Supplementary file 2

Origin of primate orphan genes: a comparative genomics approach
Toll-Riera et al.

S1.

Primate Gene	Primate Protein	Chr	Primate gene coordinates	Ka/Ks	Pseudogene accession number	Parent Gene	Pseudogene coordinates	Pseudogene type
ENSG00000214610	ENSP00000381653	9	125007105-125012941	0.86	126201	ENSG00000165533	125007204-125007370	Ambiguous
ENSG00000214819	ENSP00000382000	17	20423629-20424816	1.07	25997	ENSG00000141028	20423649-20424628	Duplicated
ENSG00000215153	ENSP00000382606	10	38777208-38780939	NA	119502	ENSG00000206055	38777774-38780174	Duplicated
					127384	ENSG00000196149	38779653-38780006	Ambiguous
ENSG00000204898	ENSP00000367037	3	40492714-40517175	NA	75422	ENSG00000322265	40517102-40517471	Ambiguous
ENSG00000188639	ENSP00000341218	7	56839246-56856131	NA	133460	ENSG00000196149	56855936-56856871	Duplicated
ENSG00000203450	ENSP00000381725	7	62869368-62871389	NA	133491	ENSG00000196149	62869448-62869996	Processed
					125307	ENSG00000196149	62870483-62870976	Processed
					125308	ENSG00000196149	62871344-62871847	Processed
ENSG00000205565	ENSP00000370103	5	69842535-69843804	NA	132341	ENSG00000188948	69843558-69843989	Processed
ENSG00000196436	ENSP00000347438	16	72969373-72983514	NA	121295	ENSG00000169203	72982790-72983173	Ambiguous
ENSG00000196648	ENSP00000351392	15	80895830-80904928	0.36	128978	ENSG00000206128	80900273-80900742	Processed
					131806	ENSG00000140478	80900649-80900807	Ambiguous
ENSG00000205976	ENSP00000371990	5	7352487-7354545	NA	132203	ENSG00000174450	7353715-7358319	Duplicated

S2.

Analysis of regulatory motifs in human gene promoters

1. Significant regulatory motifs in human gene promoters

Motif/TFBS	n_genes	%_genes	Significant Regions
GC-box/Sp1	4923	33.5%	-169 .. +19
YY1/NF-E1	3849	26.2%	-38 .. +93
GABP/ETS	2662	18.1%	-121 .. +39
CAAT-box/NF-Y	1784	12.2%	-167 .. -2
CREB/ATF	1646	11.2%	-115 .. +25
NRF1	1156	7.9%	-122 .. +29
MYF	652	4.4%	-10 .. +100
TATA-box	490	3.3%	-78 .. +11
EVI1	383	2.6%	+13 .. +83
FOXD1	210	1.4%	-10 .. +57
OCT1	179	1.2%	+45 .. +100

The table contains the number of genes with predicted regulatory motifs and the significant position ranges with respect to the transcription start site (TSS). The results were obtained by PEAKS (http://genomics.imim.es/peaks) in a dataset of 14,678 human genes, using window size 30 and p-value < 10^{-5}. The 11 motifs were obtained after clustering redundant motifs obtained using TRANSFAC and JASPAR motif libraries (see below). The presence of these motifs in gene upstream sequences (in Significant Regions) was determined for primate-specific genes in Table 3.

2. Clusters of redundant regulatory motifs

The clustering of motifs was required to provide the results of point 1 in a neat form. We clustered the initial motifs using hierarchical clustering (R package, complete hierarchical clustering). Distance between motifs was based on the proportion of overlapping motif matches along all non-redundant promoter sequences. We obtained 11 clusters, which are the motifs shown in point 1.

```
>CLUSTER_1 -> EVI1
V$EVI1_02          TRANSFAC 7

>CLUSTER_2 -> OCT1
V$OCT1_05          TRANSFAC 7

>CLUSTER_3 –> TATA-box
TBP                JASPAR CORE

>CLUSTER_4 –> YY1/NF-E1
V$NFMUE1_Q6        TRANSFAC 7
V$YY1_02           TRANSFAC 7
V$YY1_Q6           TRANSFAC 7

>CLUSTER_5 -> CAAT-box/NF-Y
NF-Y               JASPAR CORE
V$ACAAT_B          TRANSFAC 7
V$ALPHACP1_01      TRANSFAC 7
V$NFY_C            TRANSFAC 7

>CLUSTER_6 -> MYF
Myf                JASPAR CORE
```

>CLUSTER_7 -> NRF1
V$NRF1_Q6 TRANSFAC 7

>CLUSTER_8 -> GC-box/Sp1
V$SP1_Q2_01 TRANSFAC 7
V$SP1_Q4_01 TRANSFAC 7
V$SP1_Q6_01 TRANSFAC 7

>CLUSTER_9 -> FOXD1
FOXD1 JASPAR CORE

>CLUSTER_10 -> CREB/ATF
CREB1 TRANSFAC 7
V$ATF4_Q2 TRANSFAC 7
V$ATF_B TRANSFAC 7
V$CREBP1CJUN_01 TRANSFAC 7
V$CREBP1_Q2 TRANSFAC 7
V$CREB_01 TRANSFAC 7
V$E4F1_Q6 TRANSFAC 7

>CLUSTER_11 -> GABP/ETS
ELK1 JASPAR CORE
ELK4 JASPAR CORE
GABPA JASPAR CORE
V$ELK1_01 TRANSFAC 7
V$TEL2_Q6 TRANSFAC 7

3. Analysis of regulatory motifs in genes of different phylogenetic distribution

Using the motifs in point 1, we calculated the average number of regulatory motifs per gene in the different datasets Primates, Mammals, Vertebrates and Eukarya. For each of these groups, we compared the results to 1000 random datasets of similar composition, generated using three order 1 Markov models for regions of different GC content (see Bellora et al., 2007a). The p-value was calculated using the Z statistic.

