



Figures and figure supplements

Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence

Rafik Neme and Diethard Tautz

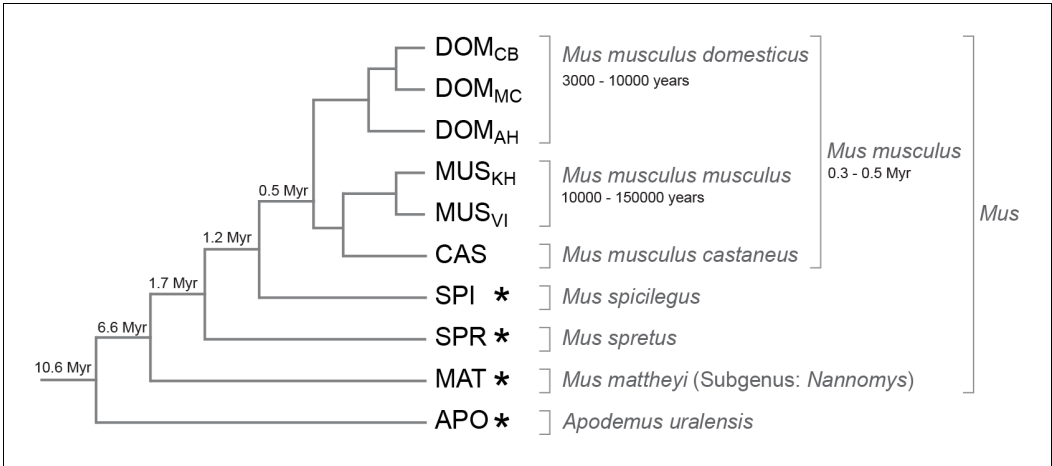


Figure 1. Phylogenetic relationships and time estimates for the taxa used in the study. New genome sequences were generated for taxa with *. A common genome was constructed across all taxa (**Figure 1—figure supplement 1**) based on a mapping algorithm that is not affected by the sequence divergence between the samples (Appendix 1). **Figure 1—figure supplement 2** shows the intersection of genome coverage between the named species.

DOI: [10.7554/eLife.09977.003](https://doi.org/10.7554/eLife.09977.003)

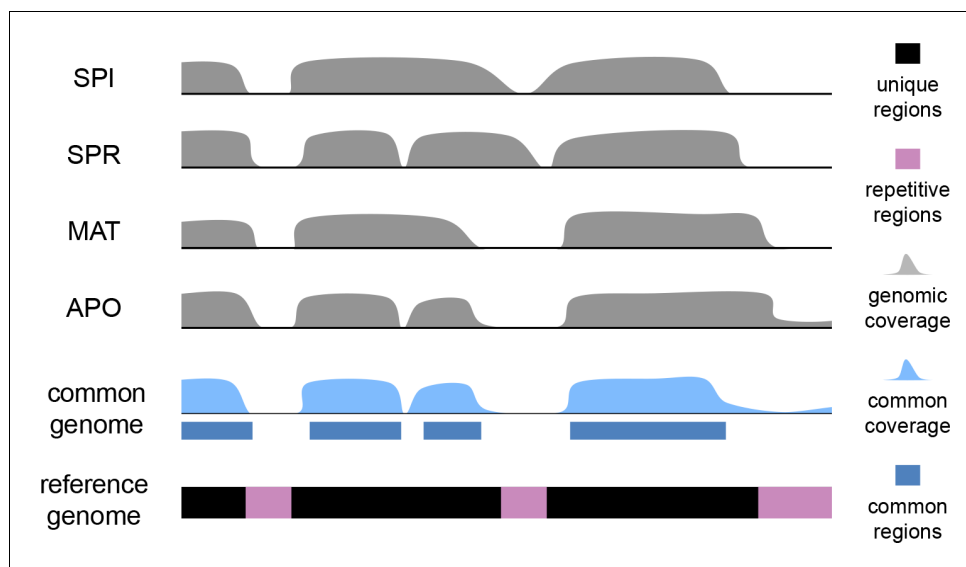


Figure 1—figure supplement 1. Scheme for the establishment of the 'common genome' using genomic reads and the mouse reference genome. The common genome represents the portion of the reference which is present and detectable across all species. The genome sequencing, processing and sequence analysis were done in the same way as for transcriptomes, effectively removing possible biases derived from sequencing and mapping. Note that the assignment of the common genome fraction was done after mapping all genomic and transcriptomic reads to the reference, i.e. the mapping process was not affected by a reduced mapping target.

DOI: [10.7554/eLife.09977.004](https://doi.org/10.7554/eLife.09977.004)

