GigaScience

GeneSeqToFamily: the Ensembl Compara GeneTrees pipeline as a Galaxy workflow

--Manuscript Draft--

Manuscript Number:	GIGA-D-17-00074R1	
Full Title:	GeneSeqToFamily: the Ensembl Compara GeneTrees pipeline as a Galaxy workflow	
Article Type:	Technical Note	
Funding Information:	Biotechnology and Biological Sciences Research Council (BBSRC strategic funds)	Mr. Anil S Thanki Dr Wilfried Haerty Dr Robert P Davey
	Biotechnology and Biological Sciences Research Council (BBSRC Biomathematics and Bioinformatics Training fund (2014).)	Dr Nicola Soranzo
Abstract:	Background Gene duplication is a major factor contributing to evolutionary novelty, and the contraction or expansion of gene families has often been associated with morphological, physiological and environmental adaptations. The study of homologous genes helps us to understand the evolution of gene families. It plays a vital role in finding ancestral gene duplication events as well as identifying genes that have diverged from a common ancestor under positive selection. There are various tools available, such as MSOAR, OrthoMCL and HomoloGene, to identify gene families and visualise syntenic information between species, providing an overview of syntenic regions evolution at the family level. Unfortunately, none of them provide information about structural changes within genes, such as the conservation of ancestral exon boundaries amongst multiple genomes. The Ensembl GeneTrees computational pipeline generates gene trees based on coding sequences and provides details about exon conservation, and is used in the Ensembl Compara project to discover gene families. Findings A certain amount of expertise is required to configure and run the Ensembl Compara GeneTrees pipeline via command line. Therefore, we have converted the command line Ensembl Compara GeneTrees pipeline into a Galaxy workflow, called GeneSeqToFamily, and provided additional functionality. This workflow uses existing tools from the Galaxy ToolShed, as well as providing additional wrappers and tools that are required to run the workflow. Conclusions GeneSeqToFamily represents the Ensembl Compara pipeline as a set of interconnected Galaxy tools, so they can be run interactively within the Galaxy's user-friendly workflow environment while still providing the flexibility to tailor the analysis by changing configurations and tools if necessary. Additional tools allow users to subsequently visualise the gene families produced by the workflow, using the Aequatus.js interactive tool, which has been developed as part of the Aequatus software project.	
Corresponding Author:	Anil S Thanki, MSc Earlham Institute Norwich, Norfolk UNITED KINGDOM	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Earlham Institute	
Corresponding Author's Secondary Institution:		
First Author:	Anil S Thanki, MSc	
First Author Secondary Information:		

Order of Authors:	Anil S Thanki, MSc
	Nicola Soranzo, PhD
	Wilfried Haerty, PhD
	Robert P Davey, PhD
Order of Authors Secondary Information:	

Response to Reviewers:

Dear Reviewers:

I am pleased to resubmit for publication the revised version of “GeneSeqToFamily: the Ensembl Compara GeneTrees pipeline as a Galaxy workflow.” I appreciated the constructive criticisms of the reviewers. I have addressed each of their concerns as outlined below.

Reviewer #1:

The authors have put together a Galaxy workflow incorporating various tools to be able to perform Ensembl Compara's pipeline for gene family identification and synteny visualization. This in itself is valuable and useful and I have no concerns with the pipeline itself, though I have not tried it.

However, I am unsatisfied with the parameter evaluation and benchmarking:

1 - First, to test the tool, the objective must be clearer defined. Is the goal to find homologs most generally? Full-length homologs? Orthologs? And if orthologs, then orthologs relative to the set of species used as input or something else? This must be explicitly stated for the evaluation and parameter search to be well-defined. The manuscript should also make clear that this goal is the one that utility for reaching is implied, instead of any other goal.

Response:
The main aim of the test is to demonstrate that we are able to identify orthogroups (groups of genes sharing ancestry by descent) across a set of species. In the revised version of the manuscript, we clarified the goals of the tool and its purpose.

2 - The data set used for parameter searches is tiny, just three species, two of which are obscure (how good are assemblies?). This should be done on a much larger set, perhaps use the Quest for Orthologs' reference dataset?

Response:
Although we only used three species, the data set includes over 63,000 genes. The reviewer raises a valid question regarding the quality of the assemblies, i.e. while the mouse genome can be considered of very high quality, both the platypus and Tasmanian devil genome assemblies are of much lower quality. Therefore, we have now run our workflow on two sets of three species, one including high quality assemblies (human, mouse and dog), the second set composed of draft genome assemblies only (pig, platypus and horse). Furthermore, we ran our analysis on all six species to assess the impact of mixed assemblies qualities on the identification of gene families.

3 - Testing six settings does not show that one of the settings consistently give any particular results. The way to do that would be to test each parameter set many times, using different species combinations each time, and check if indeed Analysis 5 parameters consistently yield desired results under a well-defined quality metric (dependent on the goal posts set above).

Response:
As per reviewer’s request, we tested six sets of parameters on the high quality, low quality, and both high and low quality datasets mentioned above. All datasets tested confirm that Analysis 5 returns the optimal results for these datasets. We acknowledge that different datasets might require further tuning, but we believe these parameters to be suitably comprehensive for most analyses.

4 - Testing performance on 23 genes is vastly insufficient. Please look into much larger benchmark sets (e.g. OrthoBench or quest for orthologs' benchmark tool), choose some larger, more diverse (functionally and phylogenetically) set to use as a gold standard, define a metric for success, then show how well the tool works under repeated applications of method settings on different subsets of the data.

Response:
We tried running our workflow with ‘Quest for orthologs’, but the recommended dataset available for this tool is very old (2011). We contacted the author of Quest for Orthologs for the latest release (2017), but there is no gold standard resultant data available to compare to, so maybe we can look into benchmarking this workflow in the future when results from other tools are available to compare with.

5 To be clear, I think this is valuable and worthy of publication, but for any claim of functionality, it must be robustly tested under clear definitions on large-scale diverse and representative data as stated above (which should also have some clear measure of accuracy relative to the gold standard).

Response:
We thank the reviewer for their comments, and we hope that the revised manuscript and associated broader analyses will help readers decide whether this workflow is suitable for their datasets. We feel that the mechanisms that the reviewer kindly suggested for assessing our workflow against a gold standard are inadequate to give a representative assessment. We will continue to monitor the available tools and standards in order to provide support for the usefulness of the workflow.

6 Similarly, please add time and memory consumption statistics for different use cases.

Response:
Time and memory consumption for the workflow depends on the infrastructure it is running on. We ran the Galaxy jobs on an High Performance Computing cluster, where measuring the amount of memory used is tricky. Also, the total time used to complete the pipeline may heavily vary depending on the time spent in the cluster scheduler waiting for resources to be available.

Anyhow, most of the time is spent on the NCBI blastp step, which for the examples in the paper varied between 16 and 24 hours for 6 species with 125,342 sequences.

Reviewer #2:
The article presents GeneSeqToFamily, the Ensembl Compara GeneTrees pipeline as a Galaxy workflow. Authors introduce the importance of studying homologous genes to understand evolution of gene families. They present the Ensembl Compara GeneTrees pipeline and the resulting conversion into a Galaxy workflow named GeneSeqToFamily. This workflow is thus the first "ready-to-use" solution providing information about structural changes within genes using Galaxy. The article is well written and I would like to see it published, I have provided very few notes, comments, and suggestions below.
I don't succeed to find typo or other little errors in the manuscript. Moreover, I don't have any major recommendation. Only few minors remarks, maybe ideas who can be of interest… or not ;)

Article

I particularly like the manner authors have presented workflow tools in the table 1. This is a good manner to summarized components, details and origins of each tool. Moreover, this translate the wish of transparency by authors and the fact they don't want to reinvent the wheel using, if relevant, existing tools.

Response:
We thank the reviewer for their kind comments and agree that reinventing the wheel is something we tried to avoid!

1) It appears to me that selecting the longest CDS per gene is a good idea and I'm wondering if this step cannot be made through the use of "classical" existing Galaxy tools as "Text manipulation" ones.

Response:
It is possible to use other tools, but users would probably have to chain together multiple tools to achieve the same result. By having our own tool we can make sure it works perfectly with different data formats, which could be difficult with other tools.

Concerning clustering steps, I want to propose some ideas, but not sure these are appropriate….

2) Not sure this is relevant but concerning clustering of similar sequences, I'm wondering if the use of algo like Swarm (https://github.com/torognes/swarm) (dedicated to metagenomics amplicon-based ngs data clustering) can be of interest…

Response:
As a first release of the workflow, we mainly concentrated on reproducing the Ensembl Compara pipeline into an easy-to-use Galaxy workflow. But we are grateful to the reviewer for the suggestion of testing alternatives for the stock tools in the workflow, and that is something we intend to look into in the future.

3) Concerning the overall clustering approach and related filtered steps, maybe it can be applied
a) a combination of parallel clustering methods (K-means/PAM/CAH/…) to compare "raw" clustering results and thus eliminate / reduce potential noise (ie sequences moving from one cluster to another when we change the clustering method, translating that this sequence is hard to assign)
b) a clustering affecting bootstrap confidence score to nodes thus giving ideas about relevance of each cluster.

Response:
We are grateful to reviewer for the suggestions for the clustering tools. As mentioned above, in first release of the workflow our aim is to reproduce the Ensembl Compara pipeline into a Galaxy workflow, which we managed to achieve. Improving the workflow by modifying the clustering approach is something we can look into in a future release of the workflow.

4) you propose the use of the Newick format. Maybe, as for the Philoviz visualization plugin, it would be interesting to add Nexus and phyloxml datatypes?

Response:
This workflow requires Newick format for the species tree because TreeBeST accepts input only in this format. There are several tools available to convert from Nexus and PhyloXML to Newick format, e.g. http://biopython.org/wiki/Phylo , so users who are producing data in these formats can perform the conversion before uploading them to Galaxy.

6) Maybe authors can give more details more aspects related to particular issues related to
a) alternative transcripts as it appears to me that ""just"" changing the max_target_seqs value from 3 to 4 seems to be quite "light" and

Response:
We agree with the reviewer that the explanation for changing the max_target_seqs value was a bit unclear in the manuscript, we have rewritten this part and reviewers will hopefully find it more satisfying.

b) working on non-model organisms.

Response:
With the significant reduction of sequencing costs, the proportion of fully sequenced genomes from non-model organisms has been steadily increasing. The analysis of the evolution of gene families across many species, including non-model organisms,

allows us to assess the quality of the genome assemblies and associated annotations. We can also derive observations linking the expansion and contraction of gene families, as well as detect signals of positive selection in genes with 1:1 orthologs to specific physiological and behavioural adaptations (e.g. association between beta-keratin and shell evolution in turtles; Li et al. 2013. Genome Biol Evol. 5(5):923-33); taste receptors in whales - Feng et al. 2014 Genome Biol Evol. 6(6):1254-65). We have added such details into the manuscript.

