Supplementary material for Stochastic gain and loss of novel transcribed open reading frames in the human lineage

Daniel Dowling, Johnathan Schmitz, and Erich Bornberg-Bauer

September 2, 2020

Contents

1 Data sets used **2**
1.1 RNAseq, Riboseq, and Mass spectrometry data used 2
1.2 Genome and GFF data used 3
1.3 Total read counts for transcriptomic data sets used 3

2 ORFs identified **5**

3 Statistical tests **5**
3.1 ORF sequence properties 5
3.2 Selection analysis 6

4 Repetitive elements found in human ORFs **7**
4.1 Repetitive elements found in human ORFs 7

5 Primate specific ORFs with proteomics evidence **8**
5.1 Great apes with mass spectrometry evidence for translation . . . 8

6 Re-analysis of sequences **9**

7 GC-content of transcribed ORFs **12**

8 Sequence identity of transcribed ORF homologs **13**

9 Expression of novel ORFs **14**

10 Analyses of new sequence properties **17**

1 Data sets used

1.1 RNAseq, Riboseq, and Mass spectrometry data used

Table S1: Transcriptomic and proteomic data sets used

species	datatype	identifiers	reference
Homo sapiens	RNAseq	SRR306840, SRX081984, SRR306845, SRX081989, SRR306848, SRX081991, SRR306852, SRR306854, SRR306855, SRR306840, SRX081984, SRR306845 SRX081989 SRR306848 SRX081991, SRR306852, SRR306854, SRR306855, SRR306858	[1]
Pan troglodytes	RNAseq	SRR306812, SRR306818, SRR306824, SRR306825, SRR306822, SRR306820	[1]
Pan paniscus	RNAseq	RR306828, SRR306830, SRR306836, SRR306834, SRR306832, SRR306837	[1]
Gorilla gorilla	RNAseq	RR306801, SRR306803, SRR306809, SRR306810, SRR306805, SRR306807	[1]
Pongo pygmaeus	RNAseq	SRR306792, SRR306793, SRR306795, SRR306799, SRR306797	[1]
Macaca mulatta	RNAseq	SRR306778, SRR306781, SRR306783, SRR306785, SRR306787, SRR306788	[1]
Mus musculus	RNAseq	RR306761, SRR306760, SRR306759, SRR306758, SRR306764,SRR306767, SRR306770, SRR306773, SRR306775, SRR306789	[1]
Homo sapiens	Riboseq	Battle15_All_ribocov_track	[2]
Homo sapiens	RNAseq	Battle15_All_mrnacov_track	[2]
Homo sapiens	Mass-spectrometry	PXD011817	[3]
Homo sapiens	Mass-spectrometry	PXD001572	[4]

1.2 Genome and GFF data used

Table S2: Genome and GFF files used for assembling transcriptomes

species	data type	source
Homo sapiens	GFF	GCF_000001405.37_GRCh38.p11_genomic.gff.gz
Homo sapiens	genome	GCF_000001405.37_GRCh38.p11_genomic.fna.gz
Pan troglodytes	GFF	GCF_000001515.7_Pan_tro_3.0_genomic.gff.gz
Pan troglodytes	genome	GCF_000001515.7_Pan_tro_3.0_genomic.fna.gz
Pan paniscus	GFF	GCF_000258655.2_panpan1.1_genomic.gff.gz
Pan paniscus	genome	GCF_000258655.2_panpan1.1_genomic.fna.gz
Gorilla gorilla	GFF	GCF_000151905.2_gorGor4_genomic.gff.gz
Gorilla gorilla	genome	GCF_000151905.2_gorGor4_genomic.fna.gz
Pongo pygmaeus	GFF	GCF_000001545.4_P_pygmaeus_2.0.2_genomic.gff.gz
Pongo pygmaeus	genome	GCF_000001545.4_P_pygmaeus_2.0.2_genomic.fna.gz
Macaca mulatta	GFF	GCF_000772875.2_Mmul_8.0.1_genomic.gff.gz
Macaca mulatta	genome	GCF_000772875.2_Mmul_8.0.1_genomic.fna.gz
Mus musculus	GFF	GCF_000001635.26_GRCm38.p6_genomic.gff.gz
Mus musculus	genome	GCF_000001635.26_GRCm38.p6_genomic.fna.gz

1.3 Total read counts for transcriptomic data sets used

Table S3: Read counts for transcriptomes

species	brain	cerebellum	testes	heart	liver	kidney
Homo sapiens	44795322	37635232	33068033	30936565	32922167	21926014
Pan troglodytes	27273899	19900399	26364419	45840208	18043692	36411794
Pan paniscus	37206658	32941383	15536191	24925667	19075280	25534408
Gorilla gorilla	32611328	19090764	21236012	32269972	36337882	29542949
Pongo pygmaeus	33527031	20085444	-	32239994	35506723	27472342
Macaca mulatta	18281486	19055736	21720133	20137441	31782210	22145222
Mus musculus	65943582	20799622	16660610	26376762	16274055	23768384