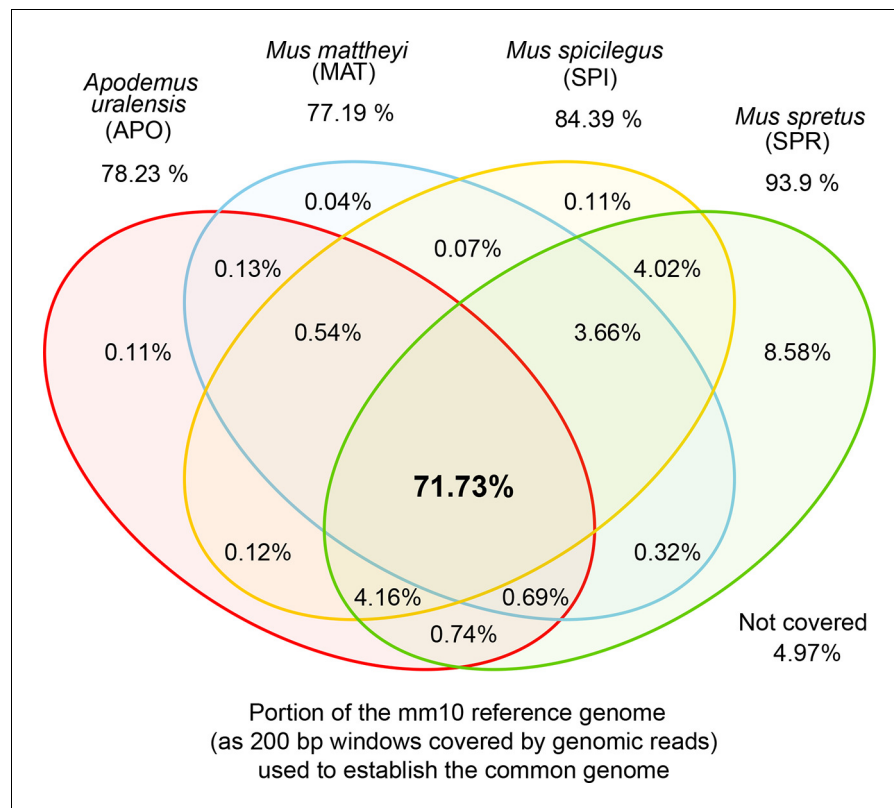


Figure 1—figure supplement 2. Venn diagrams of representation of the common genome, derived from 200bp windows covered in genomic reads in species with more than one million years divergence to the reference. Windows covered by all four species are used as the common genome (shown as the intersection of all species).
DOI: [10.7554/eLife.09977.005](https://doi.org/10.7554/eLife.09977.005)

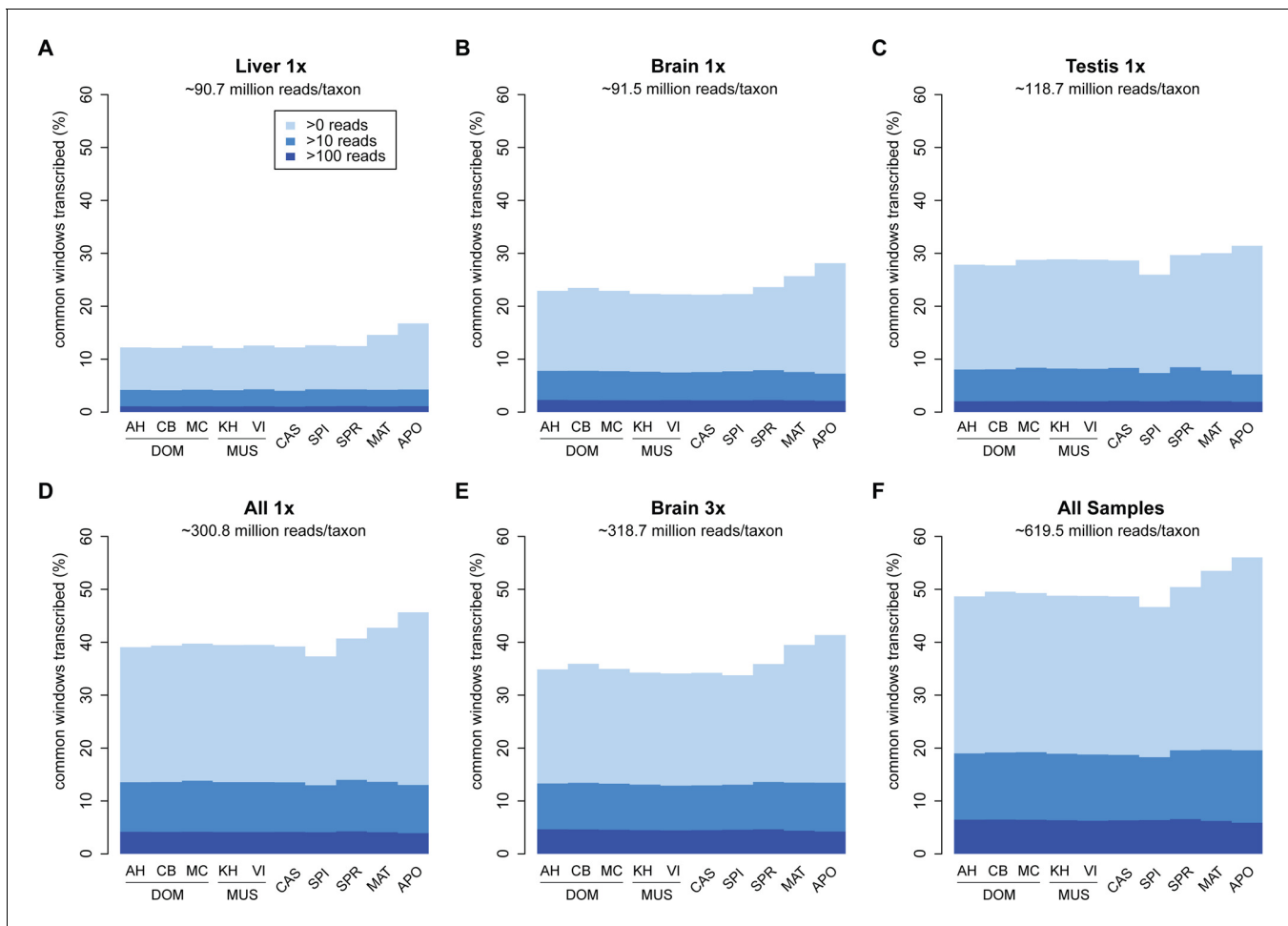


Figure 2. Transcriptome coverage of the common genome per taxon. (A–C) Liver, brain and testis, respectively, sequenced at approximately the same depth. (D) Combination of samples from A–D. (E) Additional sequencing of brain samples at 3x depth, compared to B. (F) Combination of all samples, including additional brain sequencing. Three coverage levels are represented by colors from light blue to dark blue: window coverage with at least 1, 10 and 100 reads. Taxon abbreviations as summarized in **Figure 1**, with closest to the reference genome to the left of each panel and most divergent one to the right. Note that the slight rise in low read coverage for the distant taxa could partially be due to slightly more mismapping of reads at this phylogenetic distance (see Appendix 1 for simulation of mapping efficiency), but is also affected by a larger fraction of singleton reads (compare **Figure 4—figure supplement 1**).

