GeneSeqToFamily: the Ensembl Compara GeneTrees pipeline as a Galaxy workflow

Anil S. Thanki[1], Nicola Soranzo[1], Wilfried Haerty[1], Matthieu Muffato[2], Robert P. Davey[1]

1. Earlham Institute (EI), Norwich Research Park, Norwich NR4 7UZ, UK
2. EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, 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 gene families, produced using 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 the TreeBest [3] (part of TreeFam [4]) methodology to find internal structural-level synteny for homologous genes. TreeBeST runs 5 independent phylogenetic methods on the same data, then merges the 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 [5], hcluster_sg [6], T-Coffee [7], and phylogenetic tree construction tools, including TreeBeST [3]. 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 [8][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 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 show details about the type of interrelation event that gave rise to the family, such as speciation, duplication, and gene splits.

Workflow/Method

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, T-Coffee, 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 (GAFA).* Finally, we developed helper tools that are not part of the workflow itself, but aid the preparation of input data for the workflow and these are also in the ToolShed, i.e. *gff3_to_json, ete_species_tree_generator, fasta_header_converter, get_feature_info* and *get_sequences*.

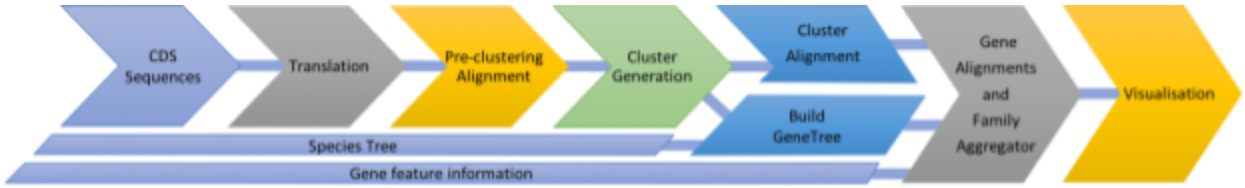

Figure 1: Overview of GeneSeqToFamily workflow

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 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 a SQLite [15] database for visualisation and downstream reuse. Each step of the workflow along with data preparation steps is explained in detail below.

