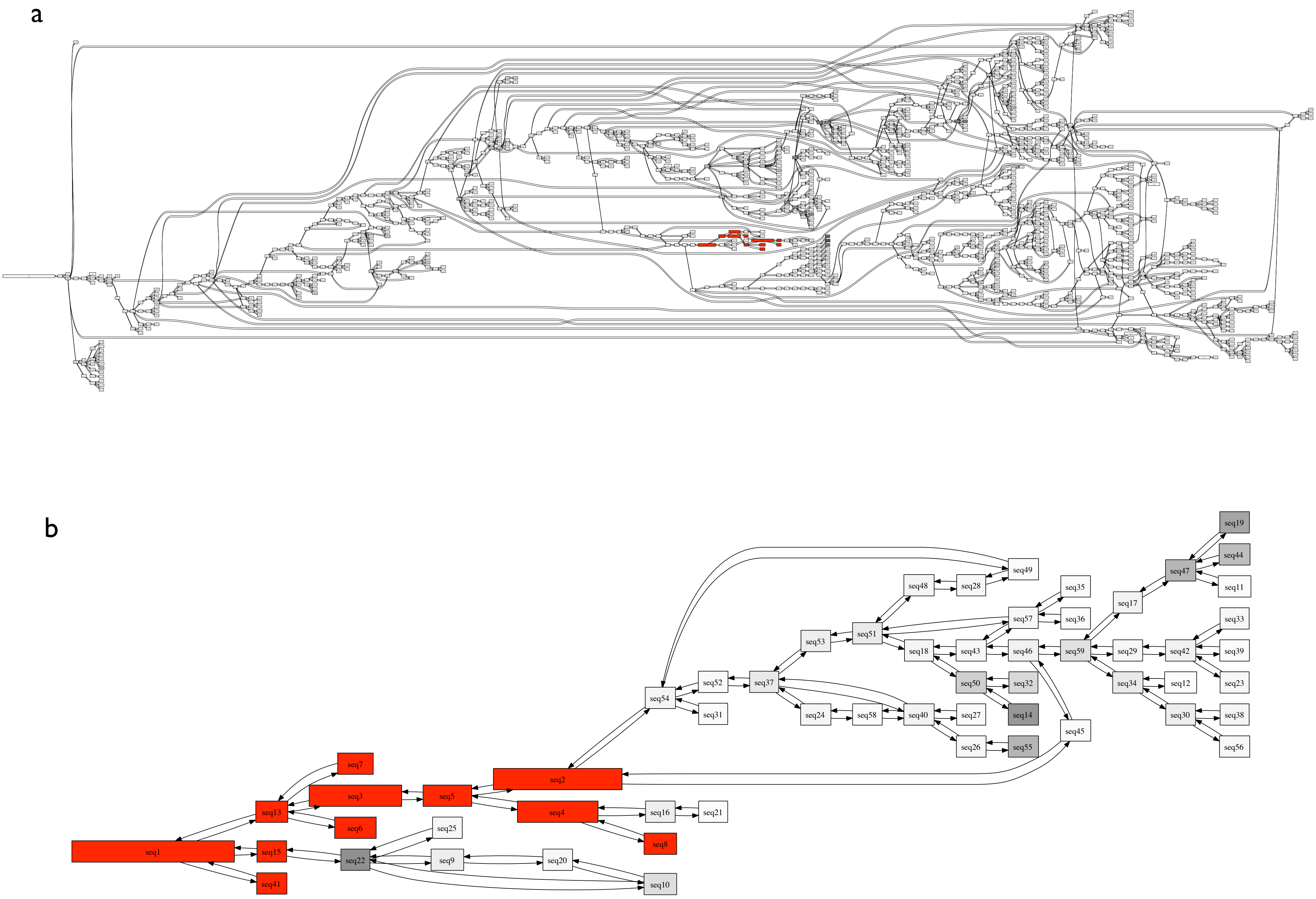
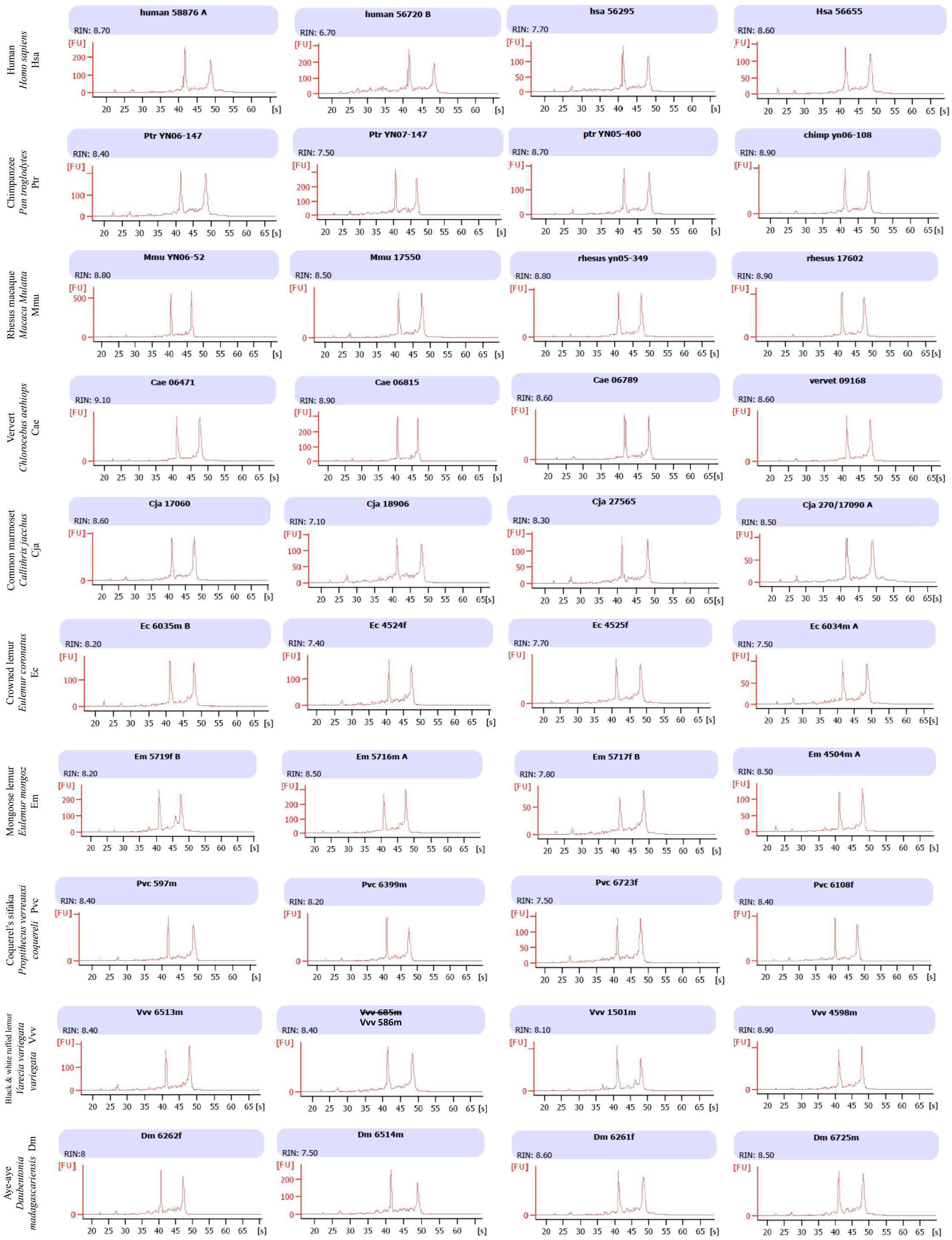
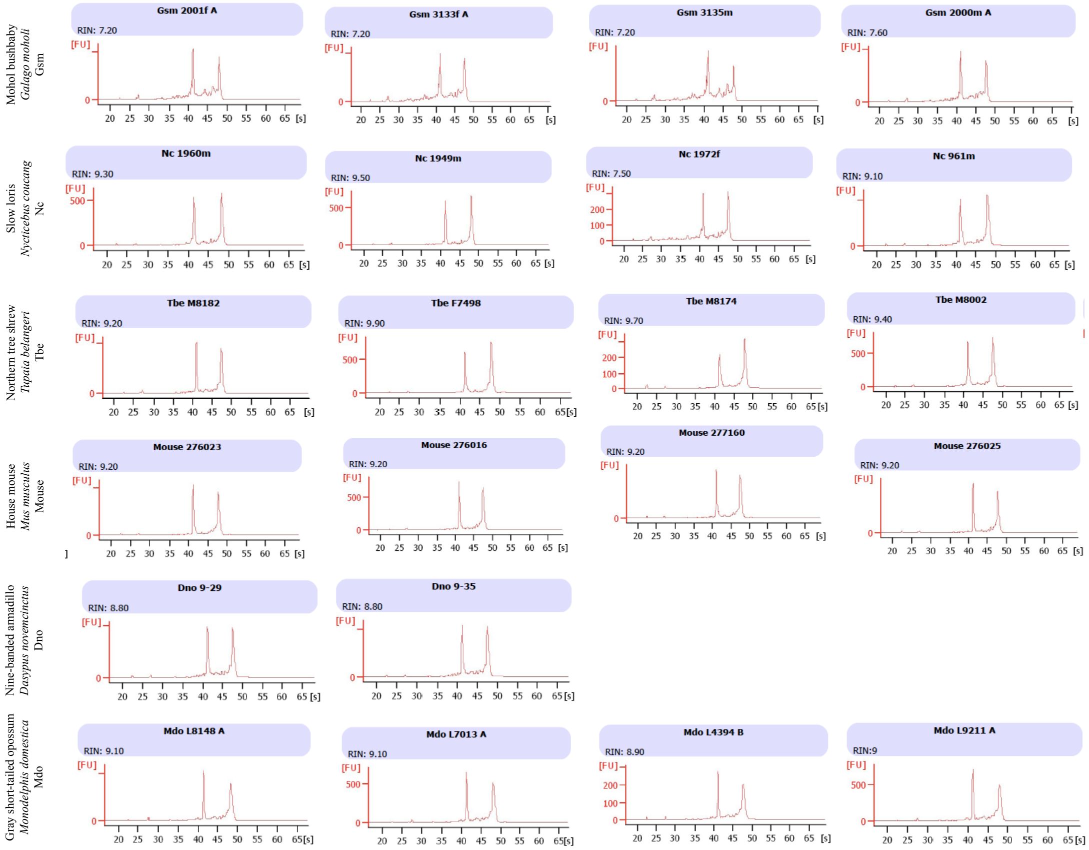