Dataset	Av. N motifs	Av. N motifs random	p-value
Primates	0.85	0.65	10^{-3}
Mammals	1.16	0.75	$< 10^{-5}$
Vertebrates	1.25	0.81	$< 10^{-5}$
Eukarya	1.70	0.97	$< 10^{-5}$

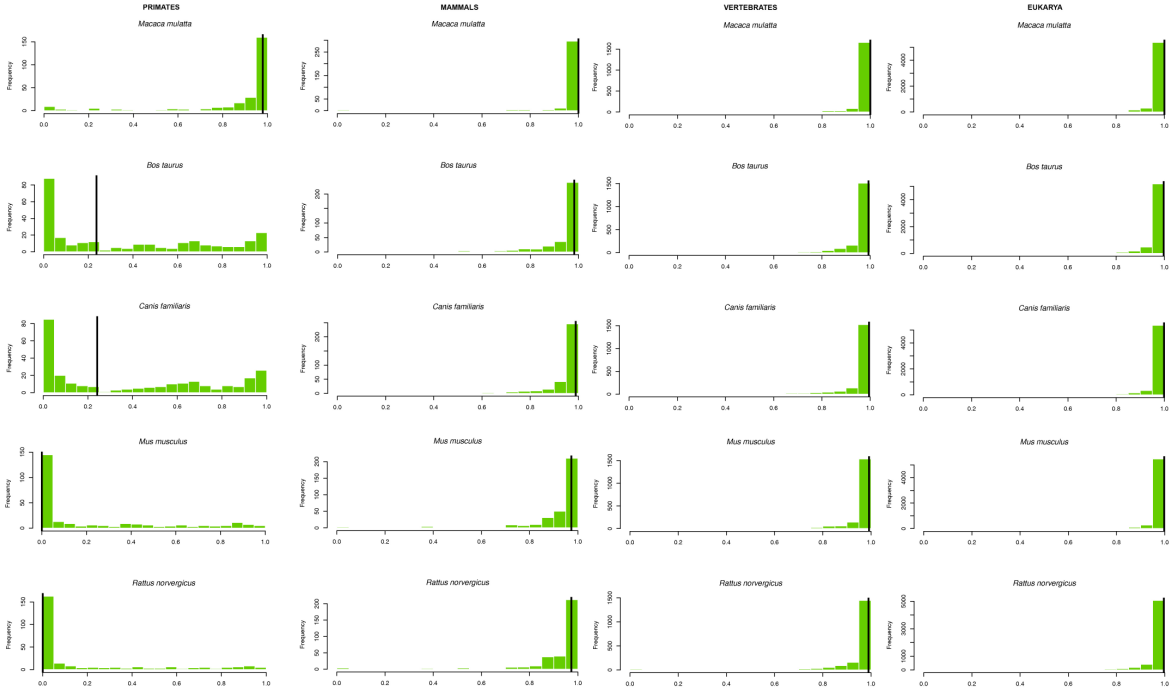
References

http://evolutionarygenomics.imim.es/datasets/nbellora/Hs_promoters/PEAKS_results/predictions.html

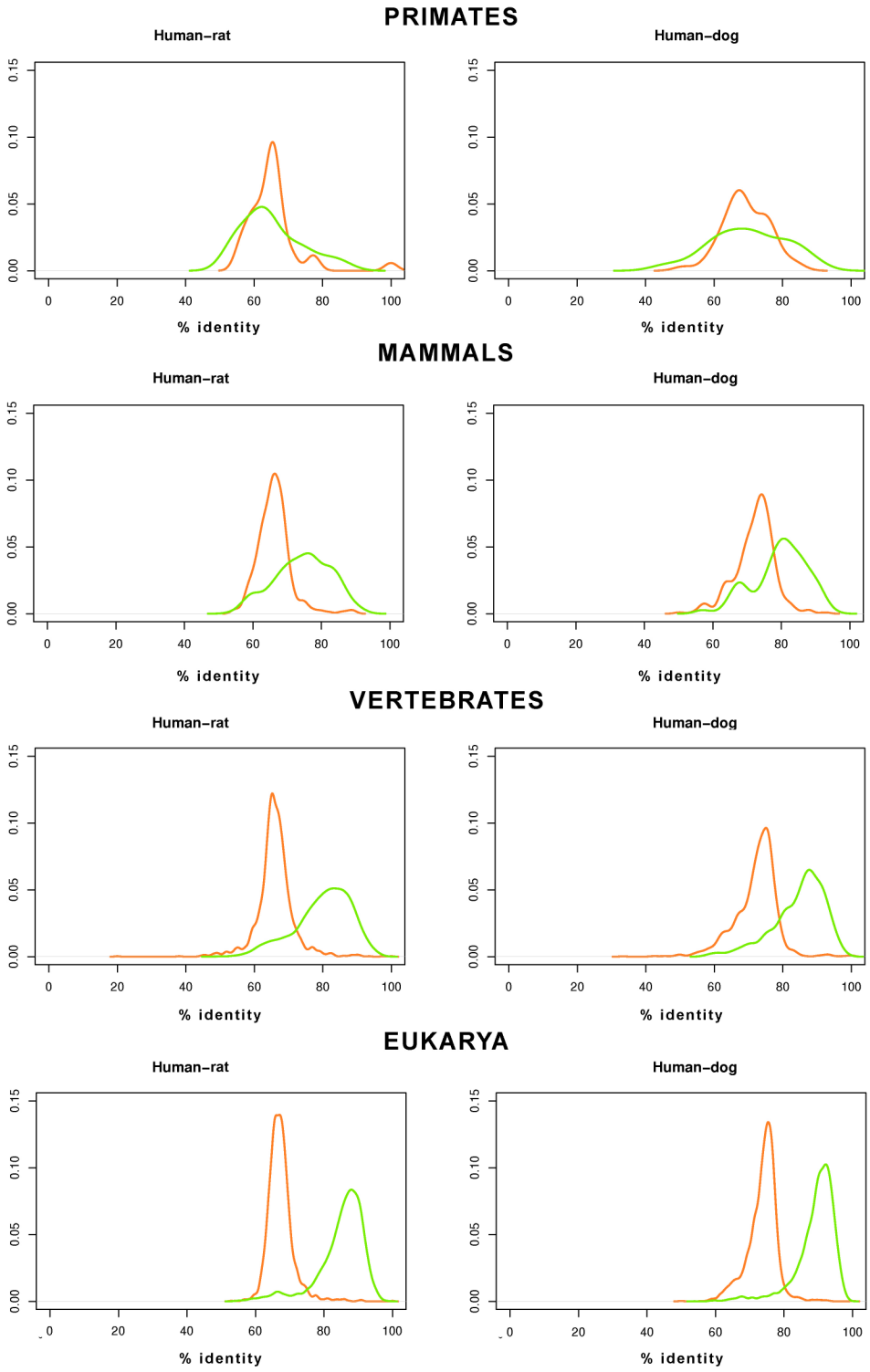
Bellora N, Farre D, Alba MM (2007a) Positional bias of general and tissue-specific regulatory motifs in mouse promoters. BMC Genomics 8:459

