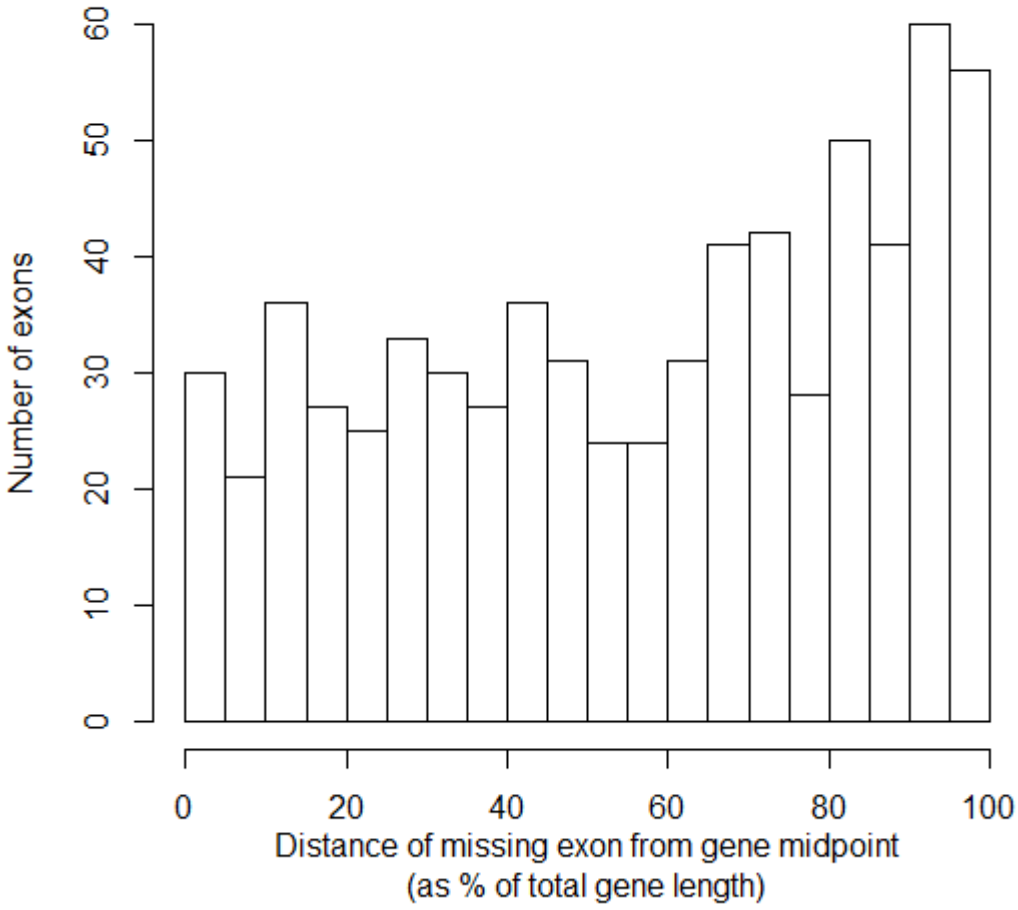
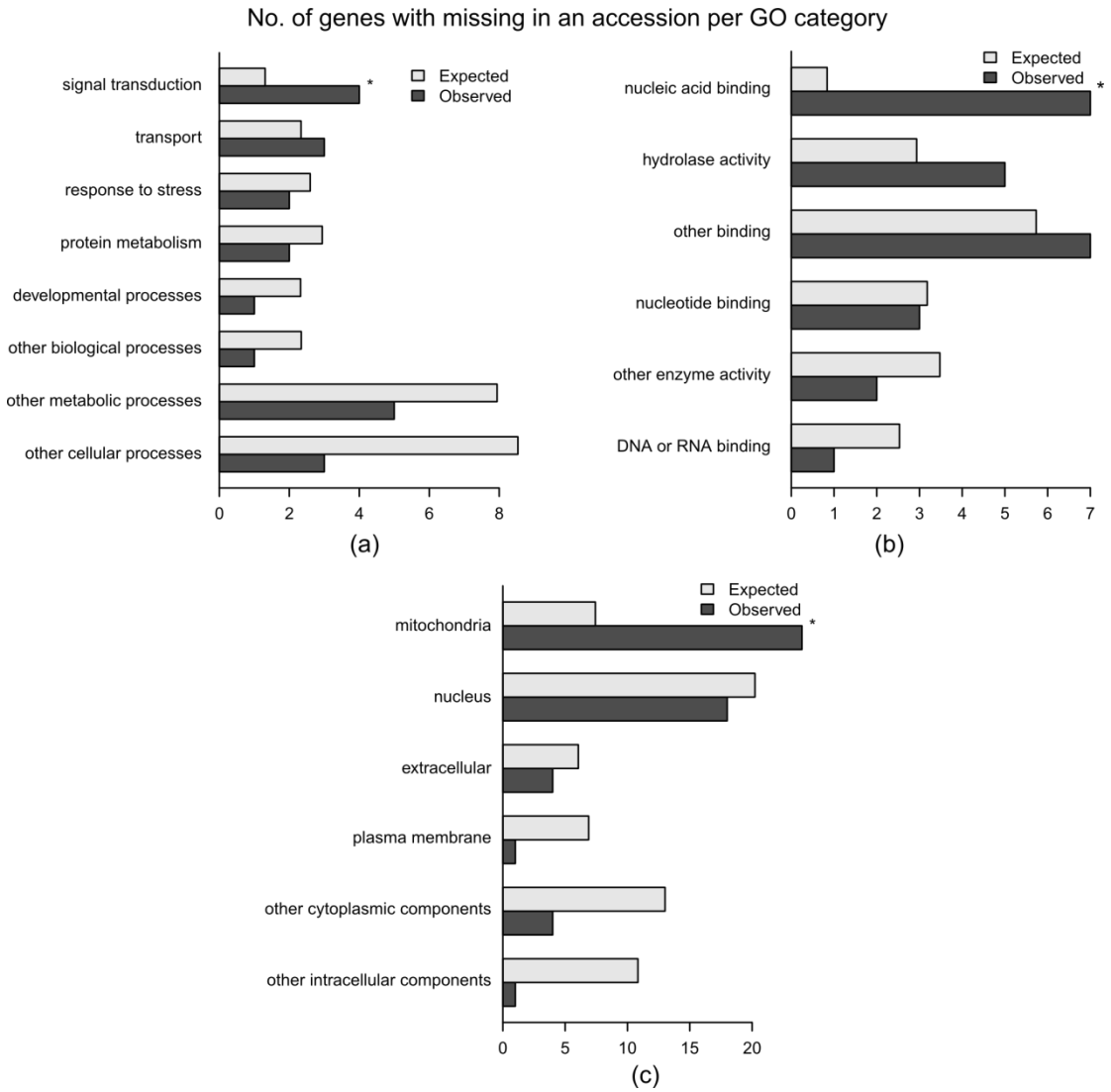
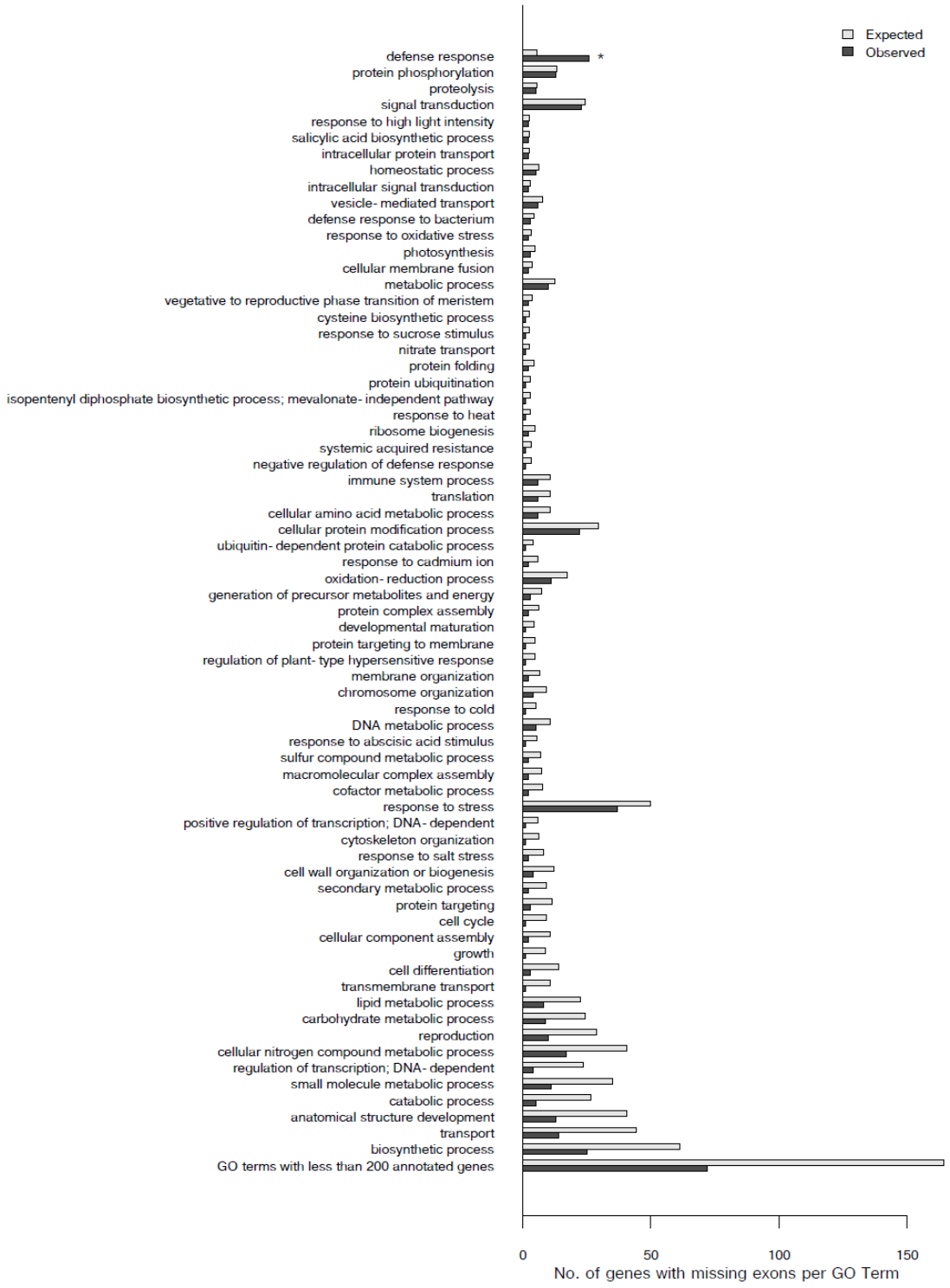
SUPPLEMENTARY FIGURES



Supplementary Figure 1. Location within a Col-0 gene of exons missing in at least one other accession. Only genes with at least one, but not all, exons missing were considered. For the regression of histogram density estimates against bin midpoints, adj. $r^2 = 0.47$, $p = 5e^{-4}$.

