Supporting Information

Perry et al. 10.1073/pnas.1211990110

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/~rsharris/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., megabases) 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. Indexed 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-trim 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.

We used the MarkDuplicates utility in the Picard toolset (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 SNP 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 the Galaxy Web site for public use to reproduce the results presented here or to conduct other analyses. Beyond the analyses presented in this article, other functionality related to ecological and conservation population genomics is also now available on the Galaxy Web site (for example, the selection of restriction enzymes and primers for restriction fragment-length polymorphism-based SNP genotyping). We expect to continue to expand these tools according to user needs and as additional analyses become possible with increases in sequencing capacity.

We evaluated population structure three ways: (*i*) with principal component analysis (PCA) using the program SMARTPCA (7); (*ii*) with model-based estimation of ancestry using ADMIXTURE (8), which produces results similar to the *structure* program (9) but is optimized for genomic-scale datasets; and (*iii*) a neighbor-joining tree based on genotype distances. For the ADMIXTURE analysis, we examined structure after specifying both $k = 2$ and $k = 3$ populations. A pairwise distance matrix was computed in 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 estimating F_{ST} values for each SNP using three different formulas: Reich et al. unbiased estimator (12), Weir and Cockerham'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 simple 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), 69% (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 of 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/~rsharris/lastz/) using the parameters "T=2 O=50 E=10 Y=200 K=300" and the following human-lemur substitution scores:

	A	C	G	T
A	10	−21	−9	−26
C	−21	15	−19	−9
G	−9	−19	15	−21
T	−26	−9	−21	10

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 that included any other alleles that we observed among the 12 aye-aye samples. We ignored any part of the contig for which the coverage from at least one aye-aye individual was less that 4 (i.e., less than four independent reads of a position) or where the total coverage for all individuals exceeded 120. For each remaining putative codons in the aye-aye 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 that 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).

1. Perry GH, et al. (2012) A genome sequence resource for the aye-aye (*Daubentonia madagascariensis*), a nocturnal lemur from Madagascar. *Genome Biol Evol* 4(2):126–135.
2. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
3. Miller W, et al. (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc Natl Acad Sci USA* 109(36):E2382–E2390.
4. Ye L, et al. (2011) A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol* 12(3):R31.
5. Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. *Genome Res* 20(9):1165–1173.
6. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
7. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
8. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
9. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959.
10. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
11. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.
12. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
13. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358–1370.
14. Wright S (1951) The genetical structure of populations. *Ann Eugen* 15(4):323–353.

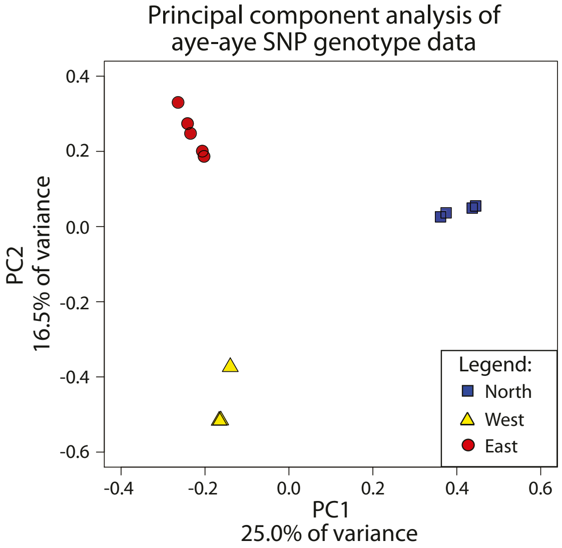


Fig. S1. Principal component analysis of aye-aye population structure. Analyses of the estimated genotypes for 666,256 SNPs with minimum 4× sequence coverage in each of the 12 individuals studied, and maximum 120× coverage in those individuals combined.

