Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis

John E. McCormack, Brant C. Faircloth, Nicholas G. Crawford, et al.

Genome Res. 2012 22: 746-754 originally published online December 29, 2011
Access the most recent version at doi:10.1101/gr.125864.111

Supplemental Material
http://genome.cshlp.org/content/suppl/2012/01/03/gr.125864.111.DC1.html

References
This article cites 62 articles, 36 of which can be accessed free at:
http://genome.cshlp.org/content/22/4/746.full.html#ref-list-1

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

Copyright © 2012 by Cold Spring Harbor Laboratory Press
Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis

John E. McCormack,1,8 Brant C. Faircloth,2 Nicholas G. Crawford,3 Patricia Adair Gowaty,4,5 Robb T. Brumfield,1,6 and Travis C. Glenn7

1Museum of Natural Science, Louisiana State University, Baton Rouge, Louisiana 70803, USA; 2Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA; 3Department of Biology, Boston University, Boston, Massachusetts 02215, USA; 4Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panamá, República de Panamá; 5Institute of the Environment, University of California, Los Angeles, California 90095, USA; 6Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA; 7Department of Environmental Health Science, University of Georgia, Athens, Georgia 30602, USA

Phylogenomics offers the potential to fully resolve the Tree of Life, but increasing genomic coverage also reveals conflicting evolutionary histories among genes, demanding new analytical strategies for elucidating a single history of life. Here, we outline a phylogenomic approach using a novel class of phylogenetic markers derived from ultraconserved elements and flanking DNA. Using species-tree analysis that accounts for discord among hundreds of independent loci, we show that this class of marker is useful for recovering deep-level phylogeny in placental mammals. In broad outline, our phylogeny agrees with recent phylogenomic studies of mammals, including several formerly controversial relationships. Our results also inform two outstanding questions in placental mammal phylogeny involving rapid speciation, where species-tree methods are particularly needed. Contrary to most phylogenomic studies, our study supports a first-diverging placental mammal lineage that includes elephants and tenrecs (Afrotheria). The level of conflict among gene histories is consistent with this basal divergence occurring in or near a phylogenetic “anomaly zone” where a failure to account for coalescent stochasticity will mislead phylogenetic inference. Addressing a long-standing phylogenetic mystery, we find some support from a high genomic coverage data set for a traditional placement of bats (Chiroptera) sister to a clade containing Perissodactyla, Cetartiodactyla, and Carnivora, and not nested within the latter clade, as has been suggested recently, although other results were conflicting. One of the most remarkable findings of our study is that ultraconserved elements and their flanking DNA are a rich source of phylogenetic information with strong potential for application across Amniotes.

[Supplemental material is available for this article.]
Available evidence indicates that several mammalian divergences occurred extremely rapidly. The basal divergence among the three placental mammal superorders—Afrotheria (e.g., elephants, tenrecs), Xenarthra (e.g., sloths, armadillos), and BHxotheria (e.g., carnivores, bats, rodents, and primates)—is thought, based on fossil-calibrated divergence times, to have been completed within 2 to 5 million years (Springer et al. 2003; Murphy et al. 2007). Even more spectacular is the explosive radiation of orders within the Laurasiatheria (e.g., dogs, bats, horses, cows, dolphins), with several basal splits occurring within 1 to 3 million years of one another (Springer et al. 2003). As predicted by coalescent theory (Degnan and Salter 2005), these are the same divergences where individual gene trees give conflicting answers about evolutionary relationships (Nishihara et al. 2006, 2009; Murphy et al. 2007; Churakov et al. 2009), underscoring the need for an approach that incorporates a model of the process at the root of this discord—coalescent stochasticity. This is especially relevant given the recent discovery of a phylogenetic anomaly zone where rapid speciation results in the most common gene tree being in conflict with the true species history (Degnan and Rosenberg 2006), a situation that will mislead phylogenetic inference unless coalescent stochasticity is taken into account (Kubatko and Degnan 2007).

