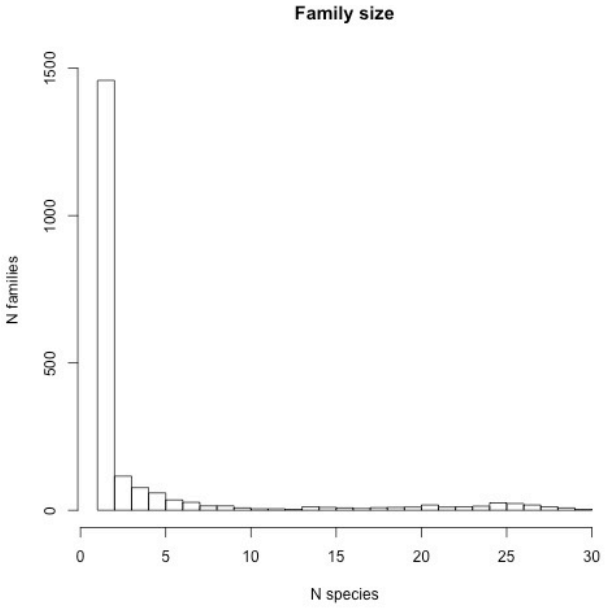
Supplementary Figures. Villanueva-Cañas et al. New genes and functional innovation in mammals.

Figure S1. Increase in family size with the use of RNA-Seq data. Family size: number of species per gene family. **A** Original Data with no RNA-Seq data, only gene annotations. **B.** Data when RNA-Seq data was employed to assemble transcripts. In the latter case we performed additional sequence similarity searches against novel transcripts. These transcripts were obtained by *de novo* transcript assembly of reads from RNA sequencing samples (RNA-Seq). Transcript assemblies were obtained from 30 mammalian species. Only families containing two or more species are considered (2-30 species).

A



B

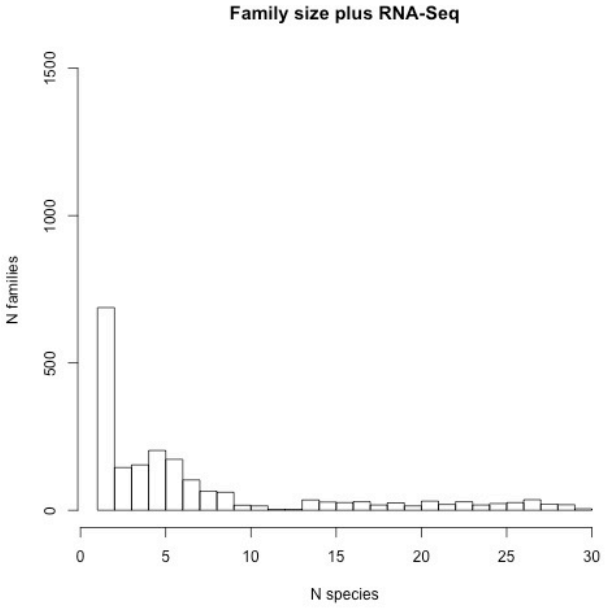


Figure S2. Mammalian species tree including node numbers. The tree depicts the phylogenetic relationships between 30 mammalian species from different major groups. The node number is indicated in each branch. We define three conservation levels: 'mam-basal' (class 2, approximately older than 100 Million years, red), 'mam-young' (class 1, green) and 'species-specific (class 0, blue). The branch length represents the approximate number of substitutions per site as inferred from previous studies (see Methods). The scale bar on the bottom left corner represents 6 substitutions per 100 nucleotides. Dotted lines have been added to some branches to improve readability.