Table S4: Mapped read counts for transcriptomes

species	brain	cerebellum	testes	heart	liver	kidney	total
Homo sapiens	28578072	28956571	25575196	23373069	26593621	16642917	129719446
Pan troglodytes	23229971	15319772	18484822	32599034	14053939	31171210	134858748
Pan paniscus	18763106	15485844	8107037	8025950	8855145	10770673	70007755
Gorilla gorilla	28072348	14933934	16194892	26312358	30352140	24568505	140434177
Pongo pygmaeus	24452145	14561503	-	19122614	26837837	20188269	105162368
Macaca mulatta	13288615	14888764	14629340	14711442	24266373	13802011	95586545
Mus musculus	55795533	14036330	13930205	19342816	13646773	17955868	134707525

2 ORFs identified

ORFs were assigned one of eight annotation statuses based on their proximity to annotated genes. These are 0 (intergenic ORF), 1 (close to gene), 2 (close to gene on opposite strand), 3 (intonic ORF), 4 (intronic ORF on opposite strand), 5 overlap with exon (out of frame), 6 overlap with exon (out of form and on opposite strand), and 7 overlap with exon (in frame).

Table S5: Number of ORFs identified using BLASTp at each conservation class and annotation status

		annotation status							
Conservation level	total	0	1	2	3	4	5	6	7
0	4429	1512	226	229	1158	1225	43	15	19
1	1048	312	67	78	246	299	25	11	10
2	1283	412	87	97	294	333	30	16	14
3	1559	502	99	127	366	371	52	22	20
4	4836	1491	300	327	1017	1184	309	88	120
5	16596	872	98	95	419	425	627	600	13460

3 Statistical tests

3.1 ORF sequence properties

Table S6: Statistical tests used. In each test the youngest ORF (conservation class 0) were compared to the oldest ORFs (conservation class 5)

	Mann Whitney U-test		Cohen's D
Features	W	P	D
ISD intergenic	755120	3.145e-09	0.2429161
ISD intronic	261730	0.01665	0.1270498
ISD exon overlap	10464000	0.0001934	0.2519427
Aggregation intergenic	637770	0.1782	0.06616157
Aggregation intronic	229680	0.0993	0.00485418
Aggregation exon overlap	9603500	0.1294	0.3158485
ORF length intergenic	1089400	2.2e-16	1.020863
ORF length intronic	393190	2.2e-16	1.202443
ORF length exon overlap	19119000	2.2e-16	0.9654509
GC-content intergenic	690610	0.05245	0.1166842
GC-content intronic	247350	0.5518	0.05137238
GC-content exon overlap	12790000	2.2e-16	0.5693152

3.2 Selection analysis

Table S7: Results of chi square tests to determine if the proportion of sequences below omega threshold differed between different conservation levels.

conservation	n	omega	df	chi square	Yates correction	p value
1 and 5	15452	<0.5	1	848.3271	846.2846	<0.00001
1 and 5	15452	<0.25	1	432.8146	431.4745	<0.00001
2 and 5	15052	<0.5	1	492.8775	490.8951	<0.00001
2 and 5	15052	<0.25	1	297.7561	296.3583	<0.00001
3 and 5	15319	<0.5	1	666.9086	664.9682	<0.00001
3 and 5	15319	<0.25	1	410.2197	408.8289	<0.00001
4 and 5	18577	<0.5	1	2529.5564	2527.7201	<0.00001
4 and 5	18577	<0.25	1	1960.6516	1959.0624	<0.00001

