Mitochondrial pseudogenes: evolution's misplaced witnesses

Douda Bensasson, De-Xing Zhang, Daniel L. Hartl and Godfrey M. Hewitt

Nuclear copies of mitochondrial DNA (mtDNA) have contaminated PCR-based mitochondrial studies of over 64 different animal species. Since the last review of these nuclear mitochondrial pseudogenes (Numts) in animals, Numts have been found in 53 of the species studied. The recent evidence suggests that Numts are not equally abundant in all species, for example they are more common in plants than in animals, and also more numerous in humans than in *Drosophila*. Methods for avoiding Numts have now been tested, and several recent studies demonstrate the potential utility of Numt DNA sequences in evolutionary studies. As relics of ancient mtDNA, these pseudogenes can be used to infer ancestral states or root mitochondrial phylogenies. Where they are numerous and selectively unconstrained, Numts are ideal for the study of spontaneous mutation in nuclear genomes.

Douda Bensasson*
Daniel L. Hartl
Dept of Organismic and Evolutionary Biology, 16 Divinity Avenue, Harvard University, Cambridge, MA 02138, USA.
*e-mail: douda@stanford.edu

De-Xing Zhang
Institute of Zoology, Chinese Academy of Sciences, Beijing, People's Republic of China.

Godfrey Hewitt
School of Biological Sciences, University of East Anglia, Norwich, UK NR4 7TJ.

The first report of the possible existence of DNA, homologous to mitochondrial DNA (mtDNA) but integrated into a nuclear genome[1], was in 1967, shortly after the discovery that organelles have separate DNA. Du Buy and Riley's study showed that the purified mtDNA of mice hybridized more strongly to mouse nuclear DNA than would be expected without regions of high homology. Since then, there have been many reports of mitochondrial PSEUDOGENES (see Glossary) integrated into the nuclear genomes of rodents, including rats[2,3], mice[4], voles[5], guinea-pigs[4], tuco-tucos (South American mouse)[6], and over 77 other eukaryotes.

The use of PCR to study mtDNA without prior purification of mtDNA, has led to many accidental amplifications of nuclear mitochondrial pseudogenes (NUMTS[7]). When a species has Numt regions PARALOGOUS to the mtDNA region of interest, PCR primers will sometimes hybridize to Numts as well as[8], or in preference to[9,10], mitochondrial sequences. The undiscovered presence of Numts can lead to robust, believable, but incorrect, genetic phylogenies[10]. As unrecognized contaminants, human Numts have been mistaken for dinosaur mtDNA (Ref. 11) and ancient monkey sequences[12]. Numts that were inadvertently PCR amplified from patients with Alzheimer's disease have also been mistaken for HETEROPLASMIC MTDNA mutations causing the disease[13].

Background

There has been recent speculation on the reasons why Numts exist[14–16]. The transfer of mtDNA to the nucleus could be part of an ongoing transfer of functional mitochondrial genes from mitochondrion to nucleus. One of the underlying causes of this process is thought to be a 'gene transfer ratchet'[14]. Because the rate of DNA transfer from mitochondrion to nucleus is thought to be much greater than in the reverse direction[17], a net movement of genes from mitochondrion to nucleus occurs[14]. Preferential DNA transfer to the nucleus has been shown experimentally in yeast[17], and no foreign DNA integrations have been observed in metazoan mitochondrial genomes. The invasion of plant organellar genomes by extraorganellar DNA, including nuclear DNA (Ref. 18), suggests that biased transfer to the nucleus might be less pronounced in plants than in animals.

Different stages of the nuclear acquisition of functioning mitochondrial genes have been documented in plants and fungi[18], and the transfer of chloroplast genes in plants is an ongoing process[19]. By contrast, there are no reports of recent transfer of functional mitochondrial sequences in metazoan nuclei[18], suggesting that this is no longer an ongoing process in metazoa. Perhaps the transfer in metazoa is limited because only a small set of vital genes remain in their mitochondrial genomes, which have a distinct mitochondrial genetic code[20,21].

Numts come in many sizes, from all types of mtDNA sequence, and bear varying degrees of similarity to their mitochondrial counterparts[8]. The chromosomal distribution of Numts is also varied. Although they typically occur in single copies at dispersed genomic locations[3,22,23], Numts are tandemly repeated at one locus in cats[7], and are telomeric, centromeric or interspersed in different grasshopper species[24]. Fragments from disparate parts of the mitochondrial genome are sometimes arranged tandemly[3,22]. In primates, cats, grasshoppers and *Sitobion* aphids, where they exist in high COPY NUMBERS, Numts have arisen through many independent transfers from mitochondrion to nucleus[25–30], as well as through amplification of a single Numt type (Box 1)[7,28,31].

The incidence of Numts in animals, their importance and potential uses in molecular ecology and evolutionary biology was fully reviewed in 1996 (Ref. 8) and there has been a recent review for birds[10]. Here we summarize the emerging taxonomic distribution of Numts across all eukaryotic taxa, review insights into the conditions for their co-amplification with mtDNA and methods for their avoidance. We also review what is understood about the mode of Numt evolution, and examples of Numt utility, including their use in rooting

http://tree.trends.com 0169–5347/01/$ – see front matter PII: S0169-5347(01)02151-6

Box 1. Mechanisms of Numt generation

Nuclear mitochondrial pseudogenes (Numts) arise both with and without RNA intermediates[a,b].Their integration into the nuclear genome was originally associated with transposable elements or short dispersed repeats[c], but close examination of many different Numt loci reveals a lack of common features at integration sites[a,b,d,e]. A possible explanation for these integrations is that they are incorporated into the nuclear genome during the repair of chromosomal breaks by nonhomologous recombination[a,b]. Recent experiments in yeast provide strong support to this hypothesis[f,g]. Such a mechanism is possible if there are mitochondrial DNA (mtDNA) fragments loose in the nucleus.This does appear to be the case, at least in yeast[h], and a bewildering array of causes and mechanisms have been proposed for the movement of mtDNA, which appears to be influenced by both genetic and environmental factors[h,i]. Factors influencing the escape of mtDNA from mitochondria include the action of mutagenic agents and other forms of cellular stress that can damage mitochondria or their membranes[h]. It has been postulated that the random insertion of mtDNA into nuclear genomes, associated with such mitochondrial stresses, could be a cause of cancer or of ageing[i–k].

