Syst. Biol. 61(5):717–726, 2012

DOI:10.1093/sysbio/sys004
Advance Access publication on January 9, 2012

Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales

BRANT C. FAIRCLOTH[1,*], JOHN E. MCCORMACK[2], NICHOLAS G. CRAWFORD[3], MICHAEL G. HARVEY[2,4], ROBB T. BRUMFIELD[2,4], AND TRAVIS C. GLENN[5]

[1]*Department of Ecology and Evolutionary Biology, 621 Charles E. Young Drive, University of California, Los Angeles, CA 90095, USA;* [2]*Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA;* [3]*Department of Biology, Boston University, Boston, MA 02215, USA;* [4]*Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; and* [5]*Department of Environmental Health Science and Georgia Genomics Facility, University of Georgia, Athens, GA 30602, USA;*
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA; E-mail: brant@faircloth-lab.org.*

Received 13 September 2011; reviews returned 2 November 2011; accepted 11 December 2011
Associate Editor: Bryan Carstens

Abstract.—Although massively parallel sequencing has facilitated large-scale DNA sequencing, comparisons among distantly related species rely upon small portions of the genome that are easily aligned. Methods are needed to efficiently obtain comparable DNA fragments prior to massively parallel sequencing, particularly for biologists working with non-model organisms. We introduce a new class of molecular marker, anchored by ultraconserved genomic elements (UCEs), that universally enable target enrichment and sequencing of thousands of orthologous loci across species separated by hundreds of millions of years of evolution. Our analyses here focus on use of UCE markers in Amniota because UCEs and phylogenetic relationships are well-known in some amniotes. We perform an in silico experiment to demonstrate that sequence flanking 2030 UCEs contains information sufficient to enable unambiguous recovery of the established primate phylogeny. We extend this experiment by performing an in vitro enrichment of 2386 UCE-anchored loci from nine, non-model avian species. We then use alignments of 854 of these loci to unambiguously recover the established evolutionary relationships within and among three ancient bird lineages. Because many organismal lineages have UCEs, this type of genetic marker and the analytical framework we outline can be applied across the tree of life, potentially reshaping our understanding of phylogeny at many taxonomic levels. [Flanking sequence; genetic markers; phylogenomics; sequence capture; target enrichment; ultraconserved elements.]

Massively parallel sequencing (MPS) has facilitated the study of evolution at the scale of the genome (Shendure and Ji 2008) by fundamentally altering the cost, efficiency, and magnitude at which we can produce and collect DNA sequences. Yet, phylogenomic studies have struggled to develop laboratory and computational protocols to scale DNA inputs efficiently to the output of MPS platforms, reducing the benefits of MPS in terms of data acquisition, time, and cost savings. For this reason, phylogenomic studies have been limited to: analyzing up to a few hundred orthologous loci with polymerase chain reaction (PCR) primers that function across the breadth of taxa under study or mining existing genomes to construct larger data sets. Multiplex PCR and methods for tagging and pooling PCR amplicons address the output of MPS platforms (Meyer et al. 2008), but these methods do not scale at the input stage, that is, they do not remove the onerous step of amplifying tens or hundreds of orthologous loci in multiple taxonomically distant species.

To address these problems, we need a large selection of hundreds to thousands of loci that can be universally characterized and are phylogenetically informative across many species within broad taxonomic categories (e.g., mammals, birds, or amniotes). Universal primers that span variable regions offer one approach (Kocher et al. 1989). Yet, after more than 20 years, the number of these markers available for phylogenomic study remains low, and the approach still requires PCR amplification that is vulnerable to contamination and can be difficult to optimize across distantly related taxa. In contrast, recently developed target enrichment strategies (Mamanova et al. 2009) offer a way to scale DNA inputs to MPS output capability without locus-specific PCR. The basic workflow of the targeted enrichment approach is to hybridize fragmented genomic DNA libraries to DNA or RNA probes that are specific to certain genomic targets, wash away non-targeted DNA present in the library, and sequence the remaining enriched DNA en masse using an MPS platform (Mamanova et al. 2009).

We introduce and test a new class of genetic marker, anchored by nuclear ultraconserved elements (UCEs; Bejerano et al. 2004), which universally enables target enrichment and large scale (>500 loci) phylogenomics across amniotes (Fig. 1). UCEs have been a genomic enigma since their discovery in humans (Bejerano et al. 2004) and other animals (Dermitzakis et al. 2005; Siepel et al. 2005; Stephen et al. 2008; Janes et al. 2011). The function of UCEs is an active area of research (Alexander et al. 2010), and UCEs may be regulators and/or enhancers of gene expression (Woolfe et al. 2004; Sandelin et al. 2004; Pennacchio et al. 2006; Lindblad-Toh et al. 2011). UCEs have several features that make them particularly appealing as anchors for molecular markers. Their level of sequence conservation makes them easy to identify and align across divergent genomes; they are found in high numbers throughout the genome (Stephen et al. 2008); they do not intersect with most types of paralogous genes (Derti et al. 2006);