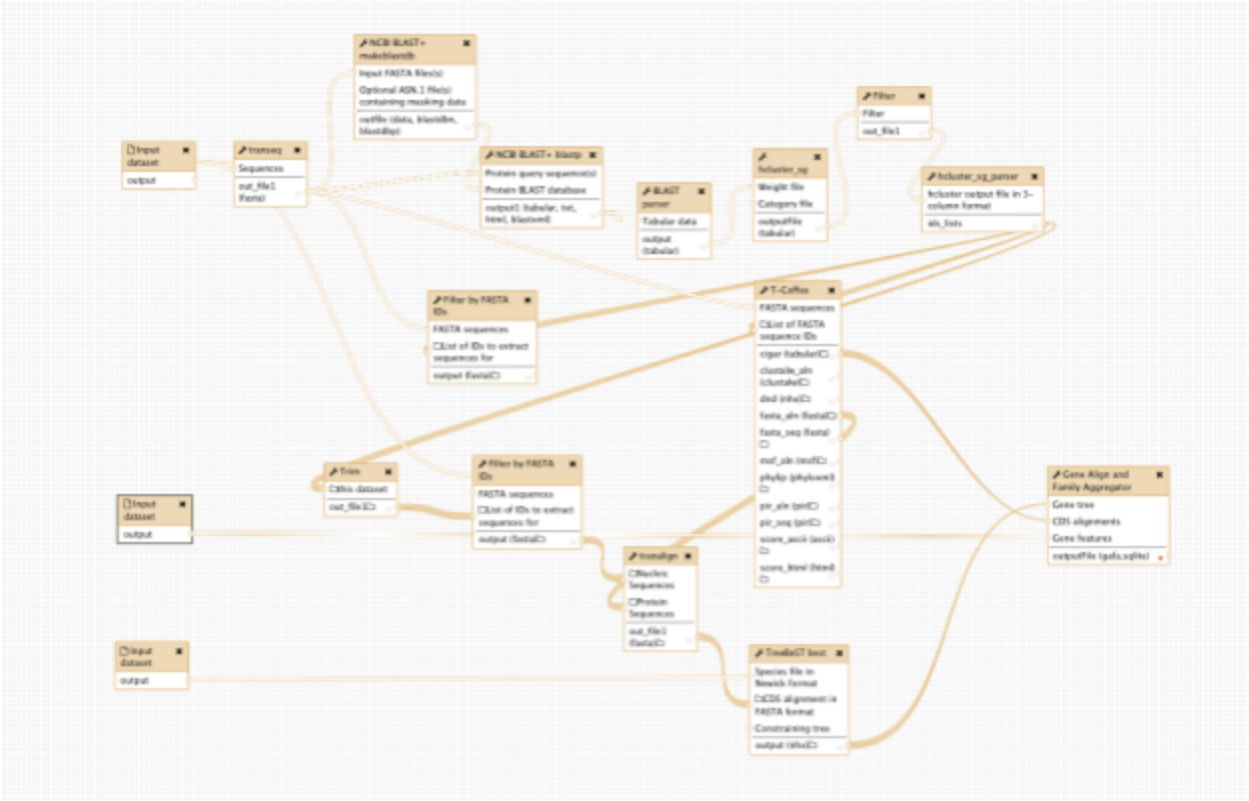

Figure 2: Screenshot from the Galaxy Workflow Editor, showing GeneSeqToFamily workflow

Table 1: Galaxy tools used in workflow

Tool Name	Tool ID	Version	Developed at EI		ToolShed Reference
			Tool	Wrapper	
NCBI BLAST+ makeblastdb	ncbi_makeblastdb	0.1.07	No	No	[16]
NCBI BLAST+ blastp	ncbi_blastp_wrapper	0.1.07	No	No	[16]
BLAST parser	blast_parser	0.1.1	Yes	Yes	[17]
hcluster_sg	hcluster_sg	0.5.1	No	Yes	[18]
hcluster_sg_parser	hcluster_sg_parser	0.1.1	Yes	Yes	[19]
T-Coffee	t_coffee	11.0.8	No	Yes	[20]
TreeBeST best	treebest_best	1.9.2	No	Yes	[21]
Gene Alignment and Family Aggregator	gafa	0.1.0	Yes	Yes	[22]

Transeq	EMBOSS: transeq101	5.0.0	No	No	[23]
Tranalign	EMBOSS: tranalign100	5.0.0	No	No	[23]
Filter by FASTA IDs	filter_by_fasta_ids	1.0	No	No	[24]
Get features by Ensembl ID	get_feature_info	0.1.2	Yes	Yes	[25]
Get sequences by Ensembl ID	get_sequences	0.1.2	Yes	Yes	[26]
GFF3 to JSON	gff3_to_json	0.1.1	Yes	Yes	[27]
ETE Species tree generator	ete_species_tree_generator	3.0.0b35	Yes	Yes	[28]

1. Translation

Transeq

Transeq, part of The European Molecular Biology Open Software Suite (EMBOSS) [29], is a tool to generate six-frame translation of nucleic acid sequences to their corresponding peptide sequences. Here we use Transeq to convert users' input CDS to protein sequences in order to run BLASTP and find protein clusters. However, downstream tools in the pipeline such as TreeBeST require nucleotide sequences to generate a gene tree. So, we use Transeq to interleave the two separate nucleotide and protein sequences into to just one CDS file.

2. Pre-clustering alignment

BLAST

This workflow uses the BLAST wrappers [30] 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 [31] 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 [32] 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 described above.

3. Cluster generation

hcluster_sg

hcluster_sg performs hierarchical clustering under mean distance 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 array of gene

clusters, each comprising a set of sequence IDs present in that cluster. This array of IDs 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 [31] converts hcluster_sg output into a list of IDs which are saved as separate files for each cluster. Each of these clusters contains at least 3 genes, which are then used to generate a gene tree via TreeBeST.

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

4. Cluster alignment

T-Coffee

T-Coffee is a multiple sequence alignment 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 [7], and we use it to align protein sequences in each cluster that are generated by hcluster_sg.

