J Mol Evol (2008) 66:362–367
DOI 10.1007/s00239-008-9088-2

Diagnosing Mitochondrial DNA Diversity: Applications of a Sentinel Gene Approach

Elizabeth L. Clare · Kevin C. R. Kerr · Taika E. von Königslöw · John J. Wilson · Paul D. N. Hebert

Received: 26 June 2007 / Accepted: 8 February 2008 / Published online: 27 March 2008
© Springer Science+Business Media, LLC 2008

Abstract Mitochondrial genomes show wide variation in their GC content. This study examines the correlations between mitochondrial genome-wide shifts in this feature and a fragment of the cytochrome *c* oxidase subunit I (COI) gene in animals, plants, and fungi. Because this approach utilizes COI as a sentinel, analyzing sequences from repositories such as GenBank and the Barcode of Life Data System (BOLD) can provide rapid insights into nucleotide usage. With this approach we probe nucleotide composition in a variety of taxonomic groups and establish the degree to which mitochondrial GC content varies among them. We then focus on two groups in particular, the classes Insecta and Aves, which possess the highest and lowest GC content, respectively. We establish that the sentinel approach provides strong indicators of mitochondrial GC content within divergent phyla (R values = 0.86–0.95, $p < 0.001$, in test cases) and provide evidence that selective pressures acting on GC content extend to non-coding regions of the plant and fungal mitochondrial genomes. We demonstrate that there is considerable variation in GC content of the mitochondrial genome within phyla and at each taxonomic level, leading to a substantial overlap zone in GC content between chordates and invertebrates. Our results provide a novel insight into the mitochondrial genome composition of animals, plants, and fungi and advocate this sentinel technique for the detection of rapid alterations in nucleotide usage as a measure of mitochondrial genome biodiversity.

Keywords Cytochrome *c* oxidase I · GC content · Nucleotide usage · Biodiversity · DNA barcoding

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9088-2) contains supplementary material, which is available to authorized users.

E. L. Clare (✉) · K. C. R. Kerr · T. E. von Königslöw ·
J. J. Wilson · P. D. N. Hebert
Department of Integrative Biology, University of Guelph,
Guelph, Ontario, Canada N1G 2W1
e-mail: eclare@uoguelph.ca

Introduction

Mitochondrial genomes show profound shifts in nucleotide usage among major taxonomic groups (Jermin et al. 1994) and deviations from stationarity can have serious impacts on phylogenetic analyses (Philippe and Laurent 1998; Foster and Hickey 1999; Mooers and Holmes 2000). Tests have been developed to detect shifts in composition enabling the exclusion of deviant taxa (e.g., Takezaki et al. 1995) and phylogenetic analyses have consequently been revised to compensate for these shifts (Mooers and Holmes 2000). However, there has also been a strong interest in understanding the environmental and biological factors that cause shifts in these parameters (reviewed by Rand 1994), particularly since GC content in the mitochondrial genome is uncoupled from such patterns in the nuclear genome (Mooers and Holmes 2000).

While the protein-coding and ribosomal genes located in the mitochondrial genome show individual variation in evolutionary rates (Lynch and Jarrell 1993), different regions of the mitochondrial genome may show congruent shifts in nucleotide composition. If this relationship is strong, the analysis of a single gene region could provide a sentinel assay for genome-wide shifts in GC content. This approach would facilitate broad-ranging surveys of life, aiding studies that seek either to probe the origins of variation in these parameters or to select lineages where variance in these factors would least complicate analyses.

The use of a fragment of the mitochondrial cytochrome *c* oxidase subunit 1 (COI) sequence as a species identification tool, namely, DNA barcoding (Hebert et al. 2003), has resulted in a recent flood of standardized COI sequences into public repositories such as BOLD (Ratnasingham and Hebert 2007). This led to the proposition of COI as a potential assay for GC shifts in mitochondrial DNA, with a high rate of success in animals (Min and Hickey 2007). Subsequent publications suggest that COI may also form the core of an identification tool for fungi (Seifert et al. 2007), which would likely result in the exponential growth of fungal COI sequences in public repositories. While DNA barcoding plants is more challenging and will doubtlessly require a multigene approach (Newmaster et al. 2007), COI sequences for plants are also actively being acquired. Using these sequences as an assay for GC content in plant and fungal mitochondrial genomes would greatly enhance our ability to monitor and explore variation in base composition.