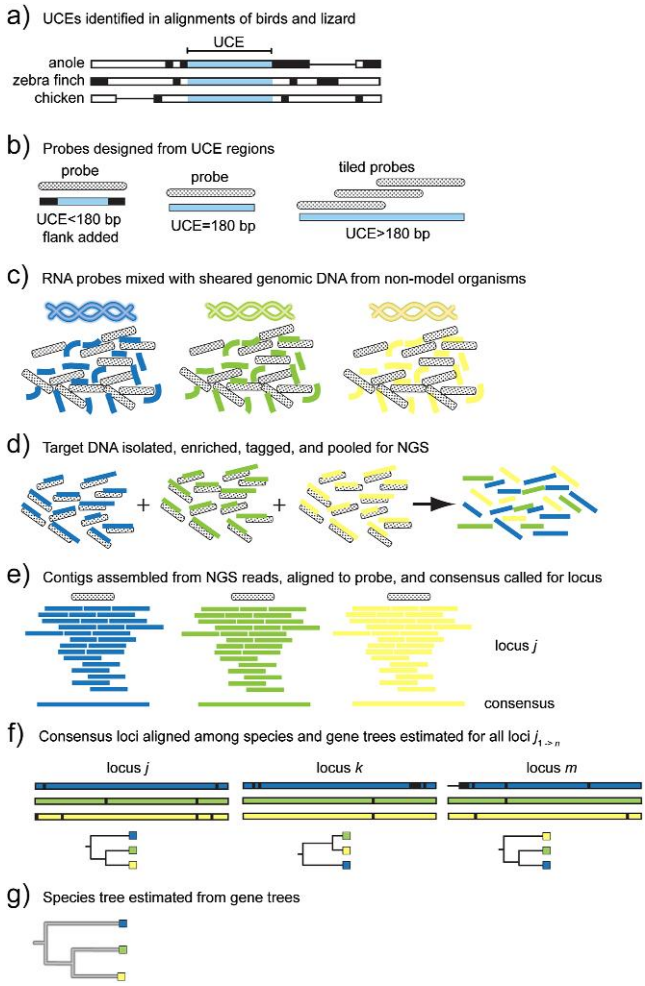

FIGURE 1. Workflow for using UCE-anchored loci in conjunction with target enrichment for phylogenomics. Note: probes = 120 bases.

they have few retroelement insertions (Simons et al. 2006); and the assumption of increasing variability in sequence flanking each UCE suggests that they might be a kind of "molecular fossil," retaining a signal of evolutionary history at many time depths, depending on distance from the core UCE region. For simplicity, we refer to all classes of DNA sequence having high sequence similarity ($\geqslant$80% identity over $\geqslant$100 bp) across divergent taxa as UCEs.

Here, we identify UCEs shared among amniotes, and we design enrichment probes targeting thousands of these UCEs. We demonstrate the practical utility of these markers for phylogenomics using a combination of in silico and in vitro target enrichment experiments that make use of variation in sequences flanking UCEs to capture thousands of independent orthologous nuclear loci suitable for downstream phylogenomic analysis.

METHODS

Identification of UCEs

We identified 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 (http://www.python.org/) script to identify runs of at least 60 bases having 100% sequence identity. We then aligned each 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 matches having an *e* value $\leqslant 1 \times 10^{-15}$ in a relational database (RDB) along with the initial screening results. We removed duplicates from the group of matches containing data from chicken, lizard, and zebra finch, and we defined the remaining set of 5599 unique sequences as UCEs. We estimated the average distance ($\pm$95% CI) between each of these UCEs using positions in the chicken genome (UCSC galgal3) because the chicken genome is currently the most complete and best assembled avian or reptile genome.

Design of Probes from UCEs

We designed target enrichment probes by selecting UCEs from the RDB, adding sequence to those UCEs shorter than 120 bp in length by selecting equal amounts of 5′ and 3′ flanking sequence from a repeat-masked chicken genome assembly, and recording the length of flanking sequence, if any, added to each. We masked all buffered UCEs containing repeat-like regions using RepeatMasker (http://www.repeatmasker.org/) prior to probe design. If UCEs were >180 bp, we tiled 120 bp probes across target regions at 2$\times$ density (i.e., probes overlapped by 60 bp). If UCEs were <180 bp total length, we selected a single probe from the center of the UCE. We used LASTZ (available at http://www.bx.psu.edu/miller_lab) to align probes to themselves and to identify and remove duplicates arising as a result of probe design. We inserted these 5561 probes into the RBD, and we updated each probe record with additional data indicating if probes contained ambiguous (N) bases, the Tm and GC content of the probe, the number of bases added to buffer a particular UCE to probe length (120 bp), the number of masked bases within designed probes, and the types of mismatches we observed for each probe's parent UCE when BLASTing chicken-anole UCEs against zebra finch.

Alignment of Designed Probes to Ten Amniote Genomes

We aligned 5561 probes to ten amniote genomes using a Python wrapper-program around LASTZ to facilitate parallel data processing. We retained only those matches having $\geqslant$92.5% identity across $\geqslant$100 bp of the 120 bp (83%) probe sequence. We used a custom Python program to screen LASTZ matches for reciprocal and non-reciprocal duplicates, and we also excluded matches where the observed number of matches was less than the number of designed probes. For example, if we tiled two probes across a UCE locus, but LASTZ only matched a single probe to the genome sequence, we dropped the parent UCE locus from further consideration.

In Silico Target Enrichment of UCE Loci from Primates

To test our putative in vitro workflow, we aligned 5561 probes to nine primate genomes and one mouse outgroup using LASTZ. As above, we retained only those matches having ⩾92.5% identity across ⩾100 bp of the 120 bp (83%) probe sequence, and we ignored reciprocal and non-reciprocal duplicates to filter out potential paralogs. We also excluded UCE loci where the number of probes matching the genome was less than expected. Across this reduced set of matches and each primate taxon, we sliced the alignment location of remaining probes from the reference genomes plus 200 bp of flanking sequence upstream (5′) and downstream (3′) of the alignment location to yield a total slice of approximately 520 bp. We chose this length of flanking sequence because preliminary calculations revealed that this length was likely close to the maximum contig size we could expect using Illumina Nextera library preparation techniques. We assembled probes + flank back into the parent UCEs they represented using a custom Python program that integrated LASTZ—to match probes to their UCE—and MUSCLE (Edgar 2004) to assemble multiple probes designed from the same parent UCE. After assembly, we refer to each UCE plus flanking sequence as a locus. For each locus, we aligned the data across primate species using a custom Python program and MUSCLE. We used a moving average across a 20-bp window to trim the ends of all alignments, ensure ends contained at least 50% sequence identity, and remove poorly aligned sequence. We allowed insertions at alignment ends as long as the insertions were present in fewer than 30% of individual taxa within a given alignment. We excluded loci that were not found in all primate species, and we dropped alignments that were missing any primate species. This resulted in a complete data matrix with no missing loci.

In Vitro Target Enrichment of UCE Loci from Birds

From the set of 5561 probes, we selected a subset of 2560 probes for synthesis where probes had <10 masked bases and <50 added bases (25 to each side). These probes represented 2386 UCEs conserved among chicken, lizard, and zebra finch. We had these probes commercially synthesized into a custom MySelect kit (Mycroarray, Inc.). We performed phenol–chloroform DNA extractions on bird tissue from vouchered museum specimens (Supplementary Table 1, available from doi:10.5061/dryad.64dv0tg1), and we prepared libraries for Illumina sequencing using Illumina Nextera library preparation kits (Epicentre Biotechnologies).