Here, we outline a generalizable framework for resolving phylogenetic discord using a novel class of markers: ultraconserved elements (UCEs) and flanking DNA (Supplemental Fig. S1). UCEs were first described in humans (Bejerano et al. 2004) and have since been found in other tetrapods (Stephen et al. 2008) and more distantly related species such as worms and yeast (Siepel et al. 2005). Many UCEs are thought to play an important role in regulation and development (Sandelin et al. 2004; Woolfe et al. 2004), and their function is a topic of intense study. There are several reasons why UCEs are expected to be good phylogenomic markers. A recent review identified key problems for phylogenomic analysis, among them incorrect identification of orthologs and saturation of nucleotide substitutions such that multiple substitutions at a given base position obscure true phylogenetic signal (Philippe et al. 2011). Addressing orthology, vertebrate UCEs have little overlap with most types of paralogous genes (Derti et al. 2006) and are found in largely transposon-free parts of the genome (Simons et al. 2006). With regard to saturation, UCEs and flanking sequence are expected to evolve slowly compared with other types of DNA, potentially making saturation less prevalent than with other marker types, a hypothesis we assess below.

After locating UCEs in amniotes and anchoring them in mammal genomes, the second part of our approach tackles the problem of coalescent stochasticity by integrating data from individual loci into recently developed phylogenetic algorithms for estimating species histories from collections of discordant gene histories (Liu et al. 2009). One of the major hurdles to estimating species trees from phylogenomic data has been the reliance of analytical methods on computationally intensive explorations of parameter space (Edwards et al. 2007), which become impractical with large phylogenomic data sets. We use a recent coalescent-based species-tree method based on ranks of coalescent events (STAR) (Liu et al. 2009) to capture relevant information about the species history from gene trees. STAR produces an analytical solution to the species tree but shows similar performance to more computationally intensive Bayesian species-tree methods (Liu et al. 2009). STAR is also robust to violations of a molecular clock, so it is particularly well suited to UCEs, which are likely to show deviations from molecular rate uniformity among taxa (for more detail on available species-tree methods, see Methods).

**Results**

**Identification of UCEs in placental mammals**

We used the 100% conserved regions of reptilian (including avian) UCEs to design 2560 in silico probes, which we aligned to existing mammal genomes (Supplemental Table S1). We excised UCE regions and 1000 bp of variable flanking sequence, which we hereafter refer to as “orthologous loci” (we expected few paralogs, but nonetheless screened for paralogs by removing probes with matches to multiple genomic locations). We minimized missing data by requiring all taxa in a data set to have all loci (although within a locus we permitted some missing data; average 6.5% of base pairs per alignment, mostly sequence at the margins of loci in outgroup taxa). Consequently, data sets with more taxa had fewer loci.

**Construction of phylogenomic data sets**

To explore the extremes of genomic and taxonomic coverage, we assembled two general-purpose data sets with different levels of genomic and taxonomic inclusion (Supplemental Table S1). One data set focused on high taxonomic coverage with 183 loci (total 94,607 bp) and 29 taxa. The other data set had high genomic coverage with 917 loci (total 908,702 bp) found in 19 taxa. To address two outstanding questions in placental mammal phylogeny that involve particularly rapid speciation, we also constructed two specialty data sets that aimed for high genomic coverage. We used the first specialty data set, 591 loci comprising six taxa and one outgroup, to assess the basal relationships of placental mammal superorders. We used the second specialty data set, 683 loci comprising six taxa and one outgroup, to investigate relationships within the rapidly radiating Laurasiatheria, focusing on the phylogenetic position of bats.

Poor taxon sampling can play an important role in phylogenetic inference (Philippe et al. 2011), although recent research by Wiens and Morrill (2011) has called into question some of the more dire predictions of poor taxonomic sampling that were found in a previous simulation study (Lemmon et al. 2009). As with all studies focusing on genome-scale data, our data sets contained many more loci than species. However, our most taxonomically inclusive data set, comprising 29 species, is the largest to date using information from vertebrates with readily available whole-genome information. In cases in which we used fewer species, we selected taxa to break up particularly long branches and mitigate the effects of long-branch attraction (i.e., we selected the two most divergent species within a group, e.g., two deeply divergent bat species).