Bellora N, Farre D, Mar Alba M (2007b) PEAKS: identification of regulatory motifs by their position in DNA sequences. Bioinformatics 23:243-244

S3.

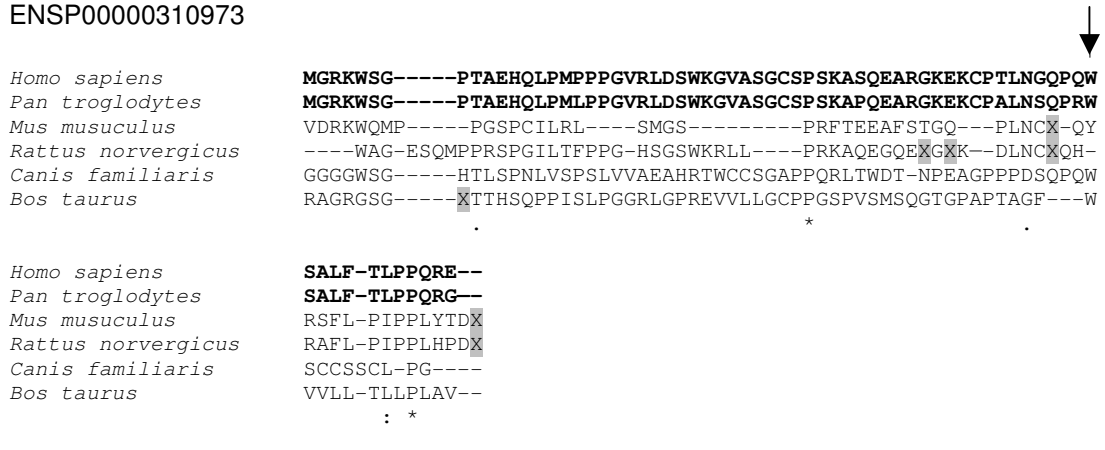

PRIMATES
MAMMALS
VERTEBRATES
EUKARYA
Macaca mulatta
Bos taurus
Canis familiaris
Mus musculus
Rattus norvergicus
Frequency

S4.

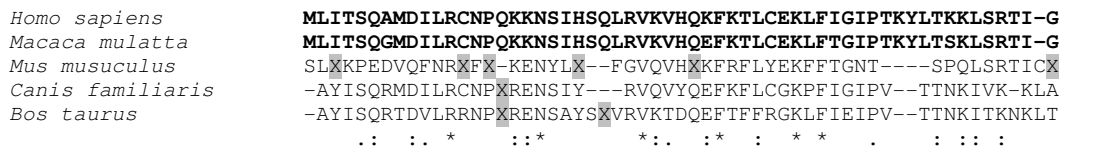

PRIMATES
Human–rat
Human–dog
MAMMALS
Human–rat
Human–dog
VERTEBRATES
Human–rat
Human–dog
EUKARYA
Human–rat
Human–dog
% identity
0.00
0.05
0.10
0.15
0
20
40
60
80
100

S5.

