

Orthologous Repeats and Phylogenetic Inference

Ali Bashir*

Chun Ye

Alkes Price

Vineet Bafna†

Abstract

Determining phylogenetic relationships between species is a difficult problem, and many phylogenetic relationships remain unresolved, even among eutherian mammals. Repetitive elements provide excellent markers for phylogenetic analysis, because their mode of evolution is predominantly homoplasy-free and unidirectional. Historically, phylogenetic studies using repetitive elements have relied on biological methods such as PCR analysis. Here, we present a purely computational method for inferring phylogenetic relationships from partial sequence data using orthologous repeats. We apply our method to the phylogeny of 28 mammals, obtaining encouraging results. With the set of species with partial sequence data available now rapidly expanding, computational analysis of repetitive elements holds great promise for the future of phylogenetic inference.

Introduction

Repetitive elements, particularly SINEs (short interspersed elements) and LINEs (long interspersed elements), provide excellent markers for phylogenetic analysis: their mode of evolution is predominantly homoplasy-free, since they do not typically insert in the same locus of two unrelated lineages, and unidirectional, since they are not precisely excised from a locus with the flanking sequences preserved (Shedlock and Okada, 2000 [21]). Indeed, the use of SINEs and LINEs to elucidate phylogeny has a rich history. SINEs and LINEs have been used to show that hippopotamuses are the closest living relative of whales (Shimamura et al., 1997 [22]; Nikaido et al., 1999 [15]), to determine phylogenetic relationships among cichlid fish (Takahashi et al., 2001 [24, 25], Terai et al., 2003 [26]), and to elucidate the phylogeny of eight primate species, providing the strongest evidence yet that chimps are the closest living relative of humans (Salem et al., 2003 [18]). In each one of these studies, the presence or absence of a repetitive element at a specific locus in a given species was determined by PCR analysis, using flanking sequences as primers: a long PCR fragment indicates that the element is present, a short PCR fragment indicates that the element is absent, and the case of no PCR amplification is inconclusive. Presence or absence of the element at informative loci can subsequently be verified via Southern hybridization or actual sequencing. With a sufficiently informative set of loci, the phylogeny of the given species can then be inferred.

It has been suggested in the past that SINE/LINE insertion studies would not make a widespread contribution to phylogenetic inference in the short term, because

The amount of time, money, and effort needed to collect data on relatively few characters will be prohibitive.
-Hillis, 1999 [8].

We agree that the biological methods described above are highly resource intensive. With the set of species with partial sequence data available now rapidly expanding, we propose instead to determine the presence or absence of a repetitive element at specific loci in each given species, and infer the resulting phylogeny, purely by computational means. Previous work has already hinted at the potential of this approach: for example, Thomas et al, 2003 [5] identified 3 repetitive elements which support a Primate-Rodent clade, and Schwartz et al., 2003 [19] identified 1 repetitive element which supports a horse-Carnivore clade. Our work extends the computational analysis of repetitive elements to elucidate phylogeny to a much larger scale. We predict that computational analysis of SINEs and LINEs will make a widespread contribution to phylogenetic inference in the years ahead.

*Bioinformatics Program, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412. Email:abashir@bioinf.ucsd.edu.

†Computer Science Department, APM 3832, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0114. Email:vbafna@cs.ucsd.edu. Ph: 858-822-4978(W), 858-534-7029(F)

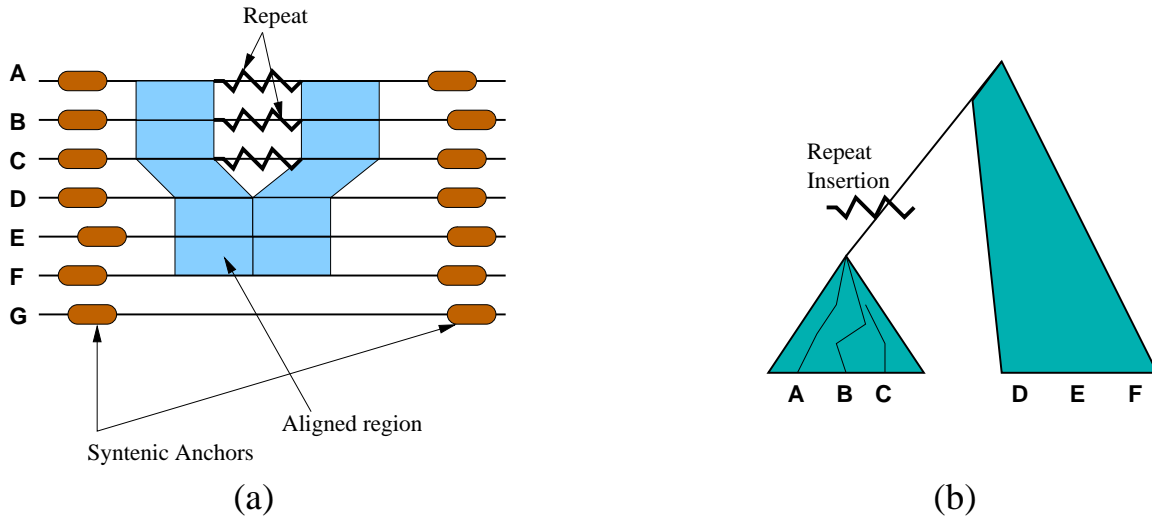


Figure 1: (a) A schematic diagram of syntenic regions in 3 species, with a repeat insertion in $A, B,$ and C . The light blue regions correspond to regions that align well, indicating that the repeat is present in A, B, C and absent in D, E, F . Neither presence, nor absence can be verified for G (b) A likely phylogeny consistent with a parsimonious explanation of the data. Species $A, B,$ and C belong in a clade that can be separated from $D, E, F,$ and the repeat was inserted in a common ancestor of the 3 species. There is no constraint on where G might occur.

We apply our method to obtain results on the phylogeny of 28 (mostly eutherian) mammals. We note that the phylogeny of eutherian mammals is largely unresolved, as there are many instances where previous studies reach conflicting conclusions (Amrine-Madsen et al., 2003 [2]). Our results resolve some of these conflicting conclusions, and are otherwise consistent with previous studies. Given the predominantly homoplasy-free, unidirectional nature of SINE/LINE insertions, we are optimistic that, with an increased amount of sequence data available in the future, our approach will conclusively resolve eutherian phylogeny and other phylogenetic problems.