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

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

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

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

Whilst suitable for a small number of families, T-Coffee generates files on disk for each family, but with large family numbers this results in increased input-output (IO) steps and a performance bottleneck in the pipeline. As such, we modified the Galaxy wrapper for T-Coffee to take advantage of T-Coffee's pipe IO facility instead of supplying file handles. The modified wrapper can now 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, via a pipe, and T-Coffee will perform MSA for that set of sequences thus removing the need to write out filtered family sequences to potentially thousands of separate files. Similarly, we added functionality to generate CIGAR alignments [33] directly as output, instead of creating individual Galaxy jobs for each family. These features drastically improve the running time of the workflow. An example of a CIGAR string for aligned sequences is shown in Figure 3, in which each CIGAR string subset changes according to other sequences.

T-Coffee to CIGAR

"T-Coffee to CIGAR" is a small script to convert the multiple sequence alignment output of T-Coffee (as FASTA) into a simple CIGAR string form. The converter takes multiple FASTA alignments as input, and returns a table whereby the first column is the gene ID held in the FASTA header, and the second column contains the CIGAR alignment. This small script is included in the new T-Coffee repository in the Galaxy Toolshed, as described above and in Table 1.

5. Gene tree construction

Nucleotide alignment using Tranalign

Tranalign [29] 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

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. TreeBeST requires the species tree in New Hampshire format and a nucleotide FASTA aligned file with at least 3 sequences.

In GeneSeqToFamily, TreeBeST uses the alignment sequences generated from Tranalign and a user-supplied species tree (either produced by a third-party software, or through the *ete_species_tree_generator* data preparation tool, described below) to produce a GeneTree for each family represented in the Newick format [34]. 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 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 [15] containing the gene tree and multiple sequence alignment, along with gene features, in order to provide a reusable, persistent data store for visualisation of synteny information with Aequatus [12]. GAFA requires a gene tree file in Newick format [34], the CIGAR alignment, and gene feature information generated with the GFF3-to-JSON tool, which are then merged into a single SQLite file.

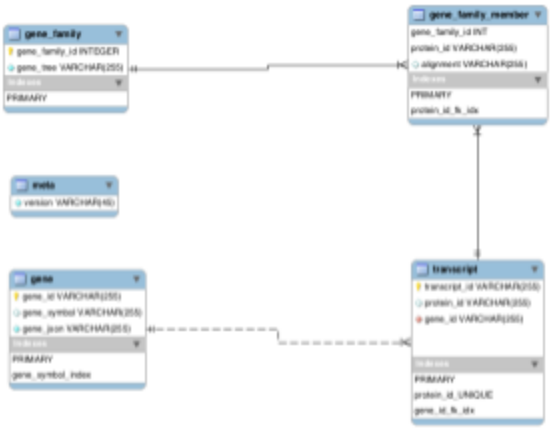

Figure 4: showing schema of GAFA SQLite database

The simple schema [31] for this database is shown in Figure 4.

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 (see Figure 3). 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 [31], 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, 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 sub-workflows to further annotate these genes by:

1) Retrieving a list of orphan genes from the GeneSeqToFamily workflow (see Figure 5) as follows:
    a) Compare the result of BLASTP, from GeneSeqToFamily workflow, with input CDS of the same workflow
    b) Find input CDS IDs which are not present in BLASTP result
    c) From input CDS of the GeneSeqToFamily workflow, retrieve respective sequence for each CDS ID (from the step above) using *filter_by_fasta_id*

    These unique CDS can be fed into the SwissProt workflow below to find homologous genes in other species.
2) Finding homologous genes for a gene(s) of interest using SwissProt (see Figure 6) as follows:
    a) Run BLASTP for CDS against the SwissProt database (from NCBI)
    b) Extract UniProt IDs from these BLASTP results, using the pre-installed Galaxy Tool *Cut1* (“Cut columns from a table”).
    c) Retrieve Ensembl IDs (representing genes and/or transcripts) for each UniProt ID using *uniprot_rest_interface* [35].
    d) Get genomic information for each gene ID and CDS for transcript ID retrieved from the core Ensembl database using the helper tools described below (*get_feature_info* and *get_sequences,* respectively).

