

Hemiplasy and homoplasy in the karyotypic phylogenies of mammals

Terence J. Robinson^{*†}, Aurora Ruiz-Herrera[‡], and John C. Avise^{†§}

^{*}Evolutionary Genomics Group, Department of Botany and Zoology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa; [†]Laboratorio di Biologia Cellulare e Molecolare, Dipartimento di Genetica e Microbiologia, Università degli Studi di Pavia, via Ferrata 1, 27100 Pavia, Italy; and [§]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697

Contributed by John C. Avise, July 31, 2008 (sent for review May 28, 2008)

Phylogenetic reconstructions are often plagued by difficulties in distinguishing phylogenetic signal (due to shared ancestry) from phylogenetic noise or homoplasy (due to character-state convergences or reversals). We use a new interpretive hypothesis, termed hemiplasy, to show how random lineage sorting might account for specific instances of seeming “phylogenetic discordance” among different chromosomal traits, or between karyotypic features and probable species phylogenies. We posit that hemiplasy is generally less likely for underdominant chromosomal polymorphisms (i.e., those with heterozygous disadvantage) than for neutral polymorphisms or especially for overdominant rearrangements (which should tend to be longer-lived), and we illustrate this concept by using examples from chiropterans and afrotherians. Chromosomal states are especially powerful in phylogenetic reconstructions because they offer strong signatures of common ancestry, but their evolutionary interpretations remain fully subject to the principles of cladistics and the potential complications of hemiplasy.

cladistics | gene trees | lineage sorting | phylogeny | species trees

In phylogenetic analyses, systematists routinely strive to distinguish homology (trait similarity due to shared ancestry) from homoplasy (trait similarity arising from evolutionary convergence, parallelism, or character-state reversals). Homology can offer valid phylogenetic signal, whereas homoplasy is regarded as evolutionary noise that, if not properly accommodated, jeopardizes phylogenetic reconstructions. Homology itself has distinct components, as first emphasized by Hennig (1) in his insightful distinction between symplesiomorphies (traits showing shared-ancestral homology) and synapomorphies (traits with shared-derived homology). From a Hennigian perspective, only valid synapomorphies properly earmark clades.

The critical distinctions between homoplasy and homology and between different kinds of homology have served the field of systematics well. However, a difficulty arises when a shared-derived genetic trait that from mechanistic considerations should be homoplasy-free nonetheless recurs in two or more taxa that seem to be unrelated. For example, suppose that a derived chromosomal inversion with presumably unique (monophyletic) endpoint breaks is present in two or more species that belong to disparate clades. Under the traditional interpretive framework outlined above, this phylogenetic dilemma could only be resolved in either of two ways: by supposing that the inversion has evolved multiple times independently, notwithstanding mechanistic karyotypic arguments to the contrary; or by supposing that the shared trait does earmark a bona-fide organismal clade, notwithstanding independent phylogenetic evidence to the contrary.

Here, we raise another potential explanation for this kind of phylogenetic enigma, and illustrate its application to karyotypic data involving chromosomal syntenies in mammals. Each synteny is a large conserved block of DNA, i.e., a linked assemblage of ordered loci. Numerous syntenies have been revealed in various mammals through chromosomal painting, principally by using flow-sorted whole human chromosomes as genetic probes

(2). These syntenic blocks, sometimes shared across even distantly related species, may involve entire chromosomes, chromosomal arms, or chromosomal segments. Based on the premise that each syntenic assemblage in extant species is of monophyletic origin, researchers have reconstructed phylogenies and ancestral karyotypes for numerous mammalian taxa (ref. 3 and references therein). Normally, the phylogenetic inferences from these cladistic appraisals are self-consistent (across syntenic blocks) and taxonomically reasonable, and they have helped greatly to identify particular mammalian clades. However, in a few cases problematic phylogenetic patterns of the sort described above have emerged.