Approach

Consider a syntenic genomic region in a set of n species. Figure 1(a) describes this schematically for $n = 7$ species. The synteny is determined by flanking orthologous regions such as single copy genes in all 7 species. Further, let n_1 ($n_1 = 3$ in Figure 1) of these n genes contain a repeat element R such that removing this repeat element results in a largely gap-free local multiple alignment of 6 of the 7 species. The multiply aligned region is depicted by the blue shaded lines in Figure 1 (a). The most parsimonious phylogeny explaining this scenario will have the 3 species in a clade with R inserted in a common ancestor (Figure 1(b)). Any other scenario would imply either that R was inserted at exactly the same location multiple times in different species, or that the insertion of R in a species was followed by a deletion event that removed only the region containing $R,$ and nothing else. Both of these are rare events, and therefore less plausible. The absence of a strong alignment (perhaps due to a deletion event) in G implies that neither presence, nor absence of R can be verified. Thus, repeat R does not impose any phylogenetic constraint on G .

As transposable repeat elements are very common, particularly in mammals, a collection of phylogenetic constraints such as the one in Figure 1(b) could be used to automatically construct a complete phylogeny. Through a multiple alignment procedure (to be described in detail in the METHODS section), we have a collection of orthologous regions containing a subset of species in which a repeat was inserted in exactly the same location, and a disjoint subset in which the repeat was not inserted. This information is computed as an *orthologous-repeats table*, O , with rows corresponding to species, and columns to repeats. The entries are given by

$$O[i, c] = \begin{cases} 1 & \text{if species } i \text{ clearly contains repeat } c \\ 0 & \text{if species } i \text{ clearly does not contain repeat } c \\ ? & \text{otherwise} \end{cases}$$

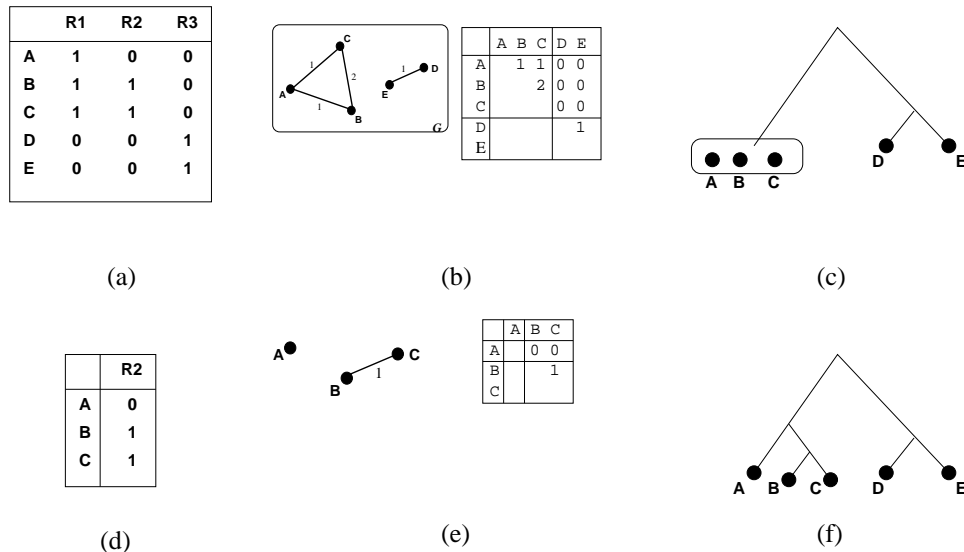


Figure 2: Sketch of phylogeny reconstruction from the orthologous-repeats table. (a) An orthologous-repeats table with 5 species and 3 repeats. (b) The resulting shared-repeat graph. We also illustrate the graph in matrix form. (NOT YET: NEED TO ADD.) Note that the connected components of the graph correspond to clades in the final phylogeny. (c) One of the two clades has 2 species and is therefore resolved. The other has 3 species, and needs to be resolved further. (d) The orthologous-repeats sub-table of species A, B, C . Only repeat $R2$ contains two 1's and one 0. (e) The resulting shared-repeat sub-graph resolves species A, B, C . (f) The final phylogeny.

For each column c , and triple (i, j, k) , where $O[i, c] = O[j, c] = 1$, and $O[k, c] = 0$, the final phylogeny must be consistent with $((i, j), k)$, with the common ancestor of i and j separated from species k . Therefore, we have the following question: Given a collection of phylogenetic constraints of the form $((i, j), k)$, does there exist a phylogeny that is consistent with all of these constraints? This problem is well-studied. Aho et al., 1981 [1] show that the tree, if it exists, can be constructed efficiently. Henzinger et al. [7] devise a more efficient algorithm for this problem, and Kannan et al., 1998 [11] consider many extensions. Another way to model this problem is by recognizing that the set of repeat insertions defines a directed and homoplasy free character set. Then, a directed perfect phylogeny, in which every character (repeat) changes from 0 to 1 (repeat insertion), exactly once in history, would also be consistent with all partial phylogenies. Recently, Pe'er et al., 2004 [16] give a near optimal algorithm to construct the directed perfect phylogeny from a partial matrix, if one exists. These algorithms only work if the data is error free, so we cannot use them directly. Instead, we use a small modification of Aho et al.'s algorithm to handle errors. The algorithm is informally described below:

The orthologous repeats are used to construct a weighted, undirected *shared-repeat* graph $G = (V, E, w)$ on n vertices. For a given repeat, consider the subset N_1 of species which clearly contain this repeat. For all pairs $(i, j) \in N_1$, increment the weight $w(i, j)$. Iterate over all repeats. At the end of this procedure, consider the subsets connected by non-zero weighted edges. Note that any pair of species (i, j) from a connected component was separated by some species k in another connected component because of a repeat that was inserted in i, j but not in k . Therefore, the correct phylogeny must put each connected component in a separate clade, each of which is connected to the root. To build the complete tree, we recurse on each connected component. While recursing on a connected component containing the subset N_c , we only consider columns which contain at least two 1's and one 0 when restricted to rows in N_c . To illustrate this, consider a case with 5 species and 3 repeats (See Figure 2(a)). The corresponding *shared-repeat graph* has two connected components $C_1 = \{A, B, C\}$, and $C_2 = \{D, E\}$. As C_2 is resolved, we recurse on C_1 . When restricted to these 3 rows, only $R2$ contains two 1's and one 0 (Figure 2(d)). The shared-repeat graph in Figure 2(e) resolves component C_1 . The final phylogeny is shown in Figure 2(f).

