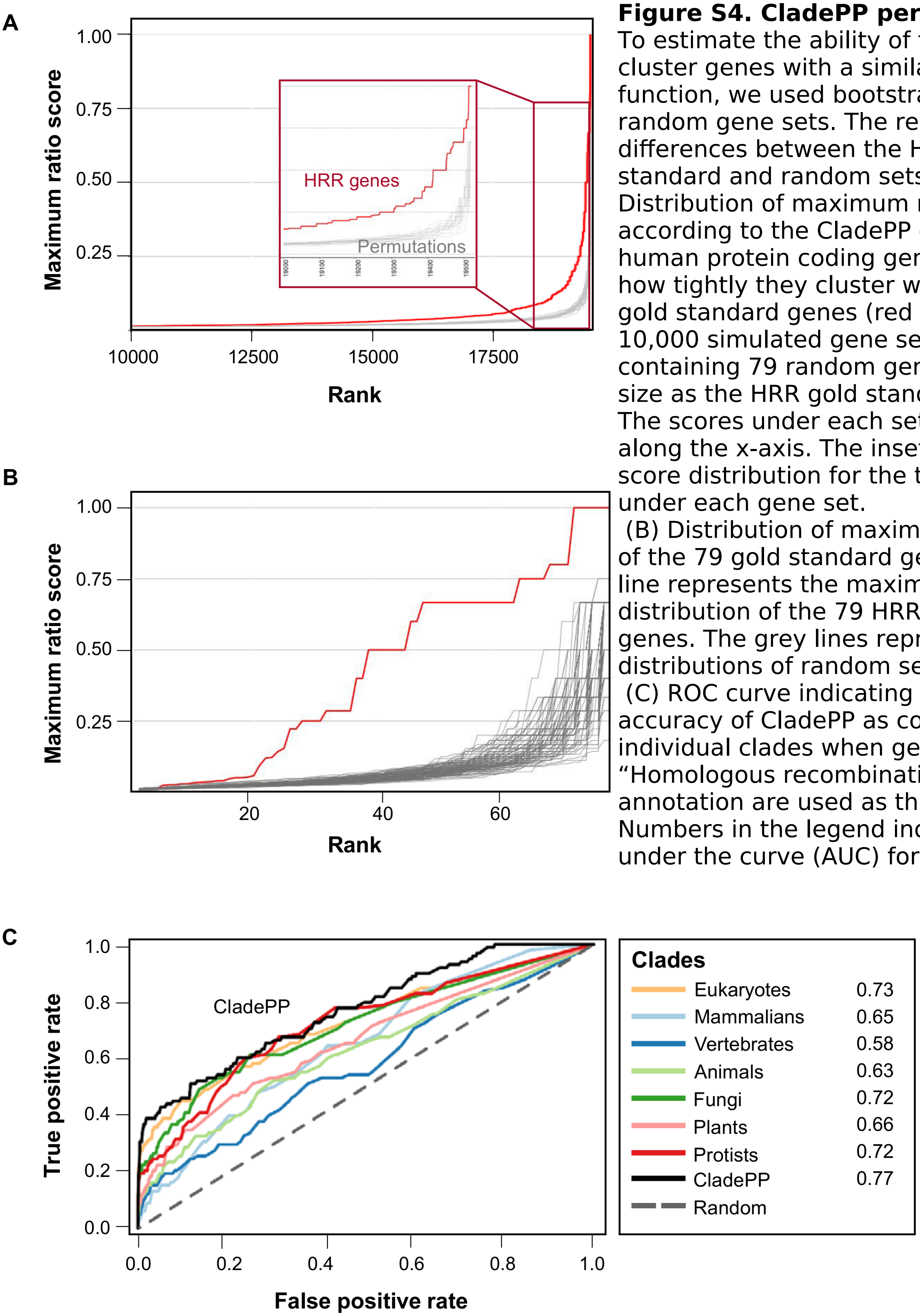

Figure S4. CladePP performance. To estimate the ability of the CladePP to cluster genes with a similar biological function, we used bootstrapping of 10,000 random gene sets. The results show differences between the HRR gold standard and random sets of genes. (A) Distribution of maximum ratio score according to the CladePP of 19,520 human protein coding genes based on how tightly they cluster with the 79 HRR gold standard genes (red line) and with 10,000 simulated gene sets, each containing 79 random genes (The same size as the HRR gold standard, grey lines). The scores under each set are sorted along the x-axis. The inset displays the score distribution for the top 500 genes under each gene set.

(B) Distribution of maximum ratio score of the 79 gold standard genes. The red line represents the maximum ratio score distribution of the 79 HRR gold standard genes. The grey lines represent the score distributions of random sets of 79 genes.

(C) ROC curve indicating the predictive accuracy of CladePP as compared to individual clades when genes with the “Homologous recombination repair ” GO annotation are used as the gold standard. Numbers in the legend indicate area under the curve (AUC) for each clade.