We recently introduced the term hemiplasy (3), which we defined as any topological discordance between a gene tree and a species tree attributable to the phylogenetic sorting of genetic polymorphisms across successive nodes in a species tree. Two fundamentally equivalent diagrammatic representations of hemiplasy are shown in Fig. 1. Therefore, hemiplasy is a genuine form of trait homology (orthology, either of alleles or of genealogical lineages) that gives the illusion of homoplasy in an organismal tree, but nonetheless is not homoplasy at the gene-tree level. Hemiplasy is also to be distinguished from other potential sources of gene-tree/species-tree discordance, including introgression, genetic transformation, or viral-mediated DNA transfer (which can be important in some biological settings, but which we do not consider further in this article). Here, we raise the possibility that hemiplasy might account for some of the phylogenetic anomalies that seem to be present in the taxonomic distributions of particular syntenic blocks of genes within mammalian karyotypes.

Background

Theory. Nei (ref. 4, pp. 401–403) quantified the theoretical probability of what we would now term hemiplasy for the simplest possible evolutionary case: neutral alleles in three related and geographically unstructured species. Under that scenario, the probability of a gene-tree/species-tree discordance is $(2/3)e^{-T/2N}$ where T is the number of generations between the successive speciation events, N is the effective population size, and e is the base of the natural logarithms. According to this formula, hemiplasy is likely under some realistic evolutionary parameters. For example, the single-locus probability of hemiplasy (or the expected percentage of loci displaying hemiplasy) is $\approx 50\%$ when $T/2N = 0.3$, as would be true, for example, if two speciation nodes were ≈ 1.0 million generations apart and the effective population size between nodes was ≈ 1.7 million individuals. Similar theory applied to more complex phylogenetic settings can be found in refs. 5 and 6.

Author contributions: T.J.R. and J.C.A. designed research; T.J.R. and A.R.-H. performed research; T.J.R. and A.R.-H. analyzed data; and T.J.R. and J.C.A. wrote the article.

The authors declare no conflict of interest.

[†]To whom correspondence may be addressed. E-mail: tjr@sun.ac.za or javise@uci.edu.