A representative sample of the orthologous-repeats table is shown in Table 1. In addition to 1's and 0's, the table also contains missing data, corresponding to species where the alignment was weak enough that neither presence nor absence of the repeat could be verified. (Our algorithm is robust to missing data; we discuss the causes of missing data in more detail below.) The interested reader can reconstruct the phylogeny using this

	repeat occurrence						
human	1	1	1	?	0	?	?
chimp	1	1	1	0	0	?	?
baboon	1	1	0	0	0	?	?
cat	?	0	0	0	1	1	0
dog	0	0	0	0	?	1	0
cow	0	0	0	0	1	0	1
pig	0	0	0	?	1	?	1
mouse	1	0	0	1	0	?	?
rat	?	0	?	1	0	?	?

Table 1: An orthologous-repeats table containing a sampling of repeats. Each column corresponds to a specific repeat. The symbol 1 corresponds to the presence, and 0 to the absence of that repeat. '?' indicates missing data, when neither presence, nor absence of the repeat could be confirmed.

set of repeats. In principle, a phylogeny of n species could be reconstructed using $n - 2$ repeats¹. In practice, however, extra repeats are useful in dealing with contradictory loci.

As described, the algorithm does not handle the case in which the shared-repeat graph yields a single connected component. This could happen if some repetitive elements lead to contradictory phylogenetic scenarios. Previous biological studies which used repetitive elements to elucidate phylogeny typically included a small number of contradictory loci. For example, in their analysis of Alu elements to determine primate phylogeny, Salem et al., 2003 [18] identified 7 loci with an Alu element clearly present in human and chimp genomes and clearly absent from gorilla, and 1 locus with an Alu element clearly present in human and gorilla and clearly absent from chimp; they concluded that the contradictory locus was not due to insertion homoplasy, but rather to incomplete lineage sorting: the Alu element at that locus was polymorphic at the time of divergence of gorilla from human and chimp, remained polymorphic at the time of divergence of chimp from human, and eventually became fixed in human and gorilla lineages but not in chimp. Incomplete lineage sorting and the contradictory loci they create can complicate any phylogenetic analysis, but generally should not pose a problem in phylogenetic analyses using repetitive elements, as long as a sufficiently large number of independent loci are examined (Shedlock and Okada, 2000).

However, these conflicts play a more significant role in automated analysis of thousands of repeats. Figure 3 illustrates that a strong alignment appears to exist for a SINE repeat in cat and rat, while the absence of this repeat is strongly supported in the other organisms, resulting in a repeat that resolves the (Laurasiatheria (L), Primate (P), Rodent (R)) clade as (P,(L,R)), which conflicts with the evidence of many such repeats. This conflicting edge between cat and rat is not an error, but accurately reflects the actual sequence data. Incomplete lineage sorting does not seem to be a plausible explanation for this example, as polymorphism of the presence or absence of the repeat would need to persist from the time of divergence of Rodents and Laurasiatheria through the time of divergence of cat and dog, which seems unlikely. We speculate instead that this may be a rare instance of insertion homoplasy.

It has been suggested that orientation and similarity of target sites is a signal that the repeat was inserted in a common ancestor [?]. While the presence of a conserved target site is a strong signal against insertion homoplasy, a lack of target site conservation is not a sufficient condition for it. Therefore, we use a different approach to detect and remove repeat loci with possible insertion homoplasy. First, as the repeats diverged before they were inserted, they show greater divergence. Thus, we can use the statistic (% SIMILARITY IN FLANKING REGION- %SIMILARITY IN REPEAT REGION), to decide if a location is an instance of insertion homoplasy. Second, repeat loci with insertion homoplasy are likely to conflict with many other repeats, and unlikely to be supported by other repeats. Finally, the presence of such repeats leads to a single connected component in the shared-repeat graph with the insertion homoplasy repeats being among the lowest weight edges. We iteratively remove minimum weight edges until the shared-repeat graph is no longer connected. In practice, we have found that the minimum weight is quite small, and the resulting phylogenies are robust (See RESULTS). Our method includes the following steps:

1. Identify repeats in all of the sequences.

¹A phylogeny containing unresolvable trichotomies could be reconstructed using less than $n - 2$ repeats.

```

human  GGGAAATCTCATAACTGATGCCAGAAGCACGT----- GGGAA-----ATCTCATAACTG
chimp  GGGAAATCTCATAACTGATGCCAGAAGCACGT----- GGGAA-----ATCTCATAACTG
baboon GGGAAATCTCATAACTGATGCCAGAAGCACGTTGCTCCAGAGCTAGCCAG -----
cat    GAGGAATCTCATAACTGACATCAGAAGCATATTGCTCTGAAGTAAACCAG gggcgcttggctggctcagtcagtagagcatgcaacgcttgaccttctggttg
dog    GAGGAATCTC--AACTGACATCTGAAGCATACTG-----
cow    GGGGAATCTTATAAGTGACATGAGAAGCACATTG-----
pig    GAGGAATCTCATATTGACACAAGAAGCAGATTG-----
rat    GAGGCATCCCATAGATGACGTGAGTGTCTCTCAGCCTAGAGCAG--Cag gggagctgaacaacacctagccatcagaaaatgtgactcataaccttatggttg
mouse -----
//
human  ---ATACCAGAAGCATGCTG-----CTCCAGA CCAGTGCTCCTGGTAGTGCCTCGAAAAGTGGCAGGCCACTGAACAAAGCGG
chimp  ---ATACCAGAAGCATGTTG-----CTCCAGA CCAGTGCTCCTGGTGGTGCCTTGAAAAGTGGCAGGCCACTGAACAAAGCGG
baboon -----TCCTGGTGGTGCCTTGAAAAGTGGCAGGCCGCTGAACAAAGCTG
cat    taaattcgagaacccatattgggtgcagagattacttaaaaataaaatctttaa CCAGTGCGTGTGGTGCCTCAAAAATGGGAGGCCACTGAATTAAGTGA
dog    -----CTCTAAA CCAGTGCTTGT---CTGCCTCAAAAATGGGAGGCCACAGTGT-AGATGG
cow    -----CTCCAGA CCAGGGCTCCTGGTGTGCCTCAAAAATGAGAGGCCACTGAACCAAGTAG
pig    -----CTCCAGA CCAGTGCTCCTGGTGTACCTCAAAAATGAGAGACAACCTGAACAGAGTGG
rat    ggggcaccacaacctgaggaggtgcagaggtaggctgagaacctgCTCTGAA CCAGTGCTCCTGA---TGCCTCATAAGTAAGAGACCACTCATTTAGATAG
mouse -----TCTAAA CCAGTGCCCCTGATGCT---TCGTAAGTAGGAGACCACGCATTTAAGCGG

```

Figure 3: Multiple alignment corresponding to a conflicting edge in the shared-repeat graph of 9 species with finished sequence. Repeats annotated by RepeatMasker are indicated in lower case.

