We performed a population genomics study of the aye-aye, a highly specialized nocturnal lemur from Madagascar. Aye-ayes have low population densities and extensive range requirements that could make this flagship species particularly susceptible to extinction. Therefore, knowledge of genetic diversity and differentiation among aye-aye populations is critical for conservation planning. Such information may also advance our general understanding of Malagasy biogeography, as aye-ayes have the largest species distribution of any lemur. We generated and analyzed whole-genome sequence data for 12 aye-ayes from three regions of Madagascar (North, West, and East). We found that the North population is genetically distinct, with strong differentiation from other aye-ayes over relatively short geographic distances. For comparison, the average $F_{ST}$ value between the North and East aye-aye populations—separated by only 248 km—is over 2.1-times greater than that observed between human Africans and Europeans. This finding is consistent with prior watershed- and climate-based hypotheses of a center of endemism in northern Madagascar. Taken together, these results suggest a strong and long-term biogeographical barrier to gene flow. Thus, the specific attention that should be directed toward preserving large, contiguous aye-aye habitats in northern Madagascar may also benefit the conservation of other distinct taxonomic units. To help facilitate future ecological- and conservation-motivated population genomic analyses by noncomputational biologists, the analytical toolkit used in this study is available on the Galaxy Web site.

conservation genomics | landscape species concept | genomics of non-model species

Madagascar maintains one of the highest levels of unique biodiversity—coupled with imminent extinction risk—in the world (1–4). The endemic primates, lemurs, are among the most diverse faunal groups on Madagascar, with ~100 distinct extant taxa (5). Because of a unique ecological and demographic profile, the lemur species with the largest geographical distribution, the aye-aye (Daubentonia madagascariensis) (6), may also be among the most sensitive to continuing degradation of Madagascar’s forests. Specifically, aye-ayes have very large individual home-range size requirements relative to other lemurs (7–11), population densities that are inferred to be very low (12), a relatively slow life history (13), and the lowest nuclear genetic diversity of any primate yet studied (14). Therefore, their ability to maintain sufficient individual numbers for long-term population viability in remaining forest patches may be at risk.

Aye-ayes are highly specialized extractive foragers, with relatively large, continuously growing incisors that are used to gnaw through decaying tree bark (deadwood) or bamboo to access wood-boring insect larvae and through the endocarp of seeds from the ramy tree (Canarium) to access endosperm (7–15). A slender, flexible, probing third digit is used to extract these foods and bring them to the mouth (18, 19). Limitations in the availability of either deadwood or Canarium resources may explain the large individual home-range requirements, but this has not yet been shown. Aye-ayes are also nocturnal, cryptic, and primarily solitary, making them difficult to study and sample in the wild. As a result, no comparative population studies of this species have been published previously.

Adequate conservation planning requires knowledge of both long-term landscape dynamics and patterns of species distribution in suitable habitats (20). Wider geographic scale assessments are particularly important for species with large-range requirements (e.g., refs. 21 and 22), such as aye-ayes. Thus, it is important to understand the patterns of genetic differentiation that exist among surviving aye-aye populations. Because of their extensive individual home-range sizes and low population densities, conservation efforts relevant to aye-ayes will need to prioritize the preservation of large and contiguous forests. Although such protected areas do exist in Madagascar [albeit many of them currently under stress (23)], prior to this study we have not been able to assess whether current protected areas and conservation strategies maximize the preservation of distinct aye-aye populations and overall genetic diversity, because the genetic relationships among aye-aye populations have been unknown.

The analysis of population-level genome sequence data offers potentially powerful insights into both demographic and evolutionary processes. Although such analyses could thus benefit behavior, conservation, and ecological research across many taxa (24–27), large-scale whole-genome sequencing population studies conducted to date have typically focused on humans and model organisms (e.g., refs. 28–30). With continued increases in sequencing capacity, genomic-scale population studies of non-human, non-model organisms are increasingly feasible. Indeed, transcriptome sequencing, other reduced representation methods, and whole-genome sequencing approaches have been used in multiple recently published studies (14, 31–35).

In this study, we have generated and analyzed intermediate-coverage whole-genome sequence data for 12 aye-aye individuals from three regions of Madagascar (Fig. 1). We investigated population structure and quantified population differentiation using an analytical toolkit that we have made available through the Galaxy Web site (36, 37) to facilitate similar, future studies.
of other species. Although the analyses in this study benefitted from our previous assembly of an aye-aye reference genome (38) and high-quality DNA samples isolated from blood and tissue, the absence of such resources would not preclude the application of a similar pipeline to other nonmodel species. Although the analyses in this study benefitted from our previous assembly of an aye-aye reference genome (38) and high-quality DNA samples isolated from blood and tissue, the absence of such resources would not preclude the application of a similar pipeline to other nonmodel species. To limit the incorporation of erroneous genotypes into our analyses, we focused on the genotypes of 666,256 SNPs (of a total of 4,555,737 SNPs) that were covered by a minimum of four sequence reads per individual (see Materials and Methods).

Aye-Aye Population Structure. To examine relationships among the aye-aye individuals from the three regions of Madagascar, we constructed a neighbor-joining tree based on genome-wide SNP genotype distances. Individuals from each of the three populations cluster, and the East and West population clusters are more similar to each other than either is to the North (Fig. 2A). In addition, we constructed a rooted neighbor-joining tree by first aligning the aye-aye and human genome sequences (hg19). Given the relatively high levels of sequence divergence in introns and intergenic regions, to ensure orthology it was necessary to focus on gene coding regions. We then analyzed only autosomal synonymous SNPs from the gene-coding regions, because such SNPs do not affect amino acid sequences and are thus presumably neutral. Although the resulting tree is based on a distance matrix constructed from magnitudes of fewer SNPs than that of the genome-wide dataset (857 vs. 666,256), the patterns of aye-aye population structure are the same and the root of the tree separates the North from the East and West populations