In general, metazoan mitochondrial genomes are structurally uniform and contain little noncoding DNA. By contrast, plant and fungal mitochondrial genomes have a fundamentally different composition, as they include a large number of introns. While the relationship between GC content of COI and the mitochondrial genome in metazoans is not surprising, there is no a priori reason to expect this relationship to extend to noncoding sequences—such as introns—within the mitochondrial genomes of species from other kingdoms. This study examines the strength of the correlation between GC content of a segment of COI and GC content of the entire mitochondrial genome in a broad sample of metazoans and tests the hypothesis that a similar relationship extends to the fungal and plant kingdoms. In addition, we use these relationships to survey mitochondrial nucleotide diversity using the 5′ section of COI, which has been selected as the primary region for DNA barcoding (Hebert et al. 2003) and is available for a wide variety of taxa at the BOLD website (316,841 sequences from 33,726 species; October 24, 2007). Using available barcode sequences we measured the variation in mitochondrial GC content among major taxonomic groups, which is only possible using the sentinel gene approach due to the limited number of available complete mitochondrial genomes. We probe the extent of diversity within several phyla, classes, and orders as an initial application of this technique for assessing genomic biodiversity.

Methods

Sentinel Gene Analysis

We retrieved all 905 complete animal mitochondrial genomes available in GenBank (as of November 2006; see accession numbers in Supplementary Table S1). These genomes represent 16 animal phyla (Acanthocephala, 1; Annelida, 5; Arthropoda, 120; Brachiopoda, 3; Bryozoa, 1; Chordata, 667; Cnidaria, 23; Echinodermata, 16; Hemichordata, 2; Mollusca, 31; Nematoda, 17; Placozoa, 1; Platyhelminthes, 13; Porifera, 3; Priapulida, 1; Xenoturbellida, 1). We assembled two datasets from these mitochondrial genomes: one containing the entire mitochondrial genome for each individual and one limited to a 660-bp region used as a sentinel section. This section consists of the 5′ region of the cytochrome *c* oxidase subunit 1 gene (COI), which codes for a product from the beginning of membrane helix 1 to the end of external loop 3 (Lunt et al. 1996), amino acids 17–235 in the human COI mitochondrial sequence (NC_001807). These mitochondrial genomes vary considerably in size, from 11,423 bp (*Paraspadella gotoi* NC_006083) to 43,079 bp (*Trichoplax adhaerens* NC_008151). Hence, the putative sentinel section (the DNA barcode) varies between 2% and 6% of the complete mitochondrial genome in the 905 species included in our analysis. We examined nucleotide composition by comparing the GC content in the DNA barcode with the GC content in the complete mitochondrial genome.

We also examined the mitochondrial genomes of 40 fungi and 19 plants available from GenBank (see accession numbers in Supplementary Table S1). Due to the large number of introns in COI from fungi and plants, we limited our choice of COI sentinel section to exons from the 5′ end of COI. We chose a gene segment that was homologous to that described above for animals (sequence alignment available from the authors upon request). The GC content from these sections was compared to the GC content of the entire mitochondrial genome.

Mitochondrial GC Diversity

Our second line of investigation examined the diversity in GC content across multiple taxonomic groups using this sentinel gene approach. COI sequences were obtained from the Barcode of Life Data System (BOLD) (Ratnasingham and Hebert 2007). We calculated GC content for members of 6 phyla (Chordata, Echinodermata, Cnidaria, Annelida, Arthropoda, and Nematoda), 8 classes (Aves, Actinopterygii, Amphibia, Mammalia, Anthozoa, Malacostraca, Secernentea, and Insecta) and 12 orders (Anura, Charadriiformes, Passeriformes, Decapoda, Amphipoda, Isopoda, Perciformes, Chiroptera, Coleoptera, Diptera, Hymenoptera, and Lepidoptera). We sought to maximize diversity in our selection of taxa, but limited analysis to groups with sequences from at least 100 species. We selected single representatives for each species based on longest sequence length and trimmed sequences to the barcode region. Incomplete (short) sequences were used when no other sequence was available. We calculated GC content for each sequence as a proportion of bases present (typically 660 bp). All sequences are available in the published projects section on BOLD (www.barcodinglife.org).