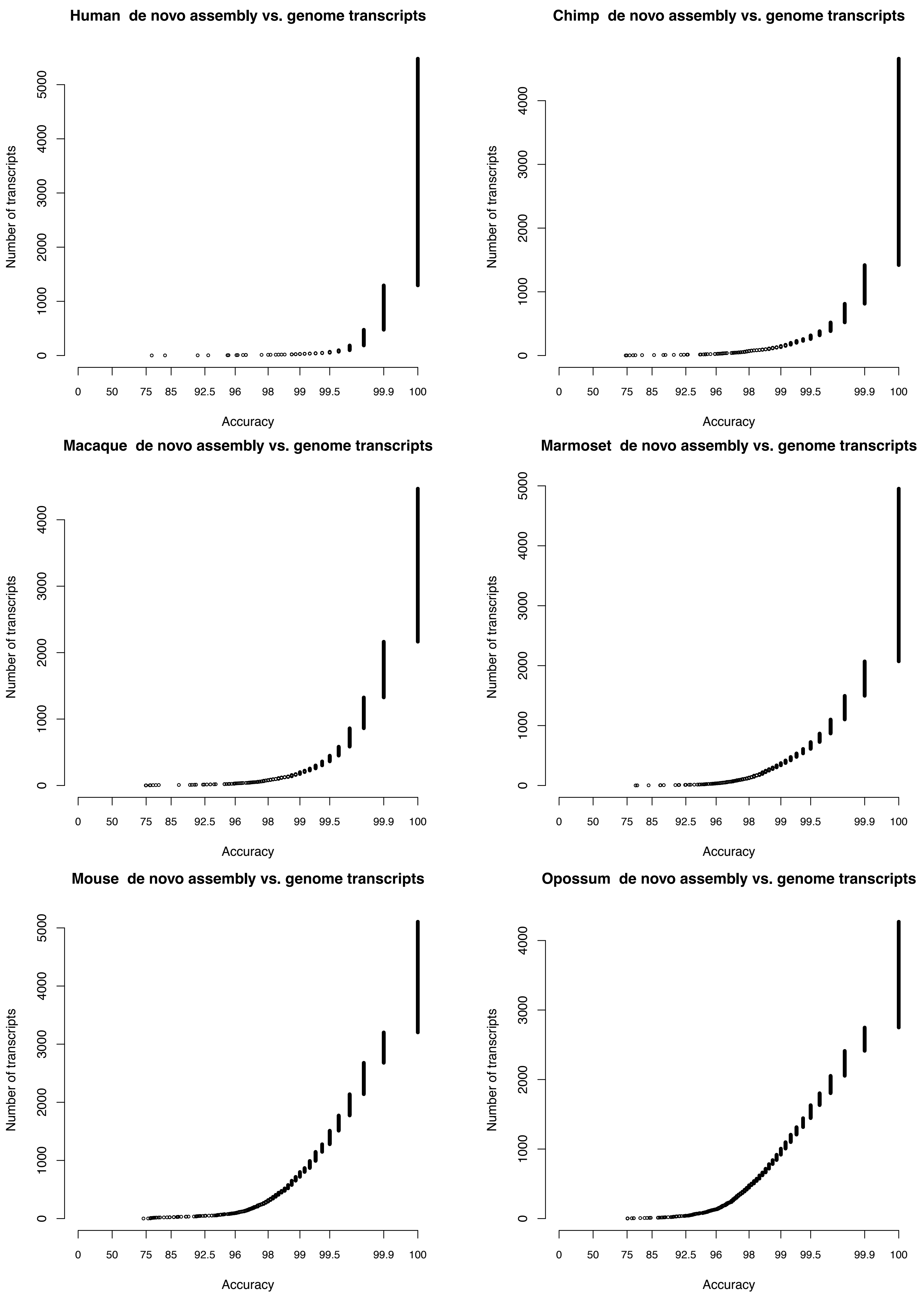
Supplemental Figure 1. De novo transcript assembly.
(**a**) The local de Bruijn graph for the SNF8 gene in rhesus macaque without filtering. This shows the inherent complexities of using the de Bruinn graph exclusively for transcriptome assembly. The de Bruijn graph built from k-mers aligned to the SNF8 reference sequence. Contigs marked red contain the aligned k-mers. Without any filtering on coverage levels the contigs for the gene SNF are glued to k-mers from other genes through a repetitive element. (**b**) Zoomed in on the contigs with aligned k-mers, shown in red. The red contig labeled seq2 is connected to two contigs (seq54 and seq45) that represent intronic sequence in rhesus macaque. These contigs are in turn connected to repetitive elements present in other regions. The two linking contigs have low coverage and thus are not considered in our assembly method, simplifying the problem of assembling the gene by isolating the true contigs of SNF8 from other genes.

Supplemental Figure 2. RNA sample quality.