![Aye-aye geographical species range and sampling locations. Remaining forests (light green) and presumed current species distribution of aye-aies (dark green) in Madagascar. Capture locations for the North (blue squares), West (yellow triangles), and East (red circles) aye-aye population samples included in this study are indicated. Species distribution is based on the selection of remaining forest (identified from classified satellite imagery, deforestation data from Harper, et al. (2007), courtesy of Cambridge University Press (51)) using a vector polygon of aye-aye distribution from Andrainarivo et al. (71), which was then further modified according to our field observations of aye-aye feeding traces and occasional sightings. Nonforested areas were not represented in the aye-aye distribution.](https://www.pnas.org/doi/10.1073/pnas.1211990110)

**Fig. 1.** Aye-aye geographical species range and sampling locations. Remaining forests (light green) and presumed current species distribution of aye-aies (dark green) in Madagascar. Capture locations for the North (blue squares), West (yellow triangles), and East (red circles) aye-aye population samples included in this study are indicated. Species distribution is based on the selection of remaining forest (identified from classified satellite imagery, deforestation data from Harper, et al. (2007), courtesy of Cambridge University Press (51)) using a vector polygon of aye-aye distribution from Andrainarivo et al. (71), which was then further modified according to our field observations of aye-aye feeding traces and occasional sightings. Nonforested areas were not represented in the aye-aye distribution.

**Results**

We analyzed intermediate-coverage whole-genome sequence data for 12 aye-aye individuals from three regions of Madagascar: North (n = 4 individuals), West (n = 3), and East (n = 5) (Fig. 1). We identified SNPs following sequence-read alignment to an aye-aye reference genome (38). To limit the incorporation of erroneous genotypes into our analyses, we focused on the genotypes of 666,256 SNPs (of a total of 4,555,737 SNPs) that were covered by a minimum of four sequence reads per individual (see Materials and Methods).

Aye-Aye Population Structure. To examine relationships among the aye-aye individuals from the three regions of Madagascar, we constructed a neighbor-joining tree based on genome-wide SNP genotype distances. Individuals from each of the three populations cluster, and the East and West population clusters are more similar to each other than either is to the North (Fig. 2A). In addition, we constructed a rooted neighbor-joining tree by first aligning the aye-aye and human genome sequences (hg19). Given the relatively high levels of sequence divergence in introns and intergenic regions, to ensure orthology it was necessary to focus on gene coding regions. We then analyzed only autosomal synonymous SNPs from the gene-coding regions, because such SNPs do not affect amino acid sequences and are thus presumably neutral. Although the resulting tree is based on a distance matrix constructed from magnitudes of fewer SNPs than that of the genome-wide dataset (857 vs. 666,256), the patterns of aye-aye population structure are the same and the root of the tree separates the North from the East and West populations

![Neighbor-joining tree from genotype distance matrix - 666,256 genome-wide SNPs.](https://www.pnas.org/doi/10.1073/pnas.1211990110)

**Fig. 2.** Aye-aye population structure. Analyses of estimated genotype SNPs with minimum 4x sequence coverage in each of the 12 individuals studied, and maximum 120x coverage in those individuals combined (Materials and Methods). (A) Neighbor-joining tree estimated from a genotype distance matrix based on all 666,256 genome-wide SNPs. Pairwise distances were calculated as total SNP genotype distance, with distance for an individual SNP the difference between two individuals’ genotypes scored as 0, 0.5, and 1 (e.g., AA, AT, and TT, respectively). (B) Rooted neighbor-joining tree estimated from a distance matrix based on 857 autosomal synonymous SNPs from gene coding regions that could be aligned to the human genome (hg19). Pairwise distances were calculated as described above. The nucleotide of the human reference sequence was different from both aye-aye alleles for 73 of the 857 SNPs; in these cases the human genotype was scored as 0.5. (C) Population structure analyses based on all 666,256 genome-wide SNPs. Cluster membership proportions for each individual are depicted for both k = 2 and k = 3 populations. Each individual is represented as a vertical bar with population origins indicated below the bars.
(Fig. 2B). A population structure analysis produced consistent results, with the North individuals distinguished from all others at \( k = 2 \) populations, and individuals from each of the three regions completely separated at \( k = 3 \) (Fig. 2C). Results from a principal component analysis are also similar. The first principal component clearly separates the North individuals from all others, and the second principal component separates West and East individuals (Fig. S1).

**Population Differentiation.** To quantify the level of genetic differentiation between aye-aye populations, or the amount of total genetic variation that can be explained by population structure, we estimated \( F_{ST} \) for each SNP that was not fixed for the same allele in each of the two populations being compared. We calculated \( F_{ST} \) values using an unbiased estimator from Reich et al. (41) that is not adversely affected by small population sample sizes (42). The average \( F_{ST} \) values were 0.169 for the North vs. East populations (596,785 SNPs), 0.194 for the North vs. West populations (517,323 SNPs), and 0.129 for the East vs. West populations (536,734 SNPs).

We next assessed the level of observed genetic differentiation between aye-aye populations in a comparative context. To do so, we created an equivalent dataset for humans based on publicly available data. Specifically, we obtained genome sequence data for a total of 12 human individuals from three populations that were, as were our aye-aye data, generated using Illumina sequenc- ing technology. The sampled human populations were sub-Saharan African agriculturalists \((n = 4)\), European \((n = 5)\), and Southeast Asian \((n = 3)\). We matched sequence coverage levels to the aye-aye data at both the individual and population levels (Fig. S2), and used the same pipeline for sequence alignment, SNP genotype estimation, data filtering (e.g., minimum fourfold sequence coverage per individual), and \( F_{ST} \) analysis.

The average \( F_{ST} \) values were 0.078 for the African vs. European populations \((1,061,254 SNPs)\), 0.091 for the African vs. Asian populations \((988,646 SNPs)\), and 0.069 for the European vs. Asian populations \((748,232 SNPs)\). Thus, the level of genetic differentiation between aye-aye populations from the North and East regions of Madagascar is more than 2.1-times greater than that between human Africans and Europeans based on an equivalent dataset (Fig. 3). The relative level of aye-aye versus human population differentiation was similar for each of the other equivalent comparisons. Furthermore, the two least-differentiated aye-aye populations, East and West (average \( F_{ST} = 0.129 \)), are likely more genetically differentiated than Africans and Asians (average \( F_{ST} = 0.091 \)), the two most differentiated human populations in our analysis. We obtained consistent results from SNP subsets generated using sequence coverage cutoffs of five-, six- and sevenfold per individual, and when using Weir and Cockerham’s unbiased estimator for \( F_{ST} \) (43) or Wright’s original \( F_{ST} \) definition (44) instead of Reich et al.’s unbiased \( F_{ST} \) estimator (41) (Fig. S3).