References

a Blanchard, J.L. and Schmidt, G.W. (1996) Mitochondrial DNA migration events in yeast and humans: Integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution pattern. *J. Mol. Evol.* 13, 537–548

b Blanchard, J.L. and Schmidt, G.W. (1995) Pervasive migration of organellar DNA to the nucleus in plants. *J. Mol. Evol.* 41, 397–406

c Gellissen, G. and Michaelis, G. (1987) Gene transfer: mitochondria to nucleus. *Ann. New York Acad. Sci.* 503, 391–401

d Zischler, H. (2000) Nuclear integrations of mitochondrial DNA in primates: Inference of associated mutational events. *Electrophoresis* 21, 531–536

e Fukuda, M. *et al.* (1985) Mitochondrial DNA-like sequences in the human nuclear genome. *J. Mol. Biol.* 186, 257–266

f Ricchetti, M. *et al.* (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402, 96–100

g Yu, X. and Gabriel, A. (1999) Patching broken chromosomes with extranuclear cellular DNA. *Mol. Cell* 4, 873–881

h Thorsness, P.E. and Weber, E.R. (1996) Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. *Int. Rev. Cytol.* 165, 207–234

i Shafer, K.S. *et al.* (1999) Mechanisms of mitochondrial DNA escape to the nucleus in the yeast *Saccharomyces cerevisiae*. *Curr. Genet.* 36, 183–194

j Corral, M. *et al.* (1989) DNA-sequences homologous to mitochondrial genes in nuclei from normal rat-tissues and from rat hepatoma-cells. *Biochem. Biophys. Res. Commun.* 162, 258–264

k Hadler, H.I. *et al.* (1998) Selected nuclear LINE elements with mitochondrial-DNA-like inserts are more plentiful and mobile in tumor than in normal tissue of mouse and rat. *J. Cell. Biochem.* 68, 100–109

phylogenetic trees, resolving branches and in studying spontaneous mutation.

A census of mitochondrial pseudogenes

In 1996, Blanchard and Schmidt[3] proposed that Numts are not equally abundant in all species. They observed that a larger proportion of the plant nuclear sequence in GenBank was of mitochondrial origin compared with the nuclear sequences of yeast or humans, and that no Numts were identified in *Plasmodium falciparum*, *Caenorhabditis elegans* or *Drosophila melanogaster* despite these being well-studied species. There might also be differences in Numt abundance among closely related species. For example, Numts are very abundant in *Sitobion* aphids, but they might be less frequent in the other four aphid genera investigated[27].

The current data still support the idea that some species have more Numts than do others. A search of the sequences that have been mapped onto chromosomes in GenBank reveals very few Numts for *P. falciparum*, *C. elegans* and *D. melanogaster* compared with *Homo sapiens* (Table 1).

Numts have been found in over 82 different species (Table 2: the full list is at http://www.pseudogene.net), including all well-studied eukaryotic groups. There are more independent discoveries of Numts in plants than would be expected considering how often their mtDNA is studied (as estimated from the number of plant mtDNA sequences deposited in GenBank, Table 2). It is not clear whether the abundance of Numts in fungi is closer to that of metazoans or plants, but plants appear to have a much larger proportion of Numts (estimated at 3–7% in 1995; Ref. 22) than do metazoans[3]. Plants also have much larger mitochondrial genomes (367 kb for *Arabidopsis thaliana*) and more mobility among cellular compartments than do metazoans[18].

Although Numts are distributed fairly evenly across the major metazoan groups, close examination within these groups reveals some differences. They are abundant in mammals and birds, but not unusually so considering how intensively these groups are studied. Numt discoveries seem to be unusually common in Orthoptera (Table 2) (more specifically, grasshoppers). Interestingly, despite Numts being checked for in fish, and the fact that fish are well studied, no Numts have been reported in these animals (Werner E. Mayer, pers. commun.).

Table 1. Abundance of Numts in different species[a]

Species	mt genome size (bp)	Database size (bp)[b]	% of genome checked	No. of Numts[b]	Total length of Numts (bp)	Numt % of genome[b]
Saccharomyces cerevisiae	85 779	12 069 247	100	17	1389	0.012
Plasmodium falciparum	5967	28 718 804	50	3	228	0.0008
Caenorhabditis elegans	13 794	106 660 070	~100	2	212	~0.0002
Drosophila melanogaster	19 496	122 655 632	70	3	724	0.0006
Homo sapiens	16 569	2 853 531 108	84	354	418 552	0.012

[a]Abbreviations: bp, base pairs; Numt, nuclear mitochondrial pseudogenes.
[b]Summary of Numts in GenBank genome projects. Mitochondrial genomes were queried against nuclear mapped genome databases using the NCBI BLAST search facilities (http://www.ncbi.nlm.nih.gov/BLAST, http://www.ncbi.nlm.nih.gov/Malaria/plasmodiumblcus, http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast) set at default parameters (16 February 2001). Only data with < 0.0001 probability of occurring by chance were included.

