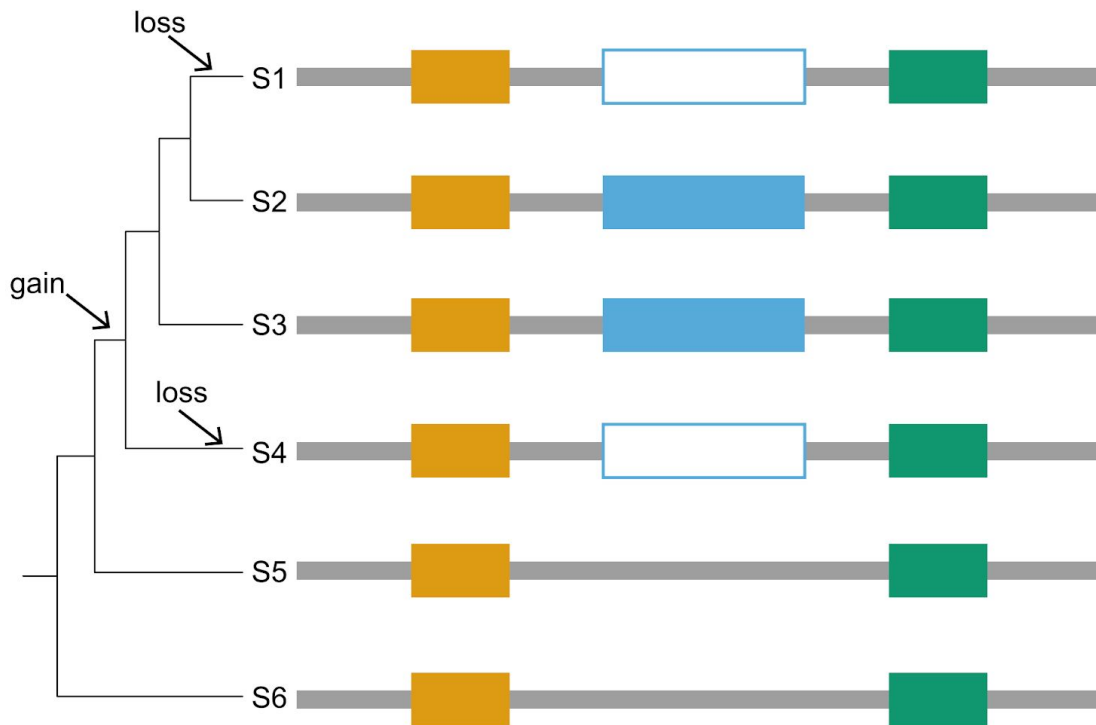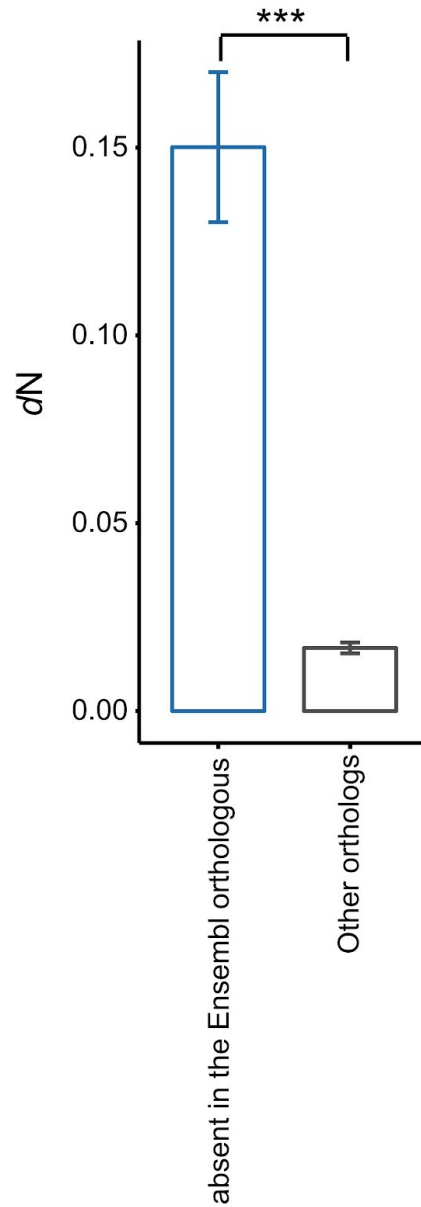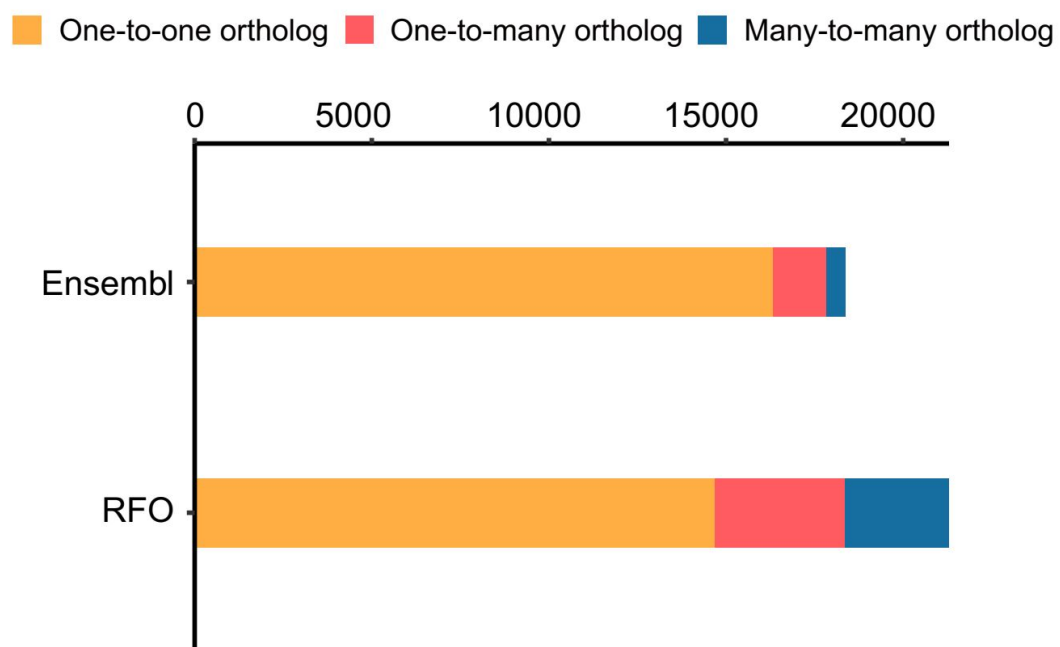# Supplementary Figures

