Supplemental Methods for:

Comparative RNA sequencing reveals substantial genetic variation in endangered primates

George H. Perry[1]*, Páll Melsted[1]*, John C. Marioni[1]*, Ying Wang[1]*, Russell Bainer[1]*, Joseph K. Pickrell[1], Katelyn Michelini[2], Sarah Zehr[3], Anne D. Yoder[3,4,5], Matthew Stephens[1,6], Jonathan K. Pritchard[1,2], Yoav Gilad[1]

[1] Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
[2] Howard Hughes Medical Institute, University of Chicago, Chicago, IL 60637, USA
[3] Duke Lemur Center, Duke University, 3705 Erwin Road, Durham, NC 27705, USA
[4] Department of Biology, Duke University, Durham, NC 27708, USA
[5] Department of Evolutionary Anthropology, Duke University, Durham, NC 27708, USA
[6] Department of Statistics, University of Chicago, Chicago, IL 60637, USA

* These authors contributed equally to this work.

OUTLINE

Page:	Section:	
4.	I.	Data access.
4.	II.	Liver tissue samples and library preparation.
5.	III.	Transcript assembly and alignment of orthologs.
13.	IV.	Estimating gene expression levels.
15.	V.	SNP genotyping methods.
24.	VI.	Analysis of changes in exon structure
30.	VII.	Evolutionary analysis of coding region nucleotide sequences.
32.	VIII.	Detecting lineage-specific changes in gene expression levels.
37.	IX.	References.

SUPPLEMENTAL FIGURES

Figure S1: *De novo* transcript assembly.

Figure S2: RNA sample quality.

Figure S3: Nucleotide sequence comparisons of assembled gene transcripts and transcripts from reference genomes.

Figure S4: Estimated expression level comparisons based on assembled gene sequences and transcripts from reference genomes.

Figure S5: Within- and between-species correlations of normalized expression level estimates.

Figure S6. Heatmap depiction of pairwise correlation distance matrix.

Figure S7: Phylogeny reconstruction from nucleotide sequence and gene expression data.

Figure S8. Selection of SNP identification criteria.

Figure S9: Principal components analysis of genome-wide SNP genotype data for the four human individuals in this study, with those from the Human Genome Diversity Panel.

Figure S10: Genetic diversity and the strength of selection.

Figure S11: *KIAA0494* exon usage divergence.

Figure S12: Directional selection on gene expression levels.

Figure S13: *SDR16C5* expression in slow loris and marmoset.

Figure S14: Flow chart of the methods used for de novo assembly and analysis of RNA-seq data.

SUPPLEMENTAL TABLES

Table S1: Species-level SNP diversity estimates.

Table S2: Individual-level heterozygosity estimates.

Table S3: SNP genotyping validation with PCR and Sanger Sequencing.

Table S4: SNP subsample analysis.

Table S5: Candidates for positive selection on amino acid sequences.

Table S6: Candidates for directional selection on gene expression levels.

Table S7: Validation of human SNPs not identified in 1000 Genomes project.

Table S8: Sample information.

I. Data access.

Paired end 76x76 bp sequencing data obtained in this study have been submitted to the NCBI Sequence Read Archive database (http://www.ncbi.nlm.nih.gov/sra; accession number SRA046085). The transcript assembly code used in this study is available at: http://pritch.bsd.uchicago.edu/software.html. The full database of assembled gene sequences, full gene multi-species alignments, orthologous coding region multi-species alignments, lineage-specific d_N/d_S results, normalized gene expression estimates, log likelihood ratios for lineage-specific expression level changes, and the identified SNPs and genotype data for each species are available as a Supplemental Database file on the Genome Research website and at http://giladlab.uchicago.edu/data.html.

II. Liver tissue samples and library preparation.

Common and scientific names for the species in the study are provided in the below table:

Common Name	**Scientific name**
Human	*Homo sapiens*
Chimpanzee	*Pan troglodytes*
Rhesus macaque	*Macaca mulatta*
Vervet (green monkey)	*Chlorocebus aethiops*
Common marmoset	*Callithrix jacchus*
Mohol bushbaby (galago)	*Galago moholi*
Slow loris	*Nycticebus coucang*
Aye-aye	*Daubentonia madagascariensis*
Black and white ruffed lemur	*Varecia variegata variegata*
Coquerel's sifaka	*Propithecus coquereli*
Mongoose lemur	*Eulemur mongoz*
Crowned lemur	*Eulemur coronatus*
Northern treeshrew	*Tupaia belangeri*
House mouse	*Mus musculus domesticus*
Nine-banded armadillo	*Dasypus novemcinctus*
Gray short-tailed opossum (laboratory opossum)	*Monodelphis domestica*

Liver samples were obtained from the National Disease Research Interchange (human), Yerkes National Primate Research Center (chimpanzee, rhesus macaque), Southwest Foundation for Biomedical Research (rhesus macaque, marmoset, opossum), Alpha Genesis

(vervet), Duke Lemur Center (slow loris, bushbaby, aye-aye, Coquerel's sifaka, black and white ruffed lemur, mongoose lemur, crowned lemur), David Fitzpatrick – Duke University (tree shrew), Michael Nachman and Matt Dean – University of Arizona (F1 offspring of wild-born house mouse), Richard Truman – National Hansen's Disease Program, Louisiana State University School of Veterinary Medicine (armadillo). See **Supplemental Table S8** for individual-level information on each sample, including age and sex where known. Liver tissue samples were harvested within 4 hours of death and flash frozen in liquid nitrogen or frozen immediately at -80° C under IRB or IACUC approval of the institutions listed above, and stored at -80° C until RNA extraction. Samples from four unrelated individuals were collected for all species except armadillo (n = 2). For each RNA extraction, 0.1 g tissue was homogenized in Trizol (Invitrogen) and purified according to manufacturer instructions. RNA sample quality was assessed with the Agilent 2100 Bioanalyzer using an RNA 6000 Nano chip (**Supplemental Fig. S2**).

RNA-seq libraries were prepared as described previously (Marioni et al. 2008) except with insert sizes ~380 bp and using Illumina paired-end adapters. Each RNA-seq library was sequenced using one lane of the Illumina Genome Analyzer IIx with paired-end, 76 bp reads (2 x 76 bp), using Cluster Generation Kit v3 and Sequencing Kit v2.

III. Transcript assembly and alignment of orthologs.

While the total number of nucleotides in the transcriptome is only a small fraction of the number of nucleotides in the whole genome, there are several challenges particular to *de novo* gene assembly from RNA-seq data. First, even among appreciably-expressed genes there is an extreme range in sequence coverage. For example, reads from a single gene (*ALB*; Albumin) comprised ~12% of the data in each species, such that coverage levels for sequencing errors at this gene (at a random ~1% sequence error rate) were often greater than those for many other genes that were ultimately assembled. Second, both pre- and mature-mRNA molecules are sequenced. While pre-mRNA represents only a small proportion of reads, this presents a problem because introns often harbor repetitive elements, which in turn can be highly similar to repetitive elements elsewhere in the genome (e.g., in introns from other genes or in 5' and 3' UTRs). Third, alternative splicing can generate multiple transcripts per gene. To overcome these specific computational challenges, we customized a *de novo* assembly approach, described below.

Several programs for transcriptome assembly have been released previously. These include Trans-ABySS (Robertson et al. 2010) and Oases

(http://www.ebi.ac.uk/~zerbino/oases/), which function by interpreting output from established whole genome assemblers, ABySS (Simpson et al. 2009) and Velvet (Zerbino and Birney 2008), respectively. A more recent program, Trinity (Grabherr et al. 2011), performs direct transcriptome assembly. All of these programs use the de Bruijn graph framework. Our assembly method differs from these existing tools in several respects, as described below. The various tools may be more or less appropriate for a particular study, depending on the specific dataset and downstream analysis goals. We have not evaluated and compared the performance of the different algorithms, as this is beyond the scope of our study. However, we note that our algorithm was specifically developed to facilitate subsequent comparative genomic analyses; it uses a sequence similarity-based comparative assembly approach, thereby establishing multi-species gene orthology as a property of the initial assembly. This unique aspect of our approach facilitates direct inter-species comparison of gene sequences and expression levels in an evolutionary framework.

Pre-assembly / correction of Ns. To facilitate the assembly we required that all reads contain only ACGT characters, since this significantly reduces the complexity of the graph assembly problem. To do this one could discard reads with Ns, or shorten them since Ns predominantly occur at the ends of reads. However, in the sequencing of one flowcell the 39th basepair was recorded as N in all reads due to machine error; it was determined that a nucleotide was skipped based on alignment to reference genomes. To recover the data from this one flowcell, we had to develop an effective error-correcting procedure, that is similar in some respects to previously-published approaches (Pevzner et al. 2001; Kelley et al. 2010; Medvedev et al. 2011). Once developed, we decided to apply the procedure for data from all flowcells, not only the flowcell with the 39th basepair error.

Specifically, for each read we recorded all kmers of length k=25, and recorded the coverage of each k-mer (i.e., the number of reads containing that k-mer and its complement) that did not overlap with an N. For each read with an N, we then replaced the N with all four possible nucleotides and recorded the total coverage (the original k-mer coverage values for k-mers without Ns, plus the 4 N-corrected k-mers) of the resulting 25-mers overlapping the N-basepair. The N-basepair was replaced with the nucleotide that produced the highest k-mer coverage in this analysis. To assess the accuracy of this error-correction procedure, we aligned reads from one human and one macaque individual to their respective reference genome transcript databases and compared our corrected nucleotide at each N position to the nucleotide in the reference transcript. We found that our corrected nucleotides in 98.5% of 5.4

million human Ns and 98.3% of 4.5 million rhesus macaque Ns were identical to the corresponding nucleotide in the respective reference genome transcript databases, demonstrating that our approach is effective and highly accurate.