Table 2. Summary of taxonomic groups in which Numts have been discovered[a]

Organisms	Number of sp. with Numts	Number of independent Numt reports[b]	Number of mtDNA sequences in GenBank	Expected number of independent Numt reports[c]	Refs[d]
Fungi	3	4	1952	2	3
Viridiplantae	10	7	1423	1	19,22
Metazoa	>69	At least 35	47 562	43	See below[e]
Onychophora	1	1	106	0	–
Annelida	0	0	162	0	–
Pseudocoelomata	0	0	398	0	–
Echinodermata	1	1	743	1	8
Mollusca	0	0	2573	2	–
Arthropoda	>23	7	9829	7	See below
Crustacea	>2	2	1253	1	30
Insecta	>21	5	7809	6	See below
Orthoptera	14	3	248	0	20,24,28,37
Hemiptera	>1	1	756	1	27
Hymenoptera	6	1	917	1	–
Other Insecta	0	0	5905	4	–
Chelicerata	0	0	751	1	–
Other Arthropoda	0	0	25	0	–
Chordata	>44	At least 26	33 121	24	See below
Sauropsida	>18	8	9462	7	See below
Aves	>18	8	5701	4	10,37,49
Other[f]	0	0	3761	3	–
Mammalia	>26	At least 18	13 535	10	2–9,11–13,23, 25,26,29,31, 34,35,38,45
Actinopterygii[g]	0	0	8160	6	–
Amphibia	0	0	1796	1	–
Other Chordates	0	0	142	0	–
Other Metazoa	0	0	634	0	–
Other Eukaryotes	0	0	420	0	–
Eukaryote TOTAL	>82	At least 46	51 357	–	–

[a]Abbreviations: mtDNA, mitochondrial DNA; Numt, nuclear mitochondrial pseudogenes.
[b]Number of references with nonoverlapping authorship in this group of organisms.
[c]Expected values are estimated assuming Numts are ubiquitous and their discovery is proportional to the number of mitochondrial sequences deposited in GenBank for that group. That is, number of mtDNA sequences in GenBank × frequency of Numt reports in Eukaryotes (46/51 357), or for Metazoa × frequency of Numt reports in Metazoa (35/47 562).
[d]These references give details of which species Numts are found in, but do not represent independent discoveries of Numts, and are not an exhaustive list. The full list is posted at http://www.pseudogene.net.
[e]See below refers to the references given for each individual group.
[f]Other Sauropsida (lizards and crocodiles).
[g]Ray-finned fish.

The distribution of Numts in metazoans (Tables 1 and 2) suggests that there may be a correlation between genome size and the number of mitochondrial sequences present in the nucleus, as there is for some other types of sequence [e.g. microsatellite DNA (Ref. 32) and rDNA (Ref. 33)]. Grasshoppers have genomes that are larger (5950–20 600 Mb) than those of mammals (1420–5680 Mb), of birds (1670–2250 Mb), or of most other insects (98–8900 Mb); humans (3400 Mb) have much larger genomes than do *D. melanogaster* (176 Mb) or *C. elegans* (86 Mb). If the frequency of Numts in noncoding regions is similar across different metazoa, then animals with more noncoding nuclear DNA would be expected to have more mitochondrial pseudogenes. A correlation with genome size is a reasonable null hypothesis, because differences among species in the Numt proportion of noncoding DNA would suggest differences in the mechanisms that govern DNA gain or loss.

A deeper understanding of the taxonomic distribution of Numts would be of broad scientific interest. It could indicate which genomes acquire extranuclear DNA that is not inherently selfish more readily, and why. Such information is of potential use in transgenic and medical research as well as being useful for evolutionary biologists and molecular ecologists.

Molecular troublemakers

Collura and Stewart[9] discovered that, under the same experimental conditions, Numts can be preferentially

Box 2. Checking and avoiding Numts

Symptoms of nuclear mitochondrial pseudogene (Numt) contamination include PCR ghost bands, extra bands in restriction profiles[a], sequence ambiguities (particularly if they are at polymorphic sites, or if they are encountered when sequencing from both strands), frameshift mutations, stop codons and unexpected phylogenetic placements. Restriction enzyme[b], SINGLE-STRANDED CONFORMATION POLYMORPHISM (SSCP) (see Box Glossary)[c], CONSTANT DENATURANT CAPILLARY ELECTROPHORESIS (CDCE)[d], and cloning and sequencing of PCR products[e,f] approaches have been used to establish if more than one mtDNA-like sequence has been amplified.

Numts can be avoided, and their nonmitochondrial location established, if the proportion of amplified mtDNA is increased. This can be done by purifying mitochondria before DNA extraction, by long PCR amplification[a], or by using tissue that is rich in mtDNA relative to nuclear DNA (e.g. muscle)[a,g]. Although all of these methods are effective, none are guaranteed[a] (http://www.pseudogene.net).

Numts can be avoided by reverse transcriptase PCR (RT–PCR)[h], but occasionally Numts are transcribed[i]. Where mtDNA and Numt sequences are known, and mtDNA is monophyletic, mtDNA-specific primers can be designed[a]; where Numts are monophyletic, they can be digested with restriction enzymes before PCR (Ref. a). Alternatively, if PCR products are cloned and sequenced, it is sometimes possible to infer, from the mode of evolution observed among individual clones, which sequence are mitochondrial and which are nuclear[f]. (For further discussion, see Refs a, j, and http://www.pseudogene.net).

Box Glossary

Single-stranded conformation polymorphism (SSCP): a technique by which DNA molecules, of different nucleotide sequence, are separated on electrophoretic gels. In their single-stranded form, DNA molecules, which differ in their nucleotide sequence, are folded differently (have different conformations). The conformation differences result in differences in electrophoretic mobility, which can be visualized on denaturing polyacrylamide gels.

Constant denaturant capillary electrophoresis (CDCE): a capillary electrophoresis method for distinguishing DNA molecules of different nucleotide sequence, because of differences in their melting points.