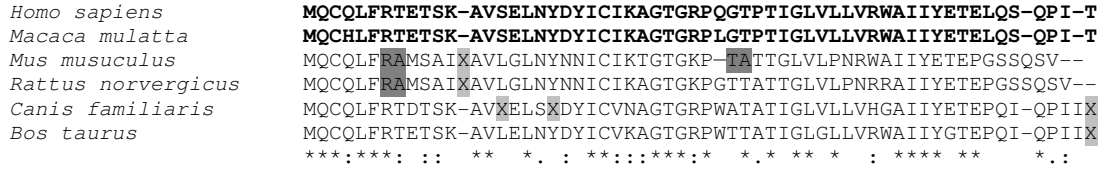
ENSP00000310973



ENSP00000317176



ENSP00000354582



ENSP00000326404

```
Homo sapiens          MIQSIWLDSQLFQLLSLPARCISDGTHQTLNSFPELESATEVFRASSVAALMETKKSLR
Macaca mulatta        MIQNIWLDSQLFQFLSVPARCISDGTHQTLNSFPELESSAEVFRASSVAALMETKRSLR
Mus musuculus         -WREISDSTHKYSALPPPP---LEGFLGTLSPXIHPKFNGRHMEKEKA----------R
Rattus norvergicus    LTGNVWQCSXMLSPLSYGK---APSEHPVPRSILRLRSSVVTTLGRK-----------R
Canis familiaris      ---AREYRVHKYSNSHLLPKGIPDGTGQSLSSFPELESSTGILRGSC----------MR
Bos taurus            --DAEYLGAQIFQLPSLLXRGIPDAADHSPSLFSEQESYAELFRGTSV----------K
                                    .           .         . .                   :

Homo sapiens          AGNLHHDQPIT-QEHSQLLLQGCT-C
Macaca mulatta        AGNLHHNQPIT-QEHSQLLLQRCT-Y
Mus musuculus         IGGLHQIQPPPGQE-A-IFCERAL--
Rattus norvergicus    LGSVACTRSSCHQDREPPFVRQPX--
Canis familiaris      VGDLYQGHLVT-QTESQILLXDGTSY
Bos taurus            VGNLHHDQPIT-QEHSQLLLQGCT-C
                       *.:   :    *     :
```

ENSP00000364363

```
Homo sapiens          MTDGRWLRRSWRGDGQCHD-TGHQPGPAALTPQDQAPLAPALAQPRYLDTDNLSPTKETRA
Macaca mulatta        MTDRGWLRRSWRGDDQRYD-TGHQPGPAALTPQDQAPLAPALARPRYLDTDNLSPTKETRA
Mus musuculus         MRDKSCLERSWTRAFQDCN-PGYYHGHFTLL--DKGYLAPVVAFPRXQDRHLESYQGNKCC
Rattus norvergicus    MRDKRSLKRSWTRACQCCD--RLPSWPLSSSR-QRS-PSPRYGSARXQDLIIWSLIKETSV
Canis familiaris      VKDRR--RRHWKGGCQSHD-PRXH---LALWPSPHKTVIRGPSSGPAXEPDTCSLTKETRA
Bos taurus            LRARSW-PRSWRGD-RRCHVPGVSLAPQP-SPLKTKTHGSALAGPRDPDTETWSLSKETGC
                      :       * *    :  .         .             .     :    *   :.

Homo sapiens          GIGANLSSASGHWVLGSQLFQRLG-QLPGLLRLPALFNKTRK
Macaca mulatta        GIGANLSSASGHWVLGSQLFQRLG-QLPGLLRLPALFNKTRK
Mus musuculus         PI------CRGHGALWSXWF------HTQTTTWGICXADPF-
Rattus norvergicus    AQC-----AGGHRAXWSRLL------PTQTVTWVSCXADPS-
Canis familiaris      GIGANLSSASGHWVLGSQLFQRLG-QLPGLLRLPALFNKTRK
Bos taurus            LGHRTLTCRAPR----DRWFRRQK-TIRDSLPTCTFXQSTER
                                 :    .  :                 ..
```

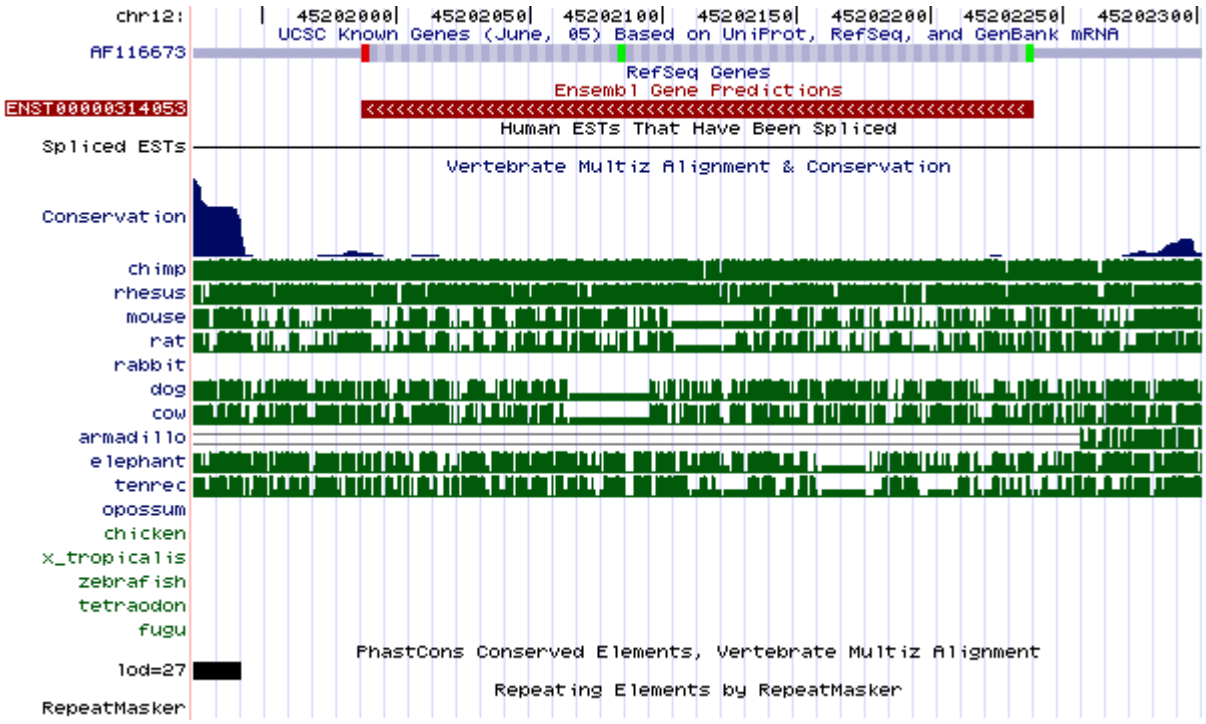
ENSP00000367499

```
Homo sapiens          MDILLDLGWH-FSNCDEDTFYSPVQNTEGDLLFFDHNLK-TDRGHVERSVMD
Macaca mulatta        IDILLDLGWH-FSNCNEDTFYSPVQNTEGDLLFSDHNINATDRGHVERSDMD
Mus musuculus         LIL--ELG---KXTC----FLRPAQN-RGDSLSSSHHQSAE---VENSE---
Rattus norvergicus    SIL--EFG---XXTC----FHRPVXN-RGDGLSSPHCQSTE--EVEKSE---
Canis familiaris      IDI-L-XSWPGLSNCNEDTFYSPAQNT-GDLLFSDHNLN-AVRXYKE-SVMD
Bos taurus            LVIL-DLHWF-FQTAKRHILX--LSSESGDPVFFVHNLN-TVREHR--SVMD
                        :          ..    :     .  ** :   *  .         .
```

