Choosing and Using Introns in Molecular Phylogenetics

Simon Creer

School of Biological Sciences, University of Wales, Bangor, Gwynedd, LL57 2UW, United Kingdom.

Abstract: Introns are now commonly used in molecular phylogenetics in an attempt to recover gene trees that are concordant with species trees, but there are a range of genomic, logistical and analytical considerations that are infrequently discussed in empirical studies that utilize intron data. This review outlines expedient approaches for locus selection, overcoming paralogy problems, recombination detection methods and the identification and incorporation of LVHs in molecular systematics. A range of parsimony and Bayesian analytical approaches are also described in order to highlight the methods that can currently be employed to align sequences and treat indels in subsequent analyses. By covering the main points associated with the generation and analysis of intron data, this review aims to provide a comprehensive introduction to using introns (or any non-coding nuclear data partition) in contemporary phylogenetics.

Keywords: introns, EPIC primers, recombination, length variant heterozygote (LVH), indel, alignment.

Introduction

Non-coding introns are now routinely used in molecular systematics as independent markers (Oakley and Phillips, 1999; van Oppen et al. 2000), or in concert with other gene partitions in an attempt to recover gene trees that are concordant with species trees in plant (Borsch et al. 2003; Guo and Ge, 2004; Oh and Potter, 2005; Shaw et al. 2005; Zhang et al. 2006), fungal (Freeman et al. 2002; Froslev et al. 2005; Cortinas et al. 2006) and animal (Palumbi and Baker, 1994; Prychitko and Moore, 1997; Flynn and Nedbal, 1998; Pitra et al. 2000; Johnson and Clayton, 2000; Rockman et al. 2001; Weibel and Moore, 2002; Rowe and Honeycutt, 2002; Birks and Edwards, 2002; Ericson et al. 2002; Braband et al. 2002; Beltrán et al. 2002; Creer et al. 2003, 2006) molecular phylogenetics. Despite the exponential rise of intron-based molecular genetic studies (Friesen et al. 1997, 1999; Friesen, 2000; Zhang and Hewitt, 2003), a number of genomic, molecular biological, and analytical issues need to be considered during the progression from project conception to data analysis. Factors such as locus selection, paralogy, the occurrence of length variant heterozygotes (LVHs), alignment, insertion/deletion (indel) or gap treatment, and the detection of recombination can all influence how data are generated and analyzed, but such issues are infrequently, or incompletely discussed in empirical studies. Accordingly, this review aims to provide a reference point related to the above issues in order to facilitate an easy introduction to working with introns, or other non-coding data partitions, in molecular phylogenetics.

What are Introns?

Introns are untranslated gene regions of genomic DNA that are spliced out in the formation of mature RNA molecules and can be conveniently divided into groups, based on their splicing mechanism. Group I and II introns are characterized by different self-splicing mechanisms and are found in some bacterial and organellar genomes (Kelchner, 2000, 2002; Hausner et al. 2006), and group I introns are also found in ribosomal RNAs (rRNAs) of protists and fungal nuclei. Conversely, spliceosomal introns (the most common insertions in eukaryotic nuclear pre-mRNA genes) require a complex of five RNAs and hundreds of proteins, known as the spliceosome, to facilitate intron excision in the formation of mature messenger RNA molecules (Bonen and Vogel, 2001; Roy and Gilbert, 2006). In eukaryotic introns, two types of spliceosome are recognized. The common U2-type splices GT-AG introns, so called because the introns start with 5' GT and end with 3' AG dinucleotides, and possess a characteristic pyridimine rich region that precedes the 3' splice site (Stryer, 1988; Senapathy et al. 1990; Friesen, 2000). The second U12-type, splices the vary rare AT-AC introns, that have a number of dinucleotides at the 3' end (Belshaw

Correspondence: Simon Creer, School of Biological Sciences, University of Wales, Bangor, Gwynedd, LL57 2UW, United Kingdom. Tel: + (0)1248 382302; Email: s.creer@bangor.ac.uk

and Bensasson, 2006). Finally, transfer RNA (tRNA) introns are found in eukaryotic nuclei and in Archaea, but are spliced enzymatically using a completely different mechanism to spliceosomal introns (Haugen et al. 2005; Roy and Gilbert, 2006).

Introns have been shown to affect eukaryotic gene expression in a number of ways, including initial transcription, editing, polyadenylation and nuclear export of the pre-mRNA, translation and decay of the mRNA product, in addition to exon shuffling, duplication and alternative splicing of discrete genes (Gasch et al. 1989; Alder et al. 1992; Kirby et al. 1995; Leicht et al. 1995; Prychitko and Moore, 2003; Le Hir et al. 2003). Thus, although introns have clear functional significance, empirical data have shown that they can be considered as neutral markers that possess a number of traits that are desirable for molecular phylogenetics (Friesen, 1997, 2000). Compared to coding regions, the non-coding nature of introns predicts the acquisition of a large number of independent parsimony informative characters from most sites equally, associated with less homoplasy and lower transition: transversion ratios (Slade et al. 1994; Prychitko and Moore, 2000, 2003). Nevertheless, depending on the splicing mechanisms that are involved in the excision process, some classes of introns may possess a mosaic-like structure involving conserved and secondary structure elements, and/or mutational hotspots, that appear to evolve under complex and different evolutionary constraints (e.g. compensating base pair changes) (Borsch et al. 2003; Quandt and Stech, 2005; Quandt et al. 2004). Moreover, it must also be acknowledged that diploid spliceosomal intron alleles have an average effective population size four times that of mtDNA and empirical "ball park" estimations in animals have shown that introns mutate at approximately one quarter the rate of animal mtDNA (Prychitko and Moore, 1997, 2000; Creer et al. 2003). Consequently, animal mtDNA haplotypes are expected to coalesce (i.e. become monophyletic) and track recent speciation events more rapidly than intron loci (Moore, 1995; Wiens, 2000).

Locus Selection and Primer Design

The amplification of introns for interspecific studies is usually facilitated by designing primers that anneal to conserved regions within exons to either side of the target intron (e.g. 50 base pairs (bp) upstream and downstream of the 5' and 3' intron splice sites for the forward and reverse primers respectively). This exon-primed, intron-crossing (EPIC) primer design strategy was introduced over ten years ago (Lessa, 1992; Slade et al. 1993), but widely applicable primers such as those that have contributed to the meteoric success of various animal mtDNA genes (Kocher et al. 1989) have yet to be realised (Zhang and Hewitt, 2003; Hughes et al. 2006).