Taxonomic Hierarchy and GC Diversity

Our final line of inquiry examined the variability in GC content within two classes, the Insecta and the Aves, which represent the lowest and the highest average GC content respectively. We obtained COI barcode sequences from BOLD (as above) and determined the GC content at each level of the taxonomic hierarchy, selecting groups that would maximize the available species. For the birds we used the order Passeriformes, families Emberizidae and Parulidae, and genera *Pipilo*, *Spizella*, *Dendroica*, and *Vermivora*. For the insects, we used the order Hymenoptera, families Apidae and Formicidae, and genera *Bombus, Euglossa, Crematogaster*, and *Myrmecocystus*.

We also extracted similar data from the whole mitochondrial genomes available from GenBank and compared these measurements to the predicted GC values of the COI sentinel gene approach.

Results

Sentinel Gene Analysis

Among animals, the GC content of whole mitochondrial genomes varied from 13% to 54%, while the barcode sequence varied from 22% to 53%. However, the GC content of the barcode was highly correlated with that of the full mitochondrial genome ($r = 0.95$, $p < 0.001$) (Fig. 1a). Both approaches identify the bee, *Meliponia bicolor* (NC_004529.1), as having the lowest GC content, with the whole mitochondrial genome indicating 13.28% GC and the barcode region estimated at 21.76% GC content. Overall rankings by GC content between the two approaches were not identical. Whole mitochondrial genome analysis indicates that the fish, *Chirocentrus dorab*, has the highest GC content, at 53.22%. The sentinel gene approach estimates *C. dorab* at 52.8% GC, ranking it as the fourth highest, while the noodle fish, *Salanx ariakensis*, is predicted to be the highest, at 53.4% GC (see Supplementary Table S2 for comparative ranking).

Results for COI sections based on fungal and plant mitochondrial genomes were similar to the 660-bp section targeted for animals. Fungal mitochondrial genomes were correlated with the COI region at $r = 0.86$, $p < 0.001$

