MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 3853–3856

NEWS AND VIEWS

COMMENT

Chloroplast DNA-based studies in molecular ecology may be compromised by nuclear-encoded plastid sequence

WOLFGANG ARTHOFER,* SILVIO SCHÜLER,† FLORIAN M. STEINER,*[1] and BIRGIT C. SCHLICK-STEINER*[1]

*Molecular Ecology Group, Institute of Ecology, University of Innsbruck, Technikerstrasse 25, 6020 Innsbruck, Austria,
†Institute of Genetics, Federal Research and Training Centre for Forests, Natural Hazards and Landscape, Vienna, Austria

Ongoing genetic transfer from mitochondria and plastids into the nucleus is a well-documented fact. While in metazoan molecular ecology the need for surveillance against pseudogene-mediated artefacts when analysing mtDNA sequences is commonly accepted, no comparable measurements have been established for plastid-based studies. We highlight the impact and management of nuclear mitochondrial insertions, argue that nuclear plastid sequences represent an underestimated but major factor in plant molecular ecology, and discuss potential avenues of remedy in chloroplast studies.

Keywords: chloroplast DNA, mitochondrial DNA, nuclear mitochondrial, nuclear plastid, phylogeny, phylogeography, population genetics, pseudogene

Received 3 May 2010; revision received 1 July 2010; accepted 7 July 2010

In a recent publication in *Molecular Ecology Resources*, Naciri & Manen (2010) report on the finding of divergent chloroplast DNA (cpDNA) sequences in the angiosperm *Eryngium alpinum* and hypothesize nuclear insertions of cpDNA as the source of their observation. Their inference is, in our mind, an important one. Ongoing genetic transfer from mitochondria and plastids into the nucleus is a well-documented fact (Ayliffe & Timmis 1992; Zhang & Hewitt 1996; Martin & Herrmann 1998; Bensasson *et al.* 2001; Richly & Leister 2004; Kleine *et al.* 2009). Several studies demonstrated that nuclear copies of mitochondrial DNA (mtDNA) may lead to erroneous phylogenetic inferences (e.g., Zhang & Hewitt 1996; Sorenson & Quinn 1998; Bensasson *et al.* 2001; Thalmann *et al.* 2004). Today, it is widely accepted that any use of mtDNA in molecular ecology requires tests to prove a true mitochondrial origin of the underlying sequences (Song *et al.* 2008 and references therein). In contrast, awareness of pseudogene-mediated artefacts is still largely lacking for chloroplasts, as argue Naciri & Manen (2010) for plant phylogeography and molecular diagnostics. Here, we go beyond their paper by putting the problem in the more general context of molecular ecology as a whole. We summarize the impact and management of nuclear mitochondrial (NUMT) insertions, make some example-based assumptions on the abundance of nuclear plastid (NUPT) sequences, and briefly discuss the problem of how to identify and eliminate NUPT-based artefacts.

MtDNA has been the major source of molecular phylogenetic data of metazoans for almost two decades (Avise 2009). In 1996, Zhang *et al.* reviewed NUMT-based errors in mtDNA studies for the first time and proposed double bands, unexpected frameshifts, sequence ambiguities, and contradictory tree topologies as major signs of NUMT contamination. In 2001, studies of 64 metazoan species had been confirmed to contain pseudogene artefacts (Bensasson *et al.* 2001), and, as an example of a more recent review, Yao *et al.* (2008) documented several NUMT-based errors in clinical disease studies. Of great practical concern, Benesh *et al.* (2006) demonstrated a case of preferential binding of universal primers to a NUMT sequence. The total frequency of NUMT detections is difficult to assess as they are infrequently reported following their disclosure (cf. Beckenbach 2009), but as of 29 April 2010, GenBank/Nucleotides contained 757 entries designated as 'NUMT'. The current toolkit for NUMT identification (see Song *et al.* 2008 for a review) includes (i) testing for PCR ghost bands after mtDNA amplification, (ii) examination of sequence chromatograms for ambiguities, including examination of PHRED scores, (iii) sequence translation and test for indels, frameshifts, and premature stop codons, (iv) sequence examination for compositional biases, (v) comparison with mtDNA sequences of closely related taxa, and (vi) BLAST analysis. Strategies for contamination avoidance rely on (a) preferred use of mtDNA rich tissues, (b) mtDNA enrichment, (c) long PCR, (d) reverse transcription PCR, and (e) use of taxon-specific primers; the latter, however, did not prove very effective in avoiding NUMT amplification in a recent study (Moulton *et al.* 2010).

