Ruiz-Orera et al. Origins of *de novo* genes in human and chimpanzee.

Supplementary Table

Ensembl gene ID	Assembly gene ID	Number of transcripts	Tissue-specificity	ORF length (aa)	Class
ENSG00000224186	XLOC_155668	2	Testis	96, 52	Overlapping antisense
ENSG00000253976	XLOC_193538	3	Testis	74, 49, 48	Overlapping Intronic
ENSG00000249016 (*)	XLOC_159345	1	Testis	34	Intergenic
ENSG00000263417	XLOC_088783	3	Testis	148, 136, 61	Intergenic
ENSPTRG00000041026	XLOC_236355	1	Testis	83	Intergenic
ENSPTRG00000041735	XLOC_047766	1	Brain	74	Overlapping Intronic
ENSPTRG00000041069 (*)	XLOC_227215	2	Testis	34, 34	Intergenic
ENSPTRG00000040082	XLOC_160846	1	Brain	35	Intergenic

Table 1. De novo genes annotated as protein-coding in Ensembl v. 75. Identification of annotated genes in the set of *de novo* genes was based on the comparison of the genomic coordinates of the assembled transcripts and the genomic coordinates of annotated genes using Cuffcompare. All these genes were hominoid-specific (expressed both in human and chimpanzee). (*) refers to the same orthologous gene in human and chimpanzee. Note that all human coding genes had been annotated as different classes of long non-coding RNAs (lncRNAs) in Ensembl v. 77.

Supplementary figures

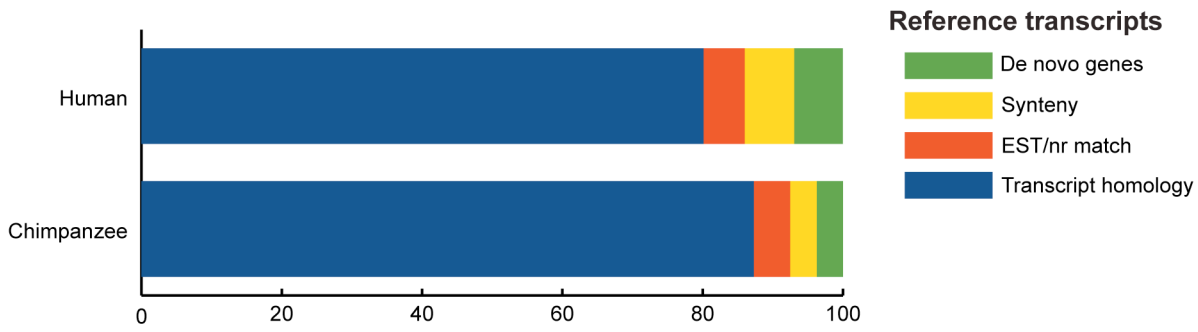

Fig. S1. Summary of the filters applied to obtain the final list of *de novo* genes specific of human or chimpanzee. Transcript homology: genes discarded because of homology to transcriptomes (assemblies or annotations) from other species using sequence similarity searches. Synteny: genes discarded because they overlapped other transcripts in genomic syntenic regions. EST/nr: genes discarded because they matched sequences from the EST or nr databases.