**Variability and saturation of core UCE and flanking DNA compared with exons**

UCE loci had an average length of 400–750 bp depending on the data set (Supplemental Fig. S2) and were generally separated by wide physical distances (>2 Mb) (Supplemental Fig. S3), indicating that UCE loci were unlikely to be physically linked and were therefore likely to be segregating independently at the timescales considered. One basic feature of UCE loci is that they show increasing variability moving away from the core UCE toward the flanking regions (Stephen et al. 2008). Even though we defined UCE probe regions as 100% conserved between lizard and bird, there was usually some variability in the core UCE section in the other taxa included in the data sets. Analysis of a subset of loci ($n = 20$) from the 183-locus data set indicates that both core and flanking UCE regions contain variability and that variability is higher in flanking regions (Fig. 1A). Additionally, flanking regions
had a higher ratio of informative sites to variable sites than the core UCE. Both core and flanking UCE regions had lower variability and a lower ratio of informative sites than the 20 loci analyzed for deep-level mammal phylogeny by Springer et al. (2007). However, this difference was largely driven by high variability at the third position of codons, whereas variability in first and second position codons is more similar to UCE flanking regions. (Fig. 1A). Additionally, core UCEs and flanking regions had significantly lower saturation indexes compared with those of exons, especially at third positions of codons (Fig. 1B), although no marker or codon position showed severe effects of saturation according to the significance tests used by Xia et al. (2003) (Supplemental Fig. S4).

Phylogeny of placental mammals from UCE-fueled species trees

Using a method of species-tree analysis (STAR) (Liu et al. 2009) in which a species history is estimated from independent, often discordant gene histories, we recovered two species trees from our general-purpose data sets that were wholly concordant with one another (Fig. 2A; Supplemental Figs. S5, S6) despite the fact that individual gene histories showed considerable discordance (Fig. 2B; Supplemental Fig. S7). Notwithstanding topological conflict at the gene level, species-tree bootstrap replicates largely agreed (Fig. 2C). Bootstrap support values for the high genomic coverage data set tended to be higher (Fig. 2A). Topologies from the STAR trees largely agreed with topologies produced from Bayesian analysis of the same data sets where genes were concatenated and assumed to share the same evolutionary history (Supplemental Figs. S5, S6). One major difference was that, in the Bayesian concatenation analysis, the tree shrew grouped with Glires instead of with primates. The STAR tree, on the other hand, placed tree shrew in its accepted position as an outgroup to primates (Janecka et al. 2007).

Assessing the need for species trees with coalescent simulation

Next, using a coalescent simulation, we assessed the specific need for a species-tree approach to resolve the basal divergence of placentals. Simulating gene histories on several possible species histories reflecting current knowledge about early placentals (Bininda-Emonds et al. 2007) revealed that discord among gene histories is expected to be high over a large span of realistic demographic values for generation time, population size, and divergence time (Fig. 3). Plotting observed values of gene-tree discord from a recent study (Table 1; Nishihara et al. 2009) on this theoretical state-space suggests that the basal divergence event exists in a region of especially high gene-tree discord and may even lie in a phylogenetic anomaly zone where concatenation will mislead phylogenetic inference (Fig. 3). Values of gene-tree discord estimated from our own data (Table 1) are even higher and thus place the divergence event even closer to the anomaly zone.

UCE species tree supports Afrotheria as the first-diverging placental mammal lineage

The STAR tree from the 683-locus data set created expressly to explore the basal relationships of placentals indicated that Afrotheria was the first superorder to diverge, with 64% bootstrap support (Fig. 4A). Alternate topologies had less than half the bootstrap support: 31% of bootstrap replicates supported a sister relationship between Afrotheria and Xenarthra, and 5% supported a scenario with Xenarthra diverging first (Table 1). As expected based on the coalescent simulations, gene trees were highly discordant for the basal relationships. Of 591 gene trees, 433 (73%) were resolved for the relationships between Afrotheria, Xenarthra, and Boreotheria. Of the resolved gene trees, a plurality supported Afrotheria as diverging first (26% compared with 25% for Afrotheria and Xenarthra as sister taxa and 23% for Xenarthra diverging first) (Table 1). In contrast to the STAR tree, the Bayesian tree built from concatenated data supported Afrotheria and Xenarthra as sister taxa with high PP.