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

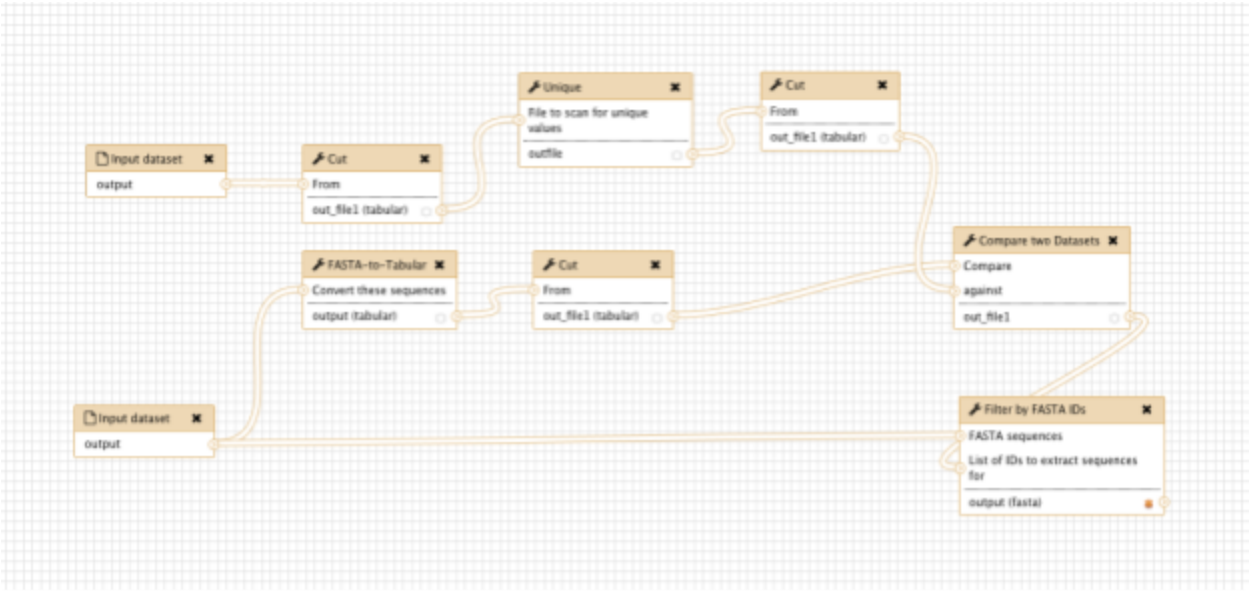

Figure 5: Screenshot from the Galaxy Workflow Editor, showing orphan gene finding workflow

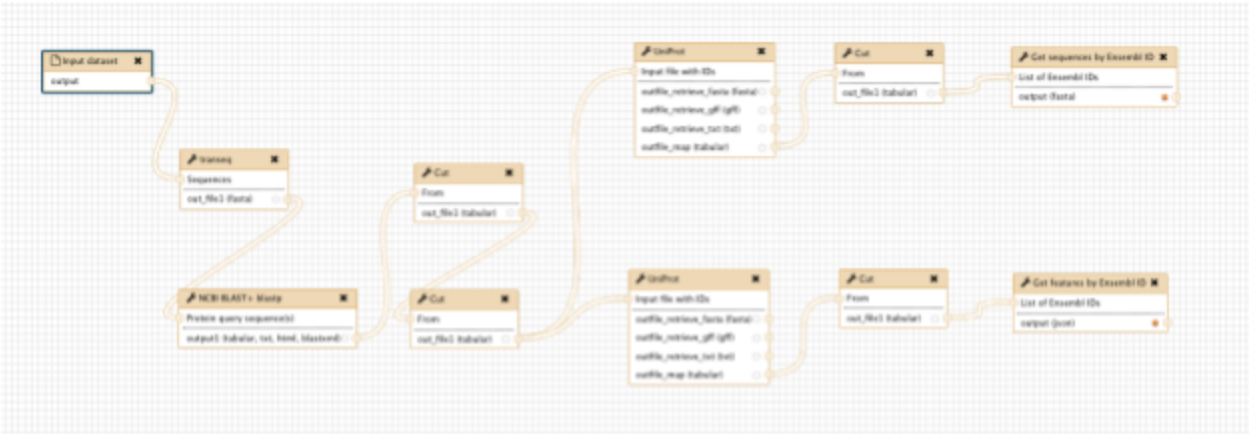

Figure 6: Screenshot from the Galaxy Workflow Editor, showing SwissProt workflow

Data preparation

We have developed a number of tools that assist in common data preparation steps that may need to be carried out on user-supplied datasets.