2. Use a genome multiple alignment tool to compute a multiple alignment of all sequences. The specific tool used, MultiPipMaker builds a multiple alignment from $n - 1$ master slave alignments of a single sequence against all others.
3. Construct an $n \times m$ orthologous-repeats table O , where m is the number of repeats in the Master sequence.
4. Repeat with each sequence as the Master sequence to construct a complete orthologous-repeats table.
5. Remove Repeats (columns in the table) corresponding to possible instances of insertion homoplasy.
6. Construct a complete phylogeny from the orthologous-repeats table.
7. Compute Bootstrap values of the phylogeny to determine robust branches.

These steps are described in detail in the METHODS section.

Results

Species with finished sequence

We first applied our method to 9 species with finished sequence data presently available, using sequence data from the 1.5Mb 7q31 region (see Methods). We constructed an orthologous-repeats table containing XXX columns, after removal of conflicting repeats (see online Supplemental Data. MUST ADD.) The resulting shared-repeat graph is displayed in Table 2(a). After omitting edges of weight 1 or 2, this shared-repeat graph splits into two connected components: a Primate-Rodent clade (human,chimp,baboon,mouse,rat) and a Laurasiatheria clade (cat,dog,cow,pig). Reapplication of the method to these clades produces the shared-repeat subgraphs displayed in Table 2(b). The Primate-Rodent subgraph is indicative of a Primate clade (human,chimp,baboon) and a Rodent clade (mouse,rat); after omitting the edge of weight 1, the Laurasiatheria subgraph is indicative of a Carnivore clade (cat,dog) and an Artiodactyl clade (cow,pig). Finally, reapplication of the method to the Primate clade produces the shared-repeat subgraph displayed in Table 2(c), which is indicative of a human-chimp clade. Combining all of these results, we obtain the phylogenetic tree displayed in Figure 4.

A larger set of species

We subsequently applied our method to a larger set of 28 (mostly eutherian) mammals with partial sequence data available, again using sequence data from the 1.5Mb 7q31 region. The resulting phylogenetic tree is displayed

(a)

	human	chimp	baboon	mouse	rat	cat	dog	cow	pig
human		929	660	3	0	1	0	0	0
chimp			620	5	1	2	1	1	1
baboon				3	0	1	0	0	0
mouse					42	1	1	1	1
rat						0	0	0	0
cat							7	27	13
dog								6	12
cow									16
pig									

(b)

	human	chimp	baboon	mouse	rat
human		234	119	0	0
chimp			111	0	0
baboon				0	0
mouse					27
rat					

	cat	dog	cow	pig
cat		14	0	0
dog			0	1
cow				2
pig				

(c)

	human	chimp	baboon
human		55	0
chimp			0
baboon			

Table 2: Shared-repeat graph and subgraphs of 9 species with finished sequence. (a) The shared-repeat graph on all 9 species is indicative of a Primate-Rodent and Laurasiatheria clade. (b) Shared-repeat subgraphs for Primate-Rodent and Laurasiatheria clades are indicative of Primate, Rodent, Carnivore and Artiodactyl clades. (c) The shared-repeat subgraph for the Primate clade is indicative of a human-chimp clade.

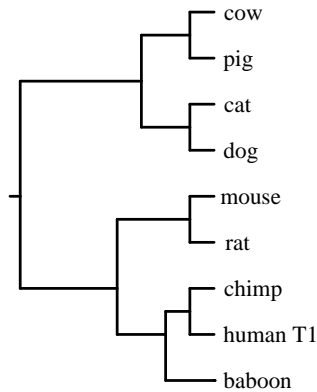


Figure 4: Phylogenetic tree of 9 species with finished sequence. NEEDS UPDATING.

