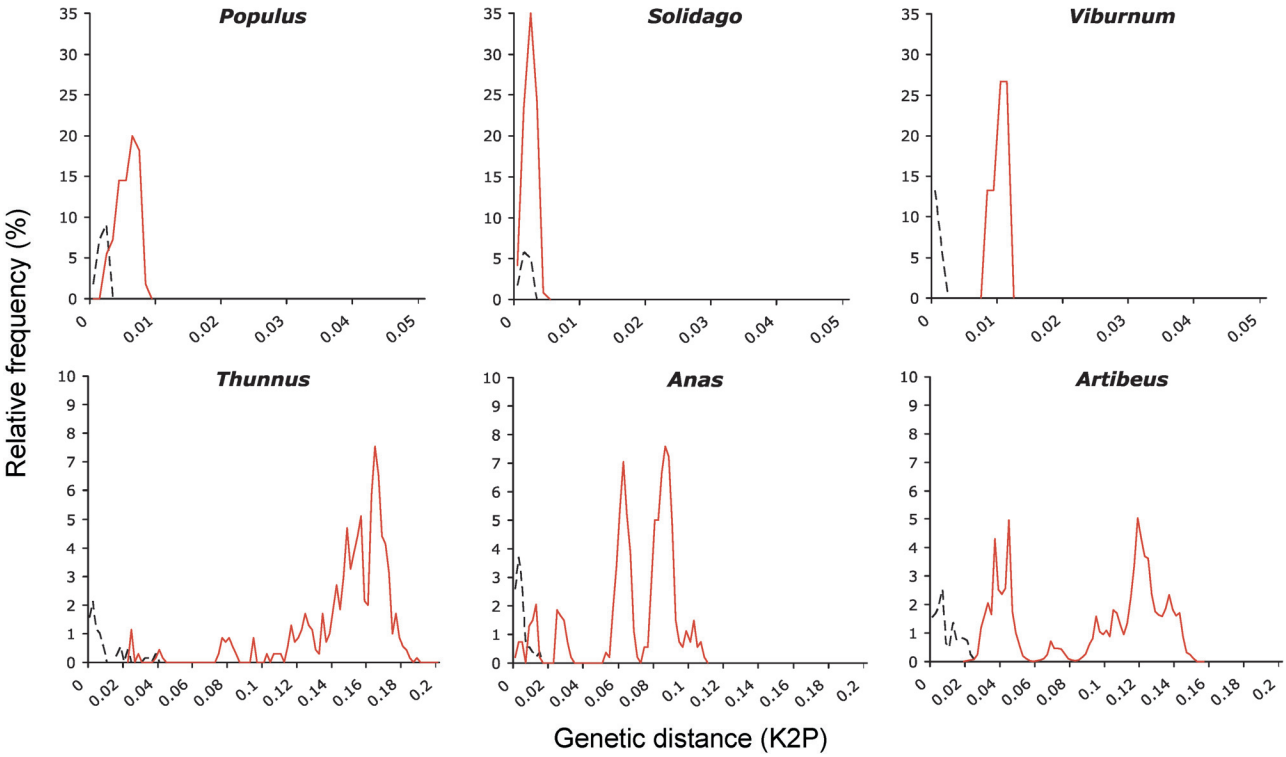

Fig. 3 Distributions of intraspecific (black broken line) and interspecific (red solid line) genetic distances (K2P) for a selection of three plant genera (from Fazekas *et al*. 2008), and three animal genera (one each from Australian fish, birds of North America, and bats of Guyana). Each genus is represented by a minimum of three species and each species by samples from multiple locations. Plant data are based on sequences from seven plastid DNA regions, whereas animal data are based on sequences from *cox1*.

haplotypes. Not all species in nature are expected to be monophyletic (e.g. Olmstead 1995), and so we may expect an upper limit in the precision of plastid-based plant DNA barcoding markers. Indeed, in plant barcoding studies that have included multiple samples per species (Fazekas *et al*. 2008; Lahaye *et al*. 2008; Newmaster *et al*. 2008) a significant proportion (up to 30%) of non-monophyletic species have been detected. A portion of these may simply reflect lack of resolution in local subsets of a gene tree; however, some of these may represent genuine gene-tree paraphyly.