**Neutral Genetic Diversity.** Aye-ayes have the lowest level of genetic diversity of any studied primate species (14, 38, 45), which is likely a function of large individual range requirements and low population densities throughout the aye-aye species distribution. Alternatively, because the previous genetic diversity estimates were based primarily on individuals with ancestry from only one region of Madagascar (equivalent to our East population), they could reflect population-specific rather than species-wide demographic processes. We sought to address this issue by estimating neutral genetic diversity separately for each of the three aye-aye populations in our study. Although we expect to underestimate true genetic diversity with the intermediate coverage sequence data (because of the undercalling of rare SNPs and heterozygous sites), these estimates should be generally comparable both across aye-aye populations and against humans (using our sample size and sequence coverage-matched human dataset).

We identified the total number of autosomal synonymous sites covered by a minimum of four sequence reads per individual for each species (aye-aye = 368,675 synonymous sites; human = 915,245 sites) (SI Materials and Methods) and computed average pairwise genetic diversity \((\pi)\) for each population: aye-aye North \( \pi = 0.054\% \), aye-aye East \( \pi = 0.057\% \), aye-aye West \( \pi = 0.049\% \), human African \( \pi = 0.093\% \), human European \( \pi = 0.070\% \), human Asian \( \pi = 0.066\% \). As expected, these estimates are lower than those from previous studies that used higher coverage sequence data (14), but the magnitude of the between-species differences are similar. Moreover, in contrast to the variability observed among human populations [as expected, with higher genetic diversity in Africa (46)], genetic diversity levels are similar among aye-aye populations suggesting that relatively low genetic diversity is a species-wide characteristic.

**Discussion**

We conducted this study to characterize patterns and levels of genetic differentiation among aye-aye populations for conservation planning purposes as well as to contribute to the general understanding of biogeographical processes in Madagascar, as this species has the widest geographical distribution of any lemur (6, 12). The aye-aye’s demographic profile suggests a particular sensitivity of this species to the rapid degradation and fragmentation of Madagascar’s forests. Specifically, aye-aye home-range sizes of 120–215 ha for males and 30–40 ha for females [with travel distances up to 4.4 km/night (47)] are very large for a solitary animal only ∼2.5 kg in size and considerably larger than those of other lemurs across a diversity of activity patterns, social systems, and body sizes [for example, 1–2 ha for woolly lemurs (8), 5 ha for fork-marked lemurs (9), 5.3 ± 5.2 ha for blue-eyed black lemurs (10), and 5.7–10.1 ha for Verreaux’s sifakas (11)]. Because there is minimal same-sex overlap in female home ranges (47), aye-aye population densities are inferred to be relatively very low (12). Perhaps unsurprisingly, the level of
estimated nuclear genomic diversity in aye-ayes is the lowest of any primate yet studied (14). Thus, the ability of aye-ayes to maintain sufficient individual numbers for long-term population viability in remaining forest patches may be at risk.

Our analyses revealed that although aye-ayes from the East and West coasts of Madagascar are distinguished readily by their genome sequences, divergence between either of these populations and the northern Madagascar population is greater. Although connected forests between the habitats of the North and East sampled populations no longer exist (Fig. 1), the level of genetic differentiation between the populations in these regions implies a longer-term reproductive barrier than that which could be attributed to the human-mediated habitat loss that began only within the past 2,300 y (48, 49) before accelerating rapidly over the past century (50, 51). We found that the level of genetic differentiation between the North and East populations is substantially greater than that between, for example, human African and European populations, based on the analysis of an equivalently curated human SNP database. Although the relative level of aye-aye genetic differentiation—across only a relatively small geographic distance (248 km), representing a small part of the total aye-aye range (Fig. 1)—is intriguing, what relevance does this result have for conservation planning? In particular, the level of genetic differentiation among human populations is not typically considered high among primates (52). In addition, the observed level of aye-aye population differentiation may not be unusual for a species with low population sizes and geographical barriers across its range. Future expectations of an expanded population genomic database that will include many additional endangered taxa will help us to better contextualize this result.

As a starting point, we also computed the average $F_{ST}$ between populations of Alaskan and Norwegian polar bears using data from a recent genomics study (34). Although the polar bear sequence coverage levels could not be matched precisely to our aye-aye and human datasets, the average North vs. East aye-aye $F_{ST}$ was more than five-times greater than that for polar bears (SI Materials and Methods), strengthening the belief that the observed aye-aye $F_{ST}$ values may be unusual for a wide-ranging animal over such a small geographic distance. For now, we are not suggesting that the North aye-aye population should necessarily be considered a distinct taxonomic unit. However, if general goals of conservation include preserving distinct populations and maximizing overall species-level genetic diversity, then this population merits particular protection, especially for a species with a demographic profile that suggests high extinction risk.

Because aye-ayes have a geographical species distribution that only excludes the central highlands and the southwest of Madagascar [the largest distribution of any lemur (6, 12)], these results are also particularly valuable for our broader understanding of the island’s biogeography. Indeed, the distinctiveness of the North aye-aye population accords well with 1 of 12 proposed cross-taxon centers of endemism, based on an analysis of elevation, the location of watersheds, and Quaternary climatic shifts (53). Multiple river systems, such as the Manambato, Bemarivo, and Ankava-nana, have the potential to limit the dispersal of aye-ayes and other species between northern Madagascar and adjacent regions. Furthermore, the Tsaratanana Massif, which includes the highest peak in Madagascar, rises above the known elevational limits of aye-ayes (54), and thus may form at least a partial barrier to gene flow. We have sampled from only a portion of the total aye-aye species distribution. The cryptic, nocturnal nature of this species and low population densities constrain sampling efforts, but future population genomic studies that include populations from further south along the west and east coasts of Madagascar would likely contribute further to biogeographic knowledge and to aye-aye conservation efforts.

