KinFin: Software for taxon-aware analysis of clustered protein sequences

Dominik R. Laetsch [1,2]* and Mark L. Blaxter[1]

[1] *Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT UK*

[2] *The James Hutton Institute, Errol Road, Dundee DD2 5DA UK*

* *Corresponding author: dominik.laetsch@gmail.com*

Supplementary results

Network representation of the clustering

The network (Figure S1) showed no consistent phylogenetic patterning between the taxa, with exception of the separation of the orders Rhabditida (grey) and Spirurida (coloured). Positioning of individual nodes within Spirurida varied for different runs of the layout algorithm (data not shown) and even nodes of the same species (*i.e.* LOA1/LOA2, WBANC1/WBANC2, OOCHE1/OOCHE2) were occasionally, spatially separated in the network which indicated non-overlapping gene predictions between the different proteomes for a given species.

Comparison of clustering behaviour of proteomes for which two assemblies exist

The different clustering behaviour for the proteins predicted for these assemblies is shown in Figure S2. For *L. loa* and *W. bancrofti*, higher number of proteins in a proteome correlates with both higher number of singleton proteins and proteins in clusters shared with other nematode species. Interestingly, genome assemblies for *W. bancrofti* differ substantially in contiguity and CEGMA completeness (see WormbaseParasite) and the higher proportion of proteins that WBANC1 shares with other species might be a result of fragmented gene predictions, since the average mean lengths of WBANC2 proteins in shared clusters is 91 residues shorter than WBANC1 proteins (436 versus 345).