To enrich the targeted UCEs, we followed the basic workflow for solution-based target enrichment (Gnirke et al. 2009) with several modifications for Nextera-prepared libraries. First, we used 1.8X AMPure XP in place of column clean up following the Nextera tagmentation reaction because the recommended Zymo column clean up yielded lower final DNA concentrations than AMPure. Second, we amplified tagmented libraries using reduced length, HPLC-purified PCR primers complementary to the adapters attached to each DNA fragment during tagmentation. We did not attach sequence tags to libraries at this point to reduce potential complications during enrichment introduced by longer, individually variable, adapters. We increased the number of PCR cycles during the post-tagmentation PCR to 16 or 19 to yield sufficient (∼500 ng) template for enrichment. Following library preparation, we substituted a blocking mix of 500 μM (each) oligos composed of the forward and reverse complements of the Nextera adapters for the kit-provided adapter blocking mix (#3), and we individually enriched species-specific libraries using the synthetic RNA probes described above. We incubated hybridization reactions at 65° C for 24 h on a thermal cycler. Following hybridization, we enriched samples by binding hybridized DNA to streptavidin-coated beads, and we eluted DNA from streptavidin-coated beads using 50 μL of NaOH, which we neutralized with an additional 50 μL of Tris–HCl. We cleaned eluted DNA by binding to 1.8X (v/v) AMPure XP Beads, and we resuspended clean enriched DNA in 30 μL nuclease-free water. We amplified 14 μL of clean, enriched DNA in a 50 μL PCR reaction combining forward, reverse, and indexing primers with either Nextera Taq or Phusion DNA Polymerase (New England Biolabs) to add a custom set of 24 Nextera indexing adapters, robust to insertions, deletions, and substitutions to each sample. Following PCR, we cleaned reactions by binding amplicons to 1.8X (v/v) AMPure XP. We quantified enriched indexed libraries using a Bioanalyzer, and we combined libraries for sequencing at equimolar concentrations assuming all adapter-ligated fragments were at the mean length across all libraries.

We sequenced libraries using the Nextera sequencing primers and a 100 bp single-end Illumina Genome Analyzer IIx run conducted by the LSU Genomics Facility. Following sequencing, the LSU Genomics Facility demultiplexed libraries using Illumina-provided software, and we used a pipeline (https://github.com/faircloth-lab/illumiprocessor) implemented in Python to trim adapter contamination from reads using SCYTHE (https://github.com/vsbuffalo/scythe), adaptively quality-trim reads using SICKLE (https://github.com/najoshi/sickle), exclude reads containing ambiguous (N) bases, and collect metadata for each cluster of reads analyzed.

After pre-processing reads, we assembled species-specific reads into contigs using VELVET (Zerbino and Birney 2008) via VelvetOptimiser (http://bioinformatics.net.au/software.velvetoptimiser.shtml), which we used with default parameters ($k_{range} = 59$–79) to optimize kmer length, coverage, and cutoff for the assembly of data from each species. Velvet resolves potentially variable sites to the majority allele (Zerbino and Birney 2008). We converted contigs output by VELVET to AMOS bank format, and we computed coverage within contigs using custom Python code to parse output from the cvgStat and analyze-read-depth programs provided within the AMOS 3.0.0 software package (http://sourceforge.net/

projects/amos). We used a custom Python program integrating LASTZ (match_contigs_to_probes.py) to align assembled contigs back to their respective probe/ UCE region while removing reciprocal and non-reciprocal duplicates and probes having fewer matches than expected, as described above.

This program (match_contigs_to_probes.py) creates a relational database of matches to UCE loci by taxon, and we used a second program (get_match_counts.py) to query this database and produce a fasta file containing only those contigs built from UCEs present in every taxon. This program also has the ability to include UCE loci drawn from existing genome sequences, for the primary purpose of adding high-quality outgroup data from genome-enabled taxa. We used this feature to include UCE loci identified in Carolina anole (UCSC anoCar2) as outgroup data for the bird phylogeny. We aligned and trimmed reads as described above. We used multimodel inference and model averaging (Burnham and Anderson 2002) of binomial-family generalized linear models (Calcagno and de Mazancount 2010; R Core Development Team 2011) to evaluate the effect of reasonable combinations of the following parameters on enrichment of UCE loci (detected, undetected): UCE GC content, UCE length, probe TM, probe count, masked bases included within probes, bases added to buffer probes, and taxon.

Estimating Substitution Models

We used custom Python programs (run_mraic.py) wrapping a modified MrAIC 1.4.4 (Nylander 2004) to estimate, in parallel, the most likely finite-sites substitution models for each of the alignments generated for primates (2030 loci) and birds (854 loci). We selected the appropriate substitution model for all loci using AIC (Akaike 1974).

Bayesian Analysis of Concatenated Data

We grouped genes with the same substitution model into different partitions, we assigned an appropriate substitution model to each partition, and we concatenated partitions and analyzed these data using MrBayes 3.1 (Huelsenbeck and Ronquist 2001). We conducted all MrBayes analyses using two independent runs (four chains each) of 5,000,000 iterations each, sampling trees every 100 iterations to yield a total of 50,000 trees. We sampled the last 25,000 trees after checking results for convergence by visualizing the log of posterior probability within and between the independent runs for each analysis, ensuring the average standard deviation of split frequencies was <0.00001, and ensuring the potential scale reduction factor for estimated parameters was approximately 1.0.

Analysis of Gene Trees and Species Trees

We estimated gene trees under maximum likelihood with PhyML 3.0 (Guindon et al. 2010) using the most likely substitution model for each tree, which we estimated as described above. We estimated species trees from these gene trees using the STAR (Species Trees based on Average Ranks of coalescences) and STEAC (Species Trees Estimated from Average Coalescent times) methods implemented in the R package Phybase (Liu and Yu 2010). STAR and STEAC calculate a species tree topology analytically based on average ranks or times of coalescent events in collections of gene trees (Liu et al. 2009). STAR and STEAC perform similarly to probabilistic coalescent-based species-tree methods (e.g., BEST), which are unsuited, from a practical perspective, for the size of 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 phylogenies we investigated; a benefit of STEAC, on the other hand, is that it provides an estimate of branch lengths, although they can be somewhat biased (Liu et al. 2009). After generating a single species tree, we used a custom Python program on 250 nodes of a Hadoop (http://hadoop.apache.org/) cluster (Amazon Elastic Map Reduce) to perform 1000 nonparametric bootstrap replicates by resampling nucleotides within loci as well as resampling loci within the data set (Seo 2008).

