A Genome Sequence Resource for the Aye-Aye (Daubentonia madagascariensis), a Nocturnal Lemur from Madagascar

George H. Perry1,*, Darryl Reeves2,3, Páll Melsted1,4, Aakrosh Ratan5, Webb Miller5, Katelyn Michelini6, Edward E. Louis Jr7, Jonathan K. Pritchard1,6, Christopher E. Mason2,3,*, and Yoav Gilad1,*

1Department of Human Genetics, University of Chicago
2The Institute for Computational Biomedicine, Weill Cornell Medical College
3Department of Physiology and Biophysics, Weill Cornell Medical College
4Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland
5Center for Comparative Genomics and Bioinformatics, Pennsylvania State University
6Howard Hughes Medical Institute, University of Chicago
7Center for Conservation and Research, Omaha’s Henry Doorly Zoo, Omaha, Nebraska
*Corresponding author: E-mail: ghp3@psu.edu; chm2042@med.cornell.edu; gilad@uchicago.edu.

Accepted: 1 December 2011

Data Deposition: The sequence data have been deposited in the National Center for Biotechnology Information short read archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) as study no. SRA043766.1. The genome assembly data have been deposited in GenBank as a whole-genome shotgun sequencing project (Accession number AGTM00000000). The scaffold sequences used in the analyses in this study, locations of identified SNPs in the aye-aye genome, and the multispecies gene alignment files are available at http://giladlab.uchicago.edu/data.html.

Abstract

We present a high-coverage draft genome assembly of the aye-aye (Daubentonia madagascariensis), a highly unusual nocturnal primate from Madagascar. Our assembly totals ~3.0 billion bp (3.0 Gb), roughly the size of the human genome, comprised of ~2.6 million scaffolds (N50 scaffold size = 13,597 bp) based on short paired-end sequencing reads. We compared the aye-aye genome sequence data with four other published primate genomes (human, chimpanzee, orangutan, and rhesus macaque) as well as with the mouse and dog genomes as nonprimate outgroups. Unexpectedly, we observed strong evidence for a relatively slow substitution rate in the aye-aye lineage compared with these and other primates. In fact, the aye-aye branch length is estimated to be ~10% shorter than that of the human lineage, which is known for its low substitution rate. This finding may be explained, in part, by the protracted aye-aye life-history pattern, including late weaning and age of first reproduction relative to other lemurs. Additionally, the availability of this draft lemur genome sequence allowed us to polarize nucleotide and protein sequence changes to the ancestral primate lineage—a critical period in primate evolution, for which the relevant fossil record is sparse. Finally, we identified 293,800 high-confidence single nucleotide polymorphisms in the donor individual for our aye-aye genome sequence, a captive-born individual from two wild-born parents. The resulting heterozygosity estimate of 0.051% is the lowest of any primate studied to date, which is understandable considering the aye-aye’s extensive home-range size and relatively low population densities. Yet this level of genetic diversity also suggests that conservation efforts benefiting this unusual species should be prioritized, especially in the face of the accelerating degradation and fragmentation of Madagascar’s forests.

Key words: genome assembly, molecular clock, primate evolution, lemur.

Introduction

The only surviving representative of the primate family Daubentoniidae is the aye-aye (Daubentonia madagascariensis). The aye-aye is a nocturnal lemur with unusual derived traits including an elongated, thin, highly flexible third middle digit on the hand, a pair of relatively huge continuously growing incisors, the largest relative brain size of any strepsirrhine primate (lemurs and lorisoids), and a relatively slow
life history including late weaning and a protracted learning period (Cartmill 1974; Martin 1990; Simons 1995; Fleagle 1999; Catlett et al. 2010). These features are likely adaptations that facilitate complex extractive foraging strategies to obtain seeds (endosperm) of hard-shelled ramy nuts or grubs from cavities gnawed in tree bark or bamboo stalks (Gibson 1986; Iwano and Iwakawa 1988; Erickson 1991; Sterling and Povinelli 1999). Aye-ayes have the largest species range of any extant lemur (fig. 1; Mittermeier et al. 2010). However, they are primarily solitary foragers, with extensive home ranges (Sterling 1993) and population densities that are likely very low (Mittermeier et al. 2010).