Unfortunately, I don't have taking time to implement the GeneSeqToFamily tools and workflow to a dedicated Galaxy instance. I would have appreciated to have a link giving access to a testing GeneSeqToFamily Galaxy instance. Maybe creating and pointing a dedicated Docker based Galaxy flavor can be of interest, thus giving an easy way to deploy a GeneSeqToFamily local Galaxy instance.

Response:
We have created a virtual image for testing the workflow, available [1], which is ready to be used with VirtualBox, and is supplied with necessary tools and plugins installed for the workflow. We will look into producing a Docker image for the same purpose in future.

8) Moreover, regarding the https://github.com/TGAC/earlham-galaxytools/tree/master/workflows/GeneSeqToFamily github link, the readme file "only" explain GeneSeqToFamily workflow without instructions on how to deploy it.

Response:
We have updated the readme.md file for the workflow with the list of tools required and instructions explaining how to install the workflow.

9) Regarding tools on the toolshed, it would be nice to have a repository called "suite_GeneSeqToFamily" or something like that to easily install all GeneSeqToFamily related Galaxy tools without installing it one by one like for the "suite_stacks" tools for example.

Response:
In the forthcoming Galaxy release 17.09 it will be possible to install a workflow together with the needed tools, so we hope that this feature is going to solve this issue without having to create a “suite” repository.

[1]Galaxy with GeneSeqToFamily
http://repos.tgac.ac.uk/vms/Galaxy_with_GeneSeqToFamily.ova

Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes

Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

GeneSeqToFamily: the Ensembl Compara GeneTrees pipeline as a Galaxy workflow

Anil S. Thanki[1], Nicola Soranzo[1], Wilfried Haerty[1], Robert P. Davey[1]

1. Earlham Institute (EI), Norwich Research Park, Norwich NR4 7UZ, UK

Abstract

Background

Gene duplication is a major factor contributing to evolutionary novelty, and the contraction or expansion of gene families has often been associated with morphological, physiological and environmental adaptations. The study of homologous genes helps us to understand the evolution of gene families. It plays a vital role in finding ancestral gene duplication events as well as identifying genes that have diverged from a common ancestor under positive selection. There are various tools available, such as MSOAR, OrthoMCL and HomoloGene, to identify gene families and visualise syntenic information between species, providing an overview of syntenic regions evolution at the family level. Unfortunately, none of them provide information about structural changes within genes, such as the conservation of ancestral exon boundaries amongst multiple genomes. The Ensembl GeneTrees computational pipeline generates gene trees based on coding sequences and provides details about exon conservation, and is used in the Ensembl Compara project to discover gene families.

Findings

A certain amount of expertise is required to configure and run the Ensembl Compara GeneTrees pipeline via command line. Therefore, we have converted the command line Ensembl Compara GeneTrees pipeline into a Galaxy workflow, called GeneSeqToFamily, and provided additional functionality. This workflow uses existing tools from the Galaxy ToolShed, as well as providing additional wrappers and tools that are required to run the workflow.

Conclusions

GeneSeqToFamily represents the Ensembl Compara pipeline as a set of interconnected Galaxy tools, so they can be run interactively within the Galaxy's user-friendly workflow environment while still providing the flexibility to tailor the analysis by changing configurations and tools if necessary. Additional tools allow users to subsequently visualise the gene families produced by the workflow, using the Aequatus.js interactive tool, which has been developed as part of the Aequatus software project.

Keywords

Galaxy, Pipeline, Workflow, Genomics, Comparative Genomics, Homology, Orthology, Paralogy, Phylogeny, Gene Family, Alignment, Compara, Ensembl

Introduction

The phylogenetic information inferred from the study of homologous genes helps us to understand the evolution of gene families, which plays a vital role in finding ancestral gene duplication events as well as identifying regions under positive selection within species [1]. In order to investigate these low-level comparisons between gene families, the Ensembl Compara GeneTrees gene orthology and paralogy prediction software suite [2] was developed as a pipeline that uses TreeBest [3] [4] (part of TreeFam [5]) to find internal structural-level synteny for homologous genes. TreeBeST implements multiple independent phylogenetic methods and can merge their results in a consensus tree whilst trying to minimise duplications and deletions relative to a known species tree. This allows TreeBeST to take advantage of the fact that DNA-based methods are often more accurate for closely related parts of trees, while protein-based trees are better at longer distances.

The Ensembl GeneTrees pipeline comprises seven basic steps, starting from a set of protein sequences and performing similarity searching and multiple large-scale alignments to infer homology among them, using various tools: BLAST [6], hcluster_sg [7], T-Coffee [8], and phylogenetic tree construction tools, including TreeBeST. Whilst all these tools are freely available, most are specific to certain computing environments, are only usable via the command line, and require many dependencies to be fulfilled. Therefore, users are not always sufficiently expert in system administration in order to install, run, and debug the various tools at each stage in a chain of processes. To help ease the complexity of running the GeneTrees pipeline, we have employed the Galaxy bioinformatics analysis platform to relieve the burden of managing these system-level challenges.

Galaxy is an open-source framework for running a broad collection of bioinformatics tools via a user-friendly web interface [9]. No client software is required other than a recent web browser, and users are able to run tools singly or aggregated into interconnected pipelines, called *workflows*. Galaxy enables users to not only create, but also share workflows with the community. In this way, it helps users who have little or no bioinformatics expertise to run potentially complex pipelines in order to analyse their own data and interrogate results within a single online platform. Furthermore, pipelines can be published in a scientific paper or in a repository such as myExperiment [10] to encourage transparency and reproducibility.

In addition to analytical tools, Galaxy also contains plugins [11] for data visualisation. Galaxy visualisation plugins may be interactive and can be configured to visualise various data types, for example, bar plots, scatter plots, and phylogenetic trees. It is also possible to develop custom visualisation plugins and easily integrate them into Galaxy. As the output of the GeneSeqToFamily workflow is not conducive to human readability, we also provide a data-to-visualisation plugin based on the Aequatus software [12]. Aequatus.js [13] is a new JavaScript library for the visualisation of homologous genes, which is extracted from the standalone

Aequatus tool. It provides a detailed view of gene structure across gene families, including shared exon information within gene families alongside gene tree representations. It also shows details about the type of interrelation event that gave rise to the family, such as speciation, duplication, and gene splits.

Methods

The GeneSeqToFamily workflow has been developed to run the Ensembl Compara software suite within the Galaxy environment, combining various tools alongside preconfigured parameters obtained from the Ensembl Compara pipeline to produce gene trees. Among the tools used in GeneSeqToFamily (listed in Table 1), some were existing tools in the Galaxy ToolShed [14], such as NCBI BLAST, TranSeq, Tranalign and various format converters. Additional tools that are part of the pipeline were developed at the Earlham Institute (EI) and submitted to the ToolShed, i.e. *BLAST parser, hcluster_sg, hcluster_sg parser, T-Coffee, TreeBeST best* and *Gene Alignment and Family Aggregator.* Finally, we developed helper tools that are not part of the workflow itself, but aid the generation of input data for the workflow and these are also in the ToolShed, i.e. *Get features by Ensembl ID*, *Get sequences by Ensembl ID*, *Select longest CDS per gene*, *ETE species tree generator* and *GeneSeqToFamily preparation*.

The workflow comprises 7 main steps, starting with translation from input coding sequences (CDS) to protein sequences, finding subsequent pairwise alignments of those protein sequences using BLASTP, and then the generation of clusters from the alignments using hcluster_sg. The workflow then splits into two simultaneous paths, whereby in one path it performs the multiple sequence alignment (MSA) for each cluster using T-Coffee, and in the other, generates a gene tree with TreeBeST taking the cluster alignment and a species tree as input. Finally, these paths merge to aggregate the MSA, the gene tree and the gene feature information (transcripts, exons, and so on) into an SQLite [15] database for visualisation and downstream reuse. Each step of the workflow along with the data preparation steps is explained in detail below.

Figure 1: Overview of the GeneSeqToFamily workflow

Figure 2: Screenshot from the Galaxy Workflow Editor, showing the GeneSeqToFamily workflow

Table 1: Galaxy tools used in the workflows

Tool Name	**Tool ID**	**Version**	**Developed at EI**		**ToolShed Reference**
			Tool	**Wrapper**	

Get sequences by Ensembl ID	get_sequences	0.1.2	Yes	Yes	[16]
Get features by Ensembl ID	get_feature_info	0.1.2	Yes	Yes	[17]
Select longest CDS per gene	ensembl_longest_cds_per_gene	0.0.2	Yes	Yes	[18]
ETE species tree generator	ete_species_tree_generator	3.0.0b35	Yes	Yes	[19]
GeneSeqToFamily preparation	gstf_preparation	0.3.0	Yes	Yes	[20]
Transeq	EMBOSS: transeq101	5.0.0	No	No	[21]
NCBI BLAST+ makeblastdb	ncbi_makeblastdb	0.1.07	No	No	[22]
NCBI BLAST+ blastp	ncbi_blastp_wrapper	0.1.07	No	No	[22]
BLAST parser	blast_parser	0.1.1	Yes	Yes	[23]
hcluster_sg	hcluster_sg	0.5.1	No	Yes	[24]
hcluster_sg parser	hcluster_sg_parser	0.2.0	Yes	Yes	[25]
Filter by FASTA IDs	filter_by_fasta_ids	1.0	No	No	[26]
T-Coffee	t_coffee	11.0.8	No	Yes	[27]
Tranalign	EMBOSS: tranalign100	5.0.0	No	No	[21]
TreeBeST best	treebest_best	1.9.2	No	Yes	[28]
Gene Alignment and Family Aggregator	gafa	0.3.0	Yes	Yes	[29]
Unique	tp_sorted_uniq	1.1.0	No	No	[30]
FASTA-to-Tabular	fasta2tab	1.1.0	No	No	[31]
UniProt ID mapping and retrieval	uniprot_rest_interface	0.1	No	No	[32]

Data generation and preparation

We have developed a number of tools that assist in preparing the datasets needed by the workflows.

Ensembl REST API tools

Galaxy tools were developed which utilise the Ensembl REST API [33] to retrieve sequence information (*Get sequences by Ensembl ID*) and feature information (*Get features by Ensembl ID*) by Ensembl ID from the Ensembl service. REST (REpresentational State Transfer) is an architecture style for designing networked applications [34] which encourages the use of standardised HTTP technology to send and receive data between computers. As such, these tools are designed to help users to retrieve existing data from Ensembl rather than requiring them to manually download datasets to their own computers and then subsequently uploading them into the workflow.