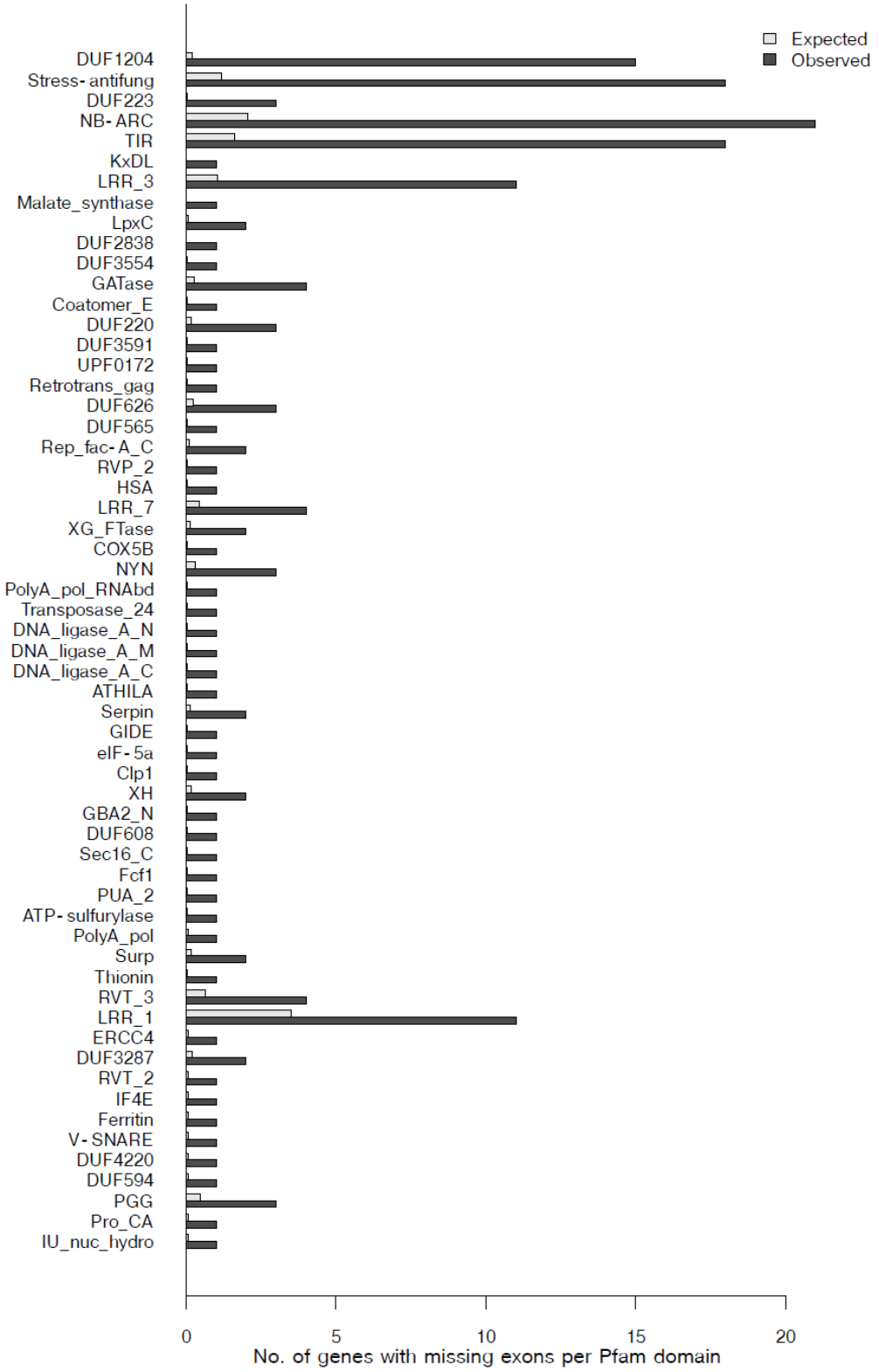


Supplementary Figure 2. Distribution of 'CDS presence/absence' (CDS-PAV) genes (n=81) – those with their entire coding region missing in at least one accession – by GOslim categories for molecular function (a), biological process (b) and cellular component (c). Both expected and observed number of CDS-PAV affected genes per category represented on each bar. Where there is a significant enrichment (p-value <= 0.05) between the amount of observed and expected CDS-PAV genes for a particular category an asterisk is shown over the bars.