RESULTS

UCEs Anchor Thousands of Loci in Amniote Genomes

We identified 5599 non-duplicated UCEs by screening genome alignments of two birds (*Gallus gallus* and *Taeniopygia guttata*) and one lizard (*Anolis carolinensis*). General characteristics of these UCEs are that they are generally short [average = 92.5 bp; range 60–742 bp], distributed throughout the genome and, on average, separated from one another by large genomic distances ($188{,}150 \pm 12{,}485$ bp). We used a unique subset of these UCEs to design 5561 enrichment probes targeting these regions.

To demonstrate universality of these loci, we identified homologous genetic regions in five taxonomically diverse amniotes as well as one amphibian (Figure 2 and Supplementary Table 2) using sequence similarity searches of enrichment probes against available genome sequences. Although we identified and designed UCE enrichment probes from reptilians, we located the same probes in high numbers across taxonomically intermediate species, like crocodile, and more distant amniote groups, such as mammals. Unsurprisingly, the number of homologous UCEs dropped with increasing phylogenetic distance from Reptilia, but even the amphibian genome (*Xenopus tropicalis*) contained over 1000 homologous UCEs. Because we used a different search algorithm and filtering parameters to identify homologous loci across all taxa, we recovered fewer (~75% to 85%) than the expected number ($n = 5561$) of enrichment probes targeting UCE loci in *Gallus gallus*, *Taeniopygia guttata*, and *Anolis carolinensis*.

UCE-Anchored Loci Recover the Primate Phylogeny

To assess the ability of UCE-anchored loci to reliably infer a known phylogeny, we used sequence similarity

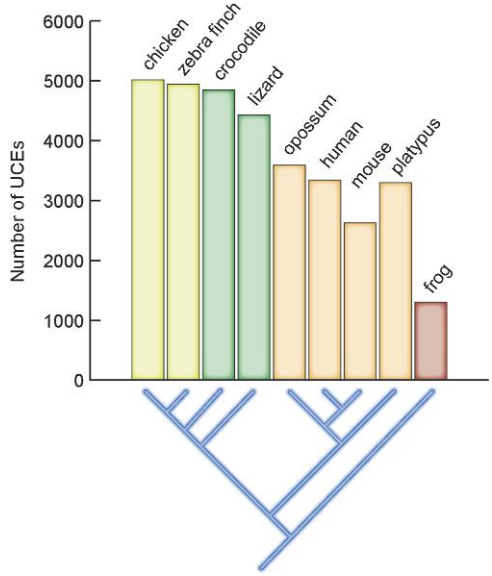

FIGURE 2. The number of UCE-anchored loci found in different amniote groups including birds (*Gallus gallus* and *Taeniopygia guttata*), reptiles (*Crocodylus porosus* and *Anolis carolinensis*), mammals (*Monodelphis domestica*, *Homo sapiens*, *Mus musculus*, and *Ornithorhynchus anatinus*), and one amphibian outgroup (*Xenopus tropicalus*).

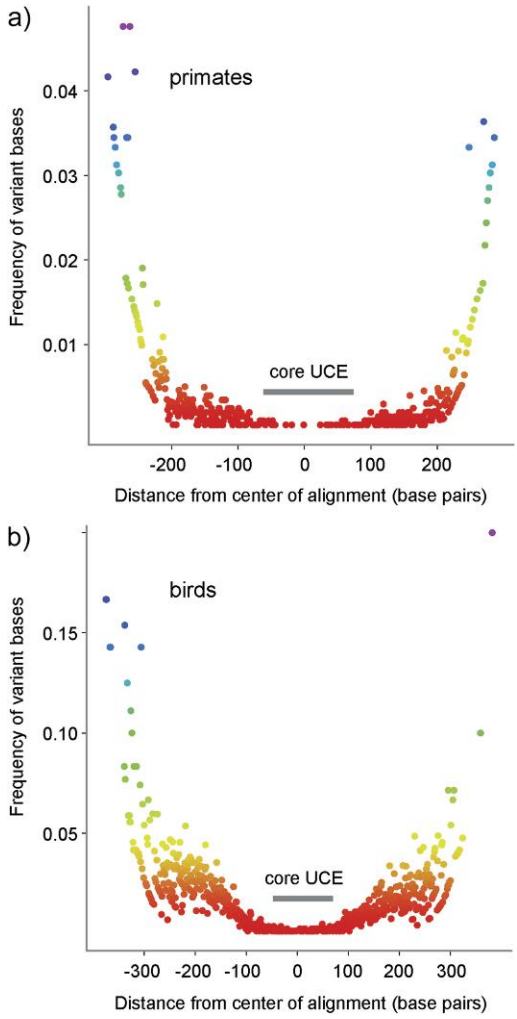

FIGURE 3. Variability increases in the regions immediately flanking the core of UCE-anchored loci in (a) primates and (b) birds. We have removed data points having no variability and outliers for clarity of presentation. Note different scales of axes between figure panels. The variability in avian flanking regions reflects the deeper divergences among bird taxa.

searches to identify regions homologous to 2030 UCE loci in nine primate genomes (Supplementary Table 3), whose relationships are not controversial as well as one rodent outgroup (*Mus musculus*). We aligned enrichment probes targeting each UCE to primate genomes, removed all duplicate matches, and excised the match location $\pm$ 200 bp of sequence flanking for each match, which we then assembled back into the UCE loci from which we designed the probe(s). We refer to these assemblies as loci. After aligning loci among the primates and mouse outgroup, we trimmed unalignable 5′ and 3′ regions, resulting in an average locus length of 432 bp. We confirmed that variable sequence exists in the flanking regions and, to a limited degree, within the core UCE (Fig. 3a). As predicted, variability increased with increasing distance from the center of the UCE.