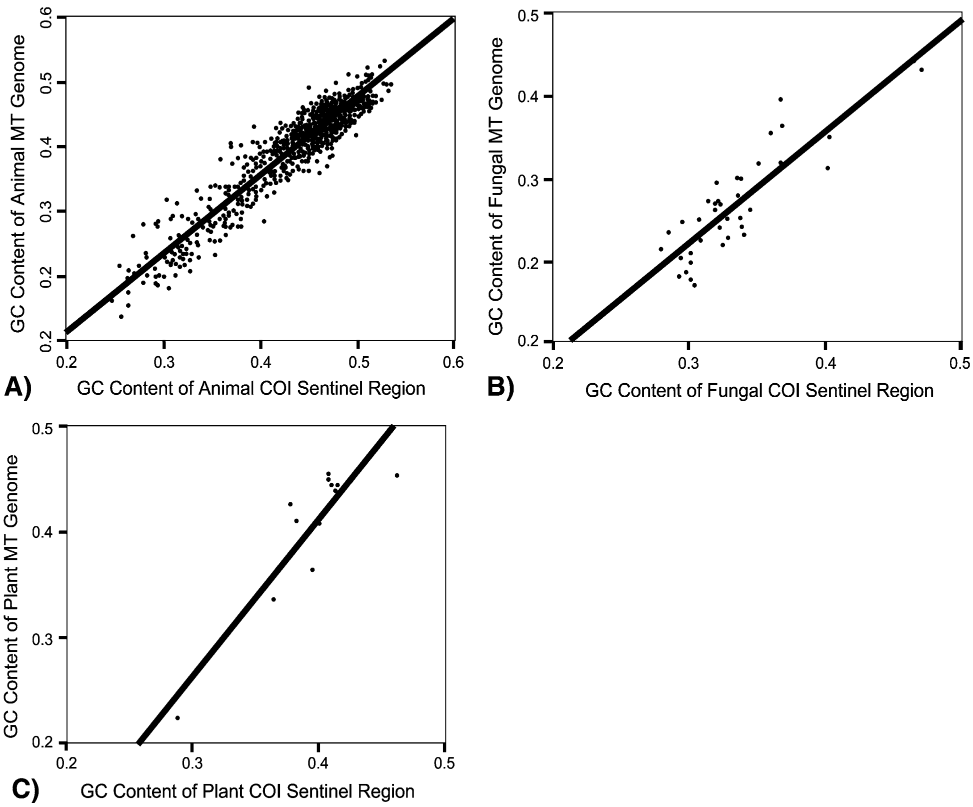
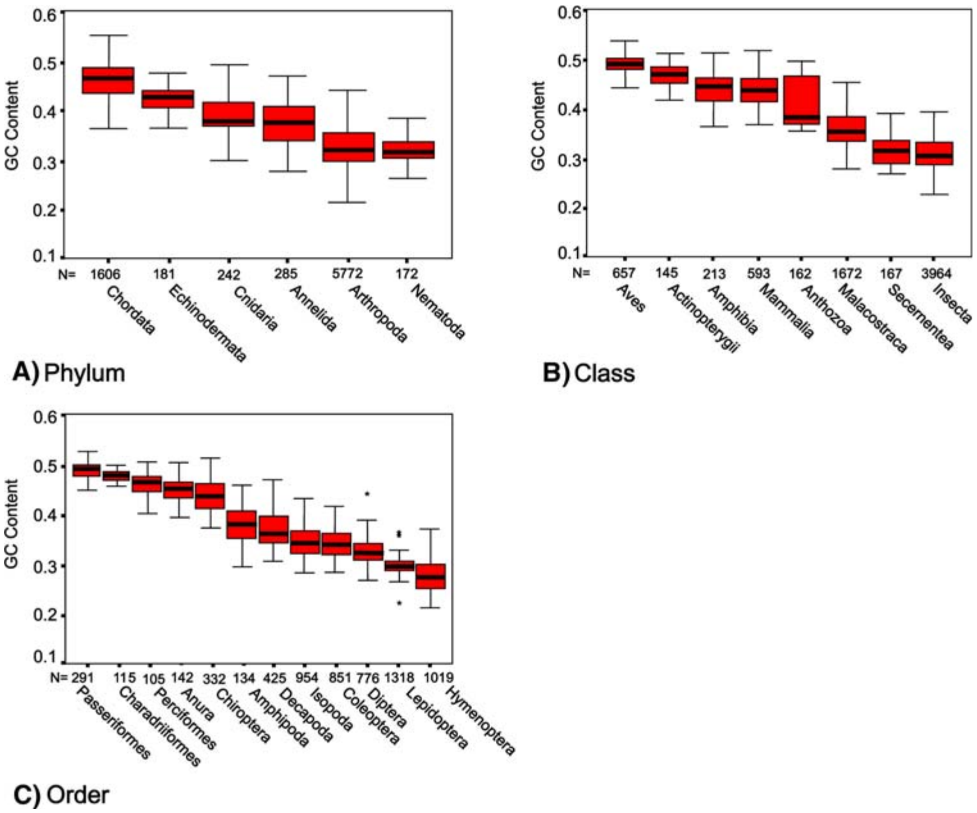


Fig. 1 Correlations between the GC content in a small region of COI and that in the entire mitochondrial genome. Strong correlations exists between the GC content in the DNA barcode "sentinel" and the GC content of the complete mitochondrial genome in (**A**) animals ($r = 0.95$, $p < 0.001$), (**B**) fungi ($r = 0.86$, $p < 0.001$), and (**C**) plants ($r = 0.90$, $p < 0.001$)

Fig. 2 A boxplot of GC content in the mitochondrial genomes of a variety of animal phyla. Probing with the sentinel sequence approach demonstrates that GC content is higher in chordates than invertebrates ($p < 0.001$) within (**A**) phyla, (**B**) class, and (**C**) order. All groups are highly variable and an overlap zone exists encompassing the most GC-rich and GC-poor phyla and classes



(Fig. 1b), and plant mitochondrial genomes were correlated with the COI region at $r = 0.90$, $p < 0.001$ (Fig. 1c).