We have also developed:

- *Select longest CDS per gene*, which filters a CDS FASTA file from Ensembl retaining only the longest CDS sequence for each gene
- *ETE species tree generator*, which uses the ETE toolkit [35] to generate a species tree from a list of species names or taxon IDs through NCBI Taxonomy.

GeneSeqToFamily workflow

0. GeneSeqToFamily preparation

Before GeneSeqToFamily can be run, a data preparation step must be carried out. We have developed a tool called *GeneSeqToFamily preparation* to preprocess the input datasets (gene feature information and CDS) for the GeneSeqToFamily workflow. It converts a set of gene feature information files in GFF3 [36] and/or JavaScript Object Notation (JSON) [37] format to an SQLite database. It also modifies all CDS FASTA header lines by appending the species name to the transcript identifier, as required by *TreeBeST best*. We decided to use an SQLite database to store the gene feature information because:

- the GFF3 format has a relatively inconvenient and unstructured additional information field (9th column)
- searching is much faster and more memory efficient in a database than in a text file like JSON or GFF3, especially when dealing with feature information for multiple large genomes

1. CDS translation

Transeq

Transeq, part of the European Molecular Biology Open Software Suite (EMBOSS) [38], is a tool to generate six-frame translation of nucleic acid sequences to their corresponding peptide sequences. Here we use Transeq to convert a CDS to protein sequences in order to run BLASTP and find protein clusters. However, since downstream tools in the pipeline such as TreeBeST require nucleotide sequences to generate a gene tree, the protein sequences cannot be directly used as workflow input and are instead generated with Transeq.

2. Pre-clustering alignment

BLAST

This workflow uses the BLAST wrappers [39] developed to run BLAST+ tools within Galaxy. BLASTP is run over the set of sequences against the database of the same input, as is the case with BLAST-all, in order to form clusters of related sequences.

BLAST parser

BLAST parser is a small Galaxy tool to convert the BLAST output into the input format required by hcluster_sg. It takes the BLAST 12-column output [40] as input and generates a 3-column tabular file, comprising the BLAST query, the hit result, and the edge weight. The weight value is simply calculated as minus $\log_{10}$ of the BLAST e-value, replacing this with 100 if this value is greater than 100. It also removes the self-matching BLAST results.

3. Cluster generation

hcluster_sg

hcluster_sg performs clustering for sparse graphs. It reads an input file that describes the similarity between two sequences, and iterates through the process of grouping two nearest nodes at each iteration. hcluster_sg outputs a single list of gene clusters, each comprising a set of sequence IDs present in that cluster. This list needs to be reformatted using the *hcluster_sg parser* tool in order to be suitable for input into T-Coffee and TreeBeST (see below).

hcluster_sg parser

hcluster_sg parser converts the hcluster_sg output into a collection of lists of IDs, one list for each cluster. Each of these clusters will then be used to generate a gene tree via TreeBeST. The tool can also filter out clusters with a number of elements outside a specified range. The IDs contained in all discarded clusters are collected in separate output dataset. Since TreeBeST requires at least 3 genes to generate a gene tree, we configured the tool to filter out clusters with less than 3 genes.

Filter by FASTA IDs, which is available from the Galaxy ToolShed, is used to create separate FASTA files using the sequence IDs listed in each gene cluster.

4. Cluster alignment

T-Coffee

T-Coffee is a MSA package, but can also be used to combine the output of other alignment methods (Clustal, MAFFT, Probcons, MUSCLE) into a single alignment. T-Coffee can align both nucleotide and protein sequences [8], and we use it to align the protein sequences in each cluster generated by hcluster_sg.

We modified the Galaxy wrapper for T-Coffee to take a single FASTA (as normal) and an optional list of FASTA IDs to filter. If a list of IDs is provided, the wrapper will pass only those sequences to T-Coffee, which will perform the MSA for that set of sequences, thus removing the need to create thousands of intermediate Galaxy datasets.

5. Gene tree construction

Tranalign

Tranalign [38] is a tool that reads a set of nucleotide sequences and a corresponding aligned set of protein sequences and returns a set of aligned nucleotide sequences. Here we use it to generate CDS alignments of gene sequences using the protein alignments produced by T-Coffee.

TreeBeST 'best'

TreeBeST (Tree Building guided by Species Tree) is a tool to generate, manipulate, and display phylogenetic trees and can be used to build gene trees based on a known species tree.

The 'best' command of TreeBeST builds 5 different gene trees from a FASTA alignment file using different phylogenetic algorithms, then merges them into a single consensus tree using a species tree as a reference. In GeneSeqToFamily, *TreeBeST best* uses the nucleotide MSAs generated by Tranalign (at least 3 sequences are required) and a user-supplied species tree in Newick format [41] (either produced by a third-party software or through the *ETE species tree generator* data preparation tool, described above) to produce a GeneTree for each family represented also in Newick format. The resulting GeneTree also includes useful annotations specifying phylogenetic information of events responsible for the presence/absence of genes, for example, 'S' means speciation event, 'D' means duplication, and 'DCS' denotes the duplication score.

6. Gene Alignment and Family Aggregation

Gene Alignment and Family Aggregator (GAFA)

GAFA is a Galaxy tool which generates a single SQLite database containing the gene trees and MSAs, along with gene features, in order to provide a reusable, persistent data store for visualisation of synteny information with Aequatus. *GAFA* requires:

- gene trees in Newick format,
- the protein MSAs in fasta_aln format from *T-Coffee* and
- gene feature information generated with the *GeneSeqToFamily preparation* tool.

Internally, *GAFA* converts each MSA from fasta_aln format to a simple CIGAR string [42]. An example of CIGAR strings for aligned sequences is shown in Figure 3, in which each CIGAR string subset changes according to other sequences.

The simple schema [43] for the generated SQLite database is shown in Figure 4.

Figure 3: Showing how CIGAR for multiple sequence alignment is generated

Figure 4: Schema of the GAFA SQLite database

7. Visualisation

Aequatus visualisation plugin

The SQLite database generated by the GAFA tool can be rendered using a new visualisation plugin, Aequatus.js. The Aequatus.js library, developed at EI as part of the Aequatus project, has been configured to be used within Galaxy to visualise homologous gene structure and gene family relationships. This allows users to interrogate not only the evolutionary history of the gene family but also the structural variation (exon gain/loss) within genes across the phylogeny. Aequatus.js is available to download from GitHub [43], as visualisation plugins cannot yet be submitted to the Galaxy ToolShed.

Finding homology information for orphan genes

Although the GeneSeqToFamily workflow will assign most of the genes to orthogroups, many genes within a species might appear to be unique without homologous relationship to any other genes from other species. This observation could be the consequence of the parameters selected, choice of species or incomplete annotations. This could also reflect real absence of homology such as for rapidly evolving gene families. In addition to the GeneSeqToFamily workflow, we also developed two associated workflows to further annotate these genes by:

1) Retrieving a list of orphan genes from the GeneSeqToFamily workflow (see Figure 5) as follows:
    a) Find the IDs of the sequences present in the input CDS of the GeneSeqToFamily workflow, but not in the result of *BLAST parser* from the same workflow
    b) Add to this list the IDs of the sequences discarded by *hcluster_sg parser*
    c) From the input CDS dataset, retrieve the respective sequence for each CDS ID (from the step above) using *Filter by FASTA IDs*

   These unique CDS can be fed into the SwissProt workflow below to find homologous genes in other species.
2) Finding homologous genes for some genes of interest using SwissProt (see Figure 6) as follows:
    a) Translate CDS into protein sequences using *Transeq*
    b) Run BLASTP for the protein sequences against the SwissProt database (from NCBI)
    c) Extract UniProt IDs from these BLASTP results, using the preinstalled Galaxy tool *Cut columns from a table* (tool id *Cut1*)
    d) Retrieve Ensembl IDs (representing genes and/or transcripts) for each UniProt ID using *UniProt ID mapping and retrieval*
    e) Get genomic information for each gene ID and CDS for each transcript ID from the core Ensembl database using *Get features by Ensembl ID* and *Get sequences by Ensembl ID* respectively.

   The results from this second workflow can be subsequently used as input to GeneSeqToFamily for familial analysis.

Figure 5: Screenshot from the Galaxy Workflow Editor, showing the orphan gene finding workflow

Figure 6: Screenshot from the Galaxy Workflow Editor, showing the SwissProt workflow

Example use cases

The main aim of our workflow is to allow the reliable and reproducible identification of gene families across a defined set of species. To assess the robustness of our workflow but also to set a benchmark and find optimum parameters to run the workflow, we analysed a large dataset of CDS from six mammal species (platypus, pig, horse, human, mouse and dog). Since BLASTP plays a crucial role in our analyses, we primarily focuses on the associated parameters.

We have run our workflow on two sets of three species, one including high quality assemblies (human, mouse and dog), the second set composed of draft genome assemblies only (pig, platypus and horse). We also ran our analysis on all six species to assess the impact of mixed assemblies qualities on the identification of gene families.

We downloaded the CDS sequences for all species from Ensembl (release 89) and filtered them to retain only the longest transcript per gene (as in the Ensembl Compara pipeline), obtaining a total of 125,342 CDS. We then ran the GeneSeqToFamily workflow on each dataset using various BLASTP parameters (as shown in Table 3), in order to identify the optimal values for the workflow to generate a gene tree in which members are possibly evolved from a single ancestor gene, and usually with identical biochemical functions.

Our results show that the number of gene families can vary quite distinctly with different BLASTP parameters. Stringent parameters (Analysis 6) result in a large number of smaller families, while relaxed parameters (Analysis 1) generate a smaller number of large families, which may include distantly related genes. By testing different parameters and comparing the analyses with third party tools such as PantherDB to validate the results against known families, we chose those parameters listed as Analysis 5. These values seem to consistently generate legitimate sets of gene families with closely related family members based on the datasets we tested.

From these results, we can also observe that the quality of the genome annotation affects the classification of genes into gene families. Genomes with high quality annotation contains fewer orphan genes compared to genomes with low quality annotation. But when genes from genomes with low quality annotation are mixed with well annotated genes, there are better chances of classifying them in a gene family, thus leaving fewer orphan genes.