4 Repetitive elements found in human ORFs

4.1 Repetitive elements found in human ORFs

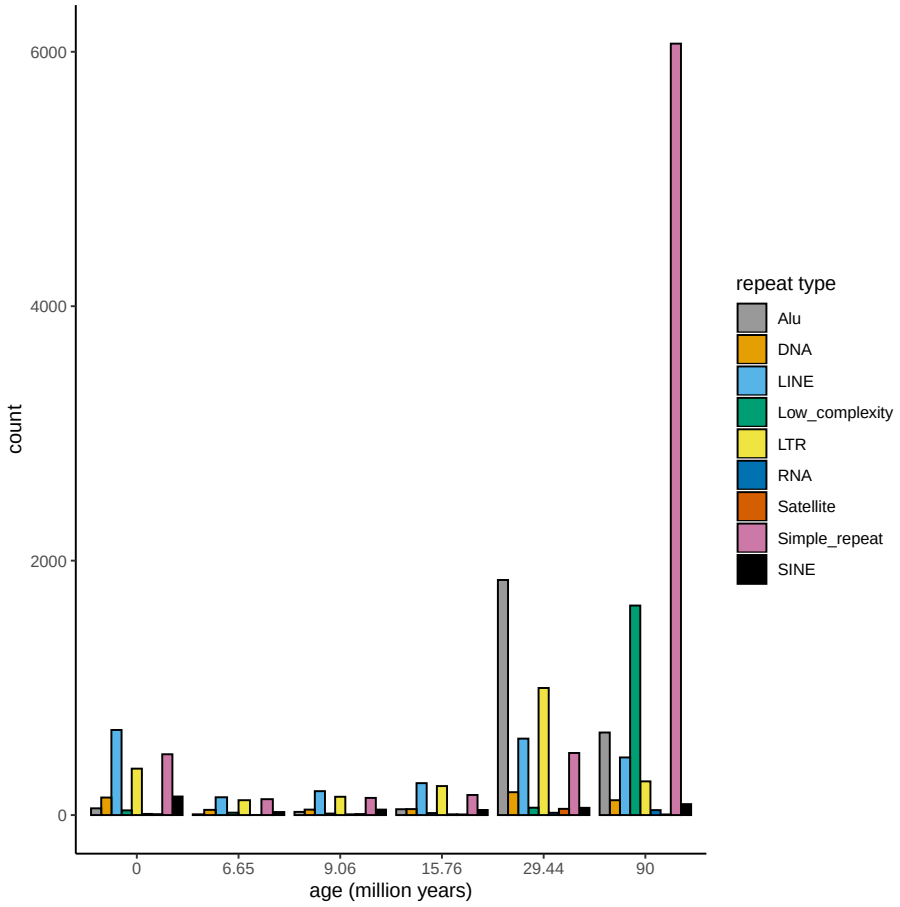

Figure S1: Repetitive elements identified in human ORFs. Age corresponds to the age of the ORF based on identifying homologous sequences transcribed in other primate taxa. SINE refers to non alu-element SINEs.

5 Primate specific ORFs with proteomics evidence

5.1 Great apes with mass spectrometry evidence for translation

Table S8: great-ape specific ORFs with mass spectrometry evidence

ORF ID	Minimum Age (MY)	reference
MSTRG.26607.1_2	0.00	[5]
MSTRG.26607.2_2	0.00	[5]
MSTRG.37003.1_1	0.00	[5]
MSTRG.9310.5_2	6.65	[5]
MSTRG.14698.17_2	9.06	[5]
MSTRG.19945.1_5	9.06	[5]
rna106715_4	9.06	[5]
MSTRG.17291.1_2	15.76	[5]
MSTRG.18707.1_1	15.76	[5]
MSTRG.19700.1_2	15.76	[5]
MSTRG.20318.13_1	15.76	[5]
MSTRG.29103.4_17	15.76	[5]
MSTRG.29865.1_1	15.76	[5]
MSTRG.31316.11_7	15.76	[5]
MSTRG.32504.1_2	15.76	[5]
MSTRG.33748.5_15	15.76	[5]
MSTRG.33754.1_3	15.76	[5]
MSTRG.39770.1_2	15.76	[5]
MSTRG.5.4_15	15.76	[5]
rna77466_7	15.76	[5]
MSTRG.20868.1_2	0.00	[3]
MSTRG.28027.1_9	0.00	[3]
MSTRG.32964.10_3	0.00	[3]
MSTRG.38646.1_9	0.00	[3]
MSTRG.39718.1_10	15.76	[3]