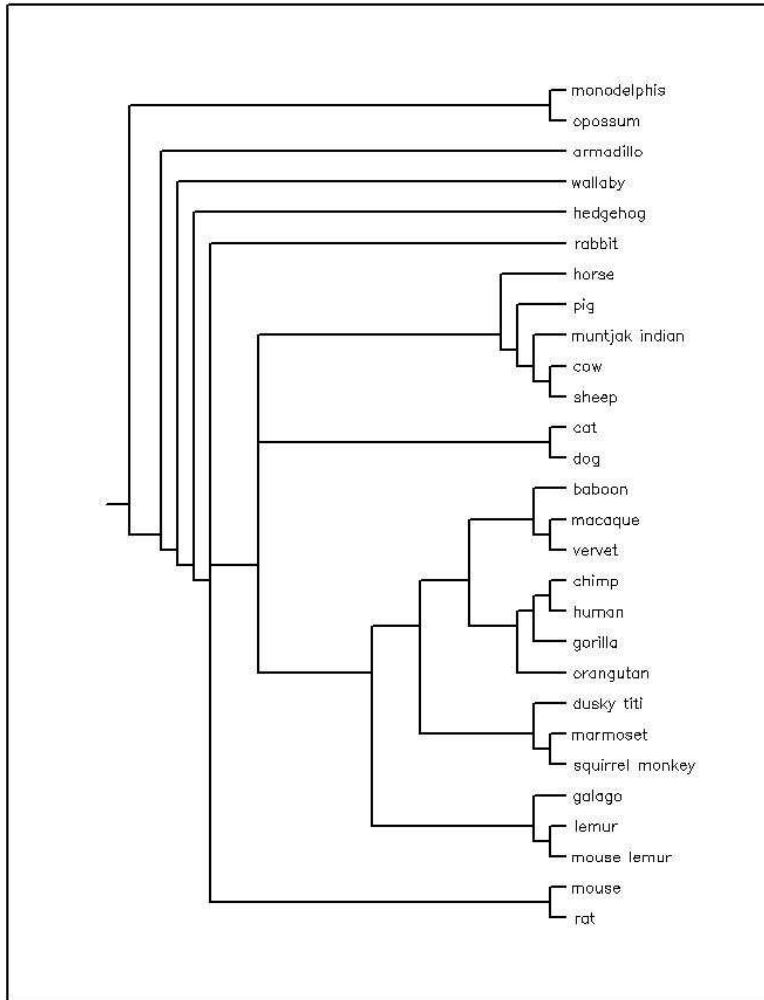


Figure 5: Phylogenetic tree of a large set of 28 species. NEEDS UPDATING. Bootstrap support values are based on 1,000 bootstrap replicates.

in Figure 5. Each node is labeled by a bootstrap support value for that clade, obtained from an analysis of 1,000 bootstrap replicates. (TO ADD: mention program Ali used, with reference?) Results for parts of the tree where previous studies reached conflicting conclusions are discussed in detail below (see DISCUSSION). Otherwise, our tree is entirely consistent with previous studies. In particular, our phylogeny of the 13 primate species in our data set agrees exactly with the widely accepted phylogeny of primates (Purvis, 1995 [17]), and nearly all primate phylogeny branches are supported by high bootstrap values. For example, we have identified hundreds of repeats which correctly separate (baboon,macaque,vervet,chimp,human,gorilla,orangutan) and (dusky titi,marmoset,squirrel monkey) from (galago,lemur,mouse lemur), and only XXX conflicting repeats which support alternate resolutions of this trichotomy. Each one of these conflicting repeats is consistent with insertion homoplasy;² the conflicting repeats are removed from the orthologous-repeats table during the conflict removal step. These numbers, and the resulting XXX% bootstrap support for the correct resolution of this trichotomy, illustrate the robustness of our approach in dealing with instances of insertion homoplasy.

²In each case, there exist two clades which contain all species with the repeat clearly present and no species with the repeat clearly absent, supporting the hypothesis of two distinct repeat insertion events in the ancestor of each clade.

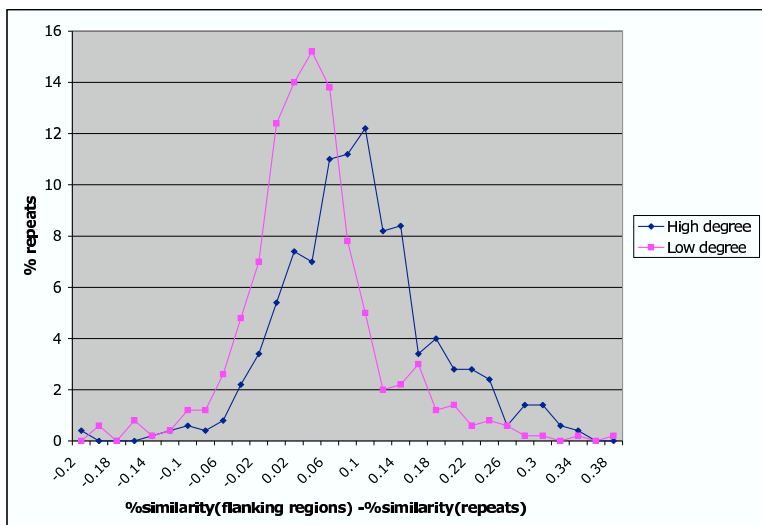


Figure 6: Distribution of the difference statistic among columns with high and low degree of incompatibility. The statistic measures the % difference in sequence similarity between flanking and repeat regions. Repeats which show incompatibility to many other repeats are likely to be instances of insertion homoplasy. These repeats show a higher degree of divergence.

Assessment of flanking region alignments

In the aforementioned analyses, the shared-repeat graph was created by tracking the presence or absence of various repeats in multiple species. In order to decide if a repeat inserted in a given organism, flanking region alignments (described in METHODS) were critical. To assess the quality of these alignments, we retrieved each upstream and downstream flanking sequence which resulted in a '1' or '0' (as defined in the APPROACH) and determined its percent identity to the reference sequence. Additionally, we randomly permuted each of these flanking sequences 10 times and determined the percent identity of the permuted sequence to the reference sequence. A histogram of these results, for our largest mammalian set, is shown in Figure 8. The original sequence overlaps with only the tail end of the permuted distribution. This supports the claim that the original flanking alignments represent significant orthology and are, therefore, useful in determining whether a repeat inserted prior to the divergence of two species.

Assessment of Conflicting Repeats

As discussed earlier, a few of the repeats are instances of insertion homoplasy, which can complicate phylogenetic analyses. Note that if there is no instance of insertion homoplasy, then each pair of columns in the orthologous repeats table must be *compatible* in the following sense: There do not exist rows that contain (0, 0), (0, 1), and (1, 0) in the two columns. Such incompatibilities are common in molecular sequence data, but should be rare for repeat insertion data. We define an *incompatibility graph* on the columns of the orthologous repeats table. Each column is a node in the graph. Two columns are connected by an edge if they are not compatible. The columns that contain an instance of insertion homoplasy should conflict with many others, and therefore, correspond to high-degree nodes. Note also that if the repeats were inserted independently, their divergence from the flanking regions should be higher than repeats that were inserted in a common ancestor of the sequence. For each of the columns in the table, we computed the difference in % similarity between the flanking regions and the repeat regions. To determine if this can be used as a statistic to detect independently inserted repeats, we looked at the distribution of this number for the 500 highest and the 500 lowest degree nodes in the incompatibility graph. See Figure 6. While the true distributions overlap, they have distinct means of 8.6% for high-degree, and 3.2% for the low-degree nodes. A *t*-test to determine if the means were equal gave a P-value of $1.1E - 32$. Based on this, we remove all columns for which the difference is 7.5% or higher. This columns removal procedure still retains some instances of insertion homoplasy, but these show up as high degree nodes in the incompatibility graph. We constructed *incompatibility graphs* for the 9 organism data-set as well as the complete 28-organism