There are caveats, however. BLASTP parameters used in Analysis 5 restrict the maximum number of target sequences per query (max_target_seqs) to 3 (the first hit when using all-versus-all BLAST will always be the query sequence itself). The minimum query coverage per High-scoring Segment Pair (HSP) (qcovhsp) is set to 90% and e-value cut-off to 1e-10, in order to find the HSP closest to the query thus allowing partial matches which could be seen in the event of gene split. If the input CDS contain alternative transcripts per gene, setting a value of 3 for the max_target_seqs parameter will restrict the search to only 3 matches per query, and the presence of alternative transcripts will decrease the likelihood of finding matches from other genes. Therefore we recommend to increase the max_target_seqs parameter to 4, so each gene will have better chances of finding also a match with other genes rather than isoforms.

Table 2: Results of the GeneSeqToFamily workflow run on 125,342 CDS from 6 species using 6 different BLAST parameter configurations, the complete list of which are shown in Table 3.

Summary						
Analysis	**1**	**2**	**3**	**4**	**5**	**6**
No of families	8,780	19,324	19,323	18,936	19,185	23,235
Filtered out (>200)	31	3	3	3	3	2
Filtered out (<3)	1,497	1,360	1,166	972	1,028	3,916
Filtered families to consider	7,252	17,961	18,154	17,961	18,154	19,317
Orphan Genes (percentage in whole input CDS)	2,914 (2.32%)	2,943 (2.35%)	13,034 (10.40%)	2,946 (2.35%)	13,041 (10.40%)	14,492 (11.56%)

Table 2.1: Results of the GeneSeqToFamily workflow run on 63,772 CDS from 3 species with low genome quality using 6 different BLAST parameter configurations, the complete list of which are shown in Table 3.

Summary						
Analysis	**1**	**2**	**3**	**4**	**5**	**6**
No of families	8,211	13,159	13,887	12,831	13,767	17,361
Filtered out (>200)	11	3	3	3	3	2
Filtered out (<3)	2,287	2,285	3,261	1,956	3,1403	6,003
Filtered families to consider	5,911	10,871	10,623	10,872	10,624	11,356
Orphan Genes (percentage in whole input CDS)	1,925 (3.02%)	1,947 (3.05%)	11,983 (18.79%)	1,948 (3.05%)	11,992 (18.80%)	13,789 (21.62%)

Table 2.2: Results of the GeneSeqToFamily workflow run on 61,570 CDS from 3 species high quality genome using 6 different BLAST parameter configurations, the complete list of which are shown in Table 3.

Summary						
Analysis	**1**	**2**	**3**	**4**	**5**	**6**
No of families	8,214	17,201	17,275	16,940	17,179	19,012
Filtered out (>200)	11	1	1	1	1	0
Filtered out (<3)	1,248	1,113	1,992	851	1,899	4,163
Filtered families to consider	6,955	16,087	15,282	16,088	15,279	14,849
Orphan Genes (percentage in whole input CDS)	1,857 (3.02%)	1,862 (3.02%)	6,795 (11.04%)	1,864 (3.03%)	6,802 (11.05%)	7,502 (12.18%)

Table 3: Complete list of BLAST parameter configurations. Analysis 5 is highlighted to denote those parameters that were chosen to be used as workflow defaults.

Analysis ID	**e-value**	**Max targets per query**	**Min coverage**
1	1e-03 (Default)	0 (Default)	0 (Default)
2	1e-03	3	0
3	1e-03	3	90%
4	1e-10	3	0
5	1e-10	3	90%
6	1e-10	2	90%

To validate the biological relevance of results from the GeneSeqToFamily workflow, we analysed a smaller set of 23 homologous genes (39 transcripts) from *Pan troglodytes* (chimpanzee), *Homo sapiens* (human), *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Sus scrofa* (pig) and *Canis familiaris* (domesticated dog). These genes are a combination of those found in four gene families, i.e. monoamine oxidases (MAO) A and B, insulin receptor (INSR), BRCA1-associated ATM activator 1 (BRAT1), and were chosen because they are present in all 6 species yet distinct from each other. Though MAO gene variants (A and B) are 70% similar, a single gene tree for all MAO genes could be generated if appropriate parameters are not selected. As such, these genes represent a reliable dataset to test whether the GenSeqToFamily workflow can reproduce already known gene families.

Before running the workflow, feature information and CDS for the selected genes were retrieved from the core Ensembl database using the helper tools described above (*Get features by Ensembl ID* and *Get sequences by Ensembl ID* respectively). A species tree was generated using *ETE species tree generator* and CDS were prepared with *GeneSeqToFamily preparation*. We ran the GeneSeqToFamily workflow on these data using the parameters shown in Analysis 5 of Table 3, but we set max_target_seqs as 4 (as described in the previous use case) to get a wider range of results from BLAST because our dataset includes alternative transcripts. This workflow generated 4 different gene trees, one for each gene family. Figure 7, 8, 9 and 10 show the resulting gene trees for MAOA, MAOB, BRAT1 and INSR gene families. Different colours of the nodes in each gene tree on the left-hand-side highlight potential evolutionary events, such as speciation, duplication, and gene splits. Homologous genes showing shared exons use the same colour in each representation, including insertions (black blocks) and deletions (red lines). The GeneTrees for these genes are already available in Ensembl and we used them to validate our findings [44] [45] [46] [47]. Our gene trees exactly matched the Ensembl GeneTrees, showing that the workflow generates biologically valid results. We have provided the underlying data for this example along with the submitted workflow in figshare [48].

Figure 7: Homologous genes of MAOA of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus and Sus scrofa.*

Figure 8: Homologous genes of MAOB of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus and Sus scrofa.*

Figure 9: Homologous genes of BRAT1 of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus and Sus scrofa*.

Figure 10: Homologous genes of INSR of *Rattus norvegicus* from *Rattus norvegicus, Mus musculus, Pan troglodytes, Homo sapiens and Sus scrofa*.

Conclusion

The ultimate goal of the GeneSeqToFamily is to provide a user-friendly workflow to analyse and discover homologous genes using the Ensembl Compara GeneTrees pipeline within the Galaxy framework, where users can interrogate genes of interest without using the command-line whilst still providing the flexibility to tailor analysis by changing configurations and tools if necessary. We have shown it to be an accurate, robust, and reusable method to elucidate and analyse potentially large numbers of gene families in a range of model and non-model organisms. The workflow stores the resulting gene families into a SQLite database, which can be visualised using the Aequatus.js interactive tool, as well as shared as a complete reproducible container for potentially large gene family datasets.

Gradually, we hope that the Galaxy community will undertake their own analyses and feedback improvements to various tools, and publish successful combinations of parameters used in the GeneSeqToFamily workflow. We encourage this process by allowing users to share their own version of GeneSeqToFamily workflow for appraisal by the community.

Future directions

In terms of core workflow functionality, we would like to incorporate pairwise alignment between pairs of genes for closely related species in addition of the MSA for the gene family, which will help users to compare orthologs and paralogs in greater detail.

We also plan to explicitly include the PantherDB resources [49]. Protein ANalysis Through Evolutionary Relationships (PANTHER) is a classification system to characterise known proteins and genes according to family, molecular function, biological process and pathway. The integration of PantherDB with GeneSeqToFamily will enable the automation of gene family validation and add supplementary information about those gene families, which could in turn be used to further validate novel genomics annotation.

Finally we intend to add the ability to query the *GAFA* SQLite database using keywords, to make it easy for users to find gene trees which include their genes of interest without needing to delve into the database itself.

Availability and requirements

Project name: GeneSeqToFamily
Project home page: https://github.com/TGAC/earlham-galaxytools/tree/master/workflows/GeneSeqToFamily
Archived version: 0.1.0
Operating system(s): Platform independent
Programming language: JavaScript, Perl, Python, XML, SQL
Other Requirements: Web Browser; for development: Galaxy
Any restrictions to use by non-academics: None
License: The MIT License

Availability of supporting data

The example files and additional data sets supporting the results of this article are available in figshare [48]. A virtual image for Galaxy with necessary tools and installed workflow also available at Earlham repos [50].

Acknowledgements

AT, WH and RPD are supported by BBSRC institute strategic programme grant funds awarded to EI. NS is funded under the BBSRC Biomathematics and Bioinformatics Training fund (2014) awarded to EI. This research was supported in part by the NBI Computing infrastructure for Science (CiS) group who provide technical support and maintenance to EI's High Performance Computing cluster and storage systems, enabling us to develop this workflow.

We would like to thank Matthieu Muffato from the European Bioinformatics Institute (EBI) for his advice during the initial stage of the project.

References

1. Jensen JD, Wong A, Aquadro CF. Approaches for identifying targets of positive selection. Trends Genet. 2007;23:568–77.

2. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2008;19:327–35.

3. Ensembl. Ensembl/treebest. GitHub. https://github.com/Ensembl/treebest. Accessed 26 Jan 2016.

4. Heng L. Constructing the TreeFam database. The Institute of Theoretical Physics, Chinese Academic of Science; 2006. http://pfigshare-u-files.s3.amazonaws.com/1421613/PhDthesisliheng2006English.pdf.

5. Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, et al. TreeFam: 2008 Update. Nucleic Acids Res. 2008;36 Database issue:D735–40.

6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

7. Li H et al. hcluster_sg: hierarchical clustering software for sparse graphs. https://github.com/douglasgscofield/hcluster. Accessed 26 Jan 2016.

8. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302:205–17.

9. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44:W3–10.

10. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, et al. myExperiment: a repository and social network for the sharing of bioinformatics workflows.

Nucleic Acids Res. 2010;38 Web Server issue:W677–82.

11. Goecks J, Eberhard C, Too T, Galaxy Team, Nekrutenko A, Taylor J. Web-based visual analysis for high-throughput genomics. BMC Genomics. 2013;14:397.

12. Thanki AS, Ayling S, Herrero J, Davey RP. Aequatus: An open-source homology browser. bioRxiv. 2016;:055632. doi:10.1101/055632.

13. TGAC. TGAC/aequatus.js. GitHub. https://github.com/TGAC/aequatus.js. Accessed 26 Jan 2016.

14. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, et al. Dissemination of scientific software with Galaxy ToolShed. Genome Biol. 2014;15:403.

15. SQLite Home Page. https://www.sqlite.org/. Accessed 18 Nov 2016.

16. Get sequences by Ensembl ID : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/ensembl_get_sequences/. Accessed 20 Dec 2016.

17. Get features by Ensembl ID : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/ensembl_get_feature_info/. Accessed 20 Dec 2016.

18. Select longest CDS per gene : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/ensembl_longest_cds_per_gene/. Accessed 8 Mar 2017.

19. ETE species tree generator : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/ete/. Accessed 20 Dec 2016.

20. GeneSeqToFamily preparation : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/gstf_preparation/. Accessed 17 Mar 2017.