Comparable to mtDNA in metazoans, cpDNA has been the most important molecular tool in phylogenetic and phylogeographic studies of plants for the past quarter century (e.g., Palmer 1987; Clegg 1993; Wu *et al.* 2007; Qiu *et al.* 2009). The maternal inheritance in most plant genera allowed using cpDNA for the inference of relationships among nearly all taxonomic units ranging from classes (e.g. among the basal angiosperms, Graham & Olmstead

Correspondence: Wolfgang Arthofer, Fax: +43 512507 6190;
E-mail: wolfgang.arthofer@uibk.ac.at
[1]These authors contributed equally to this paper.

2000) to intraspecific races and populations (Petit *et al.* 1993, 1997). Phylogenetic studies mainly focused on sequence variation patterns in coding regions, whereas phylogeographic variation patterns were widely studied based on non-coding sequences (Taberlet *et al.* 1991; Demesure *et al.* 1995).

Although evidence of cpDNA transfer to the nucleus dates back to the time when NUMTs slowly became a topic in metazoan phylogenetics (Ayliffe & Timmis 1992) and although NUPTs were early identified in genome-sequencing projects of higher plants (Richly & Leister 2004), a limited number of papers have dealt in depth with the topic until today (e.g. Shamuradov *et al.* 2003; Leister 2005; Matsuo *et al.* 2005; Kejnovsky *et al.* 2006; Kleine *et al.* 2009). Apart from the recent paper by Naciri & Manen (2010), only Meimberg *et al.* (2006) emphasized potential consequences of NUPT contamination for the reconstruction of plant phylogenies. GenBank/Nucleotide currently contains no entries designated as 'NUPT'.

We performed a MEGABLAST search of whole-chloroplast sequences against the corresponding nuclear genomes of four plant species, and an analogous MEGABLAST of mitochondrial vs. nuclear sequences in two animal species in which NUMTs have been documented. Hits with an alignment length of at least 200 bp and a similarity of at least 85% were assumed to represent pseudogenes and are summarized in Table 1. The query indicates that NUPTs and NUMTs show quite similar values for length and sequence similarity but that in three of the plant species tested here, the absolute (i.e. uncorrected for genome sizes) numbers of NUPTs are far higher than the amount of NUMTs in the inquired animals. The NUPT similarity distribution is given in Fig. 1. Except for *Arabidopsis thaliana*, in all species exact matches were found, and the similarity class of <100 to 97% similarity to the cpDNA sequence was the one with the highest number of NUPTs.

An *in silico* PCR using the primers for the *trnH-psbA* spacer region proposed for land plant barcoding (Kress & Erickson 2007) suggests the possibility of amplifying a NUPT for *Vitis vinifera* (Table 2). The assumptive NUPT is located on chromosome 6 and has 92.7% similarity to the corresponding cpDNA sequence. We suggest that the results from our exercise be viewed as a proof of concept