If a particular gene is evolving slowly relative to the speciation rate, or if too small a fragment is sequenced, there may simply be insufficient nucleotide differences to distinguish species. This problem should be straightforward to correct by simply increasing the number of loci examined. However, based on Fazekas *et al*. (2008), it does not appear that species discrimination in plants is always limited by the amount of variability. By combining up to seven plastid DNA regions, we increased the number of phylogenetically informative characters per species to above that observed in animal *cox1* sequences, but the degree of species resolution did not increase proportionally (Fig. 1a). It is important to note that the approach to a limit in species resolution with increasing PICs (Fig. 1a) may not hold for all plant genera considered individually. In some cases (e.g. *Solidago*), the number of informative characters does not increase when multiple plastid regions are combined. Across all genera, however, our data suggests that lack of monophyly is not simply the result of insufficient variation; rather it may often reflect discrepancies between the plastid gene tree and taxonomic species boundaries (Maddison 1997). If so, this may offer important insights into the nature of plant species boundaries.

Gene-tree paraphyly may be quite common in plants, reflecting three distinct phenomena: (i) gene exchange caused by hybridization and polyploidy; (ii) incomplete sorting of ancestral polymorphisms; and (iii) imperfect species definitions and taxonomy. None of these potential sources of paraphyly are mutually exclusive, and several may contribute towards reducing the power of species discrimination in particular lineages investigated in plant DNA barcoding studies. Furthermore, most published barcoding studies should underestimate instances of paraphyly (over-estimate monophyly), since they do not exhaustively sample all species within genera, or the full geographical ranges of individual species.

Historically, many botanists have argued that plant species are not as sharply defined as animals due to the incidence of reticulate evolution, facilitated by hybridization and genome duplication (Stebbins 1950; Grant 1957). These processes can cause differentiated species to share similar

© 2009 Blackwell Publishing Ltd

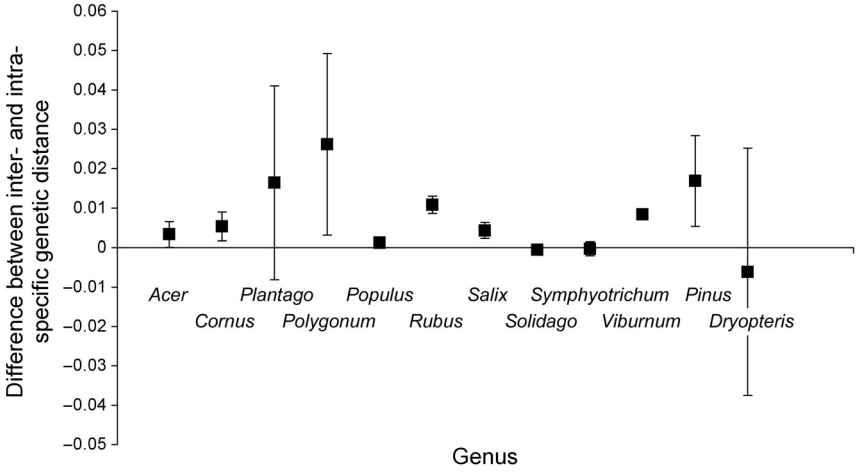

Fig. 4 Mean (+/– 95% CI) difference between interspecific and intraspecific genetic distances (K2P) for species in 12 plant genera (data from Fazekas *et al*. (2008)). Each genus was represented by a minimum of three species and each species was represented by samples from multiple locations. Distances were based on sequences from seven plastid DNA regions (see Appendix S2).

or related plastid haplotypes and result in discordance between gene and species trees (Maddison 1997; Funk & Omland 2003). To examine whether these processes can account for variation in barcode resolution, we quantified genetic discontinuity among species as the genetic distance gap (the minimum interspecific genetic distance minus maximum intraspecific genetic distance for each species, averaged across congeners) for each of 12 genera from Fazekas *et al*. (2008), and related those values to the incidence of polyploidy and hybridization (see Appendix S2, Supporting information). The 'genetic distance gap' varied widely among genera here, from –0.0063 in the fern *Dryopteris* and –0.0007 in *Solidago*, to 0.0260 in *Polygonum* (Fig. 4).