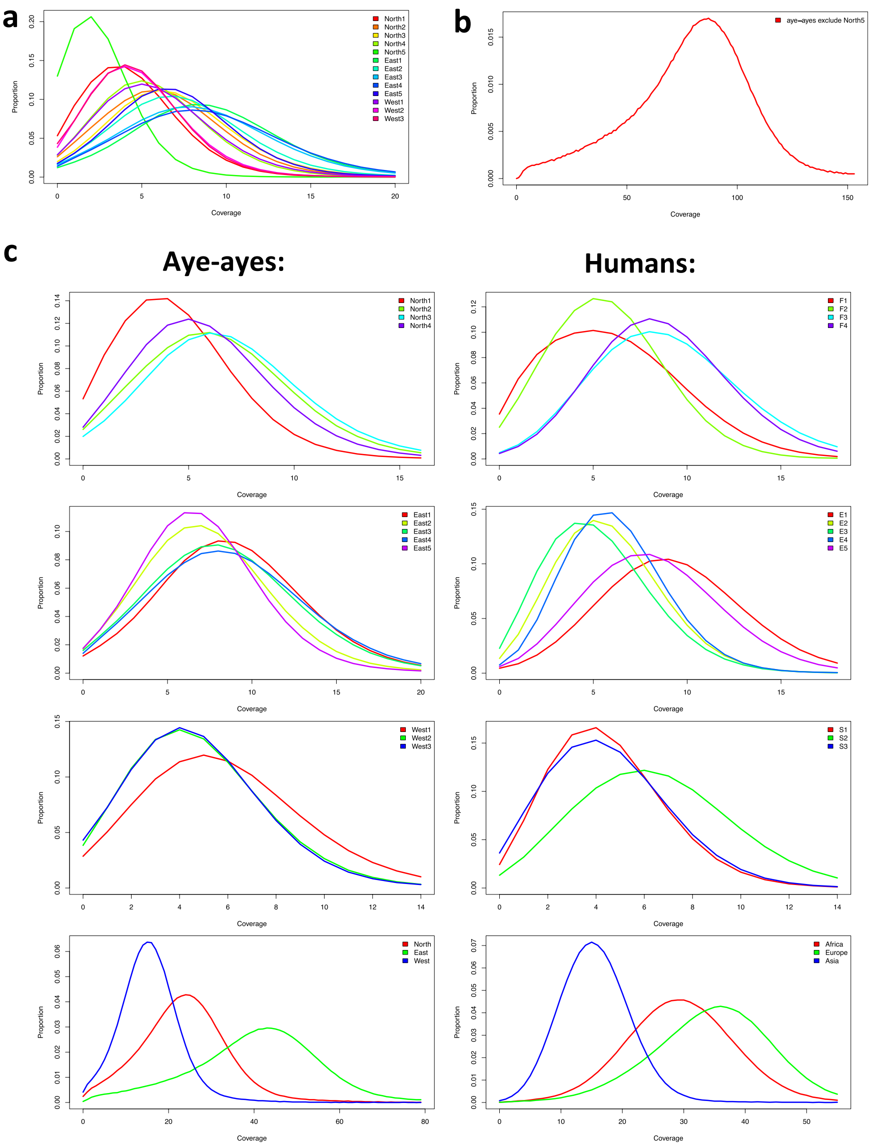


Fig. S2. Individual and population SNP coverage levels. Frequency distributions of the number of sequence reads covering SNPs for each individual, and population totals. (*A*) All aye-ayes including North5. (*B*) Aye-aye population totals (combined all three populations), excluding North5. (*C*) Individual and population totals compared between aye-ayes and matched human data.