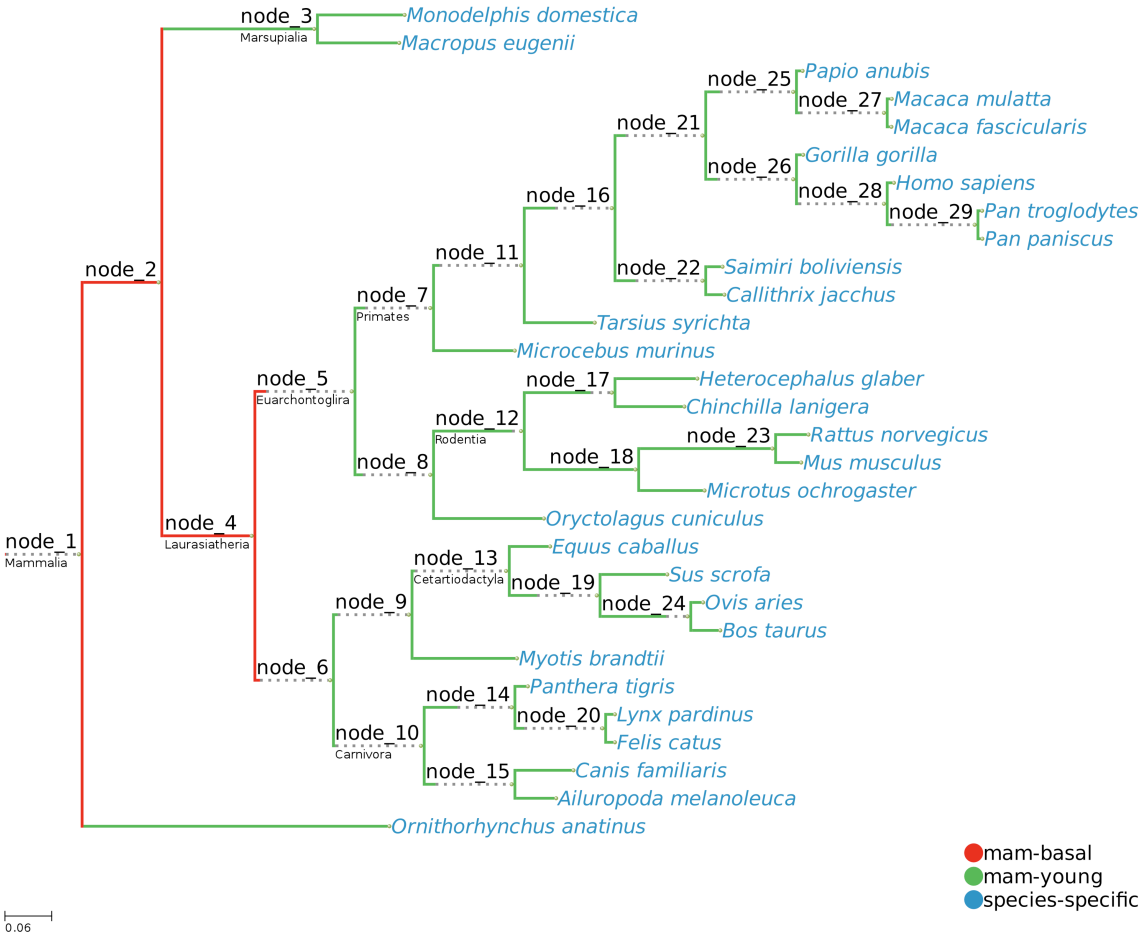
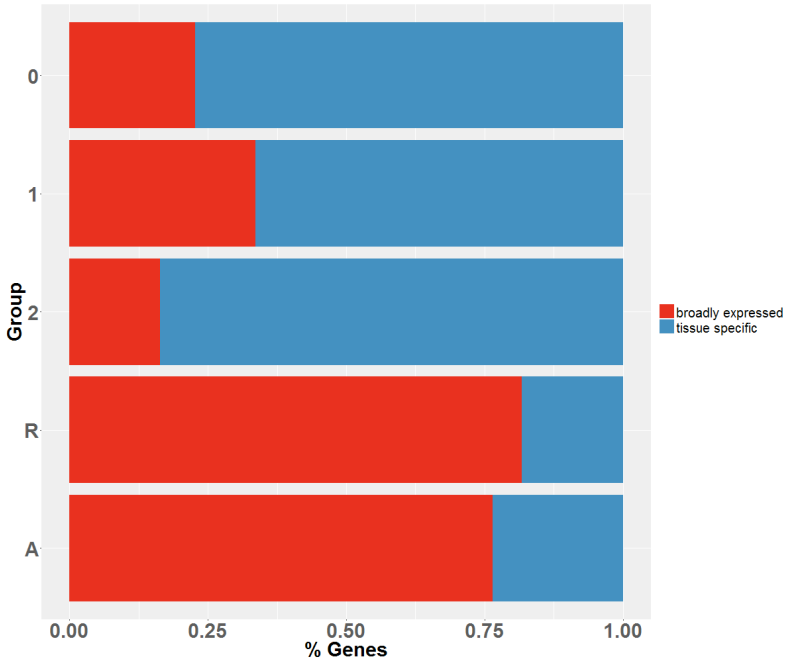