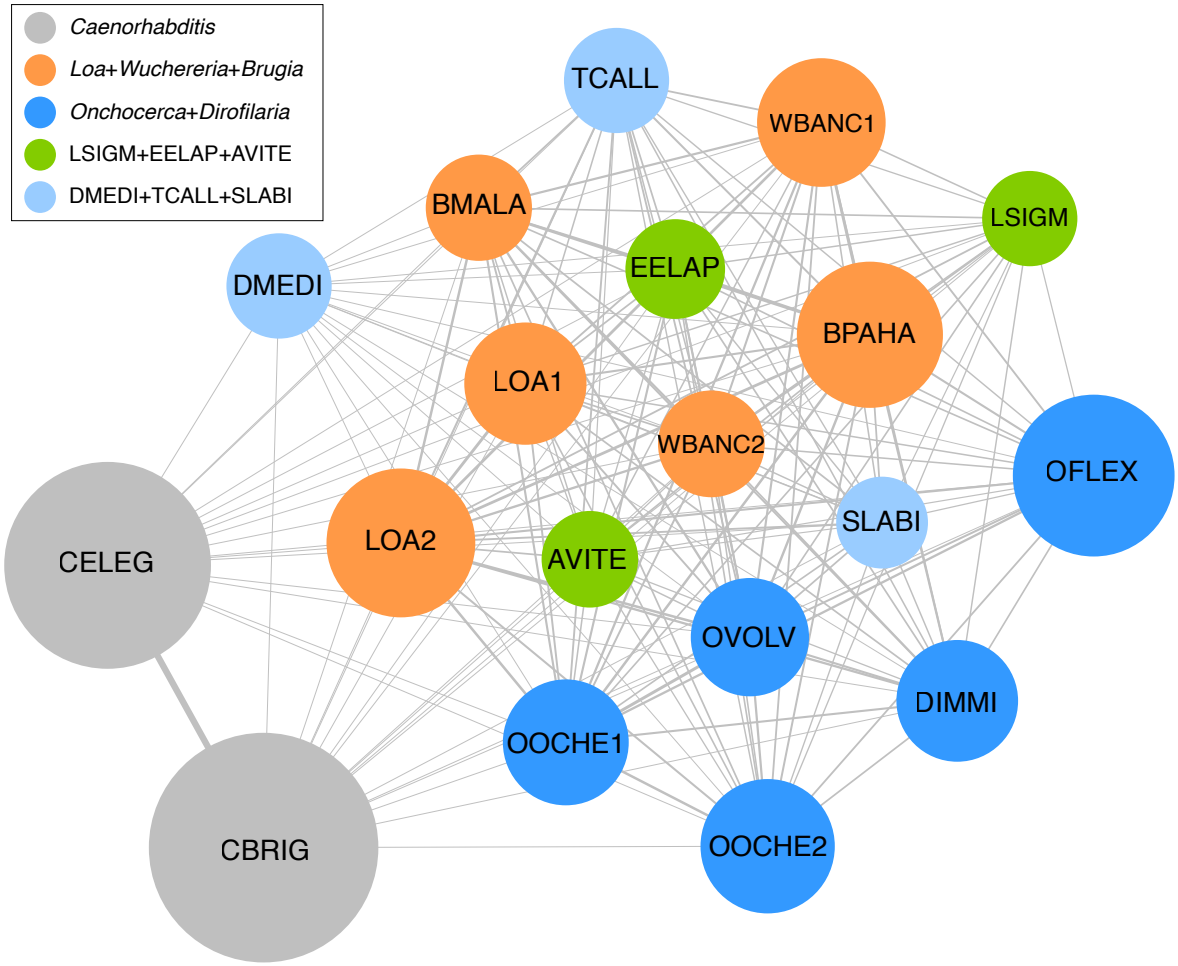

***Figure S1**: Network representation of the clustering data. Proteomes are represented by nodes, coloured by based on phylogenetic clade, scaled by count of proteins, and positioned by a force directed layout algorithm. Edges are drawn between two nodes, weighted by the number of clusters in which both proteomes occur simultaneously (excluding clusters in which all proteomes are present).*

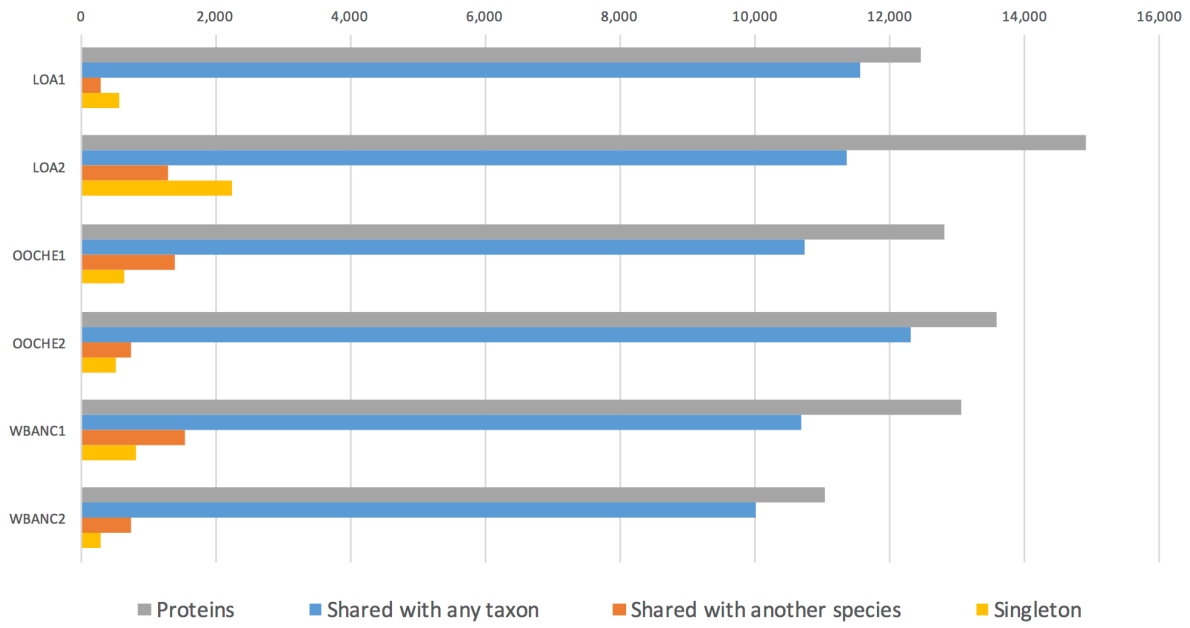

***Figure S2**: Count of proteins for proteomes of the same species. 'Proteins': total number of proteins, 'Singleton': number of proteins in singleton clusters, 'Shared with another species': proteins in clusters shared with another nematode species, 'Shared with any taxon': proteins in clusters shared with any taxon.*