Supplementary Materials for Foy et al. “A shift in aggregation avoidance strategy marks a long-term direction to protein evolution”

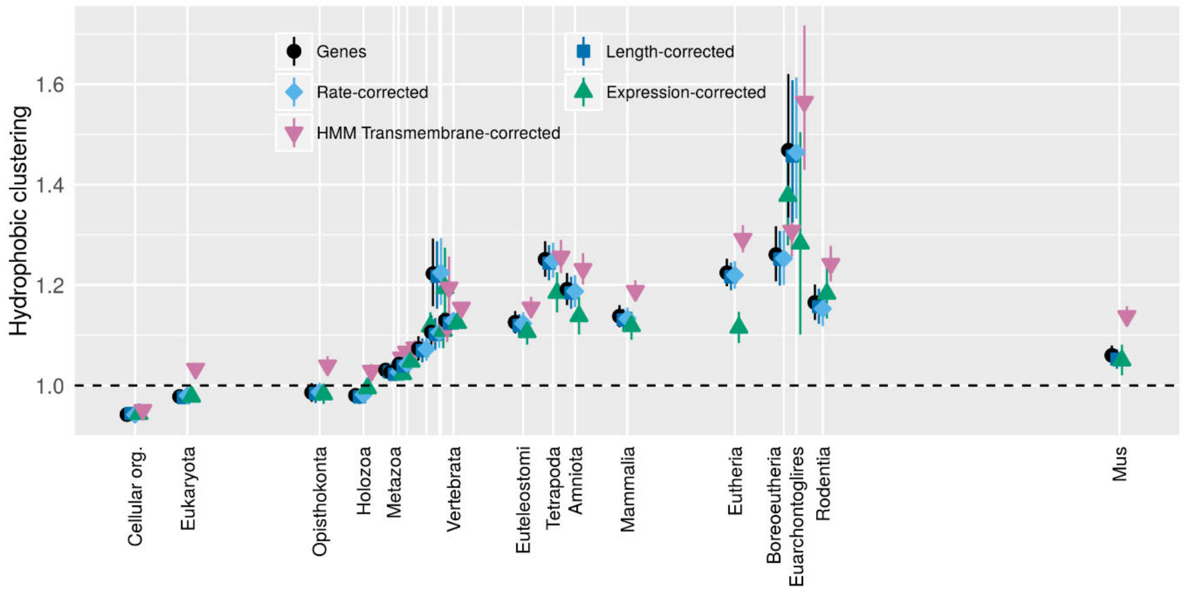

Fig. S1. Hydrophobic clustering values corrected for length, evolutionary rate, expression level, and predicted transmembrane status. Means and standard errors are estimated using mixed-effects linear regression treating phylostratum and transmembrane status as fixed terms, gene family as a random term, and evolutionary rate, length, and expression level as quantitative terms. Evolutionary rates, determined using PAML for a mouse-rat comparison, were downloaded from Ensembl BioMart (Smedley *et al.* 2015) [accessed February 18, 2016]; they are thus unavailable for Mus-specific genes. Expression levels were downloaded from the Protein Abundance Database (PaxDB) (Wang *et al.* 2015) [accessed February 2016]. Length was Box-Cox transformed with λ_1=-0.107 and λ_2=30, rate with λ_1=0.18 and λ_2=0, and expression with λ_1=-0.23 and λ_2=0.15. All three quantitative terms were individually found to be statistically significant in improving linear model fit (likelihood ratio test, p = 0.003, 0.043, and 0.015 for length, rate, and expression respectively); however, each quantitative term has only modest effect on the estimates of the mean hydrophobic clustering. Corrected values use a length of 389 amino acids, an evolutionary rate of 0.146, expression of 1.1 ppm, and probability 0.258 of being transmembrane; the first three are the backtransformed values of the means of the transformed data, the last the frequency. The x-axis is the same as for Figure 2.