© 2008 by The National Academy of Sciences of the USA

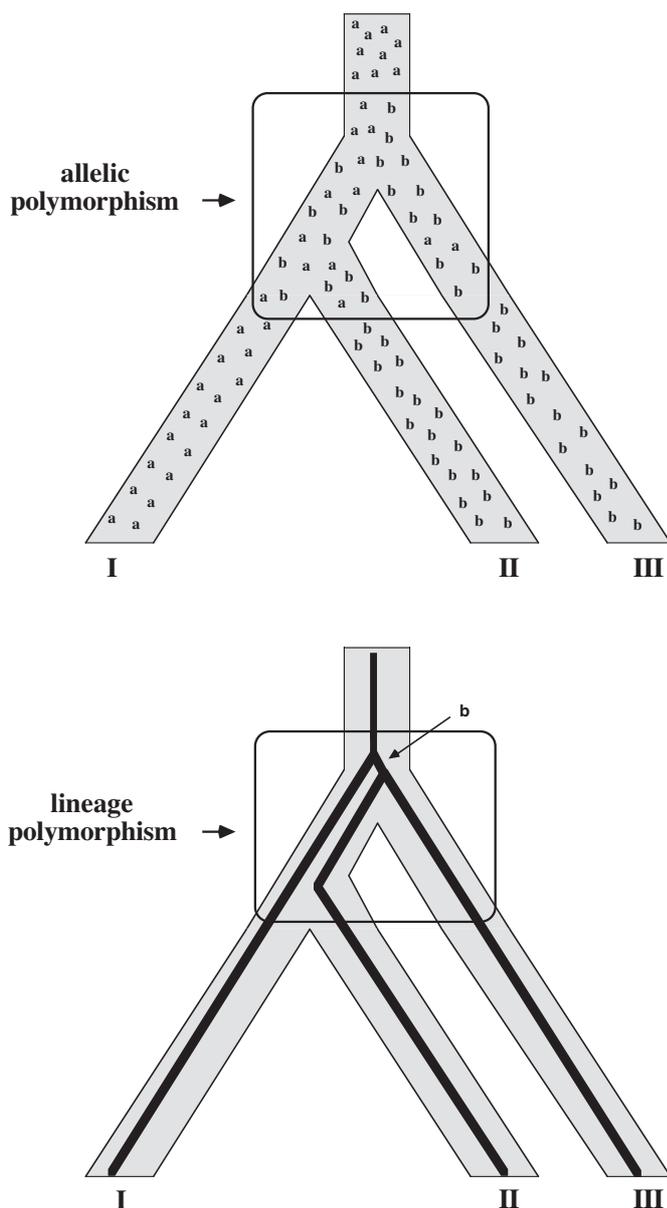


Fig. 1. Schematic representations of hemiplasy. Shown are the distributions of a genic or chromosomal polymorphism (*Upper*) and a set of genealogical lineages (*Lower*) that traversed successive speciation nodes in an organismal phylogeny (broad branches) only to become fixed, by lineage sorting, in descendant species in a pattern that appears at face value to be discordant with the species phylogeny. In these diagrams, species II and III both have the homologous and derived character “b,” so the gene tree gives the impression that species II and III are the more closely related. However, in truth species I and II are sister taxa, despite the fact that species I alone retains the ancestral genetic condition “a.”

SINE Data. Short-interspersed elements (SINEs) have long been touted as powerful phylogenetic markers, for several reasons (7): each SINE is a retropseudogene that resides at a specific chromosomal location; each occupied site is thought to represent a single (monophyletic) insertion event; SINEs are stable once inserted into a genome [in large part because no known mechanism exists for excision of an element, such that SINE absence at a chromosomal site presumably reflects the ancestral state (8)]; and large numbers of independent SINEs are dispersed in the genomes of most eukaryotic organisms. For these reasons,

any SINE shared by different species was initially assumed to be definitive in marking an organismal clade. As phrased by Nikaido *et al.* (9), with respect to SINEs “the probability that homoplasy will obscure phylogenetic relationships is, for all practical purposes, zero.”

However, it soon became clear that different SINEs sometimes (albeit rarely) disagree in the organismal clades they presumably delineate. This and other evidence led Hillis (10) to note that lineage sorting can introduce homoplasy-like outcomes (that we would now term hemiplasy) when a polymorphism becomes fixed in some but not all of the descendants of a polymorphic ancestor. Researchers quickly acknowledged this possibility for some of the occasional character conflicts among SINEs. For example, Shedlock and Okada (11) wrote that “if the time to fixation transcends species boundaries formed in rapid succession, such as can occur during explosive radiations, inconsistent patterns of SINE insertions may be observed because of ancestral polymorphism.” Similar statements now appear as standard caveats for published phylogenies on SINEs (12), and there has emerged a widespread recognition that SINEs are merely “nearly perfect” (13) rather than perfect phylogenetic markers.

Karyotypic Hemiplasy. Syntenic blocks involving entire chromosomes, chromosomal arms, or large chromosomal segments are sometimes shared across even distantly-related mammalian species. Standard wisdom is that each such syntenic block is of monophyletic origin, because the independent assembly of a shared synteny in different lineages seems mechanistically highly implausible (14). In other words, shared syntenic associations are much more useful in defining clades than chromosomal reorganizations that disrupt an ancestral synteny, because the latter might recur by means of breakpoint reuse (15–18), especially in regions of segmental duplications (19), high concentrations of repetitive elements, and fragile sites (18, 20, 21). With respect to monophyletic origin (but not with respect to the probability of later evolutionary loss), the potential phylogenetic utility of syntenic blocks thus bears a considerable analogy to the phylogenetic utility of SINE elements, both of which are viewed as rare genomic changes (22).