References

a Sorenson, M.D. and Quinn, T.W. (1998) Numts: A challenge for avian systematics and population biology. *The Auk* 115, 214–221

b Zhang, D-X. and Hewitt, G.M. (1996) Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. *Mol. Ecol.* 5, 295–300

c Sunnucks, P. *et al.* (2000) SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Mol. Ecol.* 9, 1699–1710

d Li-Sucholeiki, X-C. *et al.* (1999) Applications of constant denaturant capillary electrophoresis / high fidelity polymerase chain reaction to human genetic analysis. *Electrophoresis* 20, 1224–1232

e Bensasson, D. *et al.* (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Mol. Biol. Evol.* 17, 406–415

f Lemos, B. *et al.* (1999) Mitochondrial DNA-like sequences in the nuclear genome of the opossum genus *Didelphis* (Marsupialia: *Didelphidae*). *J. Heredity* 90, 543–547

g Greenwood, A. and Paabo, S. (1999) Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. *Mol. Ecol.* 8, 133–137

h Collura, R.V. *et al.* (1996) A quick direct method that can differentiate expressed mitochondrial genes from their nuclear pseudogenes. *Curr. Biol.* 6, 1337–1339

i Blanchard, J.L. and Schmidt, G.W. (1996) Mitochondrial DNA migration events in yeast and humans: Integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution pattern. *J. Mol. Evol.* 13, 537–548

j Zhang, D-X. and Hewitt, G.M. (1996) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends Ecol. Evol.* 11, 247–251

PCR amplified in the orang-utan, although mtDNA predominates in the PCR products of other hominoids. This discrepancy arose because the authentic orang-utan mtDNA sequence had diverged in such a way that the primers used would no longer amplify it, although they were still able to amplify an ancient mitochondrial pseudogene in the nuclear genome. If the phylogenetic placement of orang-utans was not well known, and the amplified Numts had not had a frameshift mutation and two stop codons, the 'mitochondrial' sequence phylogeny would have placed orang-utans with Old World monkeys. Such errors also occur in other animal groups[30], for example, Arctander[34] describes how inadvertent Numt amplification leads to inference of an incorrect phylogenetic relationship among passerine birds.

Undiscovered phylogenetic mistakes could also occur when Numt copy numbers vary among individuals, or when the proportion of mtDNA sequence varies relative to that of pseudogenes. Individual differences in Numt copy numbers have been established for humans (even among siblings)[35], shrimps[30], grasshoppers (D. Bensasson, PhD thesis, University of East Anglia, 1999), and for chloroplasts[19]. The relative proportion of mtDNA was also shown to differ among gall wasp populations (Antonis Rokas, pers. commun.).

The accidental co-amplification of Numts is not only influenced by the abundance of Numts in the species or individuals being studied, but also by which primers are used for the PCR reaction[8–10] and by which DNA extraction protocol[13] and tissue source are used[10,36]. PCR products obtained for aphids[27], grasshoppers[28], shrimps[30], elephants[36], gorillas (M.I. Jensen-Seaman, PhD thesis, Yale University, 2000) and gall wasps (Antonis Rokas, pers. commun.) contained many different Numt sequences in addition to the true mitochondrial sequence. Although, in all these cases, the mtDNA sequence was usually the single most common mtDNA-like region, in grasshoppers, shrimps, gorillas and elephants it sometimes represented <50% of the PCR product. If these PCR products were sequenced directly, the sequence obtained is unlikely to be mitochondrial, and would not necessarily even be ambiguous. Sequences obtained by direct sequencing of a Numt–mtDNA mixture will be ambiguous only if the amplified sequences differ from each other at a site in similar proportions.

It is possible to determine whether there are multiple mtDNA-like sequences by using restriction enzymes to cut PCR products at sites that are polymorphic among the individuals being studied, or at sites that give ambiguous sequences[26,37]. This approach allows the detection of Numts when they are multiple, but diverse, and the mtDNA proportion is too high for the visualization of single-copy Numts on

Box 3. Technical hints for Numt use

Generating a data set of paralogous Numt sequences

Paralogous Numt sequences can be generated by performing PCR and then cloning and sequencing mtDNA-like PCR products. Choosing a fast-evolving mitochondrial region for PCR should maximize the differences between mtDNA and its nuclear copies. Before PCR, the proportion of amplified mtDNA should be decreased by enriching for nuclear DNA (Ref. a), using tissue that is rich in nuclear DNA (e.g. sperm heads)[b], or by digesting total genomic DNA with a restriction enzyme that will cut mtDNA at a single unconserved site: if any nuclear mitochondrial pseudogenes (Numts) differ from the mtDNA at this site, they will predominate in the subsequent PCR product[c]. The pre-PCR digestion approach can also be used to focus on Numts of a particular evolutionary age. PCR products can be cloned and sequenced directly, screened with restriction enzymes, or screened by constant denaturant capillary electrophoresis (CDCE)[d] or single-stranded conformation polymorphism (SSCP)[e]. An alternative approach is to screen large-fragment genomic libraries by PCR (Ref. c).

Testing whether pseudogenes arose through independent transfers

Where Numts are nonfunctional, pairwise comparisons of Numts that reveal significant CODON POSITION BIAS (see Box Glossary) in the differences between them imply that the Numts are descended from different functional ancestors and are, therefore, the result of independent transfers to the nucleus[f,g]. The phylogenetic relationship among mtDNA and Numt sequences can also reveal whether Numts arose through independent transfers[h].

Distinguishing between substitutions arising in the nucleus or in mitochondria

For orthologous Numt sequences, the distinction between substitutions arising in the nucleus or in mitochondria is straightforward[i]. For paralogous sequences, substitutions on Numt branches can be identified by parsimony analysis or maximum likelihood, and noncoding Numt changes can be identified as 'unique' changes in an alignment of Numts and mtDNA (Refs h,j).

Dating transfers

Where divergence dates are known for mitochondrial lineages involved in the analysis, the date of transfer can be estimated using the method of Li *et al.*[k], as in Refs. l and m.

Box Glossary

Codon position bias: the biased accumulation of nucleotide substitutions with respect to the nucleotide position within a codon.