Supplementary Figure S1. Conceptual examples of a gene gain in the ancestor of S1, S2, S3, and S4 but lost in S1 and S4.

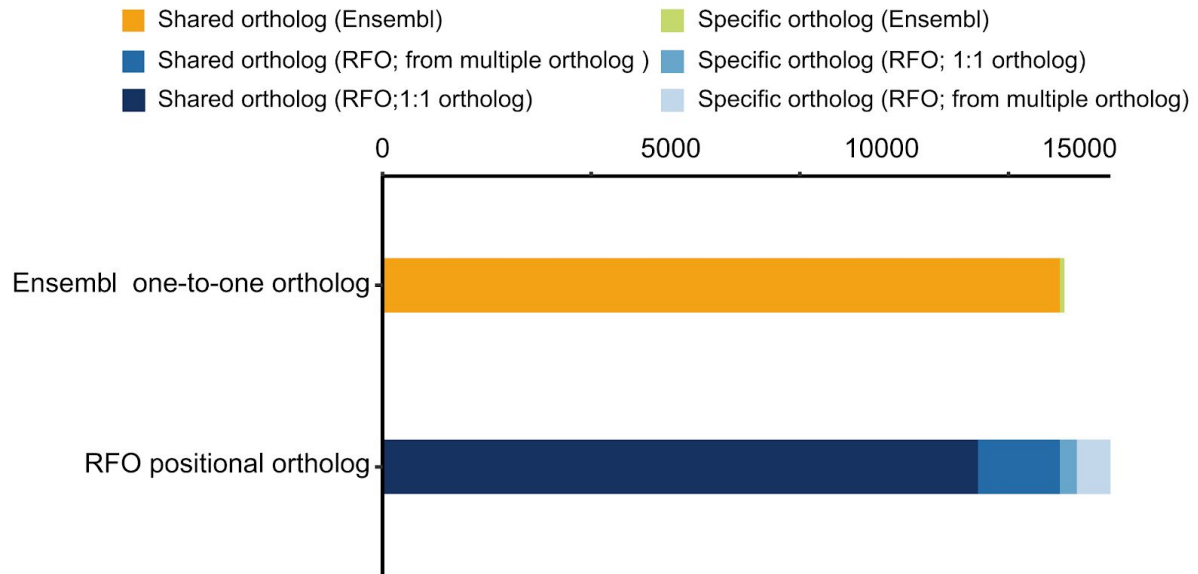**Supplementary Figure S2**. Comparison of nonsynonymous substitution rate(dN) between the novel one-to-one orthologs that were absent in the Ensembl orthologous table and other orthologs in human-gibbon. The bar indicates the mean and the error bars correspond to the SD (standard deviation) (*P*-value<0.01, Welch Two Sample t-test).