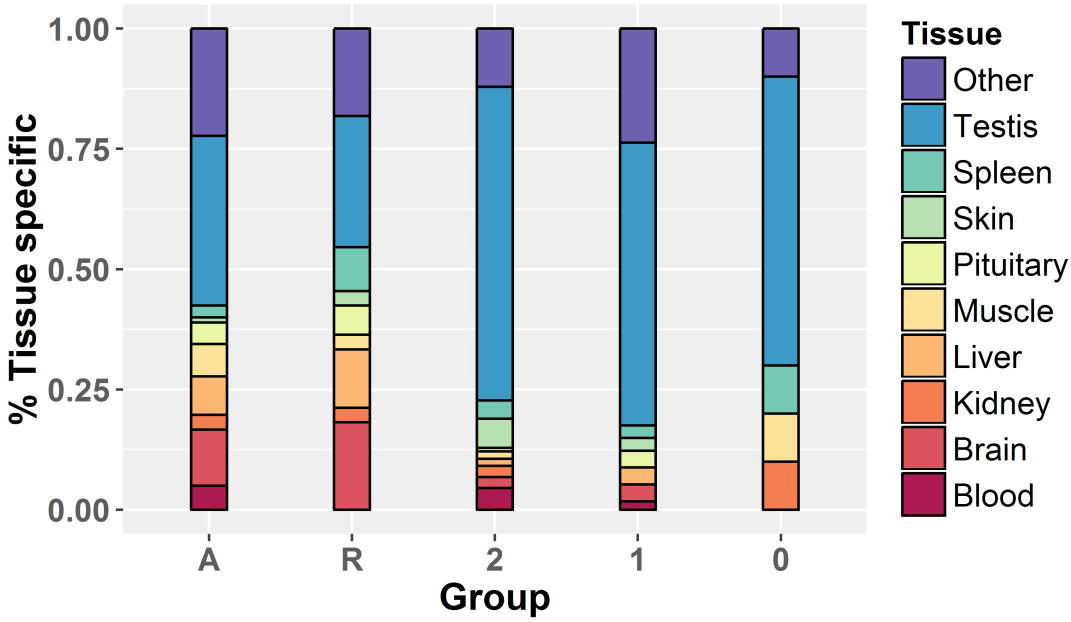