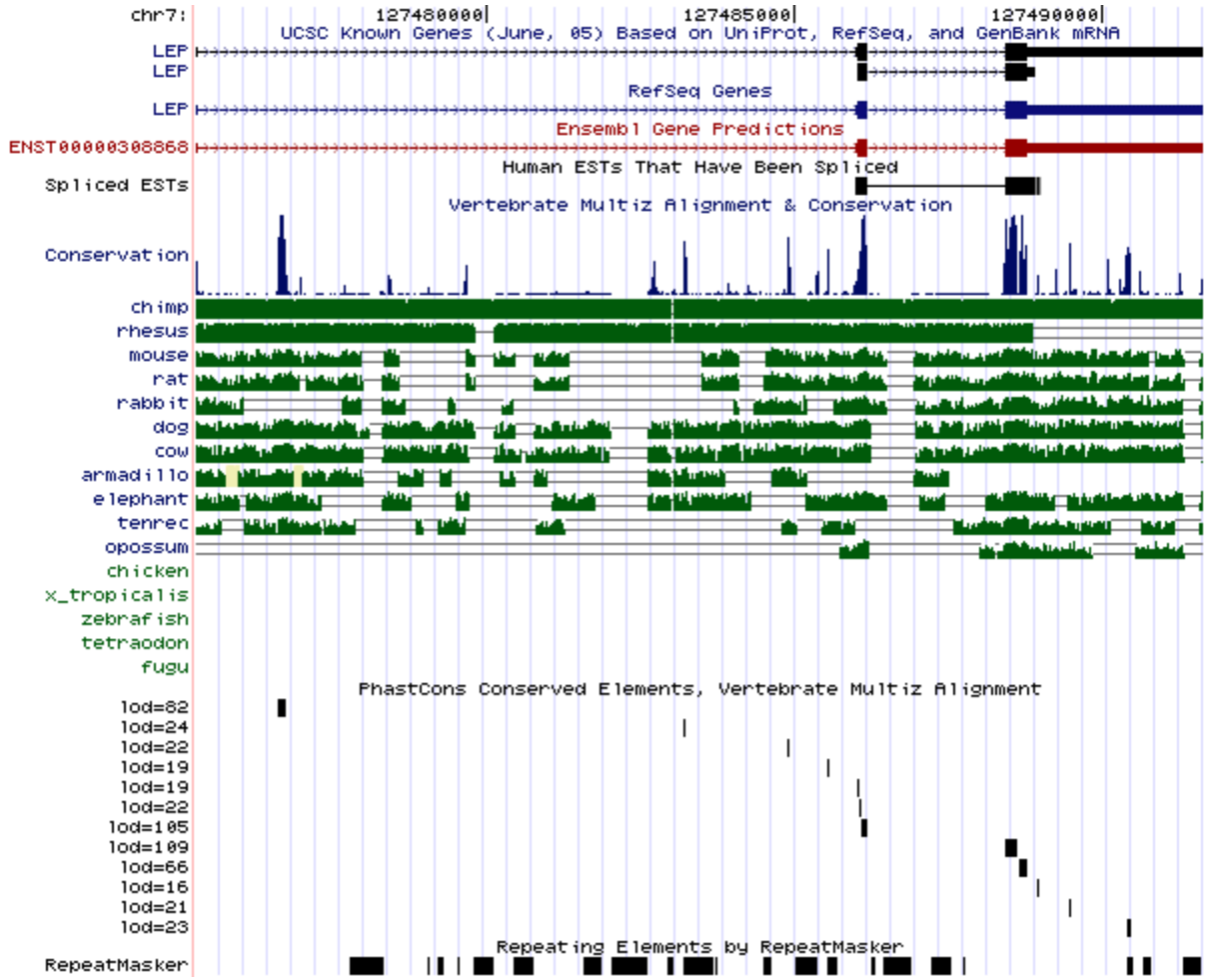
S6.