Empirical testing

Although truly universally applicable primers do not exist, a number of putatively broad, and taxon-specific EPIC primers are now available for potential use in animals (Slade et al. 1993; Palumbi and Baker, 1994; Friesen et al. 1997, 1999; Prychitko and Moore, 1997; Hassan et al. 2002; Jarman et al. 2002; Touriya et al. 2003; Sota and Vogler, 2003; Aitken et al. 2004) and vascular and non-vascular plants, (Shaw et al. 2005; Ishikawa et al. 2002). Thus, one approach that can be utilized to locate a suitable selection of markers, is to assay the performance of numerous primers from existing studies (Creer et al. 2005), or primers that have worked well in related taxa. Using a diverse array of PCR optimization strategies on representatives from a desired phylogenetic range may not result in complete success, but the approach is predicted to identify a subsection of primers that will result in successful amplifications throughout the target genetic group. If amplifications are lacking in a number of taxa for a particular locus, the sequences derived from the successful PCR reactions can, and should where possible, serve as templates for taxon-specific EPIC (Slade et al. 1993) or even intron (I) PIC primer designs (i.e. where the primers are designed within conserved regions of the actual intron).

Data mining

In August 2005, the amount of sequence data available in the GenBank repository of the National Center for Biotechnology Information website (NCBI – http://www.ncbi.nlm.nih.gov/) exceeded 100 gigabases, i.e. over 100 billion base pairs. A significant proportion of these data will correspond to model organisms and genome sequencing projects, but there is also currently an abundance of annotated whole genomic and mRNA cDNA sequences that can be mined for tailor-made EPIC priming sites in target taxa spanning appropriate

phylogenetic ranges. Bioinformatic tools such as Spidey (http://www.ncbi.nlm.nih.gov/Tools/) now make the task easier by aligning one or more mRNA sequences to single genomic sequences. Primers can then be designed using readily available programs such as Primer 3 (Rosen and Skaletsky, 1998) for single sequences, or PriFi (Fredslund et al. 2005) for multiple sequence alignments.

Dealing with Paralogy

The above approaches primarily rely on the conservation of assumed single copy nuclear exon coding regions for primer design, however, if gene duplication has occurred, the use of degenerate primers can result in the inadvertent amplification of paralogous loci (Tank and Sang, 2001; Archambault and Bruneau, 2004; Pfeil et al. 2004; Meimberg et al. 2006). Obviously, phylogenetic error can arise if paralogs (genes related by duplication) are mistakenly interpreted to be orthologs (genes derived from a single ancestral gene in the last common ancestor of the compared species) when inferring phylogenetic relationships (Sanderson and Shaffer, 2002; Koonin, 2005). There is some hope that paralogous genes can be detected by differences in molecular architecture, e.g. in size, structure, codon usage or base composition (Cotton, 2005), or via the interpretation of tree topologies that are grossly incongruent with widely held perceptions. Alternatively, attempts to overcome paralogy can be achieved via computational algorithms, or by alternative molecular biological strategies. Maddison (1997), Page and Charleston (1997) and Slowinski et al. (1997) described a procedure, termed "gene tree parsimony" (Page, 1998) that employs heuristic searches for species trees that minimizes the weighted sum of gene duplications plus losses (in addition to deep coalescences and lateral transfers) necessary to fit gene family trees to species trees (Slowinski et al. 1997). In addition to visualizing the fit between gene and known species phylogenies, the program GeneTree (Page, 1998) can be used to infer species phylogenies from duplicated genes, whereby the optimal species tree is that in which the gene trees can be embedded with the least cost. Furthermore, maximum likelihood and Bayesian methods that allow probabilistic model incorporation are now being developed for reconciling gene and species trees (Arvestad et al. 2003). Such directions show particular promise in overcoming the conceptual limitations involved in reconciliation approaches reliant solely on the principles of parsimony (Cotton, 2005).

On the other hand, the generation of cDNA libraries and designing one primer in the 3' untranslated region (3' UTR) and the other in coding regions has been suggested as a molecular-based method to overcome the amplification of non-orthologous genes. Such an approach aims to exploit the fact that divergence between paralogous loci is likely to occur more rapidly in the 3'UTR compared to adjacent exons. However, the specificity gained by using the exon-3'UTR approach is likely to come at a cost in that priming site substitutions are likely to result in PCR failure as genetic distance increases from the species from which the cDNA was sequenced (Whittall et al. 2006).

It is therefore likely that a combination of experimental, computational and bioinformatic approaches may yield a number of orthologous and potentially phylogenetically useful genes. Once the loci have been chosen, it is important that the introns should then be sequenced across an appropriate broad taxonomic range to ensure that the markers yield sufficient phylogenetic signal (Shaw et al. 2005; Hughes et al. 2006). Testing a large number of primers may seem labor-intensive, but it is certainly a cost-effective strategy. A short investment of time and money at the outset of a project vastly outweighs the disadvantages associated with working with less than optimal markers throughout a genetically disparate range.

Addressing recombination

A fundamental concept in molecular phylogenetics is that a single phylogeny can be reconstructed from the sequences under study (Posada and Crandall, 2001; Wiuf et al. 2001; Husmeier and McGuire, 2003). Nevertheless, nuclear genes can frequently experience recombination events that can create mosaic genes (Maynard-Smith, 1992) where different regions possess diverse phylogenetic histories (Posada and Crandall, 2001).

One way to avoid the potentially confounding problem of recombination is to use nuclear genes that experience very low rates of recombination, but a more likely solution is to detect recombination events and incorporate the data into models of molecular evolution during analysis, thus facilitating the fuller exploitation of nuclear markers (Zhang and Hewitt, 2003). A number of

methods based upon similarity, distance, phylogeny, compatibility/congruence, distribution of substitutions (Posada and Crandall, 2001; Posada, 2002; Posada et al. 2002) and Bayesian approaches (Husmeier and McGuire, 2003) have been developed to detect recombination (a full list of resources for detecting recombination can be found at http://www.umber.embnet.org/~robertson/recombination/index.shtml). Moreover, software such as Recombination Detection Package 2 (RDP2 available at http://darwin.uvigo.es/rdp/rdp.html) combines ten different published methods in an attempt to identify recombinant sequences and recombination breakpoints (Martin et al. 2005). Programs such as RDP2 utilize a range of approaches as empirical tests on simulated data have shown that no single method is likely to be optimal in detecting recombination under all conditions (Posada and Crandall, 2001; Posada, 2002).