Assembly. The assembly process is divided into 9 steps. The inputs are the N-free reads from the sequencer as well as a reference transcriptome. As a reference transcriptome we used the human RefSeq database (hg19), consisting of 28,098 isoforms from 18,606 genes. For each species, the assembly was performed using the combined RNA-seq reads from all individuals of that species.

1. *Create the de Bruijn graph.* For each read we recorded all k-mers of k=39. At k=39, sequencing errors tend to create tips instead of bubbles in the de Bruijn graph (which are easier to remove or correct than bubbles). Each read of 76 bp generates 38 distinct k-mers.

We found that most k-mers observed only once resulted from sequencing errors or were generated by very lowly-expressed genes that ultimately could not be assembled. These singleton k-mers contributed substantially to the memory overhead of the program. We therefore revised the algorithm to focus on all k-mers with coverage 2 or higher by filtering out uniquely occurring kmers using a Bloom Filter approach (Melsted and Pritchard 2011).

To build the de Bruijn graph, we recorded the coverage of each k-mer (i.e., the number of reads containing that k-mer and its complement), and all adjacent k-mers. Once the de Bruijn graph was constructed, we simplified the graph by correcting obvious sequencing errors appearing in the graph as short tips, in a manner similar to Velvet's error correction (Zerbino and Birney 2008).

2*. K-mer to gene alignment.* Our goal in this step was to identify regions of small-scale similarity in the de Bruijn graph generated for each species, to human RefSeq gene sequences. These homologous regions were used to set general expectations for transcript coverage levels and to isolate the portion of the graph likely to contain individual gene sequences. We note that for the ultimate assembly of a transcript, it is not necessary to identify homology across the entire gene length. The assembly process can proceed from a limited number of homologous regions for a given gene, because of the interconnectedness of de Bruijn graph contigs that correspond to a particular expressed transcript. Moreover, we note that while our approach relies on the

maintenance of some degree of sequence similarity in gene coding regions between non-human species and humans, our simulations show that our assembly method is robust to relative gains and losses of internal exons in non-human species (see **Analysis of Changes in Exon Structure**, below).

For each human reference gene, we collapsed all of the associated isoforms to a single consensus sequence that included all coding region exons arranged by their genomic order. Then, for each species, every k-mer in the graph was aligned against the reference sequence (one gene at a time) and we recorded all matches. To speed up the matching process we required an exact match of short length (i.e., a seed), similar to the heuristic used by BLAT:

(i) For all reference genes we kept a map of seeds to (gene, position) pairs.

(ii) For each k-mer we evaluated whether there were any matching seeds.

(iii) Given a match, we evaluated whether there was a full local alignment for that k-mer to the gene.

We varied the length of the seeds and the number of mismatches allowed, taking into account the sequence divergence between species. In this procedure, it is important to set the length of the seed appropriately. If the seed length is too small, then there might be too many spurious hits, which will greatly increase the computational requirements. If the seed length is too large, then we might miss regions with higher divergence from the reference sequence. We similarly controlled the maximum sequence divergence between homologous regions by limiting the number of mismatches: Allowing too few mismatches would limit our ability to assemble a large proportion of genes; allowing too many mismatches could result in improperly assembled sequences.

To set these parameters appropriately, we estimated the expected coding region sequence divergence between humans and all other species, based on available data. For species for which such data were not available, we used approximate divergence dates to estimate sequence divergence. This analysis was based on phylogenetic clades of species with similar distance from human. For example, all strepsirrhine primates were assumed to have similar divergence from humans.

We then used the sequence divergence estimates to establish the number of mismatches allowed in the search for potentially orthologous k-mers. One would expect a range of sequence divergence values among k-mers along a particular gene. Matching does not need to be complete for the appropriate contigs to be identified from the graph to achieve successful gene assembly. The table below provides the seed lengths,

mismatches allowed, and corresponding divergence cutoff, used for each species in the 'k-mer to gene' alignment process. For comparison, we also show the actual coding region sequence similarity to human assembled sequences, based on concatenation of multi-species alignments (performed with FSA (Bradley et al. 2009)) of the common set of 515 genes that were ultimately assembled for all 16 species in the study.

Species	Seed length	Mismatches	Divergence cutoff	Actual sequence similarity
Human	33	1	97.4	100.0
Chimpanzee	27	2	94.8	99.6
Macaque	21	3	92.3	98.0
Vervet	21	3	92.3	98.0
Marmoset	21	3	92.3	96.5
Slow Loris	18	3	92.3	93.0
Bushbaby	21	3	92.3	92.9
Aye-aye	18	3	92.3	94.4
Sifaka	18	3	92.3	93.9
Black & white ruffed lemur	18	3	92.3	93.9
Mongoose Lemur	18	3	92.3	93.9
Crowned Lemur	18	3	92.3	93.9
Tree Shrew	14	5	87.2	91.9
Mouse	15	4	89.7	88.9
Armadillo	15	4	89.7	91.5
Opossum	15	4	89.7	83.7

Our next goal was to identify the path through the de Bruijn graph that corresponded to each gene and to extract the transcript sequence (steps 3-9). These steps were repeated for each gene (and species) in isolation.

***3**. Determine coverage levels.* For each gene, we estimated the expression level for that gene in the species of interest, using the set of k-mers that were homologous to the reference sequence. In order to be robust to species-specific exon loss or alternative splicing, we used the 90th percentile of these coverage estimates as the representative value for the gene coverage level for later filtering steps in the assembly.

4. Graph filtering. Based on the coverage estimated in step 3, we set a coverage threshold that was 10% of the estimated value, but not less than a coverage of 3 reads. These thresholds were established by experimentation and by assessment against known reference genome transcripts for rhesus macaque and mouse.

Let *c* be the coverage threshold and *G* be the de Bruijn graph constructed in step 1. The following steps work with the filtered de Bruijn graph G_c , which consists of all nodes with coverage greater than or equal to *c*. The reason for this filtering step is to remove parts of the graph that are irrelevant, such as sequencing errors and contigs representing intronic bleed-through (from pre-mRNA). Furthermore, this step simplifies the graph, as repetitive elements are more common in introns than exons. Since the de Bruijn graph *G* is shared by all genes, we do not modify the graph *G*, but apply the filtering as part of the per-gene assembly process.

5. *Initial contig generation.* Starting from G_c we identified all contigs containing the associated k-mers from step 2. Note that contigs in G_c will generally be longer than contigs in *G*, since low coverage k-mers are removed and thus the graph has less branching.

6*. Graph exploration.* Starting from the contigs from step 5, we identified the connected components, defined as all contigs containing the associated k-mers from step 2, plus all linked contigs. We note that sequencing errors and intronic sequences have likely been removed in the step 4 coverage-based filtering. Therefore, at this point we would not expect the total length of the connected contigs to greatly exceed that of the human reference gene. We set a threshold of 10 times the length of the human reference sequence, which we established by experimentation. If the total number of observed k-mers in the connected contigs exceeded this threshold, then we rejected the graph and considered the assembly failed for that gene in this species. Otherwise, the subsequent filtering steps, described below, were generally sufficient to achieve successful transcript assembly.

7. *Error correction.* Once the local de Bruijn graph was constructed, we removed tips and popped bubbles in the local graph in a similar fashion as performed in Velvet (Zerbino and Birney 2008). That is, when there were multiple paths through the graph caused by SNPs or any remaining sequencing errors that were not removed by the

coverage-based filtering, we compared coverage levels between alternate paths and removed the one with lower coverage. Note that when popping bubbles, we required that the two alternative paths differed by at most 10 bp in length. Thus, this step will remove bubbles that are the result of sequencing errors, SNPs, and short indels, but not bubbles that reflect alternative splicing of exons. However, we acknowledge that contigs reflecting alternatively spliced exons would have been removed in step 4 if they were expressed at much lower levels than the rest of the transcript.

***8**. Path analysis.* Once the local de Bruijn graph was simplified, we considered each possible remaining path that was at least 50% of the length of the coding region of the corresponding human reference orthologous sequence. Each path was then aligned against the reference sequence using the Fast Statistical Alignment algorithm (Bradley et al. 2009), and the best match was selected as the largest number of total aligned bases.

***9**. Consensus sequence.* If the proportion of nucleotides in the remaining path through the graph (after alignment to the human reference sequence) was above a set threshold (see table below), then we used this path to generate a consensus sequence for the gene. Otherwise, we rejected the path and generated no consensus sequence, considering the assembly failed for that gene. The alignment thresholds used in this step are described in the following table:

Species	**Alignment threshold**
Human, Chimpanzee, Macaque, Vervet, Marmoset	80%
Black and White Ruffed Lemur, Mongoose Lemur, Crowned Lemur, Sifaka, Aye-Aye, Slow Loris, Bushbaby	70%
Armadillo, Tree Shrew, Mouse, Opossum	60%

Post-assembly. We removed potentially erroneously-identified paralogous genes from the assembly by aligning each sequence against all genes in the set of human reference genes using BLAST. Any consensus sequence for which the top BLAST hit was not the associated gene, was removed. The final number of genes assembled per species and the number of potentially paralogous genes that were removed from the dataset are given in the table below.

In addition, we performed a similar analysis for high sequence identity to pseudogenes, by downloading a list of human pseudogenes from the Ensembl database (version 64) with the Biomart tool, using the pseudogene filter (Flicek et al. 2011). We then repeated the BLAST-based analysis used to identify potentially erroneously-identified paralogous genes with the combined set of human RefSeq genes and pseudogenes. The number of genes identified by this analysis, between only 24 and 135 genes per species, is provided in the table below.

We confirmed that had we removed from further analysis the genes identified based on similarity to pseudogenes, none of the specific results we discuss would be affected (positive selection examples, enrichment analyses, etc.). Moreover, as the removal of such a small proportion of genes (0.4% - 2.7%, depending on the species) does not affect any of the general patterns we reported on either (e.g., based on genetic diversity estimates), in final analysis, we decided not to apply the filter for pseudogenes in the analysis presented in the paper.