The lack of a genetic gap among congeners within *Dryopteris* and *Solidago* mirrors the low species resolution of the plastid DNA barcodes (33% and 16.7% resolution respectively) in these genera. The presence or absence of polyploid variation within a genus, determined using the inferred base number of the genus and chromosome counts for each species [using the Index to Plant Chromosome Numbers, IPCN, Missouri Botanical Gardens (mobot.mobot.org/W3T/Search/ipcn/html), supplemented by some genus-specific treatments] had no significant association with the magnitude of the genetic distance gap (ANOVA, $F_{1,10} = 0.0004$, $P > 0.90$; Fig. 5a). In contrast, the incidence of hybridization was a strong predictor of gap size. Specifically, genera with published evidence of naturally occurring hybridization (for the species in our data set), had significantly less genetic discontinuity than genera lacking hybridization (ANOVA, $F_{1,10} = 17.35$, $P = 0.0019$; Fig. 5b). This association was not confounded by differences in life history (woody vs. herbaceous), which had no association with the genetic distance gap (ANOVA, $F_{1,10} = 0.0001$, $P > 0.90$, not shown).

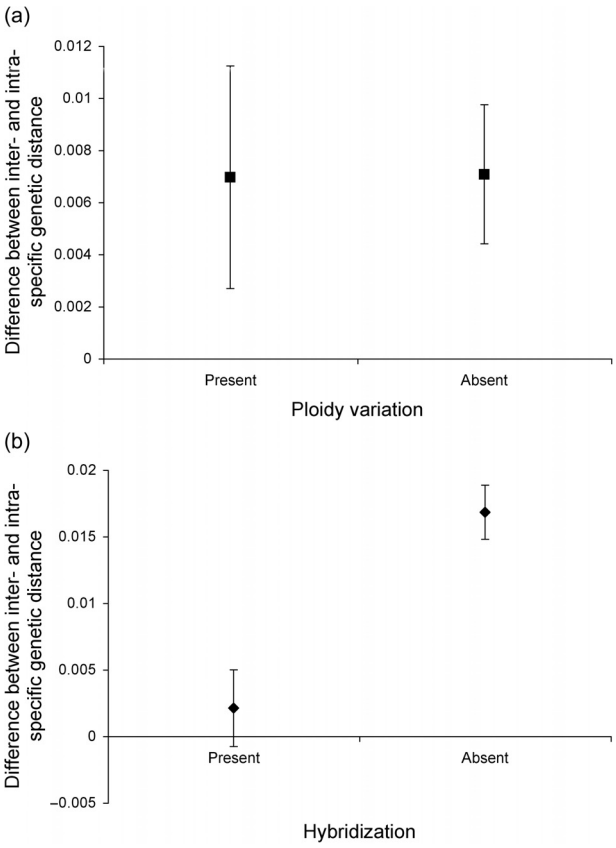

Fig. 5 Mean (± SE) difference between interspecific and intraspecific genetic distance for 12 plant genera grouped according to the incidence of: a) presence of polyploidy, and b) evidence of hybridization (barcoding data from Fazekas *et al*. (2008)). The magnitude of the genetic distance gap was significantly associated with the incidence of hybridization (ANOVA, $F_{1,10} = 17.35$, $P = 0.0019$) but not with polyploidy.