Although recombination is an integral part of meiosis, it appears paradoxical that only a limited number of empirical phylogenetic studies have attempted to detect recombination within nuclear gene datasets (Miadlikowska et al. 2003; Printzen et al. 2003; Jarvinen et al. 2004; Devos et al. 2005; Poke et al. 2006). The lack of instances of recombination detection may reflect the comparative infancy of the field of nuclear, as opposed to mtDNA gene-based molecular systematics (but see Piganeau et al. (2004) and Tsaousis et al. (2005) for recent animal mtDNA recombination surveys). Alternatively, relatively derived taxonomic lineages are precluded from recombination detection analyses as most methods have been shown to fail if sequence divergence is less than five percent (Posada and Crandall, 2001; Devos et al. 2005). Still, for datasets reflecting deeper phylogenetic levels, the above recent bioinformatic innovations suggest there are no reasons why recombination detection cannot form an integral part of sequence analyses during phylogeny reconstruction. Depending on the particular scenario, the taxa involved in the recombination event may then either be excluded from the analysis, or included, and the recombination information integrated into phylogeny reconstruction or interpretation.

Detection, separation and incorporation of LVHs

Diploid non-coding nuclear genes will either be heterozygotic or homozygotic, but it is infrequently reported that heterozygotic introns often differ in length. Heterozygotic introns of the same length are easily recognised in direct sequencing chromatograms as dual peaks of approximately equal intensity occupying the same base position, and can be detected by eye, or using software such as Polyphred (Nickerson et al. 1997). The latter ambiguous sites can be scored as Ns for phylogenetic purposes, but direct sequencing a LVH will result in the apparent corruption of the sequence reaction due to the superimposition of two separate sequence chromatograms occupying the same frame (Mallarino et al. 2005). If a LVH is suspected, many solutions exist to separate the two alleles including using the 'allele-dropout-effect', haplotype separation by single strand conformation polymorphism (SSCP) and denaturing gradient gel electrophoresis (DGGE) (Zhang and Hewitt, 2003). For most laboratories dealing with phylogenetic analyses, cloning can provide an easy solution for separating the two alleles, although separating haplotypes with denaturing high performance liquid chromatography (DHPLC) may be the future solution towards resolving LVHs that prove to be difficult to clone, or for high-throughput purposes (Underhill et al. 1996, 1997; Zhang and Hewitt, 2003).

The ocurrence of LVHs was first highlighted by Palumbi and Baker (1994), but a large proportion of intron-based molecular phylogenetic studies do not mention LVHs, and others have attempted but have failed to detect intra-individual length variation (van Oppen et al. 2000; Prychitko and Moore, 2000; Birks and Edwards, 2002). Recently however, studies have detected and incorporated LVHs into phylogenetic frameworks (Beltrán et al. 2002; Sota and Vogler, 2003; Pons et al. 2004; Creer et al. 2006) suggesting that the phenomenon is common within intron loci and should be considered as a matter of course in studies using diploid introns as phylogenetic markers. The analysis of LVHs within a single locus analysis is simple, but multiple-partition total evidence (Kluge, 1989; Nixon and Carpenter, 1996) approaches provide additional challenges. Sota and Vogler, (2003) recently employed an intuitive approach whereby LVHs are simultaneously incorporated as independent terminals in data matrices by duplication of homozygotic loci alongside heterozygotic loci. Therefore, if an individual was homozygotic for locus A and heterozygotic for locus B, the heterozygotic taxon would be represented by two identical

allele A sequences (AA) and the two length variant B loci sequences (Bb), and vice versa. If, on the other hand, the individual was heterozygotic for both loci, all four combinations of the LVHs are included in the analyses. This approach may be tractable with a limited number of partitions, but the number of independent terminals (represented by 2^n, where n is the number of loci) may prove cumbersome with combinations of multiple heterozygotic loci. An emerging solution to this problem may lie in the Phylogeny of Organisms from Allelic Data (POFAD) algorithm that converts distance matrices of alleles to organismal distance matrices from one or more genes (Joly and Bruneau, 2006), but further independent testing will be needed to confirm or refute its systematic utility.

Alignment and indel treatment approaches

Non-coding nuclear gene partitions frequently experience diverse indel events that create considerable alignment problems. In order to achieve positional homology, multiple DNA sequences are therefore either aligned "by-eye" or by using a range of algorithms implemented by computer programs such as ClustalX (Thompson et al. 1997), T-Coffee (Notredame et al. 2000), DIALIGN (Morgenstern, 1999), or MUSCLE, that is recommended for large numbers of sequences (Edgar, 2004). Proponents of algorithm-based alignments criticize the subjectivity and lack of repeatability of "by-eye" alignments (Giribet and Wheeler, 1999), although some empirical studies have shown that manual alignments are not significantly worse than computer assisted alignments (Sanchis et al. 2001; Belshaw and Quicke, 2002). On the other hand, the diversity and length of indels experienced in intron partitions frequently cause computer-based alignments to have significant proportions of misaligned taxa or gene regions. The often used and optimal strategy may therefore be to utilize appropriate programs and amend any obviously misaligned regions by hand (Freudenstein and Chase, 2001; Sanchis et al. 2001; Kawakita et al. 2003; Creer et al. 2006). The manual intervention in the latter scenario does introduce non-objectivity, but the complete removal of subjectivity in complex alignments is likely to remain a utopian goal (Lutzoni et al. 2000).