The slow life history, extensive home-range size, and low population densities of the aye-aye, when considered in the context of Madagascar's continued forest degradation and fragmentation, may have severe implications for the long-term survival of this important species. Presently, aye-ayes are considered “Near Threatened” (Red List of threatened species) but an elevation to “Endangered” status has been proposed (Mittermeier et al. 2010).

A publicly available resource of aye-aye genomic sequence data would facilitate future population studies designed to assess genetic diversity and characterize patterns of biogeographical variation to aid in aye-aye conservation efforts and could be compared with the published genomes of other primates for evolutionary analyses. With the advent of new sequencing technologies, obtaining complete genomic sequence of a new species should no longer be considered a prohibitive task. However, the full assembly
and analysis of a finished or high-quality draft genome sequence, as typically produced by large collaborative consortia, remain prohibitively challenging and would be slow for individual labs interested in advancing research with their own study organisms. In this paper, we demonstrate that readily attainable unfinished draft-quality genome sequences have high potential value and utility for many conservation, ecological, and evolutionary studies of nonmodel organisms.

Materials and Methods

DNA Isolation and Sequencing

Genomic DNA was isolated from a liver tissue sample taken at necropsy from an adult male aye-aye (Dm6514m; Goblin) at the Duke Lemur Center, Durham, North Carolina, with the Gentra PureGene purification kit (QIagen). This captive-born individual was the offspring of two wild-born parents (Dm6451m and Dm6453f), each originally captured in Northeast Madagascar near Mananara-Nord (fig. 1). Genomic DNA libraries for paired-end Illumina Genome Analyzer IIx sequencing were prepared following manufacturer instructions. Three different libraries were constructed, with modes of insert sizes at 448, 580, and 957 bp as estimated with a DNA 1000 chip on the Agilent Bioanalyzer 2100 (supplementary fig. 1, Supplementary Material online).

The 580 bp insert library was pair-end (PE) sequenced on one flowcell (eight lanes) of the Illumina Genome Analyzer IIx for 108 cycles from each end (PE 108 × 108), generating a total of 263,743,085 PE reads (an average of 33.2 million PE reads per lane). On a second flowcell, we sequenced the 448 bp insert library on one lane (33,345,067 PE reads). As a control, one flowcell, we sequenced the 957 bp insert size library, generating a total of 2,856,952,161 bp. The largest contig was 86,436 bp, and the N50 contig size was 3,650 bp (i.e., half of the total length of the contigs is contained within contigs of at least 3,650 bp).

The CLC bio Assembler does not include a scaffolding step. Therefore, we used SSPACE (Boetzer et al. 2011) to merge contigs into scaffolds, requiring a minimum of two paired reads between contigs for merging. The parameters used for SSPACE were “-x=0 -k=2 -v=1,” with the reverse compliment option turned off and minimum allowed error of 0.125, 0.083, and 0.05 for reads from the 448, 580, and 957 bp libraries, respectively. We thus produced 2,564,533 scaffolds representing a total of 2,970,026,655 bp (maximum scaffold size = 252,467 bp; N50 scaffold size = 13,597 bp). The genome assembly data have been deposited in GenBank as a whole-genome shotgun sequencing project (Accession number AGTM00000000). In addition, because there may be differences in the deposited GenBank assembly sequence and the assembly sequence used for analyses in our paper due to GenBank trimming procedures, the specific scaffold sequences that were used for analyses in this paper are available at http://giladlab.uchicago.edu/data.html.

Quality Assessment

To evaluate the quality of our draft genome assembly, we compared our sequence data with aye-aye nuclear genome nucleotide sequence data previously deposited in GenBank that were generated by polymerase chain reaction (PCR) and Sanger sequencing methods (Perry et al. 2006, 2007; Horvath et al. 2008; Perelman et al. 2011). We used BLAST (Altschul et al. 1997) to identify 96 nonoverlapping best alignments >100 bp in size with ≥92% sequence identity (to exclude small and nonorthologous alignments) to 76 of the 77 aye-aye sequences that have been deposited in GenBank (supplementary fig. 2, Supplementary Material online). In total (including indels), 68,876 of 69,013 aligned base
pairs (99.80%) were identical between the GenBank and genome sequences.