Figure S3. Gene expression patterns of genes from different conservation levels for human genes. Conservation levels A: ancestral; R: random; 2: 'mam-basal'; 1: 'mam-young'; 0: species-specific. **A.** Proportion of broadly-expressed and tissue-specific genes in different conservation classes. **B.** Number of genes with maximum expression in a given tissue for different conservation classes. **C.** Box-plot showing the distribution of FPKM gene expression values, at a logarithmic scale, in different conservation classes. Data in B and C is for tissue-specific genes.

A



B



C

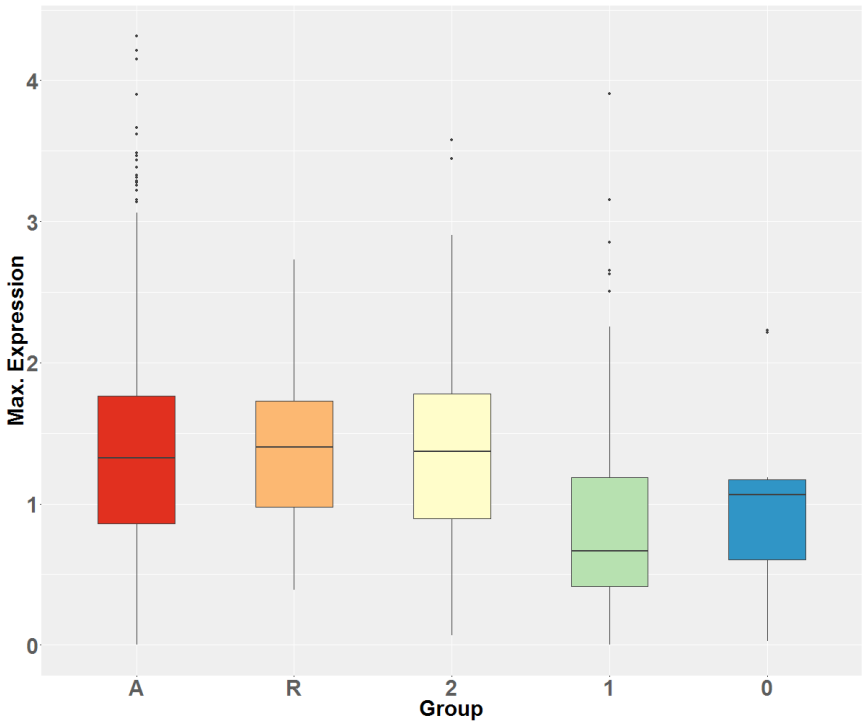
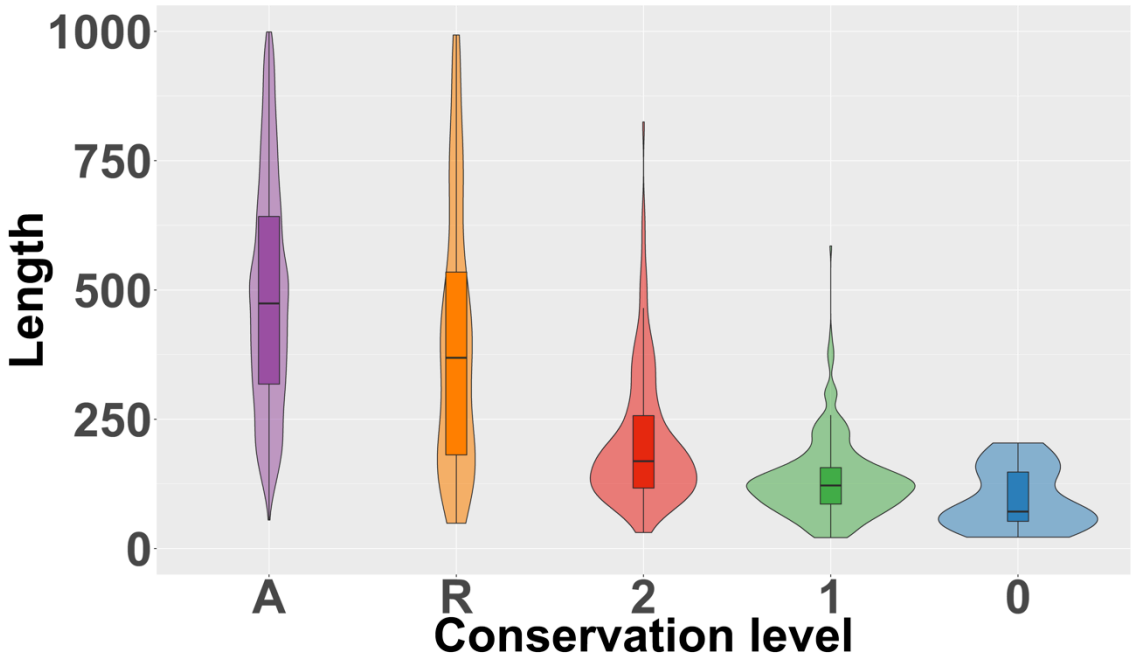