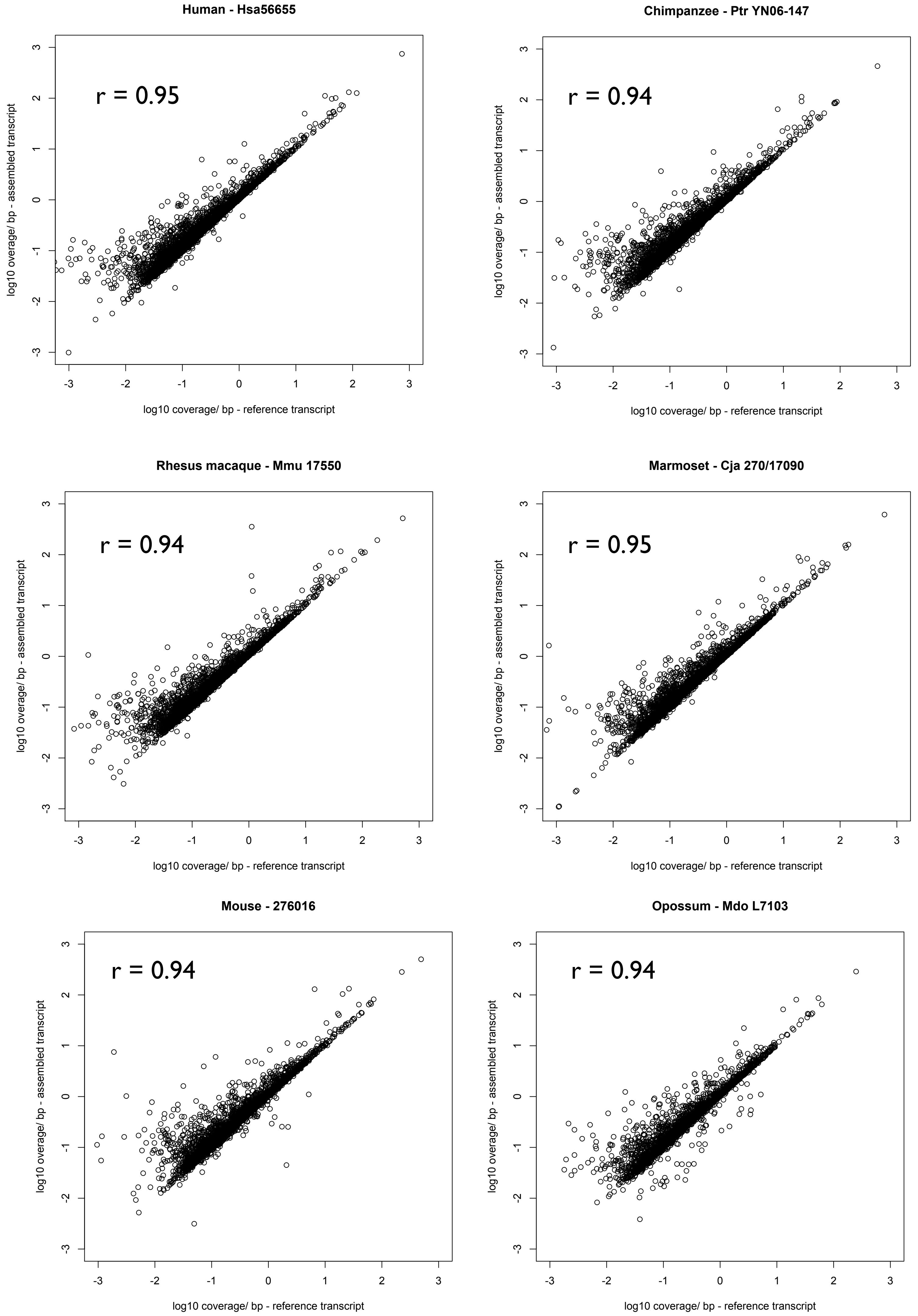


Supplemental Figure 2. RNA sample quality.
RNA Integrity Numbers (RIN) for each sample are reported. Bushbaby average RIN (7.3) was lowest for any species in the study, and there were no bushbaby individuals with RIN > 7.6, possibly affecting our ability to assemble genes for this species.



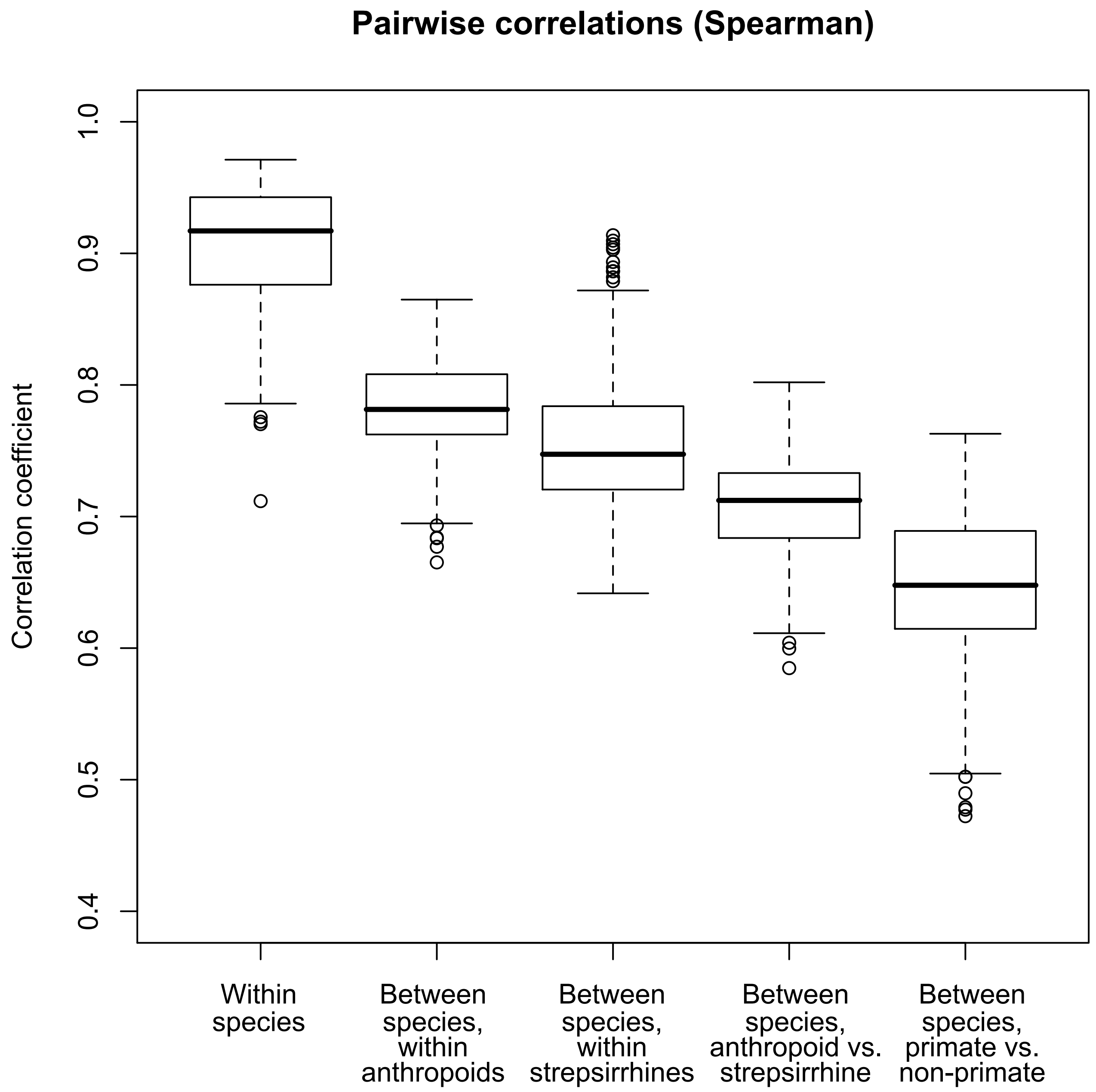
Supplemental Figure 3. Nucleotide sequence comparisons of assembled gene sequence and transcripts from reference genomes.