6 Re-analysis of sequences

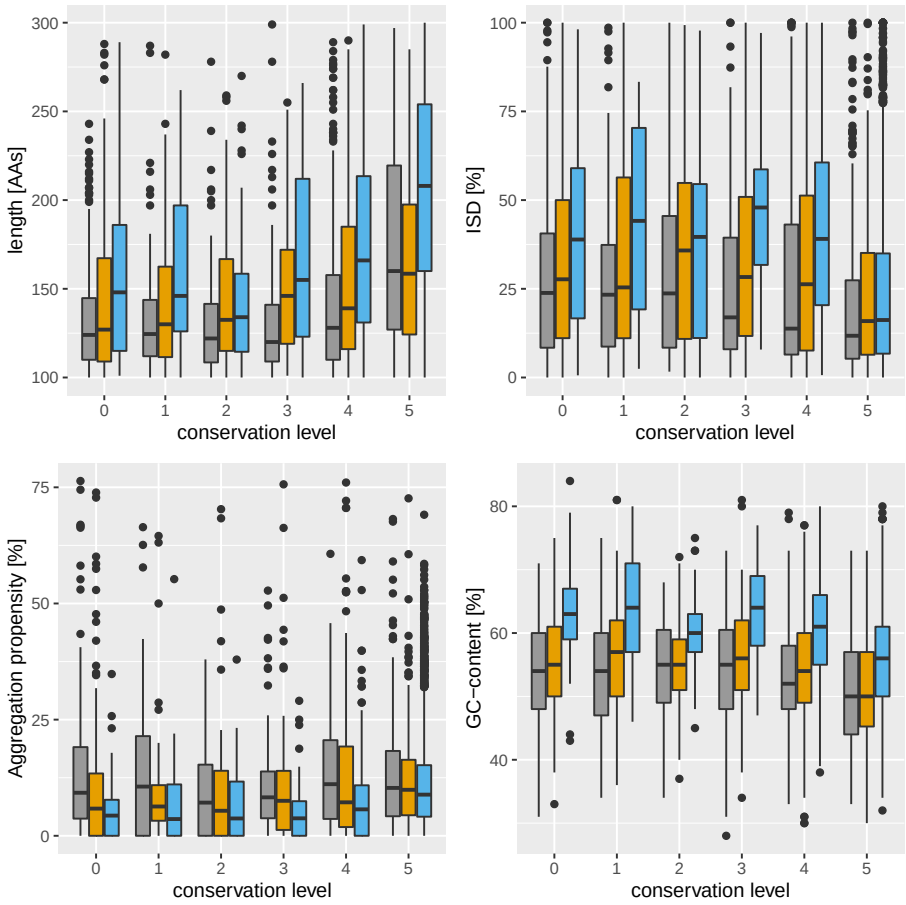

Figure S2: **Sequence properties of transcribed human ORFs.** Only ORFs with lengths between 100 and 300 amino-acids were used in this plot. a.) ORF length. b.) Intrinsic structural disorder (ISD). c.) Aggregation propensity. d.) GC-content. Coloured boxes show ORFs which overlap with intergenic regions (grey), introns (yellow), and exon overlap (blue).

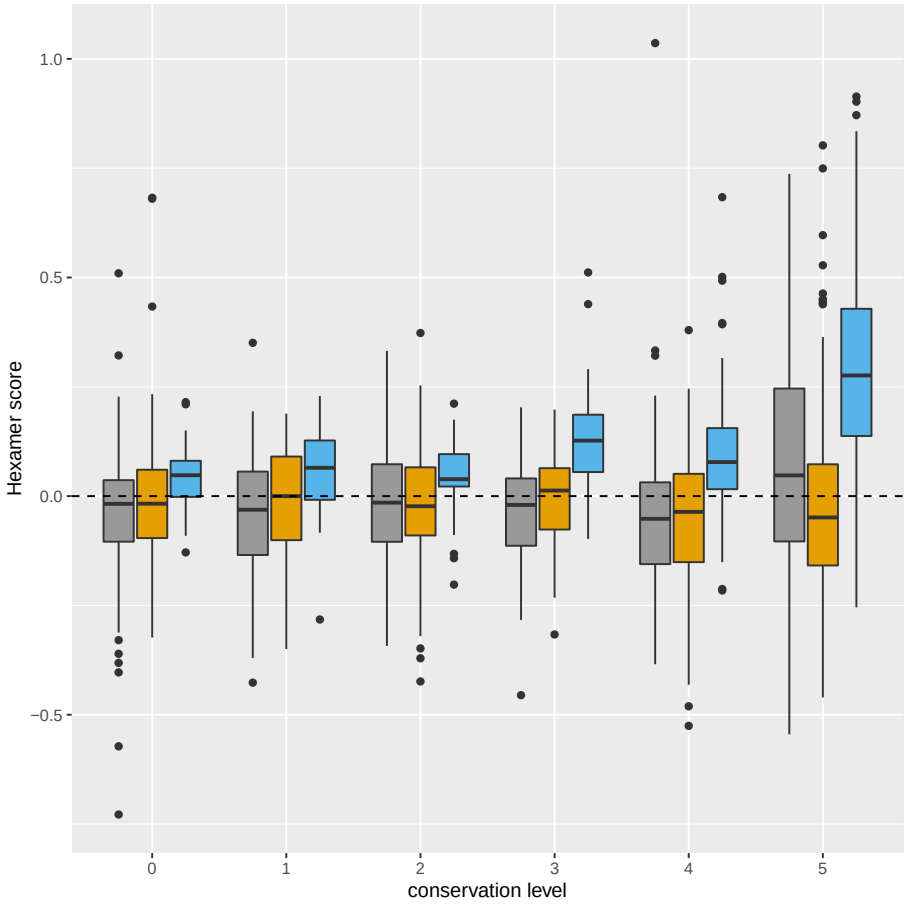

Figure S3: **Hexamer score of transcribed ORFs.** Only ORFs with lengths between 100 and 300 amino-acids were used in this plot. Grey boxed show ORFs overlap intergenic regions, yellow boxes overlap introns, and blue boxes overlap exons. Values greater than 0 (marked with a dotted line) indicate coding sequences, values below 0 indicate non-coding.