21. EMBOSS : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/devteam/emboss_5/. Accessed 21 Dec 2016.

22. NCBI BLAST plus : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/devteam/ncbi_blast_plus. Accessed 21 Dec 2016.

23. BLAST parser : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/blast_parser/. Accessed 20 Dec 2016.

24. hcluster_sg : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/hcluster_sg/. Accessed 20 Dec 2016.

25. hcluster_sg parser : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/hcluster_sg_parser/. Accessed 20 Dec 2016.

26. Filter by FASTA IDs : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/galaxyp/filter_by_fasta_ids/. Accessed 21 Dec 2016.

27. T-Coffee : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/t_coffee/. Accessed 20 Dec 2016.

28. TreeBeST best : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/treebest_best/. Accessed 20 Dec 2016.

29. Gene Align and Family Aggregator (GAFA) : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/earlhaminst/gafa/. Accessed 21 Dec 2016.

30. text_processing : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/bgruening/text_processing/. Accessed 19 Apr 2017.

31. FASTA-to-Tabular converter : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/devteam/fasta_to_tabular/. Accessed 19 Apr 2017.

32. uniprot_rest_interface : Galaxy Tool Shed. https://toolshed.g2.bx.psu.edu/view/bgruening/uniprot_rest_interface/. Accessed 20 Mar 2017.

33. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, et al. The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics. 2015;31:143–5.

34. Representational State Transfer. http://www.peej.co.uk/articles/rest.html. Accessed 4 Feb 2016.

35. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016. doi:10.1093/molbev/msw046.

36. GFF3 - GMOD. http://gmod.org/wiki/GFF3. Accessed 4 Feb 2016.

37. JSON. http://www.json.org. Accessed 4 Feb 2016.

38. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.

39. Cock PJA, Chilton JM, Grüning B, Johnson JE, Soranzo N. NCBI BLAST+ integrated into Galaxy. Gigascience. 2015;4:39.

40. National Center for Biotechnology Information (U.S.), Camacho C. BLAST(r) Command Line Applications User Manual. 2008.

41. “Newick’s 8:45” Tree Format Standard. http://evolution.genetics.washington.edu/phylip/newick_doc.html. Accessed 8 Apr 2016.

42. Sequence Alignment/Map Format Specification. http://samtools.github.io/hts-specs/SAMv1.pdf. Accessed 20 Dec 2016.

43. TGAC. TGAC/earlham-galaxytools. GitHub. https://github.com/TGAC/earlham-galaxytools. Accessed 21 Mar 2016.

44. Gene: BRAT1 (ENSG00000106009) - Gene tree - Homo sapiens - Ensembl genome browser 87. http://dec2016.archive.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG0000010600

9;r=7:2537877-2555727; Accessed 23 Dec 2016.

45. Gene: INSR (ENSG00000171105) - Gene tree - Homo sapiens - Ensembl genome browser 87. http://dec2016.archive.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG00000171105;r=19:7112255-7294034; Accessed 23 Dec 2016.

46. Gene: MAOA (ENSG00000189221) - Gene tree - Homo sapiens - Ensembl genome browser 87. http://dec2016.archive.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG00000189221;r=X:43654907-43746824; Accessed 23 Dec 2016.

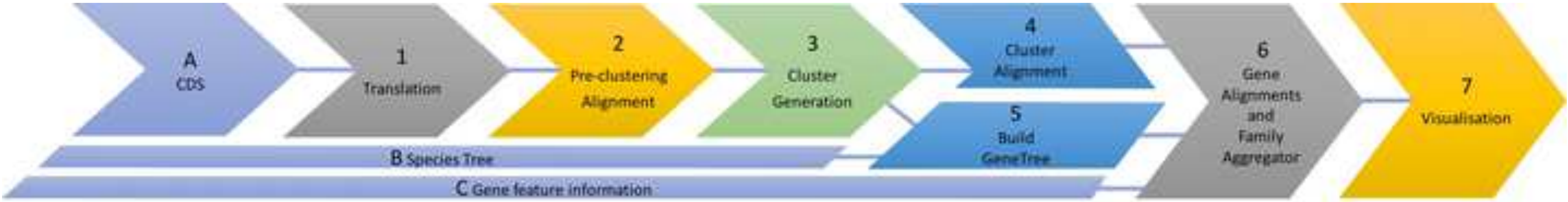
47. Gene: MAOB (ENSG00000069535) - Gene tree - Homo sapiens - Ensembl genome browser 87. http://dec2016.archive.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG00000069535;r=X:43766611-43882447; Accessed 23 Dec 2016.

48. Thanki AS, Soranzo N, Haerty W, Davey R. GeneSeqToFamily.zip. 2017. doi:10.6084/m9.figshare.4484141.v14.

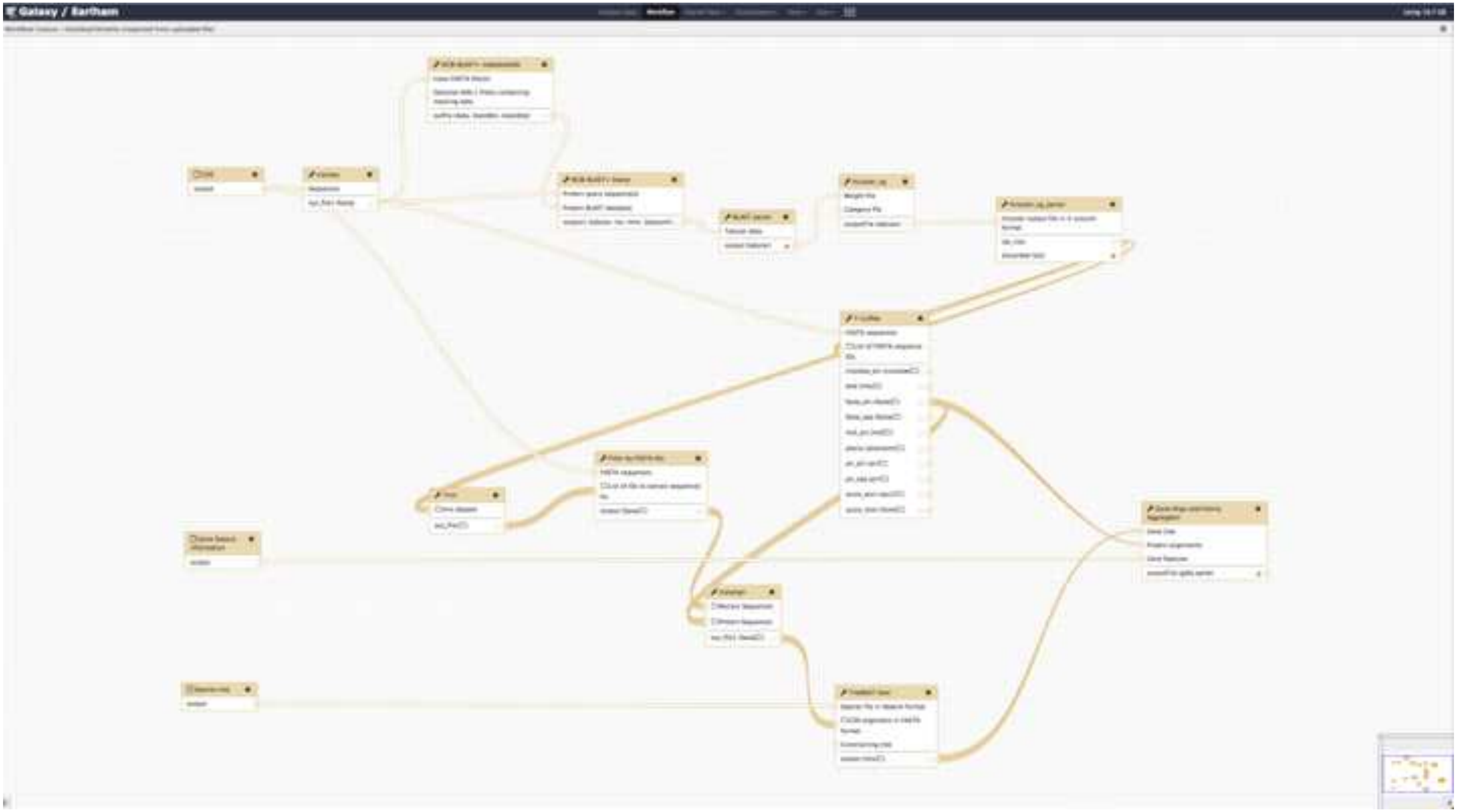
49. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, et al. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 2005;33 Database issue:D284–8.

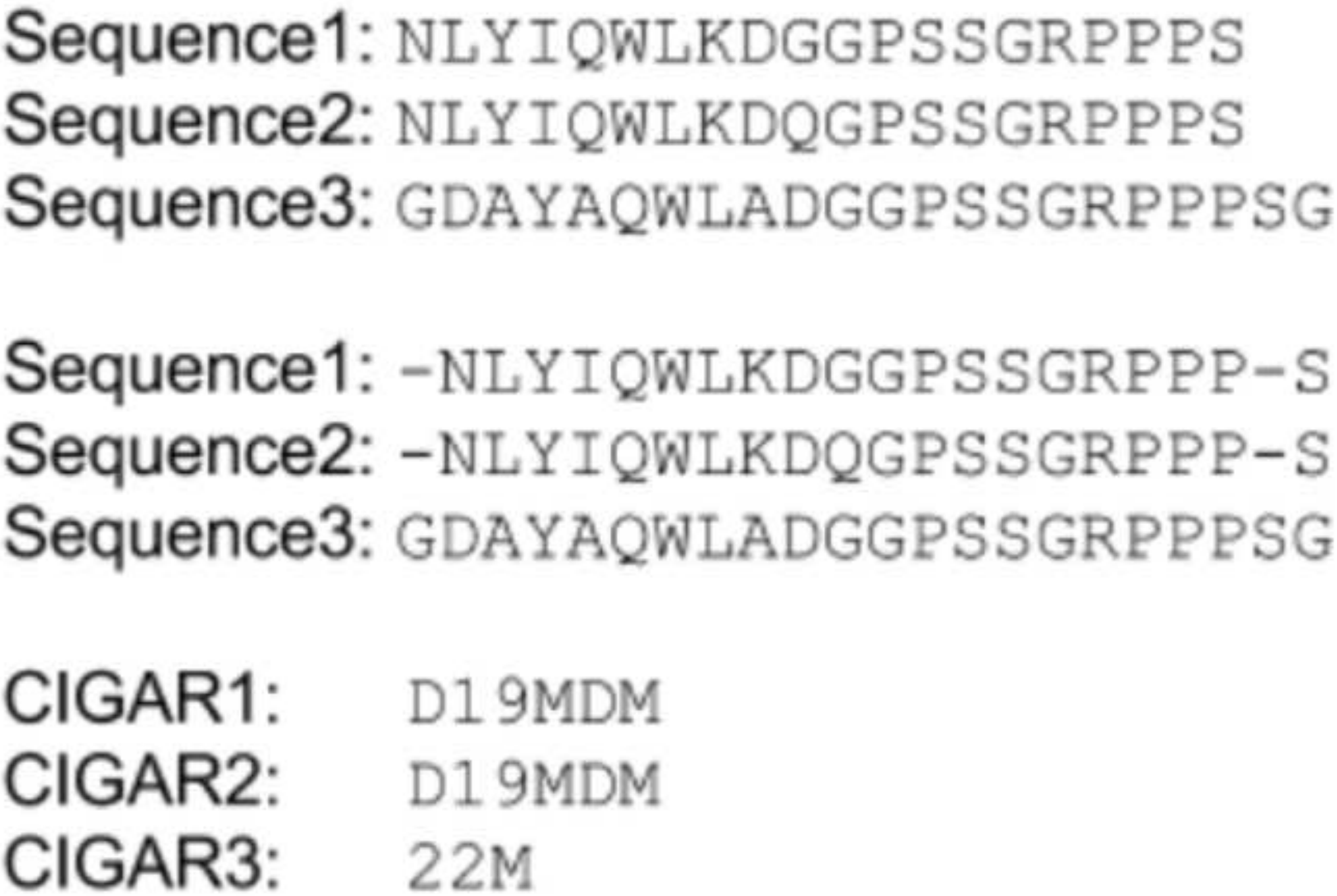
50. Galaxy Virtual Image. http://repos.tgac.ac.uk/vms/Galaxy_with_GeneSeqToFamily.ova. Accessed 28 Jul 2017.