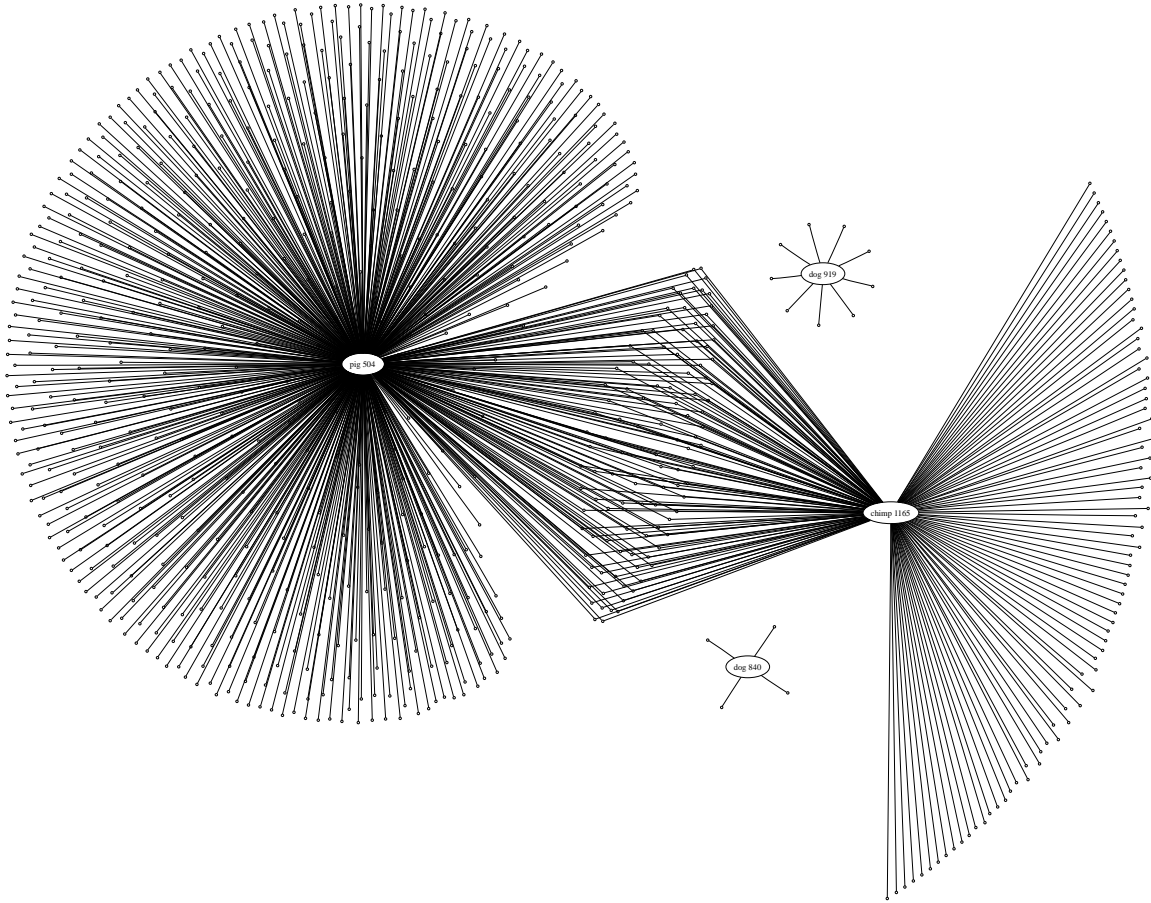


Figure 7: Incompatibility graph for the 9 organism data-set. Note that all 821 incompatibility edges are incident on one of 4 nodes.

dataset. For the 9 primate species, there were a total of 1101 columns, of which 717 nodes were connected by 821 edges. However, all edges are incident to only 4 nodes, and removing them would make the graph conflict free (Figure ??). The 28 organism data-set has similar characteristics. There were a total of XXX conflicts involving XXX columns. However, removal of the highest degree XXX columns eliminates all conflicts. In our method, we iteratively remove the highest degree node until no-conflict remains

Discussion

The phylogeny of eutherian mammals is largely unresolved, as there are many instances where previous studies reach conflicting conclusions (Amrine-Madsen et al., 2003 [2]). An example is the placement of Rodents, i.e. resolution of the trichotomy between Rodents, Primates and Laurasiatheria (which include Carnivores, Artiodactyls, horse, etc.). Some studies report a Primate-Rodent clade (Murphy et al., 2001 [14]; Amrine-Madsen et al., 2003 [2]) while others report the divergence of Rodent from a Primate-Laurasiatheria clade (Arnason et al., 2002 [3]; Misawa and Janke, 2003 [12]). In our analysis, we identified XXX repeats separating Primates and Rodents from Laurasiatheria; we report a Primate-Rodent clade with XXX% bootstrap support. Our results agree with Thomas et al, 2003 [5], who identified 3 repetitive elements which support a Primate-Rodent clade. (TO ADD: DETAILS ON WHY WE DIDN'T FIND ALL 3? OR 4?).

Another interesting example is the placement of horse in the phylogenetic tree. Early studies of horse, Carnivores and Artiodactyls reported a horse-Artiodactyl clade (Graur et al., 1997 [6]), while more recent studies report a horse-Carnivore clade (Murphy et al., 2001 [14]; Arnason et al., 2002 [3]; Amrine-Madsen et

al., 2003 [2]). In our analysis, we identified XXX repeats separating horse and Carnivores from Artiodactyls; we report a horse-Carnivore clade with XXX% bootstrap support. It is notable that our program discovers the same L1MA9 repeat that Schwartz et al., 2003 [19] used to establish the horse-Carnivore clade.

The placement of rabbit in the phylogenetic tree has been the subject of considerable debate. The resolution of the trichotomy between rabbit, Primates and Laurasiatheria has been variously reported as (Laurasiatheria,(rabbit,Primate)) (Murphy et al., 2001 [14]; Amrine-Madsen et al., 2003 [2]), or (Primate,(rabbit,Laurasiatheria)) (Arnason et al., 2002 [3]), or (rabbit,(Primate,Laurasiatheria)) (Misawa and Janke, 2003 [12]). We identified XXX repeats separating rabbit and Primates from Laurasiatheria, strongly supporting (Laurasiatheria,(rabbit,Primate)). We further note that the Murphy et al. and Amrine-Madsen et al. studies confirm the Glires hypothesis of a rabbit-Rodent clade, while the Arnason et al. and Misawa and Janke studies reject the Glires hypothesis. Although we neither confirm or reject the Glires hypothesis, due to our unresolved (rabbit,Rodent,Primate) trichotomy, our rabbit results do reject the two studies rejecting the Glires hypothesis.