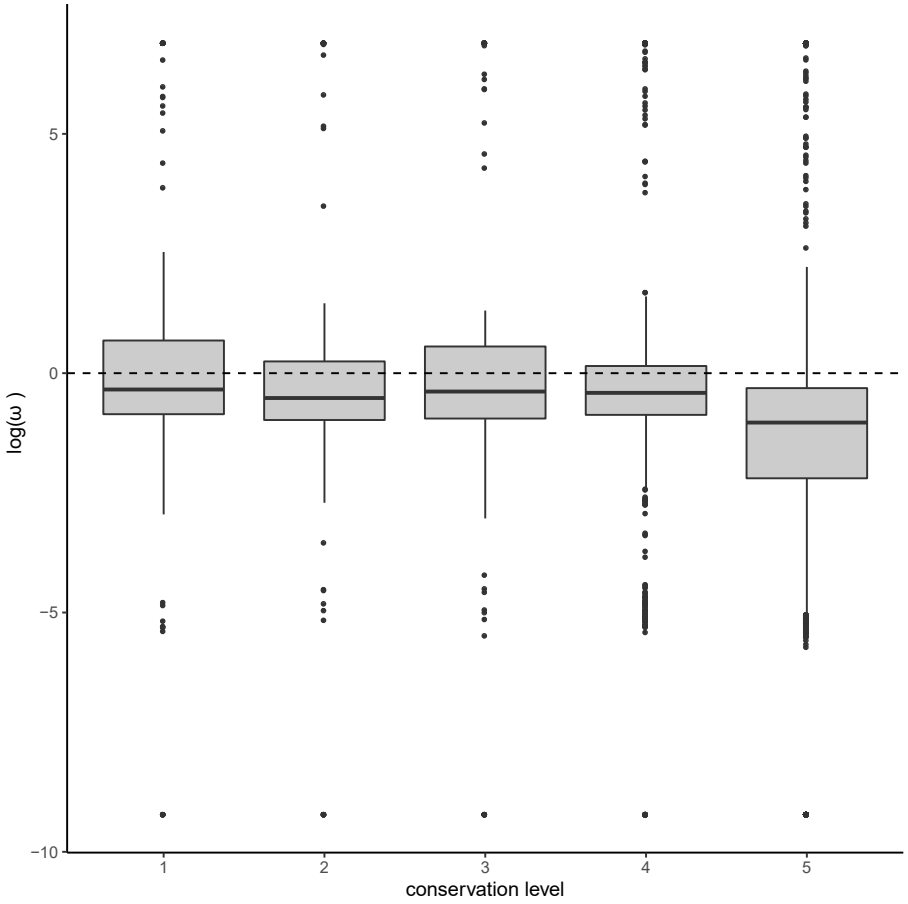

Figure S4: **dN/dS (ω) values of transcribed ORFs.** Only ORFs with lengths between 100 and 300 amino-acids were used in this plot. Dotted line indicates omega of 1, values below 1 are likely to evolving under purifying selection.

7 GC-content of transcribed ORFs

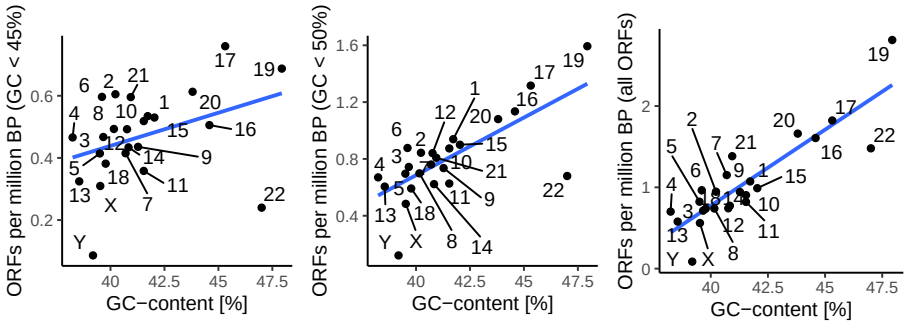

Figure S5: **Correlation between number of ORFs per chromosome and chromosome GC content.** From left to right, ORFs below 45% GC, below 50%, all ORFs.