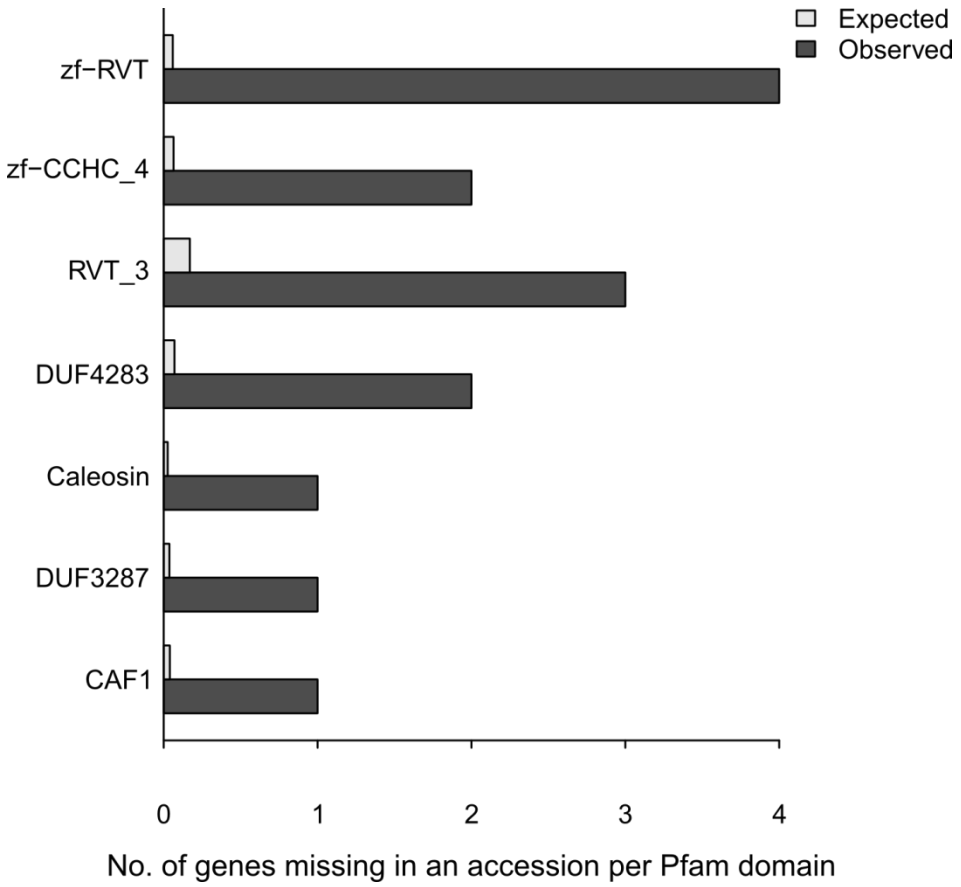
Supplementary Figure 3. Distribution of 'exon presence/absence' (E-PAV) genes – those with at least one, but not all, exons missing in at least one accession – by GO category (n=330). Only GO terms with 200 or greater genes were considered. Both expected and observed number of E-PAV genes per category represented on each bar.

Where there is a significant enrichment (p-value <= 0.05) between the amount of observed and expected E-PAV genes for a particular category an asterisk is shown over the bars.