The proportion of nonidentical sites from the GenBank BLAST analysis is similar to a published estimate of aye-aye genetic diversity: π (average pairwise sequence divergence) = 0.08% (Perry et al. 2007) that was based on sequence data from 16 intergenic regions (>25 kb in total) and two opsin genes for eight wild-born aye-ayes. The eight individuals included the two wild-born parents of the individual whose genome we sequenced for this study. Therefore, we evaluated the sequence and polymorphism data from that study (34,187 bp) to further estimate our assembly error rate. Based on the BLAST analysis against the 18 regions from Perry et al. (2007), 31,871 of 31,936 aligned bp (99.80%) were identical in the aye-aye genome scaffolds. Of the 65 total differences, 37 were found to be SNPs (including 27 sites annotated with ambiguity codes in the GenBank sequence used in the BLAST analysis because that individual was heterozygous at those positions). Five differences were inferred sequence errors in the scaffold sequence (four of the five errors were clustered near differences were inferred sequence errors in the scaffold sequence). The eight individuals included the two wild-born parents of the individual whose genome we sequenced for this study. Therefore, we evaluated the sequence and polymorphism data from that study (34,187 bp) to further estimate our assembly error rate. Based on the BLAST analysis against the 18 regions from Perry et al. (2007), 31,871 of 31,936 aligned bp (99.80%) were identical in the aye-aye genome scaffolds. Of the 65 total differences, 37 were found to be SNPs (including 27 sites annotated with ambiguity codes in the GenBank sequence used in the BLAST analysis because that individual was heterozygous at those positions). Five differences were inferred sequence errors in the scaffold sequence (four of the five errors were clustered near differences were inferred sequence errors in the scaffold sequence). The eight individuals included the two wild-born parents of the individual whose genome we sequenced for this study. Therefore, we evaluated the sequence and polymorphism data from that study (34,187 bp) to further estimate our assembly error rate. Based on the BLAST analysis against the 18 regions from Perry et al. (2007), 31,871 of 31,936 aligned bp (99.80%) were identical in the aye-aye genome scaffolds. Of the 65 total differences, 37 were found to be SNPs (including 27 sites annotated with ambiguity codes in the GenBank sequence used in the BLAST analysis because that individual was heterozygous at those positions). Five differences were inferred sequence errors in the scaffold sequence (four of the five errors were clustered near differences were inferred sequence errors in the scaffold sequence). The eight individuals included the two wild-born parents of the individual whose genome we sequenced for this study. Therefore, we evaluated the sequence and polymorphism data from that study (34,187 bp) to further estimate our assembly error rate. Based on the BLAST analysis against the 18 regions from Perry et al. (2007), 31,871 of 31,936 aligned bp (99.80%) were identical in the aye-aye genome scaffolds. Of the 65 total differences, 37 were found to be SNPs (including 27 sites annotated with ambiguity codes in the GenBank sequence used in the BLAST analysis because that individual was heterozygous at those positions). Five differences were inferred sequence errors in the scaffold sequence (four of the five errors were clustered near differences were inferred sequence errors in the scaffold sequence).
any frame-shifting indels introduced in the alignment step by normalizing the alignments relative to the human RefSeq transcript coding sequences. That is, when the frame of the aligned sequences was shifted with respect to the human reference transcript, 1–2 bp gaps were introduced in all sequences as a correction. The resulting multispecies alignment files for each analyzed gene are available at http://giladlab.uchicago.edu/data.html. The MEGA4 program (Tamura et al. 2007) was used to estimate nucleotide sequence divergence and phylogeny and to perform relative rate tests.