Preservation of the distinctive aye-aye populations in northern Madagascar would likely have indirect benefits for the conservation of other taxa within this center of endemism, the genetic diversity and structure of which are not yet fully characterized. The geographical correspondence of this northern region with patterns of species turnover in other lemur taxa (53) suggests similar underlying biogeographic processes and responses to landscape variation. Although the use of surrogate species for conservation management is strongly debated (55, 56), carefully selected individual species or groups may serve as appropriate indicators of overall biodiversity, although these applications are often highly context-dependent (57). Similarly, protecting species at the highest trophic levels and with large home ranges (e.g., apex predators) can serve ecosystem-level conservation goals and preserve disproportionate amounts of diversity in other taxa (58). Although not strictly predatory, aye-ayes maintain an ecological niche that necessitates large individual ranges, which can incorporate entire communities and populations of other endemic fauna, including other lemurs. Thus, any efforts to preserve contiguous forests large and diverse enough to support a viable population of aye-ayes would be very likely to also meet or exceed the space requirements for other taxa. Additionally, because of their relatively slow life history, aye-ayes may be particularly sensitive to rapid habitat changes. The combination of these factors would make aye-ayes a strong candidate for a focal species in a “landscape species” approach to conservation planning (59, 60).

Conclusion

In this study, we generated and analyzed complete genome sequences from 12 aye-aye individuals to characterize levels and patterns of genetic differentiation and to highlight a distinct population in northern Madagascar. This work serves as a potential model for future conservation- and ecology-motivated population genomic studies of nonmodel species; such research is expected to become more feasible with continuing advances in sequencing technology and capacity. Thus, we have made the analytical toolkit used for our analyses available on the Galaxy Web site. We argue that conservation attention should be directed toward an important center of endemism in northern Madagascar. Such efforts would preserve distinct populations of a species that is one of the world’s most unusual and highly specialized mammals, as well as other potentially distinct populations and taxa in the region.

Materials and Methods

DNA Samples. DNA samples from 13 wild-caught aye-aye individuals were initially included in this study. Genomic DNA was extracted from liver-tissue samples collected at necropsy or whole venous blood from wild-born founders at the Duke Lemur Center (Durham, NC) and from whole venous blood samples collected from free-ranging individuals in Madagascar. In Madagascar, aye-ayes were immobilized with a CO₂ projection rifle or blowgun with 10 mg/kg of Telazol (Fort Dodge Animal Health), and 1.0 cc/kg whole blood was collected and placed in storage buffer [0.1 M Tris, 0.1 M Na₂EDTA, 2% (vol/vol) SDS] at room temperature until transferred to the laboratory for storage at −80 °C. All collection and export permits were obtained from Madagascar National Parks, formerly Association Nationale pour la Gestion des Aires Protégées, the Ministère des Eaux et Forêts of Madagascar. All rules and regulations were followed according to Malagasy law. Samples were imported to the United States under Convention for International Trade in Endangered Species permits 08US121039/9, 08US121040/9, and 08US121041/9 from the US Fish and Wildlife Service. Capture and sampling procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of Omaha’s Henry Doorly Zoo and Aquarium under IACUC #12-101. Genomic DNA was isolated from the samples using a standard Phenol-chloroform extraction protocol (61).

Sequencing, Sequence Alignment, and SNP Identification. For detailed descriptions of the library preparation, sequencing, alignment, and SNP identification methods see SI Materials and Methods. Each sample was pair-end sequenced for 101 bp from each end using one lane of the Illumina HiSeq 2000 sequencing system. We obtained an average of 204,202,246 total reads...
per lane (SD = 40,714,120) (Dataset S1), or ~20 Gb of raw sequence data per individual. Sequence data have been deposited in the National Center for Biotechnology Institute short read archive under accession no. SRR066444. Sequence reads were aligned to the aye-aye reference genome sequence (38) using the Burrows-Wheeler Aligner (62). On average, we mapped 16 Gb of sequence data per individual (SD 4.1 Gb) (Dataset S1), corresponding to an average of ~5.6-fold coverage of the 2.9-Gb aye-aye reference genome sequence. We used SAMTools (63) to identify the locations of SNPs and estimate genotypes at all SNPs for each individual, regardless of sequence coverage for that SNP and individual.

Quality Control and SNP Filtering. We selected an intermediate-coverage population genomics study design, using one HiSeq lane per individual (resulting in an average of ~5.6× per-individual mapped sequence coverage). For analyses of these data, our approach was to focus on the subset of identified SNPs with a minimum of 4x sequence coverage for each individual in the study. In doing so, we ignored the majority of our data, but we have more confidence in the accuracy of the estimated genotypes of the SNPs that we do analyze. The number of SNPs remaining after this filtering step is still large (see below), sufficient for accurate population genetics inference (e.g., ref. 64). In future population genomics studies, it may become possible to extend the analysis to a larger proportion of variable sites in the genome by utilizing the intermediate-coverage sequence data for each individual. However, such a design would have been economically and computationally inefficient for our study, given current output of the Illumina HiSeq 2000 sequencing system and data storage and analysis computing resource needs.

With only 4x coverage, some error in the estimated genotypes is expected (e.g., the probability of observing both SNP alleles among four reads for a heterozygous individual = 0.875). However, the 4x threshold is the minimum coverage that we defined for an SNP; thus, the coverage levels for most individuals are higher. At this coverage threshold, overall genotype estimate qualities are expected to be high. Before applying the 4x minimum coverage filter, we considered the coverage distributions (for the 4,555,737 identified SNPs) for each individual (Fig. S2A). One individual, North5, was a coverage outlier, with a mode of only 1.6x coverage and 70% of the SNP data covered by fewer than 2x coverage. North5 also had the fewest mapped reads of any individual in the study (Dataset S1). Accordingly, we excluded North5 from further analyses. We also examined the sum total coverage distribution for the remaining 12 individuals (North1 to -4, West1 to -3, East11 to -5) (Fig. S2B). SNPs with relatively low sum total coverage will be filtered because they would not meet the 4x per-individual threshold. Some proportion of the SNPs with relatively high sum total coverage could be located in duplicated regions of the genome, and thus potential false-positives. Therefore, we chose to filter all SNPs with >120x sum total coverage. With this step, we likely excluded many true SNPs, but the remaining number of SNPs was large. Specifically, there were 666,256 SNPs with minimum 4x coverage in each of the 12 individuals, and not more than 120x coverage in those individuals combined. These SNPs were the focus of our population genomic analyses.