Our placement of hedgehog inside the Laurasiatheria clade and armadillo outside the clade containing Primates, Rodents and Laurasiatheria is consistent with Murphy et al., 2001 [14] and Amrine-Madsen et al., 2003 [2], but inconsistent with Arnason et al., 2002 [3], which places armadillo inside the Laurasiatheria clade and hedgehog outside the clade containing Primates, Rodents and Laurasiatheria. It is notable that all of our results are consistent with Murphy et al., 2001 [14] and Amrine-Madsen et al., 2003 [2], despite having been obtained via entirely unrelated means.

Overall, we consider our generation of an accurate phylogenetic tree of 28 mammalian species from only 1.5Mb of sequence data to be an extremely encouraging result. Our bootstrap support values are considerably lower than other studies, in which nearly all bootstrap support values exceed 95% (Murphy et al., 2001 [14]; Arnason et al., 2002 [3]; Amrine-Madsen et al., 2003 [2]; Misawa and Janke, 2003 [12]). However, these conflicting studies, each supported by high bootstrap values, cannot all be correct! Indeed, the Murphy et al. study has already been implicated as suffering from exaggerated support values (Misawa and Nei, 2003 [13]). We anticipate that with an increased amount of sequence data, we will obtain higher bootstrap support values and resolve the unresolved trichotomy between rabbit, Rodents and Primates. In addition, it may be possible to filter out the rare cases of insertion homoplasy we encounter by identifying and comparing target site duplications which occur at the site of SINE/LINE insertions; automating this procedure is an important direction of our ongoing research. In conclusion, we are optimistic that going forward, our approach will conclusively resolve eutherian phylogeny and other phylogenetic problems.

Methods

Data

Sequences were collected from the NIH Intramural Sequencing Center (NISC), Comparative Vertebrate Sequencing project (NIH Intramural Sequencing Center, 2004 [4]). The set of sequences used were from target reference 7q31, Encode Name Enm001, a region approximately 1.5 Mb in size. The sequences themselves ranged from 1.2Mb (pig) to 2.3 Mb (marmoset). To obtain preliminary data for organisms with unpublished 7q31 sequence, the entire 7q31 data set was scanned. Genbank files, for accession numbers from that data set, were retrieved; from these files the corresponding sequences were extracted. Contigs were joined to one another via overlap information embedded within each genbank file. Note that the concatenated sequences are not complete, and the alignment introduces gaps.

Repeat masking

Repeat were masked by RepeatMasker (Smit and Green, 2004 [23]) using a library of repeat derived from the set of mammalian repeats in Repbase (Jurka, 1998 [9], 2000 [10]). Repeat masking for the primate data-set was performed utilizing the -prim flag. For the 9 organism data-set, Repeat masked sequences were obtained from the Thomas et al. supplemental data (Thomas et al, 2003 [5]). For the remaining species, the library of repeats consisted of a concatenation of 23 complete RepeatMasker libraries, derived from RepBase9.06. RepeatMasker was run on each sequence using fast settings for speed and sensitivity.

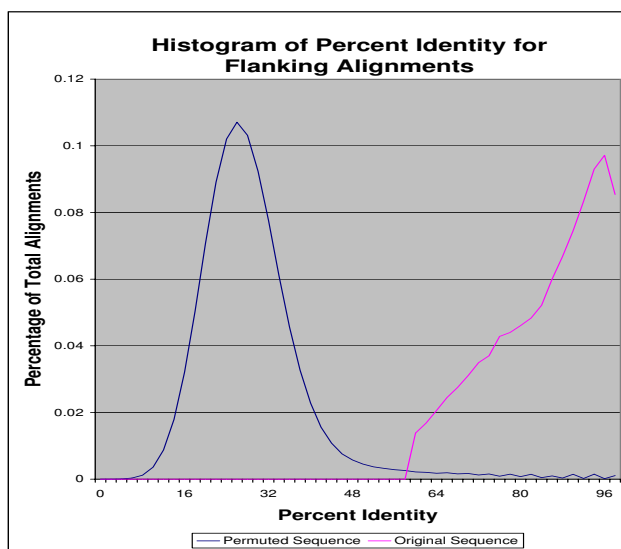


Figure 8: Histogram of percent identity for original and permuted sequences in flanking regions. The original sequence distribution appears skewed for two reasons. First, alignments cannot have percent identity greater than 100 percent (right side). Second, the cutoff for acceptable alignments required percent identity greater than 60 percent (left side).

Multiple Alignments

Multiple Alignments were generated via MultiPipMaker (Schwartz et al., 2003 [19]). MultiPipMaker is a tool for aligning multiple, long (Mb size) genomic DNA sequences quickly and with good sensitivity. The program takes as input a single reference sequence and multiple secondary sequences; additionally, one of the following options must be selected: show all matches, chaining, or single coverage. Alignments are first computed by pairwise Blatz alignments, and subsequent refinements, between the reference organism and each secondary sequence. MultiPipMaker then looks at sub-alignments within the global multiple alignment to see if modifications can be made to improve the overall score of the alignment. Since our sequences were variable in length and since the alignments generated by MultiPipMaker are most relevant as alignments to the reference sequence, it was necessary to rerun MultiPipMaker with each organism as reference sequence. Thus, for our n organisms we generated n multiple alignments (the ordering of the secondary sequences was irrelevant). Moreover, the chaining option was selected to avoid duplicate matches caused by the "show all matches" option, i.e. a single region in the reference sequence aligning to two regions in a secondary sequence. This option was selected over single coverage because: 1) the secondary sequences were assumed to be contiguous, 2) the comparisons were made with a single strand of the secondary sequence, and 3) the order of conserved regions was assumed identical in the two sequences (Schwartz et al., 2003 [20]).