Following alignment, the next step is to decide what to do with the indel data. The classical strategy is to treat alignment gaps as missing data (Kumar et al. 2004; Swofford et al. 1996; Swofford, 1998) and such an approach is attractive if the indel events are minor. Indels however, may represent Hennigian biological events (Archambault and Bruneau, 2004) and often represent a substantial percentage of sequence data in non-coding data partitions. Disregarding gap data may therefore represent the loss of a considerable proportion of phylogenetic signal (Freudenstein and Chase, 2001). In attempt to remedy this situation, a number of approaches have emerged to incorporate gap characters with substitutional data in phylogenetics.

Perhaps the simplest approach is to code gaps as fifth character states (Swofford, 1998). Alternatively, gaps with different start and/or end positions can be replaced (i.e. treat as missing data) with a coded binary matrix (based on presence/absence) that is concatenated and analysed with the normal DNA data. Simmons and Ochoterena, (2000) formalized this approach with the advent of "simple coding" and the software GapCoder (Young and Healy, 2003) facilitates the construction of the binary matrix. A newer approach, called Modified Complex Indel Coding (MCIC) has additionally been developed that aims to maximize the phylogenetic information retained from unambiguously aligned sequences that was previously not utilized by simple coding (Müller, 2006). Modified Complex Indel Coding can be performed using "IndelCoder" within the program SeqState (Müller, 2005).

Introns frequently experience combinations of indels and substitutions that result in areas that cannot be aligned equivocally. Homopolymers, pyridimine rich (both independent, and inclusive of the 3' splice site (Senapathy et al. 1990)) and A/C rich regions all appear to be commonplace. In order to overcome the homology problems associated with areas of ambiguous alignment, the software INtegrating Ambiguously Aligned SEquences INAASE (Lutzoni et al. 2000) expedites the replacement of these regions with multistate coded characters (step matrices), that are analysed alongside the DNA base characters. Thus, by replacing the area of ambiguous alignment with a step matrix, multistate coding attempts to incorporate unequivocally aligned regions of DNA without violating positional homology (Lutzoni et al. 2000). Nevertheless, Müller (2005) points out that INAASE effectively ignores some characters in delimited

multistate regions and does not address the issue of incorporating information from length mutational events in regions for which positional homology has been established. Alternatively, indel coding methods use the information from length mutational events as well as the information from substitutional data in the same region.

Simple coding, MCIC, multistate coding and coding gaps as a fifth character state all rely on posterior coding of indels that are derived from a multiple sequence alignment (Wheeler, 2001). Alternatives lie in fixed-state optimization (Wheeler, 1999) and direct optimization (Wheeler, 1996) that can be executed using the software POY (Wheeler and Gladstein, 2000; for debate, refer to Simmons, 2004 and Kluge and Grant, 2006). Both approaches differ from multiple sequence alignment based methods as the sequence data is not preprocessed, but proceed directly to cladogram optimization (Wheeler, 2001). Fixed-state optimization treats each sequence as a character state, and generates a matrix of transformation costs that relate different states to one another, in a similar fashion to multistate coding (Wheeler, 1999). Alternatively, direct optimization incorporates indel events as additional transformations during the optimization step in tree evaluation instead of trying to reconcile sequence lengths by adding gaps as additional states. Substitutions and indel events are simultaneously minimized and unique alignments are generated for each historical hypothesis (Aagesen, 2005; Hormiga et al. 2003; Wheeler, 2001).

All of the above solutions addressing alignment and gap treatment strategies have been approached using parsimony. Very recently however, a number of methods have emerged that aim to simultaneously infer multiple alignment and construct phylogenetic hypotheses using Bayesian approaches. Lunter et al. (2005) and Fleissner et al. (2005) have used the TKF1 (Thorne et al. 1991) and the TKF2 (Thorne et al. 1992) models respectively. The TKF1 model treats indels as independent single base pair events, whereas the TKF2 model permits non-nested and non-overlapping indels of several base pairs in length. Alternatively, Redelings and Suchard, (2005) have adopted a novel model and algorithm that allows multiple base pair, overlapping and nested indels, accommodating all homology structures.

Thus, indel treatment strategies can be conveniently split into static vs. dynamic and parsimony vs. model-based approaches, but it is also pertinent to acknowledge that any method that that treats indel characters as independent data points (e.g. fifth state, POY and Lunter et al.'s (2005) Bayesian approach) disregards any knowledge concerning the biological mechanisms underlying indel evolution. Indel mutation processes are unlikely to arise from the same mutational mechanisms as substitutional data (Pons and Vogler, 2006). Smaller gaps (1–30 b.p.) are hypothesized to result from slipped-strand mispairing while it is thought that larger gaps (>30 b.p.) are caused by unequal crossing over or due to transposition (Giribet and Wheeler, 1999; Freudenstein and Chase, 2001; Li, 1997). Therefore, 1—*n* b.p. gaps are often unlikely to represent 1—*n* independent mutation events and methods that treat gaps as independent characters (ie base pair by base pair) can significantly overweight larger indels, compared to smaller, equivalent indel events and can therefore generate innaccurate trees (Lutzoni et al. 2000; Freudenstein and Chase, 2001; Creer et al. 2006). According to this logic, simple coding, MCIC and the Bayesian approaches of Fleissner et al. (2005) and Redelings and Suchard (2005), differ from all the other approaches by treating indels, regardless of length, as independent events.

Finally, following a diverse combination of alignment and indel treatment approaches, a decision must be made regarding which phylogenetic hypothesis most accurately represents evolutionary history (Giribet and Wheeler, 1999; Sanchis et al. 2001). It is widely acknowledged that congruence among datasets provides an accurate estimate of phylogeny (Miyamoto and Fitch, 1995; Giribet and Wheeler, 1999; Wheeler, 2001; Sanchis et al. 2001). Thus, if no independent congruence measures are available, all hypotheses can be presented and shared topologies discussed regarding data-dependent consensus (Arnedo et al. 2004). If however, independent but incompatible datasets are available (e.g. morphology, other genes omitting significant indels, or extremely large datasets) taxonomic congruence can be used as a measure favoring treatments that maximize phylogenetic consensus (Giribet and Wheeler, 1999; Cognato and Vogler, 2001; Giribet, 2001; Belshaw and Quicke, 2002). Alternatively, if multiple data partitions are compatible, character congruence, often measured by incongruence length difference (ILD, Mickevich and Farris (1981)), can be used to objectively assess treatment associated homoplasy through

simultaneous analyses (Giribet and Wheeler, 1999; Wheeler, 2001).