Ensembl REST API tools

Galaxy tools were developed which utilise the Ensembl REST API [36] to retrieve sequence information *(get_sequences)* and feature information (*get_feature_info*) by Ensembl ID from the Ensembl service. REST (Representational State Transfer) is an architecture style for designing networked applications [37] which encourages the reuse of 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:

- *fasta_header_converter*, which trims FASTA header information and appends species information to be used in TreeBeST
- *ete_species_tree_generator,* which uses the ETE toolkit [38] to generate a species tree from a list of species names or taxon IDs through NCBI Taxonomy.

GFF3-to-JSON

GFF3-to-JSON takes GFF3 files and JSON-formated genes as input and converts them into single JSON-formatted gene files which are identical to the data supplied from Ensembl REST API. GFF (Generic Feature Format) is a simple tab-delimited text file for describing genomic features in a text file. GFF3 is a latest version of GFF, which is nine-column, tab-delimited, plain text file [39][40], and widely used in bioinformatics. JSON (JavaScript Object Notation) is a lightweight data-interchange format, which is language independent and easier to understand for both humans and machines [41]. We promote the use of gene feature data in the Gene JSON format rather than GFF due to GFF's relatively inconvenient and unstructured additional information field.

Example Use Case

We tested this workflow on a large dataset of 63,194 genes (single transcript per gene) from three vertebrate species *Sarcophilus harrisii* (Tasmanian devil), *Mus musculus* (Mouse), and *Ornithorhynchus anatinus* (Platypus) in order to set a benchmark and to find optimum parameters to run the workflow. BLAST plays a crucial role in determining gene families. We ran the GeneSeqToFamily workflow using various BLAST parameters (as shown in Table 3), in order to try and identify those that were optimal 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 such as proteins. 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.

There are caveats, however. BLAST parameters used in Analysis 5 restrict the High-scoring Segment Pairs (HSPs) to 3 hits per reference sequence (the first hit when using all-versus-all BLAST will always be the query sequence itself). The minimum query coverage per 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 multiple alternative transcripts per gene, we recommend setting the HSP parameter to 4 rather than 3 to get a wider range of results from BLAST, thereby helping to generate gene families with matching genes together with alternative transcripts. In contrast setting a value of 3 for the HSP 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, thus increasing the likelihood of splitting of a gene tree into multiple trees and adversely inflating the number of families.

Summary						
Analysis	**1**	**2**	**3**	**4**	**5**	**6**
No of genes	63,194	63,194	63,194	63,194	63,194	63,194
No of families	8,410	13,642	13,772	13,065	13,065	17,090
Filtered out (>200)	12	1	1	1	1	0
Filtered out (<3)	2,643	2,589	3,124	2,012	2,012	5,872
Filtered families to consider	5,755	11,052	10,647	11,052	11,052	11,218
Average Family size	7.19	2.91	3.31	2.69	5.1	3.65
Median Family Size	4	3	3	3	4	3
Largest Family Size	2,200	829	602	829	829	65
Smallest Family Size	1	1	1	1	1	1

Table 2: showing the results of the GeneSeqToFamily workflow ran with 6 different BLAST parameter configurations, the complete list of which are shown in Table 3.

Analysis ID	e-value	HSP	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%

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

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, which could lead to a single gene tree for all MAO genes 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_feature_info* and *get_sequences* respectively). A species tree was generated using *species_tree_generator* and CDS were prepared with *fasta_header_converter*. We ran the GeneSeqToFamily workflow on these data using the parameters shown in Analysis 5 of Table 3. Here we set HSP 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 tree for MAOA, MAOB, BRAT1 and INSR gene families. Different colours of the nodes in the 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 GeneTree for these genes is already available in Ensembl and we used this to validate our findings [42] [43] [44] [45]. The Ensembl GeneTree exactly matched our GeneTree, showing that the workflow generates biologically valid results. We have provided the underlying data for this example along with the submitted workflow in figshare [46].