Mitochondrial GC Diversity

Nucleotide composition varied greatly across taxa and at all levels of the taxonomic hierarchy. Vertebrates have a significantly higher average GC content than invertebrates ($p < 0.001$) (Fig. 2), but there is considerable overlap between these groups. All invertebrate orders examined contain species with GC contents falling within the range of the vertebrate groups' (Fig. 2c). After removing the arthropods which have an extremely large sample size, estimates of mean GC content were not correlated with sample size ($r = -0.349$, p = 0.094), while estimates for the range of GC content were significantly influenced by sample size ($r = 0.623$, $p = 0.001$).

Taxonomic Hierarchy and GC Diversity

The variation in GC content declined at each taxonomic level from class to genus (Fig. 3), though there are still large shifts (up to 8%) in GC content at the generic level. Although the range in GC content was consistently higher in insects than in birds, it is likely that this is a direct effect of the sample size (which is consistently larger in the Insecta), making the detection of biological patterns impossible.

A comparison of GC estimates based on the extensive sampling available with the sentinel gene approach versus estimates from the whole genome database of GenBank further demonstrates the effect of samples size. Estimates of mean and range of GC content among the classes Insecta and Aves based on whole mitochondrial genomes (Insecta mean = 24%, range = 13–35%; Aves mean = 45%, range = 41–50%) were smaller than those obtained using the barcode region (Insecta mean = 31%, range = 22–45%; Aves mean = 49%, range = 43–55%). Similar results were seen for other levels of the taxonomic hierarchy but further comparisons such as at the genus level are limited due to the lack of whole mitochondrial genomes available in GenBank.

Discussion

This study reaffirms that a strong correlation exists between the GC content of the 5′ region of COI (the DNA barcode) and that of the full mitochondrial genomes in all metazoans, but establishes that this relationship is not limited to animal life but extends to plants and fungi. The tight correlation between GC content of this sentinel and the entire mitochondrial genome of fungi and plants provides a significant insight into the nature of selective pressures impacting nucleotide usage, as these forces apparently act on both

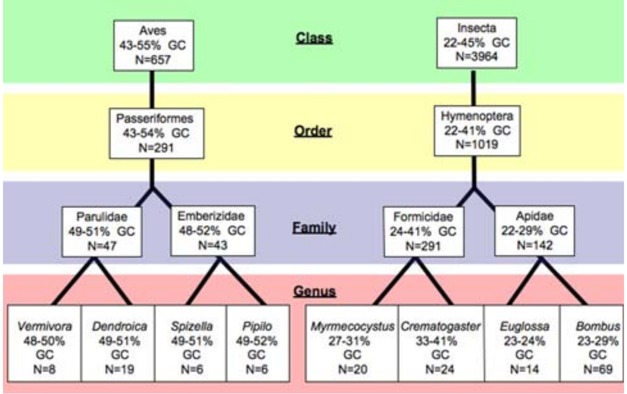


Fig. 3 Range in GC content of the mitochondrial genome at different levels in the taxonomic hierarchy. Estimates of the range in GC content decrease from class to genus but large shifts in GC content (up to 8%) are detected even at the generic level. Insects show greater variation in GC content than birds, but range estimates are influenced by sample size

coding and noncoding DNA. The number of complete mitochondrial genomes in GenBank is constantly increasing, but the cost of whole mitochondrial genome acquisition will generally be higher than that for sequencing individual genes. Hence, a database of short sequences will encompass greater diversity than whole mitochondrial genome databases. With no clear relationship between mitochondrial and nuclear GC ratios (Powell et al. 1986; Mooers and Holmes 2000), these sentinel sequences cannot be used as a probe for the base composition of the nucleus. However, the rapid accumulation of sequence data on standardized regions of the COI gene as part of the Barcode of Life Initiative (Ratnasingham and Hebert 2007) provides an unprecedented volume of data from which to estimate nucleotide usage in the mitochondrial genomes.