Figure S4. Sequence length in relation to protein conservation level. Distribution of protein sizes (in amino acids) in different groups, shown as violin plots. **A.** Data for human. **B.** Data for mouse. Conservation levels A: ancestral, conserved in 34 non-mammalian species from diverse eukaryotic groups; R: random, non-mammalian-specific random gene dataset; 2: 'mam-basal'; 1: 'mam-young'; 0: species-specific. Mammalian-specific genes (0,1,2) were in general shorter than non-mammalian-specific genes (A,R), in both human and mouse (Wilcoxon test, p-value < 10^{-5}). We also observed that proteins from level 2 tended to be longer than those from level 1, in both human and mosue (Wilcoxon test, p-value < 0.01).

A



B

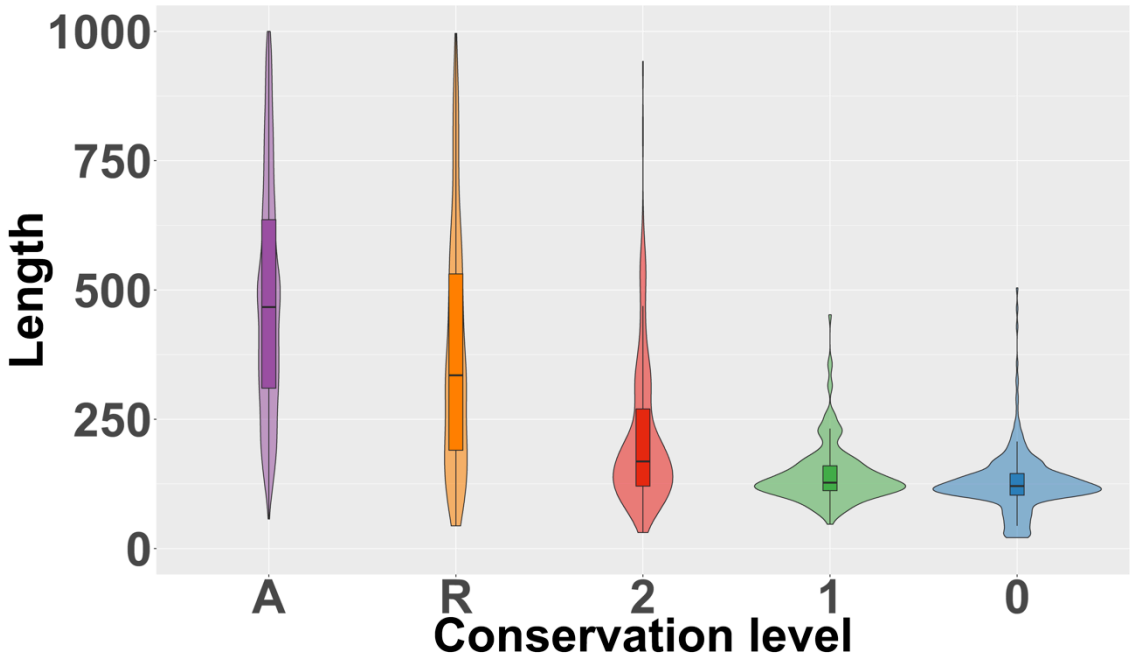
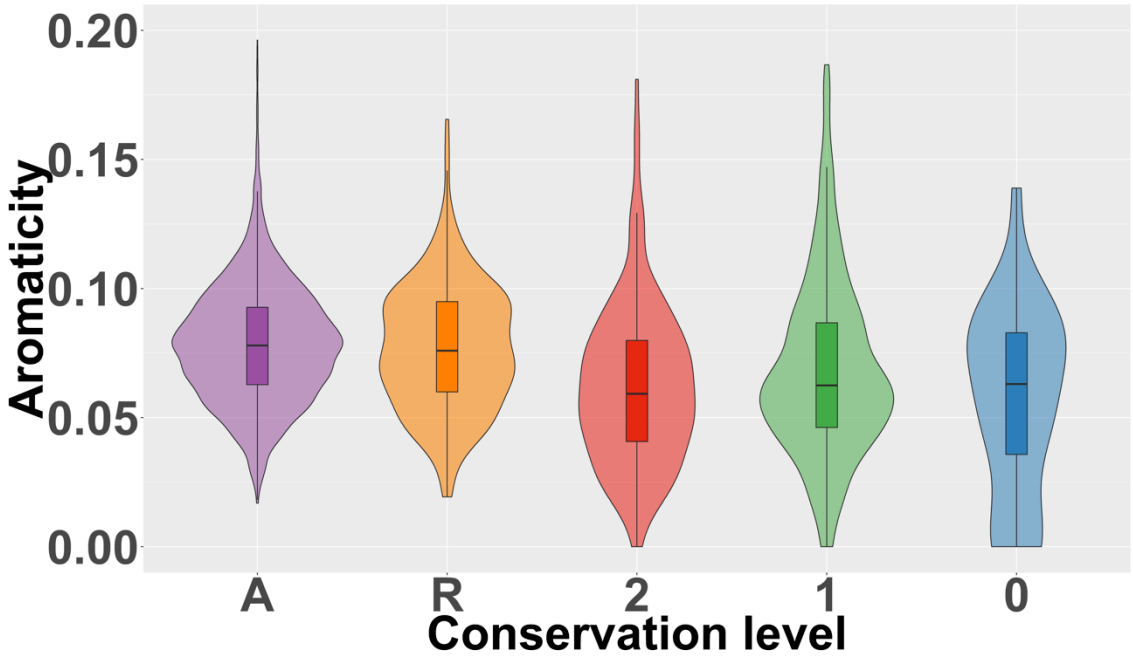


