PHYS.ORG

# How are we related? A Compara-bly easy workflow to find gene families

March 5 2018

Credit: Earlham Institute

Published in *GigaScience*, the open source Galaxy workflow allows researchers to make easier work of finding gene families; an important tool when it comes to analysing the evolution, structure and function of genes across species.

Co-author, Wilfried Haerty explained why this tool is so useful to biologists: "The software developed at the Earlham Institute enables scientists to investigate species of interest using a flexible and reproducible pipeline. The performance of our workflow was assessed on vertebrate genome assemblies of various qualities (platypus, pig, horse, dog, mouse and human). The species were selected to assess the impact of genome quality on gene families identification. The mouse, dog and human genomes are of high quality whereas the three others are at different stages of analysis completion."

Based on and expanding Ensembl's existing EnsemblCompara Gene Trees pipeline, the GeneSeqToFamily workflow removes many complex prerequisites of the process, such as having to use the command line to install a large number of separate tools, by converting the whole process into Galaxy; a much simpler platform to use.

Importantly, the workflow is highly customisable, allowing users to choose parameters, change tools and run the software on their own genes , without having to use the Ensembl database.

Not just a workflow, GeneSeqToFamily contains a number of new, standalone Galaxy tools, including TreeBeST, hcluster_sg, T-Coffee and ETE. Developed at EI by Anil Thanki and Nicola Soranzo of the Data Infrastructure Group, the software makes the process of finding and generating phylogenetic trees easier, using a range of open platforms and databases. Anil Thanki, Scientific Programmer, said: "We are excited to put our work in the open domain, where it allows biologists and bioinformaticians to use the Ensembl Compara GeneTrees Pipeline in a simple, graphical user interface and modify it if needed."

The team hopes that the new workflow will help users unfamiliar with the complexities associated with using Compara to be able to more easily analyse phylogenetic datasets, while collating a number of useful gene family tools in one Galaxy workflow. Users can either select existing Ensembl databases to use as the reference sets for their analysis, or provide their own data in the same format, and tools are provided that can help.

Earlham Institute is committed to providing tools and algorithms to support, enable and develop computational biology and life sciences research, with projects such as Galaxy helping to open access to a range of scientific tools and databases.

The Data Infrastructure group, led by Dr. Rob Davey, also supports resources such as CyVerse UK and COPO which, alongside Galaxy, expand the availability and usability of computational resources to the wider scientific community in the UK and internationally through EI's National Capability in e-Infrastructure.

**More information:** Anil S Thanki et al. GeneSeqToFamily: a Galaxy workflow to find gene families based on the Ensembl Compara GeneTrees, *GigaScience* (2018). DOI: 10.1093/gigascience/giy005

Provided by Earlham Institute

Citation: How are we related? A Compara-bly easy workflow to find gene families (2018, March 5) retrieved 20 April 2024 from https://phys.org/news/2018-03-compara-bly-easy-workflow-gene-families.html