During the course of comparing previously identified syntenic blocks in eutherian mammals against outgroup taxa (23), we noticed at least two candidate examples of hemiplasy (involving chiropterans and afrotherians) that have motivated this report. In what follows, each syntenic block in a given mammalian species is named according to the human chromosome or chromosomal arm to which its linked loci are apparently homologous, as judged by cross-species chromosome painting (CP) by using human chromosomes as genetic probes. These data are sometimes used to construct phylogenetic trees or, more frequently, the chromosomal characters themselves are mapped onto a consensus sequence-based tree.

Chiroptera. By using similar types of chromosomal analyses, cross-species CP coupled with chromosomal mapping to a consensus sequence-based tree derived from (24–26), Mao *et al.* (27) identified 10 presumed instances of homoplasy (convergent evolution in this case) in Chiroptera that they regarded as weakening the reliability of chromosomal characters for resolving interfamily relationships of bats. For example, a presumably homoplastic synteny (chromosomal block HSA 1/6/5) was shared by particular species representing two rather distinct bat families, Pteropodidae and Megadermatidae (Fig. 2). However, an alternative possibility is that this and other such cases might be due to hemiplasy.

By using calibrations from a relaxed Bayesian molecular clock (28), Eick *et al.* (24) estimated the following divergence times for pertinent nodes in the chiropteran tree (Fig. 2): 41 mya for the

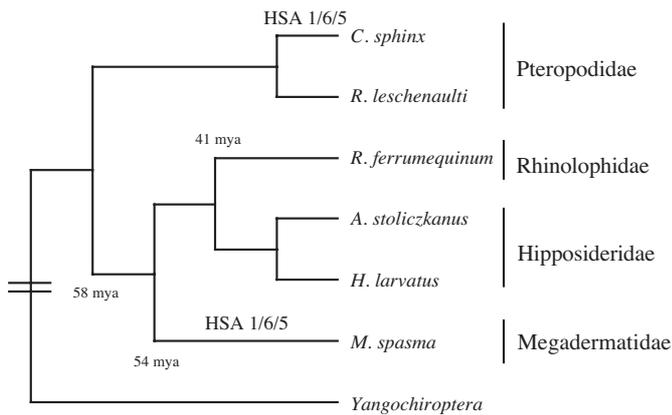


Fig. 2. Phylogenetic relationships among bat species representing four taxonomic families (redrawn from ref. 27). Also shown is the presence of the chromosomal synteny 1/6/5, which formerly was interpreted to be homoplasid (27) but might instead be an example of hemiplasy.

split of Rhinolophidae from Hipposideridae; 54 mya for the split of the Megadermatidae from Hipposideridae plus Rhinolophidae; and 58 mya for the divergence of Pteropodidae from Rhinolophidae plus Hipposideridae plus Megadermatidae. These estimates provide a temporal framework for the required persistence times (under the hemiplasy scenario) of chromosomal polymorphisms across successive speciation nodes. If the syntenic block HSA 1/6/5 was part of a polymorphism in the ancestral lineage leading to Pteropodidae and Megadermatidae, this polymorphism must have persisted for ≈ 4 million years before sorting eventually into the descendant lineages where the different chromosomal arrangements today are housed. Similar reasoning can be applied to the other “homoplasid” karyotypic states identified by Mao *et al.* (27).

Most of the rearrangements responsible for repatterning chiropteran genomes involve centric or Robertsonian fusions (29, 30), which, unless in monobrachial combinations (31), are perceived as not being particularly underdominant (i.e., possessing heterozygous disadvantage). So, perhaps such rearrangements occasionally do survive as polymorphic states for considerable lengths of time, although this would also depend on historical variables including effective population size and spatial population structure. Robertsonian polymorphisms, such as the $2n = 56, 58, 62$ series documented in *Rhinolophus hipposideros* from different geographic areas and the $2n = 42 - 44$ variation in *Rhinolophus pearsoni* from various provinces in China (32), are well known in bats, and in some other mammalian groups including rodents (33, 34), bovids (35), insectivores (36), and primates (37). In some cases (such as in the 44-chromosome gibbons), multiple related species share polymorphic chromosomal conditions, indicating that the polymorphisms have survived speciation events (38).