8 Sequence identity of transcribed ORF homologs

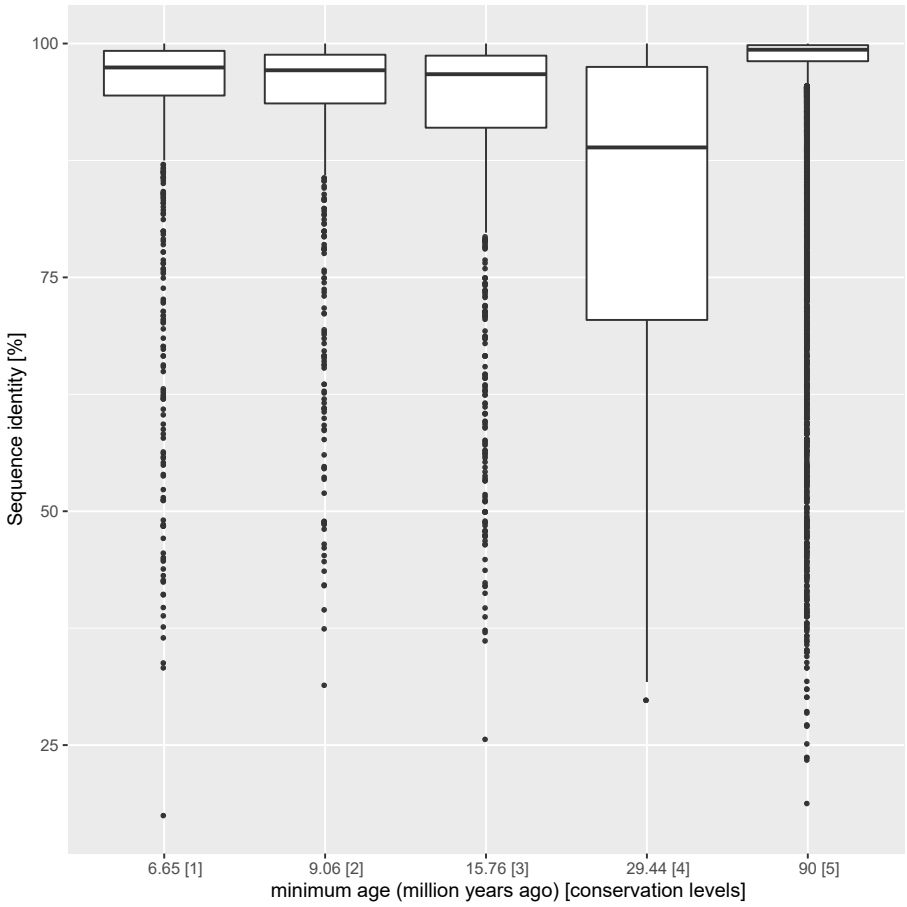

Figure S6: **Sequence identity [%] between human and chimpanzee homologs assigned to different age categories.** Conservation level shown in square brackets.

9 Expression of novel ORFs

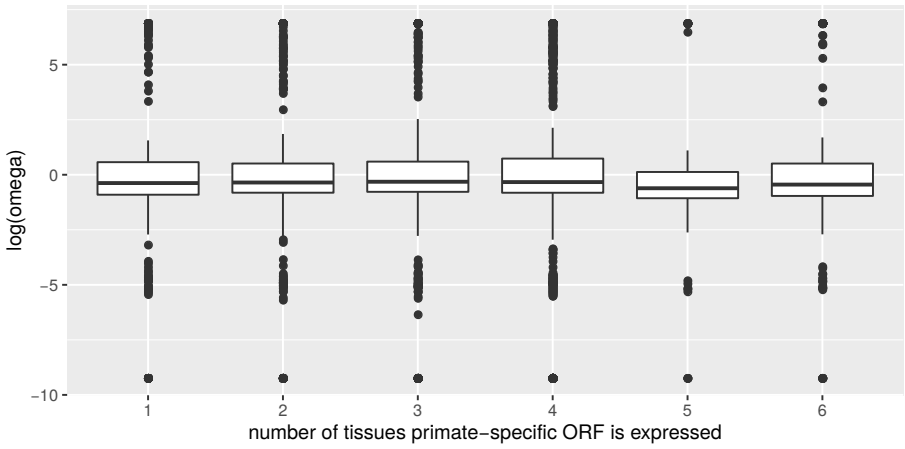

Figure S7: **Tissue expression of transcribed ORFs.** The number of different tissues ORFs of different ages are expressed in. Bottom: dN/dS values (omega) of expressed primate-specific ORFs is not correlated with the number of tissues in which the are expressed.