In summary, working with introns, or other non-coding nuclear partitions is not as straightforward as working with organellar data (Sang, 2002). Bioinformatic approaches, or assaying a large number of EPIC primers may be required to locate the most appropriate markers for non-model organisms. PCRs may need more optimization due to the degeneracy of the primers involved, and/or the single copy nature of nuclear targets, and direct sequencing may not be possible if LVHs are discovered. Once the data has been generated from orthologous loci, recombination checks can now be routinely performed and a range of parsimony and model-based analytical innovations are also available regarding alignment and the treatment of indel data. Given the growing reliance of the molecular systematic community on non-coding DNA, a key goal that remains is to identify which analytical methods most accurately recover phylogenetic history (Wheeler, 1996; Giribet and Wheeler, 1999; Simmons and Ochoterena, 2000; Lutzoni et al. 2000; Fleissner et al. 2005; Creer et al. 2006). It is therefore important that further testing of all the methods is performed on simulated and empirical datasets to establish which strategies are optimal when using introns or non-coding nuclear partitions as phylogenetic markers.

Acknowledgements

Simon Creer derived his experience on working with introns whilst working on research grants focusing on Asian Pitviper systematics awarded by The Wellcome Trust (057257/Z/99/Z; 060384/Z/00/Z) and The Leverhulme Trust (F174/O) to Roger Thorpe and Anita Malhotra, University of Wales, Bangor. Further thanks go to Anita Malhotra, Roger Thorpe, Dave Lunt and two anonymous reviewers for valuable feedback that improved an earlier version of this manuscript.

References

Archambault, A. and Bruneau, A. 2004. Phylogenetic utility of the LEAFY/FLORICAULA gene in the Caesalpinioideae (Leguminosae): Gene duplication and a novel insertion. *Sys. Bot.*, 29:609–626.

Arvestad, L., Berglund, A-C., Lagergren, J. et al. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19:i7–i15.

Aagesen, L. 2005. Direct optimization, affine gap cost, and node stability. *Mol. Phyl. Evol*., 36:641–653.

Aitken, N., Smith, S., Schwarz, C. et al. 2004. Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Mol. Ecol*., 13:1423–1431.

Alder, H., Yoshinouchi, M., Prystowsky, M.B. et al. 1992. A conserved region in intron 1 negatively regulates the expression of the PCNA gene. *Nucl. Acids. Res*., 20:1769–1775.

Arnedo, M.A., Coddington, J., Agnarsson, I. et al. 2004. From a comb to a tree: phylogenetic relationships of the comb-footed spiders (Araneae, Theridiidae) inferred from nuclear and mitochondrial genes. *Mol. Phyl. Evol.*, 31:225–245.

Belshaw, R. and Quicke, D.L.J. 2002. Robustness of ancestral state estimates: evolution of life history strategy in ichneumonoid parasitoids. *Syst. Biol*., 51:450–477.

Belshaw, R. and Bensasson, D. 2006. The rise and fall of introns. Heredity, 96:208–213.

Beltrán, M., Jiggins, C.D., Bull, V. et al. 2002. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. *Mol. Biol. Evol*., 19:2176–2190.

Birks, S.M. and Edwards, S.V. 2002. A phylogeny of the megapodes (Aves: Megapodiidae) based on nuclear and mitochondrial DNA sequences. *Mol. Phyl. Evol.*, 23:408–421.

Bonen, L. and Vogel, J. 2001. The ins and outs of group II introns. *Trends. Genet*,. 17:322–331.

Borsch, T., Hilu, K.W., Quandt, D. et al. 2003. Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. *J. Evol. Biol*., 16:558–576.

Braband, A., Richter, S., Hiesel, R. et al. 2002. Phylogenetic relationships within the *Phyllopoda* (Crustacea, Branchiopoda) based on mitochondrial and nuclear markers. *Mol. Phyl. Evol*., 25:229–244.

Cognato, A.I. and Vogler, A.P. 2001. Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Syst. Biol*., 50:758–780.

Cortinas, M.N., Crous, P.W., Wingfield, B.D. et al. 2006. Multi-gene phylogenies and phenotypic characters distinguish two species within the *Colletogloeopsis zuluensis* complex associated with Eucalyptus stem cankers. *Stud. Mycol*., 55:133–146.

Cotton, J.A. 2005. Analytical methods for detecting paralogy in molecular datasets. *Meth Enzym*., 395:700–724.

Creer, S., Malhotra, A. and Thorpe, R.S. 2003. Assessing the phylogenetic utility of four mitochondrial genes and a nuclear intron in the Asian pit viper genus *Trimeresurus*: Separate, simultaneous, and conditional data combination analyses. *Mol. Biol. Evol*., 20:1240–1251.

Creer, S., Malhotra, A., Thorpe, R.S. et al. 2005. Targeting Optimal Introns for Phylogenetic Analyses in Non-model Taxa: Experimental Results in Asian Pitvipers. *Cladistics*, 21:390–395.

Creer, S., Pook, C.E., Malhotra, A. et al. 2006. Optimal intron analyses in the *Trimeresurus* radiation of Asian pitvipers. *Syst. Biol*., 55:57–72.

Devos, N., Oh, S., Raspe, O. et al. 2005. Nuclear ribosomal DNA sequence variation and evolution of spotted marsh-orchids (*Dactylorhiza maculata* group). *Mol. Phyl. Evol*., 36:568–580.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids. Res*., 32:1792–1797.

Ericson, P.G.P., Christidis, L., Irestedt, M. et al. 2002. Systematic affinities of the lyrebirds (Passeriformes: Menura), with a novel classification of the major groups of passerine birds. *Mol. Phyl. Evol*., 25:53–62.

Fleissner, R., Metzler, D. and Haeseler, A.V. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol*., 54:548–561.

Flynn, J.J. and Nedbal, M.A. 1998. Phylogeny of the Carnivora (Mammalia): Congruence vs. incompatability among multiple data sets. *Mol. Phyl. Evol*., 9:414–426.

Fredslund, J., Schauser, L., Madsen, L.H. et al. 2005. PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucl. Acid. Res*., 33:W516–W520.