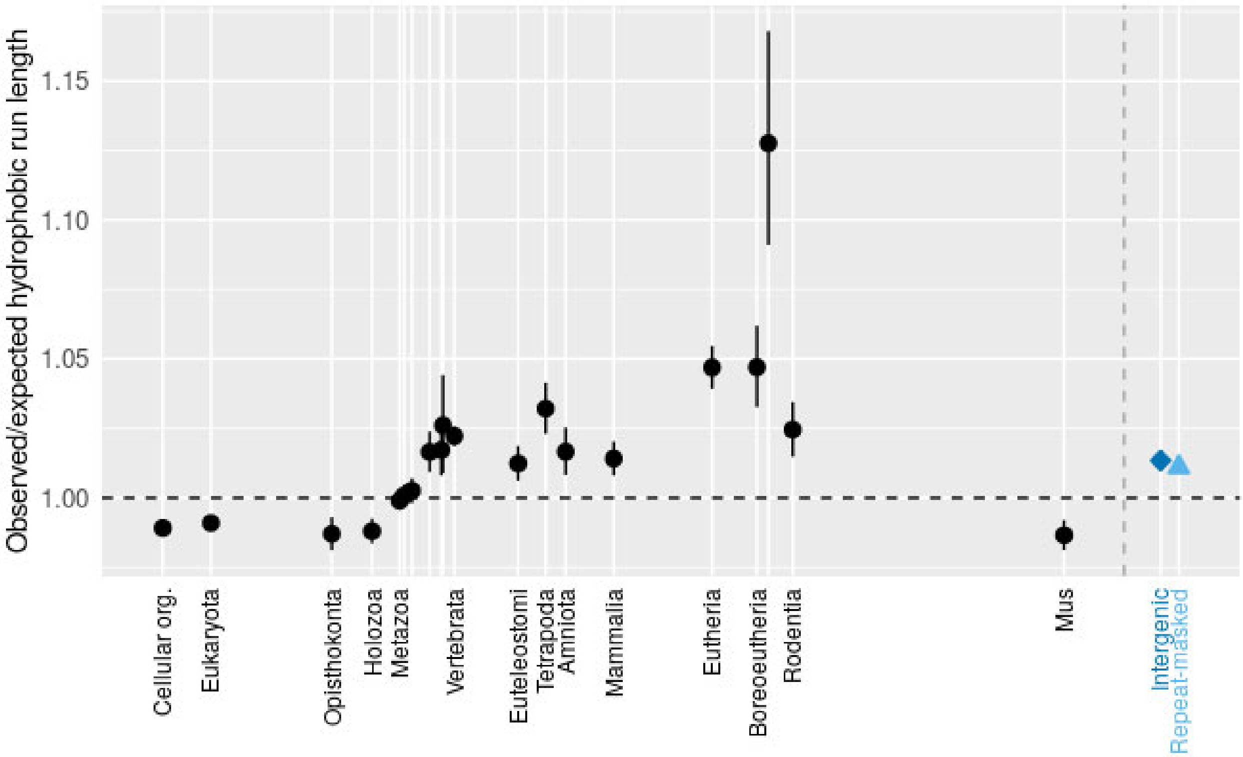

Fig. S2. The mean run length of hydrophobic amino acids behaves similarly to the clustering metric of Figure 5, but with the trend going back less far in time. Observed mean run lengths are normalized by the expected run length for a randomly ordered sequence with the same total number of hydrophobic amino acids. When the probability that a residue is hydrophobic is p, then the probability that a run of hydrophobic amino acids has length k follows a geometric distribution $p(k) = (p)^{1-k}(1-p)$ ($k = 1,2, \ldots$), and therefore the expected value of k is $1/(1-p)$.