Indeed, we are unsure that a pseudogene filter is truly effective, because there are three practical differences between filters for similarity across paralogs and pseudogenes: (i) comparative studies have shown that a large fraction of the ~22,000 pseudogenes in the human genome are not shared with more distantly-related primates including New World Monkeys and especially strepsirrhines (e.g., bushbabies, lorises, and lemurs) (Zheng et al. 2007). For more distant species especially, filtering against these genes results in the erroneous removal of orthologous genes that might be one or a few nucleotides more similar to a pseudogene, by chance. (ii) Pseudogene sequences evolve faster than functional genes. Thus, ancient pseudogenes, shared across species, are not likely to be recognized by our initial assembly approach in the first place. (iii) Finally, pseudogenes, recognizable or not, are only a problem for our analysis if they are expressed, and only a minority of pseudogenes are expressed at detectable levels.

Species	Assembled genes	Paralogous genes (removed)	Pseudogene filter (not removed)
Human	5,523	239	24
Chimp	5,294	333	44
Macaque	5,497	478	65

Vervet	5,646	431	76
Marmoset	5,313	437	71
Slow Loris	4,789	549	83
Bushbaby	2,680	311	64
AyeAye	5,339	588	89
Sifaka	5,646	630	98
Black & white ruffed lemur	5,627	640	106
Mongoose Lemur	5,443	637	96
Crowned Lemur	5,487	668	99
Tree Shrew	5,924	883	122
Mouse	5,239	792	129
Armadillo	5,625	971	107
Opossum	4,851	658	135

IV. Estimating gene expression levels.

To estimate the expression level of each gene, for each sample we first aligned the sequenced reads against a reference containing the sequences of the set of assembled genes for the appropriate species using BWA (Li and Durbin 2009) with default parameters, considering only uniquely mapped reads. For this analysis, we analyzed separately the two reads of each pair.

Individual reads not aligned in the first step were evaluated using a gapped alignment approach (Pickrell et al. 2010b), to account for potential alternative splicing. To do so, we independently aligned the first and last 20 bp of each remaining read to all transcripts using BWA, and performed the following steps to determine whether the alignment supports an exon-junction read.

- If both ends of a read mapped to the same transcript, then we internally extended each alignment. If the internal sequence could not be fully extended, then we discarded the read.
- If both ends of a read mapped to only one transcript and could be fully extended, then we included the read in the expression estimate for that gene. If there were full extensions for the read to multiple transcripts, then we discarded the read (this would be considered a non-unique alignment).
- If only one end of a read mapped to the transcript, then we internally extended that alignment as far as possible and searched the transcript for a perfect match to the remainder of the read if the remainder of the read was at least 10 bp. If there was a perfect match, then the read was kept. If the remainder of the read

was less than 10 bp or could not be perfectly matched if at least 10 bp, then the read was discarded.

For our evolutionary analysis of gene expression levels, we chose to consider orthologous gene regions across species rather than the fully assembled gene sequence from each species. That is, if the full gene sequence was not assembled for every species, then we restricted our analysis to the specific region of the gene that *was* commonly assembled across species. This approach makes it less likely that our inter-species comparison of gene expression levels would be affected by sequencing biases or the inclusion of alternatively spliced exons in some species only. To do so, we performed a multi-species alignment (Bradley et al. 2009) and identified the maximum orthologous region that was fully aligned across all species. Reads contributing to a gene's expression level were restricted to those falling in the maximum orthologous region, which was itself constrained to exclude non-coding regions (i.e., UTRs were not included in the gene expression analysis).

We used the total number of reads mapping to the identified orthologous region of a transcript (including exon-junction reads) as a measure of its expression level. Having done this, we next normalized these data *within* each species using the following steps. We first divided the expression level of each transcript by the transcript's length. Second, since it has previously been shown that gene expression levels measured in different lanes of RNA-sequencing can show systematic differences that are correlated with the gene's GC content, we accounted for this using the following procedure, motivated by the method described by Pickrell et al. (Pickrell et al. 2010a): Within each species, let z_{ij} denote the expression level of transcript *j* in individual *i.* Let $y_{ij} = z_{ij}/\Sigma_i z_{ij}$ be the proportion of reads mapping to transcript *j* in individual *i*. Subsequently, we regressed y_{ij} against g_j, where g_j is the GC content (measured as a proportion) of transcript j. If f_{ij} is the fitted value from the loess regression for transcript *j* of individual *i,* we calculated the normalized expression value for transcript *j* for individual *i* as: $x_{ij} = (\text{mean}_j(f_{ij})/f_{ij}) * z_{ij}$

Having performed the within-species corrections, we next normalized the data to account for differences *between* species. This is necessary to ensure that estimates of gene expression levels are comparable across species. The challenge, however, is that different gene sets were assembled in different species. Our approach proceeded as follows. First, using expression measurements for all assembled transcripts, we calculated the average expression value for each gene, across all species. Second, we ranked these "average" expression values to create a synthetic distribution, and calculated the median of this distribution (medSyn). Third, for each sample, we calculated a normalization factor NormSamp = medSyn/(sampleMedian) and adjusted the expression of all assembled transcripts using this factor so that the median

was equal to medSyn for all samples (Bullard et al. 2010).

We took advantage of the availability of high quality sequenced genomes for 6 of the species in our study to evaluate the quality of our gene expression estimates based on the *de novo* assembly approach. To do so, we compared the *de novo* assembly-based estimates to corresponding gene expression estimates based on a more typical read alignment analysis against transcript sequences that were predicted from the reference genome sequences of human, chimpanzee, rhesus macaque, marmoset, mouse, and opossum (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/multiz46way/alignments/). We found that gene expression estimates based on the two approaches were highly correlated (**Supplemental Fig. S4**).

We also evaluated pairwise correlations of normalized gene expression levels between all pairs of individuals and species in the study. As expected, we observed higher correlation estimates for within-species comparisons than for comparisons of individuals between species, with lower correlations for pairs of more distantly-related species (anthropoid versus strepsirrhine primates, primates versus non-primates; **Supplemental Fig. S5**). The estimated average gene expression levels across species are also more strongly correlated between more closely-related pairs of species (**Supplemental Fig. S6**).

V. SNP genotyping.

Special considerations for SNP identification from RNA-seq data. We interrogated the RNA-seq reads from each individual to infer the positions of single nucleotide polymorphisms (SNPs) in the corresponding genomic DNA. A genomic position for which an individual is heterozygous may theoretically be identified from either high-coverage genomic DNA or RNA-seq data, when the proportion of sequencing reads with each of the two variant nucleotides is ~0.5, allowing for sampling variation. With RNA-seq data, however, SNP identification is more complex than with genomic DNA sequencing data. First, there is the issue of allele-specific expression (Pastinen 2010), when transcripts are not expressed at equal levels from each of an individual's two chromosomes. Second, one needs to consider the possible effects of RNA editing (Li et al. 2011).

It should be noted that since the primary goal of our analysis was to compare the estimates of genetic diversity among the 16 species in our study, these two issues would not be expected to affect the observed *relative* differences between species, unless rates of allele-specific expression or RNA editing varied significantly across taxa. However, our preference was to establish parameters for SNP identification such that neither allele-specific expression nor RNA editing had substantial adverse effects on our ability to accurately call SNPs from

RNA-seq data. To assess the accuracy of our approach (described below), we made the following assessments and validations:

(i) Compared the human SNPs we identified to genome-wide SNP genotype data collected from genomic DNA for the same human samples in this study using the Illumina 1M-Duo SNP array platform.

(ii) Compared the human SNPs we identified to SNP position calls from the 1000 Genomes project (The 1000 Genomes Project Consortium 2010) that were based on genomic DNA resequencing data from 179 individuals from four human populations.

(iii) Used PCR and Sanger sequencing, performed on genomic DNA for the same samples used in this study, to validate a subset of SNPs in four species: human, rhesus macaque, Coquerel's sifaka, and black and white ruffed lemur.

The results of these comparisons demonstrated that our SNP identification approach was highly accurate (see *SNP genotyping validation*, below).

SNP identification. We first aligned all reads from each individual to the database of consensus sequence transcripts that we assembled for that species (see above), using the default parameters of BWA (Li and Durbin 2009). In the final preparation step of the RNA-seq libraries, there is a PCR amplification step that uses the ligated adapter sequences as primer sites for consistent amplification. To help limit any bias from PCR amplification in the SNP identification process, we performed a read-filtering step to consider only one read pair from each uniquely-aligned starting position and strand. Specifically:

(i) If two paired reads each had the same start position for read 1, but different start positions from read 2, then these reads were considered to have originated independently and were both kept in the analysis.

(ii) When more than one paired read had identical aligned start positions from each read, we kept one read at random and excluded the remaining reads from further analysis. For this filtering decision, we ignored the alignment quality score, as single nucleotide differences from the consensus sequence due to true SNPs could have subtle effects on that score. We did not consider any base call with a Phred-scaled quality score < 30.

To establish SNP identification criteria, we systematically assessed genotyping accuracy as a function of multiple different per-strand coverage requirements and "SNP call definitions" based on the proportion of the most common nucleotide at each site. By "SNP call definition",

we mean the threshold at which a heterozygous site would be called, when the proportion of reads with the most common nucleotide at a given position was at or below that threshold (for reads aligning to *both* strands). By requiring the SNP definition to be met by reads mapped to each strand, we limited the effects of potential strand-specific sequencing biases (Nakamura et al. 2011). Examples of SNP call definitions that we considered were ≤ 0.6, ≤ 0.65, ≤ 0.7, ≤ 0.75, etc.