Freeman, A.B., Duong, K.K., Shi, T.L. et al. 2002. Isolates of *Microbotryum violaceum* from North American host species are phylogenetically distinct from their European host-derived counterparts. *Mol. Phyl. Evol*., 23:158–170.

Freudenstein, J.V. and Chase, M.W. 2001. Analysis of mitochondrial *nad1* b-c intron sequences in Orchidaceae: utility and coding of length-change characters. *Syst. Bot*., 26:643–657.

Friesen, V. 2000. Introns. In: Baker AJ, ed. Molecular Methods in Ecology. Blackwell Science Ltd. p 274–294.

Friesen, V.L., Congdon, B.C., Kidd, M.G. et al. 1999. Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. *Mol. Ecol*., 8:2141–2152.

Friesen, V.L., Congdon, B.C., Walsh, H.E. et al. 1997. Intron variation in marbled murrelets detected using analyses of single-stranded conformational polymorphisms. *Mol. Ecol*., 6:1047–1058.

Froslev, T.G., Matheny, P.B. and Hibbett, D.S. 2005. Lower level relationships in the mushroom genus *Cortinarius* (Basidiomycota, Agaricales): A comparison of RPB1, RPB2, and ITS phylogenies. *Mol. Phyl. Evol*., 37:602–618.

Gasch, A., Hinz, U. and Renkawitz-Pohl, R. 1989. Intron and upstream sequences regulate expression of the Drosophila β3-tubulin gene in the visceral and somatic musculature, respectively. *Proc. Natl. Acad. Sci. USA*., 86:3215–3218.

Giribet, G. 2001. Exploring the behavior of POY, a program for direct optimisation of molecular data. *Cladistics*, 17:S60–S70.

Giribet, G. and Wheeler, W.C. 1999. On gaps. *Mol. Phyl. Evol*., 13:132–143.

Guo, Y.L. and Ge, S. 2004. The utility of mitochondrial nad1 intron in phylogenetic study of *Oryzeae* with reference to the systematic position of *Porteresia*. *Acta. Phytotaxon. Sin*., 42:333–344.

Hassan, M., Lemaire, C., Fauvelot, C. et al. 2002. Seventeen new exon-primed intron-crossing polymerase chain reaction amplifiable introns in fish. *Mol. Ecol. Notes*, 2:334–340.

Haugen, P., Simon, D.M. and Bhattacharya, D. 2005. The natural history of group I introns. *Trends Genet*., 21:111–119.

Hausner, G., Olson, R., Simon, D. et al. 2006. Origin and evolution of the chloroplast trnK (matK) intron: A model for evolution of group II intron RNA structures, *Mol. Biol. Evol*., 23:280–391.

Hormiga, G., Arnedo, M. and Gillespie, R.G. 2003. Speciation on a conveyor belt: sequential colonisation of the Hawaiian islands by *Orsonwelles* spiders (Araneae, Linyphiidae). *Syst. Biol*., 52:70–88.

Hughes, C.E., Eastwood, R.J. and Bailey, C.D. 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Phil. Trans. Roy. Soc. B*., 361:211–225.

Husmeier, D. and McGuire, G. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol. Biol. Evol*., 20:315–337.

Ishikawa, H., Watano, Y., Kano, K. et al. 2002. Development of primer sets for PCR amplification of the PgiC gene in ferns. *J. Plant. Res*., 115:65–70.

Jarman, S.N., Ward, R.D. and Elliot, N.G. 2002. Oligonucleotide primers for PCR amplification of coelomate introns. *Mar. Biotechnol*., 4:347–355.

Jarvinen, P., Palme, A., Morales, L. et al. 2004. Phylogenetic relationships of *Betula* species (Betulaceae) based on nuclear ADH and chloroplast matK sequences. *Am. J. Bot*., 91:1834–1845.

Johnson, K.P. and Clayton, D.H. 2000. Nuclear and mitochondrial genes contain similar phylogenetic signal for pigeons and doves (Aves: Columbiformes). *Mol. Phyl. Evol*., 14:141–151.

Joly, S. and Bruneau, A. 2006. Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from *Rosa* in North America. *Syst. Biol*., 55:623–636.

Kawakita, A., Sota, T., Ascher, J.S. et al. 2003. Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (*Bombus*). *Mol. Biol. Evol*., 20:87–92.

Kelchner, S.A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals. Missouri. Bot. Gard*., 87:482–498.

Kelchner, S.A. 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *Am. J. Bot*., 89:1651–1669.

Kirby, D.A., Muse, S.V. and Stephan, W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc. Natl. Acad. Sci*., USA, 92:9047–9051.

Kluge, A.G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool*., 38:7–25.

Kluge, A.G. and Grant, T. 2006. From conviction to anti-superfluity: old and new justifications of parsimony in phylogenetic inference. *Cladistics*, 22:276–288.

Kocher, T.D., Thomas, W.K., Meyer, A. et al. 1989. Dynamics of mitochondrial DNA evolution in animals : amplification and sequencing with conserved primers. Proc. *Natl. Acad. Sci*., USA, 86:6196–6200.

Koonin, E.V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*., 39:309–338.

Kumar, S., Tamura, K. and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinf*., 5:150–163.

Le Hir, H., Nott, A. and Moore, M.J. 2003. How introns influence and enhance eukaryotic gene expression. *Trends. Biochem. Sci*., 28:215–220.

Leicht, B.G., Muse, S.V., Hanczyc, M. et al. 1995. Constraints on intron evolution in the gene encoding the myosin alkali light chain in *Drosophila Genetics*, 139:299–308.

Lessa, E.P. 1992. Rapid surveying of DNA sequence variation in natural populations. *Mol. Biol. Evol*., 9:323–330.

Li, W.H. 1997. Molecular evolution. Sinauer, Sunderland, MA.

Lunter, G., Miklos. I., Drummond, A. et al. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6:83.

Lutzoni, F., Wagner, P., Reeb, V. et al. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Syst. Biol*., 49:628–651.

Maddison, W.P. 1997. Gene trees in species trees. *Sys. Biol*., 46:523–536.

Mallarino, R., Bermingham, E., Willmott, K.R. et al. 2005. Molecular systematics of the butterfly genus *Ithomia* (Lepidoptera: Ithominae): a composite phylogenetic hypothesis based on seven genes. *Mol. Phyl. Evol*., 34:625–644.