DOI: [10.7554/eLife.09977.006](https://doi.org/10.7554/eLife.09977.006)

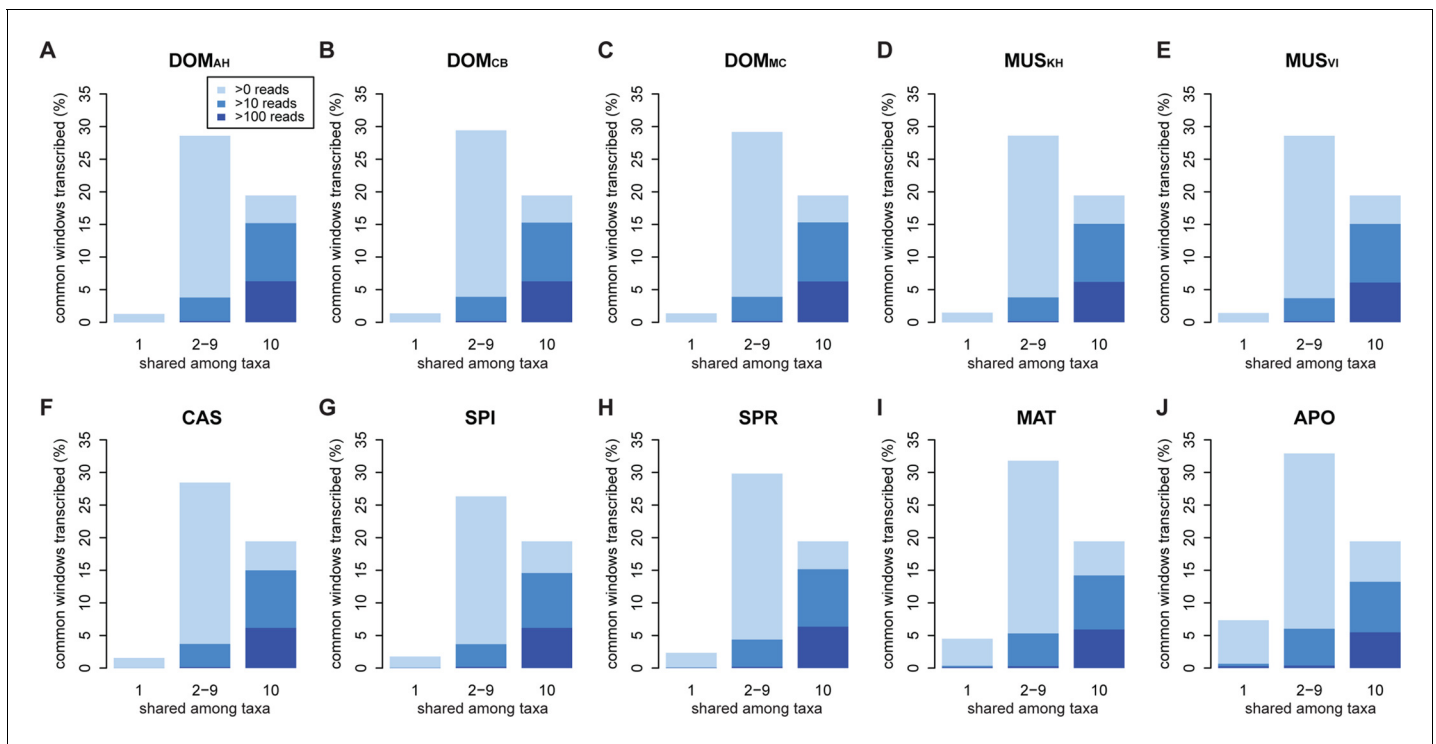


Figure 3. Distribution of shared and non-shared windows with transcripts for each taxon, based on the aggregate dataset across all three tissues. Three classes are represented: i) windows that are found in a single taxon only, ii) windows found in 2–9 taxa and iii) windows shared among all 10 taxa (from left to right in each panel). Windows with transcripts were first classified as belonging to one of the three classes, independent of their coverage, and were then assigned to the coverage classes represented by the blue shading (from light blue to dark blue: window coverage with at least 1, 10 and 100 reads). Taxon names as summarized in **Figure 1**. **Figure 3—figure supplement 1** shows an extended version where class ii) is separated into each individual group. Relative enrichment of annotated genes in the conserved class is shown in **Figure 3—figure supplement 2**.

DOI: [10.7554/eLife.09977.007](https://doi.org/10.7554/eLife.09977.007)