For each species with an available reference genome sequence, a cumulative distribution function plot of percent nucleotide similarity between assembled and reference transcripts following alignment with FSA.



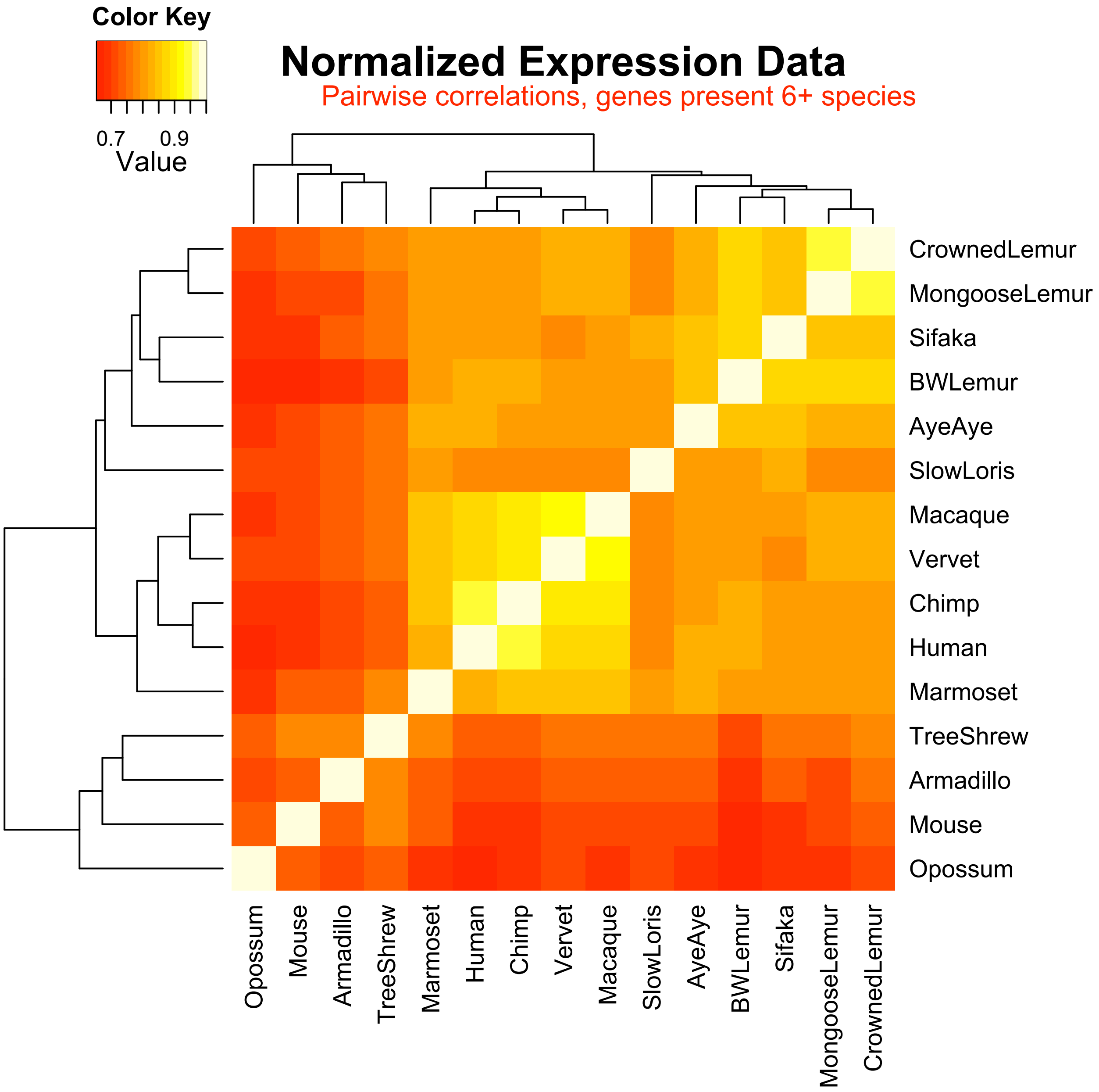
Supplemental Figure 4. Estimated expression level comparisons based on assembled gene sequences and transcripts from reference genomes.

For one representative individual from each species with an available reference genome sequence, a scatter plot of expression level estimates based on read counts.



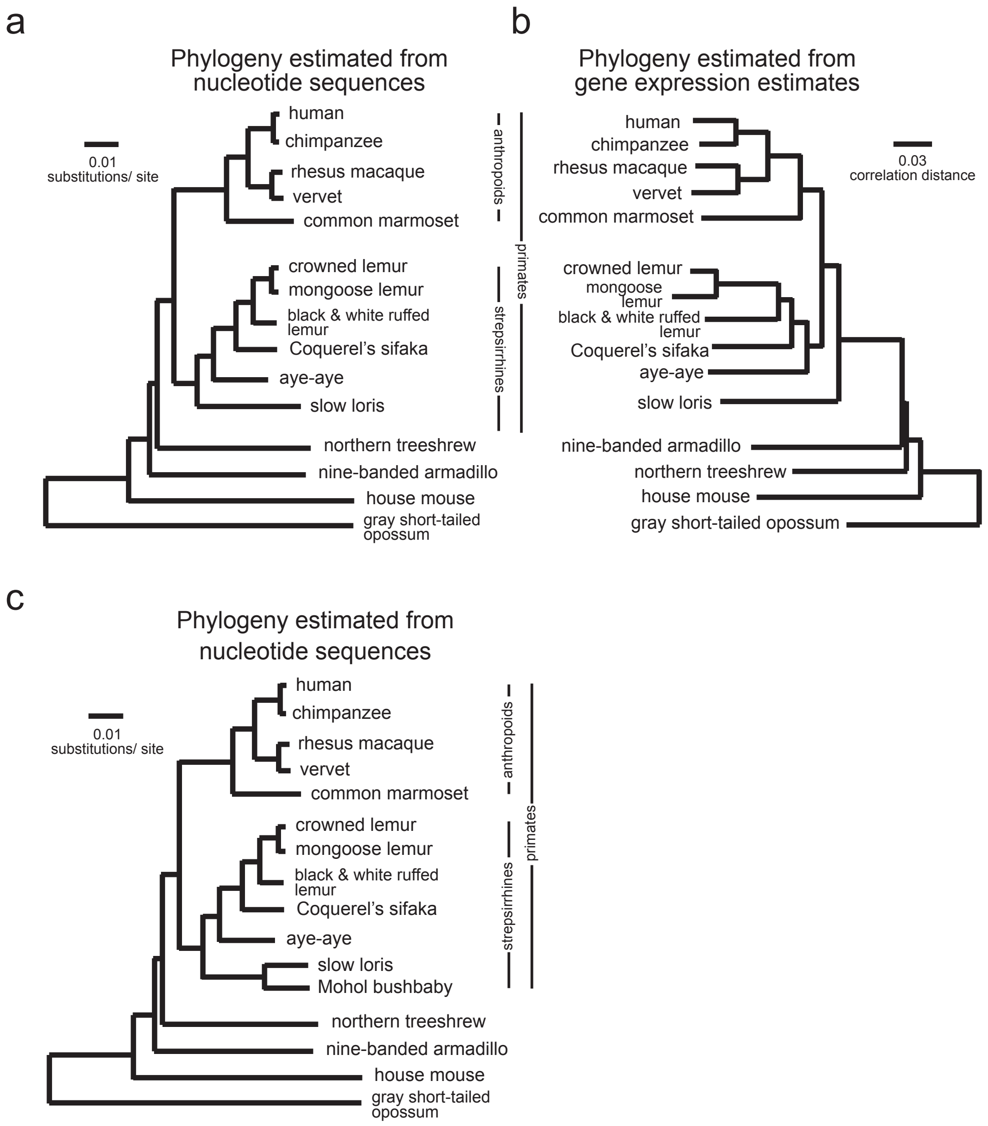
Supplemental Figure 5. Within- and between- species correlations of normalized expression level estimates.