UCE species tree and the phylogenetic placement of bats

With respect to the phylogenetic position of bats within the Laurasia, our highest genomic coverage data set produced a STAR tree that supports bats as sister to a group containing Perissodactyla, Cetartiodactyla, and Carnivora, with 64% bootstrap support (Fig. 4B). The Bayesian tree with concatenated data produced an unusual topology with high PP for all nodes showing bats grouping sister to Perissodactyla, with this group sister to a monophyletic group containing Cetartiodactyla and Carnivora.

Discussion

Accurate phylogenomic inference has two important facets: estimating gene histories and inferring species histories from gene histories. The phylogenomic framework we use here addresses both of these challenges to infer the evolutionary history of placental mammals. With regard to gene trees, we introduce a novel class of genetic marker anchored by ultraconserved elements (UCEs) and show that UCE flanking regions have similar variability to first

**Figure 1.** Variability and saturation of UCE core and flanking regions compared to exons. (A) High variability in exons from Springer et al. (2007) is driven largely by third positions of codons, whereas variability in first and second position codons is more similar to UCE flanking regions. (B) UCEs have low saturation indexes, whereas saturation is highest among third positions of codons of exons. Box plots show the mean (line within box), 25th to 75th percentiles (box), fifth to 95th percentiles (whiskers), and outliers (dots).
and second base positions of codons in exons (if slightly less information per locus) and significantly less saturation (Fig. 1B). Most of the information content of exons lies in the third positions of codons (Fig. 1A). UCEs have less informative variability (the same is true of first/second codon positions), but core UCEs and flanking regions also have much lower saturation scores. So although the lower variability of UCEs does not offer a strong advantage over exons, low saturation does make UCEs appealing markers for inferring gene trees for ancient evolutionary divergences, where multiple hits can create homoplasy, which obscures phylogenetic signal (Whitfield and Lockhart 2007).

With regard to species trees, we note that species-tree analysis, like the STAR method we use, can be conducted using most types of genetic markers. Retroposons are an exception, although there is no reason why the theoretical framework could not be extended to accommodate them. However, the principal benefit of using UCEs in species-tree analysis, compared with other types of markers like exons, is that the core region is highly conserved. This conservation allows UCEs to be rapidly characterized in high numbers in a broad array of species across the Tree of Life and with few of the problems of paralogy that plague other types of phylogenomic markers (Philippe et al. 2011). The sheer quantity of discrete UCEs shared among evolutionarily distant species thus addresses the issue of coalescent stochasticity, which requires many independent loci, in a way that would be more difficult to do with other marker types like exons. Additionally, purifying selection on UCEs (Katzman et al. 2007) could reduce the incidence of incomplete lineage sorting during short speciation intervals by reducing the effective population size (e.g., Hobolth et al. 2007; McVicker et al. 2009), which would also be an advantage over other marker types, but this remains to be tested.
UCEs may provide more loci for specific research questions. For example, Stephen et al. (2008) found a large pool of UCEs from mammal alignments. This set of UCEs was largely nonoverlapping with UCEs from this study: We found 897 UCEs shared between data sets, which corresponds to 7% of the Stephen et al. (2008) UCE loci or 38% of the UCE loci identified here. Our study thus represents a major, novel source of UCEs in amniotes (in addition to those of Jans et al. 2011). When we processed the UCEs from Stephen et al. (2008) using our pipeline for probe design and alignment (including the requirement that all loci are present in all 29 species), we identified 261 additional loci, which we analyzed separately and in combination with our 183 locus data set (Supplemental Fig. S8). The combined data set of 444 (183 + 261) loci resulted in topologies nearly identical to our earlier analyses, but with higher bootstrap support at most nodes (Fig. 2A; Supplemental Fig. S8).

