Supplementary Materials for "Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of *De Novo* Gene Birth" by Wilson et al.

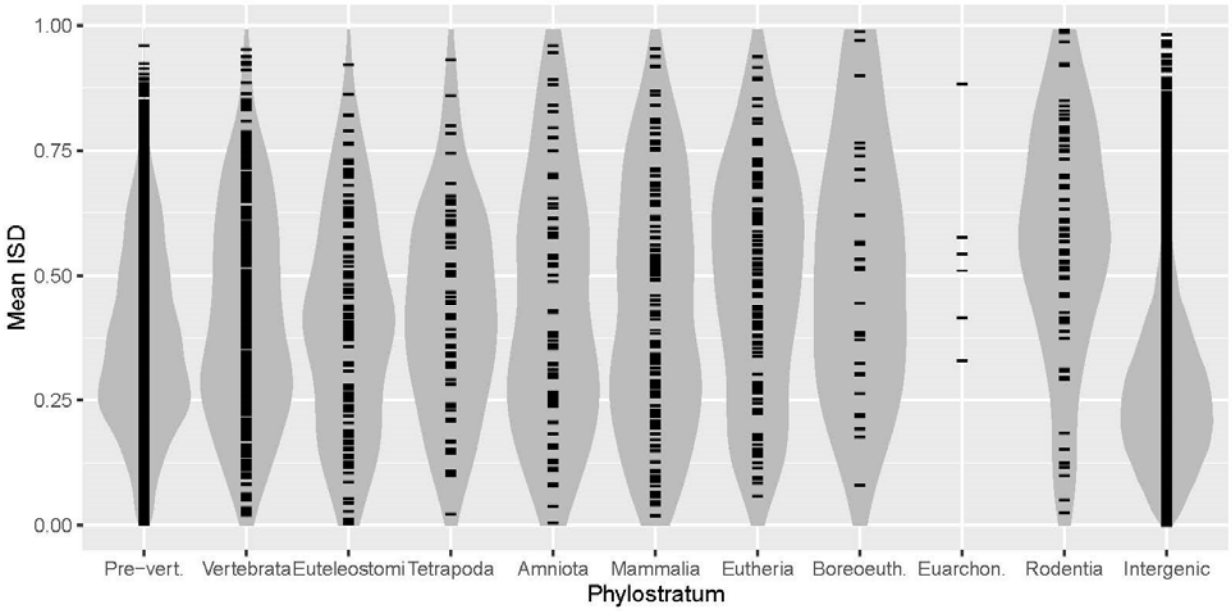

Fig. S1: Variance among gene families in their ISD. Each small horizontal bar represents the mean of a gene family, with violin plots fitted to distributions containing more than 6 gene families. Similarly to previous work[1], genes show higher variance than that expected from random sequences, presumably due to functional diversification. Here we also see that older gene families exhibit less variance than young gene families.

Table S1. A key to the phylostratum numbers used in Table S2.

Phylostratum	Gene shared among	Number of Gene Families	Number of Genes in the Gene Families
1	Cellular_Organisms	2958	8142
2	Eukaryota	2507	4873
3	Opisthokonta	122	220
4	Holozoa	212	385
5	Metazoa	456	1354
6	Eumetazoa	330	531
7	Bilateria	235	382
8	Deuterostomia	106	217
9	Chordata	56	76
10	Olfactores	18	36
11	Vertebrata	478	718
12	Euteleostomi	139	218
13	Tetrapoda	67	75
14	Amniota	69	75
15	Mammalia	144	179
16	Eutheria	111	145
17	Boreoeutheria	31	34
18	Euarchontoglires	6	6
19	Rodentia	79	84
20	Unique to *M. musculus*	NA	789

Legends for large Supplementary data tables.

Table S2. *M. musculus* proteins, listing Ensembl gene identifier, gene family membership, original phylostratum assigned by BLASTp, phylostratum assigned via gene family membership at the end of our processing (as keyed in Table S1), untransformed ISD, evolutionary rate, and the complete protein sequence (including cysteines). The GeneFamilyNumber column identifies members of the same gene family via a shared numeric identifier. Proteins with a BLASTpPhylostratum labelled as "0" failed in the initial BLASTp and were excluded from our analysis. Genes missing both any non-mouse BLASTp hits and Ensembl rat orthologs were assigned GeneFamilyPhylostratum = 20, while those assigned older phylostrata in the initial analysis but which failed to match an Ensembl rat ortholog were assigned a "NULL" GeneFamilyPhylostratum. Evolutionary rates are missing for around a quarter of the proteins, and are marked with "\N". Genes with an ExcludedGeneBinary of "1" were excluded from our analysis either because Ensembl lacks rat-ortholog-based dN/dS values or because their gene family was split among multiple phylostrata.

Table S3. Nucleotide sequences from intergenic regions of *M. musculus* genome, listing the Ensembl gene identifier of the gene to which they were taken in proximity, and the ISD that would be translated from this sequence.

Table S4. Nucleotide sequences from intergenic regions of the masked *M. musculus* genome, listing the Ensembl gene identifier of the gene to which they were taken in proximity, and the ISD that would be translated from this sequence.

Table S5. Randomly generated nucleotide sequences with a GC content matched to that of each coding gene, listing Ensembl gene identifier, and the ISD that would be translated from this sequence.

Table S6. Scrambled amino acid sequences, listing Ensembl gene identifier of the protein that was scrambled, and ISD of the scrambled version.

Table S7. *Saccharomyces cerevisiae* proteins from Table 1, listing the SGD[2] systematic gene name, "Conservation Levels" assigned by Carvunis *et al.*[3] where "NA" indicates that the gene lacked such assignment and was omitted from Fig. 6B, gene family membership, phylostratum both as initially assigned by BLASTp and after reconciliation within gene families, the phylostratum identified in terms of the taxonomic group that is most distant to *S. cerevisiae* but that still contains a homolog of the gene, binaries to indicate gene annotations that are unique to *S. cerevisiae*, classified by SGD dubious, and unclassifiable by our phylostratigraphic pipeline, untransformed ISD values with and without cysteines removed, and the complete protein sequence (including cysteines). Unlike the phylostrata of *M. musculus*, the conservation levels assigned by Carvunis *et al*. ascend as the genes age, i.e. the conservation levels 1,2,3…10 correspond to oldest hits found by Carvunis *et al.*[3] in *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. castellii*, *K. waltii*, *C. albicans*, *Y. lipolytica*, *N. crassa*, and *S. pombe* respectively. The taxa listed in Figure 6A correspond to the GeneFamilyPhylostratum in the following manner: Pre-Ascomycota is 1-5, *S. pombe* is 6, *Y. lipolytica* is 7-8, *K. waltii* is 9, and *S. kudriavzevii* is 9

(subset distinguished in the binary *S. kudriavzevii* column). Those genes with a GeneFamilyPhylostratum of 10 were found in our analysis to be unique to *S. cerevisiae* and were excluded.

1 Tartaglia, G. G., Pellarin, R., Cavalli, A. & Caflisch, A. Organism complexity anti-correlates with proteomic β-aggregation propensity. *Protein Science* **14**, 2735-2740 (2005).

2 Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700-D705 (2012).

3 Carvunis, A.-R. *et al.* Proto-genes and *de novo* gene birth. *Nature* **487**, 370-374 (2012).