Figure S5. Aromaticity in relation to protein conservation level. Distribution of the values in each groups is is shown as violin plots. **A.** Data four human. **B.** Data for mouse. Conservation levels A: ancestral, conserved in 34 non-mammalian species from diverse eukaryotic groups; R: random, non-mammalian-specific random gene dataset; 2: 'mam-basal'; 1: 'mam-young'; 0: species-specific.

A



B

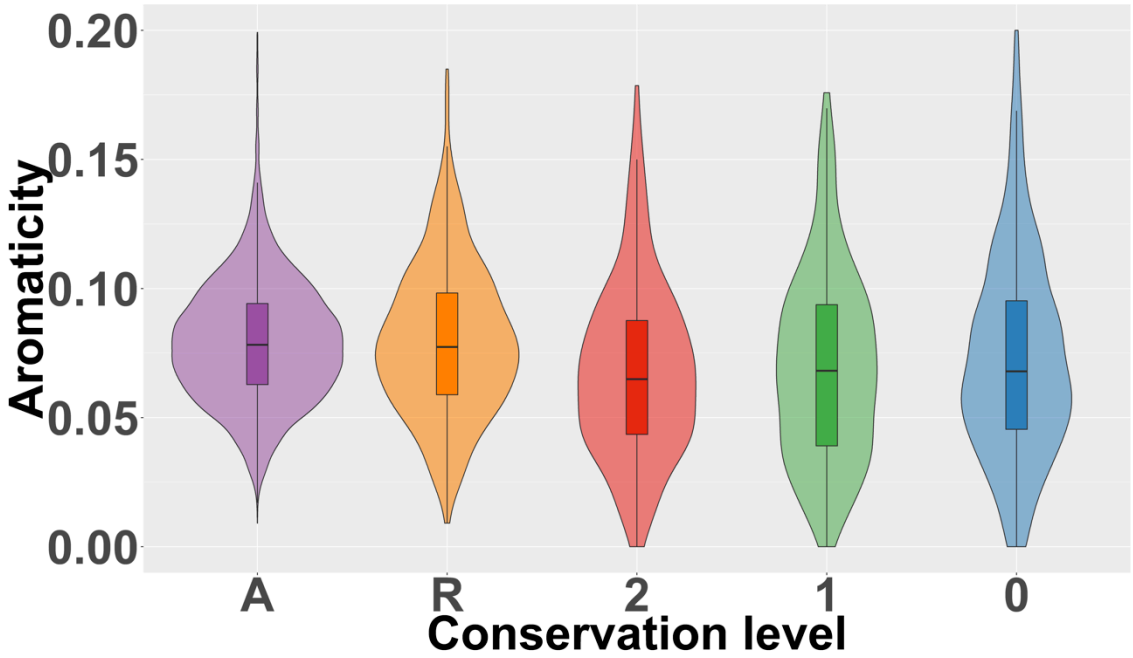