We used PAML (Yang 2007) to estimate ancestral sequences and the numbers and rates of nonsynonymous and synonymous substitutions on each branch. The ratio of the rates of nonsynonymous to synonymous substitution (dN/dS) can be examined to make inferences about long-term selective pressures on amino acid sequences. We removed genes for which dS > 1 on any branch or dS > 0.5 on any branch excepting mouse, as these outlying values may reflect alignment or sequence artifacts. We also limited our analysis to genes with >100 synonymous sites that were aligned and analyzed across all species.

We calculated two dN/dS ratios for each remaining gene in each lineage. First, the conventional dN/dS value, where dS is based on the synonymous substitution rate for the individual gene on that branch. The second dN/dS value is based on a single genome-wide estimate of dS for each lineage; calculated as the total number of synonymous substitutions summed across all genes divided by the total number of synonymous sites summed across all genes. The second dN/dS value, which uses the genome-wide dS value in the denominator, may be valuable for interpreting results on shorter branches, where the number of gene-specific synonymous substitutions can be low or zero. We considered genes and branches that meet the following conditions as potential candidates to have evolved under positive selection: dN/dS > 1, dN/dS genome > 1, and ≥2 synonymous substitutions (on short branches, stochastic variation in the number of synonymous substitutions can lead to large numbers of genes with dN/dS > 1). Although it is unlikely that all genes and branches meeting these conditions were subjected to positive selection at the amino acid sequence level, the set of candidates we generated is likely enriched for such genes. Functional enrichment analyses using gene ontology annotations (Ashburner et al. 2000) were conducted using GeneTrail (Keller et al. 2008).

**Results and Discussion**

Our goal was not to produce a finished aye-aye genome assembly. With short read data only and currently available assembly algorithms, a high-quality “finished” assembly of a new mammalian genome requires truly extensive sequence coverage (on the order of 100-fold coverage) and considerable computational power that is not yet commonly available, extensive library preparation and sequencing efforts, or both (e.g., Gnerre et al. 2011). The complete annotation of the aye-aye genome (e.g., the annotation of genomic regions, functional units, duplicated regions, etc.) was also beyond the scope of this project, as this would require considerable resources, typically achieved only through consortium-led projects.

Rather, we sought to efficiently generate a genomic resource that would be sufficient to facilitate most downstream applications relevant to an endangered nonmodel species. These applications might include evolutionary and phylogenetic analyses or the development of primers and markers for conservation- and behavior-related population genetic studies. The results of the initial analyses presented below illustrate this point.

**Aye-Aye Genetic Diversity**

We generated ~19× sequence coverage of the assembled portions of the genome of one aye-aye individual. With these data, we estimated aye-aye genetic diversity using magnitudes more nucleotide sites than were previously available for such an analysis. Among the 581,674,772 sites with coverage sufficient for reliable SNP identification (see Materials and Methods), we identified 293,800 heterozygous sites (π = 0.051%). This genetic diversity estimate is slightly lower than those from each of the two previous aye-aye nuclear genome population genetic studies (π = 0.081% from 25,649 bp of intergenic region nucleotide sequence data generated by PCR and Sanger sequencing for eight individuals; Perry et al. 2007 and π = 0.073% from 197,784 synonymous sites in the coding regions of 1,175 genes based on RNA-seq data for two individuals; Perry et al. 2012). These previous studies included individuals with ancestry from multiple regions of Madagascar, whereas both parents of the donor individual for our genome reference sequence were captured in a single region. This difference may explain, at least in part, the slight variation in results. In addition, our estimate was computed from multiple magnitudes more sites over a broader representation of the genome but from one individual only.