Galaxy Tools. We created tools on the Galaxy Web site usegalaxy.org (36, 37) to facilitate ecological- and conservation-motivated analyses of population genomics datasets such as ours, by noncomputational biologists. The user uploads SNP genotype calls with coverage and genotype quality information (alternatively, SAMtools functionality is also available through Galaxy). From the uploaded SNP table, the user may specify populations, compute and display coverage distributions at the individual and population levels, filter SNPs based on individual and population minimum and maximum sequence coverage levels, as well as minimum genotype quality, examine population structure, and perform analyses based on FST. Several popular population genetics and genomics analysis programs, including SMARTPCA (65) and ADMIXTURE (66), have been integrated into the Galaxy functionality. For detailed descriptions of the population genomic analyses used in this study, see SI Materials and Methods.

Human SNP Comparative Data. To understand the significance of the average FST value between the North and East aye-aye populations, we created a similar dataset of human sequence data. Specifically, given the relatively small sample sizes for each population in our study and the use of intermediate-coverage sequence data, unknown false-positive and false-negative SNP call error rates and genotyping errors may have affected the accuracy of our aye-aye average FST estimate. Therefore, it was inappropriate to compare our result for the two aye-aye populations to published reports of average FST from other species that were based on microarray-based SNP genotyping data or Sanger sequencing data (or high-coverage massively-parallel sequencing data, but to our knowledge such datasets have not yet been published). To address this issue, we downloaded illumina sequence reads for four African agriculturalist (e.g., Bantu-speaking) individuals, five individuals of European descent, and three Southeast Asian individuals from published and publicly available genome-sequencing studies (28, 29, 67–70) (Dataset S2). For our comparative purposes, we considered the African sample equivalent to the aye-aye North population sample (as described in Results, the aye-aye North population is more distinct from the West and East populations than either of the West and East populations from each other; this is equivalent to the relationship among human African, European, and Southeast Asian populations), the European sample equivalent to the East, and the Southeast Asian sample to the West. We included varying numbers of reads for each individual (Dataset S3) to match the aye-aye coverage distributions, at both the individual and population levels (Fig. S2C).

We aligned reads, identified SNPs, and estimated genotypes in a manner identical to that used for the aye-aye data, except using the human genome (hg19) as a reference. A total of 8,598,051 SNPs were identified. There are at least two reasons why the number of identified human SNPs was greater than the number of identified aye-aye SNPs. First, nuclear genetic diversity is lower in aye-ayes than in humans (14). Second, although the total size (~2.9 Gb) of the aye-aye reference genome sequence (38) is similar to that of humans, it is comprised of ~2.6 million scaffolds. Thus, the effective size of the genome for SNP analysis is lower for aye-ayes because of expected reductions in mapping and coverage levels near scaffold ends. We filtered the human SNPs to a subset of 1,146,658 SNPs with a minimum of 4x sequence coverage per individual and a maximum of 120x total coverage for all individuals. Population structure analyses performed on these genotype data produced the expected results (Fig. S4).

ACKNOWLEDGMENTS. We thank Sarah Zehr and the Duke Lemur Center for providing captive aye-aye samples; Shannon Engberg and Carolyn Bailey for sample preparation; Lynn Tomsho and John McGraw for library preparation and sequencing; Tracy Wyman for assistance with Fig. 1; Emily Davenport and Kate Thompson for comments on an earlier draft of the manuscript; the Madagascar Biodiversity Partnership for assistance in sample collection and field logistics in Madagascar; and the Madagascar National Parks, formerly Association Nationale pour la Gestion des Aires Protégées, and the Ministère des Eaux et Forêts of Madagascar for sampling permission. Funding for aye-aye DNA collection was obtained by Conservation International and the Margot Marsh Biodiversity Foundation (E.E.L.), along with logistical support from the Ahmanzon Foundation and the Theodore F. and Claire M. Hubbard Family Foundation. Genome sequencing was funded by the College of Liberal Arts, Pennsylvania State University (G.H.P.). The development of the Galaxy tools to analyze intermediate-coverage sequence data from multiple individuals was supported by the National Institutes of Health Grant UI 1R033184-01 (to the Pennsylvania State College of Medicine) and by the National Science Foundation Award DEB 0733029 and a grant from the Pennsylvania Department of Health using Tobacco Commonwealth Universal Research Enhancement Funds. This is Duke Lemur Center publication #1241.


SI Text

Although the analyses in this study benefitted from our previous assembly of an aye-aye reference genome (1), the absence of such a resource would not preclude the application of a similar pipeline to other nonmodel species. Two approaches could have been used to align our aye-aye sequence reads to a reference sequence so that SNP differences could be identified. First, as in the article, one could align reads to a previously reported genome assembly for the species using an aligner, Burrows–Wheeler Aligner (BWA) (2), that requires high sequence similarity (about 97%), allowing it to run quickly. This is the approach that we used for this study. An alternate approach would be to align reads to annotated exons from the genome of a related species. In the case of aye-ayes, this would be human gene coding regions. These exons are easier to correctly align between aye-aye and human than the bulk of the genome because they are generally under purifying selection; then, to study neutral patterns of genetic diversity, analyses could focus on synonymous sites and synonymous SNPs. However, such an approach still requires an alignment program that tolerates lower identity (about 80%) between the two sequences being aligned (and consequently runs much slower than BWA). The program LASTZ, which is freely available and thoroughly documented (www.bx.psu.edu/~rsahrris/lastz/), can be used for this approach. Although the SNP identifications that we made in this study were based on the first approach, we did use the LASTZ alignment procedure to identify orthologous regions from the aye-aye genome to human coding regions, to identify ancestral and derived states and estimate rooted phylogenetic trees (see above). The alignment of reads directly to a more distantly related genome sequence would perform similarly (3) (see below).