Martin, D., Williamson, C. and Posada, D. 2005. RDP2: recombination detection and analysis from sequence alignments. *Bionformatics*. 21:260–262.

Maynard-Smith, J. 1992. Analyzing the mosaic structure of genes. *J. Mol. Evol*., 34:126–129.

Meimberg, H., Thalhammer, S., Brachmann, A. et al. 2006. Comparative analysis of a translocated copy of the trnK intron in carnivorous family Nepenthanceae. *Mol. Phyl. Evol*., 39:478–490.

Miadlikowska, J., Lutzoni, F., Goward, T. et al. 2003. New approach to an old problem: Incorporating signal from gap-rich regions of ITS and rDNA large subunit into phylogenetic analyses to resolve the *Peltigera canina* species complex. *Mycologica*, 95:1181–1203.

Mickevich, M.F. and Farris, J.S. 1981. The implications of congruence in *Menidia*. *Syst. Zool*., 30:351–370.

Miyamoto, M.M. and Fitch, W.M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol*., 44:64–76.

Moore, W.S. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, 49:718–726.

Morgenstern, B. 1999. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218.

Müller, K. 2006. Incorporating information from length-mutational events into phylogenetic analysis. *Mol. Phyl. Evol*., 38:667–676.

Müller, K. 2005. SeqState - primer design and sequence statistics for phylogenetic DNA data sets. *Appl. Bioinformatics*, 4:65–69.

Nickerson, D.A., Tobe, V.O. and Taylor, S.L. 1997. Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucl. Acids. Res*., 25:2745–2751.

Nixon, K.C. and Carpenter, J.M. 1996. On simultaneous analysis. *Cladistics*, 12:221–241.

Notredame, C., Higgins, D. and Heringa, J. 2000. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol*., 302:205–217.

Oakley, T.H. and Phillips, R.B. 1999. Phylogeny of Salmonine fishes based on growth hormone introns: Atlantic (*Salmo*) and Pacific (*Oncorhynchus*) salmon are not sister taxa. *Mol. Phyl. Evol*., 11:381–393.

Oh, S.H. and Potter, D. 2005. Molecular phylogenetic systematics and biogeography of tribe neillieae (Rosaceae) using DNA sequences of cpDNA, rDNA, and LEAFY. *Am. J. Bot*., 92:179–192.

van Oppen. M.J.H., Willis, B.L., van Vugt, H.W.J.A. et al. 2000. Examination of species boundaries in the *Acropora cervicornis* group (Scleractina, Cnidaria) using nuclear DNA sequence analyses. *Mol. Ecol*., 9:1363–1373.

Page, R.D.M. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14:819–820.

Page, R.D.M. and Charleston, M.A. 1997. Reconciled trees and incongruent gene and species trees. In: Mirkin B McMorris FR Roberts and FS Rzhetsky, eds. Mathematical hierarchies in biology. *Am. Math. Soc., Providence*, RI. p. 57–71.

Palumbi, S.R. and Baker, C.S. 1994. Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Mol. Biol. Evol*., 11:426–435.

Pfeil, B.E., Brubaker, C.L., Craven, L.A. et al. 2004. Paralogy and orthology in the Malvaceae *rpb2* gene family: investigation of gene duplication in *Hibiscus*. *Mol. Biol. Evol*., 21:1428–1437.

Piganeau, G., Gardner, M. and Eyre-Walker, A. 2004. A broad survey of recombination in animal mitochondria. *Mol. Biol. Evol*., 21:2319–2325.

Pitra, C., Lieckfeldt, D. and Alonso, J.C. 2000. Population subdivision in Europe's great bustard inferred from mitochondrial and nuclear DNA sequence variation. *Mol. Ecol*., 9:1165–1170.

Poke, F.S., Martin, D.P., Steane, D.A. et al. 2006. The impact of intragenic recombination on phylogenetic reconstruction at the sectional level in Eucalyptus when using a single copy nuclear gene (cinnamoyl CoA reductase). *Mol. Phyl. Evol*., 39:160–170.

Pons, J., Barraclough, T.G., Theodorides, K. et al. 2004. Using exon and intron sequences of the gene Mp20 to resolve basal relationships in *Cicindela* (Coleoptera: Cicindelidae). *Syst. Biol*., 53:554–570.

Pons, J. and Vogler, A.P. 2006. Size, frequency, and phylogenetic signal of multiple-residue indels in sequence alignment of introns. *Cladistics*, 22:144–156.

Posada, D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol*., 19:708–717.

Posada, D. and Crandall, K.A. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci*., USA, 98:13757–13762.

Posada, D., Crandall, K.A. and Holmes, E.C. 2002. Recombination in evolutionary genomics. *Ann. Rev. Genet*., 36:75–97.

Printzen, C., Ekman, S. and Tonsberg, T. 2003. Phylogeography of *Cavernularia hultenii*: evidence of slow genetic drift in a widely disjunct lichen. *Mol. Ecol*., 12:1473–1486.

Prychitko, T.M. and Moore, W.S. 1997. The utility of DNA sequences of an intron from the β-fibrinogen gene in phylogenetic analysis of woodpeckers (Aves:Picidae). *Mol. Phyl. Evol*., 8:193–204.

Prychitko, T.M. and Moore, W.S. 2000. Comparative evolution of the mitochondrial cytochrome b gene and nuclear ß-fibrinogen intron 7 in woodpeckers. *Mol. Biol. Evol*., 17:1101–1111.

Prychitko, T.M. and Moore, W.S. 2003. Alignment and phylogenetic analysis of β-fibrinogen intron 7 sequences among avian orders reveal conserved regions within the intron. *Mol. Biol. Evol*., 20:762–771.

Quandt, D., Müller, K., Stech, M. et al. 2004. Molecular evolution of the chloroplast *TRN*L-F region in land plants. *Monographs. Syst. Bot. Missouri. Bot. Gard*., 98:13–37.

Quandt, D. and Stech, M. 2005. Molecular evolution of the *trn*L_{UAA} intron in bryophytes. *Mol. Phyl. Evol*., 36:429–443.

Redelings, B.D. and Suchard, M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol*., 54:401–418.