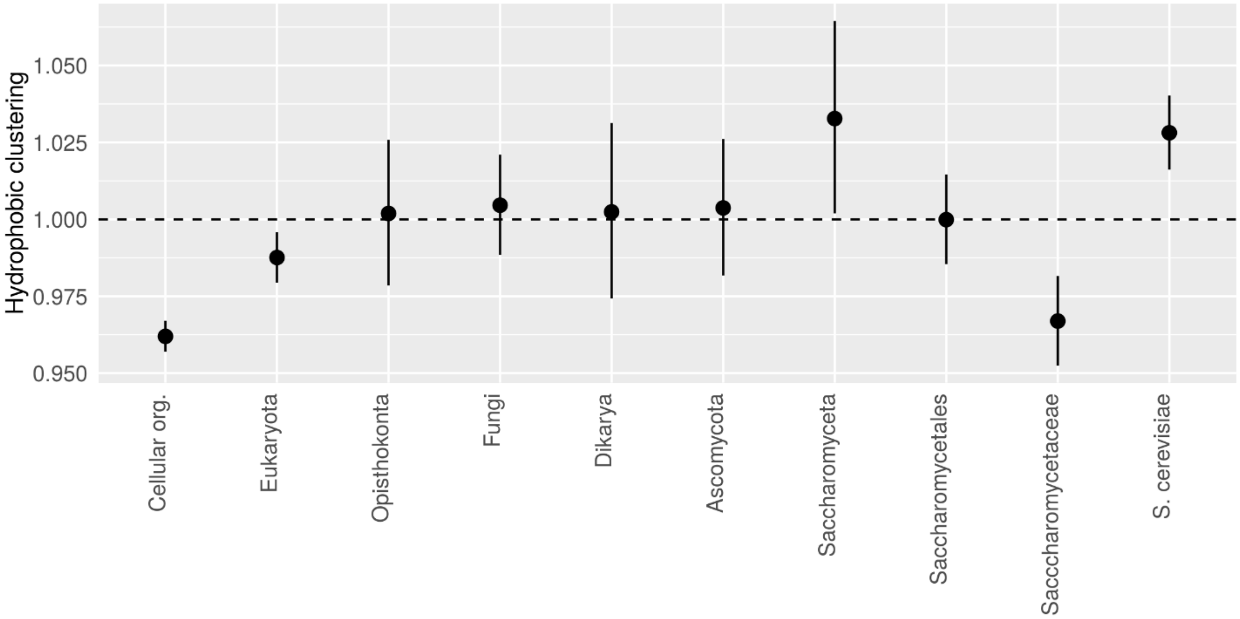

Fig. S3. The trend in clustering proceeds faster for *S. cerevisiae* gene families than for the mouse genes illustrated in Figure 5, and falls to lower clustering values. Back-transformed central tendency estimates +/- one standard error come from a linear mixed model, where gene family and phylostratum are random and fixed terms respectively. Gene family and phylostratum annotations are taken from Wilson et al. (2017).

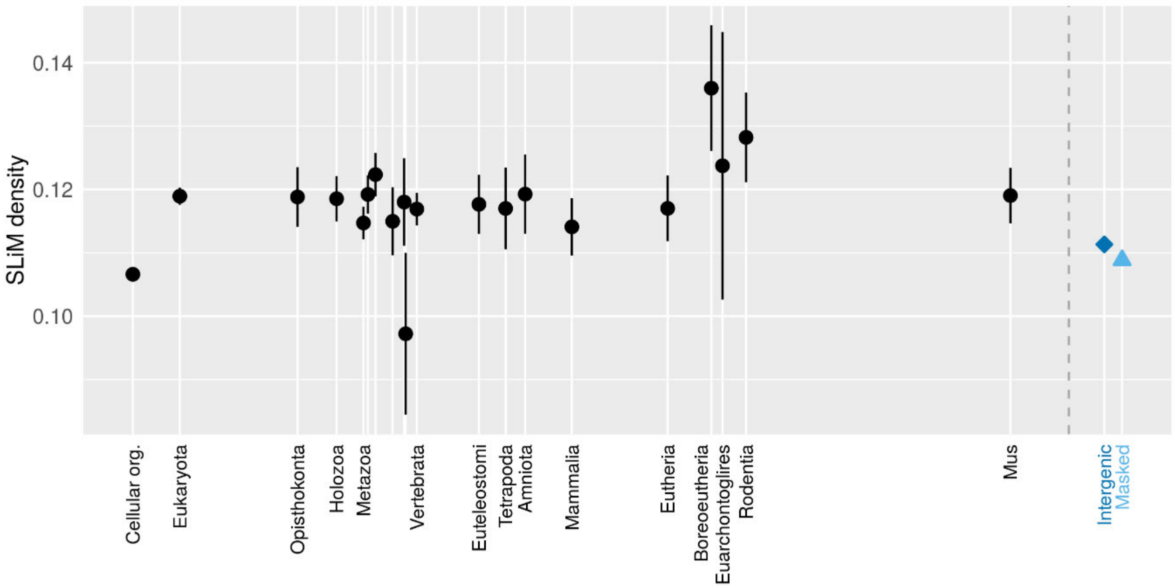

Fig. S4. With the exception of fewer short linear motifs (SLiMs) in genes shared with prokaryotes, SLiM density does not depend significantly on gene age. To calculate SLiM density, we compiled the 254 regular expressions from http://elm.eu.org/elms (Dinkel *et al.* 2016) [accessed October 18, 2016], excluded 13 that clearly did not apply to mouse, and excluded 37 due to having too low an information content, i.e. being too likely to appear by chance alone. This was guided by the heuristic of containing at least 3 high-information sites; a list of the 204 regular expressions used is given in Supplementary Table 1. We then calculated the number of matches in a sequence, divided by protein length. The probability distribution of this SLiM measure, across all proteins, resembled a Gaussian plus a fat right tail. Given the possibility that the Gaussian portion might represent a null expectation while the right tail represents the minority of SLiMs that are functional and under selection to be retained, we did not transform the data. Central tendency estimates +/- one standard error come from a linear mixed model, where gene family and phylostratum are random and fixed terms respectively. The x-axis is the same as for Figure 2.