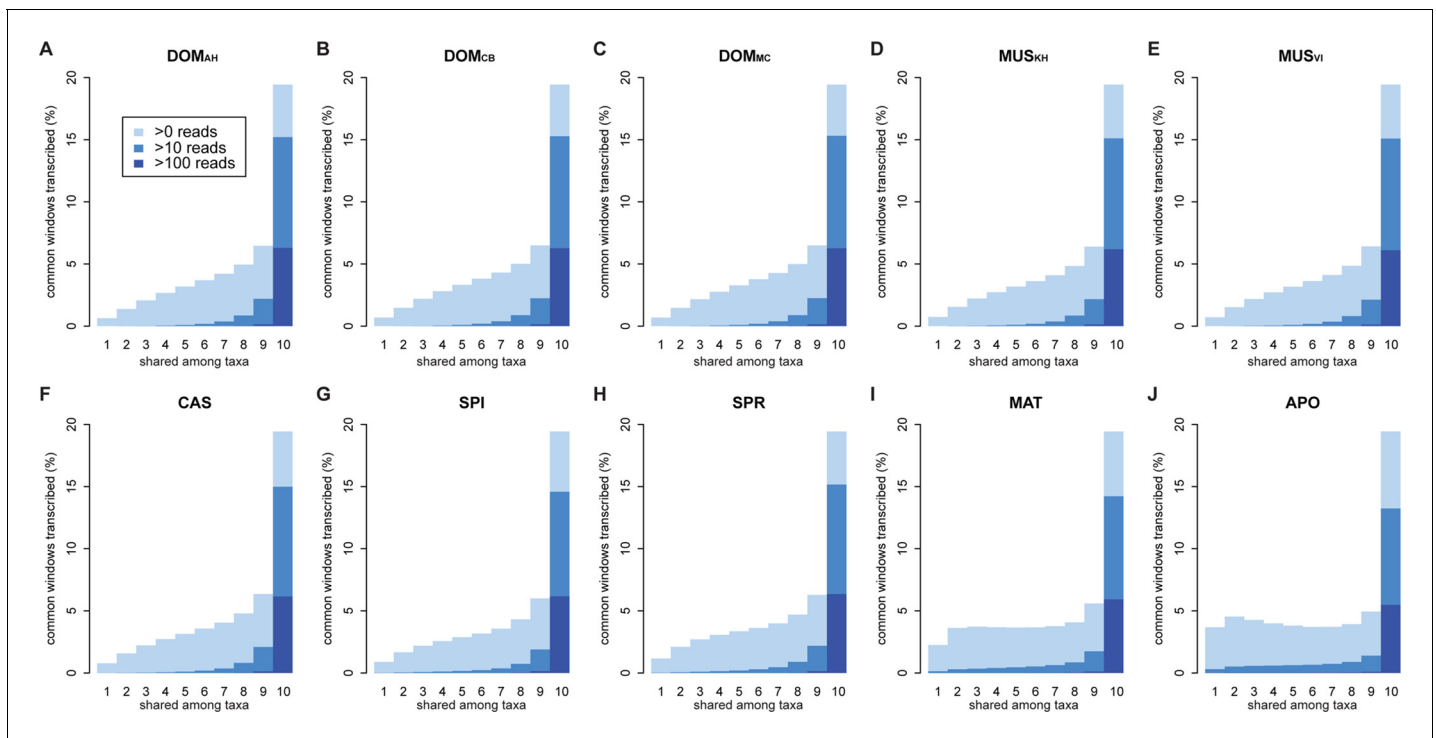


Figure 3—figure supplement 1. Distribution of shared transcripts according to the number of taxa shared, based on the aggregate dataset across all three tissues. Windows with transcripts were first classified as belonging to each of the sharing categories (from 1 to 10), independent of their coverage, and were then assigned to the coverage classes represented by the blue shading (from light blue to dark blue: window coverage with at least 1, 10 and 100 transcripts). Taxon names as summarized in **Figure 1**.

DOI: [10.7554/eLife.09977.008](https://doi.org/10.7554/eLife.09977.008)