When we adapted our probe design and alignment pipeline for exons (min-
Results from our coalescent simulation provide an explanation for why retroposon studies have found highly heterogeneous signal among gene trees bearing on the divergence of placental mammal superorders (Murphy et al. 2007; Churakov et al. 2009; Nishihara et al. 2009). Our finding that the divergence events lie close to or within a phylogenetic anomaly zone (Degnan and Rosenberg 2006) also cautions that concatenated data sets are probably not appropriate for answering this particular question, and concatenated data sets may even produce misleading results. We note that although retroposons are less likely to show homoplasny than DNA sequence data (although they may not always be free of homoplasny [Han et al. 2011]) and therefore are excellent markers for accurately recording gene trees (Shedlock et al. 2004), retroposons are not immune to the effects of coalescent stochasticity. Unfortunately, retroposons are also rarely found in high enough numbers to inform very rapid speciation events if, as coalescent simulation studies suggest, greater than 500 loci may be necessary to accurately estimate divergences in the anomaly zone (Liu et al. 2009). For comparison, the number of loci used in studies investigating placental mammal divergence with retroposons has ranged from the single digits (Kriegs et al. 2006; Nishihara et al. 2006) to as many as 68 (Nishihara et al. 2009).

Our results are also relevant to one of the most enduring mysteries of mammalian phylogenetics: the evolutionary affinities of bats (Chiroptera) in the superorder Laurasiatheria. Twenty years ago, even the monophyly of Chiroptera was debated, and most researchers thought bats’ evolutionary affinities were more with primates (Novacek 1992). Later molecular phylogenies rejected this hypothesis and defined four major groups of mammals, with bats placed in the Laurasiatheria (Madsen et al. 2001) among carnivores, some ungulates, and some insectivores, although the rapid radiation of this group made further phylogenetic resolution difficult. Recently, a surprising clade (Pegasoferae) uniting Chiroptera with Carnivora (e.g., dog) and Perissodactyla (e.g., horse) to the exclusion of Cetartiodactyla (e.g., cow, alpaca, dolphin) emerged from a study of retroposons (Nishihara et al. 2006). However, this result was based on only three retroposon gene trees, with one retroposon supporting a conflicting topology. Our species tree based on 693 loci lends some support to a more traditional placement of Chiroptera as sister to a clade containing Carnivora, Perissodactyla, and Cetartiodactyla, not nested within it (Fig. 4B). At 64%, bootstrap support for this topology was not high, and the 183-locus and 444-locus STAR trees supported an arrangement favoring Pegasoferae, albeit with even lower bootstrap support (Fig. 2A). The mixed support from various data sets and analytical techniques suggests that basal relationships in the Laurasiatheria are a particularly difficult phylogenetic problem that will likely require even more loci and taxa to address. Our results do not find strong support for Pegasoferae, but neither do they find

---

**Figure 4.** Species trees and concatenated trees from high genomic coverage data sets for two rapid radiations in placental mammals. (A) Species tree from a 591-locus analysis identifies Afrotheria as the first-diverging lineage of placental mammals, whereas alternate topologies had less than half the bootstrap support (Table 1). A Bayesian analysis based on concatenated data places Afrotheria and Xenarthra together with high PP. Data on gene-tree discordance from Figure 2 suggest that this may be because the basal divergence of placental mammals lies close to the phylogenetic anomaly zone. (B) Species tree of taxa in the Laurasiatheria based on 683 loci places bats (Chiroptera) in a traditional location sister to Perissodactyla, Cetartiodactyla, and Carnivora. Bayesian analysis based on concatenated data produced an unusual tree with bats grouping sister to Perissodactyla, but with Carnivora grouping with Cetartiodactyla. The species tree and concatenated analysis of the 183 locus data set produced a different topology, more supportive of the hypothesized clade Pegasoferae (see text), suggesting that a robust understanding of this divergence event will require further investigation incorporating additional taxa and loci. Note that STAR trees do not contain branch length information.
strong support for any particular placement of bats with the Laurasiatheria.