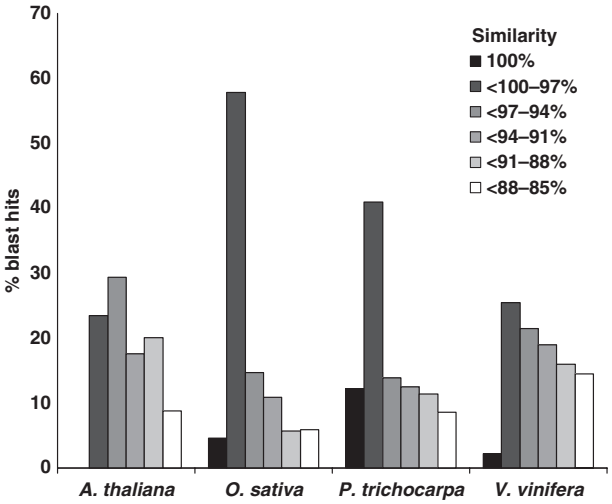

Fig. 1 Histograms of similarity distribution of BLAST hits in four plant species. Except for *Arabidopsis thaliana*, the similarity class of <100 to 97% similarity to the cpDNA sequence was the one with the highest number of potential NUPTs.

of a significant relevance of NUPTs to cpDNA-based studies and argue that NUPT contamination is likely to have a similarly distractive impact on such studies as NUMTs are established to have on mtDNA-based ones. This is underlined by recent findings on NUPT distribution and replication number in rice, where NUPTS were found on all 12 chromosomes and where 53 of 60 known plastid genes were found to occur in multiple copies throughout the nuclear genome (Akbarova *et al.* 2010). Future studies may focus on patterns of length and similarity variation among NUPT sequences, the distribution of NUPTs in the nuclear genome of more species, and whether NUPTs were replicated in repetitive regions of the chromosomes.

Accepting NUPTs as of general relevance in molecular ecology means that for reliable inferences of species and population histories, procedures for NUPT management need to be implemented. What strategies are available? On the short term, strategies successfully applied against NUMT contamination, like alertness for ghost bands and

Table 1 MEGABLAST hits of organelle vs. nuclear genomes in four plant (*Arabidopsis thaliana*, *Oryza sativa japonica*, *Populus trichocarpa*, *Vitis vinifera*) and two animal species (*Canis lupus familiaris*, *Pan troglodytes*). Only hits with an alignment size ≥200 bp and a similarity ≥85% were considered

	n hits	Max length (bp)	Mean length (bp)	Median length (bp)	Mean similarity (%)	Median similarity (%)
Arabidopsis thaliana	34	3638	638.8	495	92.9	94
Oryza sativa japonica	543	56200	2167.3	451	95.9	98
Populus trichocarpa	1011	3180	671.9	485	95.2	97
Vitis vinifera	1012	4818	563.7	334	92.8	93
Canis lupus familiaris	47	3427	907.0	535	92.1	93
Pan troglodytes	16	1055	521.3	386	87.7	87

Table 2 Results from *in silico* PCR with *trnH-psbA* primers in *Vitis vinifera*

	cpDNA	NUPT
Amplicon length (bp)	389	398
Mismatches in fwd primer region	0	0
Mismatches in rev primer region	2	2
CG content (%)	29.0	26.1
Gaps	3	0
Matches		369

double peaks, and circumvention of preferential pseudogene amplification by the use of multiple primers and loci, should be a good starting point. In many studies, use of plastid-rich leaf tissue is the rule. Particular attention, however, is required if herbarium specimens (e.g. Savolainen *et al.* 1995) or ancient wood and seed tissues (Parducci & Petit 2004) are being analysed as is frequently the case with, for example, phylogenetic studies. In such instances, the sequences of the envisaged cpDNA regions should be compared with those derived from fresh leaf material. Identification and report of any artefact will help in fine-tuning of NUPT surveillance tools. Two important exceptions, however, make it more challenging to approach NUPTs when compared to NUMTs. First, the relatively small size of most metazoan NUMTs allows their elimination by long PCR (Arthofer *et al.* 2006). The reported existence of giant NUPTs (Matsuo *et al.* 2005) with sizes >10 kbp in some species discourages this approach in plants. In line with this, also in all species examined here, NUPTs exceeding 2000 bp were found, albeit the majority of NUPT sequences were shorter than 1000 bp (Fig. 2). It is noteworthy that also in plants, the length of NUMT inserts sometimes exceeds the range known from metazoan

