Molecular systematics: **Perfect SINEs of evolutionary history?**

Michael M. Miyamoto

Short interspersed repetitive elements – SINEs – are being championed as near-perfect phylogenetic characters; they have recently been used with notable success to resolve some phylogenetic conundrums, but they do have certain limitations that restrict their use as 'perfect' characters for molecular systematics.

Address: Department of Zoology, P.O. Box 118525, University of Florida, Gainesville, Florida 32611-8525, USA.
E-mail: miyamoto@zoo.ufl.edu

Current Biology 1999, 9:R816–R819

0960-9822/99/$ – see front matter

The ultimate goal of phylogenetic systematics is to provide an accurate estimate of the true history of life [1]. Towards this goal, systematists are constantly searching for new characters that are free of the various shortcomings — such as parallel changes in different lineages or reversals to a more primitive state, collectively known as 'homoplasies' — that can mask the true phylogenetic history [2]. In molecular systematics, DNA sequences, as generally obtained these days by the polymerase chain reaction (PCR) and automated sequencing, remain the principal source of new information for phylogeny estimation. But both DNA and protein sequence data are vulnerable to the confounding effects of parallel and back substitutions, thereby necessitating the use of statistical methods to distinguish signal from noise [3]. At this time, no single source of comparative information — not even complete genome sequences [1] — appears to be entirely free of error and ambiguity for the determination of species' phylogenies.

This situation is changing, according to a recent series of articles by Okada and his colleagues [4,5]. They are championing the use of short interspersed repetitive elements — SINEs — as near-perfect phylogenetic characters for molecular systematics. SINEs are retroposons of 70–500 bases in length, which occur at more than 10^4 copies per eukaryotic genome. These retroposons have multiplied and inserted at various loci throughout the ancestral genomes of different lineages via RNA intermediates and reverse transcription. These ancestral retropositions have generated an extensive record of SINE insertions that varies across extant species according to their evolutionary histories. For phylogenetic purposes, this variable record is read as a series of two-state characters, the states being 'presence' or 'absence' (of the SINE). The presence of a specific SINE insertion is thought to be evidence that the species carrying it form a 'monophyletic' group, and in that sense the SINE is an example of what is known in the trade as a shared derived character or 'synapomorphy': in other words, the insertion event is assumed to have occurred in the stem lineage from which the various species are descended.

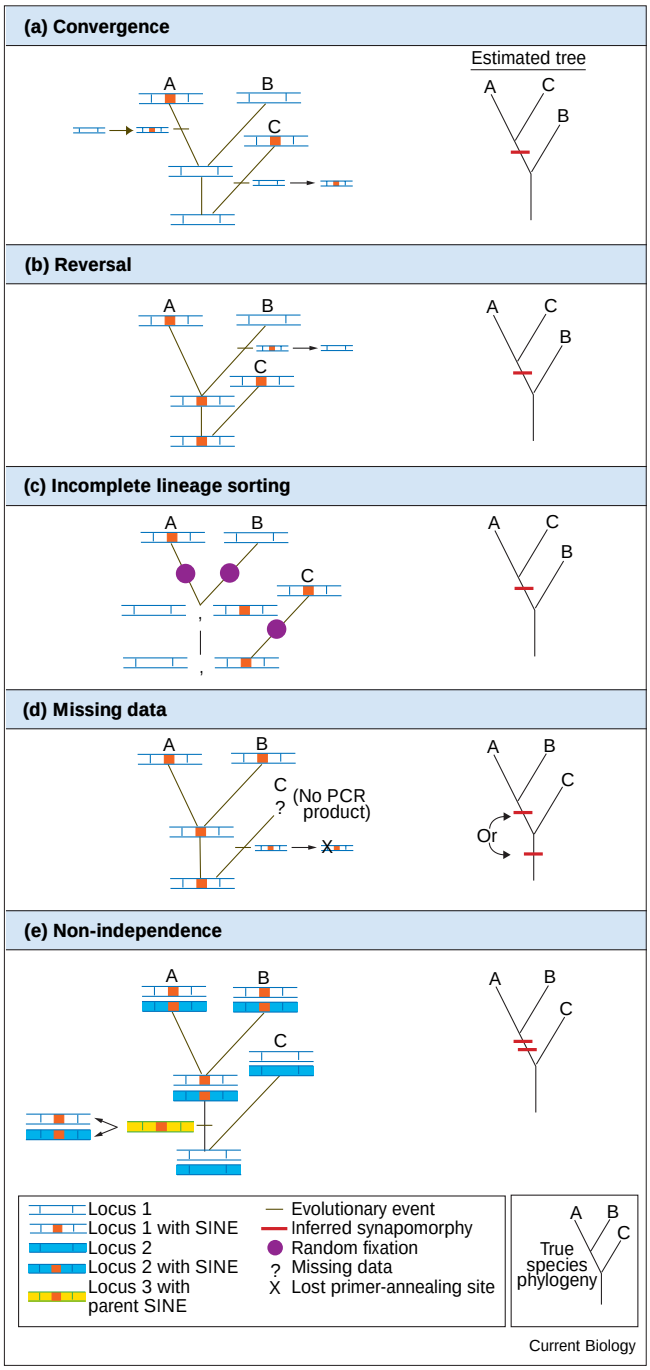
The use of SINE insertions as phylogenetic markers starts with the identification and characterization of new retroposons from a few target species of the study group [4,6]. Locus-specific primers that match the flanking regions of these elements are then designed for PCR amplification of the corresponding orthologous regions from the other species. The amplified products of the different species are resolved by gel electrophoresis, with long and short fragments corresponding to those with and without the targeted SINE, respectively. These experiments are followed by Southern hybridizations, first with a SINE-specific probe and then with one for the flanking regions, to confirm the presence or absence of the targeted element and the fidelity of the amplifications. In most cases, final confirmation of the presence or absence of a SINE is obtained by sequencing the PCR products. The entire procedure is repeated for other SINEs and loci to generate a data set for the study group with many presence/absence characters.