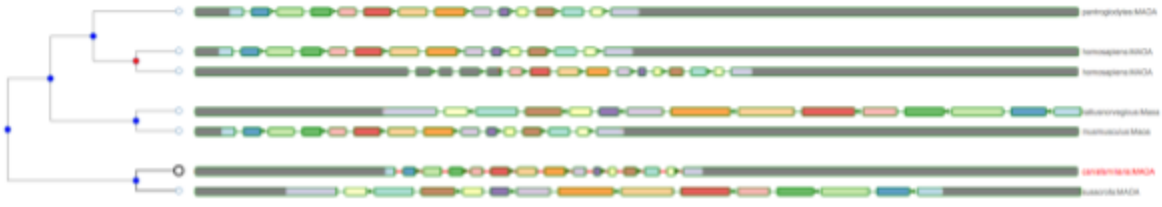

Figure 7: Showing homologous genes of MAOA of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus, Sus scrofa and Canis familiaris.*

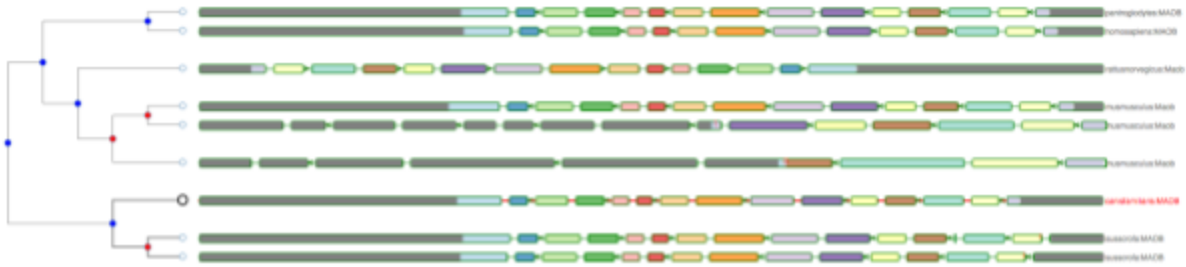

Figure 8: Showing homologous genes of MAOB of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus, Sus scrofa and Canis familiaris.*

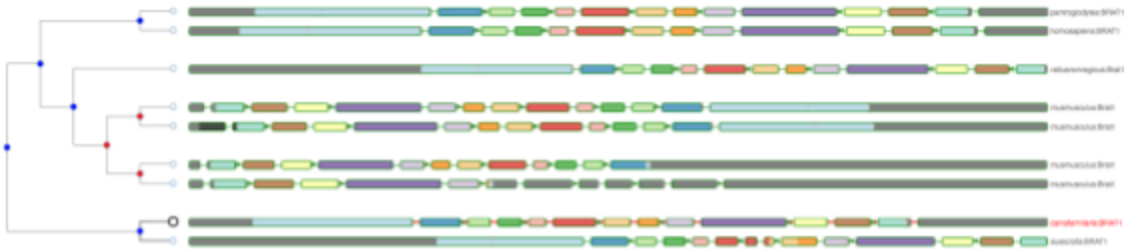

Figure 9: Showing homologous genes of BRAT1 of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus, Sus scrofa and Canis familiaris.*

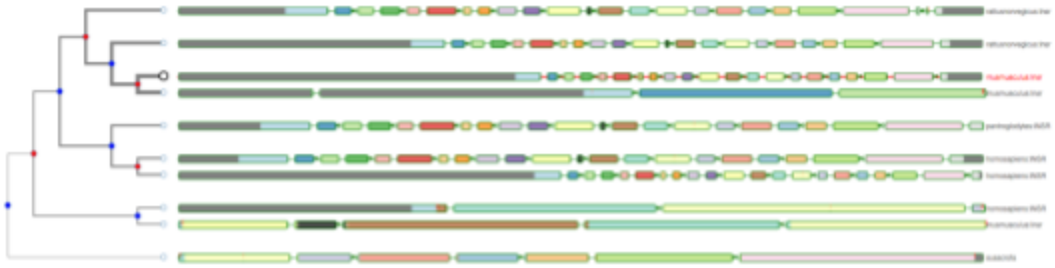

Figure 10: Showing homologous genes of INSR of *Canis familiaris* from *Mus musculus, Pan troglodytes, Homo sapiens, Rattus norvegicus, Sus scrofa and Canis familiaris.*

Conclusion

The ultimate goal of the GeneSeqToFamily is to provide a user-friendly workflow to analyse and find homologous genes using the Ensembl Compara GeneTrees pipeline within the Galaxy system, where user can analyse 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 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 include explicit integration of the PantherDB resources [47]. The PANTHER (Protein ANalysis Through Evolutionary Relationships) is a classification system to characterise known proteins and genes in order to certify genomic annotation. Association of PantherDB with GeneSeqToFamily will enable the automation of gene family validation and add supplementary information about those gene families, which could then be used in turn to further validate novel genomics annotation.