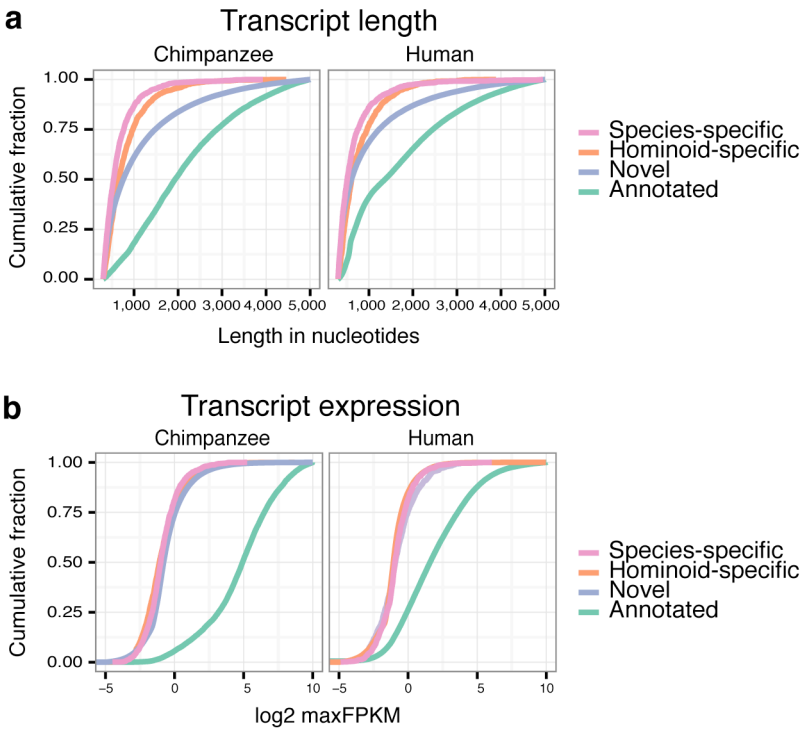

Fig. S2. Properties of *de novo* transcripts when compared to all annotated and novel transcripts. A) Cumulative density of length in species-specific, hominoid-specific, annotated and novel assembled transcripts. **B)** Log2 cumulative density of expression values in species-specific, hominoid-specific, annotated and novel assembled transcripts. Expression is measured in fragments per kilobase per million mapped reads (FPKM) values, selecting the maximum value across all samples. Collectively, *de novo* genes had a median size of 595 nucleotides and median expression of 0.31 FPKM. Species-specific transcripts are significantly shorter (Wilcoxon test, p-value $<10^{-16}$) than hominoid-specific transcripts, but no differences in expression levels are observed.

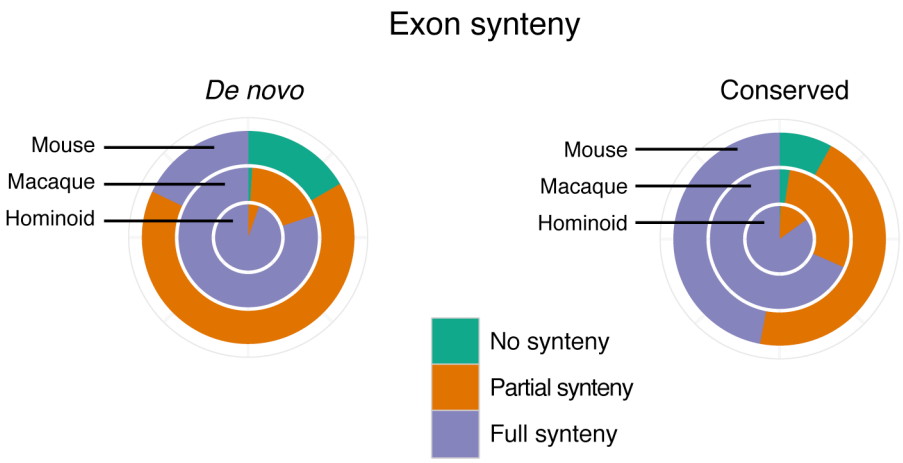

Fig. S3. Conservation of syntenic genomic regions corresponding to *de novo* or conserved genes. The existence of full or partial synteny was assessed using pairwise genomic alignments from UCSC. Hominoid (inner circle) refers to human when chimpanzee is the reference species and to chimpanzee when human is the reference species. The proportion of *de novo* and conserved transcripts with full or partial synteny decreases with phylogenetic distance. The proportion of transcripts from *de novo* genes with complete genomic synteny in macaque was comparable to that of transcripts from conserved genes.