We used a Bayesian analysis of concatenated data created from 2030 UCE-anchored loci to recover the established phylogeny of these primate species with 1.0 posterior probability for every node (Fig. 4a). We recovered the same phylogeny with high bootstrap support using two methods of species tree analysis, a technique which estimates a species history from independent, often discordant, gene histories (Edwards 2008) (Fig. 4b). To evaluate whether UCE loci follow neutral coalescent processes similar to other types of molecular markers, we determined whether UCE-anchored loci showed discordance in gene histories at levels similar to those previously described for the divergence between human, chimpanzee, and gorilla (Chen and Li 2001). Of 2030 gene trees, 777 (38%) showed a monophyletic group containing only human, gorilla, and chimpanzee. Of these 777 gene trees, 560 had unresolved relationships among human, chimpanzee, and gorilla. Of the 217 resolved gene trees, 152 (70%) supported the species tree grouping human and chimpanzee as sister species, 36 (17%) grouped human and gorilla as sister species, and 29 (13%) grouped gorilla and chimpanzee as sister species. These are very similar proportions to those described by Chen and Li (2001) (72% human/chimp, 21% human/gorilla, and 7% chimp/gorilla), suggesting that UCE-anchored loci follow coalescent processes.

UCE-Anchored Loci Recover the Phylogeny of Several Non-Model Birds

To test whether in vitro target-enrichment and assembly of UCE-anchored loci enable recovery of an established phylogeny, we used a slightly modified version of a commercially available target enrichment protocol

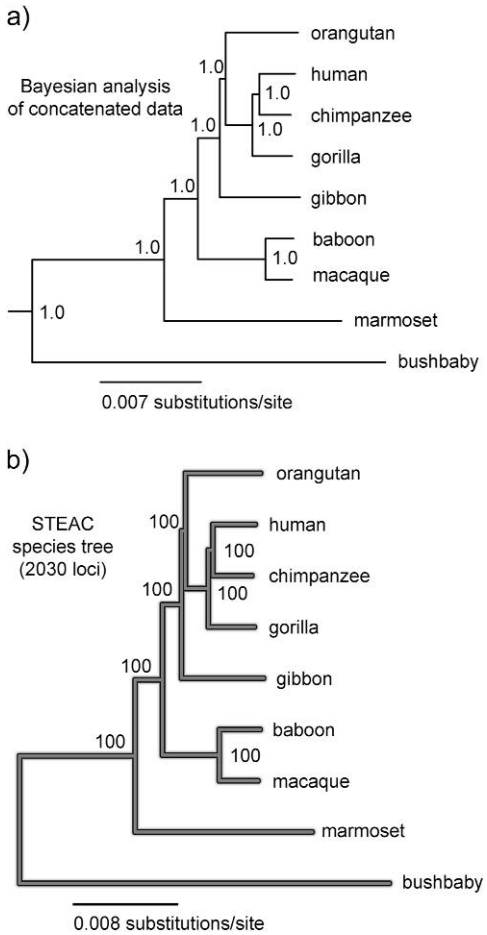

FIGURE 4. Phylogeny (a) of nine primate species based on a Bayesian analysis of concatenated data from 2030 UCE-anchored loci; A species-tree (b) based on average coalescent times within individual gene trees (STEAC). The STAR tree topology was identical. We rooted trees with mouse (not shown).

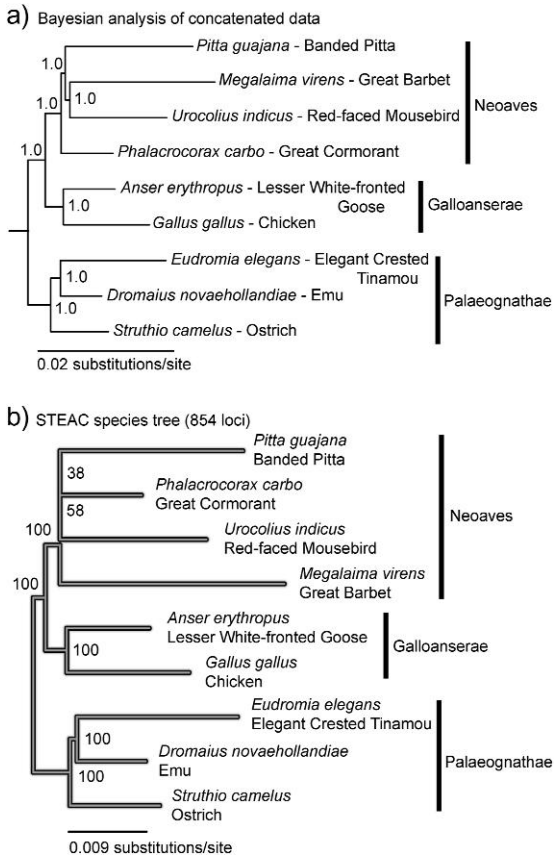

FIGURE 5. Phylogeny (a) of nine bird species representing the three basal lineages of birds based on a Bayesian analysis of concatenated data from 854 loci collected by target enrichment of UCE-anchored loci. (b) STEAC species-tree that estimates a single species history from the 854 independent gene histories. We rooted trees with lizard (not shown).

(Gnirke et al. 2009) with a subset (n = 2560; 46%) of our enrichment probes to collect data from nine bird species (Supplementary Table 1) drawn from three basal lineages of birds whose relationships are well established (Hackett et al. 2008)—the Palaeognathae, Galloanserae, and Neoaves (Table 1). Following target enrichment and sequencing, we obtained an average of 2.68 million reads per species, which we assembled into an average of 1969 contigs per species having an average contig length of 393 bp (Table 1, Supplementary Figure 1). Target UCE GC content (Supplementary Figure 2), probe T_m (Supplementary Figure 3), target taxon, and number of masked bases within each probe negatively affected capture success; target UCE length and bases added to buffer each probe did not affect capture success; and the number of probes targeting each locus (Supplementary Figure 4) increased capture success (Supplementary Table 4 and Supplementary Figure 5). After removing contigs matching more than one targeted UCE ($\bar{x} = 31.4$; Table 1), an average of 1480 (75%) of the remaining contigs matched targeted UCEs. After filtering contigs not matching UCE-anchored loci and loci not present in all bird species, we aligned contigs across the remaining UCE-anchored loci, which resulted in 854 alignments having an average post-trimming length of 412 bp. Similar to primates, variability within the alignments increased with increasing physical distance from the core UCE, and, likely as a result of deeper divergence among the species examined, birds showed a higher degree of variability in both the core UCE region and the sequence flanking the UCE (Fig. 3b).