Supplementary Figure S3. Comparison between orthologs assigned by RFO and Ensembl.
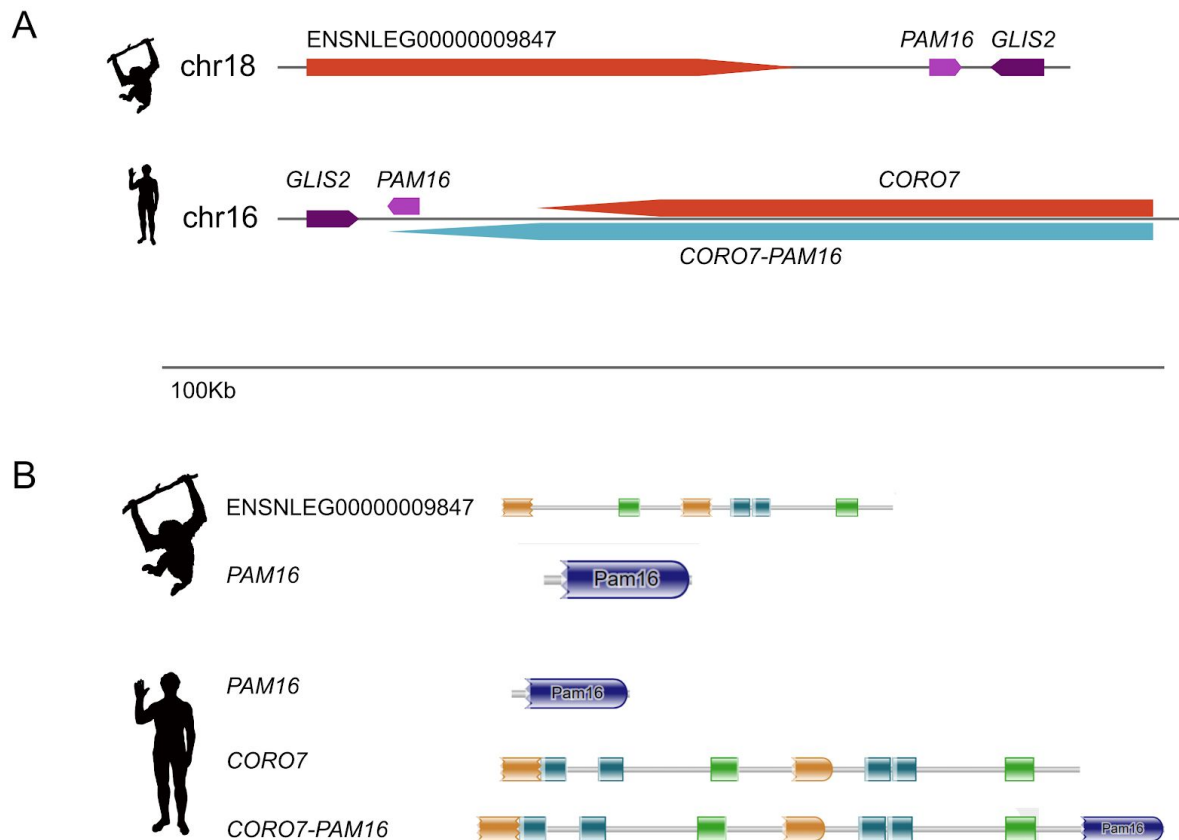
Supplementary Figure S4. An example of a complete homolog assignment of the RFO method. (A) The gene order of *PAM16*, *CORO7-PAM16* in human and gibbon. The same color represents the orthologous relationship. *CORO7-PAM16* is a homologous gene of *PAM16* and ENSNLEG00000009847 in gibbon (release 97 of Ensembl database). (B) The domain of *PAM16* and ENSNLEG00000009847 in gibbon and *PAM16*, *CORO7* and *CORO7-PAM16* in humans. The domain data was retrieved from the Hummer database (https://www.ebi.ac.uk/Tools/hmmer/). Animal illustrations are from phylopic.org, and are credited to NASA (*H. sapiens*, https://creativecommons.org/publicdomain/mark/1.0/)