The current assembly methods for so-called “next-generation” sequence data typically produce tens or hundreds of thousands (or more) of consensus sequences (4, 5), each covering a small part of the genome. For example, our aye-aye reference assembly contains 2.6 million such sequences (1). For the questions that we addressed using the aye-aye assembly (phylogeny, population differentiation), such an assembly is adequate. Indeed, the only requirement is for a very rudimentary assembly. In outline, the first step in a typical assembly process is to look for overlaps among sequence fragments, align the fragments, and produce consensus “contigs” (contiguous sequences). Subsequently, read-pairs with a known separation in the genome (because they come from the ends of a sized DNA molecule) are used to create ordered and oriented “scaffolds” of contigs. This second step requires that the sequenced library be prepared in a controlled way, and may well use several libraries with different separation distances. However, for simply producing a large number of SNPs, the scaffolds provide little if any information of value, so that part of the assembly process can be omitted.

Moreover, there are many uses of whole-genome sequence data that require more than just a set of, for example, 50,000 scaffolds, such as those requiring longer contiguity (e.g., mega-bases) for the SNPs and accurate gene annotation. For example, some of our analyses (e.g., rooting the phylogenetic tree) required the reliable identification of the orthologous human nucleotide at the positions of aye-aye SNPs. We did this by aligning the aye-aye sequence to human gene coding regions, which are relatively conserved, which facilitates alignment and orthology identification. Within these regions, we then focused on synonymous (amino acid preserving) SNPs as a neutral proxy. This approach of course assumes availability of gene annotation. Thus, an often very useful addition or alternative to creating a genome assembly is to use the preexisting assembly and gene annotation of a related species.

The effectiveness of the alternative approach is illustrated by a whole-genome analysis of the polar bear and other bear species (3). One of the polar bears was sequenced to 100-fold average coverage and assembled into 1.2 million scaffolds and “orphan” contigs. Aligning reads from 23 polar bear individuals, three brown bears, and a black bear to the assembly yielded 13,038,705 genomic positions with a nucleotide variant; although the individuals are not all of the same species, we will call those positions SNPs. Applying the alternative of mapping onto the dog assembly, we identified 12,023,192 SNPs. In general, the use of a de novo assembly of the species of interest can be expected to provide more SNPs than use of the assembly of another species, but at least in this case the gain from using the de novo assembly is not impressive.

Importantly, most of the main analyses reported in the polar-bear report could not have been obtained from an unannotated, highly fragmented de novo bear assembly. For example, the key observations indicating admixture between polar bears and certain brown bears, as well as tracking historical changes in population size, require knowing long stretches of contiguous SNPs, and the reported identification of genes showing signs of adaptive evolution naturally requires gene annotations. (The bear data and tools for those analyses can be found on Galaxy.) Although using the dog data to infer SNP-contiguity and genes in the bear genome is imperfect, few groups have the funding and expertise needed to raise the quality of the bear assembly and annotate its genes so accurately that the analysis would improve on use of the excellent dog data. Moreover, the dog gene annotation would probably be a major ingredient in the recipe for bear gene annotations, as the human genes would be for an attempt to annotate aye-aye genes.

Still, using genomic resources from another species requires that such resources exist for a “sufficiently similar” species. For species evolving at rates considered typical for mammals or birds, this condition can be quantified somewhat. According to the Web site www.timetree.org, the ancestors of bear and dog separated 45 million y ago, which thus seems to be comfortably within the span of “sufficiently similar,” except perhaps for species evolving at unusually high rates (e.g., rodents). The human–lemur separation date quoted by timetree.org is 74 million y ago. Although we didn’t attempt to identify noncoding aye-aye SNPs by mapping aye-aye reads to the human assembly, because our goal was to use the de novo assembly to perform limited kinds of analysis, we were successful in using the human gene annotations to help analyze the aye-aye, as reported in the main text.

SI Materials and Methods

Sequencing. Genomic DNA libraries were created manually from each DNA sample using the Illumina TruSeq DNA Sample Preparation v2 Low Throughput Protocol. DNA samples were quantified using a Picogreen assay on a TBS-380 Mini-Fluorometer. From each sample, 200–1,000 ng of DNA were sheared to ~300-bp fragments using the Covaris Model S2 System. Index adapters from the TruSeq DNA Sample Prep Kit v2 were ligated to the sheared DNA. Ligated products were purified with Agencourt AMPure XP beads (Beckman Coulter). The purified fragments were enriched by PCR (10 cycles) and then size selected using a Pippen Prep (Sage Science) 2% EtBr Gel Cassette (CSD-210). Libraries were visualized using the Agilent
2100 Bioanalyzer with a High-Sensitivity DNA chip. Insert sizes ranged from 300 to 350 bp.

Library concentrations were determined by quantitative PCR on an Applied Biosystems 7300 Real-Time PCR System using KAPA SYBR FAST Universal 2× qPCR master mix (Kapa Biosystems) and Illumina’s Sequencing Library qPCR Quantification Guide. Samples were normalized to 2-nM concentration, denatured, and loaded onto the Illumina cBot with a HiSeq Paired-End Flow Cell following the TruSeq PE Cluster Kit v2-cBot-HS protocol. Each library was pair-end sequenced for 101 bp from each end (recipe: 101,7,101 Paired-End Index Read) using one lane of the Illumina HiSeq 2000 sequencing system with TruSeq SBS Kit v1 sequencing chemistry and software version HCS v1.4.8, RTA v1.12.4.2.

Sequence Alignment and SNP Identification. Sequence reads were aligned to the aye-aye reference genome sequence 1 (using the BWA (2) v0.5.9. The BWA default parameters were used, with the exception of the “-q 20” option, which was applied to soft-clip low-quality 3’ ends of reads before alignment. On average, we mapped 16 Gb of sequence data per individual (SD 4.1 Gb) (Dataset S1), corresponding to an average of ∼5.6-fold coverage of the 2.9-Gb aye-aye reference genome sequence.