Afrotheria. This well supported mammalian clade of Afro-Arabian origin is comprised of elephants, sireneans, elephant shrews, golden moles, tenrecs, and the aardvark. One of the most enduring phylogenetic problems within the group concerns the position of aardvark and elephant shrews with respect to other afrotherians (39). Strong support exists for a clade (the Afroinsectiphillia) comprising elephant shrews, aardvarks, tenrecs, and golden moles (40–42), but sister group relationships within the Afroinsectiphillia remain vague. For example, Amrine-Madsen *et al.* (40) and Murphy *et al.* (42), among others, find evidence for a phylogenetic association of elephant shrew, tenrec, and golden mole to the exclusion of aardvark (Fig. 3A), whereas Waddell and Shelley (41) conclude from an analysis of a

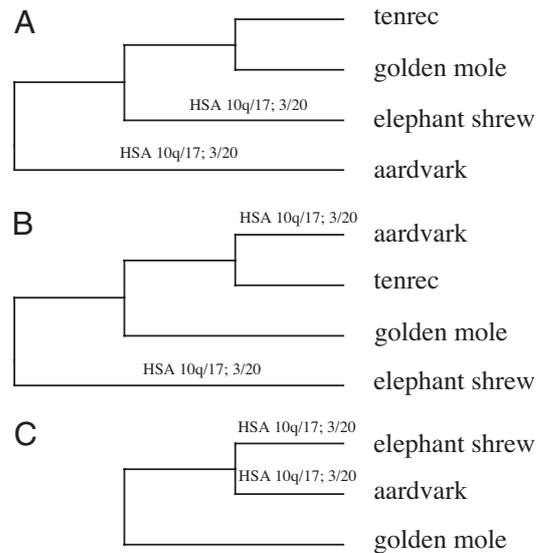


Fig. 3. Competing phylogenetic hypotheses for species comprising the Afroinsectiphillia clade (aardvark, elephant shrew, golden mole, and tenrec). (A) Clade comprising the elephant shrew, tenrec, and golden mole to the exclusion of aardvark based on concatenations of nuclear and mitochondrial DNA sequences (40, 42). (B) Clade composed of aardvark plus tenrec and a weaker grouping of these taxa and the golden mole to the exclusion of elephant shrew based on a different concatenation (41). (C) Clade composed of aardvark and elephant shrew to the exclusion of golden mole based on cross-species CP (43).

different concatenation of nuclear and mitochondrial sequences that an association exists of aardvark, tenrec, and golden mole to the exclusion of elephant shrew (Fig. 3B). To further complicate matters, cytogenetic data (43) seem at face value to unite elephant shrew and aardvark to the exclusion of golden mole (Fig. 3C). Underpinning this latter suggestion are two syntenic blocks (HSA 10q/17 and HSA 3/20) that are present in elephant shrew (43, 44) and aardvark (45) but absent in golden mole (43).

To reconcile these seemingly incompatible phylogenetic interpretations, at least three possibilities exist (apart from independent convergent evolution of the syntenic blocks, which seems highly unlikely on mechanistic grounds). First, perhaps the phylogeny in Fig. 3A is correct, in which case 10q/17 and 3/20 could be long-retained symplesiomorphies (stemming from the ancestor of Afroinsectiphillia) in the extant elephant shrew and aardvark lineages. In principle, hemiplasy for 10q/17 and 3/20, as diagrammed in Fig. 1, could also account for this phylogeny. Second, perhaps the phylogeny in Fig. 3C is correct, but this would contradict much other phylogenetic evidence. Last, perhaps Fig. 3B is correct, in which case the 10q/17 and 3/20 syntenies must have been polymorphic in the ancestor to Afroinsectiphillia and the polymorphisms were later lineage-sorted in a way that produced the gene tree/species tree discordances. This latter possibility would be an example of hemiplasy in a phylogeny for which the alternative explanation of symplesiomorphy (as traditionally defined) would not apply.