Identifying Orthologous Insertions

For each MultiPip alignment, our algorithm iterated through the reference organism's RepeatMasker generated repetitive element list, ignoring all non-transposable element based repeats (such as LTRs and simple repetitive repeats). For each considered repeat, the corresponding orthologous region in each secondary organism, as well as a 50 nucleotide upstream and downstream flanking region, was retrieved. For a repeat to be considered present in a secondary organism's sequence it must strongly align in the repeat region and within both flanking regions. For a repeat to be considered absent from a secondary organism's sequence it must strongly align within both flanking regions, while gapping out the repeat region. If neither set of requirements are satisfactorily met, the

presence of the repeat is considered uncertain for that secondary organism's sequence. Using this methodology, an orthologous-repeats table is generated. Each row of the repeat represents an organism, and each column represents a given repeat. The presence of a repeat is indicated with a '1', the absence with a '0', and uncertainty with a '?'.

Graph Generation and Tree Building Algorithm

The following procedure is an implementation of the algorithm presented by Aho et al., with modifications for dealing with conflicting edges (Aho et al. 1981 [1]).

1. A subset of the orthologous-repeats table is created, in which only "relevant" rows (organisms) are considered (initially all rows, since all organisms are being considered). Within this subset of rows, only those columns in which at least two rows have a 1 and one row has a 0 are considered.
2. Utilizing this subset of the original repeat occurrence table, a graph is created by iterating through the columns. If two rows both have a 1 in given column an edge of weight 1 is created between the two corresponding organisms. If an edge already exists between those two organisms its weight is incremented by 1.
3. Multiple connected components are sought within the graph. If the graph contains a single connected component, weak edges must be eliminated. This is accomplished by removing edges, beginning with those of weight 1 and incrementally removing edges of greater weight, until multiple connected components arise.
4. Steps 1-3 are repeated on each connected component containing greater than two organisms. The "relevant" rows in each run will be the clade of organisms within the connected component.

References

- [1] A.V. Aho, S.Y. Sagiv, T.G. Szymanski, and J.D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *Siam Journal of Computing*, 10(3):405–421, 1981.
- [2] H. Amrine-Madsen, K. Koepfli, R.K. Wayne, and M.S. Springer. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Molecular Phylogenetics and Evolution*, 28:225–40, 2003.
- [3] U. Arnason, J.A. Adegoke, K. Bodin, E.W. Born, Y.B. Esa, A. Gullberg, M. Nilsson, R.V. Short, X. Xu, and A. Janke. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proceedings of the National Academy of Sciences*, 99:8151–8156, 2002.
- [4] NIH Intramural Sequencing Center. Comparative Vertebrate Sequencing. http://www.nisc.nih.gov/open_page.html?/projects/comp_seq.html.
- [5] J.W. Thomas et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.
- [6] D. Graur, M. Gouy, and L. Duret. Evolutionary Affinities of the Order Perissodactyla and the Phylogenetic Status of the Superordinal Taxa Ungulata and Altungulata. *Molecular Phylogenetics and Evolution*, 7:195–200, 1997.
- [7] M. Henzinger, V. King, and T. Warnow. A fast algorithm for constructing rooted trees from constraints. (Unpublished Manuscript).
- [8] D.M. Hillis. SINEs of the perfect character. *Proceedings of the National Academy of Sciences*, 96:9979–81, 1999.
- [9] J. Jurka. Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, 8:333–337, 1998.

- [10] J. Jurka. Repbase Update: a database and an electronic journal of repetitive elements. *TRENDS in Genetics*, 9:418–420, 2000.
- [11] S. Kannan, T. Warnow, and S. Yooseph. Computing the local consensus of trees. *SIAM Journal of Computing*, 27(6):1695–1724, 1998.
- [12] K. Misawa and A. Janke. Revisiting the Glires concept—phylogenetic analysis of nuclear sequences. *Molecular Phylogenetics and Evolution*, 28:320–327, 2003.
- [13] K. Misawa and M. Nei. Reanalysis of Murphy et al.’s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. *Journal of Molecular Evolution*, 57 Suppl 1:S290–6, 2003.
- [14] W.J. Murphy, E. Eizirik, S.J. O’Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. de Jong, and M.S. Springer. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294:2348–51, 2001.
- [15] M. Nikaido, A. Rooney, and N. Okada. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. *Proceedings of the National Academy of Sciences*, 96:10261–66, 1999.
- [16] I. Pe’er, T. Pupko, R. Shamir, and R. Sharan. Incomplete Directed Perfect Phylogeny. *Siam Journal of Computing*, 33(3):590–607, 2004.
- [17] A. Purvis. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London Series B*, 348:405–421, 1995.
- [18] A.H. Salem, D.A. Ray, J. Xing, P.A. Callinan, J.S. Myers, D.J. Hedges, R.K. Garber, D.J. Witherspoon, L.B. Jorde, and M.A. Batzer. Alu elements and hominid phylogenetics. *Proceedings of the National Academy of Sciences*, 100:12787–91, 2003.
- [19] S. Schwartz, L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, NISC Comparative Sequencing Program, E.D. Green, R.C. Hardison, and W. Miller. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Research*, 31(13):3518–3524, 2003.
- [20] S. Schwartz, Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker- A Web Server for Aligning Two Genomic DNA Sequences. *Genome Research*, 10(4):577–86, 2003.
- [21] A.M. Shedlock and N. Okada. SINE insertions: powerful tools for molecular systematics. *BioEssays*, 22:148–60, 2000.
- [22] M. Shimamura, H. Yasue, K. Ohshima, H. Abe, H. Kato, T. Kishiro, M. Goto, I. Munechika, and N. Okada. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature*, 388:666–70, 1997.
- [23] A. Smit. RepeatMasker. <http://www.repeatmasker.org/>.
- [24] K. Takahashi, M. Nishida, M. Yuma, and N. Okada. Retroposition of the AFC family of SINEs before and during the adaptive radiation of cichlid fishes in Lake Malawi and related inferences about phylogeny. *Journal of Molecular Evolution*, 53:496–507, 2001.
- [25] K. Takahashi, Y. Terai, M. Nishida, and N. Okada. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. *Molecular Biology and Evolution*, 18:2057–66, 2001.
- [26] Y. Terai, K. Takahashi, M. Nishida, T. Sato, and N. Okada. Using SINEs to probe ancient explosive speciation: "hidden" radiation of African cichlids? *Molecular Biology and Evolution*, 20:924–30, 2003.