References

a Zhang, D-X. and Hewitt, G.M. (1996) Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. *Mol. Ecol.* 5, 295–300

b Zischler, H. *et al.* (1995) A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* 378, 489–492

c Yuan, J.D. *et al.* (1999) Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. *Cell Res.* 9, 281–290

d Li-Sucholeiki, X-C. *et al.* (1999) Applications of constant denaturant capillary electrophoresis / high fidelity polymerase chain reaction to human genetic analysis. *Electrophoresis* 20, 1224–1232

e Sunnucks, P. *et al.* (2000) SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Mol. Ecol.* 9, 1699–1710

f Bensasson, D. *et al.* (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Mol. Biol. Evol.* 17, 406–415

g Mundy, N.I. *et al.* (2000) Multiple nuclear insertions of mitochondrial cytochrome *b* sequences in Callitrichine primates. *Mol. Biol. Evol.* 17, 1075–1080

h Sunnucks, P. and Hales, D.F. (1996) Numerous transposed sequences of mitochondrial cytochrome oxidase I–II in aphids of the genus. *Sitobion* (Hemiptera: Aphididae). *Mol. Biol. Evol.* 13, 510–524

i Arctander, P. (1995) Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proc. R. Soc. London B Biol. Sci.* 262, 13–19

j Bensasson, D. *et al.* (2001) Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol. Biol. Evol.* 18, 246–253

k Li, W-H. *et al.* (1981) Pseudogenes as a paradigm of neutral evolution. *Nature* 292, 237–239

l Lopez, J.V. *et al.* (1994) *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* 39, 174–190

m DeWoody, J.A. *et al.* (1999) A translocated mitochondrial cytochrome *b* pseudogene in voles (Rodentia: *Microtus*). *J. Mol. Evol.* 48, 380–382

electrophoresis gels (e.g. in the case of grasshoppers[28]: Box 2). Extra mtDNA-like sequences in a PCR product could represent heteroplasmic mtDNA, duplications within the mitochondrial genome, or nonmitochondrial and nonnuclear EPISOMAL DNA (Refs 6,8,27). Some methods for establishing the location of extra mtDNA-like regions are discussed in Box 2 and more are given at http://www.pseudogene.net.

Numts as molecular fossils

Early studies of human Numt sequences revealed that these resembled 'ancestral' mitochondrial sequences[25], or 'molecular fossils'[21]. Because the human nuclear mutation rate is so much lower than that of mtDNA, Numts usually appear 'frozen' in comparison with their functional mitochondrial counterparts[25]. This is despite these nuclear sequences having lost their function when they were transferred to the nucleus and, therefore, not being selectively constrained[21]. Similar observations have been made for other mammals[9] and birds[34], although there are exceptions where mitochondrial genes that are under very strong selective constraints evolve more slowly than do their unconstrained nuclear paralogs[38].

Rates of nuclear and mitochondrial mutation do not differ as extremely in insects as they do in mammals and birds[39]. Depending on the level of selective constraint operating on the mitochondrial paralog of a nuclear sequence, an insect Numt could evolve faster, slower, or at a similar rate to mtDNA (D. Bensasson, PhD thesis, University of East Anglia, 1999). Nevertheless, the characteristics of Numt evolution are very different from those for mtDNA. Because metazoan Numts lose their function as a result of their transfer to the nucleus[20] and are, therefore, probably DEAD-ON-ARRIVAL PSEUDOGENES[40], they evolve in a very different way compared with functional mtDNA. Their substitution rates are equal with respect to codon position, stop codon generation, or original protein, RNA or DNA function[20]. Without purifying selection, they also readily accumulate

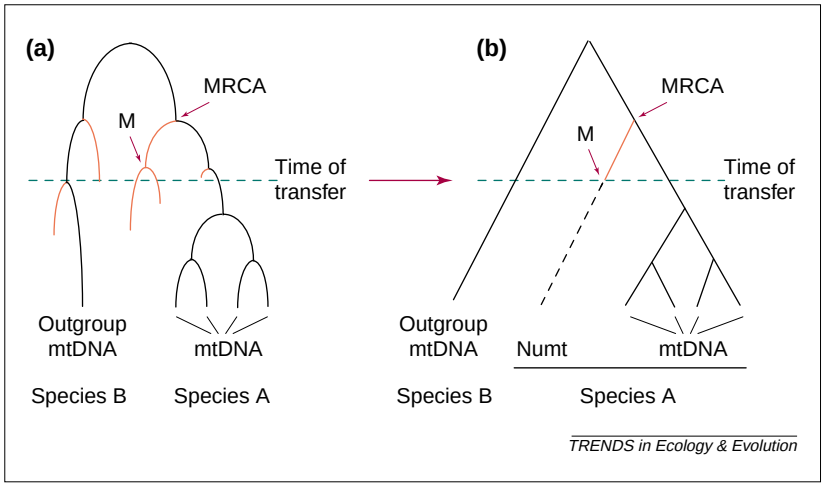

Fig. 1. Schematic representation of why nuclear mitochondrial pseudogene (Numt) branches can yield evidence of mitochondrial evolution. (a) Mitochondrial DNA (mtDNA) evolution showing extinct lineages (red lines) as well as extant lineages (black lines). Where mtDNA evolves fast, within-species differences (polymorphisms) can exist throughout mtDNA evolution. Because of an ancient polymorphism, the mtDNA sequence that migrated to the nucleus (M) is not identical to the ancestor of the modern-day mtDNA sequences of species A. Although the mtDNA lineage to which M belongs became extinct, in sexual species, the nuclear lineage to which M was transferred [shown in (b)] may still be inherited by present-day individuals of species A. (b) Cladogram of extant mtDNA sequences and the Numt sequence descended from M. The solid line represents sequence divergence with the characteristics of mitochondrial evolution where it is red, this occurred in one of the extinct mitochondrial lineages; the dashed line represents nuclear evolution. If M is not identical to a direct ancestor of one of the mtDNA sequences (as in this figure) changes that occurred in an unsampled mitochondrial ancestor may be represented on the Numt lineage. MRCA is the most recent common ancestor of mtDNA and Numts.

frameshift mutations[25]. Knowledge of this difference in selective constraint can be used to distinguish the 'ancient' mitochondrial sequence that migrated to the nucleus from the mutations that it accumulated once in the nucleus[27,41] (Box 3).