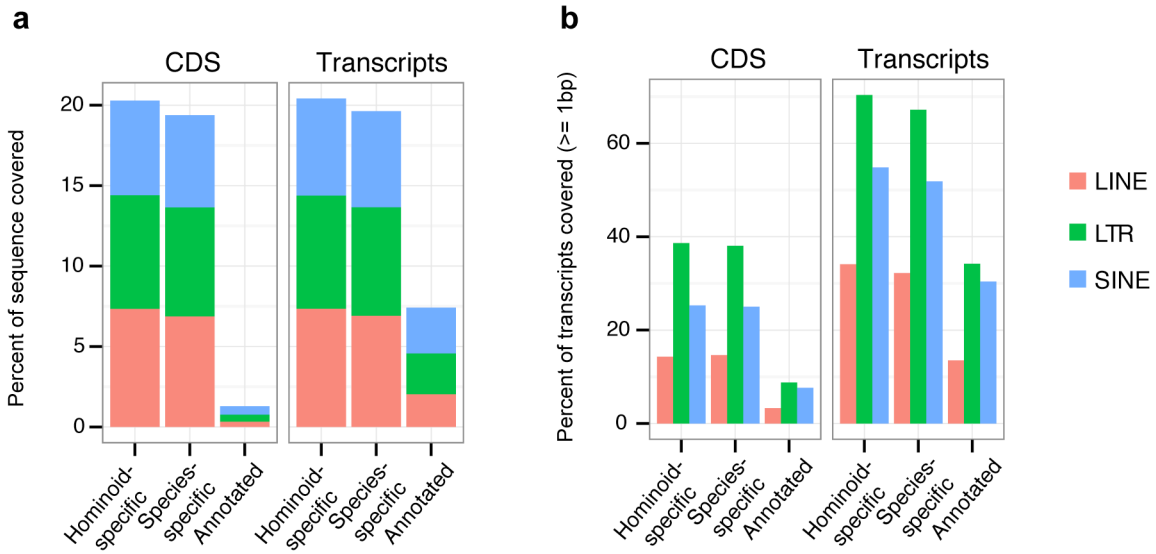

Fig. S4. *De novo* genes are enriched in transposable elements. Transcrips covered by transposable elements (TEs) considering all annotated transcripts, hominoid-specific genes or species-specific genes (human- or chimpanzee-specific genes). CDS is the annotated coding sequence in annotated protein-coding transcripts and the longest ORF in *de novo* transcripts. Classes of TEs: LINEs; long interspersed elements; LTRs, long terminal repeats; SINEs, short interspersed elements. **A)** Average fraction of transcript length covered by TEs. **B)** Number of transcripts covered by TEs (>= 1bp overlap).