Boxplots of within- and between-species pairwise correlations, from the database of genes with normalized expression estimates for at least six species (6,494 genes).

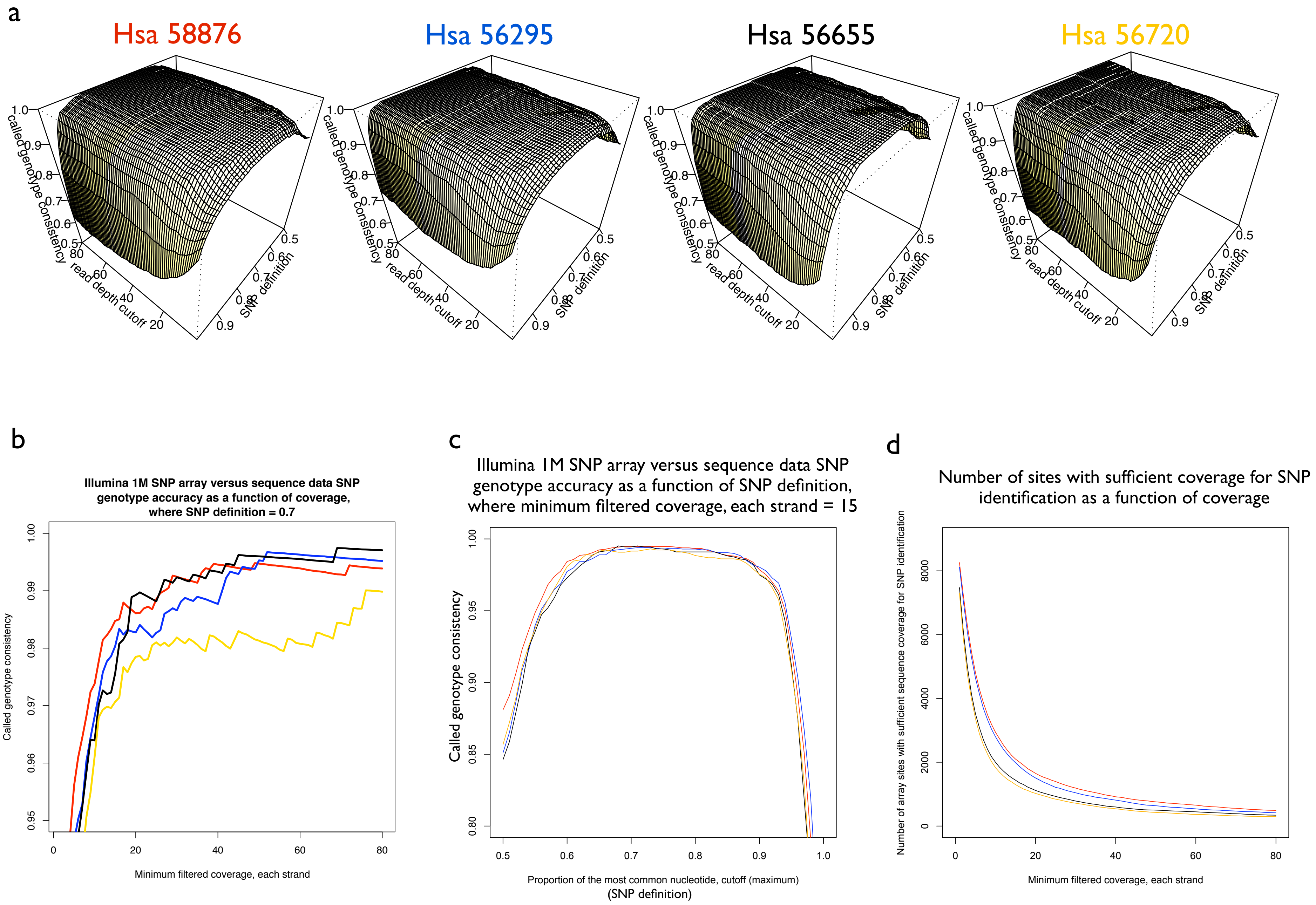


Supplemental Figure 6: Pairwise spearman correlation matrix of species mean expression estimates.

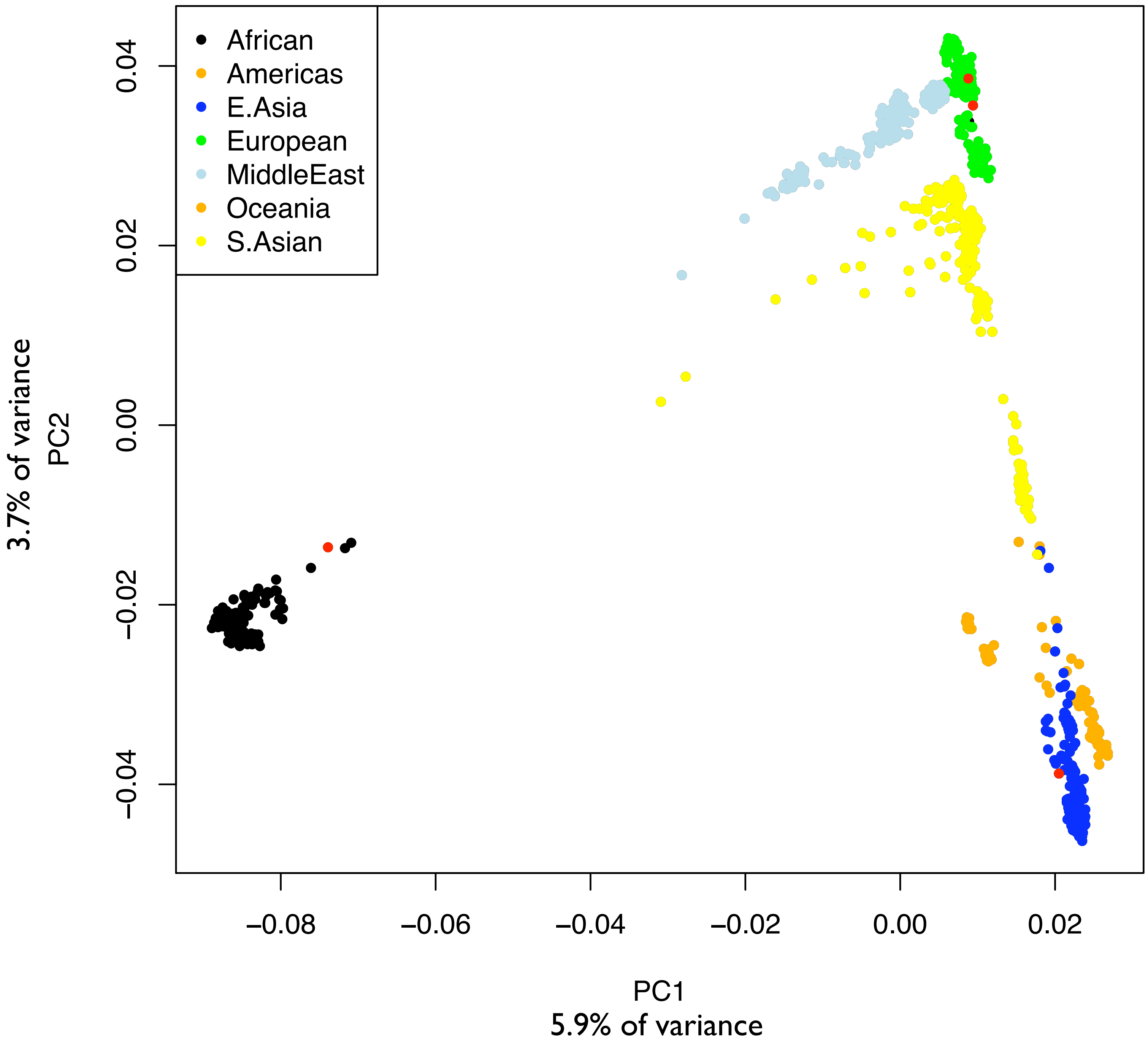
For genes with expression estimates available from a minimum of six species (6,494 genes).