USCS Genome Browser results for 8 primate-specific genes + GAPDH

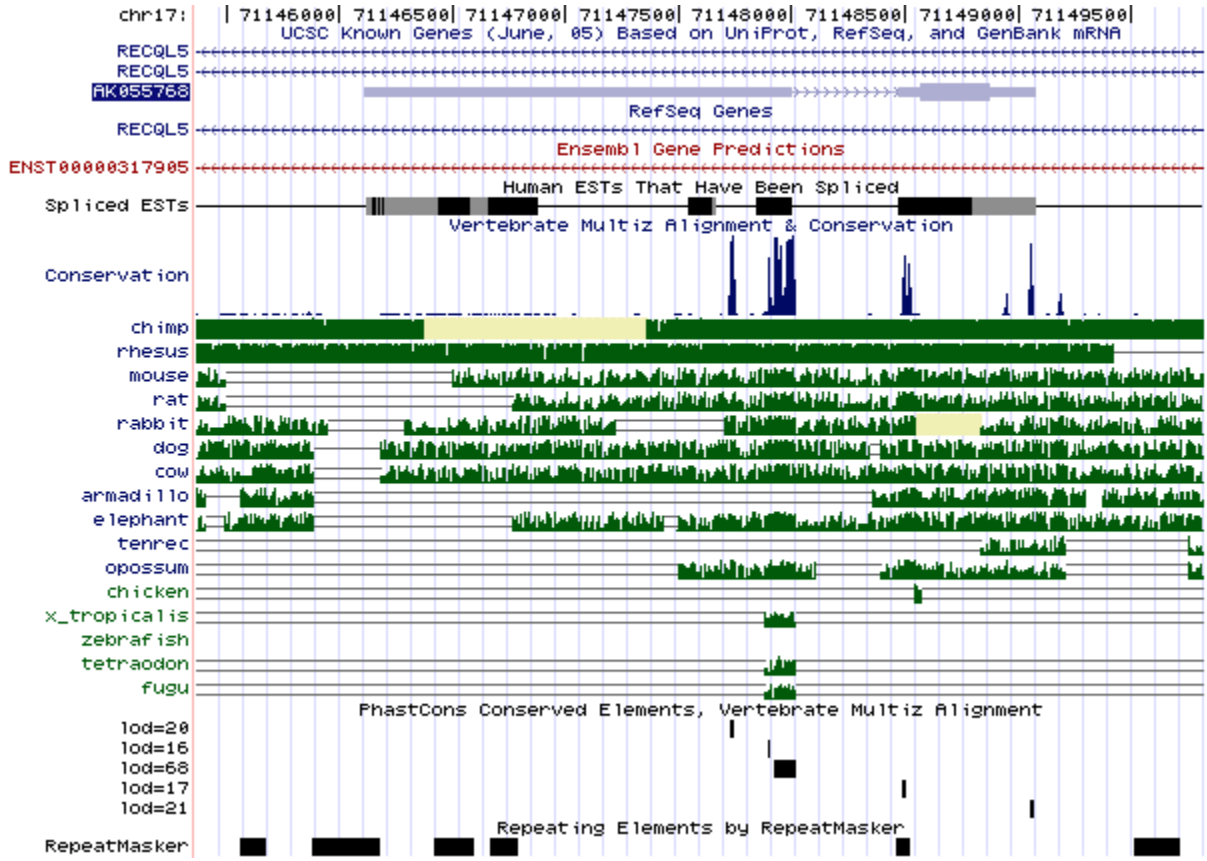
ENSG00000180547
chr12:45,201,988-45,202,239
human genome assembly NCBI36



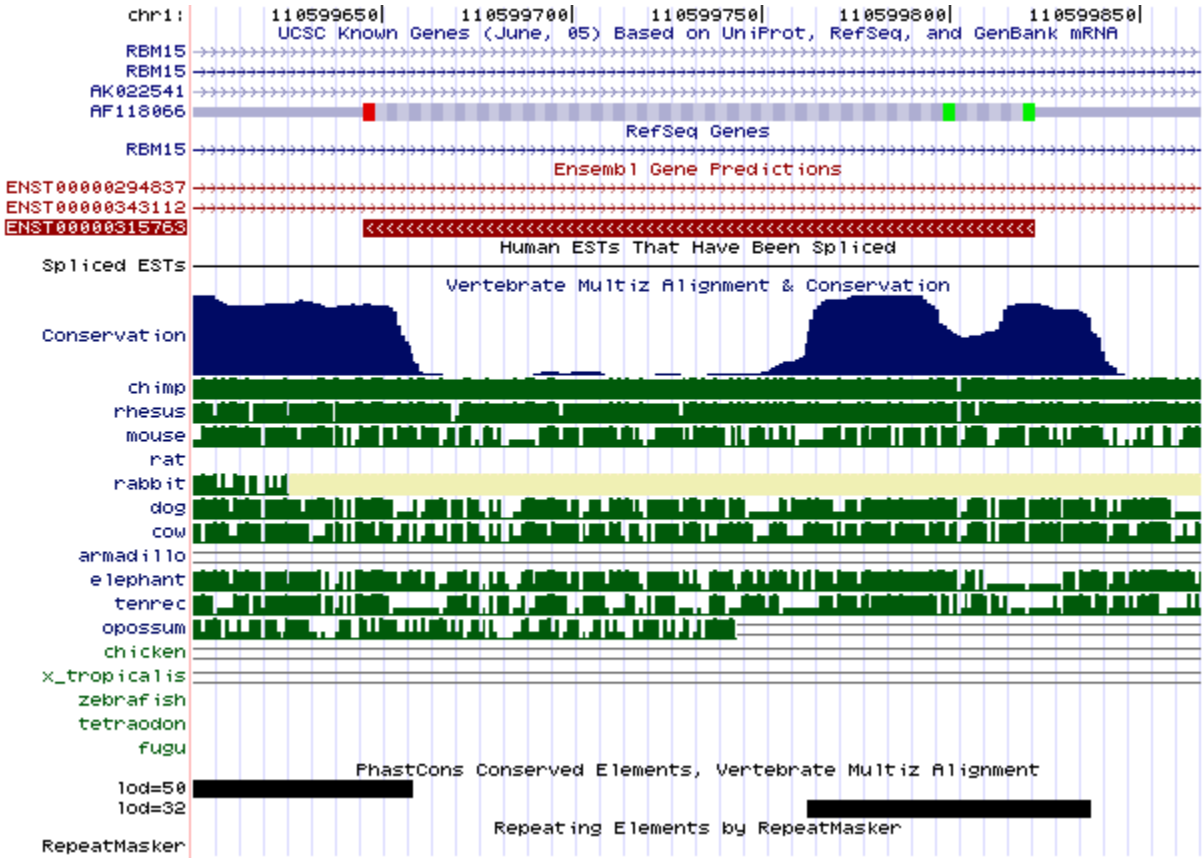
ENSG00000204537
chr7:127,475,283-127,491,632
human genome assembly NCBI36