Loss of Numt function needs to be confirmed for plants and fungi because these taxa sometimes have functional nuclear copies of mitochondrial sequence. The mutation rates in plant nuclei are also much higher than those in mitochondria[18], and so mtDNA sequences would resemble the molecular fossils of their paralogous Numts.

Mutation also differs between nucleus and mitochondrion in ways other than rate, for example, in the pattern of nucleotide substitution[42], or in the size of insertions or deletions. Once in the nucleus, mtDNA-like sequences must be subject to nuclear mutation and its attendant biases. Mammalian[42] and insect[43] mutation show a higher TRANSITION:TRANSVERSION RATIO in mitochondria than in the nucleus; nuclear GC sites are frequently methylated in mammals, so GC → GT mutations are unusually common[44], but this is not the case for mtDNA (Ref. 44). Strand asymmetry effects have also been reported for mitochondrial evolution, although no such bias is expected from nuclear evolution[27].

Thorns in the evolutionary tree

The differences between nuclear and mitochondrial molecular evolution complicate phylogenetic analysis when it is applied to both nuclear and mitochondrial paralogs. PARSIMONY ANALYSIS, in particular, can underestimate branch-length differences between Numt and mtDNA lineages. Although mammalian, bird and plant nuclear rates of evolution can be an order of magnitude different from the mitochondrial rate, parsimony analysis makes no allowance for differing rates of evolution and will ascribe nuclear substitutions with equal probabilities to nuclear or mitochondrial branches. Where the mitochondrial mutation rate is greater than the nuclear mutation rate, mtDNA sequences will reach saturation sooner than will nuclear ones[45], but as they are usually implemented, most parsimony, distance or MAXIMUM LIKELIHOOD approaches will treat multiple hits equally in nuclear and mitochondrial lineages. These factors should be considered when drawing conclusions based on branch lengths.

Long Numt branches are occasionally encountered in hominoids, and are perhaps too long to be explained by the biases of phylogenetic analysis[9,45]. These branches could be the result of DNA damage sustained during or before nuclear integration[9]. Alternatively, new mtDNA-like arrivals could experience a mutation rate that is initially high[45]. For example, there might be an initial GC → GT mutation as a result of the methylation of GC sites[45], or, if a Numt contains one or more nucleotide deletion hotspots, these will mutate at a high rate until the deletion hotspots are deleted[41].

The extra length on a Numt branch could also be the result of an 'ancient polymorphism'[9] (Fig. 1). Sunnucks and Hales conducted a detailed analysis of the changes accumulated on the Numt branches of a parsimony tree for aphids, and their findings support such a hypothesis[27]. They discovered that many of the changes were similar to changes expected under mitochondrial (selectively constrained) evolution. Sunnucks and Hales also proposed that, in sexual species, the mtDNA sequence that migrated to the nucleus (M in Fig. 1) was not necessarily identical to the most recent common ancestor of Numt and modern-day mtDNA (MRCA in Fig. 1). In sexual species, Numts can occur in individuals with mtDNA lineages that differ from the lineage from which they are derived because they are not linked to the mtDNA lineage that generated them. The mtDNA sequence that was integrated might have belonged to a mitochondrial lineage that is now extinct or unsampled[8,27]. Because some mitochondrial evolution on metazoan Numt branches is probable, approaches that use knowledge of a loss of function to distinguish nuclear and mitochondrial changes[27,41] could be more reliable than are phylogenetic approaches.

The many uses of Numts

Even without knowledge of whether Numts are functional, and irrespective of relative rates of nuclear and mitochondrial evolution, Numts can be used as genetic markers. Where the nonfunctionality of Numts can be established, they can also be used in other ways.

Genetic markers

The lack of unifying features in Numt-flanking sequences[4,23] suggests that hotspots for Numt

Glossary

'Dead-on-arrival' pseudogene: a pseudogene that loses its function immediately upon its arrival at a locus, as opposed to a gene that evolved under relaxed selective constraints at its current locus before losing its function.
Codon usage bias: the bias in favor of some three-nucleotide sequences (codons) over others that code for the same amino acids.
Copy number: the number of copies of a DNA sequence per nuclear genome or cell.
Episomal DNA: DNA that can exist freely in the cell or integrated into a nuclear or organelle genome.
Heteroplasmic mtDNA: mtDNA that differs in sequence from one organelle to another within the same individual.
Maximum likelihood analysis: a method of phylogenetic reconstruction that estimates which possible tree topology is most likely to be the true phylogeny given the data and a specified model of DNA sequence evolution.
Numt: a copy of mitochondrial DNA (mtDNA) that is integrated into the nuclear genome; a nuclear mitochondrial DNA sequence.
Orthologous: genes derived from a common ancestral sequence by evolutionary divergence.
Paralogous: genes derived from a common ancestral sequence by duplication.
Parsimony analysis: a method of phylogenetic reconstruction that assumes the minimum number of evolutionary steps.
Pseudogene: a recognizable copy of a gene that has lost its function.
Transition:transversion ratio: the rates of transition:transversion nucleotide substitutions ratio.The four bases of DNA are classified into purines (adenine and guanine) and pyramidines (cytosine and thymine).Transversions are the types of nucleotide substitution involving changes between purines and pyramidines, whereas transitions are nucleotide changes within the nucleotide base classes (e.g. purine to purine).Transitions occur more often than do transversions, but the extent of this bias varies.
Transposable element: a DNA sequence capable of moving (transposing) from one location to another in a genome.

insertion are uncommon[23]. Therefore, the presence of a Numt at a particular locus in more than one taxon indicates their common ancestry[23]. The presence or absence of Numts at specific loci can be used to determine the phylogenetic branching order of different species[23,45], and as a population genetic marker for humans[46] and *A. thaliana*[47].

There are intraspecific differences in the copy numbers of nuclear copies of organellar DNA found in peas, barley, wheat, spinach and sugar beet[19]. Variability in copy number can be detected using Southern blots and could be useful for distinguishing commercial plant varieties[19]. Because humans[35] and grasshoppers (D. Bensasson, PhD thesis, University of East Anglia, 1999) show intraspecific Numt copy number differences, Numt copy numbers could also be used as population genetic markers in these species, through the application of Southern blot or PCR approaches, as done for TRANSPOSABLE ELEMENTS[47].