The overall approach to inferring phylogenies using UCEs is especially promising because the in silico probes used in our study can be adapted easily to an in vitro design to capture DNA (Gniške et al. 2009) from virtually any species across the amniote Tree of Life. For example, probes synthesized from the in silico set we designed successfully captured >1000 loci in various reptile species, including sufficient flanking sequence to resolve deep-level phylogenetic relationships (Faircloth et al. 2012). Similar probe sets could be designed for other phylogenetic groups (Spiel et al. 2005). Increasingly sophisticated hierarchical methods for bar-coding individuals (Kenny et al. 2011) will enable targeted capture and sequencing of orthologous DNA from many individuals at hundreds or thousands of loci in a partial, massively parallel sequencing run without the laborious intermediate steps of marker discovery, variability screening, individual PCRs, and haplotype phasing. We envision that this approach, when applied to non-model organisms, could stimulate a shift in the way researchers collect and analyze broad-scale phylogenetic data, which will build from and complement whole-genome data produced by the Genome 10K Project (http://genome10k.soe.ucsc.edu/).

Methods

Identification of UCEs

We identified ultraconserved elements (UCEs) by screening whole-genome alignments of the chicken (Gallus gallus) and Carolina anole (Anolis carolinensis) prepared by the UCSC genome bioinformatics group using a custom Python script to identify runs of at least 60 bases having 100% sequence identity. We stored metadata for these regions in a relational database (RDB). Because the zebra finch to chicken genome–genomic alignment was not yet available from UCSC, we aligned each 100% conserved region from the chicken–lizard alignments to the zebra finch (UCSC taeGut1) genome using a custom Python program and BLAST (Altschul et al. 1997), and we stored metadata for each match having an e-value \( \leq 1 \times 10^{-15} \) in the RDB. We removed duplicates from the group of matches containing data from chicken, lizard, and zebra finch, and we defined the remaining set of 3154 sequences as UCEs.

Design of in silico probes from UCEs

Our approach to designing molecular probes from UCEs was that our in silico workflow should be rapidly adaptable to in vitro designs for maximal applicability to organisms without existing genomes. Therefore, instead of simply aligning UCEs to genome-enabled organisms and extracting the whole UCE and flanking sequence, our in silico approach mimicked a commercially available sequence capture workflow (Gniške et al. 2009) by tiling probes across UCEs and then reassembling UCEs and flanking sequence on a per-species basis (described below).

We designed in silico probes by selecting UCEs from the RDB, adding sequence to those shorter than 120 bp in length to make them 120 bp by selecting equal amounts of 3′ and 3′ flanking sequence from a repeat-masked genome chromosome assembly, and recording the length of flanking sequence, if any, added to each. When UCEs were >180 bp, we tiled 120 bp in silico probes across UCEs at 2× density (i.e., probes overlapped by 60 bp). If UCEs were <180 bp in total length, we selected a single probe from the center of the UCE. We conducted a BLAST search of in silico probes against themselves to identify and remove duplicates arising as a result of probe design, and we selected a reduced set of 2560 in silico molecular probes from the RDB having zero duplicate matches, <10 masked bases, and <50 added bases (25 to each side) for downstream use. These probes represented 2386 UCEs in chicken, lizard, and zebra finch.

Alignment of in silico probes to amniote genomes

We aligned in silico probes to available genomic sequence from placental mammals and several outgroups (Supplemental Table S1) using LASTZ (available at http://www.bx.psu.edu/miller_lab). We retained only those probes aligning to genomes having \( \geq 92.5\% \) identity across \( \geq 100 \) bp of the 120-bp probe sequence, and we ignored probes that matched in multiple locations within any genomic sequence, to filter out potential paralogs. We created a table of unique in silico probes located in each species and stored these data in a separate relational database (RDB2). From RDB2, we selected the sets of probes present in all members of our data sets described below.