Click here to download Figure 1.JPG



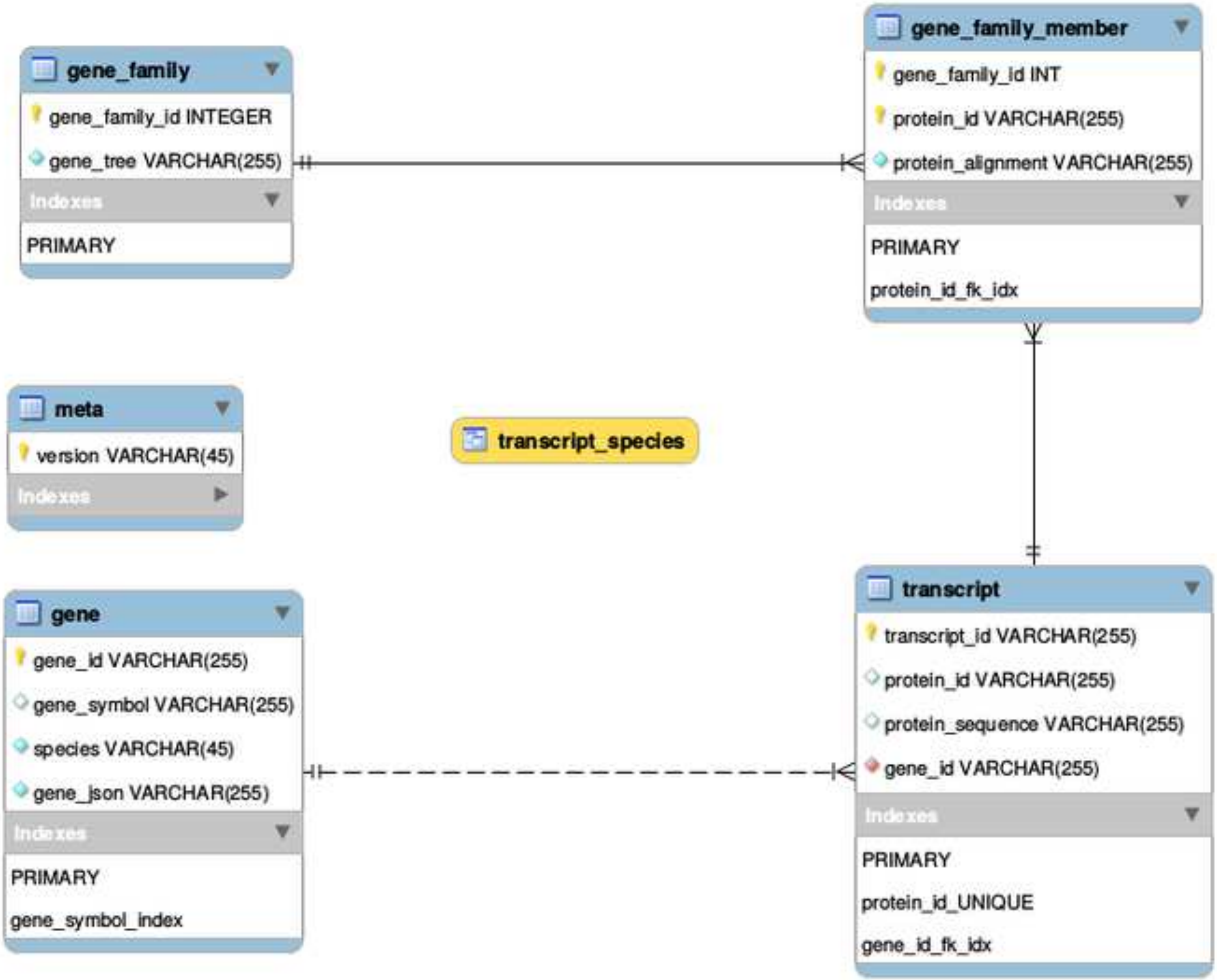
Click here to download Figure 2.png



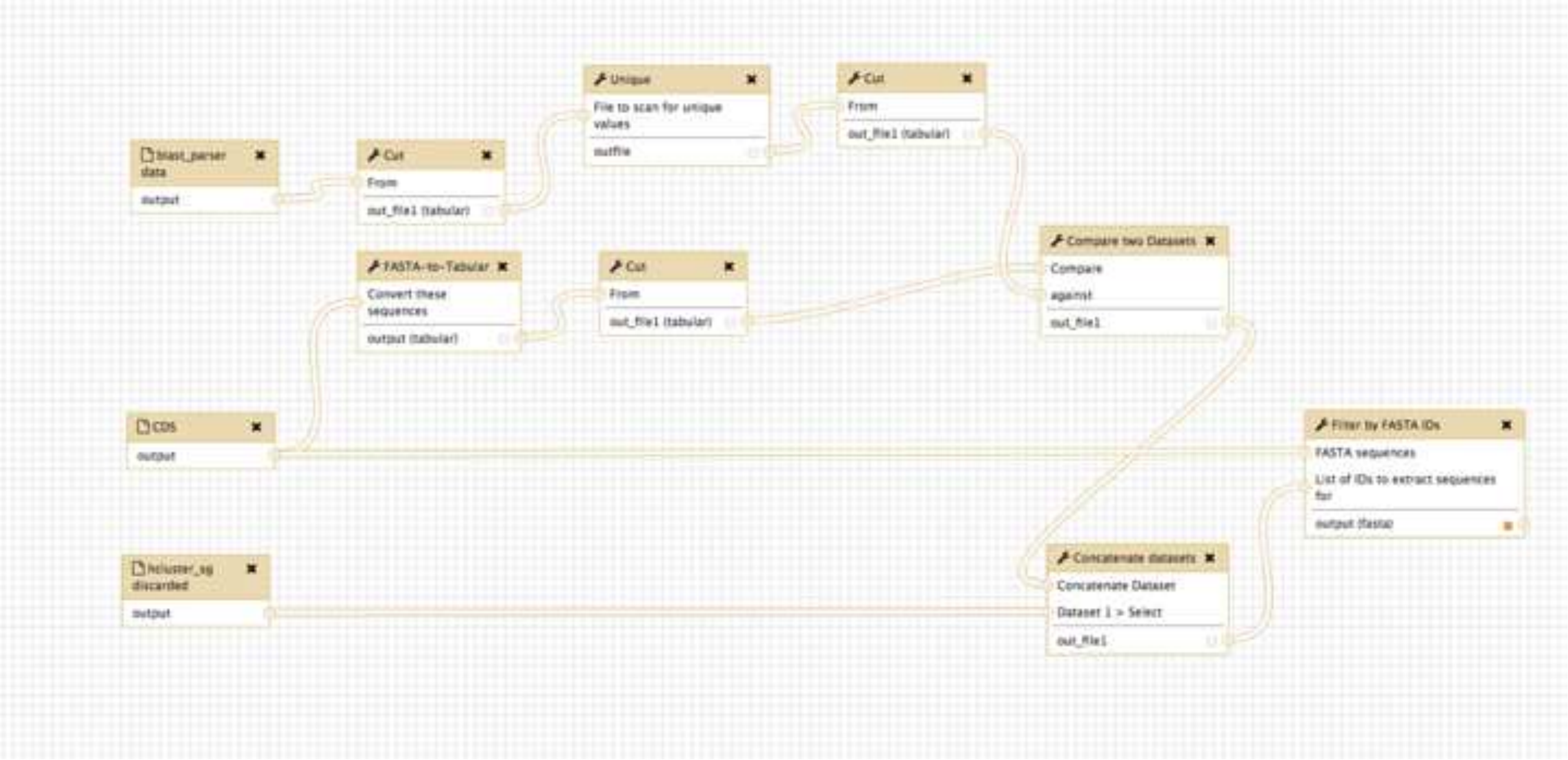


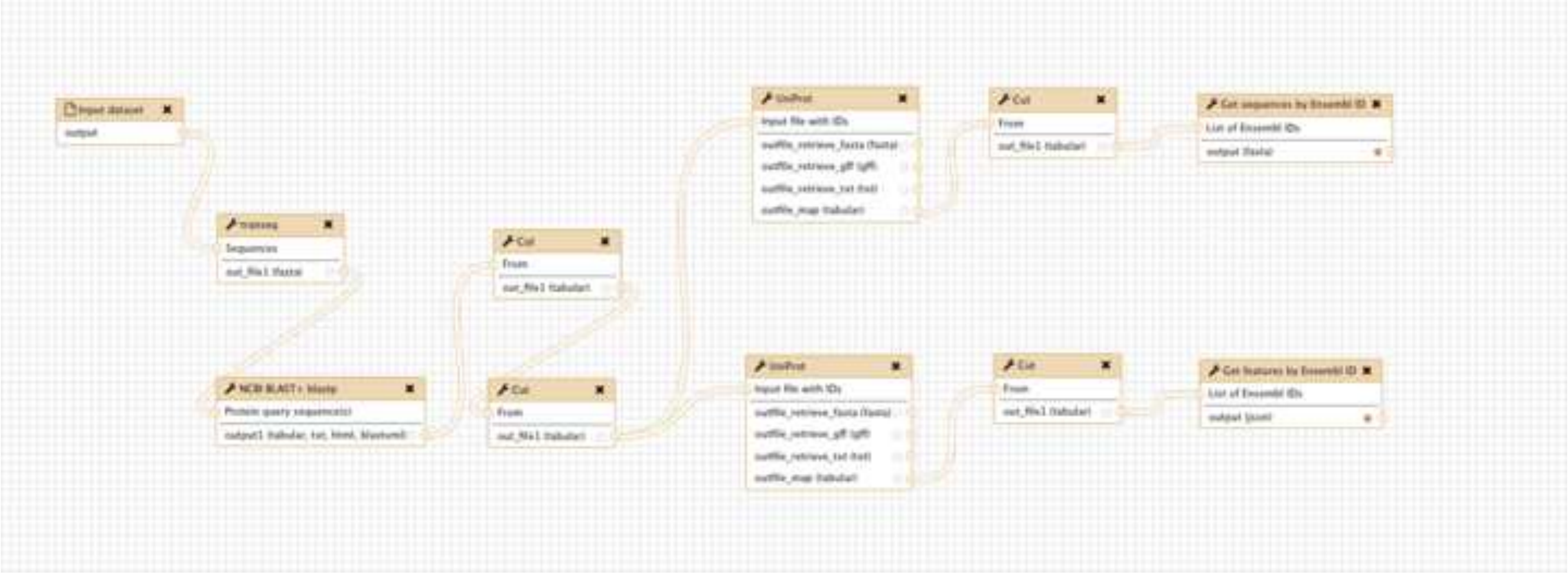
```
Sequence1: NLYIQWLKDGGPSSGRPPPS
Sequence2: NLYIQWLKDQGPSSGRPPPS
Sequence3: GDAYAQWLADGGPSSGRPPPSG

Sequence1: -NLYIQWLKDGGPSSGRPPP-S
Sequence2: -NLYIQWLKDQGPSSGRPPP-S
Sequence3: GDAYAQWLADGGPSSGRPPPSG

CIGAR1:     D19MDM
CIGAR2:     D19MDM
CIGAR3:     22M
```

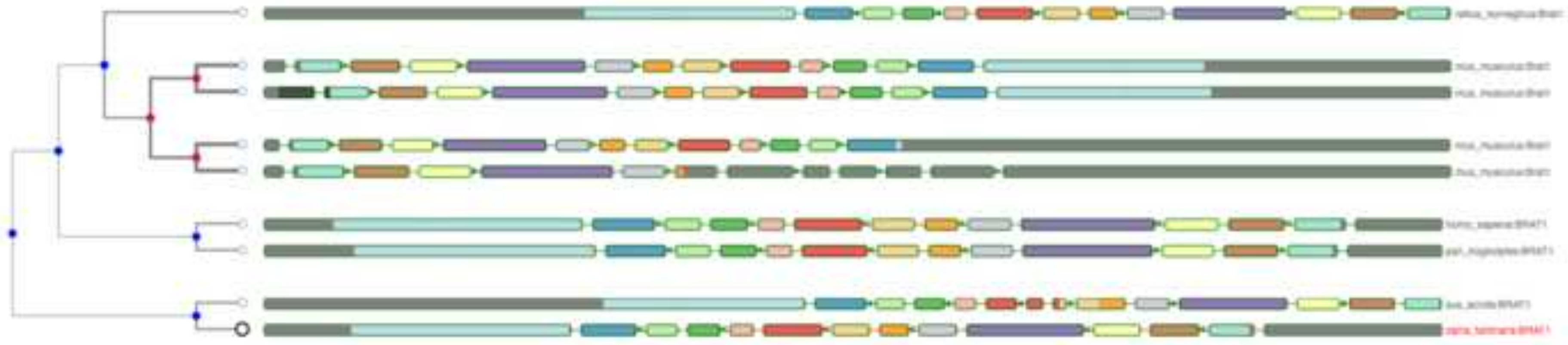
Click here to download Figure 4.png

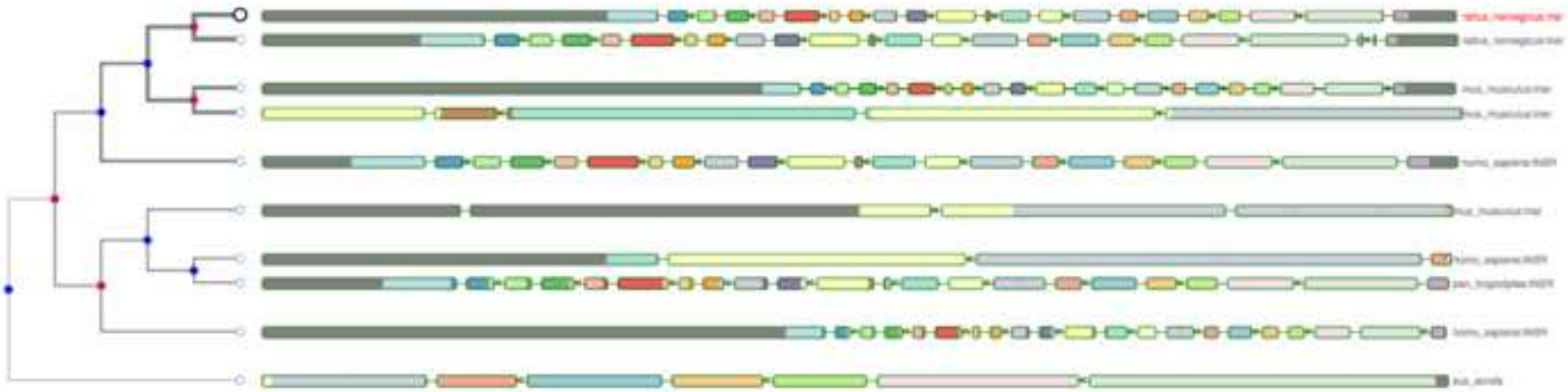


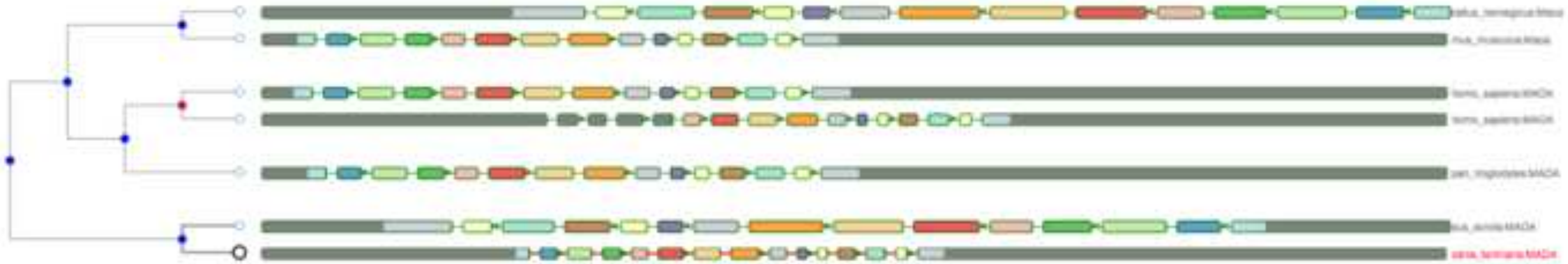
Click here to download Figure 5.png



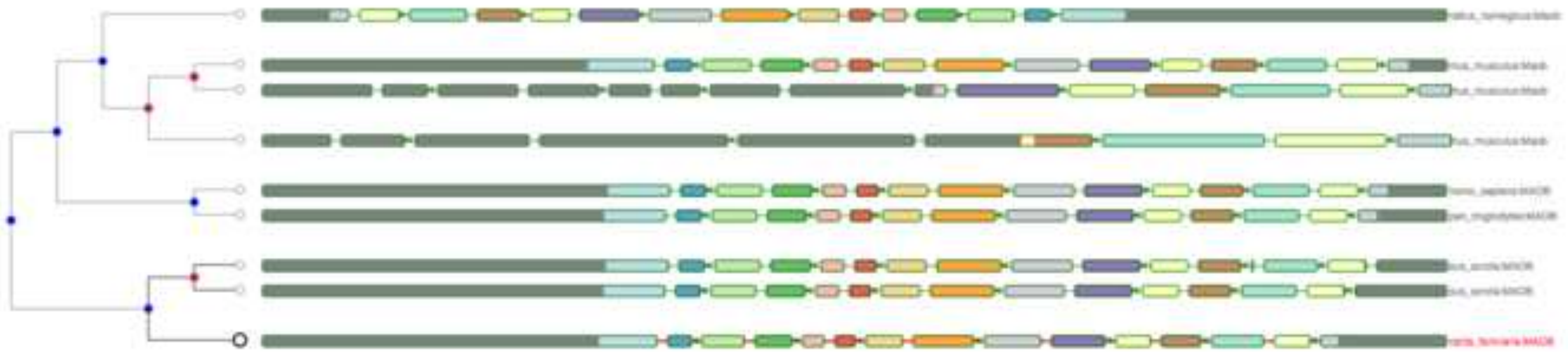

Input dataset
output
transeq
Sequences
out_file1 (fasta)
NCBI BLAST+ blastp
Protein query sequence(s)
output1 (tabular, txt, html, blastxml)
Cut
From
out_file1 (tabular)
Cut
From
out_file1 (tabular)
UniProt
Input file with IDs
outfile_retrieve_fasta (fasta)
outfile_retrieve_gff (gff)
outfile_retrieve_txt (txt)
outfile_map (tabular)
UniProt
Input file with IDs
outfile_retrieve_fasta (fasta)