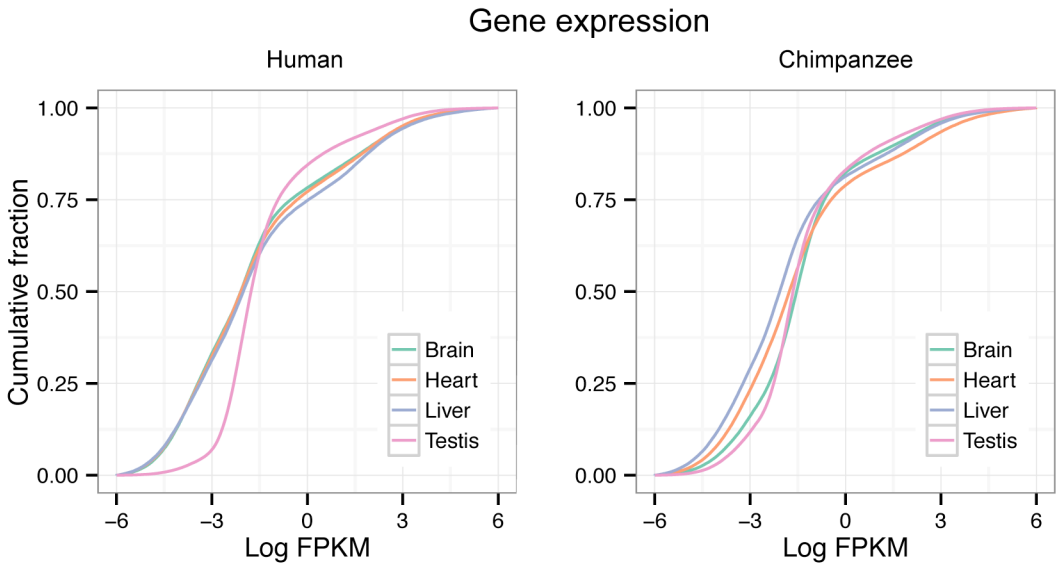

Fig. S5. Distribution of expression values in assembled genes across tissues. Log10 cumulative density of expression values in assembled genes. Expression is measured in fragments per kilobase per million mapped reads (FPKM) values, selecting the maximum value across all samples. Testis does not show a lack of highly expressed transcripts (actually the opposite is observed for human) that could explain why we detect so many transcripts being expressed in this tissue.

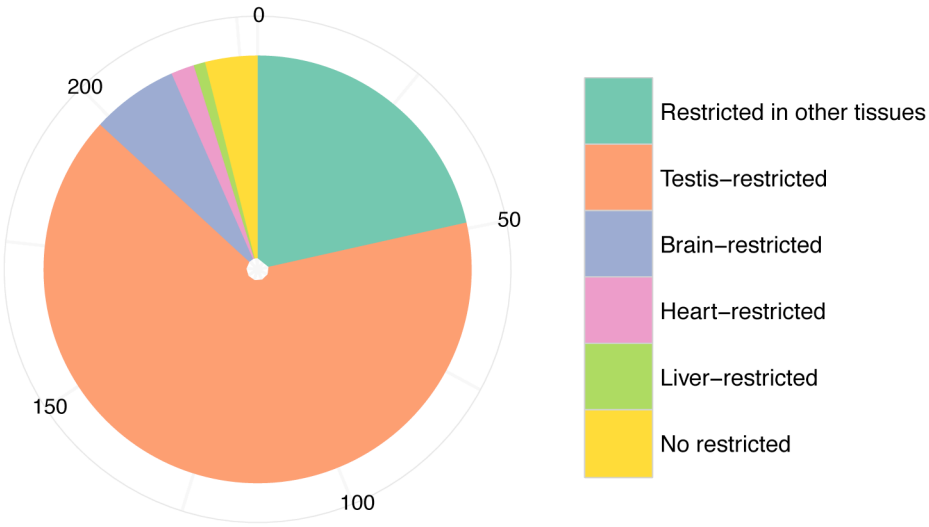

Fig. S6. Human annotated transcripts from *de novo* genes are enriched in testis according to GTEx data. Data is for annotated transcripts in the GTEx catalog which are preferentially expressed in one tissue, as measured by a tissue preferential expression index higher than 0.85 (see Methods online for more details on this index).

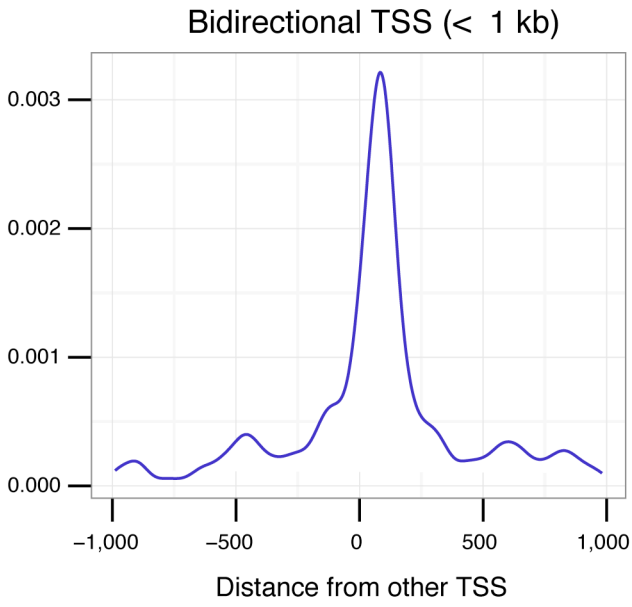

Fig. S7. Distance between the transcription start site (TSS) of transcripts from *de novo* genes and the nearest TSS from another transcript, for genes with divergent transcription. These were defined as antisense genes with the TSSs separated by less than 1 kb, potentially sharing a bidirectional promoter. Negative values imply overlap between the transcripts. There is a strong peak at around 100 nucleotides.