Regardless, all three of these independently generated genetic diversity estimates place aye-ayes at the extreme low end of primate nuclear genome diversity, at least among the 14 other species for which comparable estimates are available (Yu et al. 2003, 2004; Fischer et al. 2004, 2006, 2011; The Chimpanzee Sequencing and Analysis Consortium 2005; Voight et al. 2005; Hernandez et al. 2007; Wall et al. 2008; Perry et al. 2010; Locke et al. 2011; Perry et al. 2012). The low genetic diversity estimate could be explained, at least in part, by aye-aye demography. Behavioral observational data have suggested that aye-ayes have both extensive home ranges and low population densities (Sterling 1993; Mittermeier et al. 2010). The low genetic diversity estimate could also be explained by an ancient bottleneck.
event. Because genetic diversity is an important variable to consider when assessing extinction risk (Frankham 2005; Palstra and Ruzzante 2008), when combined with our knowledge of aye-aye demography and in the face of habitat loss and forest fragmentation in Madagascar, our result supports the recommendation of Mittermeier et al. (2010) that the conservation status of this important Malagasy species be elevated from Near Threatened to Endangered.

Genomic Divergence and Aye-Aye Branch Length Analysis

For third codon position sites in gene coding regions, the genome-wide aye-aye versus human nucleotide sequence divergence estimate = 0.152 (maximum composite likelihood model; Tamura et al. 2004). Estimated nucleotide sequence divergence varies by chromosome and is lowest on chromosome X (fig. 2), a finding consistent with those of previous primate comparative genomic studies (e.g., The Chimpanzee Sequencing and Analysis Consortium 2005).

As depicted in figure 2, the human versus aye-aye sequence divergence estimate is relatively lower than that for rhesus macaque versus aye-aye. The same pattern is true for the comparisons with other species, including dog (fig. 2). This difference is likely explained by the “hominoid slowdown” effect, which describes a lower substitution rate in hominoids (including humans) relative to Old World monkeys (including macaques) and is hypothesized to result from variation in life-history patterns among lineages, for example, in generation time (Yi et al. 2002; Steiper et al. 2004; Kim et al. 2006).

Unexpectedly, however, these data also suggest an even lower substitution rate in the aye-aye lineage relative to the human and macaque lineages. Specifically, the aye-aye versus dog nucleotide sequence divergence estimate is lower than the estimates for both human and macaque versus dog (fig. 2). Under the null model of equal substitution rates among branches, these estimates are expected to be similar (Tajima 1993). Yet, of the 4,261,562 total aligned third position sites among human, aye-aye, and dog, 275,396 are inferred to have substitutions that occurred along the human lineage, whereas 244,690 are inferred to have substitutions that occurred along the aye-aye lineage (relative-rate test; $P < 0.00001$). Thus, the nucleotide substitution rate in the lineage leading to humans from the primate common ancestor is more than 10% higher than that of the aye-aye lineage. This difference in substitution rates is readily visualized by a neighbor joining phylogeny estimated from the third position nucleotide sequence data (fig. 3).

We considered two potential explanations for the unexpected finding of a relatively low substitution rate in the aye-aye lineage. First, while the majority of third codon position substitutions are synonymous (having no effect on a translated protein’s amino acid sequence) and therefore presumably neutral, a small proportion of third codon position substitutions are nonsynonymous. Therefore, theoretically, a markedly stronger strength of purifying selection in the aye-aye lineage could drive the pattern we observed. However, the maximum likelihood–estimated genome-wide rate of synonymous substitution (Yang 2007) in the aye-aye lineage (0.093; 260,589 substitutions at 2,811,311 sites) is also considerably lower than that of the human lineage following divergence from the primate common ancestor (0.140; 394,179 substitutions at 2,811,311 sites; substitutions summed across ancestral and extant lineages; summary in supplementary table 1, Supplementary Material online). This observation suggests that the estimated branch
The relatively short aye-aye branch length is consistent, in some respects, with the hypothesis that life-history traits such as generation time are inversely correlated with nucleotide substitution rate (for a recent review, see Tsantes and Steiper 2009). Specifically, aye-ayes have the most extended life-history profile of any extant lemur, with a weaning age of 1.5 years compared with a maximum of 0.68 years in any other studied species and an interbirth interval of 2.5 years versus a maximum of 2 years and the more typical 1 year in other lemurs (Catlett et al. 2010). The age at first reproduction of aye-aye females, 3.5 years (Ross 2003; males show behavioral evidence of sexual maturity at 2.5 years; Winn 1994), is obviously much earlier than that of modern humans. However, the modern human life-history pattern is likely very different from that of our longer term ancestors starting from the time that we last shared a common ancestor with aye-ayes ~70 Ma.