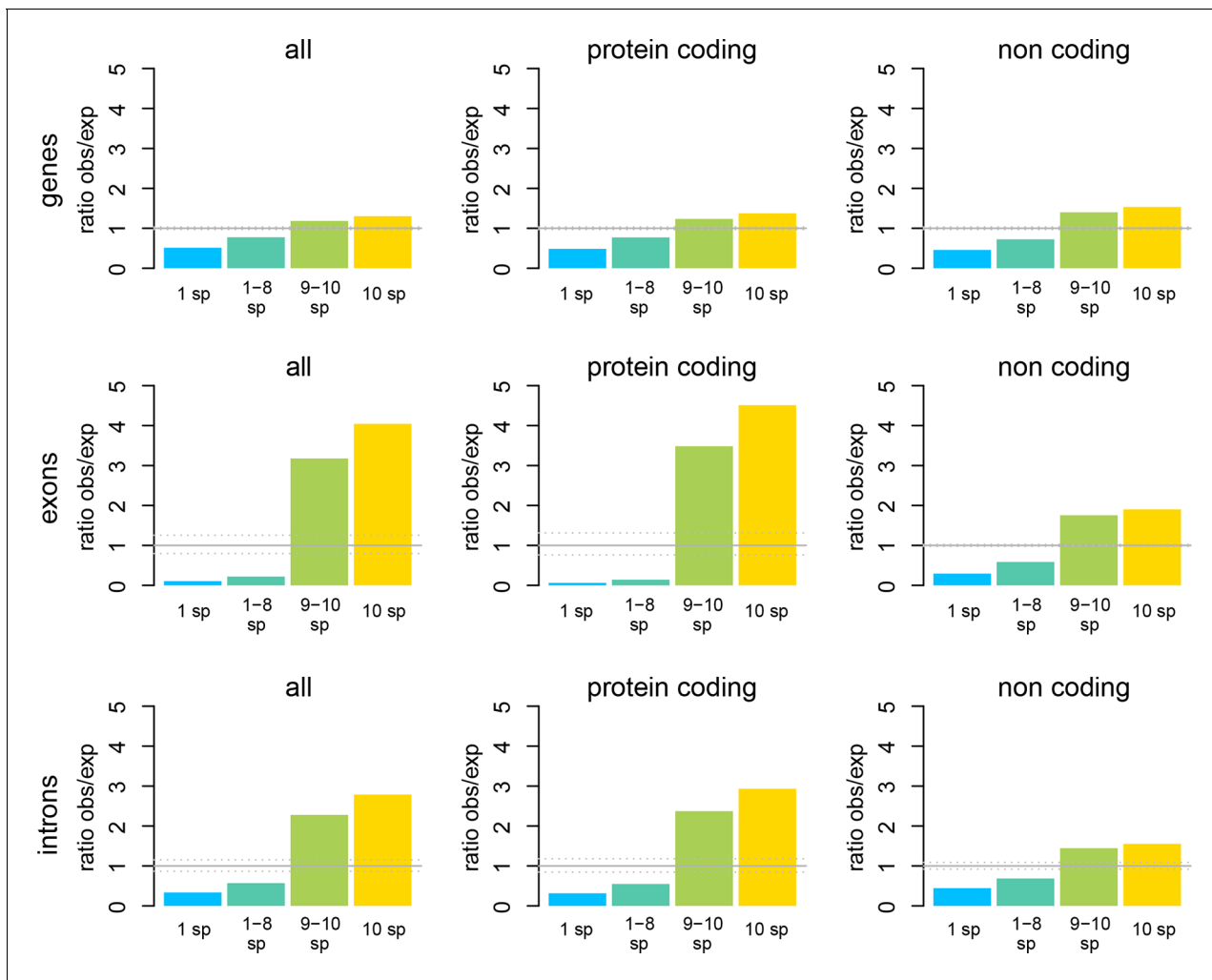


Figure 3—figure supplement 2. Windows transcribed across most species (9 or more) are strongly enriched in genes known from the reference genome, while windows transcribed in some taxa (8 or less) are strongly depleted from known genes. The effect is most evident for protein-coding genes, but still present for non-coding genes.

DOI: [10.7554/eLife.09977.009](https://doi.org/10.7554/eLife.09977.009)

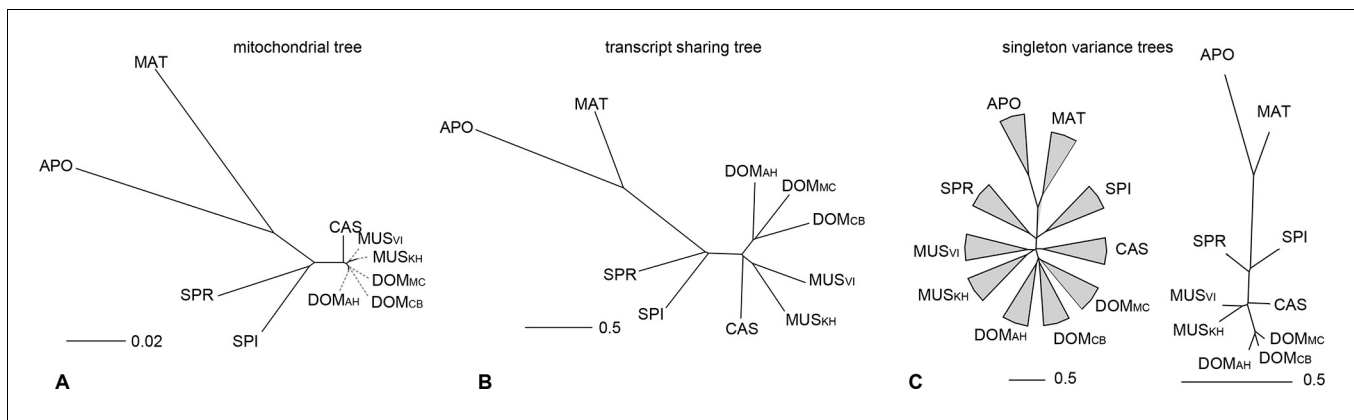


Figure 4. Distance tree comparisons based on molecular and transcriptome sharing data. (A) Molecular phylogeny based on whole mitochondrial genome sequences as a measure of molecular divergence (black lines represent the branch lengths, dashed lines serve to highlight short branches). (B) Tree based on shared transcriptome coverage of the genome, using correlations of presence and absence of transcription of the common genome. All nodes have bootstrap support values of 70% or more ($n = 1000$). (C) Tree based on shared transcriptome coverage of singleton reads only from subsampling of the extended brain transcriptomes. Left is the consensus tree with the variance component between samples depicted as triangles, right is the same tree, but only for the branch fraction that is robust to sampling variance. Taxon names as summarized in **Figure 1**. **Figure 4—figure supplement 1** shows the fraction of singletons in dependence of each sample in each taxon, **Figure 4—figure supplement 2** in dependence of read depth. **Figure 4—figure supplement 3** shows an extended version of the analysis shown in 4C for higher coverage levels.

DOI: [10.7554/eLife.09977.010](https://doi.org/10.7554/eLife.09977.010)