To determine the coverage requirement and SNP call definition thresholds, we compared SNP genotypes from the 1M-Duo Illumina SNP array platform data collected for each of the four human samples in the study, to the variants inferred from the RNAseq data using our method (**Supplemental Fig. S8**). Based on this analysis, we chose to assess all sites covered by a minimum of 15 sequence reads per strand (minimum of 30 total reads) and, of such sites, we classified as heterozygous those for which the proportion of the most common nucleotide was ≤ 0.7 on each strand. This general approach for SNP calling is similar to that which we previously used with genomic DNA sequencing data and found to result in highly accurate SNP identification (Perry et al. 2010).

Finally, we performed a sub-sampling analysis with the reads from each individual. For this analysis, reads were randomly distributed into two subsets. SNPs were identified from each subset of the data using the coverage and SNP call definition threshold criteria described above. We then determined the consistency of SNP inferences in the subsampled data within each individual. We removed three samples, one chimpanzee and two aye-ayes, from further SNP analysis due to relatively low concordance in heterozygous site identification in the subsample analysis (**Supplemental Table S4**).

For each individual, we classified each analyzable site as either (i) non-coding (UTR) / undefined, (ii) nonsynonymous (amino acid changing), or (iii) synonymous by inferring the corresponding codon positions for each base within the transcript (see **Section VII**, below). Sites could be assigned fractional synonymous and nonsynonymous values (totaling 1), when change to some of the 3 possible alternative nucleotides would affect an amino acid substitution and others would not.

All identified SNPs were classified as noncoding/undefined, nonsynonymous, or synonymous, with no fractional assignments, based on the two variable nucleotides. Genes on the human X chromosome were filtered from further analysis for all species. Heterozygous site summary statistics (including heterozygosity estimates) for each individual sample in the study are provided in **Supplemental Table S2**.

For each species, we estimated genotypes for all sites with sufficient coverage for SNP identification in all individuals (n = 2 for armadillo and aye-aye, n = 3 for chimpanzee, n = 4 for all other species). We classified all heterozygous positions as well as any sites with homozygous differences between individuals as SNPs. SNP position and genotype data for each species are available at: http://giladlab.uchicago.edu/data.html. Species-level estimates of genetic diversity π (average pairwise genetic distance) and θ (sample-size corrected proportion of segregating sites) were computed for all genes with at least 100 sites with sufficient coverage for SNP identification in each individual of that species and are provided in **Supplemental Table S1**. For the six species with published reference genome sequences (human, chimpanzee, rhesus macaque, marmoset, mouse, opossum), we asked whether the accuracy of our assembled transcript sequences compared to the gene sequences from the reference genomes (**Supplemental Fig. S3**) affected SNP diversity estimates. To do so, we estimated genetic diversity for each of these six species separately for all genes with accuracy ≥ 99% and for all genes with accuracy < 99% (for genes with at least 100 sites with sufficient coverage ofr SNP identification in each individual of that species). As reported in the below table, diversity estimates are similar for the two datasets, supporting the inclusion of all genes in the analysis of genetic diversity.

Species	**genes with >= 99% accuracy between assembled and reference transcripts**				**genes with < 99% accuracy between assembled and reference transcripts**			
	Genes	**Synonymous sites**	**Syn SNPs**	**π (syn)**	**Genes**	**Synonymous sites**	**Syn SNPs**	**π (syn)**
Human	1047	166506.0	533	0.117%	5	1250.3	11	0.360%
Chimpanzee	1168	181993.3	892	0.204%	33	6969.3	45	0.253%
Rhesus macaque	1042	169485.0	1262	0.288%	37	5910.0	43	0.263%
Marmoset	978	155894.7	514	0.122%	68	11009.7	41	0.140%
Mouse	1199	191742.7	1911	0.370%	136	25239.0	238	0.353%
Opossum	768	128597.0	494	0.161%	124	21260.7	97	0.183%

For several of the species in our study, estimates of neutral genetic diversity from the nuclear genome (autosomal chromosomes) have been published previously. Our estimates for these species are generally comparable to those previous estimates, as shown in the below table:

Species	π (synon sites), this study	Comments	π (presumably neutral sites), previous studies	Previous study refs.
Human	0.119%	Our sample includes one individual who is primarily of African descent (synon. site heterozygosity = 0.126%; versus 0.086%, 0.087%, and 0.091% for non-African individuals)	0.121% (Biaka, Africa) 0.087% (Basque, Europe) 0.081% (Han, SE Asia) 0.110% (Hausa, Africa) 0.085% (Italian, Europe) 0.079% (Chinese, SE Asia)	(Wall et al. 2008) (Wall et al. 2008) (Wall et al. 2008) (Voight et al. 2005) (Voight et al. 2005) (Voight et al. 2005)
Chimpanzee	0.206%	Synon. site heterozygosity estimates for 3 chimpanzee individuals = 0.089%, 0.090%, and 0.216%. The ancestry of the third individual is at least partially, and perhaps wholly, from the central chimpanzee subspecies, based on mtDNA HVI sequencing (see below).	0.082% (western) 0.130% (central) 0.12% (western) 0.15% (central) 0.081% (western)	(Yu et al. 2003) (Yu et al. 2003) (Fischer et al. 2004) (Fischer et al. 2004) (Perry et al. 2010)
Rhesus macaque	0.288%		0.174% (Chinese) 0.141% (Indian)	(Hernandez et al. 2007) (Hernandez et al. 2007)
Aye-aye	0.073%		0.081%	(Perry et al. 2007)
Mouse	0.368%	Our samples are from F1 mice of unrelated wild-born parents caught around Tuscon, AZ.	0.325% (Iran) 0.260% (France) 0.126% (Germany)	(Baines and Harr 2007) (Baines and Harr 2007) (Baines and Harr 2007)

The chimpanzee mtDNA HVI type was classified with PCR and Sanger sequencing of genomic DNA extracted from the same liver sample used in the RNA-seq analysis. Primers used were (both 5'-3') CTCTGTTCTTTCATGGGGAAGC and CGGGATATTGATTTCACGGAGG. The obtained sequence was included in an analysis with HVI sequences from wild-born chimpanzees of known capture location and subspecies (Stone et al. 2002). Specifically, a Neighbor Joining tree was estimated using MEGA4 (Tamura et al. 2007); this individual's mtDNA HVI sequence was located within a cluster of individuals from the central chimpanzee subspecies (*Pan troglodytes troglodytes*). We were unable to determine whether this chimpanzee individual was wild-born or captive-born.

SNP genotyping validation. We specifically wanted to assess whether allele-specific expression and RNA editing, or some unknown source of error, may have affected our SNP calls. Ultimately, we expect that heterozygous positions in transcripts with extreme allele-specific expression bias (e.g., imprinted genes) would have not been identified by our approach. However, based on our external comparison to the genome-wide SNP genotype data from

genomic DNA of the same individuals, this issue did not have a large effect on our ability to call true SNPs, at least in humans (overall 99% accurate identification of heterozygous sites; range 98.1% to 99.7% for the four human individuals; see table below). This result indicates that our false-negative SNP calling rate is low. The following table shows genotype consistency between RNA-seq SNP calls and 1M-Duo Illumina genotypes, for sites with sufficient coverage for SNP identification in the RNA-seq data in each individual human sample:

Sample	Total sites	Consistent genotypes	% consistent	Heterozygous sites	Consistent genotypes	% consistent
Hsa58876	2099	2088	99.5%	291	290	99.7%
Hsa56295	1986	1973	99.3%	336	333	99.1%
Hsa56655	1016	1010	99.4%	177	175	98.9%
Hsa56720	1298	1287	99.2%	216	212	98.1%

In addition, our expectation was that the majority of SNPs identified among the four human samples in our study would also have been observed in the larger sample of the 1000 Genomes project, which is comprised of 59 Yoruba individauls from Ibadan, Nigeria (YRI), 60 European-Americans from Utah (CEPH), 30 Han Chinese individuals from Beijing (CHB), and 30 Japanese individuals from Tokyo (JPT). For this analysis, we again focused on the 1,272 SNPs for which there was sufficient coverage to identify variants in all four human individuals. To translate the transcript-relative coordinates in our assembly to the genomic coordinates used in the 1000 Genomes database, we used BLAST to identify the location of the sequences surrounding each variant in the 1000 Genomes assembly. Specifically, for each SNP site, we separately localized the flanking 80bp immediately upstream and downstream of the variant, and excluded from further analysis any sites for which both sequences did not localize to a single position, or for which there was no single perfect match in either of the flanking sequences. Of the resulting 833 variant sites that were identified in our RNA sequencing data with corresponding unambiguous positions in the 1000 Genomes assembly, 704 (84.5%) were annotated as SNPs in the 1000 Genomes database (2010_07 release) (The 1000 Genomes Project Consortium 2010).

The remaining 129 SNPs (15.5%) may be:

- False-positives in our dataset.

- False-negatives in the 1000 Genomes data, perhaps related to the low per-individual sequence coverage for most individuals in the 1000 Genomes project and the potential under-identification of low frequency SNPs
- Rare SNPs present in one of the four individuals in our study but not in any of the 1000 Genomes individuals. There are likely to be few SNPs falling into this category, according to neutral population genetic theory.

To distinguish among these possibilities, we used genomic DNA-based PCR and Sanger sequencing to validate 11 putative SNPs, from 11 different genes, that we identified from our RNA-seq data but that were not annotated in the 1000 Genomes database. Genomic DNA primers were designed such that at least one primer was intronic. Primers and results are provided in **Supplemental Table S7**. Of the 11 putative SNPs, nine (82%) were confirmed by Sanger sequencing, suggesting that these SNPs were likely rare SNPs or false negatives in the 1000 Genomes data. Therefore, on the basis of the broader comparison to the 1000 Genomes Project database we conclude that our false-positive SNP calling rate is likely low, and, indirectly, that RNA editing does not appreciably affect our ability to accurately identify SNPs or our genetic diversity estimates.