It is clear from our comparison between insects and birds, and between GenBank and the sentinel sequence, that samples size has a strong influence on estimates of the range of GC content. The current scientific focus reflected in GenBank is on sequencing a small number of genomes, but from a wide variety of taxa rather than increased sampling of sister taxa and duplicate members of the same species. Despite this attempt to include diverse groups, the ranges of GC content among insects and birds were smaller when measured using GenBank's mitochondrial genomes than when acquired by the sentinel gene approach. A sentinel region will act as a proxy measure by serving as an indicator of shifts in nucleotide usage in taxa without sequenced mitochondrial genomes (for example, between congeneric taxa); a research avenue that is not otherwise possible given the prohibitive cost of genome sequencing.

Our second line of inquiry involved using the GC content in the barcode region to probe the GC content in diverse lineages as a measure of nucelotide biodiversity. Despite the commonly held view that invertebrate mitochondria are AT rich, while chordate mitochondria are GC rich (Hebert et al. 2003; Mooers and Holmes 2000), our observations reveal considerable variation in the range of GC values within each group. This variation leads to a zone of overlap covering even the most GC-rich and GC-poor phyla and classes. There is also considerable overlap between orders. The wide range of GC content in some groups is quite intriguing, but we must treat these values cautiously as most sequences originated in GenBank and contaminations and misidentifications in public repositories have been well documented (Harris 2003) and would create outliers that cannot be confirmed.

Among the class Insecta, base composition was highly variable and ranged from genomes that are AT rich (*Meliponia bicolour*) to genomes with only slightly skewed AT/GC ratios (*Coleobthrus germeauxi*). The detection of up to 8% variation in GC content among species in the genus *Crematogaster* (Hymenoptera) suggests rapid alterations in nucleotide usage. However, investigating GC content variation between closely related species is beyond the scope of this paper and would require a detailed phylogeny and subsequent phylogenetic independent contrasts (PICs) to reliably reveal patterns in variation and the associated variables. PIC was not implemented in our analysis because we have limited our investigation to the range of a single variable over distantly related taxa.

There is a variety of explanations for genome shifts in nucleotide composition, which can be categorized into theories of mutational biases and natural selection (Mooers and Holmes 2000). Martin (1995) identified a bias in directional nucleotide substitution that favored AT substitutions in animals. Other correlations have been found that implicate environmental temperature (Bernardi and Bernardi 1986), heavy-strand/light-strand specific composition biases (Asakawa et al. 1991), relative availability of nucleotides during synthesis (Jukes and Bhushan 1986), the energetic costs of GC vs AT synthesis in species with short generation times, and suggestions that amino acid content correlates with these AT/GC shifts (Foster et al. 1997). Many of these studies are severely limited by the number of mitochondrial genomes available for analysis. For example, Foster et al. (1997) compared a single chicken mitochondrial genome to a single honeybee mitochondrial genome to establish correspondence between amino acid usage and nucleotide shifts. Mooers and Holmes (2000) argue that mutational biases are the most probable explanation for shifts in mitochondrial GC content and that natural selection is more likely to work on fine-scale modification rather than genome-wide selection for base pair composition. The sentinel gene approach will facilitate further investigation of these hypotheses and others by providing higher-density sampling.