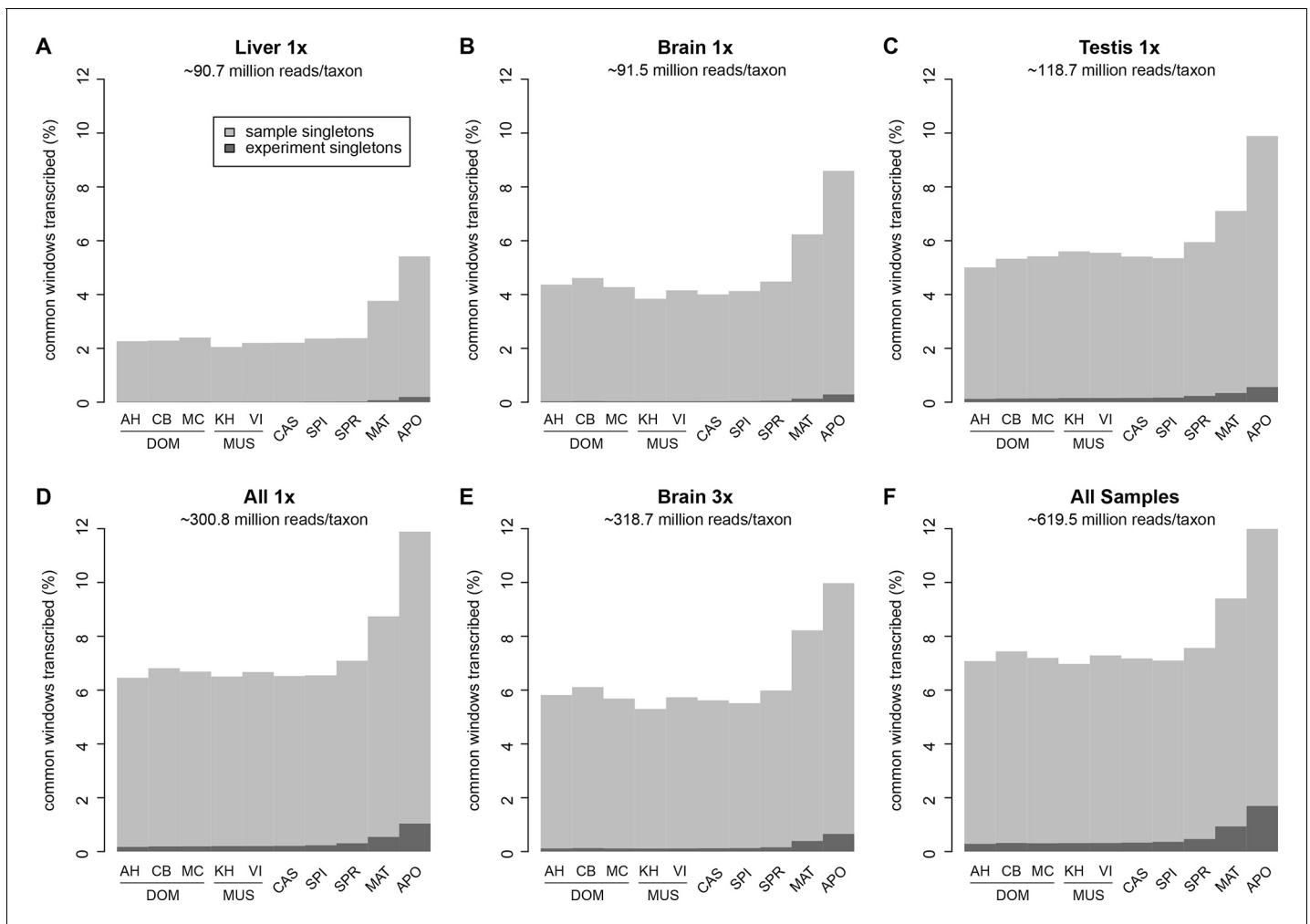


Figure 4—figure supplement 1. Fraction of windows with singletons (one paired read) of the common genome per taxon. (A–C) Liver, brain and testis, respectively, sequenced at approximately the same depth. (D) Combination of samples from A–D. (E) Additional sequencing of brain samples at 3x depth, compared to B. (F) Combination of all samples, including additional brain sequencing. Light gray indicates singletons observed in each individual sample/taxon combination. Dark gray indicates singletons across the whole experiment, i.e. not re-detected in any other tissue or taxon. Taxon abbreviations as summarized in **Figure 1**, with closest to the reference genome to the left of each panel and most divergent one to the right. Note that the rise in singleton number for the distant taxa can be ascribed to the longer branch length, i.e. absence of closely related taxa in which the singleton could have been re-detected.

DOI: [10.7554/eLife.09977.011](https://doi.org/10.7554/eLife.09977.011)

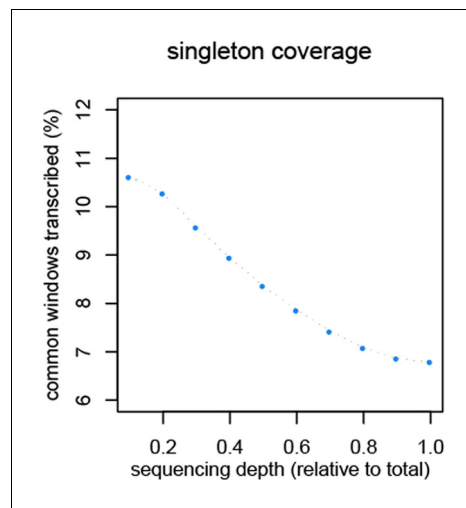


Figure 4—figure supplement 2. Reduction of singletons in dependence of aggregate sequencing depth.
DOI: [10.7554/eLife.09977.012](https://doi.org/10.7554/eLife.09977.012)