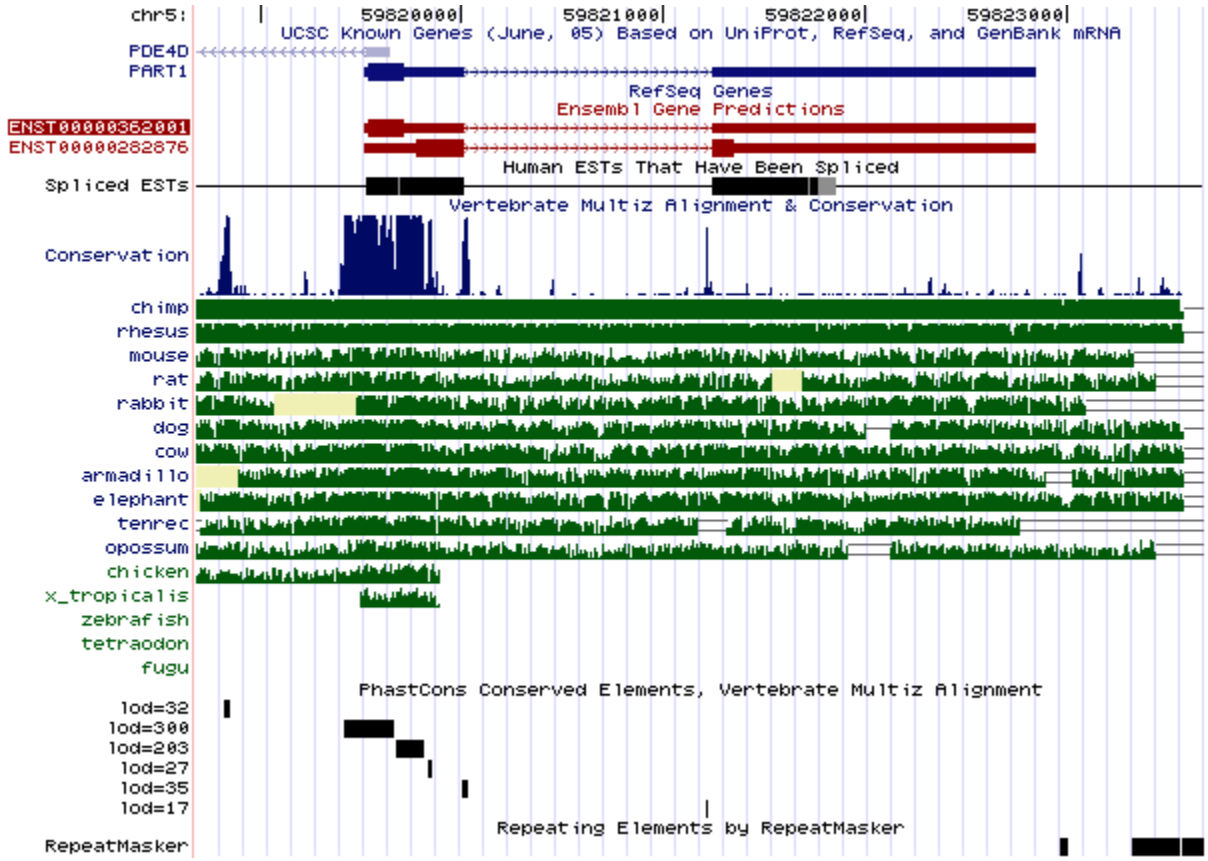
ENSG00000204323
chr17:71,145,366-71,149,822
human genome assembly NCBI36