It has been suggested that pressures influencing GC content act on noncoding regions (such as the D-loop) as well as coding areas (Jukes and Bhushan 1986). Our results support this conclusion by establishing the strong correlation between GC content in the protein-coding COI region and that in mitochondrial genomes, which vary in size, gene content, gene order, and the presence of introns (particularly in plants and fungi). This also suggests that factors responsible for shifts in nucleotide composition are independent of genome organization, gene content, and codon specific factors but are instead acting on the molecule as a whole. There is increasing evidence that within the nuclear genome, selection on GC content works at a fine scale—gene by gene—rather than on a genome-wide basis (reviewed by Mooers and Holmes 2000). Because the mitochondrial genome is inherited as a single molecule, mutational biases or selective events would likely act on it as a whole, providing a basis for the success of the sentinel gene approach.

Foster et al. (1997) also examined the correspondence of GC content in the 13 protein coding genes of the chicken and the honeybee mitochondria and commented that "the GC content of the genes individually was similar to the content of the genomes from which they came." This precursor to the idea of a sentinel sequence was not pursued until recently (Min and Hickey 2007), but it does make clear that the COI region is not especially effective in predicting nucleotide composition. Any subset of the mitochondrial genome could be used as a sentinel sequence. However, the growing interest in the development of identification systems centered on COI for animals (Hebert et al. 2003) and its potential use in other kingdoms such as fungi (Seifert et al. 2007) promises that the generation of COI sequence libraries for these groups will be comprehensive and will include multiple replicates and dense sampling schemes currently unavailable through whole mitochondrial genome databases. These developments will provide an unprecedented opportunity to capitalize on these sequence libraries for other research avenues in molecular evolution.

Acknowledgments We gratefully acknowledge the assistance of our colleagues at the Biodiversity Institute of Ontario. This work was supported by Genome Canada through the Ontario Genomics Institute, by a Natural Sciences and Engineering Research Council of Canada grant to P.D.N.H., and by a NSERC postgraduate scholarship to E.L.C. We thank M. A. Smith, R. M. Floyd, and two anonymous reviewers for their comments, which strengthened the content of this article.

References

Asakawa S, Kumazawa Y, Araki T, Himeno H, Miura K, Watanabe K (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. J Mol Evol 32:511–520

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48: 284–290

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 44:282–288

Harris DJ (2003) Can you bank on GenBank? Trends Ecol Evol 18:317–319

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond [Biol] 270:313–322

Jermin LS, Graur D, Lowe RM, Crozier RH (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. J Mol Evol 39:160–173

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44

Lunt DH, Zhang DX, Szymura JM, Hewitt GM (1996) The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. Insect Mol Biol 5:153–165

Lynch M, Jarrell PE (1993) A method for calibrating molecular clocks and its application to animal mitochondrial DNA. Genetics 135:1197–2208

Martin AP (1995) Metabolic rate and directional nucleotide substitution in animal mitochindrial DNA. Mol Biol Evol 12:1124–1131

Min XJ, Hickey DA (2007) DNA barcodes provide a quick preview of mitochondrial genome composition. PLoS ONE 2(3):e325

Mooers AO, Holmes EC (2000) The evolution of base composition and phylogenetic inference. Trends Ecol Evol 15:365–369

Newmaster SJ, Fazekas AJ, Steeves RAD, Janovec J (2007) Testing candidate plant barcode regions in the Myristicaceae. Mol Ecol Notes (in press). doi:10.1111/j.1471-8286.2007.02002.x

Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genetics Dev 8:616–623

Powell JR, Caccone A, Amato GD, Yoon C (1986) Rates of nucleotide substitution in *Drosophila* mitochondrial DNA and nuclear DNA are similar. Proc Natl Acad Sci USA 83:9090–9093

Rand DM (1994) Thermal habit, metabolic rate, and the evolution of mitochondrial DNA. Trends Ecol Evol 9:125–131

Ratnasingham S, Hebert PDN (2007) BOLD: the Barcode of Life Data system. Mol Ecol Notes 7:355–364

Seifert KA, Samson RA, deWaard JR, Houbraken J, Levesque CA, Moncalvo J, Louis-Seize G, Hebert PDN (2007) Prospects for fungus identification using COI DNA barcodes, with *Penicillium* as a test case. Proc Natl Acad Sci USA 104:3901–3906

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823–833