Paraphyly in plant species may also arise through incomplete sorting of ancestral polymorphisms (incomplete lineage sorting or 'deep coalescence': Maddison 1997). Within any given species, haplotypes will differ in their coalescence time (time since their common ancestral haplotype diverged). Chance sorting events during speciation mean that haplotypes in one species may be more closely related to those in sister species than to other haplotypes in

© 2009 Blackwell Publishing Ltd

their own species (i.e. the species is not monophyletic for the gene under consideration). The probability of retention of an ancestral polymorphism through a speciation event depends on the effective population size (~4*Ne*) of the parental species (in turn dependent on its demographic history and mode of inheritance), and the time in generations, *T*, between the two most recent speciation events (Pamilo & Nei 1988). Both larger N_e and smaller *T* lead to an increased probability of incomplete lineage sorting (missorting of polymorphisms). As a result, rapidly diverged species often contain paraphyletic gene trees (i.e. within-species haplotype diversity that is not consistent with species monophyly). If the time since the most recent speciation is also short (i.e. a young species) or the modern effective population size large, there is also a greater chance of confounding ancestral polymorphisms being retained to the present day. Ancestral polymorphisms may therefore have a profound influence on the ability to discriminate species in barcoding studies, at least in a subset of cases.

There are few methods for reliably distinguishing the effects of retained ancestral polymorphisms from gene flow. The isolation-with-migration model (Wakeley 1996; Nielsen & Wakeley 2001), distinguishes these processes (under restrictive assumptions) based on the variance in pairwise nucleotide differences among haplotypes (alleles) (see also Sang & Zhong 2000). All else being equal, taxa that have diverged genetically (i.e. gene trees are monophyletic) but experience occasional gene flow, would be expected to contain alleles that vary more widely in pairwise distance, representing haplotypes of the same species and another species. In contrast, alleles within a species that reflect retained ancestral polymorphisms should have few mutational differences and a narrower distribution of pairwise differences.

Several genera analysed in Fazekas *et al*. (2008) may bear the signature of incomplete lineage sorting. *Solidago* and *Dryopteris* both exhibit little or no genetic discontinuity among species but the values of the variance of this gap suggests a role for different mechanisms (Fig. 4). *Solidago* exhibits a uniformly small number of nucleotide differences among plastid DNA haplotypes within multiple species, reflecting a pattern of rapid successive speciation and recent divergence relative to coalescence time. In contrast, haplotype differences observed in *Dryopteris* (Fazekas *et al*. 2008) are more distinct and the gene tree supports paraphyly quite strongly, as might be expected with hybridization. Without more analyses of this kind, it is difficult to know whether incomplete lineage sorting is more likely in plants than animals.

Finally, beyond any biological attributes of plant species or technical limitations of barcoding, it is conceivable that past taxonomic practices may have contributed to the discordance between current taxonomy and genetic discontinuities in some plant groups. Incongruence between the taxonomic circumscription and historical patterns of gene flow can occur when species limits are either too inclusive (lumping) or too limited (splitting) (Funk & Omland 2003). For example, lumping of taxa into single species creates strong polyphyly, especially when taxa are not sister species. Incorrect splitting of a single species creates gene trees that are intermingled among taxa. To a certain extent, plant and animal taxonomists have adhered to different species concepts. While it may be argued that both disciplines have been reluctant to relinquish the typological representation of species (Mayr 1992), it appears that animal taxonomists have more generally embraced reproductive criteria, including the biological species concept, than plant taxonomists. In practice, species definitions in both cases are usually based on perceived morphological discontinuities rather than on measurements of gene flow and reproductive isolation. Nonetheless, operational differences in how species are defined may have led to delineation of plant taxa that do not correspond as well to genetic discontinuities.

Conclusion and future prospects

Many biologists have held the view that plant species are less well defined than vertebrate animals, due to a higher incidence of attributes such as asexual reproduction, polyploidy and hybridization (Stebbins 1950). While these phenomena may indeed be more widespread in plants, their importance has largely been inferred from case studies rather than using large-scale comparative analyses. Based on current barcoding data, it appears that plant species may be genetically less discrete than animals, although plant barcoding studies are still relatively limited in number and scope. We find that well-supported species monophyly is less common in plants and that the gap between intra- and interspecific genetic distances is less pronounced than in animal studies. As a result, discriminating plant species using single or multilocus barcodes from a single linkage group (the plastid genome) is likely to be a more challenging endeavour. Our main result for plants is likely to be robust to increased species sampling (logically, the upper limit to resolution can only decrease with improved species and population sampling).