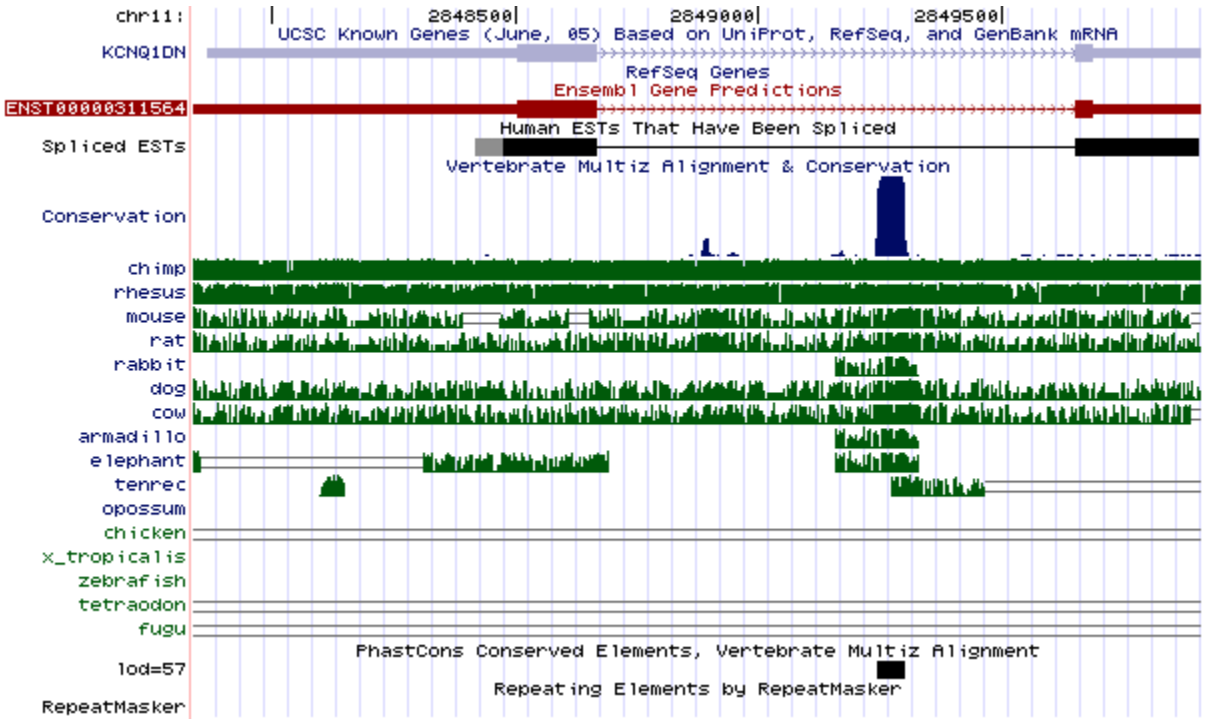
ENSG00000180441
chr1:110,599,646-110,599,822
human genome assembly NCBI36



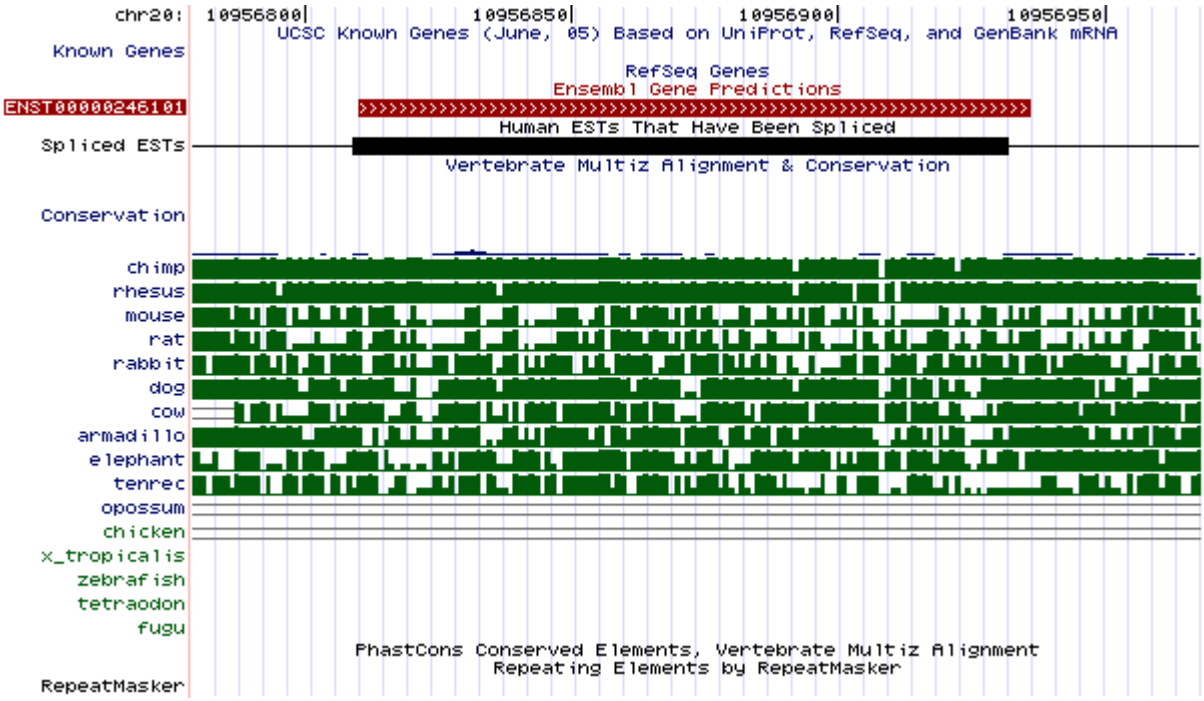
ENSG00000152931
chr5:59,819,516-59,822,848
human genome assembly NCBI36



ENSG00000174613
chr11:2,847,839-2,849,909
human genome assembly NCBI36



ENSG00000125899
chr20:10,956,780-10,956,968
human genome assembly NCBI36



GAPDH (mammalian conserved gene)
chr12:6,513,918-6,517,797
human genome assembly NCBI36