Finally, we used PCR and Sanger sequencing to validate small subsets of SNPs identified in each of four species: human, rhesus macaque, Coquerel's sifaka, and black and white ruffed lemur. These validations were performed using as template genomic DNA extracted from the same liver samples used for RNA-seq analysis. The selected SNPs were each from a different coding region, and located at least 50 bp from each end of an exon to aid in primer design, and a maximum of one SNP per gene was validated. Otherwise, the SNPs were chosen at random, without respect to properties of the SNP call. We successfully validated 21/23, 15/16, 21/23, and 15/19 SNPs for human, rhesus macaque, Coquerel's sifaka, and black and white lemur, respectively (the unequal number of validation attempts per species is due to different numbers of failed PCR primers across the species), demonstrating the high accuracy of our SNP calling approach. The assayed SNPs, primers used, and results are provided in **Supplemental Table S3**. These validation results demonstrate the accuracy of our SNP calling approach, even for species relatively divergent from humans and with high levels of estimated genetic diversity (i.e., black and white lemur and Coquerel's sifaka).

We do not believe that biases potentially associated with phylogenetic distance from humans (our assembly process was based on homology to human RefSeq gene sequences) have adversely affected our ability to make relative comparisons of SNP diversity among the species in our study. If SNP diversity estimates were strongly affected by such a bias, then we

would expect species of similar distance to humans (e.g., all lemurs) to be similarly affected. Yet, lemurs include both aye-aye, with the lowest genetic diversity estimate of any primate species (and similar to that reported in a previous study based on PCR and Sanger sequencing, as reported in the table above), and Coquerel's sifaka, with the highest diversity estimate of any primate, and for which we successfully validated 21/23 SNPs to demonstrate the accuracy of the high genetic diversity estimate for this species.

Population structure. To analyze the ancestry of the human samples in the study, we used principal components analysis (PCA) to compare our samples to those from a sample of worldwide humans with known ancestry. We first combined the Illumina 1M-Duo SNP genotype data of our samples with that from the Human Genome Diversity Panel (Li et al. 2008), and then thinned the data by removing all SNPs with an r^2 value greater than 0.1 in a sliding window of 50 SNPs. This was done using the option "--indep-pairwise 50 10 0.1" in PLINK (Purcell et al. 2007). We then performed principal components analysis (PCA), using the implementation in smartpca (Price et al. 2006). Based on the PCA, two of the individuals are primarily of European ancestry, one individual is primarily of East Asian ancestry, and one individual primarily of African ancestry (**Supplemental Fig. S9**). As expected, synonymous site heterozygosity was highest for the individual primarily of African ancestry, π = 0.126%, versus 0.086% for the individual primarily of East Asian ancestry and 0.087% and 0.091% for the two European American individuals (**Supplemental Table S2**).

We also asked whether the relatively high estimates of genetic diversity for Coquerel's sifaka and the black and white ruffed lemur, as well as the relatively low genetic diversity estimate for aye-ayes, might have been affected by population structure and sampling, or by outbreeding strategies in captive populations. While we sampled unrelated individuals from each species, some individuals were born in the wild and some individuals were born in captivity. Therefore, if founder individuals for the sampled captive sifaka and ruffed lemur were captured from very different populations with substantial between-population genetic differentiation, then our high estimates of genetic diversity in these species could reflect population structure rather than the genetic diversity from a typical population sample. For aye-ayes, because they have the largest species range of any extant lemur (Mittermeier et al. 2010), we asked whether the observed low genetic diversity estimate was reflective of the species as a whole, or due to restrictive sampling of founder individuals. To address these issues, we considered the capture locations of founders and pedigrees for the individual samples included in our study, based on information maintained by the Duke Lemur Center, along with individual-

level estimates of synonymous site heterozygosity (**Supplemental Table S2**).

Our estimate of synonymous site genetic diversity (π) for Coquerel's sifakas is 0.681%. Of the four individuals from this species in our study, one was wild-born and the other three had two wild-caught parents, each from different capture locations. These capture locations are not from opposite ends of the species range, but with little *a priori* understanding of population structure among populations in this species, we cannot automatically equate physical distance to levels of population differentiation. The individual-level synonymous site heterozygosity estimate for the wild-caught individual is 0.533%, compared to 0.615%, 0.642%, and 0.671% for the three captive born individuals. Under neutrality and in the absence of population structure, species-level genetic diversity and individual-level heterozygosity estimates are expected to be similar. Therefore, the higher species-level estimate of genetic diversity, combined with the higher estimates of individual-level heterozygosity in the captive-born individuals, suggests non-negligible population structure in Coquerel's sifakas. Such population structure could be of interest in future conservation genetic studies and conservation planning. However, even the synonymous site heterozygosity estimate of the single wild-born individual (0.533%) is nearly five times the estimated genetic diversity in humans (which itself is slightly elevated due to some structure in that population sample, as discussed above), and it is greater than or similar to the estimated nucleotide diversity levels of any other species in this study (the second largest species estimate is for tree shrew; synonymous site π = 0.532%). Furthermore, microsatellite data from a different endangered sifaka species, the golden-crowned sifaka (*Propithecus tattersalli*), also indicate relatively high genetic diversity (Quemere et al. 2010), suggesting that high genetic diversity might not be unusual for sifakas and providing indirect support for our finding.

Two of the black and white ruffed lemur individuals in our sample were wild-born, and these individuals have synonymous site heterozygosity estimates of 0.258% and 0.309%. These estimates are similar to those of the two captive-born individuals in the study: 0.222% and 0.294%. All individual-level values are high compared to the species-level estimates of most of the other primates. However, while the original capture locations for the two wild-caught individuals and the wild-caught ancestors of the two captive individuals are not well known, black and white lemur species-level synonymous site π is 0.375%, higher than the estimated heterozygosity of any individual. Similar to our analysis in Coquerel's sifakas, this result also suggests the possibility of non-negligible wild population structure in black and white ruffed lemurs that may be of interest to conservation biologists.

Of the two aye-aye individuals included in the SNP analysis, one was wild-born,

captured from Northeast Madagascar. The other individual was captive-born with two wild-born parents, one parent from a similar region in Northeast Madagascar as the wild-born individual in our sample, but the other parent from Northwest Madagascar. Synonymous site heterozygosity is 0.067% for the wild-born individual and 0.080% for the captive-born individual. Thus, while the parents of the captive-born individual are not from the extreme ends of the aye-aye species distribution, this result suggests that diversity of aye-ayes is quite low, and that the results we have reported for this species are not likely to reflect sampling from a single region in which genetic diversity happens be unusually low for the species.

Variation at functional versus putatively neutral sites. For each species, we observed substantially higher levels of genetic diversity at synonymous sites than at nonsynonymous, or amino acid changing, sites (**Supplemental Table S1**). This observation is expected, given that most nonsynonymous mutations are thought to be deleterious and therefore would be removed or kept below intermediate frequencies by purifying selection, while most synonymous sites are presumed to be neutral. We did observe variation in the ratio of nonsynonymous to synonymous site diversity across species. This ratio could be considered a measure of the efficiency of purifying selection, where a smaller ratio is consistent with stronger purifying selection against nonsynonymous mutations. We observed a negative relationship between synonymous site diversity and the ratio of nonsynonymous to synonymous site diversity (**Supplemental Fig. S10A**), as predicted by population genetic theory (Kimura et al. 1963). Humans have the second lowest level of synonymous site diversity (next to aye-aye) of any species in the study, and the largest ratio of nonsynonymous to synonymous site diversity. Conversely, the Coquerel's sifaka has the highest level of synonymous site diversity of any primate in the study, and the lowest ratio of nonsynonymous to synonymous site diversity.

VI. Analysis of changes in exon structure.

Exon structure differences based on multi-species alignment of assembled transcripts. After multi-species alignment (Bradley et al. 2009) of each gene, we identified 308 internal gaps of ≥ 50 bp, which consisted of assembled sequence in at least one species but no sequence in at least one other species. These potential between-species exon structure differences were then evaluated in greater detail:

(i) To determine whether there were single or multiple paths (i.e., alternative splicing) through the de Bruijn graphs for these genes for the different species around the gapped positions.

(ii) To consider any known evidence of alternatively-spliced exons based on the AltEvents database from the UCSC genome browser (hg19).

(iii) To visually inspect the FSA alignments for quality.

We found that 304 of the 308 gaps were either associated with evidence for alternative splicing or could be explained by alignment error. Alignment errors were especially prevalent in the UTR regions for opossum versus other species. Thus, there were only four remaining and potentially fixed inter-species exon structure differences, observed in the following genes and species: *CAST* (exon skipped in opossum; consistent with N-SCAN prediction), *CD46* (exon skipped in black and white ruffed lemur), *FAM149B1* (exon skipped in Armadillo), *TRIM35* (exon skipped in opossum).

We selected one gene for further study, *KIAA0494*, which exhibited a multi-species alignment gap in a clear phylogenetic pattern, but for which there was also evidence for alternative splicing. Specifically, *KIAA0494* exon 8 was not assembled for any strepsirrhine primate but was assembled for non-strepsirrhine primates. However, multiple paths through the de Bruijn graph around exon 8 were observed for both strepsirrhines and non-strepsirrhine primates. Starting with amplification primers designed from exon 8 regions that were conserved across non-lemur species in our study, we PCR amplified and Sanger sequenced exon 8 from each of the five lemurs in the study, using genomic DNA extracted from the liver of one individual per species. The sequences for *KIAA0494* exons 6, 7, 8, and 9 of each species (lemurs and non-lemurs) were then used to construct a database of all potential junction read sequences, comprised of 60 bp from the 3' end of the upstream exon + 60 bp from the 5' end of the downstream exon. We then used BWA (Li and Durbin 2009) to align all 76 bp reads for each species against this database to generate precise exon junction read counts (**Fig. 3A**).