Rockman, M.V., Rowell, D.M. and Tait, N.N. 2001. Phylogenetics of *Planipapillus*, lawn-headed onychophorans of the Australian Alps, based on nuclear and mitochondrial gene sequences. *Mol. Phyl. Evol*., 21:103–116.

Rosen, S. and Skaletsky, H.J. 1998. Primer 3. Code available at http://www-genome.wi.mit.edu/genomesoftware/other/primer3.html

Rowe, D.L. and Honeycutt, R.L. 2002. Phylogenetic relationships, ecological correlates, and molecular evolution within the Cavioidea (Mammalia, Rodentia). *Mol. Biol. Evol*., 19:263–277.

Roy, S.W. and Gilbert, W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet*., 7:211–221.

Sanchis, A., Michelena, J.M., Latorre, A. et al. 2001. The phylogenetic analysis of variable-length sequence data: elongation factor-1a introns in European populations of the parasitoid wasp genus *Pauesia* (Hymenoptera: Braconidae: Aphidiinae). *Mol. Biol. Evol*., 18:1117–1131.

Sanderson, M.J. and Shaffer, H.B. 2002. Troubleshooting molecular phylogenetic analyses. *Annu. Rev. Ecol. Sys*., 33:49–72.

Sang, T. 2002. Utilility of low-copy nuclear gene sequences in plant phylogenetics. *Critl. Rev. Biochem. Mol*., 37:121–147.

Senapathy, P., Shapiro, M.B. and Harris, N.L. 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and application to genome project. *Meth. Enzymol*., 183:252–278.

Shaw, J., Lickey, E.B., Beck, J.T. et al. 2005. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American. J. Bot*., 92:142–166.

Simmons, M.P. 2003. Independence of alignment and tree search. *Mol. Phyl. Evol*., 31:874–879.

Simmons, M.P. and Ochoterena, H. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol*., 49:369–381.

Slade, R.W., Moritz, C. and Heideman, A. 1994. Multiple nuclear-gene phylogenies: Application to pinnipeds and comparison with a mitochondrial DNA gene phylogeny. *Mol. Biol. Evol*., 11:341–356.

Slade, R.W., Moritz, C., Heidemann, A. et al. 1993. Rapid assessment of single-copy nuclear DNA variation in diverse species. *Mol. Ecol*., 2:359–373.

Slowinski, J.B., Knight, A. and Rooney, A.P. 1997. Inferring species trees from gene trees: A phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Mol. Phyl. Evol*., 8:349–362.

Sota, T. and Vogler AP 2003. Reconstructing species phylogeny of the carabid beetles *Ohomopterus* using multiple nuclear DNA sequences: heterogenous information content and the performance of simultaneous analyses. *Mol. Phyl. Evol*., 26:139–154.

Stryer, L. 1988. Biochemistry. Third Edition. W.H. Freeman and Company. New York.

Swofford, D.L. 1998. "PAUP": Phylogenetic analysis using parsimony (and other methods), version 4.0. Sinauer Associates, Sunderland, Massachussets.

Swofford, D.L., Olsen, G.J., Waddell, P.J. et al. 1996. Phylogenetic inference. In Hillis DM, Moritz C, and Mable BK, eds. Molecular Systematics, 2nd Ed. Sinauer Associates, Inc. Sunderland, Massachusetts, USA. p 407–446.

Tank, D.C. and Sang, T. 2001. Phylogenetic utility of the glycerol-3-phosphate acyltransferase gene: evolution and implications in *Paeonia* (Paeoniaceae). *Mol. Phyl. Evol*., 19:421–429.

Thompson, J.D., Gibson, T.J., Plewniak, F. et al. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids. Res*., 24:4876–4882.

Thorne, J.L., Kishino, H. and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol*., 33:114–124.

Thorne, J.L., Kishino, H. and Felsenstein, J. 1992. Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol*., 34:3–16.

Touriya, A., Rami, M., Cattaneo-Berrebi, G. et al. 2003. Primers for EPIC amplification of intron sequences for fish and other vertebrate population genetic studies. *Biotechniques*, 35:676–+.

Tsaousis, A., Martin, D., Ladoukakis, E. et al. 2005. Widespread recombination in published animal mtDNA sequences. *Mol. Biol. Evol*., 22:925–933.

Underhill, P.A., Jin, L., Lin, A.A. et al. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genet. Res*., 7:996–1005.

Underhill, P.A., Jin, L., Zemans, R. et al. 1996. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc. Natl. Acad. Sci*., USA, 93:196–200.

Weibel, A.C. and Moore, W.S. 2002. A test of a mitochondrial gene-based phylogeny of Woodpeckers (Genus *Picoides*) using an independent nuclear gene β-Fibrinogen intron 7. *Mol. Phyl. Evol*., 22:247–257.

Whittall, J.B., Medina-Marino, A., Zimmer, E.A. et al. 2006. Generating single-copy nuclear gene data for a recent adaptive radiation. *Mol. Phyl. Evol*., 39:124–134.

Wheeler, W. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics*, 12:1–9.

Wheeler, W. 1999. Fixed character states and the optimization of molecular sequence data. *Cladistics*, 15:379–385.

Wheeler, W. 2001. Homology and the optimization of DNA sequence data. *Cladistics*, 17:S3–S11.

Wheeler, W.C. and Gladstein D 2000. POY: The optimisation of alignment characters. American Museum of Natural History, New York. Program and documentation available at ftp://ftp.amnh.org/pub/molecular/poy/

Wiens, J.J. 2000. Decoupled evolution of display morphology and display behaviour in phyrosomatid lizards. *Biol. J. Linn. Soc*., 70:597–612.

Wiuf, C., Christensen, T. and Hein, J. 2001. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol*., 18:1929–1939.

Young, N.D. and Healy, J. 2003. GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics*. 4:6.

Zhang, D-H. and Hewitt, G.M. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Mol. Ecol*., 12:563–584.

Zhang, L.B., Simmons, M.P., Kocyan, A. et al. 2006. Phylogeny of the Cucurbitales based on DNA sequences of nine loci from three genomes: Implications for morphological and sexual system evolution. *Mol. Phyl. Evol*., 39:305–322.