Moreover, there is indirect evidence suggesting the possibility that ancestral aye-ayes had longer generation times than modern aye-ayes. Specifically, there were at least 17 now-extinct “subfossil” lemur species that survived on Madagascar until ~2,300 to 1,000 years ago; the timing of extinctions is coincident with the arrival of humans to the island (Godfrey et al. 2006). The reconstructed body sizes of all subfossil lemurs are substantially larger than those of their extant relatives (Jungers et al. 2008), and dental histology patterns suggest that ages of weaning and interbirth intervals for many of these taxa were relatively protracted as well (Catlett et al. 2010). A giant aye-aye, *Daubentonia robusta* (Lamberton 1934; Macphee and Raholimavo 1988; Simons 1994), is believed to have had a body mass greater than five times that of the extant aye-aye, *Daubentonia madagascariensis* (Jungers et al. 2008). While life-history trait estimates are not yet available for the subfossil aye-aye, it is possible that the generation time of this species was also considerably greater than that of extant aye-ayes. If so, then it would be interesting to speculate and reconstruct the generation time of ancestral aye-ayes and consider the extent to which a protracted ancestral aye-aye generation time may have contributed to, at least in part, the shortest aggregate branch lengths of any extant primate studied to date.

Evolutionary Analysis of Gene Coding Regions

This aye-aye genome sequence resource is the first such high-coverage resource available for lemurs, thus providing an important step toward more comprehensive primate comparative genomic analyses. For our analysis, we first constructed a database of aye-aye gene sequences on the basis of homology to human exons (see Materials and Methods). We then generated a multispecies alignment of the aye-aye genes and the orthologous gene sequences of human, chimpanzee, orangutan, rhesus macaque, mouse, and dog (representing four other published...
We identified 11 candidate genes for the aye-aye lineage (supplementary table 2, Supplementary Material online). Even with such a small number of genes, our analysis may help us to advance our understanding of primate-specific aspects of the immune system that may have played critical roles in our evolutionary history.

Conclusion

We have produced a genome sequence resource of the aye-aye, *Daubentonia madagascariensis*. The assembly was generated in a cost-effective computationally efficient manner and cannot be considered finished quality (being comprised of ~2.6 million scaffolds). However, the assembly is sufficient to facilitate subsequent comparative and population genetic studies that will benefit our understanding of lemur evolutionary biology in general and aye-aye conservation efforts in particular. For example, the sequence and SNP data that we produced will be useful for subsequent biogeographic studies of aye-aye genetic variation that use either PCR or DNA capture methods as well as for studies of gene expression profiles using RNA sequencing. In our analysis, we unexpectedly found that the substitution rate on the aye-aye lineage is relatively slow compared with other primates, which may reflect a protracted ancestral aye-aye life-history pattern. Additionally, our heterozygosity estimate for the aye-aye donor individual places this important species at the extreme low end of primate genetic diversity. Given the aye-aye’s demographic profile of large home ranges and low population densities in the face of continuing deforestation and habitat fragmentation in Madagascar, this result leads us to emphasize the need to prioritize efforts that will benefit aye-aye conservation.

Supplementary Material

Supplementary figures 1 and 2 and tables 1 and 2 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Sarah Zehr and the Duke Lemur Center for providing the aye-aye sample used for this study, Ying Wang for assistance with the PAML analysis, Lisa Kimmel for producing figure 1, Steve Salzberg and David Kelley for early access.
to the Quake sequence error correction program, and Walter Pirovano and Marten Boetzer for early access to the SSPACE scaffolding program. We are grateful to Anne Yoder and Nathaniel Dominy for discussions and comments on the manuscript. This work was funded by National Institutes of Health (NIH) grant GM084996 to Y.G., and G.H.P. was supported by NIH fellowship F32GM085998.

**Literature Cited**


Associate editor: Dmitri Petrov