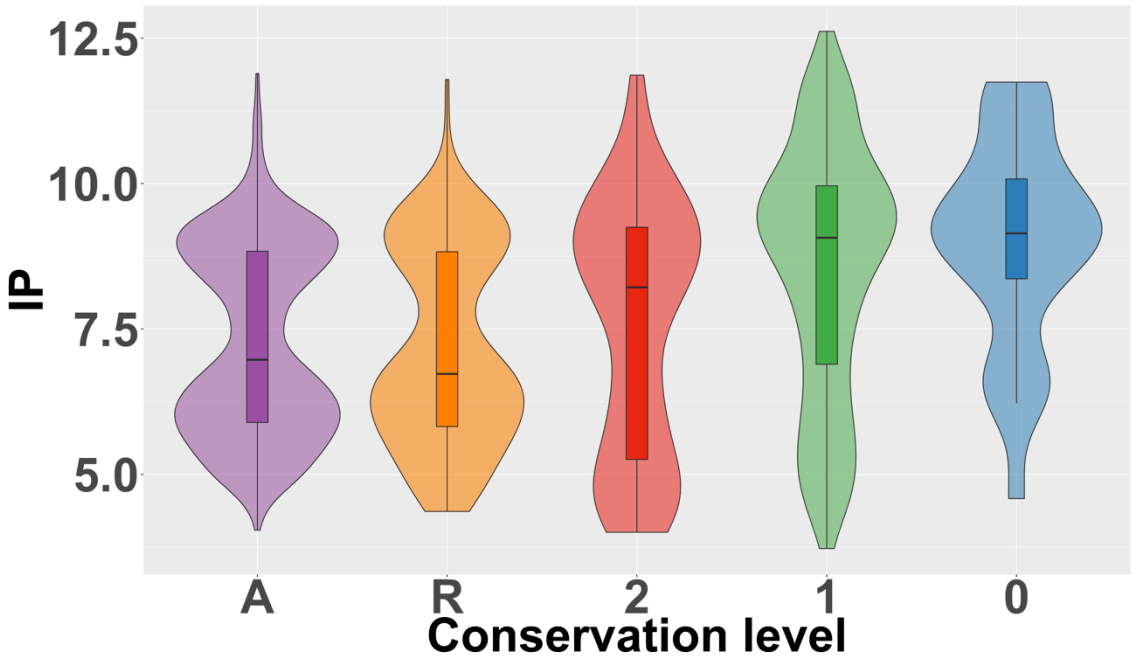


Figure S6. Isoelectric point (IP) in relation to protein conservation level. Distribution of values in each group is shown as violin plots. **A.** Data four human. **B.** Data for mouse. Conservation levels A: ancestral, conserved in 34 non-mammalian species from diverse eukaryotic groups; R: random, non-mammalian-specific random gene dataset; 2: 'mam-basal'; 1: 'mam-young'; 0: species-specific.