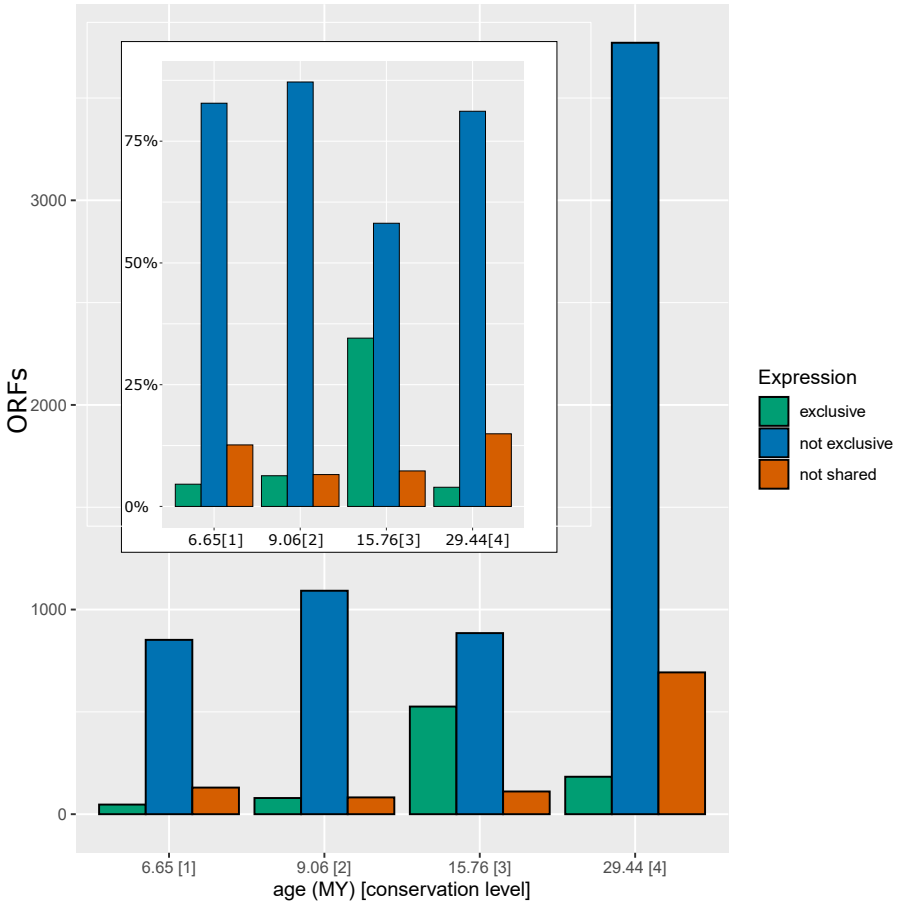

Figure S8: **Expression of human ORFs and homologous ORFs .** Expression of ORFs was compared between human and other primates in six tissues (brain, cerebellum, liver, heart, testes, and kidney). ORFs are categorized into three groups. 1) ORFs which are expressed in the exact same tissues in human and other taxa (exclusive, green bar), for example homologs expressed in brain in human and brain in chimpanzee. 2) ORFs which share expression in at least one tissue, but not all tissues (not exclusive, blue bar), for example homologs expressed in brain in human but brain and heart in chimpanzee. 3) ORFs which are expressed in different tissues in human and other taxa (not shared, red bar), for example homologs expressed in brain in human but heart in chimpanzee. The inset shows the proportion of ORFs belonging to each category as a percentage. Each age/conservation levels sums to 100%.

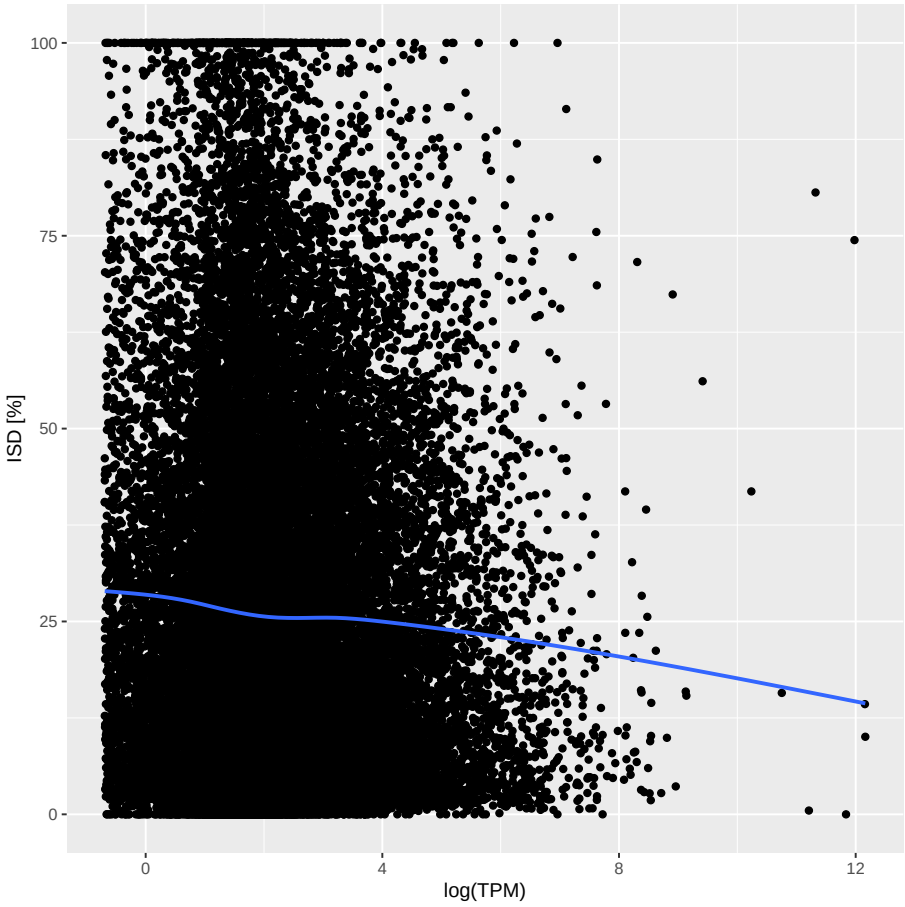

Figure S9: **Intrinsic structural disorder (ISD) and expression level.** ISD of expressed human ORFs does not correlate strongly with expression.