Generating the data sets

Because data sets that were taxonomically broader resulted in fewer aligning loci, we constructed two data sets for general phylogenetic reconstruction of placental mammals, each having differing levels of taxonomic and genomic coverage (Supplemental Table S1). To obtain high genomic coverage data sets for two particularly difficult phylogenetic hypotheses, we also created a seven-taxon data set to elucidate the phylogenetic position of bats within Laurasiatheria and a seven-taxon data set to address the early radiation of placental mammals into superorders Afrotheria, Xenarthra, and Boreotheria. We favored species for inclusion in a data set if they allowed for more loci and if they were as divergent as possible from other species in the same group, to minimize long branches (e.g., we chose dog and human for the seven-taxon data set exploring basal placental mammal relationships because they are representatives of the two divergent groups within Boreotheria).

Assembling orthologous loci from UCEs, variability, and saturation

For each species within a data set, we excised the alignment of each in silico probe, plus 500 bp of flanking sequence upstream (5′) and downstream (3′) for a total of 1000 bp of flanking sequence because preliminary investigation revealed that this distance would likely contain alignable regions with variation. Using these sequences, we assembled in silico probes back into their respective UCEs using a custom Python program that integrated LASTZ—to match probes to their UCE—and MUSCLE (Edgar 2004) to assemble multiple probes designed for the same UCE. After assembly, we referred to each UCE as a locus. For each locus, we aligned the data across species within a data set using a custom Python program and MUSCLE. We used a moving average across a 20-bp window to trim the ends of all alignments ensuring ends contained at least 50% sequence identity and that nonaligning sequence was removed. We culled loci with missing species within each data set.

To illustrate that loci are likely to be independently segregating, we computed the physical distances between loci in chicken, mouse, and human, because these genome builds (UCSC galGal3, mm9, hg19) are likely the most accurate that also span the taxonomic range we investigated. We calculated physical distance between loci (x-bar ± 95% CI) as the difference in start and stop positions between adjacent loci on each chromosome using a custom Python program.

We compared saturation in 20 randomly chosen UCE loci drawn from the 183-locus data set and the 20 nuclear loci used in Springer et al. (2007) using the saturation index \( (I_S) \) of Xia et al. (2003). We analyzed the core UCE region and UCE flanking regions
separately. For the 16 coding exons from the Springer et al. (2007) data set, we also calculated saturation of third positions of codons separately. We determined the proportion of variable sites and the ratio of informative to variable sites for UCEs and the Springer et al. (2007) loci using MEGA5 (Tamura et al. 2011). Here, we used the 13 Springer et al. (2007) loci that had complete data for a subset of lineages shared with our 29-taxon data set (opossum, sloth, elephant, hedgehog, cow, alpaca, bat, dog, horse, rabbit, mouse, marmoset, and human). For the exons that were unambiguously in-frame for their entire length (n = 6), we also calculated variability statistics separately for third versus first/second codon positions.