SNPs were identified using the Picard toolkit (http://picard.sourceforge.net) to flag potential duplicate reads (i.e., with identical aligned positions between both reads of a read pair with one or more other read pairs) resulting from the library PCR-enrichment step, that could otherwise affect the quality of the SNP calls. Of each set of potential duplicate read pairs, only the pair with the highest sum of base quality scores for bases with quality ≥15 (not mapping quality scores, as these would be affected by SNPs) was used in subsequent SNP analysis. Considering data from all individuals simultaneously, we used SAMtools v0.1.16 (6) to identify the locations of SNPs, using the option “-C 50” to reduce the mapping quality of the reads with multiple mismatches. SNP locations in the nuclear genome were filtered to maintain SNPs for which coverage in every sample was less than 30 reads for that position and the total coverage was less than 250 reads (to limit the erroneous calling of variant positions in repetitive or duplicated regions), and the rms mapping quality was greater than or equal to 20. In total, we identified 4,555,737 SNPs in the nuclear genome. Once the SNPs locations were identified, we then used SAMTools (using the pileup command) to estimate genotypes at all SNPs for each individual, regardless of sequence coverage for that SNP and individual.

Population Genomic Analyses. We conducted the majority of the analyses for this study (unless otherwise specified below) using the Galaxy tools, based on the individual genotype estimates for the filtered subset of 666,256 aye-aye SNPs. The full input table of 4,555,737 aye-aye SNPs is available as a Shared Data Library on Galaxy based on genotypes, with genotypes scored as 0, 0.5, or 1 (e.g., AA, AT, and TT) for each SNP and the distance for a given pairwise comparison the difference between the values for the two individuals. The output value for each pairwise comparison is the average difference among all considered SNPs. After downloading the genotype distance matrix from Galaxy (generated in MEGA format), we estimated and plotted a neighbor-joining tree (10) using MEGA5 (11).

We evaluated the level of differentiation between populations by computing $F_{ST}$ value for each SNP using three different formulas: Reich et al. unbiased estimator (12), Weir and Cockerman’s unbiased estimator (13), or Wright’s original definition (14). We computed the average $F_{ST}$ value among all SNPs that were not fixed for the same allele in both populations.

Polar Bear $F_{ST}$ Analyses. We conducted this study to characterize patterns and levels of genetic differentiation among aye-aye populations for conservation planning purpose. The level of observed genetic differentiation between aye-aye populations was compared with that from an equivalent dataset for humans. However, comparable analyses of additional taxa, once the data are available, are needed to place our results in a firm context.

Our recently published data for polar bears (3) provide a glimpse of what the forthcoming data may show. That study reported low-coverage whole-genome sequence data for 23 polar bears. Five of the individuals came from Alaska, with the remainder from Svalbard, an island north of Norway, roughly 2,000 miles away. Because the sequence coverage per individual was generally less than for the aye-aye data, we cannot simply discard some of the polar-bear sequences to create a dataset that matches the aye-aye coverage, as we did for the human data. For the aye-aye data, the fraction of SNPs where the four northern and five eastern individuals were covered by fewer than four reads was 40% (North 1), 21% (North 2), 17% (North 3), 25% (North 4), 9% (East 1), 15% (East 2), 12% (East 3), 12% (East 4), and 16% (East 5). With the polar bear data, the five Alaskan bears (AK1 to AK5) all had much lower coverage, and the coverage for four of the Svalbard bears (PB1, PB2, etc.) was roughly comparable to that of the eastern aye-aye population. Specifically, the fraction of SNPs where the coverage was less than four were 57% (AK1), 61% (AK2), 62% (AK3), 63% (AK4), 89% (AK5), 8% (PB3), 10% (PB4), 11% (PB6), 11% (PB8), and 14% (PB9). Thus, although we can use the bear data to mimic the aye-aye $F_{ST}$ analysis in most respects, including sequencing protocol and the Galaxy commands, sequences coverage differs substantially for one of the populations.

With the aye-aye, we had 3,919,891 SNPs whose estimated genotypes for the North1 to -4 and East1 to -5 were not identical. The restriction that each of those nine individuals has coverage at least 4 and the total coverage not exceeding 120 yields 1,340,685 SNPs. For the polar-bear data, there were 1,564,199 SNPs whose estimated genotypes of AK1-5, PB3 PB4, PB6, PB8, and PB9 were not identical, and 71,301 of them had coverage at least 4 for each individual and total coverage at most 120. Although the two aye-aye populations had $F_{ST} = 0.167$ [with the estimator for Reich et al. (12)], the two polar-bear populations have $F_{ST} = 0.029$, and similarly small values of $F_{ST}$ are obtained for other choices of two polar bear “populations” from the 23 sequenced individuals. Although having deeper coverage data for the Alaskan bears would increase the number of SNPs meeting our bounds on coverage, we see no reason that the resulting computed $F_{ST}$ would change appreciably, much less come close to the aye-aye $F_{ST}$ for the north and east populations. This result provides only a single datapoint for comparison with the aye-aye results, but we feel it strengthens the belief that the observed aye-aye $F_{ST}$ values are somewhat surprising for two populations that live so close together.
Estimating a Rooted Phylogeny and Genetic Diversity from a Synonymous Site Genotype Distance Matrix. To study genetic diversity at synonymous sites and construct rooted neighbor-joining trees with an outgroup sequence (i.e., the human reference genome sequence for aye-aye SNPs, and the aye-aye reference genome for human SNPs), we first aligned all human (hg19) exons (plus 10 bp flanking sequence on each end; according to Ensembl annotations as of July 2012) to the aye-aye reference genome (1) using the LASTZ program (freely available at www.bx.psu.edu/~rsbharris/lastz/) using the parameters “T=2 O=50 E=10 Y=200 K=500” and the following human-lemur substitution scores:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>-21</td>
<td>-9</td>
<td>-26</td>
</tr>
<tr>
<td>C</td>
<td>-21</td>
<td>15</td>
<td>-19</td>
<td>-9</td>
</tr>
<tr>
<td>G</td>
<td>-9</td>
<td>-19</td>
<td>15</td>
<td>-21</td>
</tr>
<tr>
<td>T</td>
<td>-26</td>
<td>-9</td>
<td>-21</td>
<td>10</td>
</tr>
</tbody>
</table>

For each exon we selected the alignment with the most matching nucleotides. We ignored all positions within 20 bp of any alignment gap and we discarded exons where the alignment predicted an internal stop codon in the aye-aye or for which either the nucleotide or inferred amino acid identity was less than 75%. Finally, we discarded cases where the same aye-aye sequence was aligned to more than one human exon. Of 193,176 annotated human exons, this process produced alignments for 144,972 (75.0%).