A specific Numt could also serve as a phylogenetic marker if there is sequence divergence among the taxa that share it. However, where nuclear mutation rates are slow (e.g. in mammals or birds) short Numt regions might show little or no intraspecific sequence variation[23,45].

Molecular roots and ancestors

When many paralogous Numt sequences are known, they can be used to trace the ancestral states of mtDNA at particular nucleotide sites, as has been done for mammals[26] and insects[27,41]. Independent knowledge of past ancestral states could help resolve branches of mitochondrial phylogenies where such ambiguities exist.

Numts have been used to root phylogenies for human populations[48] and birds[49]. This is particularly useful in humans, where the lack of a suitable outgroup is often limiting. The Numt chosen arose more recently than did the human divergence from chimps and, because of the low nuclear mutation rate, this Numt has changed little from the ancestral state of the variable mtDNA sequence, and is therefore ideal for rooting a fast-evolving mitochondrial phylogeny[48].

Relative rates of evolution

A 7.9-kb noncoding Numt in cats was used to compare the strength of purifying selection among mitochondrial regions[38]. Assuming that nuclear mutations accumulate at an equal rate along a Numt sequence, the regions where the mitochondrial sequence is most diverged from the Numt are the least selectively constrained[38]. Numts have also been used to compare the rate of nonfunctional nuclear sequence evolution to that of functional mtDNA (Refs 8,34).

The study of nuclear mutation

A promising application of Numt study is its use in the study of nuclear mutation. Patterns of mutation set the baseline on which other evolutionary factors operate, yet spontaneous mutation is poorly characterized in all but a few species. Spontaneous mutation is too slow, and the generation time of most metazoans too long, for the direct experimental study of spontaneous germline mutations. Patterns of spontaneous mutation can be inferred from substitutions accumulated in evolutionary time, but the choice of which types of sequence to analyze is critical. The use of functional sequences for the study of mutation is complicated by the selective effects of CODON USAGE BIAS on synonymous sites of coding sequence[50], and by possible functional constraints on introns (e.g. length constraints). Too little or too much sequence divergence among available taxa, and discontinuities between species, limit the utility of comparing ORTHOLOGOUS nonfunctional sequences. In addition, multiple unrelated pseudogene loci have been characterized for very few species that have been well studied at the molecular level (e.g. humans, mice[40] or other species for which genome projects are advanced).

By contrast, data sets consisting of many paralogous nuclear mitochondrial pseudogenes have been generated relatively quickly by cloning and sequencing PCR products without the need for prior molecular work in the species involved[27,28,30] (M.I. Jensen-Seaman, PhD thesis, Yale University, 2000). Using restriction enzymes, it is possible to enrich for pseudogenes of the desired evolutionary age. In metazoans, Numts are dead-on-arrival[40], so nuclear mutation can be distinguished from mitochondrial changes, thus enabling the study of nucleotide substitution, insertion and deletion[41]. The homology of paralogous Numt sequences allows tests for local sequence effects on mutation[41], and where Numts have been mapped, the effects of chromosomal position on mutation could also be tested for.

Conclusions and future research

The census that is emerging from GenBank and the published literature shows that there are large

Acknowledgements
We thank Vicki Friesen, Matthew Hare, Nick Harvey, Michael Jensen-Seaman, Werner E. Meyer, Kirstine Nielsen, Antonis Rokas, Paul Sunnucks and SteveTrewick for information on Numts in their study organisms, and to three reviewers for their comments on this article. Work was supported in part by a Royal Society–Fulbright Postdoctoral Fellowship to DB, and an NIH grant GM58423 to DLH.

differences among organisms in the numbers of mitochondrial pseudogenes that they harbor in their nuclear genomes. There is a growing literature on the conditions of Numt contamination and its avoidance in evolutionary studies and there are also many different ways in which these pseudogenes can now be used in evolutionary biology.

Improved characterization of the taxonomic distribution of Numts could allow their incidence in mtDNA studies to be better predicted. Furthermore, the taxonomic distribution of Numts could shed light on the plasticity of genomes in their long-term maintenance of foreign DNA that is not inherently selfish, repetitive or coding. Perhaps the taxonomic distribution could also show Numts to be broadly applicable as unconstrained sequences suitable for the study of spontaneous mutation. An understanding of spontaneous mutation would not only provide a useful baseline on which to model other evolutionary processes, but also taxonomic differences in mutational patterns would provide a basis for comparing molecular mechanisms that might result in such differences.

References

1 Du Buy, H.G. and Riley, F.L. (1967) Hybridization between the nuclear and kinetoplast DNA's of *Leishmania enriettii* and between nuclear and mitochondrial DNA's of mouse liver. *Proc. Natl. Acad. Sci. U. S. A.* 57, 790–797

2 Corral, M. *et al.* (1989) DNA-sequences homologous to mitochondrial genes in nuclei from normal rat-tissues and from rat hepatoma-cells. *Biochem. Biophys. Res. Commun.* 162, 258–264

3 Hadler, H.I. *et al.* (1998) Selected nuclear LINE elements with mitochondrial-DNA-like inserts are more plentiful and mobile in tumor than in normal tissue of mouse and rat. *J. Cell. Biochem.* 68, 100–109

4 Blanchard, J.L. and Schmidt, G.W. (1996) Mitochondrial DNA migration events in yeast and humans: Integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution pattern. *J. Mol. Evol.* 13, 537–548

5 DeWoody, J.A. *et al.* (1999) A translocated mitochondrial cytochrome *b* pseudogene in voles (Rodentia: *Microtus*). *J. Mol. Evol.* 48, 380–382

6 Mirol, P.M. *et al.* (2000) Multiple nuclear pseudogenes of mitochondrial cytochrome b in *Ctenomys* (Caviomorpha, Rodentia) with either great similarity to or high divergence from the true mitochondrial sequence. *Heredity* 84, 538–547