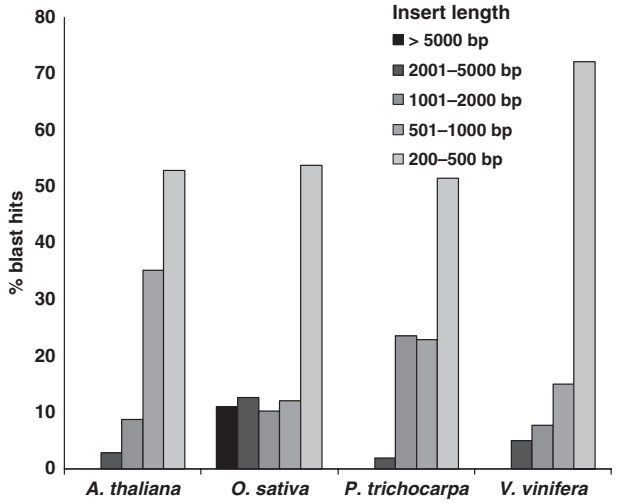

Fig. 2 Histograms of length distribution of BLAST hits in four plant species. The majority of potential NUPTs were shorter than 500 bp, but long insertions exceeding 2000 bp were found in all species.

species by far (extreme example: a 620-kbp insertion in *A. thaliana*, Stupar *et al.* 2001). Second, a typical 150- to 162 -kbp-sized chloroplast genome consists of 40–50% non-coding regions, including intergenic spacers and introns (Palmer 1987). In such regions, translation-based tests are not applicable. On the medium term, a remedy to the two problems may arise from the increasing ease of next-generation sequencing of individual genomes. While still facing some cost and bioinformatics constraints today, promising studies using these technologies in molecular ecology are at hand already (Tautz *et al.* 2010; chloroplast example: Whittall *et al.* 2010). We should not shrink back from expecting the full realization of their effects (Gilad *et al.* 2009) for use in everyday and every laboratory routine to include new approaches to identifying pseudogenes in general.

Acknowledgements

We thank Astrid Haara for assistance in data mining, and News and Views Editor Nolan Kane and three anonymous referees for constructive criticism.

References

Akbarova YY, Solovyev VV, Shahmuradov IA (2010) Possible functional and evolutionary role of plastid DNA insertions in rice genome. *Applied and Computational Mathematics*, **9**, 19–33.

Arthofer W, Avtzis DN, Riegler M, Miller W, Stauffer C (2006) 'Pitfalls in Applying Mitochondrial Markers onto the scolytid species *Pityogenes chalcographus*'. In:*Proceedings from the Third Workshop on Genetics of Bark Beetles and Associated Microorganism* (eds Bentz B, Raffa K). pp. 15–19, U.S. Department of Agriculture, Rocky Mountain Research Station.

Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Ayliffe MA, Timmis JN (1992) Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA. *Theoretical and Applied Genetics*, **85**, 229–238.

Beckenbach AT (2009) Numts and mitochondrial pseudogenes. *Myrmecological News*, **12**, 217–218.

Benesh DP, Hasu T, Suomalainen LR, Tellervo Valtonen E, Tiirola M (2006) Reliability of mitochondrial DNA in an acanthocephalan: the problem of pseudogenes. *International Journal for Parasitology*, **36**, 247–254.

Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, **16**, 314–321.

Clegg MT (1993) Chloroplast gene sequences and the study of plant evolution. *Proceedings of the National Academy of Sciences of the USA*, **90**, 363–367.

Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Molecular Ecology*, **4**, 129–131.

Gilad Y, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics*, **25**, 463–471.

Graham SW, Olmstead RG (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany*, **87**, 1712–1730.

Kejnovsky E, Kubat Z, Hobza R *et al.* (2006) Accumulation of chloroplast DNA sequences on the Y chromosome of *Silene latifolia*. *Genetica*, **128**, 167–175.