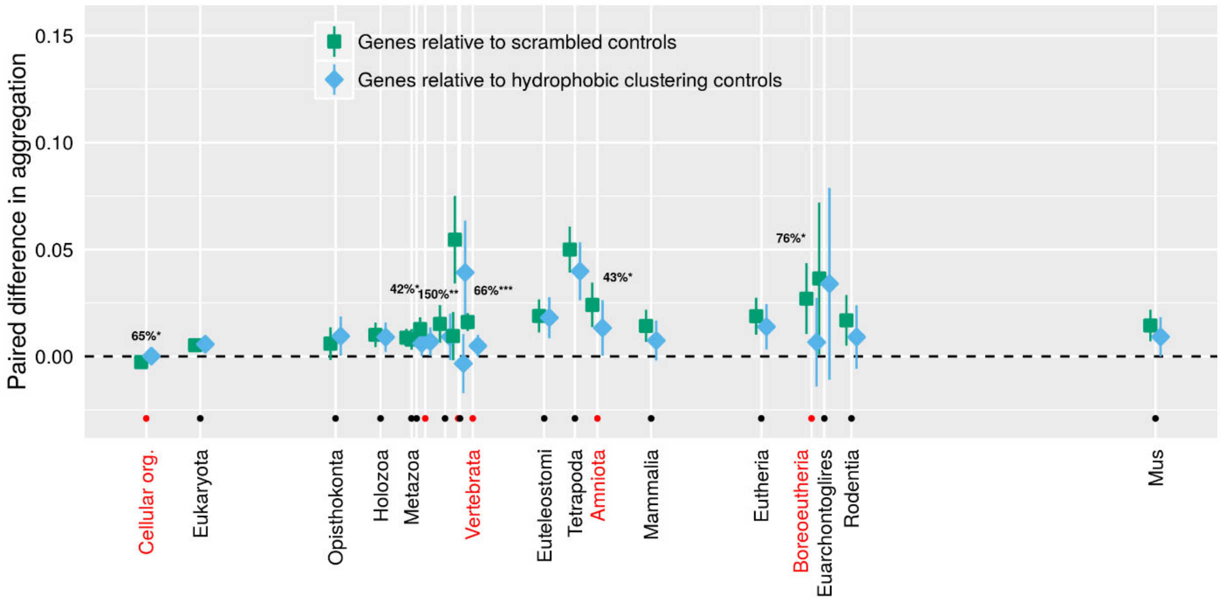

Fig. S5. As with TANGO (Fig. 4), only very old genes have aggregation propensities (as assessed by waltz) lower than that expected from their amino acid composition alone (orange). Controlling for dispersion (blue) explains a smaller proportion of the sequence contribution for young gene waltz scores than it does for TANGO in Fig. 4. 95% confidence intervals are shown, based on a linear mixed model where gene family and phylostratum are random and fixed terms respectively, and significance for the difference between green and blue is shown by red dot and text, with percentage of deviation from 0 given, as described in the Fig. 4 legend. The x-axis is the same as for Figure 2.