10 Analyses of new sequence properties

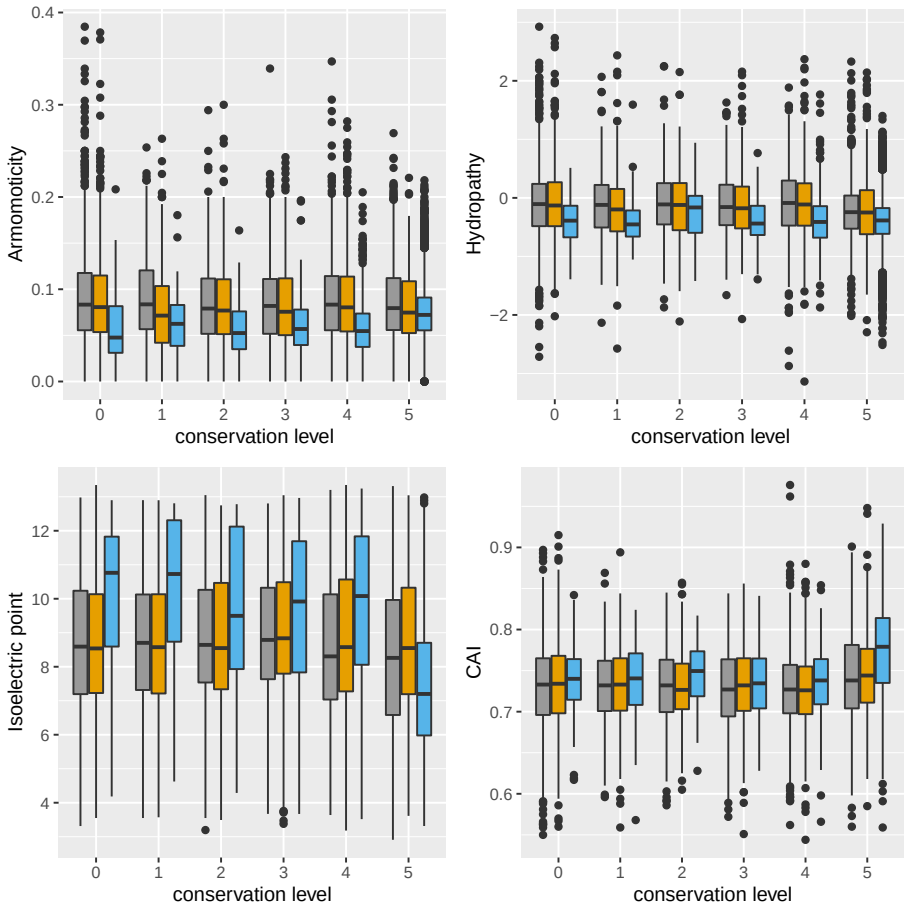

Figure S10: **Sequence properties of transcribed human ORFs** Top left: Proportion of aromatic amino-acids. Top right: Proportion of hydrophilic and hydrophobic amino-acids. Bottom left: Isoelectric point. Bottom right: Codon adaptation index. Box colours correspond to ORFs overlapping intergenic regions (grey), introns (yellow), and CDSs (blue).

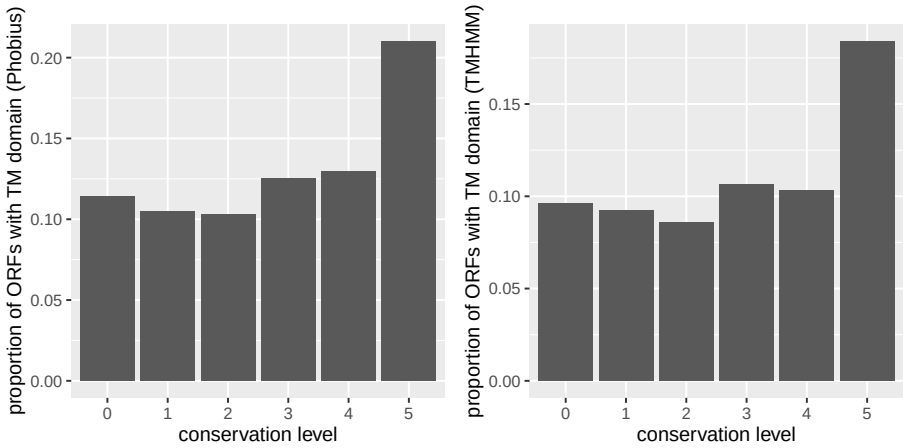

Figure S11: **Transmembrane domains in transcribed human ORFs.** Proportion of ORFs with at least one transmembrane domains predicted by Phobius (left) or TMHMM (right).

References

[1] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, October 2011.

[2] Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, 2015.

[3] Massimo Alfano, Filippo Pederzoli, Irene Locatelli, Silvia Ippolito, Erika Longhi, Pietro Zerbi, Maurizio Ferrari, Andrea Brendolan, Francesco Montorsi, Denise Drago, Annapaola Andolfo, Manuela Nebuloni, and Andrea Salonia. Impaired testicular signaling of vitamin A and vitamin K contributes to the aberrant composition of the extracellular matrix in idiopathic germ cell aplasia. *Fertility and Sterility*, 111(4):687–698, April 2019.

[4] B. A. Diner, T. Li, T. M. Greco, M. S. Crow, J. A. Fuesler, J. Wang, and I. M. Cristea. The functional interactome of PYHIN immune regulators reveals IFIX is a sensor of viral DNA. *Molecular Systems Biology*, 11(2):787–787, February 2015.

[5] Yajing Hao, Lili Zhang, Yiwei Niu, Tanxi Cai, Jianjun Luo, Shunmin He, Bao Zhang, Dejiu Zhang, Yan Qin, Fuquan Yang, and Runsheng Chen. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Briefings in Bioinformatics*, 19(4):636–643, 01 2017.