Our analysis, albeit restricted in taxonomic breadth, suggests that species discrimination is not always limited by inadequate variability at the chosen locus. Rather, plant species resolution here appears constrained at a maximum of ~70% over a wide range of variability (parsimony-based estimates; Fazekas *et al*. 2008). This analysis is only based on variation at plastid loci for genera from temperate North America ($N = 32$). Nevertheless, it is consistent with values of discrimination reported from another plant barcoding study (Kress & Erickson 2007). Species resolution is also similar to estimates by Rieseberg *et al*. (2006) of the degree to which recognized plant species reflect reproductively

© 2009 Blackwell Publishing Ltd

independent lineages (70%). Therefore, it seems unlikely that adding more plastid DNA sequences would significantly improve this situation (although it would be of interest to test the robustness of this limit to species discrimination using taxa from other geographical regions, where evolutionary history may differ). Arguably, discrimination success may be even lower when more sister-species pairs and populations within species are included (Fazekas *et al*. 2008). Nonetheless, at this point, our barcoding data provide evidence that plant species boundaries are inherently less well defined than animals.

The difficulty in discriminating among some plant species in the study of Fazekas *et al*. (2008) may be related, in part, to hybridization in the genera examined, as has been suggested for poorly defined plant species boundaries in general (Stebbins 1950; Grant 1957). In contrast, Rieseberg *et al*. (2006) showed that polyploidy, and not hybridization, was statistically and negatively related to the degree of phenetic discontinuity among plant species. This difference in the importance of hybridization may reflect our smaller sample size, taxonomic bias or, alternatively, the higher likelihood of observing effects of gene flow at plastid loci compared to phenotypic characters. These factors may also explain why Rieseberg *et al*. (2006) found no differences in phenetic discontinuities between plant and animal species. It will be particularly important to evaluate genetic discontinuities among species across a wider range of taxa, and to test whether differences in hybridization can explain the observed disparity in genetic discontinuity between plant and animal species. Hybridization is widely viewed as being more common among plants; however, robust estimates of the incidence of hybridization are difficult to obtain. Some authors suggest that its influence on plant species delineation has been exaggerated (Mayr 1992; McDade 1995), or that hybridization in animals may have been underestimated (Arnold 1997).

What are the future prospects for improving barcoding success in plants? As mentioned, our results suggest that the problem may not be resolved simply by adding additional plastid DNA sequence data. This may certainly improve discrimination in some taxa — particularly those genera in which variability was lacking (even with multiple plastid regions sequenced). Additional sequence variation may help to resolve these cases or it may not, depending on the cause(s) of paraphyly. However, based on the asymptotic relationship in Fig. 1a, we would predict that the improvement in resolution would be incremental rather than transformative.

Other genomic regions such as nuclear encoded DNA may provide improved species resolution when used in combination with plastid DNA. Multiple nuclear genes from different linkage groups may offer some clear advantages. Synonymous substitution rates of nuclear genes are generally several times greater than plastid genes, which are three times greater than plant mitochondrial genes (Wolfe *et al*. 1987; Gaut *et al*. 1996; Hajibabaei *et al*. 2006c). Moreover, nuclear genes can provide a more reliable assessment of hybridization than uniparentally inherited plastid DNA (Chase *et al*. 2005). However, there are several challenges ahead in developing nuclear gene barcoding strategies. First, it may be difficult to find nuclear genes that are not only universally amplified but are also single copy across a wide range of plant taxa (e.g. Tank & Sang 2001; Mitchell & Wen 2004). Second, the effective population size of nuclear genes is four times higher than plastid DNA, making incomplete lineage sorting more likely. It may be possible to compensate for this performance advantage of organellar genes relative to individual nuclear markers by collecting multiple nuclear loci (accounting for ancestral polymorphisms; Pamilo & Nei 1988). However, it is not clear how in practice this would be implemented in a DNA barcoding context.