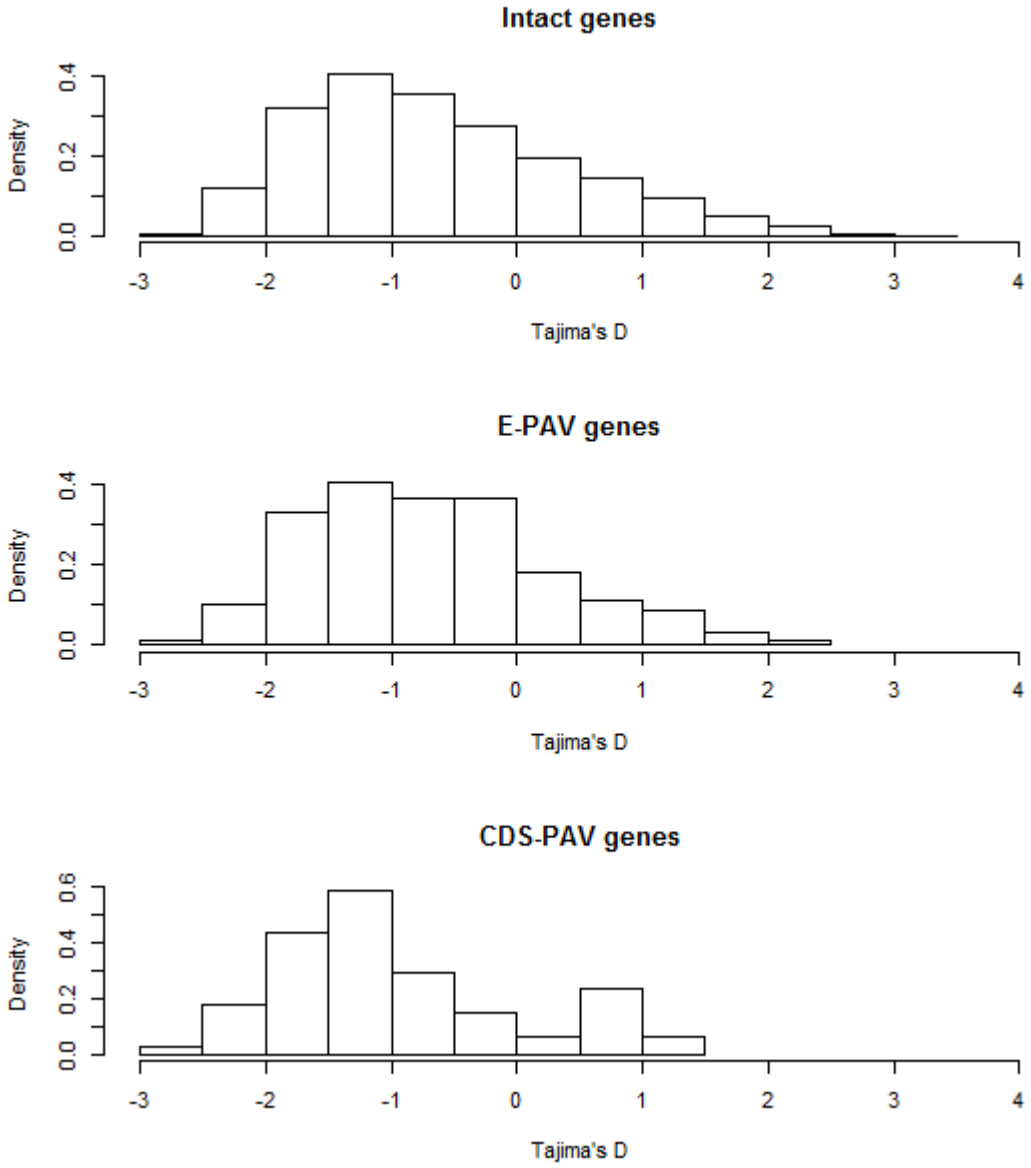


Supplementary Figure 4. Distribution of 'exon presence/absence' (E-PAV) genes – those with at least one, but not all, exons missing in at least one accession – by Pfam category (n=330). Both expected and observed number of E-PAV genes per category

represented on each bar. Only categories with a significant enrichment are shown. Note that there is a significant enrichment of E-PAV genes in the category of 'no Pfam annotation' (adjusted p-value = $3.96e^{-5}$).



Supplementary Figure 5. Distribution of 'CDS presence/absence' (CDS-PAV) genes – those with their entire coding region missing in at least one accession – by Pfam category (n=81). Both expected and observed number of CDS-PAV genes per category represented on each bar. Only categories with a significant enrichment are shown.



Supplementary Figure 6. Distribution of Tajima's *D* values for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes.