GeneSeqToFamily: a Galaxy workflow to find gene families based on the Ensembl Compara GeneTrees pipeline

Anil S. Thanki, Nicola Soranzo, Wilfried Haerty, Robert P. Davey

Earlham Institute, Norwich Research Park, Norwich NR4 7UH, UK
Email: nicola.soranzo@earlham.ac.uk

Project Website: https://github.com/TGAC/earlham-galaxytools
License: MIT

The phylogenetic information inferred from the study of homologous genes helps us to understand the evolution of gene families and plays a vital role in finding ancestral gene duplication events as well as identifying genes that are under positive selection within species.

The Ensembl GeneTrees pipeline generates gene trees based on coding sequences and provides details about exon conservation, and is used in the Ensembl Compara project to discover homologous gene families. Since substantial expertise is required to configure and run the pipeline via the command line, we created GeneSeqToFamily, an open-source Galaxy workflow based on Ensembl GeneTrees. GeneSeqToFamily helps users to run potentially large-scale gene family analyses without requiring command-line usage while still allowing parameters, configurations, and the tools themselves to be modified.

Gene families discovered with GeneSeqToFamily can be explored using the Aequatus.js interactive tool, which was integrated into Galaxy as a visualisation plugin.

We validated the workflow results against the Ensembl GeneTrees database using a small set of genes from known families. We also studied the impact of the most important tool parameters on the reconstructed gene families by running it on the reference proteomes established by the Quest for Orthologs consortium (754,149 sequences from 66 species).

We are further developing the Galaxy workflow implementing the ETE toolkit for the downstream analysis of the discovered gene families, allowing splitting gene trees by duplication events, contributing to the identification of orthologs and paralogs and setting the ground for selection analysis.

Citation: Thanki AS, Soranzo N, Haerty W, Davey RP. GeneSeqToFamily: a Galaxy workflow to find gene families based on the Ensembl Compara GeneTrees pipeline. GigaScience, giy005, https://doi.org/10.1093/gigascience/giy005