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 methods

KinFin analysis protocol for filarial nematode analyses

Clustering and functional annotation of protein sequences

For the proteomes detailed in Table 1, protein FASTA and GFF3 files were downloaded from WormBase parasite (WBPS8) (Howe, Bolt, Shafie, et al. 2016; Howe, Bolt, Cain, et al. 2016) and ngenomes.org.

Protein FASTA files were filtered and sequences shorter than 30 residues or containing non-terminal stops were excluded (`filter_fastas_before_clustering.py`, this generates single-line FASTA files that can easily be partitioned based on a list through the UNIX command `grep`). Non-longest isoforms were removed (`filter_isoforms_based_on_gff3.py`) and the resulting FASTA files were functionally annotated through `InterProScan v5.22-61.0`

(Jones et al 2014) using the `Pfam-30.0` database (Finn et al. 2016) and `SignalP-EUK-4.1` (Petersen et al. 2011) and converted to KinFin compatible format (`iprs_to_table.py`).

`OrthoFinder v1.1.4` (Emms and Kelly 2015) was used to generate the commands for BLASTp analyses. The BLASTp commands were further modified by adding the following options `-seg yes`, `-soft_masking true` and `-use_sw_tback` as suggested by (Moreno-Hagelsieb and Latimer 2008). BLASTp analyses were run on the EDDIE supercomputing cluster at the University of Edinburgh using `BLASTp v2.3.0+` (Camacho et al. 2009). Proteome clustering was carried out at default MCL inflation value of 1.5.

Basic KinFin analysis and phylogenetic inference

To construct a phylogenetic tree of the taxa that are included in the analysis, an initial, basic KinFin analysis was run by supplying input files contained in supplementary dataset 1. This initial analysis revealed 781 "true" and 3887 "fuzzy" single-copy orthologues (where 75% of species displayed single-copies and the remaining taxa varied between 0 and 100 copies). Proteins for each of the "true" single-copy orthologues clusters were extracted (`get_protein_ids_from_cluster.py` and `GNU grep`), and aligned using `mafft v7.267` (E-INS-i algorithm) (Katoh and Standley 2013). Alignments were trimmed using `trimal v1.4` (Capella-Gutiérrez et al. 2009) and concatenated using `FASconCAT v1.0` (Kück and Meusemann 2010), prior to phylogenetic tree reconstruction using `RAxML v8.1.20` (Stamatakis 2014) under the PROTGAMMAGTR model of sequence evolution and 20 alternative runs on distinct starting trees. Non-parametric bootstrap analysis was carried out for 100 replicates.

Advanced KinFin analysis

KinFin was run by providing the input files in supplementary dataset 2. In brief, taxon sets were defined for the taxonomic rank of "order" by supplying NCBI TaxIDs for each proteome, for the attribute "clade" by grouping taxa into taxon-sets for the major filarial clades, and for the attribute "host" by separating human parasites from those of other animals and outgroups. For the attribute of "clade", only one proteome per species was allocated to its respective taxon set (LOA2,

OOCHE1, and WBANC2) and unique labels were specified for the remaining taxa. The Mann-Whitney-U test was selected for pairwise protein count representation tests and the required number of proteomes in a taxon-set to be used in rarefaction/representation-test computations was set to 2.

The topology of the tree inferred through phylogenetic analysis was provided in Newick format and the two *Caenorhabditis* species were specified as outgroups for rooting the tree by setting the attribute "OUT" in the config file to 1, and to 0 for all other taxa.

Visualisation of the clustering and calculation of metrics

The distribution of cluster sizes was generated using `plot_cluster_sizes.py` and specifying the colour map "viridis". Counts of proteins by cluster type were extracted from `TAXON.attribute_metrics.txt` (folder "TAXON").

Inference of representative functional annotation of clusters

Using the script `filter_functional_annotation_of_clusters.py`, representative functional annotation of clusters was inferred for all clusters (`--domain_taxon_cov 0.75`, `--domain_protein_cov 0.75`) and for synapomorphic clusters (`--node_taxon_cov 0.75`, `--domain-taxon-coverage 0.75`, `--domain-protein-coverage 0.75`).

Analysis of clusters specific to and shared between taxon-sets and assessment of protein space

Analyses on clusters were performed on the following files:

- `order.Rhabditida.cluster_metrics.txt` (folder "order")
- `order.Spirurida.cluster_metrics.txt` (folder "order")
- `clade.pairwise_representation_test.txt` (folder "clade")
- `cluster_counts_by_taxon.txt`

and the representative functional annotation inferred in the previous step. The plot of the rarefaction curves was taken directly from the KinFin output (folder "clade").

Querying clustering and functional annotation using target genes

The output of KinFin was analysed using the script `get_count_matrix.py` to obtain protein counts by species for genes involved in heme homeostasis and biosynthesis. Presence/absence of unpredicted genes was confirmed through using `TBLASTn v2.3.0+` (Camacho et al. 2009) against the respective genomes. Presence of paralogues was confirmed by manual inspection of gene models on WormBase ParaSite.

Network representation of the clustering

A network representation of the clustering was generated using `generate_network_representation.py` and by ignoring clusters in which all taxa are present (`--exclude_universal`), and visualised using `Gephi v0.9.1` (Bastian et al. 2009). Starting from a random layout, nodes in the graph were positioned using the force directed ForceAtlas2 layout algorithm (Jacomy et al. 2014) using the parameters: `Tolerance=0.2, Approximation=1.2, Approximate Repulsion=False, Scaling=5000, Stronger Gravity=False, Gravity=1.2, LinLog mode=True, Dissuade hubs=True, Prevent overlap=True, Edge Weight Influence=1.0`. Under this layout algorithm nodes repulse each other like charged particles, while edges attract their nodes like springs. Nodes were coloured by phylogenetic clade and scaled proportional to the size of the proteome.

Comparison of clustering behaviour of proteomes for which two assemblies exist

Clustering behaviour was analysed by consulting the `TAXON.*.cluster_metrics.txt` files for each of the proteomes in question.

Bibliography

Bastian M., Heymann S., Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10, p. 421.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15), pp. 1972–1973.

Emms, D.M. and Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, p. 157.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. and Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44(D1), pp. D279-85.

Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., Harris, T.W., Kishore, R., Lee, R., Lomax, J., Li, Y., Muller, H.-M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D. and Sternberg, P.W. 2016. WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Research* 44(D1), pp. D774-80.

Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P. and Berriman, M. 2016. WormBase ParaSite - a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology*.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *Plos One* 9(6), p. e98679.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y.,

1 Lopez, R. and Hunter, S. 2014. InterProScan 5: genome-scale protein function classification.
2 *Bioinformatics* 30(9), pp. 1236–1240.

3 Katoh, K. and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7:
4 improvements in performance and usability. *Molecular Biology and Evolution* 30(4), pp. 772–780.

5 Kück, P. and Meusemann, K. 2010. FASconCAT: Convenient handling of data matrices. *Molecular*
6 *Phylogenetics and Evolution* 56(3), pp. 1115–1118.

7 Moreno-Hagelsieb, G. and Latimer, K. 2008. Choosing BLAST options for better detection of
8 orthologs as reciprocal best hits. *Bioinformatics* 24(3), pp. 319–324.

9 Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. 2011. SignalP 4.0: discriminating signal
10 peptides from transmembrane regions. *Nature Methods* 8(10), pp. 785–786.

11 Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
12 phylogenies. *Bioinformatics* 30(9), pp. 1312–1313.