Supplementary Figure S5. Comparison between orthologs assigned by RFO and Ensembl. The shared ortholog denotes a human gene that was assigned as ortholog by both the Ensembl method and the RFO method. The specific ortholog representing a human gene that was an assigned ortholog only in Ensembl 1:1 ortholog or RFO positional ortholog.
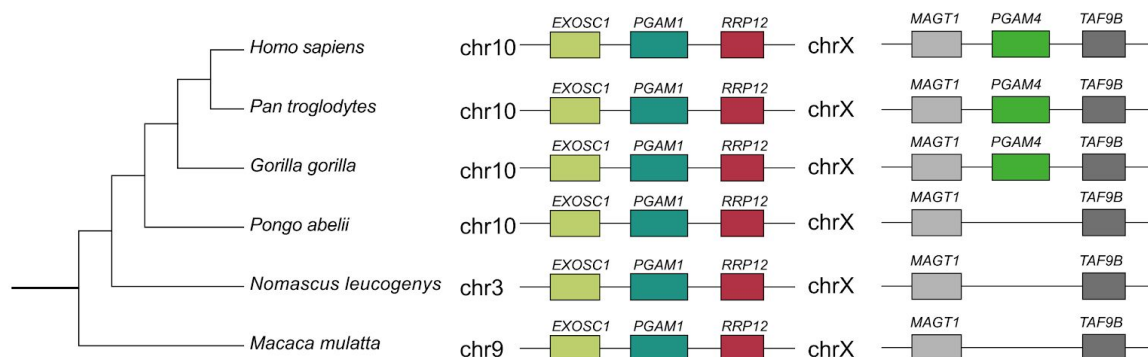
Supplementary Figure S6. *PGAM4* is a new gene generated by the duplication at the common ancestor of Hominidae.

Genome 1	Genome 2	Genome 3	Genome4
A	-	C	D

Re-annotated the gene A in Genome 2

Check whether the gene A from genome 1 can be annotated as gene B in genome 2

NO — Genomic loss

YES

Check whether gene A has frameshift and premature stop codon

YES — pseudogene

NO

Check whether gene B and gene A has a good gene synteny between genome 1 and genome 2