We used a Bayesian analysis of concatenated data from 854 loci to recover both the established evolutionary relationships among the three bird lineages, in addition to relationships within these three groups, with 1.0 posterior probability for every node (Fig. 5a). We estimated a species tree from independent gene histories, which recovered an identical topology having high bootstrap support, except for relationships within the Neoaves, which remained unresolved (Fig. 5b). A lack of resolution in Neoaves in the species tree is not surprising given the sparse taxonomic sampling in this study and previous results documenting rapid radiation (Hackett et al. 2008). Better taxonomic sampling

TABLE 1. Bird species used for target enrichment and sequence assembly summary statistics

			All contigs				Contigs aligned to UCE loci						
Species	Scientific name	Total reads[a]	Count	Average size	Average coverage	Reads in contigs	Count	Average size	Average coverage	Reads in contigs	Contigs matching > 1 locus	Contigs "on-target"[b]	Reads "on-target"[c]
Banded Pitta	*Pitta guajana*	2,723,264	2370	386	63	535,467	1588	454	71	476,750	32	67%	18%
Great Barbet	*Megalaima virens*	2,302,531	1370	341	59	263,638	1179	351	63	250,128	10	86%	11%
Red-faced Mousebird	*Urocolius indicus*	2,822,685	2208	398	74	621,917	1517	462	84	563,846	43	69%	20%
Great Cormorant	*Phalacrocorax carbo*	4,892,448	1601	521	134	1,060,533	1390	553	138	1,006,224	10	87%	21%
Lesser White-fronted Goose	*Anser erythropus*	3,187,004	2292	354	78	615,257	1617	388	91	555,348	45	71%	17%
Chicken	*Gallus gallus*	1,051,295	1852	352	32	205,950	1452	383	34	186,330	30	78%	18%
Elegant-crested Tinamou	*Eudromia elegans*	2,683,794	1920	391	79	550,578	1509	425	86	513,073	52	79%	19%
Emu	*Dromaius novaehollandiae*	1,361,505	1987	405	41	315,010	1530	449	45	293,804	24	77%	22%
Ostrich	*Struthio camelus*	1,632,540	2119	388	48	370,462	1534	443	54	342,224	37	72%	21%

[a]Count of reads following removal of adapter contamination, quality trimming, and removal of reads containing ambiguous (N) bases.
[b]Contigs "on-target" is the count of contigs matching UCE loci divided by the total number of assembled contigs.
[c]Reads "on-target" is the count of reads assembled into contigs matching UCEs divided by the number of total reads.

will be required to address the evolutionary radiation of Neoaves.

DISCUSSION

We show that UCEs anchor homologous molecular markers across amniotes and are well suited to generating novel and expansive phylogenomic data sets consisting of many hundreds to thousands of loci, simultaneously. Using a subset of the probes we designed, we captured data from an average of 1480 loci in several bird species, which yielded 854 alignments for phylogenetic analysis under strict filtering conditions of no loci missing from any species. The benefits of this approach in terms of the number of loci interrogated, the universality of the method, and the cost and time savings over PCR-based methods are clear—for example, a recent phylogenomic study of birds was based on 19 loci (Hackett et al. 2008). Given the relative ease of collecting this quantity of phylogenomic data across diverse species and the development of multiplexed library preparation protocols for MPS (Kenny et al. 2011), the markers and target enrichment approach we describe have the potential to reshape the way phylogenomic data are collected and analyzed.

There are substantial benefits to focusing on informative portions of the genome instead of sequencing entire genomes of many taxa. Although full genome sequencing is increasingly affordable, full genome assembly requires construction of multiple libraries that are each sequenced to high depth, resulting in vast amounts of data for analysis. Thus, reasonably assembling a full genome sequence across many species is not likely to be practical for some time. Once these data are obtained and aligned across the taxa of interest, only a small portion of the data is useful for phylogenomic analysis. Here, we focus on collecting data that are easily aligned among species within large phylogenetic groups. We prepare a single library per species of interest and we combine libraries representing many species in a single lane of an Illumina flow cell. For example, because 1.5 million paired-end reads of 100 nt is at least as powerful as 3 million single-end reads, a single lane of paired-end 100 nt reads on an Illumina HiSeq using version 3 chemistry (typically yielding >150 million reads) is sufficient to characterize ~1500 UCEs from 100 amniotes. Increasing the efficiency of enrichment or honing in on the most informative subset of loci will allow even more species to be sequenced in each lane.

The benefit of UCE-anchored probes over other possible target-enrichment probe sets and other MPS-based methods of identifying genetic markers is that UCEs are uniquely suited to collect data from a broad range of species, enabling phylogenomics at time depths that depend only on the needs of the research question (McCormack et al. 2012). For example, a probe set designed from a cDNA library is useful for capturing the exome of the source species and closely related species, but specificity likely drops more quickly with decreasing relatedness compared with UCEs because protein

coding regions are not as conserved (Stephen et al. 2008). The same is true for methods of SNP generation using MPS on reduced representation libraries (Davey et al. 2011), where mutations in restriction sites eventually reduce the ability to align collected data across increasingly unrelated species (McCormack et al. 2011).

In contrast, UCEs are well represented in amniotes that are taxonomically intermediate to birds and squamates, such as the crocodile, and they are also found in high numbers in more distantly related amniote groups, such as mammals (Fig. 2). Though the number of similarly conserved UCEs drops considerably in one representative of the outgroup to amniotes, the amphibian *Xenopus tropicalus*, we identified more than 1000 homologous, UCE-anchored loci in *Xenopus*, despite the phylogenetic distance. Additionally, depletion of UCEs in segmental duplications and other copy number variants (Derti et al. 2006) results in few paralogs. Beyond amniotes, UCEs have been described in other animal groups including fish (Lee et al. 2011), invertebrates, and fungi (Siepel et al. 2005); species separated for hundreds of millions of years. Identifying and targeting UCEs as phylogenomic markers in similarly divergent group of organisms, such as plants, will enable the use of UCE-anchored loci for phylogenomics across much of the tree of life with only a modest number of independent probe sets.