The retroposition process that generates a new SINE insertion is duplicative: the parent SINE is retained at its original location, while giving rise to copies that insert at new sites throughout the genome [4,5]. SINE insertions are accepted to be free of the problem of convergence (Figure 1a), as the vast number of available target sites within a eukaryotic genome and the near-random integration pattern mean that the chance of precisely the same insertion occurring more than once is negligible. They are similarly accepted to be free of reversals (Figure 1b), as no specific mechanism is known for their precise deletion. SINEs can be lost through non-specific deletions, but the chance that such a loss will exactly match the boundaries of an element is again thought to be negligible. As a final bonus, the duplicative nature of retroposition confers an evolutionary directionality onto the characters, from the primitive 'absence' state to the derived 'presence' state. In view of these features, SINE insertions are being championed as near-perfect characters for phylogeny estimation.

Overall, the supporting evidence for this claim appears solid, with the least convincing assumption being that SINEs insert randomly into the genome. Some

Figure 1

Hypothetical examples illustrating potential sources of phylogenetic error and ambiguity with SINE insertions. The true phylogeny for species A, B and C is shown at the bottom right. **(a)** An independent insertion at the same site in two different species is a convergence that introduces error into the phylogenetic analysis. **(b)** A precise deletion of an existing SINE is a reversal that also introduces error. **(c)** An ancestral polymorphism that spans successive speciations can be randomly fixed across species in a pattern that conflicts with their true relationships. **(d)** The loss of a primer-annealing site in a species by mutational decay prevents the PCR amplification of its targeted locus, thereby introducing missing data and ambiguity into the study. **(e)** The number of independent synapomorphies for a group can be over-estimated when different SINE insertions are derived from the same amplification event.



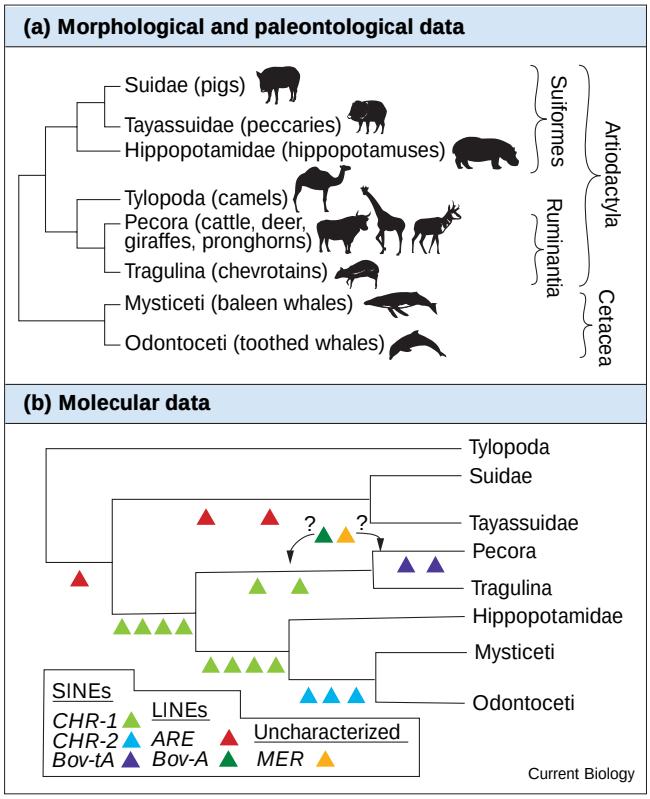
retroelements do show specific target-site selection [7], and there is at least one known example of convergent SINE insertions — in deer mice ([8], reported in [2]). These cases can, however, be regarded as exceptions to the rule of minimum insertion-site specificity by SINEs and other retroelements [4,5,7]. Although not entirely random, SINE insertions appear sufficiently so to justify their acceptance as essentially convergence-free.