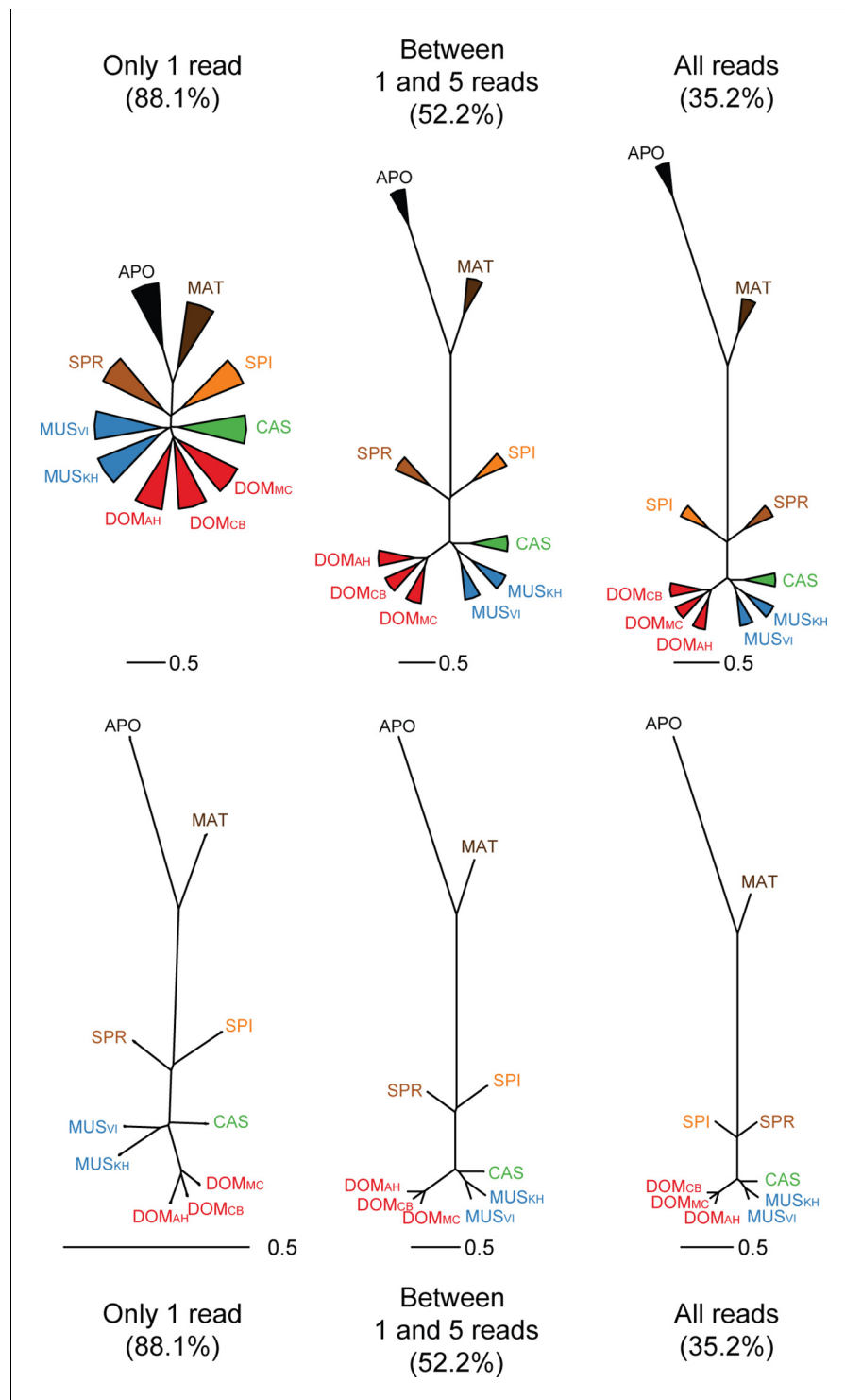


Figure 4—figure supplement 3. Trees based on shared transcriptome coverage of the genome, using binary correlations. We used the deep sequenced brain samples to estimate the proportion of sampling artifacts in terminal branches, and effectively subtracted the proportion of artifacts to obtain reliable phylogenetic signals. Each brain sample was split in three completely independent samples of 100 million reads. Top: Trees constructed using: regions covered only with one read in each taxon, regions covered by 1 and 5 reads (very low expression), regions covered by any reads, regions above 10 reads (mid expression) and regions above 100 reads (high expression). The percentage shown indicates the average level of sampling artifacts for each threshold, derived from the length of the terminal branches not found in all replicates of each taxon, i.e. the uncorrelated portion

Figure 4—figure supplement 3 continued on next page

Figure 4—figure supplement 3 continued

across samples of the same origin. These numbers are highest for the lowly expressed regions, and are lowest for the highly expressed regions, and are more or less constant within comparisons. Once subtracted, the phylogenetic signal remains robust. Taxon names as summarized in **Figure 1**. The figure part with the 1 read fraction corresponds to **Figure 4C**.

DOI: [10.7554/eLife.09977.013](https://doi.org/10.7554/eLife.09977.013)

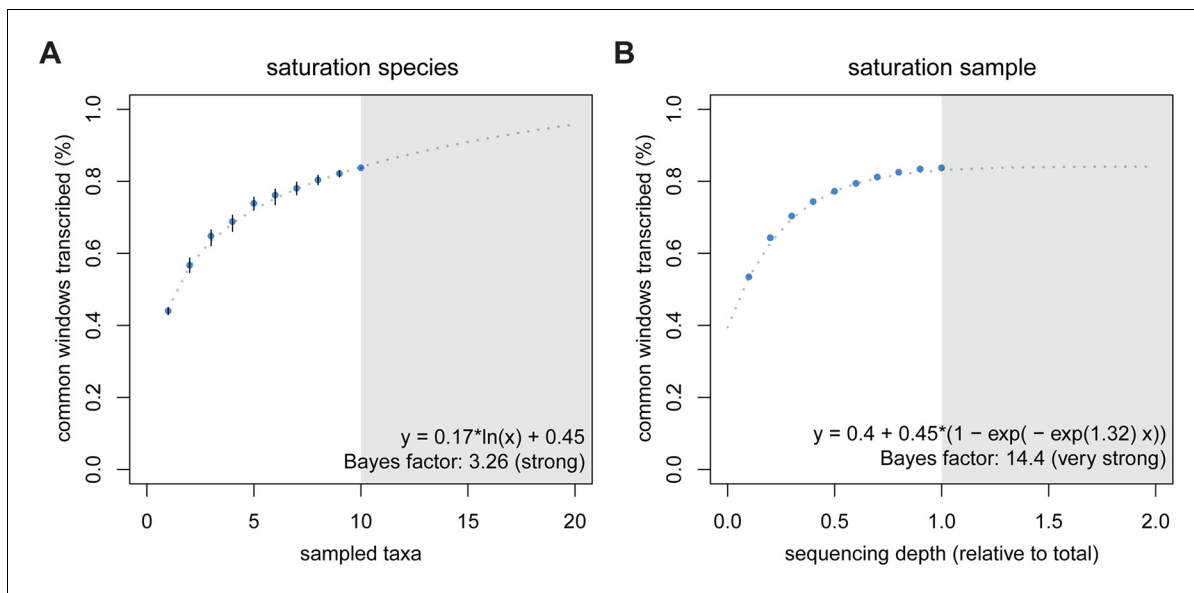


Figure 5. Rarefaction, subsampling and saturation patterns using all available samples and reads. **(A)** Sequencing depth saturation as estimated from an increase in the number of taxa. **(B)** Sequencing depth saturation as estimated from increasing read number. Blue dots indicate increases per sub-sampled sequence fraction or taxon added from our dataset. Gray dotted line indicates the predicted behavior from the indicated regression, and gray area shows the prediction after doubling the current sampling either by additional taxa **(A)** or in sequencing effort **(B)**. Each analysis was tested for logarithmic and asymptotic models. Best fit was selected from ΔBIC , with Bayes factor shown and qualitative degree of support shown. Standard deviations are shown as black lines in **A**, and are too small to display in **B** (note that due to the sampling scheme for this analysis, the values above 50% are not statistically independent and that the 100% value constitutes a single data point without variance measure).