outfile_retrieve_gff (gff)
outfile_retrieve_txt (txt)
outfile_map (tabular)
Cut
From
out_file1 (tabular)
Cut
From
out_file1 (tabular)
Get sequences by Ensembl ID
List of Ensembl IDs
output (fasta)
Get features by Ensembl ID
List of Ensembl IDs
output (json)


rattus_norvegicus:Brat1
mus_musculus:Brat1
mus_musculus:Brat1
mus_musculus:Brat1
mus_musculus:Brat1
homo_sapiens:BRAT1
pan_troglodytes:BRAT1
sus_scrofa:BRAT1
capra_hircus:BRAT1


rattus_norvegicus Insr
rattus_norvegicus Insr
mus_musculus Insr
mus_musculus Insr
homo_sapiens INSR
mus_musculus Insr
homo_sapiens INSR
pan_troglodytes INSR
homo_sapiens INSR
sus_scrofa


rattus_norvegicus.Mtacd
mus_musculus.Mtacd
homo_sapiens.MADA
homo_sapiens.MADA
pan_troglodytes.MADA
sus_scrofa.MADA
rana_temporaria.MADA

Click here to download Figure 8.png



Anil S. Thanki

Earlham Institute
Norwich Research Park
Norwich
NR4 7UZ
UK

31st July 2017

Dear Editor,

We wish to submit a revised manuscript entitled “**GeneSeqToFamily: the Ensembl Compara GeneTrees pipeline as a Galaxy workflow”** for re-consideration by GigaScience.