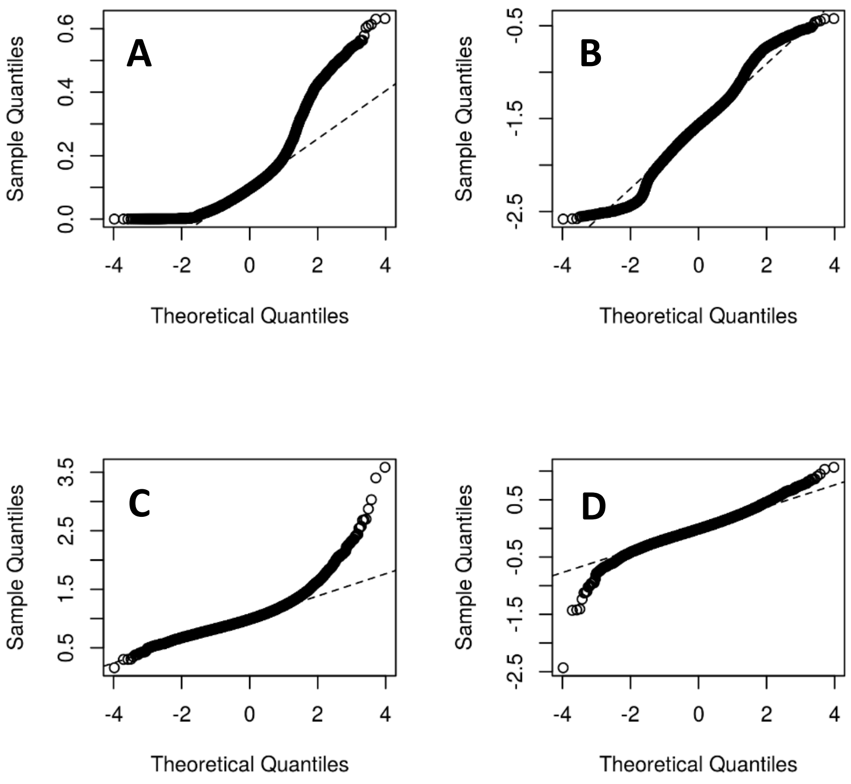

Fig. S6. Quantile-quantile (Q-Q) plots showing that non-normal distributions (A,C are non-linear) become approximately normal (B, D are approximately linear) following Box-Cox transformation. A and B show aggregation scores from TANGO and C and D show clustering. Theoretical quantiles are from a standard normal distribution.

Supplementary Table Legends

Table S1. ELM identifier and Perl regular expression of the high-information SLiM classes from the ELM database (Dinkel *et al.* 2016).

Table S2. *M. musculus* proteins, listing the Ensembl gene identifier, gene family membership as listed in Wilson et al. (2017) or reclassified here, phylostratum either as assigned to the gene family by Wilson et al. (2017) or as re-assigned to *Mus* here, the TANGO and Waltz aggregation propensity scores, normalized index of dispersion, frequency of hydrophobic residues FLIVMW as classified in Irbäck et al. (1996), intrinsic structural disorder, frequency of thermophilic residues as classified in Boussau et al. (2008) and shown in Fig. 3, the number of SLiMs in the protein, transmembrane designation (with "yes" indicating transmembrane status), normalized mean hydrophobic run length, and the protein sequence. The GeneFamilyNumber column identifies members of the same gene family via a shared numeric identifier. As specified in Wilson et al. (2017), the gene family phylostratum number is assigned based upon the estimated de novo origin of the gene family, with the oldest genes being 1, and Mus-specific genes, as classified in this paper via the *M. pahari* data, being 20.

Table S3. Intergenic controls. Nucleotide sequences from intergenic regions of M. musculus genome as described by Wilson et al. (2017), listing the Ensembl gene identifier of the gene to which they were taken in proximity, the TANGO and Waltz aggregation propensity scores, normalized index of dispersion, frequency of hydrophobic residues as classified in Irbäck et al. (1996), intrinsic structural disorder, frequency of thermophilic residues as classified in Boussau et al. (2008), the number of SLiMs in the translated sequence, and the nucleotide sequence.

Table S4. RepeatMasked (Smit *et al.* 2015) intergenic controls. Nucleotide sequences from intergenic regions of the masked M. musculus genome as described by Wilson et al. (2017), listing the Ensembl gene identifier of the gene to which they were taken in proximity, the TANGO and Waltz aggregation propensity scores, normalized index of dispersion, frequency of hydrophobic residues as classified in Irbäck et al. (1996), intrinsic structural disorder, frequency of thermophilic residues as classified in Boussau et al. (2008), the number of SLiMs in the translated sequence, and the nucleotide sequence.

Table S5. Mean TANGO and Waltz aggregation propensity scores across 50 scrambled versions of the protein corresponding to the Ensembl gene identifier.

Table S6. Dispersion-controlled scrambled sequences, listing the Ensembl gene identifier of the protein that was scrambled, the TANGO and Waltz aggregation propensity scores, the normalized hydrophobic dispersion index for the original protein, the normalized hydrophobic dispersion index for the scrambled amino acid sequence, and the dispersion-controlled scrambled amino acid sequence.

Supplementary References

Boussau, B., S. Blanquart, A. Necsulea, N. Lartillot and M. Gouy, 2008 Parallel adaptations to high temperatures in the Archaean eon. Nature 456**:** 942-945.

Dinkel, H., K. Van Roey, S. Michael, M. Kumar, B. Uyar *et al.*, 2016 ELM 2016—data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. 44**:** D294-D300.

Irbäck, A., C. Peterson and F. Potthast, 1996 Evidence for nonrandom hydrophobicity structures in protein chains. Proc. Natl. Acad. Sci. USA 93**:** 9533-9538.

Smedley, D., S. Haider, S. Durinck, L. Pandini, P. Provero *et al.*, 2015 The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43**:** W589-W598.

Smit, A., R. Hubley and P. Green, 2015 RepeatMasker Open-4.0 version 4.0.5. url=http://www.repeatmasker.org.

Wang, M., C. J. Herrmann, M. Simonovic, D. Szklarczyk and C. von Mering, 2015 Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15**:** 3163-3168.

Wilson, B. A., S. G. Foy, R. Neme and J. Masel, 2017 Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. 1**:** 0146.