7 Lopez, J.V. *et al.* (1994) *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* 39, 174–190

8 Zhang, D-X. and Hewitt, G.M. (1996) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends Ecol. Evol.* 11, 247–251

9 Collura, R.V. and Stewart, C-B. (1995) Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature* 378, 485–489

10 Sorenson, M.D. and Quinn, T.W. (1998) Numts: A challenge for avian systematics and population biology. *The Auk* 115, 214–221

11 Zischler, H. *et al.* (1995) Detecting dinosaur DNA. *Science* 268, 1192–1193

12 van der Kuyl, A.C. *et al.* (1995) Nuclear counterparts of the cytoplasmic mitochondrial 12SrRNA gene: a problem of ancient DNA and molecular phylogenies. *J. Mol. Evol.* 40, 652–657

13 Wallace, D.C. *et al.* (1997) Ancient mtDNA sequences in the human nuclear genome: A potential source of errors in identifying pathogenic mutations. *Proc. Natl. Acad. Sci. U. S. A.* 94, 14900–14905

14 Doolittle, W.F. (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* 14, 307–311

15 Berg, O.G. and Kurland, C.G. (2000) Why mitochondrial genes are most often found in nuclei. *Mol. Biol. Evol.* 17, 951–961

16 Blanchard, J.L. and Lynch, M. (2000) Organellar genes: why do they end up in the nucleus? *Trends Genet.* 16, 315–320

17 Thorsness, P.E. and Weber, E.R. (1996) Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. *Int. Rev. Cytol.* 165, 207–234

18 Palmer, J.D. *et al.* (2000) Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6960–6966

19 Ayliffe, M.A. *et al.* (1998) Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. *Mol. Biol. Evol.* 15, 738–745

20 Gellissen, G. and Michaelis, G. (1987) Gene transfer: mitochondria to nucleus. *Ann. New York Acad. Sci.* 503, 391–401

21 Perna, N.T. and Kocher, T.D. (1996) Mitochondrial DNA: molecular fossils in the nucleus. *Curr. Biol.* 6, 128–129

22 Blanchard, J.L. and Schmidt, G.W. (1995) Pervasive migration of organellar DNA to the nucleus in plants. *J. Mol. Evol.* 41, 397–406

23 Zischler, H. (2000) Nuclear integrations of mitochondrial DNA in primates: Inference of associated mutational events. *Electrophoresis* 21, 531–536

24 Vaughan, H.E. *et al.* (1999) The localization of mitochondrial sequences to chromosomal DNA in orthopterans. *Genome* 42, 874–880

25 Fukuda, M. *et al.* (1985) Mitochondrial DNA-like sequences in the human nuclear genome. *J. Mol. Biol.* 186, 257–266

26 Hu, G. and Thilly, W.G. (1994) Evolutionary trail of the mitochondrial genome as based on human 16S rDNA pseudogenes. *Gene* 147, 197–204

27 Sunnucks, P. and Hales, D.F. (1996) Numerous transposed sequences of mitochondrial cytochrome oxidase I–II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). *Mol. Biol. Evol.* 13, 510–524

28 Bensasson, D. *et al.* (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Mol. Biol. Evol.* 17, 406–415

29 Mundy, N.I. *et al.* (2000) Multiple nuclear insertions of mitochondrial cytochrome *b* sequences in Callitrichine primates. *Mol. Biol. Evol.* 17, 1075–1080

30 Williams, S.T. and Knowlton, N. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. *Mol. Biol. Evol.* (in press)

31 Hu, G. and Thilly, W.G. (1995) Multi-copy nuclear pseudogenes of mitochondrial DNA reveal recent acute genetic changes in the human genome. *Curr. Genet.* 28, 410–414

32 Hancock, J.M. (1995) The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* 41, 1038–1047

33 Baker, R.J. *et al.* (1992) Reduced number of ribosomal sites in bats: Evidence for a mechanism to contain genome size. *J. Mammol.* 73, 847–858

34 Arctander, P. (1995) Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proc. R. Soc. London B Biol. Sci.* 262, 13–19

35 Yuan, J.D. *et al.* (1999) Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. *Cell Res.* 9, 281–290

36 Greenwood, A. and Paabo, S. (1999) Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. *Mol. Ecol.* 8, 133–137

37 Zhang, D-X. and Hewitt, G.M. (1996) Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. *Mol. Ecol.* 5, 295–300

38 Lopez, J.V. *et al.* (1997) Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. *Mol. Biol. Evol.* 14, 277–286

39 Sharp, P.M. and Li, W-H. (1989) On the rate of DNA sequence evolution in *Drosophila*. *J. Mol. Evol.* 28, 398–402

40 Graur, D. *et al.* (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.* 28, 279–285

41 Bensasson, D. *et al.* (2001) Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol. Biol. Evol.* 18, 246–253

42 Brown, W.M. *et al.* (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* 18, 225–239

43 Petrov, D.A. and Hartl, D.L. (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1475–1479

44 Bulmer, M. (1986) Neighbouring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.* 3, 322–329

45 Zischler, H. *et al.* (1998) A hominoid-specific nuclear insertion of the mitochondrial D-loop: implications for reconstructing ancestral mitochondrial sequences. *Mol. Biol. Evol.* 15, 463–469

46 Thomas, R. *et al.* (1996) Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. *Hum. Biol.* 68, 847–854

47 Ullrich, H. *et al.* (1997) Mitochondrial DNA variations and nuclear RFLPs reflect different genetic similarities among 23 *Arabidopsis thaliana* ecotypes. *Plant Mol. Biol.* 33, 37–45

48 Zischler, H. *et al.* (1995) A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* 378, 489–492

49 Quinn, T.W. (1992) The genetic legacy of Mother Goose: phylogeographic patterns of lesser snow goose *Chen caerulescens caerulescens* maternal lineages. *Mol. Ecol.* 1, 105–117

50 Moriyama, E.N. and Hartl, D.L. (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* 134, 847–858