YES — False positive gene loss

NO

Check whether gene B is overlap with other genes

NO — Translocation
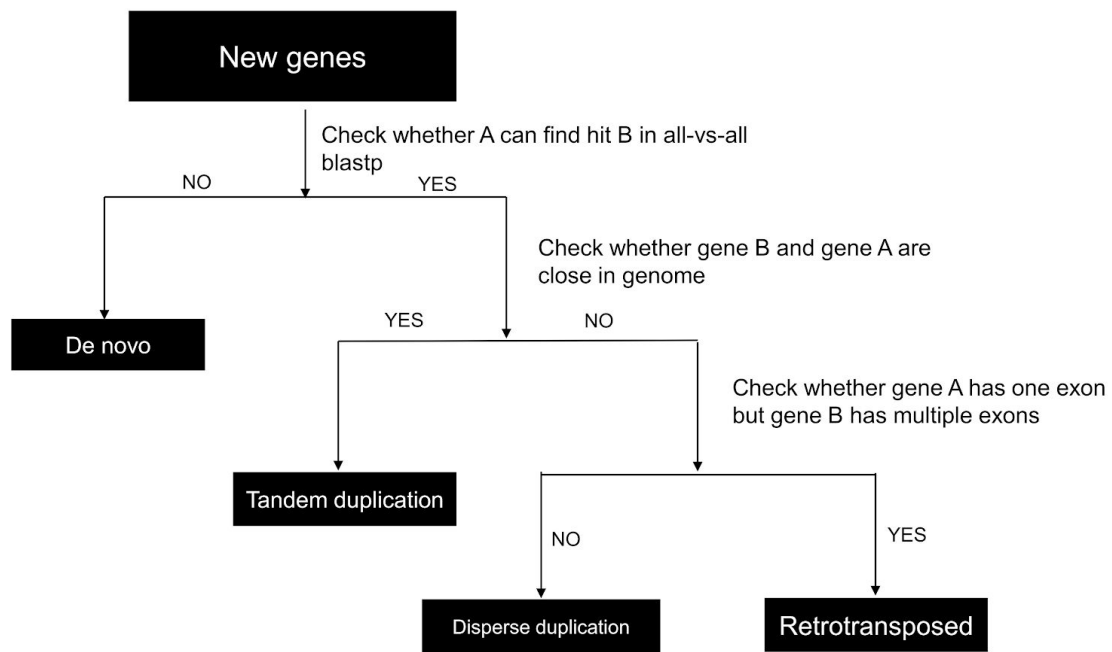
YES — Discard

Supplementary Figure S7. A hypothetical example shows the re-annotation and inference of gene loss mechanism. Genes with frameshifts and premature stop codons without assembly gaps in coding regions were considered as pseudogenes. The genes that had no gene synteny between species and had an overlap with existing genes were considered as redundant genes and were discarded.
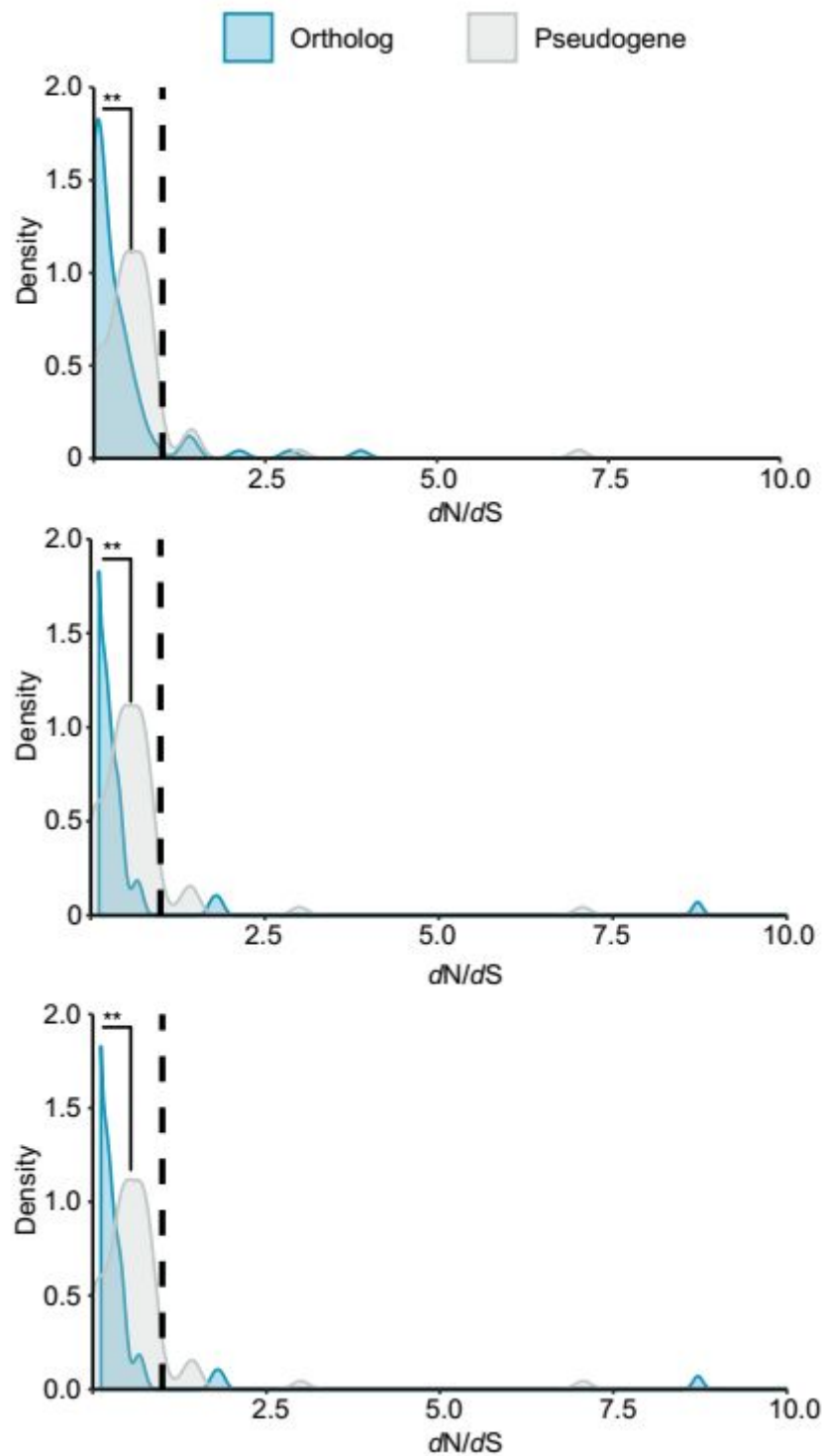
Supplementary Figure S8. Comparison of nonsynonymous to synonymous substitution rate ratio (*d*N/*d*S) of human-branch between human-specific pseudogene and 1:1 ortholog. *d*N/*d*S of pseudogenes are significantly higher than orthologs' (*P*-value < 2.2e-16,Wilcoxon rank-sum test). Dashed lines represent *d*N/*d*S=1.

```
                   ┌─────────────────┐
                   │    New genes    │
                   └─────────────────┘
                            │ Check whether A can find hit B in all-vs-all
                            │ blastp
        NO                  │     YES
   ┌────────────────────────┴──────┐
   │                               │ Check whether gene B and gene A are
   │                               │ close in genome
   ▼                     YES       │     NO
┌──────────┐         ┌─────────────┴────┐
│ De novo  │         │                  │
└──────────┘         │                  │ Check whether gene A has one exon
                     │                  │ but gene B has multiple exons
                     ▼              ┌────┴───────────────────┐
              ┌──────────────────┐  │                        │
              │Tandem duplication│  │   NO               YES │
              └──────────────────┘  ▼                        ▼
                          ┌────────────────────┐  ┌─────────────────┐
                          │ Disperse duplication│ │ Retrotransposed │
                          └────────────────────┘  └─────────────────┘
```