We used UCEs to capture phylogenetically informative content from just above the genus level in primates (∼5 myr of divergence between human and chimp) to the very deepest branches of the bird tree of life (∼65 myr of divergence). Data from the intersection of our UCE-anchored probes with dbSNP maps in humans demonstrate that DNA immediately flanking UCEs captures genetic variation at the intraspecific level (Supplementary Figure 6), but further work is needed to determine the amount of information in SNPs at such shallow time depths. Because variation tends to increase moving away from the core UCE, the question of whether UCEs will capture variation at recent timescales (and therefore be useful for population genomics) will likely turn on how much flanking sequence can be obtained around each UCE. This question will potentially be addressed as very long read (>800 bp) MPS technologies (e.g., Roche-454 FLX+ and Pacific Biosciences) become more widespread. Here, we collected data using 100-bp single-end reads. Paired-end reads of similar length would likely yield somewhat longer contigs at higher coverage, which would recover additional flanking sequence. Using Illumina libraries of maximum length (∼600 bp) would also increase the size of contigs obtained.

Several refinements to the process we present will further enhance the benefits of UCEs as phylogenomic markers for target enrichment. First, experimentally deriving the optimum tiling density for capture of UCE-anchored loci will likely increase the number of reads on-target and the sequencing coverage of captured regions flanking UCEs. Generally, we targeted UCEs using a single probe because we did not want to excessively buffer the targeted area with additional flanking sequence (i.e., 4×, 6×). However, target enrichment probes are generally robust to mismatches and length differences during hybridization and added bases did not negatively affect capture performance (Supplementary Figure 5). Because prior research suggests that optimum tiling density is ∼2× (Tewhey et al. 2009), it may be more effective to buffer targeted regions to a length supporting 2X tiling density (∼180 bp).

Second, we applied a strict criterion to our final phylogenomic data matrices, which was that we did not include loci if data were missing from any species for that locus, which still resulted in 854 loci from the bird target-enrichment data. The effect of missing data in phylogenetic analysis is not well known. One recent study suggested that missing data can mislead both topology and branch lengths (Lemmon et al. 2009), but another study came to the contradictory conclusion that including more data was beneficial, even if it led to higher levels of missing data (Wiens and Morrill 2011). In our case, the issue of dropping loci compounds as more taxa are added to the group of species under analysis (Supplementary Figure 7), largely as a result of the failure of probes to enrich loci in certain species and/or because some loci have been lost from the genomes of some species. It is important to note that many of these losses putatively result from the same set of problematic loci (Supplementary Figure 8), and there appears to be a lower bound to the loss accumulation curve as new species are added to an analysis (Supplementary Figure 7). Even including failures, the number of loci returned under the strictest of filtering conditions exceeds, by at least an order of magnitude, the data returned by currently available conserved PCR primer pairs.

Finally, UCE-anchored loci are found in high numbers, separated by wide genomic distances, and are likely to be independently sorting. These properties make UCE-anchored loci particularly well-suited for use with an emerging suite of methods for estimating species trees from collections of gene trees (Edwards 2008). Our results from primates indicate that UCEs have levels of gene-tree discordance similar to those expected by coalescent stochasticity at the well-described divergence between human, chimpanzees, and gorillas (Chen and Li 2001). This suggests that these loci conform to neutral coalescent expectations and are likely appropriate for use in phylogenetic analyses that model the coalescent, including species-tree analysis (Edwards et al. 2007). Renewed recognition of the difference between species trees and gene trees is arguably changing the face of systematics (Edwards 2008). Species tree methods are especially critical for uncovering rapid radiations where discordant gene histories result from random gene sorting during short speciation intervals (Liu et al. 2009). Species trees have yet to be fully incorporated into phylogenomics, in part because species-tree methods are computationally intensive and generally impractical for large data sets. Here, we used species-tree methods that handle many loci by calculating summary statistics from collections of gene trees

(Liu et al. 2009), which necessarily results in some information loss compared with Bayesian or likelihood methods. Faster probabilistic methods that leverage the full information content in thousands of independent loci to estimate a singular history of life are a much-needed mathematical advance for appropriately utilizing with the scale of phylogenomic data we describe here.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository (doi:10.5061/dryad.64dv0tg1). Avian DNA sequences used to build character matrices are available from NCBI Genbank (accession JQ328245-JQ335930). Character matrices and tree files for concatenated analyses are available from TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S12156). Computer code used within this manuscript is available from https://github.com/faircloth-lab/2011-faircloththetal-systbiol-uce and https://github.com/ngcrawford/CloudForest under open-source licenses.

Author Contributions

B.C.F., J.E.M., N.G.C., R.T.B., and T.C.G. designed the study; B.C.F. located UCEs, designed enrichment probes, created data sets, developed laboratory protocols, and performed phylogenetic analyses; J.E.M., M.G.H., and T.C.G. developed laboratory protocols and conducted laboratory work; N.G.C. performed phylogenetic analyses; J.E.M. performed gene-tree frequency analysis; J.E.M., B.C.F., N.G.C., R.T.B., and T.C.G. wrote the manuscript. All authors discussed results and commented on the manuscript. The authors declare no competing financial interests.

Funding

National Science Foundation [DEB-0614208 to T.C.G., DEB-0841729 to R.T.B., DEB-0956069 to R.T.B.] provided partial support for this work. Funds from an Amazon Web Services education grant to T.C.G., N.G.C., and B.C.F. supported computational portions of this work.

Acknowledgments

We thank S. Herke and the LSU Genomic Facility, R. Nilsen, S. Hubbell, P. Gowaty, M. Reasel, M. Alfaro, and C. Schneider. B. Carstens and two anonymous reviewers provided helpful comments that improved this manuscript. D. Ray, E. Green, and J. St. John allowed us access to a very early build of the crocodile genome (croPor0.0.1; http://crocgenomes.org/). D. Dittmann and C. Duffie processed the loan of avian tissue samples from the Collection of Genetic Resources at the LSU Museum of Natural Science. We also thank the many scientists, institutions, funding agencies, and individuals who have contributed to the UCSC Genome Browser (http://genome.ucsc.edu/), Ensembl (http://ensembl.org), NCBI (http://www.ncbi.nlm.nih.gov/), the Broad Institute (http://www.broadinstitute.org/), The Genome Institute at Washington University Sequencing Center (http://genome.wustl.edu/), and Baylor College of Medicine's Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu/), data from all of which helped to accomplish this work.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19:716–723.