We appreciated the constructive criticisms of the reviewers. We have addressed each of their concerns as outlined below.

Please address all correspondence concerning this manuscript to me at Anil.Thanki@earlham.ac.uk or Robert Davey at Robert.Davey@earlham.ac.uk .

Thank you for your consideration of this manuscript.

Sincerely,

Anil S. Thanki

Reviewer #1:

The authors have put together a Galaxy workflow incorporating various tools to be able to perform Ensembl Compara's pipeline for gene family identification and synteny visualization. This in itself is valuable and useful and I have no concerns with the pipeline itself, though I have not tried it.

However, I am unsatisfied with the parameter evaluation and benchmarking:

1 - First, to test the tool, the objective must be clearer defined. Is the goal to find homologs most generally? Full-length homologs? Orthologs? And if orthologs, then orthologs relative to the set of species used as input or something else? This must be explicitly stated for the evaluation and parameter search to be well-defined. The manuscript should also make clear that this goal is the one that utility for reaching is implied, instead of any other goal.

Response:
The main aim of the test is to demonstrate that we are able to identify orthogroups (groups of genes sharing ancestry by descent) across a set of species. In the revised version of the manuscript, we clarified the goals of the tool and its purpose.

2 - The data set used for parameter searches is tiny, just three species, two of which are obscure (how good are assemblies?). This should be done on a much larger set, perhaps use the Quest for Orthologs' reference dataset?

Response:
Although we only used three species, the data set includes over 63,000 genes. The reviewer raises a valid question regarding the quality of the assemblies, i.e. while the mouse genome can be considered of very high quality, both the platypus and Tasmanian devil genome assemblies are of much lower quality. Therefore, we have now run our workflow on two sets of three species, one including high quality assemblies (human, mouse and dog), the second set composed of draft genome assemblies only (pig, platypus and horse). Furthermore, we ran our analysis on all six species to assess the impact of mixed assemblies qualities on the identification of gene families.

3 - Testing six settings does not show that one of the settings consistently give any particular results. The way to do that would be to test each parameter set many times, using different species combinations each time, and check if indeed Analysis 5 parameters consistently yield desired results under a well-defined quality metric (dependent on the goal posts set above).

Response:
As per reviewer's request, we tested six sets of parameters on the high quality, low quality, and both high and low quality datasets mentioned above. All datasets tested confirm that Analysis 5 returns the optimal results for these datasets. We acknowledge that different datasets might require further tuning, but we believe these parameters to be suitably comprehensive for most analyses.

4 - Testing performance on 23 genes is vastly insufficient. Please look into much larger benchmark sets (e.g. OrthoBench or quest for orthologs' benchmark tool), choose some larger, more diverse (functionally and phylogenetically) set to use as a gold standard, define a metric for success, then show how well the tool works under repeated applications of method settings on different subsets of the data.

Response:
We tried running our workflow with ‘Quest for orthologs’, but the recommended dataset available for this tool is very old (2011). We contacted the author of Quest for Orthologs for the latest release (2017), but there is no gold standard resultant data available to compare to, so maybe we can look into benchmarking this workflow in the future when results from other tools are available to compare with.

5 To be clear, I think this is valuable and worthy of publication, but for any claim of functionality, it must be robustly tested under clear definitions on large-scale diverse and representative data as stated above (which should also have some clear measure of accuracy relative to the gold standard).

Response:
We thank the reviewer for their comments, and we hope that the revised manuscript and associated broader analyses will help readers decide whether this workflow is suitable for their datasets. We feel that the mechanisms that the reviewer kindly suggested for assessing our workflow against a gold standard are inadequate to give a representative assessment. We will continue to monitor the available tools and standards in order to provide support for the usefulness of the workflow.

6 Similarly, please add time and memory consumption statistics for different use cases.

Response:
Time and memory consumption for the workflow depends on the infrastructure it is running on. We ran the Galaxy jobs on an High Performance Computing cluster, where measuring the amount of memory used is tricky. Also, the total time used to complete the pipeline may heavily vary depending on the time spent in the cluster scheduler waiting for resources to be available.

Anyhow, most of the time is spent on the NCBI blastp step, which for the examples in the paper varied between 16 and 24 hours for 6 species with 125,342 sequences.

Reviewer #2:

The article presents GeneSeqToFamily, the Ensembl Compara GeneTrees pipeline as a Galaxy workflow. Authors introduce the importance of studying homologous genes to understand evolution of gene families. They present the Ensembl Compara GeneTrees pipeline and the resulting conversion into a Galaxy workflow named GeneSeqToFamily. This workflow is thus the first "ready-to-use" solution providing information about structural changes within genes using Galaxy. The article is well written and I would like to see it published, I have provided very few notes, comments, and suggestions below.
I don't succeed to find typo or other little errors in the manuscript. Moreover, I don't have any major recommendation. Only few minors remarks, maybe ideas who can be of interest… or not ;)

Article

I particularly like the manner authors have presented workflow tools in the table 1. This is a good manner to summarized components, details and origins of each tool. Moreover, this translate the wish of transparency by authors and the fact they don't want to reinvent the wheel using, if relevant, existing tools.

Response:
We thank the reviewer for their kind comments and agree that reinventing the wheel is something we tried to avoid!

1) It appears to me that selecting the longest CDS per gene is a good idea and I'm wondering if this step cannot be made through the use of "classical" existing Galaxy tools as "Text manipulation" ones.

Response:
It is possible to use other tools, but users would probably have to chain together multiple tools to achieve the same result. By having our own tool we can make sure it works perfectly with different data formats, which could be difficult with other tools.

Concerning clustering steps, I want to propose some ideas, but not sure these are appropriate….

2) Not sure this is relevant but concerning clustering of similar sequences, I'm wondering if the use of algo like Swarm (https://github.com/torognes/swarm) (dedicated to metagenomics amplicon-based ngs data clustering) can be of interest…

Response:
As a first release of the workflow, we mainly concentrated on reproducing the Ensembl Compara pipeline into an easy-to-use Galaxy workflow. But we are grateful to the reviewer for the suggestion of testing alternatives for the stock tools in the workflow, and that is something we intend to look into in the future.

3) Concerning the overall clustering approach and related filtered steps, maybe it can be applied

a) a combination of parallel clustering methods (K-means/PAM/CAH/…) to compare "raw" clustering results and thus eliminate / reduce potential noise (ie sequences moving from one cluster to another when we change the clustering method, translating that this sequence is hard to assign)

b) a clustering affecting bootstrap confidence score to nodes thus giving ideas about relevance of each cluster.

Response:
We are grateful to reviewer for the suggestions for the clustering tools. As mentioned above, in first release of the workflow our aim is to reproduce the Ensembl Compara pipeline into a Galaxy workflow, which we managed to achieve. Improving the workflow by modifying the clustering approach is something we can look into in a future release of the workflow.

4) you propose the use of the Newick format. Maybe, as for the Philoviz visualization plugin, it would be interesting to add Nexus and phyloxml datatypes?

Response:
This workflow requires Newick format for the species tree because TreeBeST accepts input only in this format. There are several tools available to convert from Nexus and PhyloXML to Newick format, e.g. http://biopython.org/wiki/Phylo , so users who are producing data in these formats can perform the conversion before uploading them to Galaxy.

6) Maybe authors can give more details more aspects related to particular issues related to

a) alternative transcripts as it appears to me that ""just"" changing the max_target_seqs value from 3 to 4 seems to be quite "light" and

Response:
We agree with the reviewer that the explanation for changing the max_target_seqs value was a bit unclear in the manuscript, we have rewritten this part and reviewers will hopefully find it more satisfying.

b) working on non-model organisms.

Response:
With the significant reduction of sequencing costs, the proportion of fully sequenced genomes from non-model organisms has been steadily increasing. The analysis of the evolution of gene families across many species, including non-model organisms, allows us to assess the quality of the genome assemblies and associated annotations. We can also derive observations linking the expansion and contraction of gene families, as well as detect signals of positive selection in genes with 1:1 orthologs to specific physiological and behavioural adaptations (e.g. association between beta-keratin and shell evolution in turtles; Li et al. 2013. Genome Biol Evol. 5(5):923-

33); taste receptors in whales - Feng et al. 2014 Genome Biol Evol. 6(6):1254-65). We have added such details into the manuscript.

Unfortunately, I don't have taking time to implement the GeneSeqToFamily tools and workflow to a dedicated Galaxy instance. I would have appreciated to have a link giving access to a testing GeneSeqToFamily Galaxy instance. Maybe creating and pointing a dedicated Docker based Galaxy flavor can be of interest, thus giving an easy way to deploy a GeneSeqToFamily local Galaxy instance.

Response:
We have created a virtual image for testing the workflow, available [1], which is ready to be used with VirtualBox, and is supplied with necessary tools and plugins installed for the workflow. We will look into producing a Docker image for the same purpose in future.

8) Moreover, regarding the https://github.com/TGAC/earlham-galaxytools/tree/master/workflows/GeneSeqToFamily github link, the readme file "only" explain GeneSeqToFamily workflow without instructions on how to deploy it.

Response:
We have updated the readme.md file for the workflow with the list of tools required and instructions explaining how to install the workflow.

9) Regarding tools on the toolshed, it would be nice to have a repository called "suite_GeneSeqToFamily" or something like that to easily install all GeneSeqToFamily related Galaxy tools without installing it one by one like for the "suite_stacks" tools for example.

Response:
In the forthcoming Galaxy release 17.09 it will be possible to install a workflow together with the needed tools, so we hope that this feature is going to solve this issue without having to create a “suite” repository.

[1] Galaxy with GeneSeqToFamily
http://repos.tgac.ac.uk/vms/Galaxy_with_GeneSeqToFamily.ova