The benefits of nuclear DNA barcodes certainly may exist (see Small *et al*. 1999) but they could take some time to develop. In the meantime, one might consider identifying appropriate nuclear genes specific to each major order/family of land plants. This approach will likely yield benefits for taxonomically complex groups that may often lack sufficient variation at plastid DNA. However, in the longer term we agree with Chase *et al*. (2005) and Cowan *et al*. (2006) that the best strategy for nuclear genes may be to target a large number of short regions. Modern platforms such as pyrosequencing (Pacey-Miller & Henry 2003) allow many genes to be targeted and sequenced in one reaction (Margulies *et al*. 2005). Such high-throughput methods may facilitate the use of nuclear regions for future DNA barcoding efforts in plants.

This review uses plant barcoding data to explore the question of how prevalent non-monophyly is in plants and whether it may be inherently more common than in animals. In conducting our analyses, it became apparent that relatively few molecular data sets, outside of barcoding research, exist for plants in which individuals from multiple populations from several species within a genus are sampled. This sampling design falls between typical studies in plant systematics, which usually consist of many species each with low (or no) population-level replication, and population genetic sampling, which often focuses on one or rarely a few species with more intensive population sampling. Funk & Omland (2003) identified a similar problem with animals and recognized that only by more extensive population sampling can the hypothesis of species-level monophyly be adequately tested. In plants, future DNA barcoding studies with denser species sampling, more intensive geographical sampling of species, and perhaps the use of nuclear DNA sequences, will help to fill this void. In doing so, these approaches have the potential to offer powerful insights into the prevalence of non-monophyly and the very nature of species boundaries in plants and animals.

© 2009 Blackwell Publishing Ltd

Acknowledgements

This research was funded by a grant from Genome Canada through the Ontario Genomics Institute to the Canadian Barcode of Life Network.

Conflict of interest statement

The authors have no conflict of interest to declare and note that the funders of this research had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Arnold ML (1997) *Natural Hybridization and Evolution. (Oxford Series in Ecology and Evolution)*. Oxford University Press, Oxford, UK.

Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology*, **83**, 481–491.

Chase MW, Salamin N, Wilkinson M *et al*. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1889–1895.

Clare EL, Kim BK, Engstrom MD, Eger JL, Hebert PDN (2006) DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes*, **7**, 184–190.

Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300 000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon*, **55**, 611–616.

Crisp MD, Chandler GT (1996) Paraphyletic species. *Telopea*, **6**, 813–844.

Cywinska A, Hunter FF, Hebert PDN (2006) Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*, **20**, 413–424.

Fazekas AJ, Burgess KS, Kesanakurti PR *et al*. (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *Public Library of Science ONE*, **3**, e2802. doi:10.1371/journal.pone.0002802.

Foottit RG, Maw HEL, Von Dohlen CD, Hebert PDN (2008) Species identification of aphids (Insecta: Hemiptera: Aphididae) through DNA barcodes. *Molecular Ecology Resources*, **8**, 1189–1201. doi: 10.1111/j.1755-0998.2008.02297.x.

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution and Systematics*, **34**, 397–423.

Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences, USA*, **93**, 10274–10279.

Grant V (1957) In: *The Species Problem* (ed. Mayr E), pp. 38–90. AAAS, Washington, DC.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006a) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences, USA*, **103**, 968–971.

Hajibabaei M, Singer GAC, Hickey DA (2006b) Benchmarking DNA barcodes: an assessment using available primate sequences. *Genome*, **49**, 851–854.

Hajibabaei M, Xia J, Drouin G (2006c) Seed plant phylogeny: gnetophytes are derived confers and a sister group to Pinaceae. *Molecular Phylogenetics and Evolution*, **40**, 208–217.

Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, **270**, S96–S99.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *Public Library of Science Biology*, **2**, e312.

Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, Hebert PDN (2007) Comprehensive DNA Barcode coverage of North American Birds. *Molecular Ecology Notes*, **7**, 535–543. doi: 10.1111/j.1471-8286.2006.01670.x.

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *Public Library of Science ONE*, **2**, e508. doi: 10.1371/journal.pone.0000508.

Lahaye R, van der Bank M, Bogarin D *et al*. (2008) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences, USA*, **105**, 2923–2928. doi: 10.1073/pnas.0709936105.