Analysis of gene trees and species trees

We estimated gene trees under maximum likelihood in PhyML 3.0 (Guindon et al. 2010) using their most likely substitution model as estimated with MrAIC 1.4.4 (Nylander 2004). We estimated species trees from these gene trees using the STAR (species trees based on average ranks of coalescences) method implemented with the R package Phybase (Liu and Yu 2010). STAR calculates a species tree topology analytically based on average ranks of coalescent events in a collection of gene trees (Liu et al. 2009). STAR performs similarly to probabilistic coalescent-based species-tree methods (e.g., BEST), which are unsuited from a practical perspective for the large data sets used here. STAR also performs well when gene trees deviate from equal evolutionary rates, likely the case in the deep and taxonomically diverse phylogeny we investigated (Liu et al. 2009). In initial explorations, STAR and another analytical species-tree estimation method that uses average coalescence times—STEAC—produced identical topologies with similar bootstrap support. After generating a single STAR tree, we performed 1000 nonparametric bootstrap replicates by resampling nucleotides within loci as well as resampling the loci within the data set (Seo 2008) using a custom Python program, and we generated a cluodogram of gene trees and species-tree bootstraps for the 183 locus data set with DensiTree (Bouckaert 2010). For the visualization of species-tree bootstrap replicates (Fig. 3C), we used STEAC trees because, unlike STAR trees, they contain branch length information, which aided visualization. We analyzed concatenated alignments using MrBayes 3.1 (Huang and Ronquist 2001), grouping genes with the same substitution model as estimated with MrAIC 1.4.4 into different partitions. We tried several partitioning schemes: (1) no partitioning with one GTR+I+T substitution model; (2) partitioning according to the MrAIC substitution model; and (3) partitioning according to the MrAIC substitution model with unlinked molecular rates. We observed no topological differences among results from these partitioning schemes. However, scheme 3 had trouble reaching convergence for all data sets except the 183-locus data set. We thus present results from partitioning scheme 2. All MrBayes analyses consisted of two independent runs (four chains each) of 10,000,000 iterations each, with trees sampled every 100 iterations, for a total of 100,000 trees, from which we sampled the last 50,000 after checking for convergence with the log of posterior probability within and between the independent runs for each analysis.

Calculation of gene-tree probabilities for basal divergence of superorders

We calculated gene-tree probabilities for a five-taxon species tree representing the divergence of the placental mammalian lineages Afrotheria, Xenarthra, Euarchontoglires, and Laurasiatheria, and a marsupial outgroup (opossum) with COAL (Degnan and Salter 2005). We created a species tree in coalescent units (t/2N, where t is the divergence time in generations and N is the effective population size) under a range of possible demographic parameters (generation time = 3–20 yr, population size = 10,000–1,000,000 individuals) and divergence times (0.5, 1, and 5 Mya) (Bininda-Emonds et al. 2007). We summarized gene-tree discordance with the mathematical variance in the frequencies of different gene-tree topologies, which we then rescaled to range between 0 and 1 (low to high discord, respectively). We compared these theoretical values with empirical estimates of gene-tree discord from our 591-locus, seven-taxa data set and retroposon results from the literature (Table 1; Nishihara et al. 2009).

Data access

We provide all Python code, RDB and RDB2, BED files representing both the UCEs and tiled probes, LASTZ alignments, instructions, etc., under open-source (BSD and Creative Commons) licenses at http://dx.doi.org/10.5060/D21N7ZZZ. Users should be aware that we maintain updated code/workflows for many of the steps outlined in the Methods.

Acknowledgments

We thank M. Springer for sharing the 20-locus mammal data set and J. Mattick for sharing UCE and exon data. S.P. Hubbell, S. Edwards, J. Degnan, M. Sheehan, M. Alfaro, B. Carstens, and three anonymous reviewers provided helpful comments. One reviewer suggested the point about selection potentially improving the phylogenetic utility of UCEs. H. Hoekstra provided access to the Odyssey cluster supported by the Harvard FAS Sciences Division Research Computing Group to conduct phylogenetic analysis. A research grant from Amazon Web Services (Amazon.com) also supported phylogenetic computation. We thank the many scientists, institutions, and funding agencies that have contributed genomic data available via the UCSC Genome Browser (see complete list at http://genome.ucsc.edu/goldenPath/credits.html and Supplemental Material).

Authors’ contributions: J.E.M., B.C.F., N.G.C., R.T.B., and T.C.G. designed the study; B.C.F. designed ultraconserved probes and created data sets and performed phylogenetic analysis; N.G.C. performed phylogenetic analysis; J.E.M. performed gene-tree frequency analysis; P.A.G. provided analytical resources; J.E.M., B.C.F., N.G.C., R.T.B., and T.C.G. wrote the manuscript. J.E.M., B.C.F., N.G.C., and T.C.G. contributed equally to the study. All authors discussed the results and commented on the manuscript.

References


Received May 4, 2011; accepted in revised form December 9, 2011.