Supplementary material

Supplementary table 1

Phylogenetic extent of disease protein set versus 10000 random sets of the same size.

Phylogenetic extent	Number of Disease proteins	Average number of proteins in random sets	Z score	p-value
--B---ILM	48	17	7.91	<1e-04
-AB-PFILM	29	8	7.79	<1e-04
--B--FILM	28	8	7.03	<1e-04
--B-PFILM	20	6	5.50	<1e-04
------ILM	175	112	6.11	<1e-04
-ABTPFI-M	74	43	5.07	<1e-04
-ABTPFILM	112	65	6.08	<1e-04
-------LM	88	52	4.99	<1e-04
-AB--FI-M	13	5	3.69	1.3e-03
--B-PFI-M	22	13	2.33	0.02
------I-M	89	136	-4.44	1e-04
--------M	132	221	-6.68	1e-04
---------	11	181	-13.75	<1e-04

V: Viruses, A: Archaea, B: Bacteria, T: Protista, P: Plants, F: Fungi, I: Invertebrates, L: Vertebrates no mammals, M: Mammals

Supplementary table 2

Statistical analysis of phylogenetic extent of disease genes using only fully sequenced genomes; 15 Archaea, 61 Bacteria, 2 Fungi (*S. Cerevisiae*, *S. Pombe*), 1 Plants (*A. Thaliana*) and 2 Metazoa (*D. melanogaster*, *C. Elegans)*

	Phylogenetic extent	
	Z score	P-value
Archaea	14.3	<1e-04
Bacteria	15.3	<1e-04
Fungi	4.2	<1e-04
Plants	3.7	<1e-04
Metazoa	13.7	<1e-04

Z score of number of proteins conserved in each taxonomic group between 10000 random sets and disease set [$Z_x = (X - \square_x)/\square_x$] and p-value [$p_x = \sum(n_x>X)/N$]

Supplementary figure 1

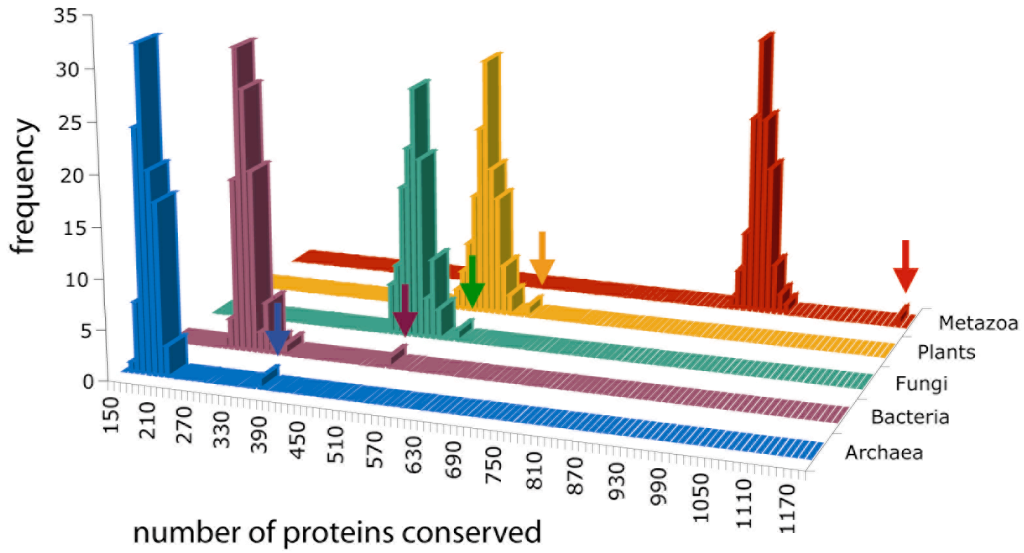

Fig 1. Phylogenetic extent of human disease proteins using complete genomes. Frequency distributions of disease proteins (bars indicated by vertical arrows) with homologs in fully sequenced genomes of Archaea (15 genomes), Bacteria (61 genomes), Fungi (*S. cerevisiae* and *S. pombe*), Plants (*A. thaliana*) and Metazoa (*C. elegans* and *D. melanogaster*), versus 100 control sets of equal size containing randomly selected human proteins. Number of proteins with homologs (from a maximum of 1567) is shown in the x-axis and the frequency of the sets on the y-axis.

Supplementary figure 2

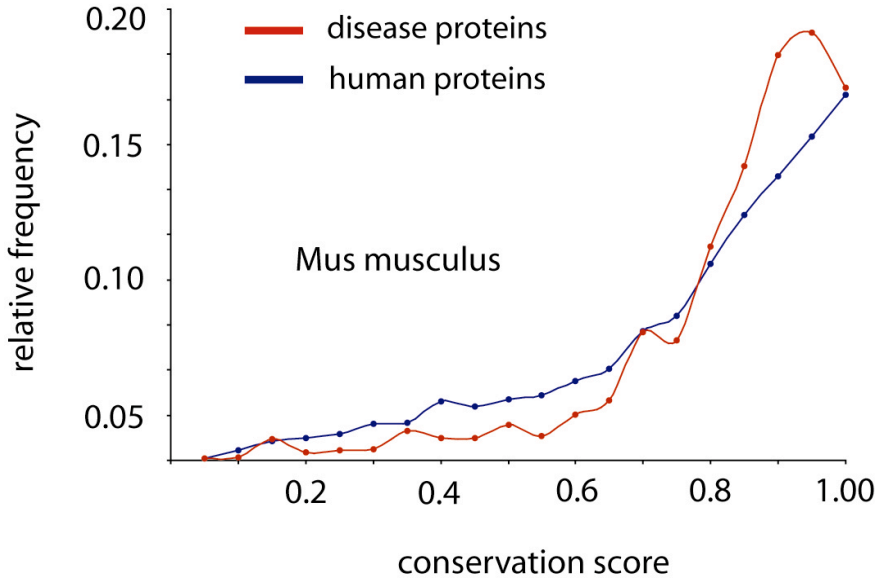

Fig 2. Conservation of disease proteins in *M. musculus* complete genome. Distribution of conservation score of disease (red line) and all human proteins (blue line) in *M. musculus*. The conservation score gives an estimation of the mutation rate that the protein has been subjected to during evolution that is independent of the length of the protein: it is calculated as the BLAST score of the closest homolog in one taxonomic group divided by the BLAST score of the protein against itself, ranging from 0 to 1 (when the closest homolog is 100% identical). The maximum distance (D) between the two distributions (Kolmogorov-Smirnov test) is 13.74 and the P-value is <2.2e-16.