Additionally, we performed quantitative PCR on cDNA synthesized from total RNA using two pairs of primers that were 100% conserved among the consensus sequences of human, rhesus macaque, vervet, black and white ruffed lemur, and Coquerel's sifaka. The forward primer of one pair spanned the junction between exons 7 and 9, while the reverse primer was located wholly in exon 9, which exhibited no signs of alternative splicing (F 5' TCCTTCAGCATGAAAGAAGATA 3'; R 5' GCCTGTTGCTCTCAGGTTTG 3'). The forward and reverse primers of the second pair were both in exon 9 (F 5' CAAACCTGAGAGCAACAGGC 3'; R TGAAAATTTTGGCAATGCTG). Samples were run in triplicate in 25 μL reactions using iQ SYBR Green Supermix (Bio-Rad) with a BioRad iCycler Thermal Cycler (**Supplemental Fig. S11**).

Power to detect between-species exon differences from assembled transcripts. Our *de novo* assembly strategy was designed to isolate portions of the de Bruijn graph based on anchors of homology to human RefSeq gene sequences, while still allowing for accurate reconstruction of internal between-species exon usage differences. To estimate our power to detect exon structure changes in the data, we simulated an inclusion and loss of an exon with respect to the Human reference.

To simulate a novel exon (with respect to human) we selected 1,132 assembled genes in rhesus macaque and mouse, and masked the anchors of homology for each exon by excluding all k-mers that overlapped the exon in the human sequence and 10 basepairs into the adjacent exons. We then attempted to assemble the gene with the reduced set of anchors and count the number of successful assemblies as those that correctly included the exon without human anchor sequence. We reconstructed 10,762 out of 11,017 (97.7%) rhesus macaque exons and 9,400 out of 9,993 (94.1%) mouse exons. In the following table, we further break down the fraction of reconstructed exons by expression quantiles from the normalized expression levels. These results suggest slightly reduced power to detect exon structure differences for lower-expressed transcripts:

Expression quintile	Reconstructed exons	
	Rhesus macaque	Mouse
Lowest 20%	4887 of 5020 (97.3%)	4097 of 4409 (92.9%)
20-40%	2558 of 2629 (97.3%)	1908 of 2003 (95.2%)
40-60%	2462 of 2509 (98.2%)	1766 of 1861 (94.9%)
60-80%	754 of 758 (99.4%)	1230 of 1298 (94.8%)
80-100%	101 of 101 (100%)	399 of 422 (94.5%)

To simulate the loss of an exon in non-human species, we took 1,000 genes in the human reference sequence and included a 150 basepair sequence from the intron of the gene. The sequence of the intron was screened for repetitive elements using RepeatMasker, resulting in 733 novel exons inserted. To test the robustness of the method, we included the modified sequence in the search for anchors (step 2 in the assembly methods, described above) and attempted to assemble the modified genes. We correctly assembled 567 of 590 rhesus macaque genes (96.6%) and 434 out of 441 mouse genes (98.4%).

Our estimate of the power to detect a novel or absent exon indicates that the observation of a lack of fixed inter-species gene structure changes cannot be attributed to the use of the human reference transcript sequence to identify anchors for assembly.

Genome assembly-based exon structure analysis. The inference that inter-species fixed changes in exon structure are rare is intriguing. We wanted to provide further support for this inference. Specifically, since we were not able to assemble each gene in all 16 species, or necessarily the entire gene in each species in which it was assembled, we also performed a second, reference genome-based analysis, to evaluate exon structure conservation. Here, we considered our RNA-seq data against the human, chimpanzee, rhesus macaque, and mouse reference genome sequences to identify and compare annotated and novel exon splice junctions between human and each non-human species with an available high quality sequenced genome. In each species, exons were annotated based on the sequencing data in two steps:

(i) Reciprocal alignment of human Ensembl exons across the human and non-human genomes.

(ii) The identification of novel exons spliced to Ensembl exons based on the presence of one or more junction-spanning reads in our RNA-seq data.

Once exons were annotated, we counted the number of spliced-junction reads entering, leaving, and skipping each exon, and identified significant between-species exon skip rate differences using a logistic regression model. To do so, we first compiled, for each human gene with at least three exons in Ensembl, a set of non-overlapping exons (taking the union of exons in each annotated transcript). Then, for each species: human (using genome assembly hg18), chimpanzee (panTro2), rhesus macaque (rheMac2), and mouse (mm9), we performed the following steps:

1. For non-human species, we used liftOver to map the coordinates of each exon in humans to the coordinates in the other species. We then used liftOver to map these coordinates back to human. We removed all exons for which the liftOver analysis to the non-human species did not return a unique match for either splice site or for which the reciprocal liftOver analysis back to humans did not return the original coordinates. We then removed all genes for which more than 50% of the exons were filtered. Using this approach, we kept 20,872, 19,751, and 16,829 genes for analysis in chimpanzee, rhesus macaque, and mouse, respectively.

2. We mapped the sequencing reads from each lane to the corresponding genome and used the algorithm in Pickrell et al. (Pickrell et al. 2010b), with a small modification, to identify splice junctions from the reads that did not map to the genome. In Pickrell et al. (Pickrell et al. 2010b), the authors used 20 base pairs from each end of a read as seed alignments to identify splice junctions. Here, because our original reads were of longer length (76 bp) than in the previous study, we used 20, 35, and 50 bases from each end of a read as seed alignments, and merged the results, selecting the longest successful seed alignment. We then filtered these alignments as in Pickrell et al. (Pickrell et al. 2010b). We kept all alignments where GT-AG or GC-AG dinucleotides appeared immediately intronic of the putative junction, and merged reads corresponding to the same putative junction. This procedure resulted in a classification of a set of splice junctions in each species. The numbers of identified splice junctions were similar across species.

3. We used the splice junctions classified above to identify new exons, which were either unannotated in human or specific to the non-human species. We first identified all 5' and 3' splice sites that appeared between 20 and 700 bases of each other, and considered these as putative exons. Testing this procedure in humans, we were able to rediscover >90% of internal exons in genes expressed above the median expression level. Additionally, we identified all regions of the genome with putative expression (one or more mapped reads), and considered these as putative exons. These definitions of putative exons are purposefully liberal. We then identified all putative exons that (i) did not overlap an annotated exon and (ii) showed evidence of splicing to an annotated exon, and then filtered these newly identified exons by the criteria in step 1 above. All exons meeting these criteria were included in our subsequent analysis, resulting in the addition of 1,833, 9,983, and 7,695 exons to the sets of analyzed chimpanzee, rhesus macaque, and mouse exons, respectively (out of 216,308, 213,610, and 192,859 total exons).

4. For each exon in each gene, we counted, for each individual, the number of reads covering each end of the exon and the number of reads skipping the exon. These counts were our primary data for the analysis of changes in alternative splicing.

We then compared levels of alternative splicing between humans and the non-human species using the following procedure:

1. We simulated 20 base pair reads tiling each exon in both species, and mapped these back to the corresponding genome. In each species, the fraction of reads mapping back uniquely to the correct exon were considered the "mappability" score of the exon. All exons with a difference in mappability of >10% were excluded from further analysis.

2. To limit our initial analysis to exons that showed reliable evidence of alternative splicing, we removed all exons with fewer than 10 reads covering junctions from both the 5' and 3' end of the exon summed across all individuals of both species, or fewer than a total of 8 reads either entering, exiting, or skipping the exon in each species

3. We defined exon skip rate as the number of reads skipping the exon divided by the number of reads skipping the exon plus half the number of reads covering either end. This is the fraction shown in **Fig. 3B**.

We used a generalized linear mixed effects model to identify exons with significant between-species skip rate differences:

$$\text{logit}(f_{ijk}) = \alpha_i + \beta_{ij} + \gamma_{ik}$$

where *i* indexes exon, *j* indexes species, and k indexes individual. α is the intercept term, β is the fixed effect for the species, and γ is a random effect for each individual. f_{ijk} is the inclusion fraction of the exon in each individual, estimated as the mean of the reads entering and exiting the exon, divided by that mean plus the number skipping the exon. The model was fit using the lme4 package in R.

Consistent with the results from our analysis of the 16-species assembled gene transcripts, we observed very few exons that were skipped always in one species but never in the other (only three, five, and 19 exons in the chimpanzee-human, rhesus macaque-human, and mouse-human comparisons, respectively). While evidence of alternative splicing of these exons could still be uncovered in future studies using higher-coverage RNA-seq data, the general relationship between absolute exon skip rate differences and gene expression levels (**Fig. 3C**) also raises another possibility – that within-tissue splicing levels may simply be less conserved for lower-expressed genes.

Our finding of only a small number of fixed changes in gene structures across the 16 species in our study is somewhat unexpected, given that a previous human-mouse bioinformatic and transcript comparison by Modrek et al. (Modrek and Lee 2003) suggested a relatively much larger number of genome-wide fixed differences in exon usage. Moreover, previous analyses of within-species RNA-seq data have shown that there are many mutations that create new exons or destroy existing ones (Pickrell et al. 2010b). Thus, it appears that the raw material for exon structure differences exists in the form of random mutations that affect splicing, yet such mutations are only rarely fixed. We note that both of our analyses – using the *de novo* assembled transcript alignments and the genome-based exon skip rates – are focused on well-aligned, orthologous genes. Furthermore, our genome-based analysis only considers exons that can be reciprocally aligned between two genomes. Thus, a wholly deleted exon, or one no longer used and thus potentially not subject to selective constraint in one species (resulting in considerable nucleotide sequence divergence – a potential consideration especially for our human-mouse analysis given the high neutral substitution rates), may not have been considered. *We note, however, that such internal exon structure differences would have been apparent in the de novo assembly alignment analysis. In addition, our simulations suggested that we would have the power to detect most such differences (as described above).* In contrast, in the Modrek et al. study, if mouse sequences orthologous to an expressed human exon could not be identified, then those exons would have been considered present in one species and absent in the other (Modrek and Lee 2003), potentially explaining their findings. Finally, repetitive element-based exons, which may be the predominant mode of novel exon generation (Keren et al. 2010), may not have been reliably detected by either of our analyses. The presence of such exons may cause our de novo assembly process to fail, and in the genome-based analysis such exons may fail in reciprocal alignment.