Both the 10q/17 and 3/20 syntenies in the aardvark and elephant shrew appear to result from Robertsonian fusions (43–45) that probably arose ≈ 75 mya in a common ancestor to Afroinsectiphillia (these syntenies are not present in the closest outgroups, elephant and manatee; refs. 45–47). Likely origination dates for the elephant shrew lineage and for the tenrec/golden mole/elephant shrew clade (Fig. 3) are reportedly ≈ 73 mya and ≈ 65 mya, respectively (42). Thus, if either of the hypothetical scenarios involving lineage sorting and hemiplasy is correct, then these derived chromosomal syntenies must have

persisted as polymorphisms (with their respective counterpart arrangements) for at least ≈ 2 million years, and maximally for more than ≈ 10 million years to temporally encompass the relevant speciation nodes.

Discussion

Distinguishing phylogenetic signal (due to shared ancestry) from phylogenetic noise or homoplasy (e.g., due to character-state convergence) is the perennial challenge in systematics. The complications of homoplasy are widely appreciated in DNA sequence analyses (as well as in morphology-based systematics), but they can also attend efforts to reconstruct phylogenies from other categories of data including chromosomal characters (48–51). Another type of potential phylogenetic complication is hemiplasy, as defined in *Background* and Fig. 1. Although the distinction between gene trees and species phylogenies has long been appreciated (52–56), the phylogenetic ramifications of hemiplasy (in contrast with those of homoplasy) have often been overlooked.

Hemiplasy as a theoretically plausible evolutionary phenomenon is not in doubt. It is a logical consequence of several biological realities: lineage turnover by means of differential organismal reproduction, the fact that species are composed of populations of individuals, the fact that every new mutation originates in one individual or family, and, therefore, the fact that any evolutionary change of state necessitates a transitional phase of genetic polymorphism at the population level. Hemiplasy can also affect any and all classes of phylogenetic marker, regardless of their mechanistic susceptibilities to homoplasy. The outstanding question is whether hemiplasy materially affects real datasets.

Some of the most favorable opportunities for identifying instances of hemiplasy involve datasets for which homoplasy can be essentially disregarded as a complicating evolutionary factor. For example, all extant occurrences of a SINE element at a particular chromosomal site presumably trace back to a single (monophyletic) insertion event, so convergent evolution can often be quite safely eliminated as an explanation for why particular species might share a particular SINE; and this feature in turn has also helped researchers identify some probable (but formerly overlooked) instances of hemiplasy in SINE data (see *Background*).

Here, we have adopted a similar rationale to address the possibility that some chromosomal synteny, each thought to be of monophyletic origin, might likewise be present in seemingly unrelated species because of lineage sorting from polymorphic conditions. The term hemiplasy formalizes and extends some of the pioneering cytogenetic observations by Dutrillaux and co-workers involving idiosyncratic lineage sorting in primates (57), and it offers an alternative explanation for some chromosomal states that conventionally were interpreted to have arisen convergently in different lineages, or were subject to evolutionary reversals each requiring the precise disruption of two adjacent synteny.