There are, however, two other potential sources of error and ambiguity for phylogenetic analyses based on SINE insertions, which have been emphasized as of greater concern [4,5]. The first involves the problem of incomplete sorting of ancestral polymorphisms into the progeny lineages after speciation (Figure 1c). This source of phylogenetic error becomes a primary concern when two or more successive speciation events occur close in time [9]. When two successive speciations are separated by N generations — where N is the effective population size — the probability of such an error for a single SINE insertion can be as high as 0.3 [6]. Alleviation of this problem requires the use of multiple SINE insertions to corroborate the monophyly of individual groups.

The second problem involves the loss over evolutionary time of the primer-annealing sites for specific loci as a result of mutational decay (Figure 1d). This loss precludes the amplification of each targeted locus from every species. If a PCR fragment is missing for a species, the corresponding character state is considered to be 'missing data', thereby introducing ambiguities into the final character matrix. For example, in the study of cetartiodactyl relationships considered below [4], 18 of the 20 insertion characters for the peccary — a pig-like artiodactyl — are missing data. As mutational decay is correlated with time, this source of ambiguity is of greatest concern for the more distantly related species. This approach is thus effectively limited to taxa with proportional sequence differences of less than about 25%, and to phylogenetic questions where lineages are separated by less than about 50 million years [4,5].

One further possible source of error in SINE-based phylogenetic analysis involves the non-independence of SINE insertions that arose from the same multiplication event (Figure 1e). Although the separate insertions are independent, the co-occurrence of these SINEs across taxa is not, because of their common origins from the same amplification event. Thus, each insertion from the amplification cannot be counted as a truly independent character or synapomorphy. This problem was not discussed by Okada and colleagues [4,5], but it may be significant for their work on cetartiodactyl relationships,

Figure 2



Phylogenies of artiodactyls and cetaceans as determined from (a) morphological and paleontological data [15,16] or (b) molecular information [4,11,12]. In (b), the SINE and LINE insertions for the twenty parsimony-informative characters [4] are represented along the internodes by arrowheads and are identified to family by their colors [10]. The published most-parsimonious phylogeny of Nikaido *et al.* [4] does not show the *Bov-A* and *MER* insertions, presumably as neither can be unambiguously assigned to either Pecora or Ruminantia because of missing data for the Tragulina. Nevertheless, as for the other eighteen informative characters, these two insertions require no homoplasy in their support of the most-parsimonious phylogeny.

where groups were defined by multiple synapomorphies derived from a single family of SINEs — the *CHR-1* family (see below). These concerns are best addressed by relying on insertions from different SINE families and subfamilies [10].

These theoretical arguments about SINE insertions are complemented by new empirical evidence from a recent phylogenetic study [4] of the superorder Cetartiodactyla, which comprises the Cetacea — whales, dolphins and porpoises — and the Artiodactyla — the even-toed ungulates. The order Artiodactyla is monophyletic according to morphological and paleontological data, but according to a large body of molecular data it is paraphyletic (that is, not a natural grouping related by evolutionary descent). The latter data rather support the existence of a monophyletic whale/hippo clade — that is, grouping of the Cetacea with the Hippopotamidae (Figure 2). These molecular data include DNA and protein sequences for many nuclear and mitochondrial genes, and for complete mitochondrial DNA genomes [11,12].

The available insertion sequence data for Cetartiodactyla consist of twenty informative characters for fifteen SINEs, four long interspersed repetitive elements (LINEs) and one uncharacterized retroposon [4]. On the basis of these twenty characters, the 'most-parsimonious' phylogeny — the one requiring the fewest distinct character changes — requires no homoplasy and is congruent with the large body of other molecular data (Figure 2b). The whale/hippo clade is supported by four synapomorphies from the *CHR-1* family, and the paraphyly of the Artiodactyla is further implied by four additional SINE and one LINE insertions from the *CHR-1* and *ARE* families, respectively.