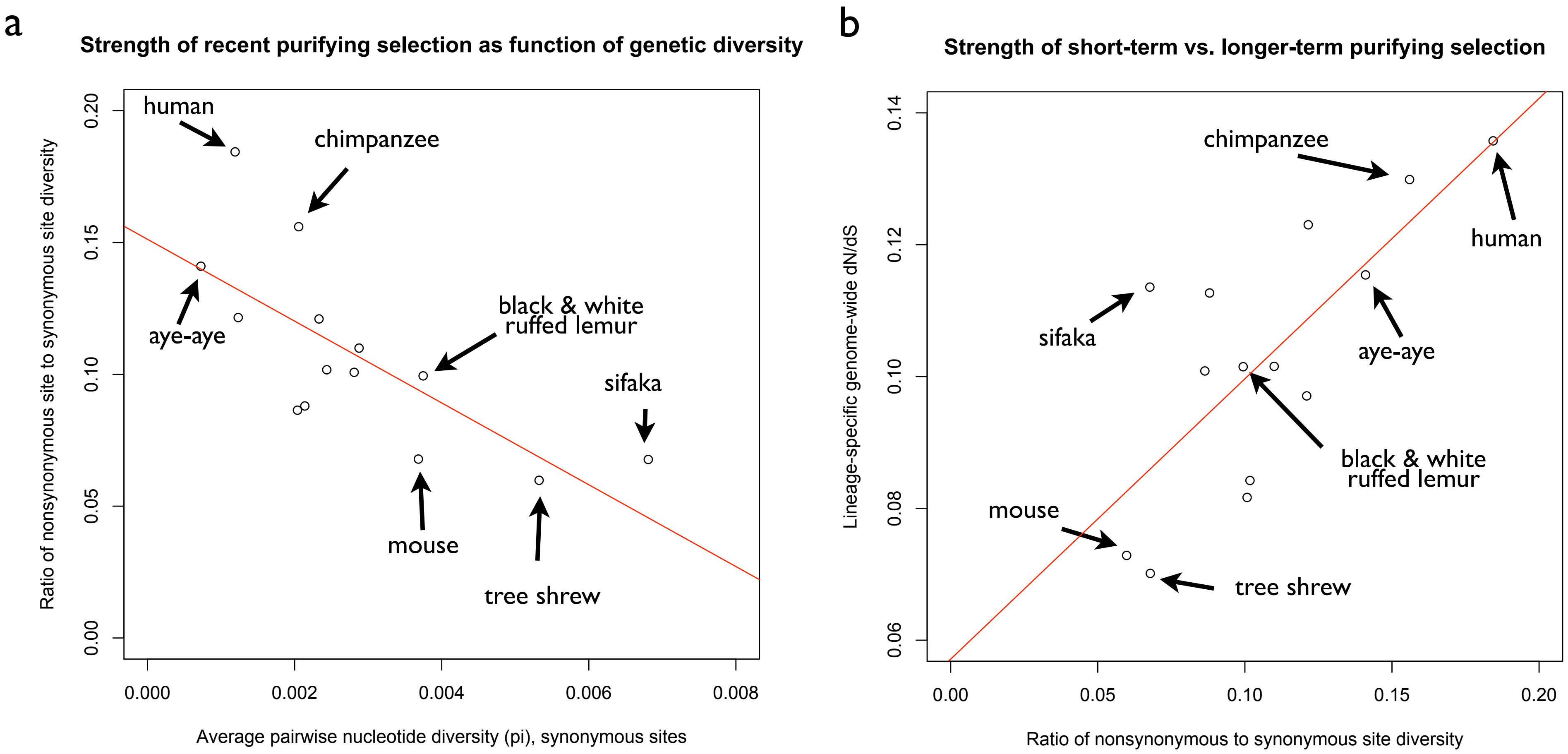
Supplemental Figure 7. Phylogeny reconstruction from nucleotide sequence and gene expression data.
(a) Neighbor-joining tree estimated from nucleotide sequence data. Nucleotide sequence distance matrix was computed from concatenated multi-species alignments of coding sequences of 515 genes that were assembled for all 16 species. (b) Neighbor-joining tree estimated from gene expression data. Gene expression pairwise correlation distance matrix was computed for species mean expression estimates using all genes assembled in at least six species (6,494 genes). (c) Neighbor-joining tree estimated from nucleotide sequence data, identical to (a) but also including bushbaby. As expected, the known primate phylogeny was recapitulated perfectly from the nucleotide sequence data, with the only discrepancy among non-primate mammals being the juxtaposition of the mouse and armadillo branches, likely explained by long branch attraction that is a common issue in phylogenetic analyses that include rodents. Variation in the expression data also follows a phylogenetic pattern, but with slow loris erroneously placed outside all other primates, and the misplacement of armadillo.



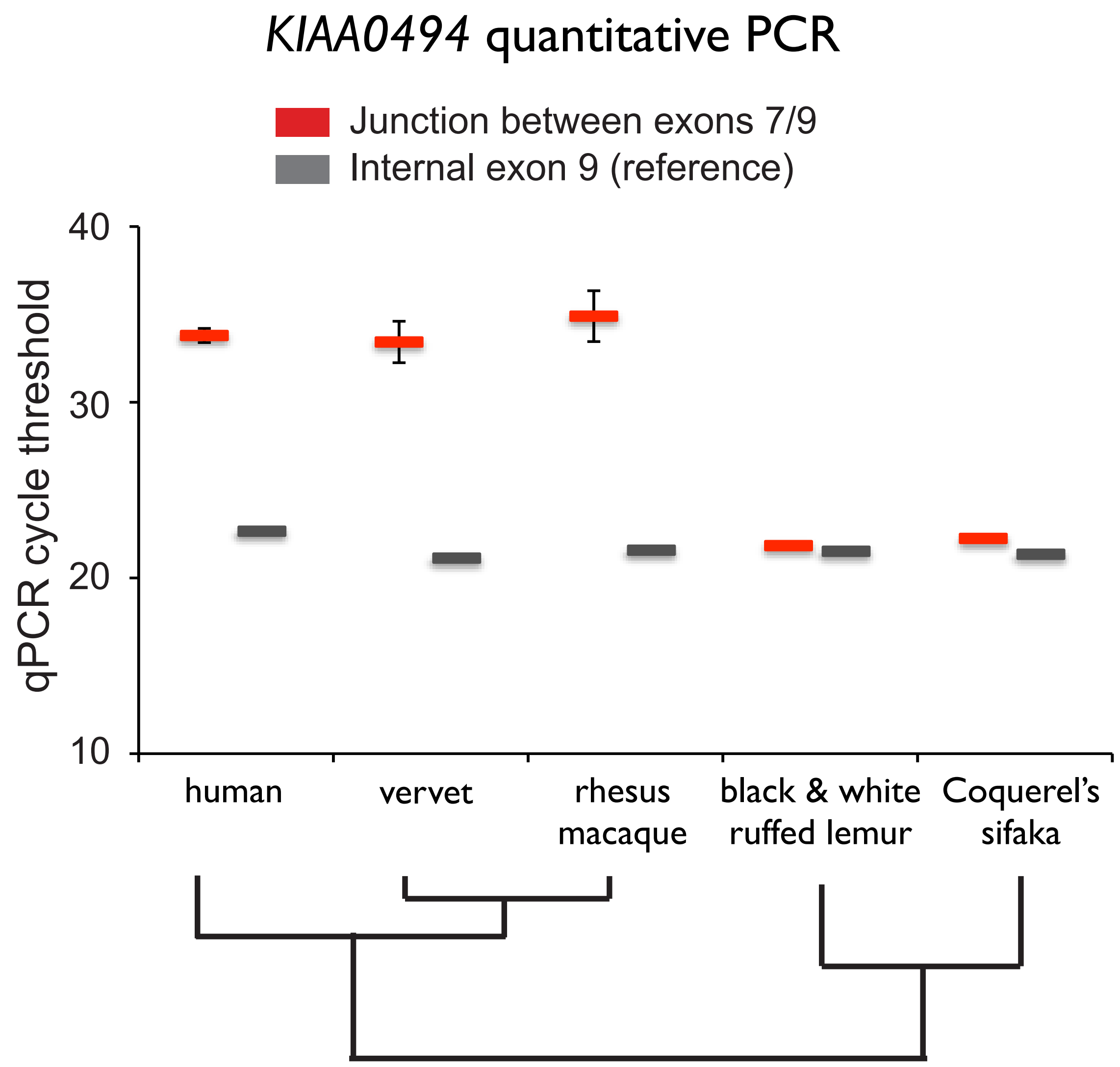
Supplemental Figure 8. Selection of SNP identification criteria.
From each of the 4 human liver samples in the study, we extracted DNA and performed genome-wide genotyping of 1 million SNPs using Illumina HumanOmni-1 Quad Bead Chips. The genotypes in liver-expressed coding regions were used to select appropriate SNP identification criteria. **(a)** 3-dimensional plots for each sample of SNP genotype accuracy as a function of the per-strand read depth (filtered sequence coverage) and the most common nucleotide proportion (SNP definition) cutoffs used to determine whether there was sufficient coverage at a site for analysis and for calling a SNP (i.e., if the proportion of the most common nucleotide at a site was less than or equal to the specified cutoff, the site was considered heterozygous), respectively. **(b)** SNP genotype accuracy as a function of coverage, where SNP definition = 0.7 (the SNP definition used in the final analysis). Line colors correspond to sample names in (a). **(c)** SNP genotype accuracy as a function of SNP definition, where minimum coverage = 15 filtered reads per strand (the coverage level used in the final analysis). **(d)** Number of analyzable sites from the Illumina array genotype as a function of coverage.