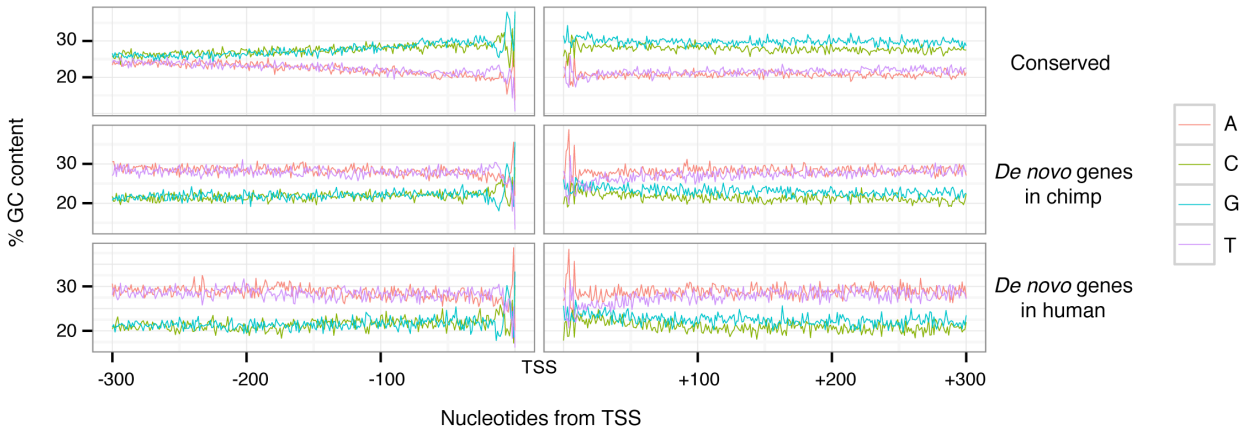

Fig. S8. De novo genes have a low GC content when compared to conserved annotated genes. Nucleotide frequencies 300 bp upstream and 300 downstream of the transcription start site (TSS) were calculated for different sets of transcripts. Conserved: 4,323 randomly taken human and chimpanzee annotated transcripts not classified as *de novo*.