Alexander R.P., Fang G., Rozowsky J., Snyder M., Gerstein M.B. 2010. Annotating non-coding regions of the genome. Nat. Rev. Genet. 11:559–571.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W., Mattick J., Haussler D. 2004. Ultraconserved elements in the human genome. Science. 304:1321.

Burnham K., Anderson D. 2002. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer.

Calcagno V., de Mazancount C. 2010. GLMULTI: An R package for easy automated model selection with (generalized) linear models. J. Stat. Softw. 34:1–29.

Chen F.C., Li W.H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68: 444–456.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510.

Dermitzakis E.T., Reymond A., Antonarakis S.E. 2005. Conserved nongenic sequences—an unexpected feature of mammalian genomes. Nat. Rev. Genet. 6:151–157.

Derti A., Roth F.P., Church G.M., Wu C.-T. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat. Genet. 38:1216–1220.

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

Edwards S.V. 2008. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U. S. A. 104:5936.

Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., Fennell T., Giannoukos G., Fisher S., Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27:182–189.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hackett S., Kimball R., Reddy S., Bowie R., Braun E., Braun M., Chojnowski J., Cox W., Han K., Harshman J. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320:1763.

Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Janes D.E., Chapus C., Gondo Y., Clayton D.F., Sinha S., Blatti C.A., Organ C.L., Fujita M.K., Balakrishnan C.N., Edwards S.V. 2011. Reptiles and mammals have differentially retained long conserved noncoding sequences from the Amniote ancestor. Genome Biol. Evol. 3:102–113.

Kenny E.M., Cormican P., Gilks W.P., Gates A.S., O'Dushlaine C.T., Pinto C., Corvin A.P., Gill M., Morris D.W. 2011. Multiplex target

enrichment using DNA indexing for ultra-high throughput SNP detection. DNA Res. 18:31–38.
Kocher T.D., Thomas W.K., Meyer A., Edwards S.V., Pääbo S., Villablanca F.X., Wilson A.C. 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. U. S. A. 86:6196.
Lee A.P., Kerk S.Y., Tan Y.Y., Brenner S., Venkatesh B. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol. Biol. Evol. 28:1205–1215.
Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130.
Lindblad-Toh K., Garber M., Zuk O., Lin MF., Parker B.J., Washietl S., Kheradpour P., Ernst J., Jordan G., Mauceli E., Ward L.D., Lowe C.B., Holloway A.K., Clamp M., Gnerre S., Alföldi J., Beal K., Chang J., Clawson H., Cuff J., Di Palma F., Fitzgerald S., Flicek P., Guttman M., Hubisz M.J., Jaffe D.B., Jungreis I., Kent W.J., Kostka D., Lara M., Martins A.L., Massingham T., Moltke I., Raney B.J., Rasmussen M.D., Robinson J., Stark A., Vilella A.J., Wen J., Xie X., Zody M.C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin J., Bloom T., Chin C.W., Heiman D., Nicol R., Nusbaum C., Young S., Wilkinson J., Worley K.C., Kovar C.L., Muzny D.M., Gibbs R.A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree A., Dihn H.H., Fowler G., Jhangiani S., Joshi V., Lee S., Lewis L.R., Nazareth L.V., Okwuonu G., Santibanez J., Warren W.C., Mardis E.R., Weinstock G.M., Wilson R.K., Genome Institute at Washington University, Delehaunty K., Dooling D., Fronik C., Fulton L., Fulton B., Graves T., Minx P., Sodergren E., Birney E., Margulies E.H., Herrero J., Green E.D., Haussler D., Siepel A., Goldman N., Pollard K.S., Pedersen J.S., Lander E.S., Kellis M. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 478:476–482.
Liu L., Yu L. 2010. PHYBASE: an R package for phylogenetic analysis. Bioinformatics. 26:962–963.
Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58: 468–477.
Mamanova L., Coffey A.J., Scott C.E., Kozarewa I., Turner E.H., Kumar A., Howard E., Shendure J., Turner D.J. 2009. Target-enrichment strategies for next-generation sequencing. Nat. Methods. 7:111–118.
McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. forthcoming 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. Genome Research. doi:10.1101/gr.125864.111.
McCormack J.E., Maley J.M., Hird S.M., Derryberry E.P., Graves G.R., Brumfield R.T. 2011. Next-generation sequencing reveals population genetic structure and a species tree for recent bird divergences. Molecular Phylogenetics and Evolution 62:397–406.
Meyer M., Stenzel U., Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. Nat. Protoc. 3:267–278.
Nylander J.A.A. 2004. MrAIC.pl. Program distributed by the author. Uppsala (Sweden): Evolutionary Biology Centre, Uppsala University.
Pennacchio L.A., Ahituv N., Moses A.M., Prabhakar S., Nobrega M.A., Shoukry M., Minovitsky S., Dubchak I., Holt A., Lewis K.D., Plajzer-Frick I., Akiyama J., De Val S., Afzal V., Black B.L., Couronne O., Eisen M.B., Visel A., Rubin E.M. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 444: 499–502.
R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Sandelin A., Bailey P., Bruce S., Engstrom P.G., Klos J.M., Wasserman W.W., Ericson J., Lenhard B. 2004. Arrays of ultraconserved noncoding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics. 5: 99.
Seo T.K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol., 25: 960–971.
Shendure J., Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26:1135–1145.
Siepel A., Bejerano G., Pedersen J.S., Hinrichs A.S., Hou M., Rosenbloom K., Clawson H., Spieth J., Hillier L.D.W., Richards S. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Simons C., Pheasant M., Makunin I.V., Mattick J.S. (2006) Transposon-free regions in mammalian genomes. Genome Res. 16: 164–172.
Stephen S., Pheasant M., Makunin I.V., Mattick J.S. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. 25:402–408.
Tewhey R., Nakano M., Wang X., Pabón-Peña C., Novak B., Giuffre A., Lin E., Happe S., Roberts D.N., LeProust E.M. 2009. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol. 10:R116.
Wiens J.J., Morrill M.C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.
Woolfe A., Goodson M., Goode D.K., Snell P., McEwen G.K., Vavouri T., Smith S.F., North P., Callaway H., Kelly K. 2004. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3: e7.
Zerbino D.R., Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821.