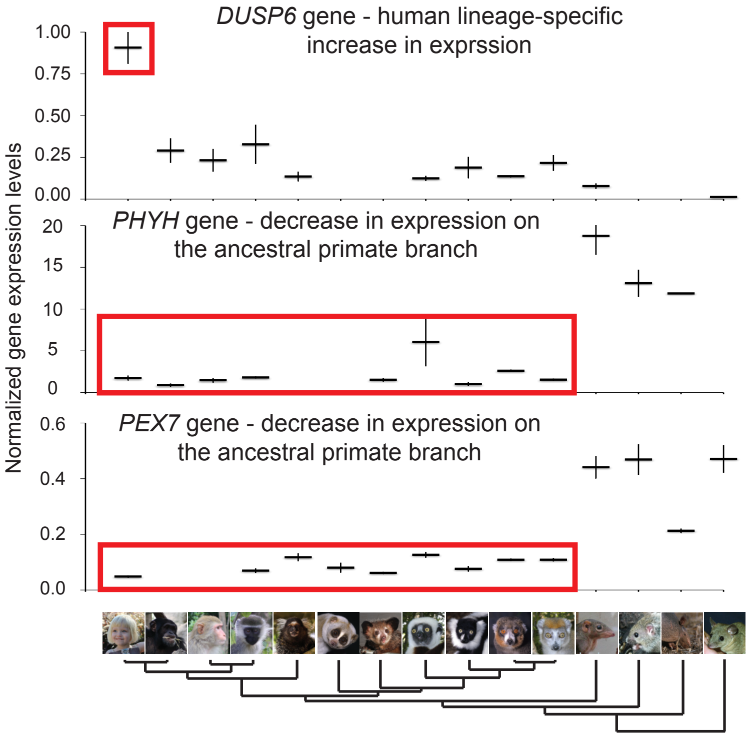
Supplemental Figure 9: Principal components analysis of genome-wide SNP genotype data for the four human individuals in this study, with those from the Human Genome Diversity Panel.
The first two principal components from genotype data for the four human individuals in our study (red circles) based on data collected with the 1M SNP genotype Human-Omni1 quad bead chips (Illumina), using SNPs for which genotypes were also available from a sample of worldwide individuals in the Human Genome Diversity panel. The sample clustering results suggest that two of the humans in our study have predominantly European ancestry, while the remaining two individuals have predominantly African and East Asian ancestries, respectively.



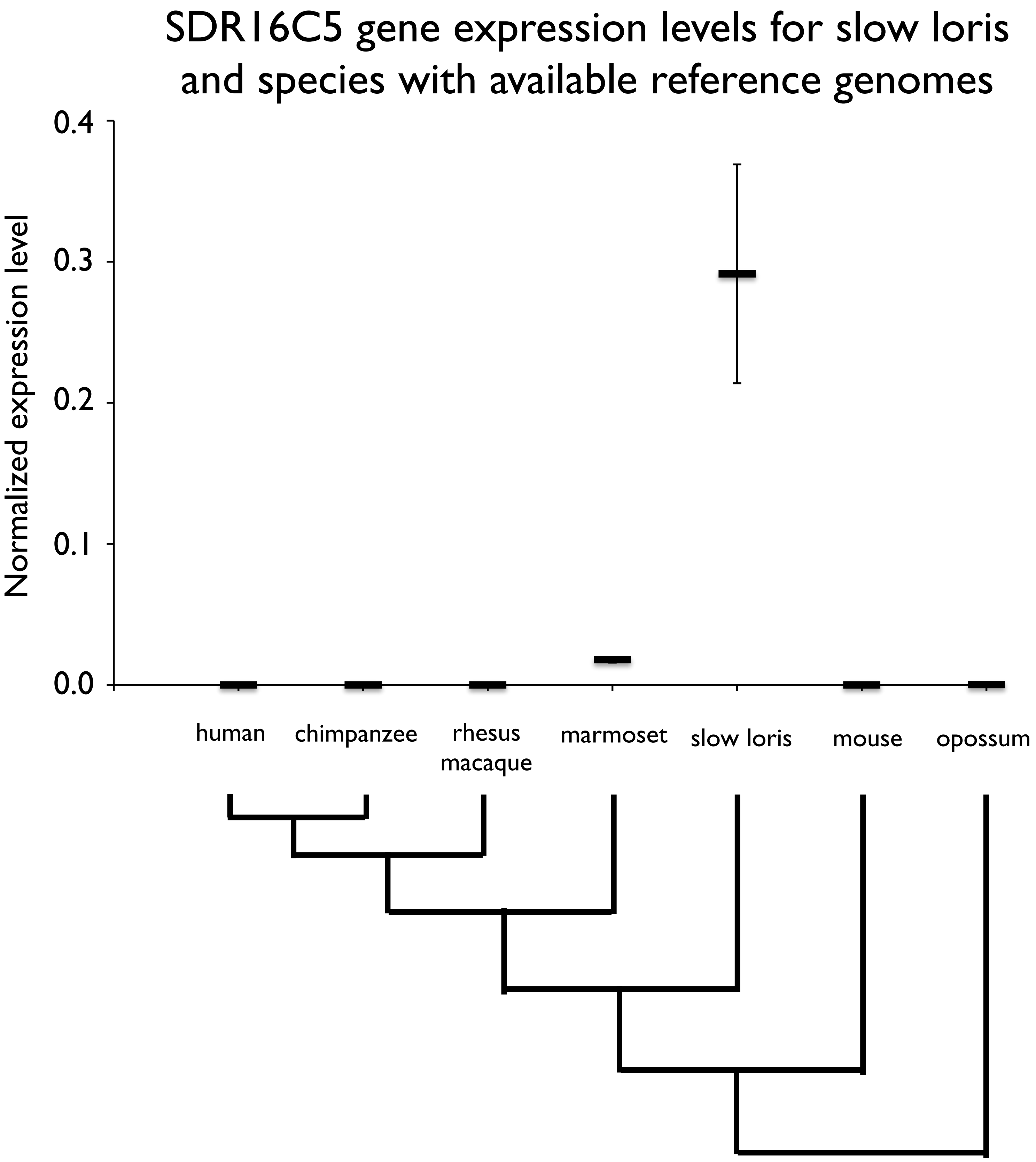
Supplemental Figure 10. Genetic diversity and the strength of selection.
(a) The strong relationship between within-species genetic diversity (pi) at synonymous site and the ratio of nonsynonymous to synonymous site genetic diversity is consistent with the notion that purifying selection is more efficient when effective population sizes are greater. Of the primates in our study, Coquerel's sifaka has the highest level of synonymous site genetic diversity and the smallest ratio of synonymous:synonymous genetic diversity. Human, chimpanzee, and aye-aye have among the lowest synonymous site diversity estimates and the highest ratios of nonsynonymous:synonymous site diversity. (b) There is also a relationship between the ratio of nonsynonymous:synonymous site genetic diversity and lineage-specific dN/dS (genome-wide dN/dS, summed over all genes). The observed variation in this relationship likely reflects, at least in part, changes in effective population sizes over the history of a lineage.