A



B

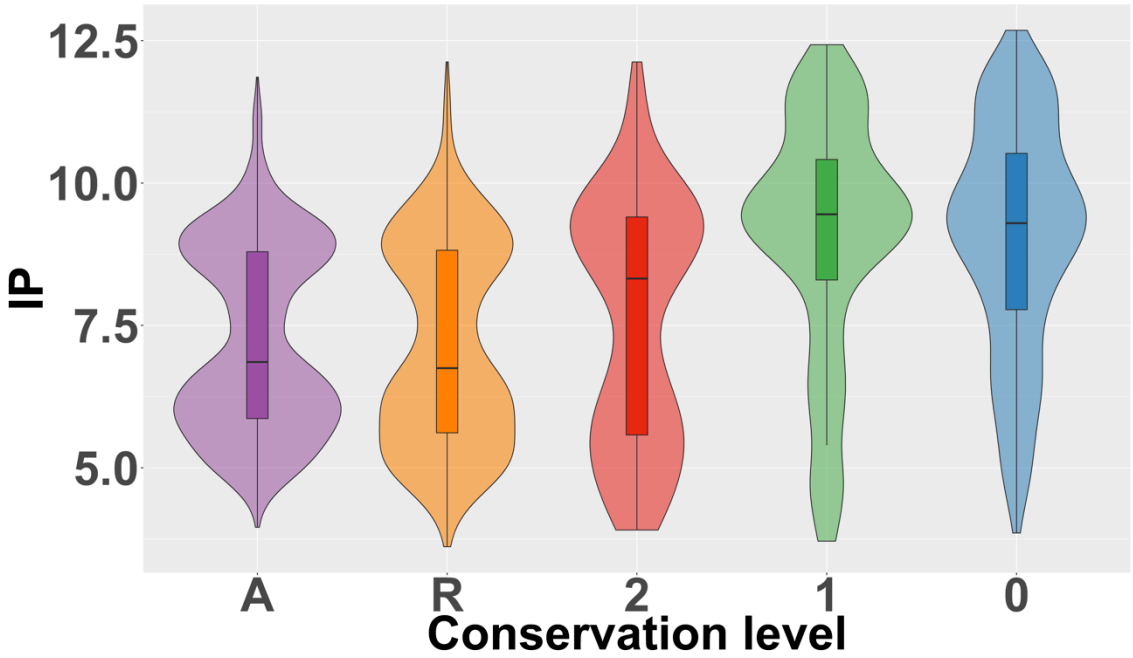


Figure S7. Discovery of a new sequence with antimicrobial activity in MUCIN-7. A. Sequence of MUCIN-7 main isoform (Uniprot ID Q8TAX7). The sequence stretch with antimicrobial activity predicted by AMPA and subsequently validated experimentally is indicated in red. **B.** Minimal inhibitory concentration (MIC, µg/mL) of the antimicrobial stretch against several Gram+ and Gram- bacteria. Under no effect values higher than 128 are expected.

A

>sp|Q8TAX7|MUC7_HUMAN Mucin-7 OS=Homo sapiens GN=MUC7 PE=1 SV=2MKTLPLFVCICALSACFSFSEGRERDHELRHRRHHHQSPKSHFELPHYPGLLAHQKPFIR KSYKCLHKRCRPKLPPSPNNPPKFPNPHQPPKHPDKNSSVVNPTLVATTQIPSVTFPSASTKIT TLPNVTFLPQNATTISSRENVNTSSSVATLAPVNSPAPQDTTAAPPTPSATTPAPPSSSAPPETT AAPPTPSATTQAPPSSSAPPETTAAPPTPPATTPAPPSSSAPPETTAAPPTPSATTPAPLSSSA PPETTAVPPTPSATTLDPSSASAPPETTAAPPTPSATTPAPPSSPAPQETTAAPITTPNSSPTTL APDTSETSAAPTHQTTTSVTTQTTTTKQPTSAPGQNKISRFLLYMKNLLNRIIDDMVEQ

B

	Gram -		Gram +	
	E. coli ATCC 25922	*P. aeruginosa* ATCC 27853	*E. faecalis* ATCC 29212	*S. aureus* ATCC 29213
MIC (µg/mL)	64	128	16	16