**Supplementary Figure S9**. The inferred new gene origination mechanism.
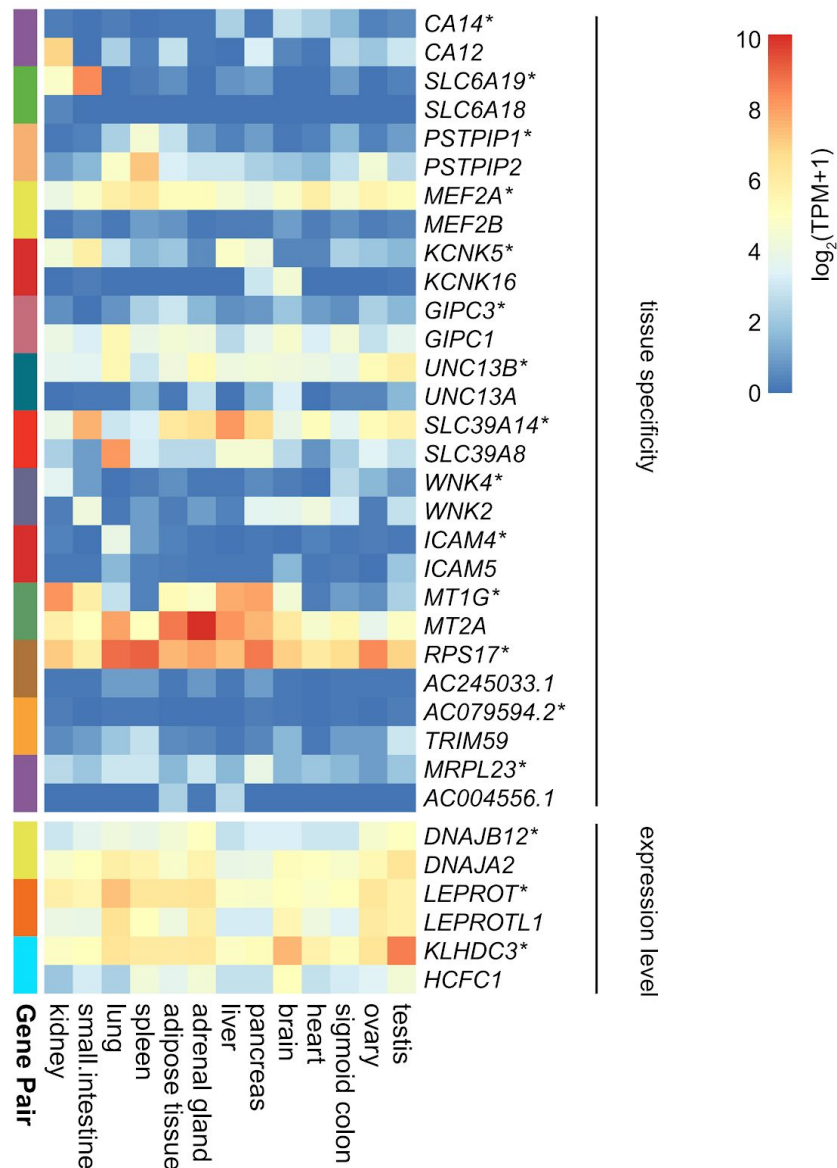
Supplementary Fig S10. Comparison of gene expression levels between PSGs and their closest Non-PSGs. Thirteen of 30 PSGs in humans had the closest Non-PSG. The row annotation depicts the gene pair of PSG and its closest Non-PSG. Asterisks denote PSGs, and other genes are Non-PSGs. Each PSG is next to its corresponding Non-PSG. The upper "tissue specificity" indicates that PSG and Non-PSG have different tissue specificity inside each gene pair. The lower "expression level" indicates PSG and Non-PSG expression across all the tissues in comparison but with different expression levels inside each gene pair.