The phenomenon of hemiplasy is most plausible when the internodal distances in a phylogenetic tree are short (relative to effective population sizes) and/or when the persistence time of a polymorphism is long. The latter, in turn, is more likely for neutral polymorphisms (such as some Robertsonian fusions that

have little or no impact on fertility) than it is for polymorphisms that are underdominant, and it is especially likely for balanced polymorphisms. Hemiplasy is also more likely in species that are subdivided geographically, such that the multiple genetic elements of a collective polymorphism are buffered against extinction by virtue of being housed and perhaps fixed in different populations. Conversely, hemiplasy is least likely under circumstances where the fixation of new genetic variants occurs rapidly in each nonstructured species through genetic drift, inbreeding, selection in favor of homozygotes, or meiotic drive.

Even in genuinely homoplasy-free datasets, the distinction between a gene tree and species tree and the possibility of hemiplasy mean that no single genetic character can be deemed definitive in earmarking a clade. In recognition of this fact, Waddell *et al.* (58) proposed a statistical framework for testing clades by using data from SINEs. The particular tests apply to any rooted tree, with three taxa, under a Wright–Fisher coalescent model and assumptions of panmixia, nonoverlapping generations, and constant population size. From their analyses (which in effect take lineage sorting and the possibility of hemiplasy into account), at least three SINE (or chromosomal) characters (none being contradictory) would be required to reject alternative phylogenetic groupings at the 95% confidence level.

In practice, even this small number of synapomorphic traits might be difficult to attain in various types of karyotypic data. Consider, for example, the Afroinsectiphillia, where only two shared synteny (HSA 10q/17 and HSA 3/20) currently support a sister association between elephant shrew and armadillo (the probability of this outcome by chance lineage sorting is 0.11; see table 3 in ref. 58). Thus, strict rejection of the alternate hypotheses [(armadillo plus golden mole) plus elephant shrew, or (elephant shrew plus golden mole) plus armadillo] is not yet possible. Such considerations highlight a general caveat about the use of otherwise powerful chromosomal rearrangements in constructing phylogenies: the number of informative characters at a particular node must be above the critical threshold required for statistical support (given the reality of the distinction between gene trees and a species tree).

In closing, we do not claim to have proven any instances of hemiplasy in the current karyotypic datasets on syntenic blocks. Rather, our intent has been to raise consciousness about the hemiplasy phenomenon in the field of chromosomal research, and thereby stimulate further discussion along these lines. The impact of hemiplasy is a function of several historical variables (including population demographics, selection mode and intensity, and internodal times) that traditionally have been almost entirely neglected in phylogenetic reconstructions based on chromosomal characters. It would be useful to evaluate the possibility of karyotypic hemiplasy in a wide variety of phylogenetic settings.

ACKNOWLEDGMENTS. We thank Peter Waddell for drawing our attention to the applicability of statistical analyses developed for SINEs in testing of the validity of phylogenetic affinities based on chromosomal data. Gauthier Dobigny, Lutz Froenicke, and Darren Griffin provided insightful comments on various drafts of this manuscript. T.J.R. was supported by the National Research Foundation (South Africa), and A.R.-H. was supported by a Spanish Ministry of Education and Science (MEC) Postdoctoral Fellowship. J.C.A. was supported by funds from the University of California at Irvine.

- Hennig W (1950) *Grundzüge einer Theorie der Phylogenetischen Systematik* (Deutscher Zentralverlag, Berlin); (1966) *Phylogenetic Systematics* (University of Illinois Press, Urbana, Illinois) (English).
- Ferguson-Smith MA, Trifonov V (2007) Mammalian karyotype evolution. *Nat Rev Genet* 8:950–962.
- Avise JC, Robinson TJ (2008) Hemiplasy: A new term in the lexicon of phylogenetics. *Syst Biol* 57:503–507.
- Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).

- Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. *Theor Pop Biol* 61:225–247.
- Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet* 2:e68.
- Shedlock AM, Okada N (2000) SINE insertions: Powerful tools for molecular systematics. *BioEssays* 22:148–160.
- Nishihara H, *et al.* (2005) A retroposon analysis of Afrotherian phylogeny. *Mol Biol Evol* 22:1823–1833.