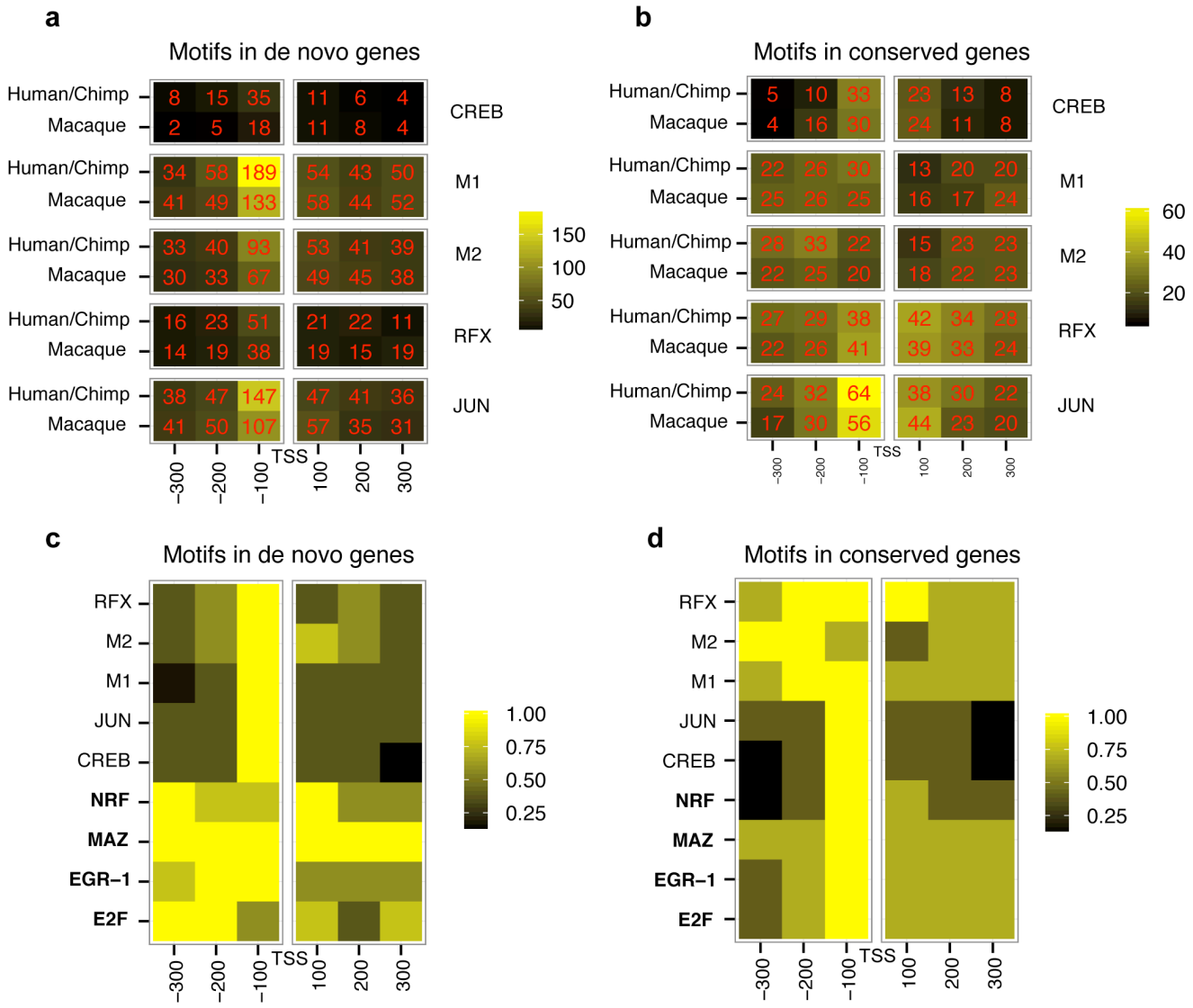

Fig. S9. Regulatory motif frequencies around the TSS. A Number of matches of overrepresented motifs in 100 bp windows in *de novo* genes and in the corresponding macaque syntenic regions (corresponds to Figure 3a in main manuscript file). **B** Same data for conserved annotated genes. **C** Relative motif frequencies in *de novo* genes including motifs overrepresented in conserved annotated genes in general but not in *de novo* genes (NRF, MAZ, EGR-1, E2F). **D** Data for the same motifs for conserved annotated genes.

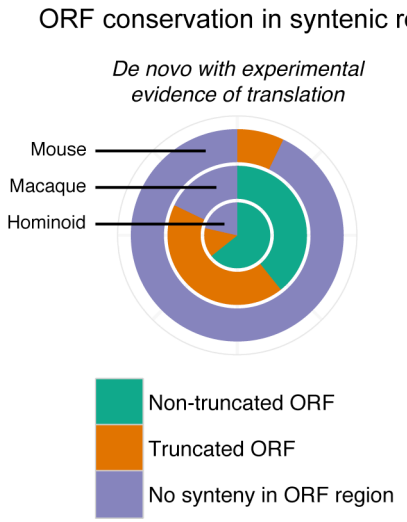

Fig. S10. Conservation of ORFs in syntenic genomic regions corresponding to *de novo* genes with experimental evidence of translation. The existence of full or partial synteny was assessed using pairwise genomic alignments from UCSC. Hominoid (inner circle) refers to human when chimpanzee is the reference species and to chimpanzee when human is the reference species. Only ORFs in *de novo* genes with evidences of preoteogenomics or ribosome profiling are displayed. Non-truncated ORFs are the ones in which the frame, the start codon and the stop codon are conserved in the other syntenic genomic region; otherwise the ORF is truncated.