VII. Evolutionary analysis of coding region nucleotide sequences.

For each gene, we used FSA (Bradley et al. 2009) to generate a multi-species alignment. We next removed 5' and 3' UTRs and start and stop codons, and corrected any out-of-frame indels introduced in the alignment step. To do so, we evaluated the trimmed alignments against the human RefSeq transcript coding sequence. When the frame of the aligned sequences was shifted with respect to the human reference transcript, we introduced 1-2 bp gaps in all sequences as a correction. The software we used for evolutionary analysis of the coding region nucleotide sequences ignores any 3 bp codons that contain indels in any species in the tree (Yang 2007); therefore, this adjustment does not produce any analysis artifacts.

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 (d_N/d_S) can be examined to make inferences about long-term selective pressures on amino acid sequences. Small d_N/d_S ratios (significantly < 1) are consistent with long-term purifying selection (functional constraint) on amino acid sequences (i.e., nonsynonymous mutations have been fixed at a much lower rate than relatively neutral synonymous mutations, presumably because most such mutations had detrimental effects on fitness). A d_N/d_S ratio ~1 is consistent with neutral evolution of amino acid sequences, while a d_N/d_S ratio significantly > 1 may reflect a past, long-term history of positive selection on amino acid substitutions (i.e., more nonsynonymous mutations were fixed than expected under neutrality, presumably because a subset of these mutations were advantageous).

For each lineage, we considered only those genes for which the branch in question could be reconstructed and studied properly in our evolutionary model, given the availability of the orthologous gene sequence for other species in the tree (recall that we were not able to assemble all genes from all species). We also required that sequence data be available for at least two outgroup species relative to the branch in question, in order to increase the confidence in ancestral sequence reconstruction. For example:

- For analysis of sequence changes on the human-specific branch, we required the availability of sequence from human and chimpanzee, as well as any other species in the tree.
- For analysis of sequence changes on the ancestral primate branch, we required the availability of sequences from at least one strepsirrhine primate, at least one anthropoid primate, tree shrew, and at least one other non-primate outgroup species besides tree shrew (mouse, armadillo, or opossum).

We also restricted our analysis to genes with at least 150 assembled and aligned synonymous sites, and then removed genes with one or more branches for which the gene-specific d_S value was ≥ 99% of the genes for that branch, as such cases may result from alignment artifacts.

We calculated two d_N/d_S ratios for each remaining gene in each lineage. First, the conventional d_N/d_S value, where d_S is based on the synonymous substitution rate for the individual gene on that branch. In addition, we calculated a single genome-wide d_S value for each lineage, as the total number of synonymous substitutions summed across all genes available for analysis on that lineage divided by the total number of synonymous sites summed across all genes. The second d_N/d_S value uses the genome-wide d_S value, which may be

valuable for interpreting results on shorter branches, on which there may have not been any synonymous substitutions for some genes.

We considered genes and branches that meet any of the following conditions as potential candidates to have evolved under directional selection: $d_N/d_S > 1$ and d_N/d_Sgenome > 1, $d_N/d_S > 1$ and ≥ 2 synonymous substitutions (on short branches, stochasticity in the number of synonymous substitutions can lead to large numbers of genes with $d_N/d_S > 1$), or d_N/d_Sgenome > 1.5 (**Supplemental Table S5**). While it is probable that not all genes and branches meeting these conditions were subjected to positive selection at the amino acid sequence level, this set of candidates is likely enriched for such genes. However, we note that the power of this test diminishes with increasing branch lengths, where the large number of synonymous substitutions likely overwhelms even the strongest signals of repeated nonsynonymous fixations. For example, based on our criteria, we did not identify candidate genes for the mouse and treeshrew branches, and only two genes were identified for the aye-aye lineage (*C11orf24*, *FOXRED1*). These are three of the longest branches in our phylogeny excluding armadillo and opossum, which were not included in this analysis because they could not meet the requirements for the presence of outgroup sequences.

We assessed patterns of variation in genome-wide d_N/d_S rates across lineages (namely, summing across all genes in each lineage). Lower genome-wide d_N/d_S values reflect greater efficiency of long-term purifying selection. We compared this measure of the strength of long-term purifying selection against the shorter-term measure, the ratio of nonynonymous to synonymous site diversity, for each lineage/species. We observed a positive relationship (**Supplemental Fig. S10B**). We further examined the data for indications of changes in effective population size. For example, the Coquerel's sifaka has a higher-than-expected estimate of genome-wide d_N/d_S, given the low ratio of nonsynonymous to synonymous site diversity observed for this species. This observation is consistent with a long-term effective population size that was smaller than that of more recent times. Humans have the highest nonsynonymous to synonymous genetic diversity ratios and the highest genome-wide d_N/d_S estimates of any species/ lineage in the study, followed by chimpanzee and aye-aye, consistent with relatively low efficiencies of purifying selection throughout the histories of these lineages.

VIII. Detecting lineage-specific changes in gene expression levels.

The evolution of expression levels of orthologous genes from multiple species can be modeled as a Brownian motion process along the branches of the phylogeny (e.g., Bedford and Hartl 2009). In the null model, we consider that all genes evolve by Brownian motion along the same

tree (i.e., with shared branch lengths), but that the rate of evolution (i.e., the accumulated variance along each branch) varies across genes, since genes may differ in their level of constraint or the mutational target size of regulatory regions. Specifically, let v_i be the length of branch i, and let σ_j^2 be the variance parameter for gene j. Then the variance of the Brownian motion process for gene j on branch i is defined as $v_i\sigma_j^2$. Note that there is statistical nonidentifiability between σ^2 and the branch lengths v (specifically, multiplying σ^2 by a constant c and dividing all the branch lengths by c produces the same model). To deal with this we estimated the branch lengths subject to the constraint that the mean value of σ_j^2 across all genes was 1.

We next used the property that Brownian motion along the branches of a tree is equivalent to a multivariate normal process, in which the covariance matrix Σ is a simple function of the tree topology and branch lengths (Felsenstein 1973). Then, for each gene, if μ_j denotes the expression level for the common ancestor of all species in the tree (i.e., the root of the expression tree), and y_j denotes the vector of gene expression levels for gene j from all sampled individuals, the expression data y_j are distributed as a multivariate normal distribution:

$$y_j = \mathrm{MVN}_s(\mu_j 1, \sigma_j^2 \Sigma(v)), \qquad (1)$$

where s denotes the number of sampled individuals, and Σ is a simple function of the branch lengths v.

For our analysis, we first estimated the branch lengths of a common overall gene expression tree against which expression changes for individual genes were later evaluated. These branch lengths were estimated using the known phylogeny of the species in our study. We modeled the data from the separate individuals in each species as star trees at the tips of each branch of the species tree. In this context, the branch lengths leading to individuals are estimates of the variance of each individual relative to the species means, and these reflect both genetic variation within species as well as random technical or environmental inter-individual variation. We used this approach to model inter-individual variation because it allows us to incorporate the uncertainty in the species means in a straightforward way. The branch lengths were fitted using all genes for which expression estimates were available for all 15 species (i.e., excluding bushbaby), with gene-specific μ and σ^2.

In the above equation, σ^2 is assumed to be constant across the expression tree for each gene. However, we are specifically interested in identifying genes whose expression patterns across the phylogeny deviate from the model shown in equation 1. In particular, we are

interested in genes with unusually large changes in expression on particular branches. Such changes might be detected either as individual species whose expression level for a particular gene is very different from the other most closely related species (in the case of selection on terminal branches) or as clades of species whose expression levels differ greatly from the rest of the tree. These are genes for which the standard model with a single evolutionary rate σ_j^2 does not fit the data well. While we do not expect all such large deviations from the overall gene expression tree to reflect directional selection on gene expression levels, genes ranked at the top of the list in each branch – namely, genes with the largest lineage-specific shifts in expression levels – are likely to be enriched for genes whose expression levels were affected by directional selection.

Specifically, our alternative model is that if gene *j* has undergone selection on branch *i*, then the variance on branch *j* is $Cv_i\sigma_j^2$, where C=1 corresponds to the null model, and C>1 implies that the gene has evolved faster than expected on branch *i*. Under this alternative model,

$$y_j = \mathrm{MVN}_s(\mu_j 1, \sigma_j^2 \Sigma(v, Z_j, C_j)), \tag{2}$$

where Z indicates the index of the branch involved for gene *j*, and C indicates the constant factor multiplied to the length of Z_j. The likelihood ratio of the alternative and null models is

$$LR(Z_j) = \frac{L(\hat{\mu}_j, \hat{\sigma}_j^2, \hat{C}_j;\ y_j, Z_j)}{L(\hat{\mu}_j, \hat{\sigma}_j^2;\ y_j)} \tag{3}$$

Large values of the likelihood ratio indicate support for the alternative model that there has been a change in rate on branch Z_j. Standard likelihood theory suggests that twice the log likelihood ratio on a particular branch should be approximately X^2-distributed so that, for example, when the null hypothesis is true we might expect that the log LR will exceed 2.0 with probability ~0.05. However, since we recognize that the Brownian motion model is an imperfect representation of reality, we prefer to treat the likelihood ratio as a method of ranking the genes with the most unusual patterns of evolution on each branch, rather than as a formal test of significance.