Kleine T, Maier UG, Leister D (2009) DNA transfer from the organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual Review of Plant Biology*, **60**, 115–138.

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. *PLoS ONE*, **6**, e508.

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends in Genetics*, **21**, 655–663.

Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiology*, **118**, 9–17.

Matsuo M, Ito Y, Yamauchi R, Obokata J (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *The Plant Cell*, **17**, 665–675.

Meimberg H, Thalhammer S, Brachmann A, Heubl G (2006) Comparative analysis of a translocated copy of the *trnK* intron in carnivorous family Nepenthaceae. *Molecular Phylogenetics and Evolution*, **39**, 478–490.

Moulton MJ, Song H, Whiting MF (2010) Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). *Molecular Ecology Resources*, **10**, 615–627.

Naciri Y, Manen JF (2010) Potential DNA transfer from the chloroplast to the nucleus in *Eryngium alpinum*. *Molecular Ecology Resources*, **10**, 728–731.

Palmer JD (1987) Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *The American Naturalist*, **130**, S6–S29.

Parducci L, Petit RJ (2004) Ancient DNA – unlocking plants' fossil secrets. *New Phytologist*, **161**, 335–339.

Petit RJ, Kremer A, Wagner DB (1993) Geographic structure of chloroplast DNA polymorphisms in European oaks. *Theoretical and Applied Genetics*, **87**, 122–128.

Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A (1997) Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences of the USA*, **94**, 9996–10001.

Qiu Y-X, Guan B-C, Fu C-X, Comes HP (2009) Did glacials and/or interglacials promote allopatric incipient speciation in East Asian temperate plants? Phylogeographic and coalescent analyses on refugial isolation and divergence in Dysosma versipellis. *Molecular Phylogenetics and Evolution*, **51**, 281–293.

Richly E, Leister D (2004) NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution*, **21**, 1081–1084.

Savolainen V, Cuénoud P, Spichiger R, Martinez MDP, Crèvecoeur M, Manen JF (1995) The use of herbarium specimens in DNA phylogenetics: evaluation and improvement. *Plant Systematics and Evolution*, **197**, 87–98.

Shamuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *Plant Molecular Biology*, **52**, 923–934.

Song H, Buhay JE, Whiting MF, Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the USA*, **105**, 13486–13491.

Sorenson MD, Quinn TW (1998) Numts: a challenge for avian systematics and population biology. *The Auk*, **115**, 214–221.

Stupar RM, Lilly JW, Town CD *et al.* (2001) Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. *Proceedings of the National Academy of Sciences of the USA*, **98**, 5099–5103.

Taberlet P, Gielly L, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, **17**, 1105–1109.

Tautz D, Ellegren H, Weigel D (2010) Next generation molecular ecology. *Molecular Ecology*, **19**, 1–3.

Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L (2004) Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, **13**, 321–335.

Whittall JB, Syring J, Parks M *et al.* (2010) Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology*, **19**, 100–114.

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 slants. *Molecular Biology and Evolution*, **24**, 1366–1379.

Yao YG, Kong QP, Salas A, Bandelt HJ (2008) Pseudomitochondrial genome haunts disease studies. *Journal of Medical Genetics*, **45**, 769–772.

Zhang DX, Hewitt GM (1996) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends in Ecology & Evolution*, **11**, 247–251.

W Arthofer, FM Steiner and BC Schlick-Steiner focus on terrestrial animals (insects, harvestmen, millipedes, spiders) and their symbionts (ascomycete fungi, bacteria) of the Alpine environment. Their research topics include Alpine endemism, biogeography, climate-change and conservation biology, morphology and integrative taxonomy and the development of markers for population genetics.

S Schüler is interested in geno- and phenotypic variation of forest trees and the analysis of natural and anthropogenic influences shaping this variation. He utilizes molecular methods with the goal of developing guidelines for sustainable forest management.

---

doi: 10.1111/j.1365-294X.2010.04787.x