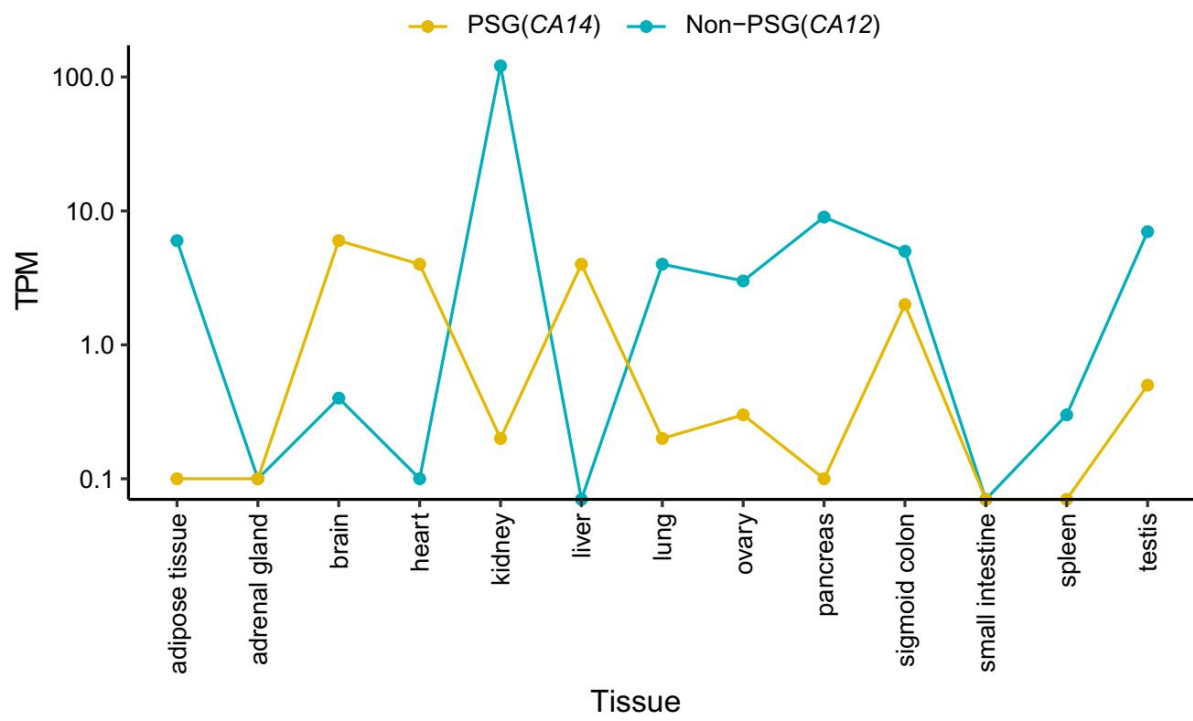
Supplementary Fig S11. Patterns of *CA14* and *CA12* expression in human tissues.
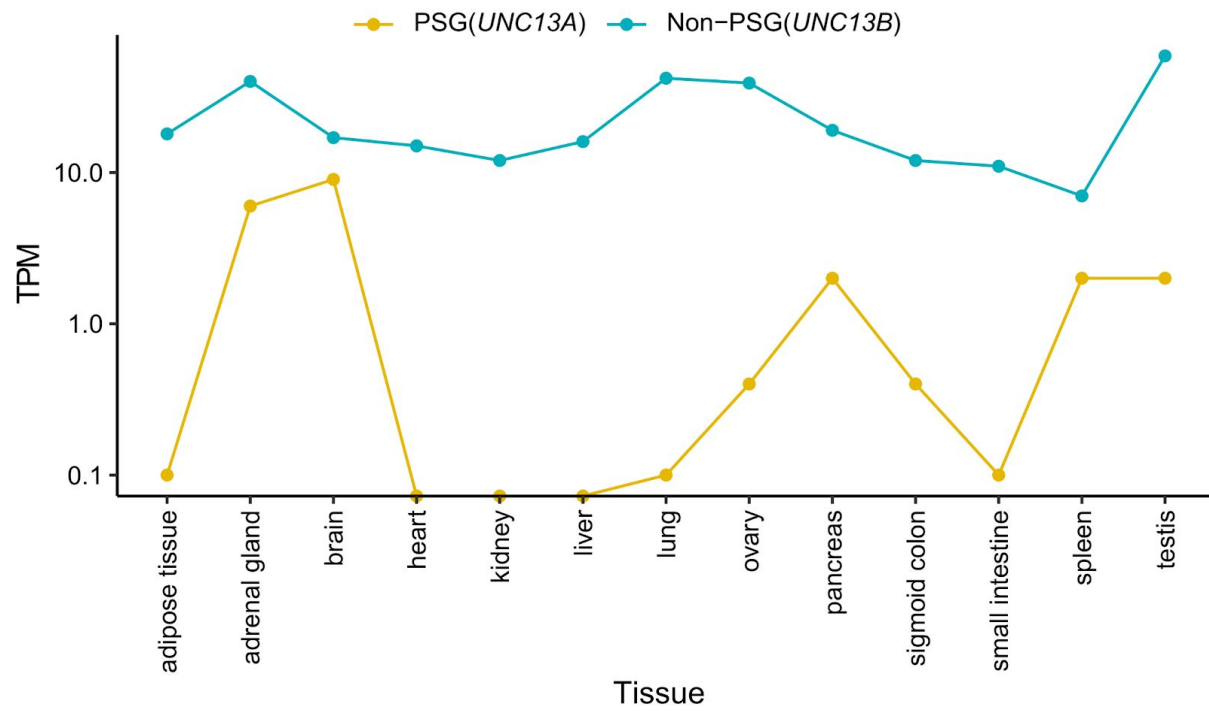
Supplementary Fig S12. Patterns of *UNC13A* and *UNC13B* expression in human tissues.
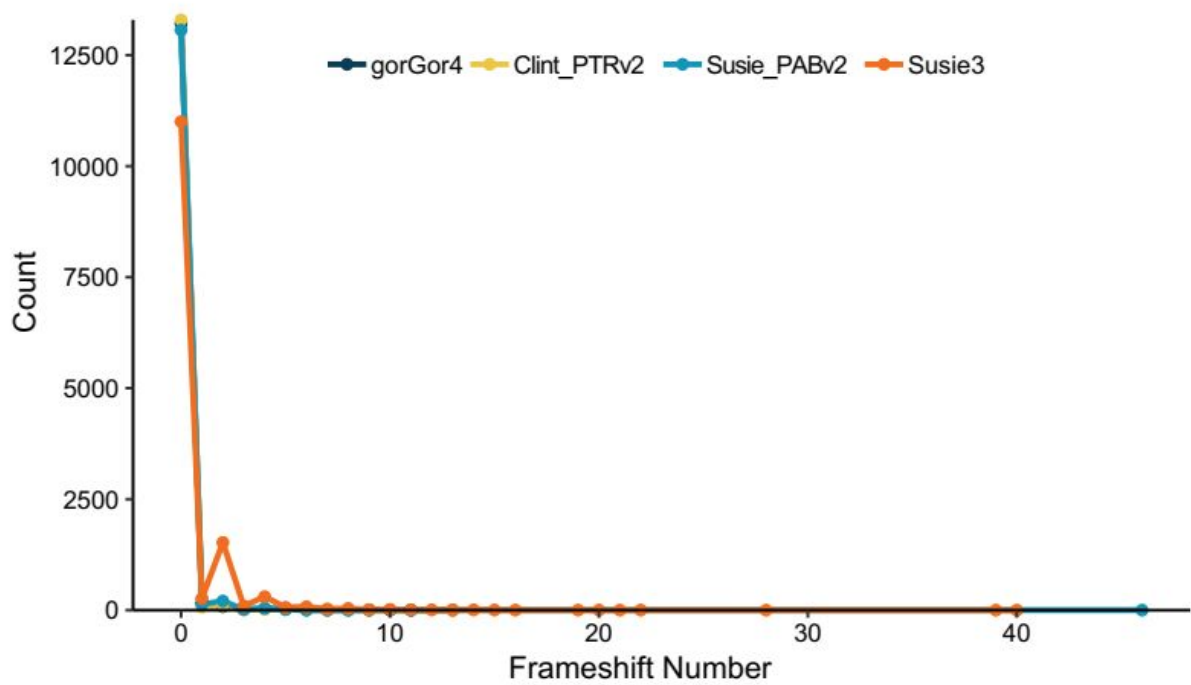
Supplementary Figure S13. Distribution of annotated frameshift numbers for a set of 13,480 genes that are conserved between human and mouse/rat/dog/cow in four different great ape genomes with Exonerate (exonerate –model protein2genome –n 1;v2.2.0). The color of lines and dots indicate the genome versions in the NCBI database.

Supplementary Table

Supplementary Table S1. Summary of assembly and annotation statistics for the primates genomes

Specie(Latin name)	Genome version	Genome size	Gene annotation source	#Protein coding gene
Homo sapiens	GRCh38.p12	3,099,750,718	Ensemb l(release 97)	19,971
Pan troglodytes	Clint_PTRv2	3,024,031,013	NCBI (release 105)	21,448
Gorilla gorilla	Susie3	3,084,595,669	Homolog annotation[a]	17,036
Pongo abelii	Susie_PABv2	3,065,035,716	NCBI (release 103)	20,562
Nomascus leucogenys	Nleu_3.0	2,962,077,449	Ensembl (release 97)	20,644
Macaca mulatta	Mmul_8.0.1	3,236,207,768	Ensemb l(release 97)	21,085

a. See detail in Materials and Methods section.