Likelihood ratios were calculated over all branches, for all genes with expression estimates in at least 6 species (6,494 genes). We analyzed genes based on criteria similar to those described in the d_N/d_S analysis methods. For each lineage, we considered only genes for which the branch in question could be isolated properly, given data availability for other species in the tree. We also required data to be available for at least two outgroup species (relative to

the branch in question) in order to increase our confidence that gene expression changes could be isolated to the correct branches. For each branch, we defined directional selection candidates as those with likelihood ratio values $\geq$ 10 (and estimated C > 1) on that branch and < 5 on all other branches (**Supplemental Table S6**). We emphasize that these genes should be considered candidates for directional selection. While the predominant signal in our data is consistent with the known phylogeny (**Figure 1b, Supplemental Figure S6**) even at deep lineage divergences, and thus overall, the patterns observed support a robust genetic influence on gene regulation, our data do not allow us to distinguish between gene expression differences reflecting environmental versus genetic change on an individual gene basis. Specifically, individuals from some species in our study are more likely to have shared a similar environment (e.g., diets; Somel et al. 2008) than individuals from other species. Moreover, while sexes are known for most of the individuals in the study, ages and causes of death are not (**Supplemental Table S8**), other factors that also may influence gene expression levels (Franz et al. 2005; Fraser et al. 2005). If these environmental differences influence gene expression levels, then they may be indistinguishable from genetic effects in our study.

As mentioned in the main text, we excluded the bushbaby data from the gene expression analysis because because fewer genes (2,680) were assembled for bushbaby than for any other species. This possibly reflects the lower RNA quality in the bushbaby samples (**Supplemental Fig. S2**).

We performed functional enrichment analyses for the sets of directional selection candidates on each branch using the Gene Ontology database of functional annotations (Ashburner et al. 2000) and the Gene Set Analysis Toolkit V2 (Duncan et al. 2010), with the background set of genes that met the criteria for analysis described above. We observed highly significant enrichments for genes that function in the peroxisome in the ancestral primate lineage (9 genes observed; 0.5 genes expected; FDR = 7 x 10^{-9}; genes *PEX7, HACL1, IDE, SCP2, PEX13, LONP2, ACOX3, MGST1*, and *PHYH*) and for those with oxygen transport functions in the marmoset lineage (3 genes observed; 0.03 genes expected; FDR = 0.0001; genes *HBA2, HBB*, and *CYBG*). We note that there is not yet strong literature support for the peroxisome function of *MGST1*, as otherwise recorded in the Gene Ontology database. It is notable that *PHYH* is a candidate positive selection gene at the amino acid sequence level in several ancestral lemur lineages (**Supplemental Table S5**), suggesting the possibility of adaptive evolution at multiple levels and in multiple lineages in this gene.

Other genes with marked lineage-specific changes in expression levels identified in **Supplemental Table S6**, or others in the dataset with unusual between-species expression

patterns, may also underlie species and lineage-specific adaptations of evolutionary significance, especially as related to diet and the metabolism and detoxification processes of the liver. For example, while our inability to assemble a given gene in a given species does not necessarily indicate that this gene is not appreciably expressed (see **Transcript assembly and alignment of orthologs**, above), we were interested to observe that we successfully assembled the *SDR16C5* gene for slow loris only, and not any other species in the study. To examine this observation in more detail, we also considered the number of reads aligned to predicted *SDR16C5* transcripts based on the reference genome sequences of human, chimpanzee, rhesus macaque, marmoset, mouse, and opossum. We utilized the same approach that was used to compare gene expression estimates based on read alignments to assembled genes versus alignment reference genome transcripts, presented in **Supplemental Fig. S5**. Besides slow loris, we found that *SDR16C5* is only expressed at appreciable levels in the marmoset, but still at considerably lower levels than slow lorises (**Supplemental Fig. S13**).

SDR16C5, an epidermal retinol dehydrogenase, is involved in the first, rate-limiting step of retinol (Vitamin A) metabolism (Matsuzaka et al. 2002; Lee et al. 2009). Retinol is a derivative of isoprene, the monomer of latex. Slow lorises feed extensively on tree exudates (Tan and Drake 2001; Swapna et al. 2010), which may include gums, saps, and latex. Exudativory is relatively rare among non-primate mammals, but among primates, several independent taxa including marmosets (and slow lorises), have apparent craniofacial adaptations for tree gouging, and specialize on exudates (Nash 1986; Vinyard et al. 2003). It is not known how exudates are digested in the primates, but this process is thought to be aided by bacterial fermentation in the gut (Power and Myers 2009). In this case, there may be large quantities of the digestive products, such as retinol, absorbed through the large intestine, which may then be filtered by the liver. The expression of *SDR16C5* in the liver tissues of slow loris and mamoset could represent convergent adaptation against the fitness-reducing effects of vitamin A toxicity. Such hypotheses based on single-gene observations should be considered highly tenuous. Still, this information may be valuable if it ultimately leads to further study and a better understanding of diet-related adaptations and evolutionary ecology among non-human primates, especially, in this case, incorporating the activity and importance of the gut microbiome to diet.

IX. References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. *Nat Genet* **25**(1): 25-29.

Baines JF, Harr B. 2007. Reduced X-linked diversity in derived populations of house mice. *Genetics* **175**(4): 1911-1921.

Bedford T, Hartl DL. 2009. Optimization of gene expression by natural selection. *Proc Natl Acad Sci U S A* **106**(4): 1133-1138.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol* **5**(5): e1000392.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**: 94.

Duncan DT, Prodduturi N, Zhang B. 2010. WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. *BMC Bioinformatics* **11**(Suppl 4): P10.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet* **25**(5): 471-492.

Fischer A, Wiebe V, Paabo S, Przeworski M. 2004. Evidence for a complex demographic history of chimpanzees. *Mol Biol Evol* **21**(5): 799-808.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S et al. 2011. Ensembl 2011. *Nucleic Acids Res* **39**(Database issue): D800-806.

Franz H, Ullmann C, Becker A, Ryan M, Bahn S, Arendt T, Simon M, Paabo S, Khaitovich P. 2005. Systematic analysis of gene expression in human brains before and after death. *Genome Biol* **6**(13): R112.

Fraser HB, Khaitovich P, Plotkin JB, Paabo S, Eisen MB. 2005. Aging and gene expression in the primate brain. *PLoS Biol* **3**(9): e274.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*.

Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, Nazareth L, Indap A, Bourquin T, McPherson J et al. 2007. Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* **316**(5822): 240-243.

Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* **11**(11): R116.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**(5): 345-355.

Kimura M, Maruyama T, Crow JF. 1963. The Mutation Load in Small Populations. *Genetics* **48**: 1303-1312.

Lee SA, Belyaeva OV, Kedishvili NY. 2009. Biochemical characterization of human epidermal retinol dehydrogenase 2. *Chem Biol Interact* **178**(1-3): 182-187.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866): 1100-1104.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**(6038): 53-58.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**(9): 1509-1517.

Matsuzaka Y, Okamoto K, Tsuji H, Mabuchi T, Ozawa A, Tamiya G, Inoko H. 2002. Identification of the hRDH-E2 gene, a novel member of the SDR family, and its increased expression in psoriatic lesion. *Biochem Biophys Res Commun* **297**(5): 1171-1180.

Medvedev P, Scott E, Kakaradov B, Pevzner P. 2011. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics* **27**(13): i137-i141.

Melsted P, Pritchard JK. 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* **in press**.

Mittermeier RA, Louis EE, Richardson M, Schwitzer C, Langrand O, Rylands AB, Hawkins F, Rajaobelina S, Ratsimbazafy J, Rasoloarison R et al. 2010. *Lemurs of Madagascar*. Conservation International, Arlington, VA.

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* **34**(2): 177-180.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H et al. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**(13): e90.

Nash LT. 1986. Dietary, behavioral, and morphological aspects of gummivory in primates. *Yrbk Phys Anthropol* **29**: 113-137.

Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**(8): 533-538.

Perry GH, Marioni JC, Melsted P, Gilad Y. 2010. Genomic-scale capture and sequencing of endogenous DNA from feces. *Mol Ecol* **19**(24): 5332-5344.

Perry GH, Martin RD, Verrelli BC. 2007. Signatures of functional constraint at aye-aye opsin genes: the potential of adaptive color vision in a nocturnal primate. *Mol Biol Evol* **24**(9): 1963-1970.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**(17): 9748-9753.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010a. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**(7289): 768-772.

Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010b. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**(12): e1001236.

Power ML, Myers EW. 2009. Digestion in the common marmoset (Callithrix jacchus), a gummivore-frugivore. *Am J Primatol* **71**(12): 957-963.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8): 904-909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Quemere E, Crouau-Roy B, Rabarivola C, Louis EE, Jr., Chikhi L. 2010. Landscape genetics of an endangered lemur (Propithecus tattersalli) within its entire fragmented range. *Mol Ecol* **19**(8): 1606-1621.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ et al. 2010. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**(11): 909-912.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**(6): 1117-1123.

Somel M, Creely H, Franz H, Mueller U, Lachmann M, Khaitovich P, Paabo S. 2008. Human and chimpanzee gene expression differences replicated in mice fed different diets. *PLoS One* **3**(1): e1504.

Stone AC, Griffiths RC, Zegura SL, Hammer MF. 2002. High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc Natl Acad Sci U S A* **99**(1): 43-48.

Swapna N, Radhakrishna S, Gupta AK, Kumar A. 2010. Exudativory in the Bengal slow loris (Nycticebus bengalensis) in Trishna Wildlife Sanctuary, Tripura, northeast India. *Am J Primatol* **72**(2): 113-121.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**(8): 1596-1599.

Tan CL, Drake JH. 2001. Evidence of tree gouging and exudate eating in pygmy slow lorises (Nycticebus pygmaeus). *Folia Primatol (Basel)* **72**(1): 37-39.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.

Vinyard CJ, Wall CE, Williams SH, Hylander WL. 2003. Comparative functional analysis of skull morphology of tree-gouging primates. *Am J Phys Anthropol* **120**(2): 153-170.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* **102**(51): 18508-18513.

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF. 2008. A novel DNA sequence database for analyzing human demographic history. *Genome Res* **18**(8): 1354-1361.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8): 1586-1591.

Yu N, Jensen-Seaman MI, Chemnick L, Kidd JR, Deinard AS, Ryder O, Kidd KK, Li WH. 2003. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**(4): 1511-1518.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**(5): 821-829.

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M et al. 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res* **17**(6): 839-851.