DOI: [10.7554/eLife.09977.014](https://doi.org/10.7554/eLife.09977.014)

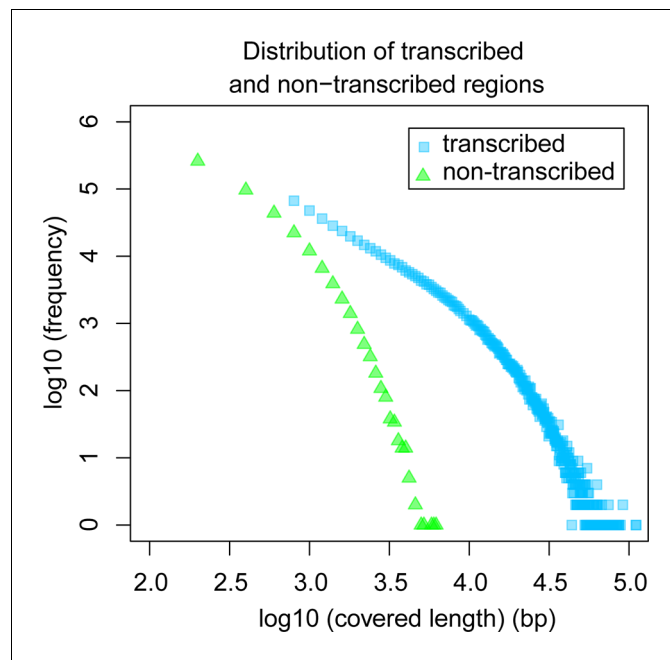
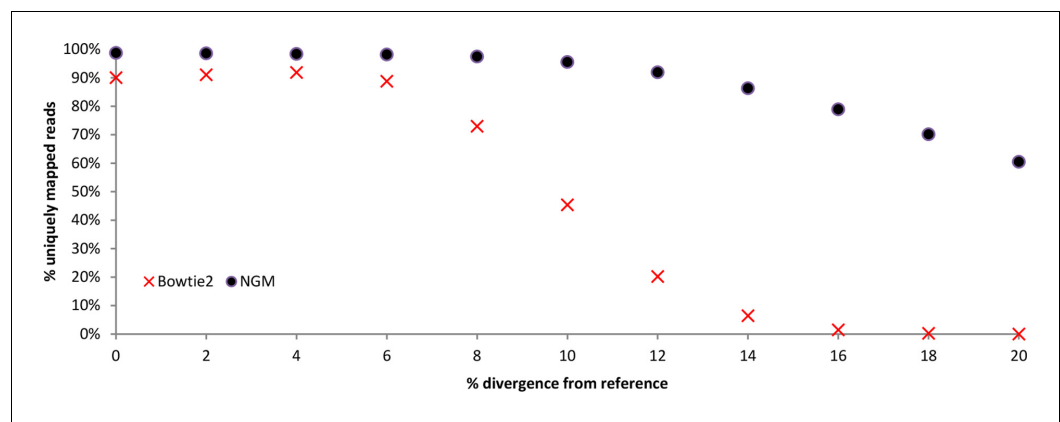


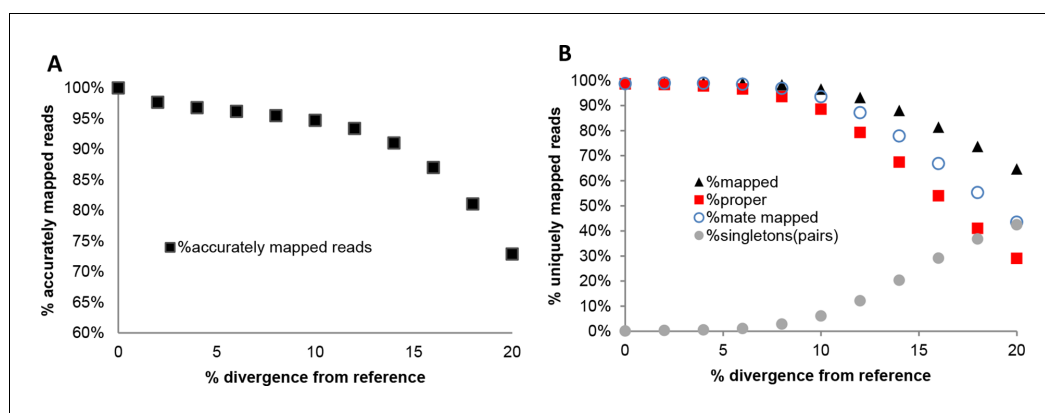
Figure 6. Comparative analysis of lengths of regions transcribed or not transcribed across all data (including deeper brain sequencing) in all samples. Size distribution of regions not covered in any transcript (green) versus size distribution of regions with at least one transcript (blue).

DOI: [10.7554/eLife.09977.015](https://doi.org/10.7554/eLife.09977.015)



Appendix figure 1. Performance of NextGenMap compared to Bowtie2.

DOI: [10.7554/eLife.09977.021](https://doi.org/10.7554/eLife.09977.021)



Appendix figure 2. Performance of NextGenMap in terms accuracy of mapping using the same set of reads and increasingly divergent versions of the reference genome (A), and paired-end mapping statistics (B).

DOI: [10.7554/eLife.09977.022](https://doi.org/10.7554/eLife.09977.022)