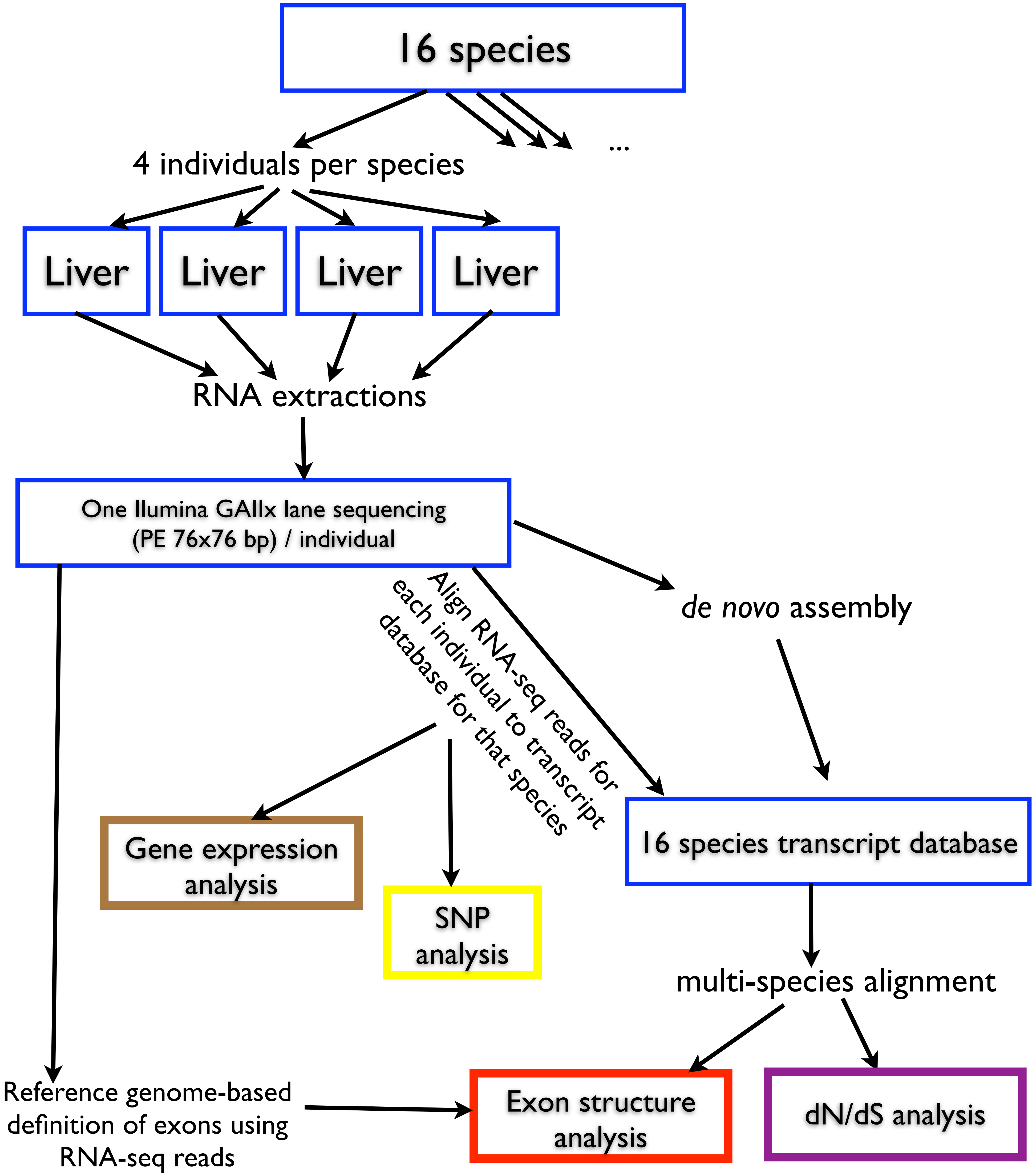
Supplemental Figure 11. *KIAA0494* exon usage divergence.
qPCR cycle threshold values from two pairs of primers specific to the *KIAA0494* gene (a higher cycle threshold reflects fewer starting copies; i.e., more PCR cycles are required to accumulate a certain amount of amplified product). For the first pair (shown in Red), the forward primer spanned the junctions of exons 7 and 9 and the reverse primer was in exon 9. For the second pair (grey), both the forward and reverse primers were in exon 9. There is no evidence of any alternative splicing of exon 9, in any species. Across all species, the cycle thresholds with the internal exon 9 primer pair were similar. In contrast, for amplifications with the exon 7/9 primer pair, cycle thresholds for the two lemur species (black and white ruffed lemur and Coquerel's sifaka) were considerably lower than those for the anthropoid primates. This result is consistent with the exon junction read count data (Figure 3), suggesting that the exon 7/9 junction is nearly always used in lemurs (exon 8 is nearly always skipped) but rarely used in other species. All primers were 100% conserved among the consensus sequences of the studied species.



Supplemental Figure 12. Directional selection on gene expression levels. Examples of genes with expression patterns (mean species estimated expression levels ± s.e.) consistent with the action of directional selection, for species-specific (DUSP6, human lineage) and ancestral (PHYH and PEX7, ancestral primate lineage) branches. PHYH and PEX7 are peroxisomal genes.



Supplemental Figure 13. ***SDR16C5*** **gene expression in slow loris and marmoset.**
SDR16C5 was assembled only for on species, slow loris. This gene is involved in the first, rate-limiting step of retinol metabolism. Since retinol is a derivative of isoprene, the monomer of latex, and slow lorises consume tree exudates including latex as an important component of their diet, we asked whether the reason that SDR16C5 was assembled only for slow loris might be explained by between-species differences in expression, by estimating SDR16C5 expression levels using genome reference sequence-based transcripts, for the six species in our study for which this is possible (human, chimpanzee, rhesus macaque, marmoset, mouse, opossum). Among these species, only marmoset also has appreciable levels of SDR16C5 in the liver. Marmosets have numerous craniofacial adaptations that they use to gouge injuries in trees, from which they also forage extensively on tree exudates, potentially including latex.


16 species
...
4 individuals per species
Liver
Liver
Liver
Liver
RNA extractions
One Ilumina GAIIx lane sequencing (PE 76x76 bp) / individual
de novo assembly
Align RNA-seq reads for each individual to transcript database for that species
16 species transcript database
Gene expression analysis
SNP analysis
multi-species alignment
Reference genome-based definition of exons using RNA-seq reads
Exon structure analysis
dN/dS analysis

Supplemental Figure 14. Flow chart of the methods used for *de novo* assembly and analysis of RNA-seq data.