Levin D (1979) The nature of plant species. *Science*, **204**, 381–384.

Lynch JD (1989) The gauge of speciation: on the frequencies of modes of speciation. In: *Speciation and its Consequences* (eds Otte D, Endler JA). Sinauer & Associates Inc, Sunderland, Massachusetts.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Margulies M, Egholm M, Altman WE *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Mayr E (1992) A local flora and the biological species concept. *American Journal of Botany*, **79**, 222–238.

McDade LA (1995) Species concepts and problems in practice: insight from botanical monographs. *Systematic Botany*, **20**, 606–622.

Meier R, Zhang G, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the 'Barcoding Gap' and leads to misidentification. *Systematic Biology*, **57**, 809–813.

Mitchell A, Wen J (2004) Phylogenetic utility and evidence for multiple copies of Granule-Bound Starch Synthase I (GBSSI) in Araliaceae. *Taxon*, **53**, 29–44.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, **8**, 480–490. doi: 10.1111/j.1471-8286.2007.02002.x.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Olmstead RG (1995) Species concepts and plesiomorphic species. *Systematic Botany*, **20**, 623–630.

Pacey-Miller T, Henry R (2003) Single-nucleotide polymorphism detection in plants using a single-stranded pyrosequencing protocol with a universal biotinylated primer. *Analytical Biochemistry*, **317**, 166–170.

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Molecular Biology and Evolution*, **5**, 568–583.

Rieseberg LH, Brouillet L (1994) Are many plant species paraphyletic? *Taxon*, **43**, 21–32.

Rieseberg LH, Wood TE, Baack EJ (2006) The nature of plant species. *Nature*, **440**, 524–527.

© 2009 Blackwell Publishing Ltd

Sang T, Zhong Y (2000) Testing hybridization hypotheses based on incongruent gene trees. *Systematic Biology*, **49**, 422–434.

Small RL, Ryburn JA, Wendel JF (1999) Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). *Molecular Biology and Evolution*, **16**, 491–501.

Smith MA, Poyarkov NA Jr, Hebert PDN (2008) *CO1* DNA barcoding amphibians: take the chance, meet the challenge. *Molecular Ecology Resources*, **8**, 235–246.

Stebbins GL Jr (1950) *Variation and Evolution in Plants*. Columbia University Press, New York.

Tank DC, Sang T (2001) Phylogenetic utility of the glycerol-3-phosphate acyltransferase gene: evolution and implications in *Paeonia* (Paeoniaceae). *Molecular Phylogenetics and Evolution*, **19**, 421–429.

Wakeley J (1996) Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology*, **49**, 369–386.

Ward RW, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **360**, 1847–1857.

Wolfe KH, Li WH, Sharpe PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA*, **84**, 9054–9058.

Supporting information

Additional supporting information may be found in the online version of this article:

Appendix S1 List of genera, number of species and source of data from which we calculated intra- and interspecific distance (K2P) (interspecific distance only for congeneric species). We included only genera with multiple species and species with multiple samples.

Appendix S2 Summary of the mean gap in genetic distance, incidence of polyploidy and hybridization, and dominant life history (woody; perennial; herbaceous) for 12 genera of land plants. We used the data to determine the relation between polyploidy, hybridization or life history and the degree of genetic discontinuity. We estimated the genetic distance gap from data in Fazekas *et al*. (2008) as the difference between the minimum interspecific genetic distance and maximum intraspecific genetic distance for each species within a genus, averaged across congeners. *N* refers to the number of species per genus; all species were represented by at least three barcode sequences. We determined the incidence of polyploidy by estimating the base chromosome number per genus and the ploidy (number of copies of the base chromosome set) in each species using the Index to Plant Chromosome Numbers (IPCN; Missouri Botanical Garden). We considered a genus variable in ploidy if variation existed either within or between species in our sample. Incidence of hybridization reflected the evidence of hybridization in the published literature on the species in our sample. We scored hybridization as present if any species within a genus was of confirmed hybrid origin or was known to hybridize with other congeners.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

© 2009 Blackwell Publishing Ltd