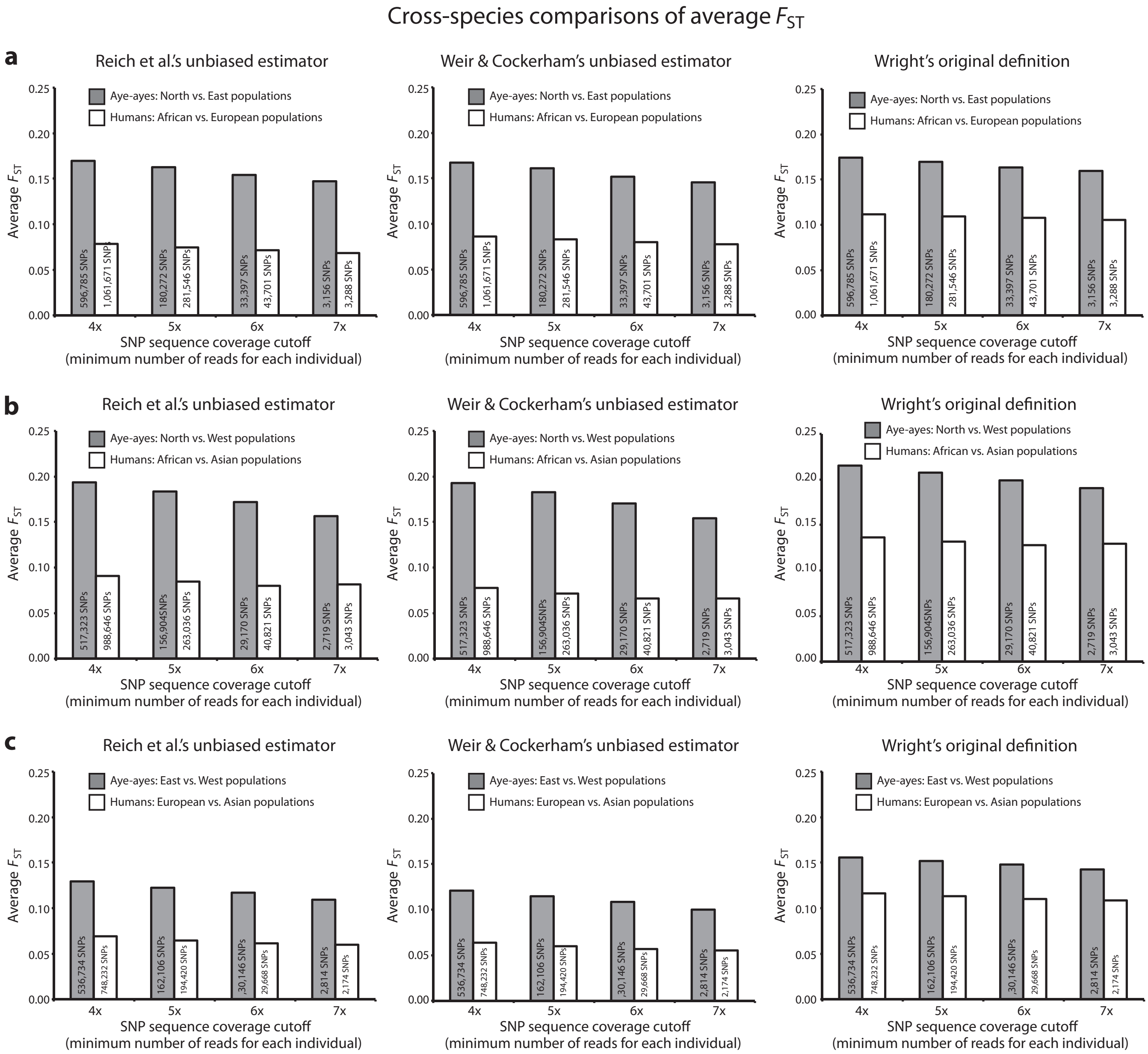
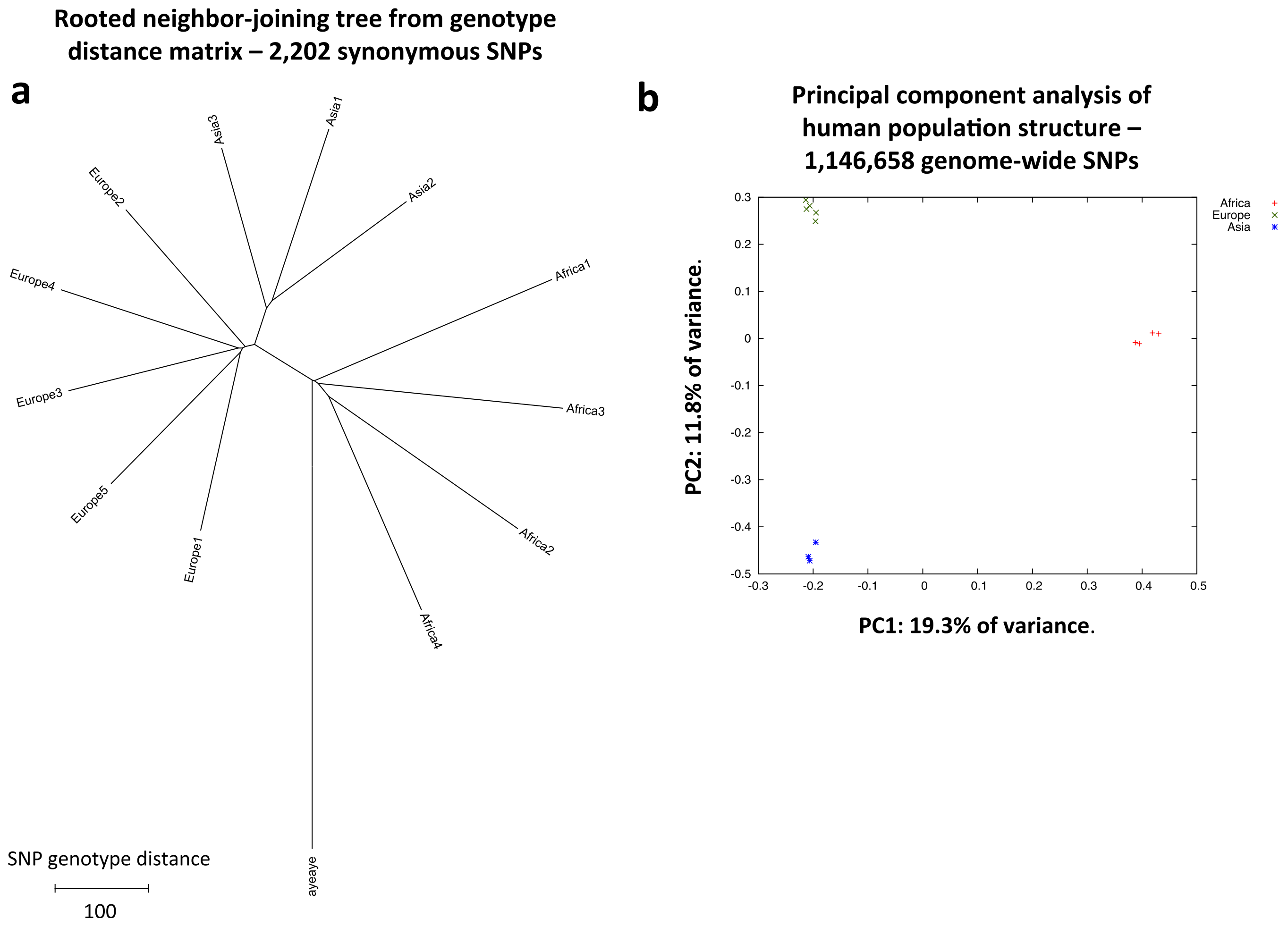


Fig. S3. Aye-aye/ human F_{ST} comparisons at various cut-offs and with different methods for estimating F_{ST}. Average F_{ST} values for aye-aye populations and comparative human populations at various minimum per-individual SNP coverage cut-offs and using three different F_{ST} estimators. SNPs that 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 West populations and the East or North population, would not be included in the North vs. East aye-aye comparison). The numbers of SNPs analyzed in each comparison are listed. (*A*) Aye-aye North (n = 8 chromosomes) vs. East (n = 10 chr) compared with human Africa (n = 8 chr) vs. Europe (n = 10 chr). (*B*) Aye-aye North (n = 8 chr) vs. West (n = 6 chr) compared with human Africa (n = 8 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).



Rooted neighbor-joining tree from genotype distance matrix – 2,202 synonymous SNPs

Principal component analysis of human population structure – 1,146,658 genome-wide SNPs

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