We evaluated codon sequences to estimate the number of nonsynonymous and synonymous sites for each species and to identify synonymous SNPs. To estimate the number of synonymous sites, we treated each human coding exon as follows. We created an alignment of the human reference sequence to both the most similar aye-aye contig (if any) and a variant of that contig created an alignment of the human reference sequence to both the most similar aye-aye contig (if any) and a variant of that contig included any other alleles that we observed among the 12 human samples. We ignored any part of the human exon sequence for which the coverage from at least one human individual was less than 4 or where the total coverage for all individuals exceeded 120. For each remaining codons in the human reference assembly we performed the following operations. (i) We counted how many of the nine variant codons were synonymous (eventually the total was divided by three to get the number of synonymous sites). (ii) If the codon had one or more differing nucleotides between human and aye-aye but the amino acids were identical, then we counted it as synonymous between human and aye-aye and noted whether that amino acid varied within our aye-aye samples. (iii) If the amino acid differed between the human and aye-aye reference sequences, we counted it as such and noted whether the amino acid varied within our aye-aye samples.

In addition, if a nucleotide but not the corresponding amino acid varied within our aye-aye samples, we used the putatively orthologous human nucleotide to classify the reference aye-aye nucleotide as ancestral (if it was identical to the human reference nucleotide), derived (if the human reference nucleotide was identical to the observed variant aye-aye nucleotide), or “de novo” (otherwise).

For the analogous examination of the human data, we treated each human coding exon as follows. We created an alignment of the human reference sequence to both the most similar aye-aye contig (if any) and a variant of the human exon sequence that included any other alleles that we observed among the 12 human samples. We ignored any part of the human exon sequence for which the coverage from at least one human individual was less than 4 or where the total coverage for all individuals exceeded 120. For each remaining codons in the human reference assembly we performed the following operations. (i) We counted how many of the nine variant codons were synonymous (eventually the total was divided by three to get the number of synonymous sites). (ii) If the codon had one or more differing nucleotides between human and aye-aye but the amino acids were identical, then we counted it as synonymous between human and aye-aye and noted whether that amino acid varied within our human samples. (iii) If the amino acid differed between the human and aye-aye reference sequences, we counted it as such and noted whether the amino acid varied within our human samples.

In addition, if a nucleotide but not the corresponding amino acid varied within our human samples, we used the putatively orthologous aye-aye nucleotide to classify the reference human nucleotide as ancestral (if it is identical to the aye-aye reference nucleotide), derived (if the aye-aye reference nucleotide is identical to the observed variant human nucleotide), or “de novo” (otherwise).

Fig. S1. Principal component analysis of aye-aye population structure. Analyses of the estimated genotypes for 666,256 SNPs with minimum 4× sequence coverage in each of the 12 individuals studied, and maximum 120× coverage in those individuals combined.
Fig. S2. Individual and population SNP coverage levels. Frequency distributions of the number of sequence reads covering SNPs for each individual, and population totals. (A) All aye-ayes including North5. (B) Aye-aye population totals (combined all three populations), excluding North5. (C) Individual and population totals compared between aye-ayes and matched human data.
Cross-species comparisons of average $F_{ST}$

Fig. S3. Aye-aye/ human $F_{ST}$ comparisons at various cut-offs and with different methods for estimating $F_{ST}$. Average $F_{ST}$ values for aye-aye populations and comparative human populations at various minimum per-individual SNP coverage cut-offs and using three different $F_{ST}$ estimators. SNPs that were fixed for the same allele in both populations of a species were excluded (e.g., SNPs that were variable only among the West aye-aye population sample, or between the Aye-aye North ($n = 8$ chr) vs. West ($n = 6$ chr) compared with human Africa ($n = 8$ chr) vs. Asia ($n = 6$ chr). (B) Aye-aye North ($n = 8$ chr) vs. West ($n = 6$ chr) compared with human Europe ($n = 10$ chr) vs. Asia ($n = 6$ chr). (C) Aye-aye East ($n = 10$ chr) vs. West ($n = 6$ chr) compared with human Europe ($n = 10$ chr) vs. Asia ($n = 6$ chr).
Fig. S4. Human population structure. Analyses of estimated SNP genotypes with minimum 4× sequence coverage in each of the 12 human individuals studied, and maximum 120× coverage in those individuals combined. (A) Rooted neighbor-joining tree estimated from a distance matrix based on 2,202 autosomal synonymous SNPs from gene coding regions that could be aligned to the aye-aye reference genome. Pairwise distances were calculated as total SNP genotype distance, with distance for an individual SNP the difference between two individuals’ genotypes scored as 0, 0.5, and 1 (e.g., AA, AT, and TT, respectively). The nucleotide of the aye-aye reference sequence was different from both aye-aye alleles for 163 of the 2,202 SNPs; in these cases the aye-aye genotype was scored as 0.5. As expected, the individuals from each human population cluster, and European and Asian populations cluster most closely. The root separates African from European and Asian populations. (B) Principal component analysis of all 1,146,658 genome-wide human SNPs meeting the coverage requirements. As expected, PC1 clearly distinguishes African individuals from European and Asian individuals. PC2 distinguishes individuals from the latter two populations.

Dataset S1. Aye-aye sequence reads and mapping summary

Dataset S1

Dataset S2. Human sequence reads and mapping summary

Dataset S2

Dataset S3. Human sequence reads detailed accession numbers

Dataset S3