Further analysis at the subfamily level is required to show whether the various insertions on which this analysis was based were truly independent [10]. The conclusions have, however, received further support recently from a couple of sources. The first is a recent study [13] of the key morphological and paleontological data supporting artiodactyl monophyly — primarily based on tarsal anatomy — which concluded that the cetaceans may be nested within the order Artiodactyla. And the second is the convincing rebuttal of a recent criticism of SINE insertions [14], which has been shown to be based on a misinterpretation of failed PCR amplifications as reflecting the absence of an element rather than missing data (N. Okada, personal communication). These considerations, together with the consistency and congruence of the molecular data, add up to compelling support for the view that SINE elements are a powerful tool in phylogenetic analysis (see also [6]).

SINE insertions are currently being championed as near-perfect phylogenetic characters, because of their distinct molecular and evolutionary properties and their success in recent empirical studies [4–6]. But the history of molecular systematics is littered with examples of promising approaches that were eventually revealed to be distinctly less than perfect [2]. The recent claims about the advantages of SINE insertions promise to generate much attention from the systematics community. Such scrutiny is to be welcomed, as it will lead to more critical tests of the power of SINE insertions for the resolution of difficult phylogenetic problems.

Acknowledgements
I thank David M. Hillis and Norihiro Okada for sending me pre-prints of their articles in press and in review, David M. Hillis and Michele R. Tennant for their comments and suggestions about the research, Jennifer Piascik for drafting the two figures of my manuscript, and the Department of Zoology, University of Florida for their financial assistance.

References

1. Doolittle WF: Phylogenetic classification and the universal tree. *Science* 1999, 284:2124-2128.
2. Hillis DM: SINEs of the perfect character. *Proc Natl Acad Sci USA* 1999, 96:9979-9981.
3. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM: Phylogenetic inference. In *Molecular Systematics*, 2nd edn. Edited by Hillis DM, Moritz C, Mable BK. Sunderland, MA: Sinauer; 1996:407-514.
4. Nikaido M, Rooney AP, Okada N: Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. *Proc Natl Acad Sci USA* 1999, 96:10261-10266.
5. Shedlock AM, Okada N: SINE insertions: powerful tools for molecular systematics. *BioEssays* 1999, in press.
6. Murata S, Takasaki N, Saitoh M, Tachida H, Okada N: Details of retropositional genome dynamics that provide a rationale for a generic division: the distinct branching of all the Pacific salmon and trout (***Oncorhynchus***) from the Atlantic salmon and trout (***Salmo***). *Genetics* 1996, 142:915-926.
7. Craig NL: Target site selection in transposition. *Annu Rev Biochem* 1997, 66:437-474.
8. Cantrell MA, Wichman HA: Multiple independent SINE insertions at identical loci in deer mouse species. *Soc Syst Biol Meeting*, Madison, Wisconsin; 1999:12.
9. Tachida H, Iizuka M: A population genetic study of the evolution of SINEs. I. Polymorphism with regard to the presence or absence of an element. *Genetics* 1993, 133:1023-1030.
10. Shimamura M, Abe H, Nikaido M, Ohshima K, Okada N: Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of $tRNA^{Glu}$-derived families of SINEs. *Mol Biol Evol* 1999, 16:1046-1060.
11. Gatesy J, Milinkovitch M, Waddell V, Stanhope M: Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa. *Syst Biol* 1999, 48:6-20.
12. Ursing BM, Arnason U: Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. *Proc R Soc Lond B Biol Sci* 1998, 265:221-225.
13. Thewissen JGM, Madar SI: Ankle morphology of the earliest cetaceans and its implications for the phylogenetic relations among ungulates. *Syst Biol* 1999, 48:21-30.
14. Luckett WP, Hong N: Phylogenetic relationships between the orders Artiodactyla and Cetacea: a combined assessment of morphological and molecular evidence. *J Mammal Evol* 1998, 5:127-182.
15. Gentry AW, Hooker JJ: The phylogeny of Artiodactyla. In *The Phylogeny and Classification of the Tetrapods, Volume 2. Mammals.* Edited by Benton MJ. Oxford: Clarendon Press; 1988:235-272.
16. Geisler JH, Luo Z: Relationships of Cetacea to terrestrial ungulates and the evolution of cranial vasculature in Cete. In *The Emergence of Whales.* Edited by Thewissen JGM. New York: Plenum Press; 1998:163-212.