Availability and requirements

Project name: GeneSeqToFamily
Project home page:
https://github.com/TGAC/earlham-galaxytools/tree/master/workflows/GeneSeqToFamily
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.

Availability of supporting data

The example files and additional data sets supporting the results of this article are available in figshare [46].

Acknowledgements

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.

References

1. Jensen JD, Wong A, Aquadro CF: **Approaches for identifying targets of positive selection**. *Trends Genet* 2007, **23**:568–577.

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**:327–335.

3. **Ensembl/treebest** [https://github.com/Ensembl/treebest]

4. Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, Hériché J-K, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R: **TreeFam: 2008 Update**. *Nucleic Acids Res* 2008, **36**(Database issue):D735–40.

5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403–410.

6. **Hcluster_sg: hierarchical clustering software for sparse graphs** [http://sourceforge.net/p/treesoft/code/HEAD/tree/]

7. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**:205–217.

8. Goecks J, Nekrutenko A, Taylor J, Galaxy Team: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol* 2010, **11**:R86.

9. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists**. *Curr Protoc Mol Biol* 2010, **Chapter 19**:Unit 19.10.1–21.

10. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, De Roure D: **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.

13. **TGAC/aequatus.js** [https://github.com/TGAC/aequatus.js]

14. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy Team, Taylor J, Nekrutenko A: **Dissemination of scientific software with Galaxy ToolShed**. *Genome Biol* 2014, **15**:403.

15. **SQLite Home Page** [https://www.sqlite.org/]

16. **NCBI BLAST plus : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/devteam/ncbi_blast_plus]

17. **BLAST parser : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/blast_parser/]

18. **hcluster_sg : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/hcluster_sg/]

19. **hcluster_sg parser : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/hcluster_sg_parser/]

20. **T_Coffee : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/t_coffee/]

21. **TreeBeST best : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/treebest_best/]

22. **Gene Align and Family Aggregator (GAFA) : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/gafa/]

23. **EMBOSS : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/devteam/emboss_5/]

24. **Filter by FASTA IDs : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/galaxyp/filter_by_fasta_ids/]

25. **Get features by Ensembl ID : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/ensembl_get_feature_info/]

26. **Get sequences by Ensembl ID : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/ensembl_get_sequences/]

27. **GFF3 to JSON converter : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/gff3_to_json/]

28. **ETE species tree generator : Galaxy Tool Shed** [https://toolshed.g2.bx.psu.edu/view/earlhaminst/ete/]

29. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**:276–277.

30. Cock PJA, Chilton JM, Grüning B, Johnson JE, Soranzo N: **NCBI BLAST+ integrated into**

Galaxy. *Gigascience* 2015, **4**:39.

31. **TGAC/earlham-galaxytools** [https://github.com/TGAC/earlham-galaxytools]

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

33. **Sequence Alignment/Map Format Specification**
[http://samtools.github.io/hts-specs/SAMv1.pdf]

34. **"Newick's 8:45" Tree Format Standard**
[http://evolution.genetics.washington.edu/phylip/newick_doc.html]

35. **bgruening/galaxytools** [https://github.com/bgruening/galaxytools]

36. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, Ruffier M, Taylor K, Vullo A, Flicek P: **The Ensembl REST API: Ensembl Data for Any Language**. *Bioinformatics* 2015, **31**:143–145.

37. **Representational State Transfer** [http://www.peej.co.uk/articles/rest.html]

38. Huerta-Cepas J, Serra F, Bork P: **ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data**. *Mol Biol Evol* 2016.

39. **GFF3 - GMOD** [http://gmod.org/wiki/GFF3]

40. **The Sequence Ontology - Resources - GFF3**
[http://www.sequenceontology.org/gff3.shtml]

41. **JSON** [http://www.json.org]

42. **Gene: BRAT1 (ENSG00000106009) - Gene tree - Homo sapiens - Ensembl genome browser 87**
[http://dec2016.archive.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG00000106009;r=7:2537877-2555727;]

43. **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;]

44. **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;]

45. **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;]

46. Anil S. T, Nicola S, Wilfried H, Robert D: **GeneSeqToFamily.zip**. *figshare* 2016.

47. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD: **The PANTHER database of protein families, subfamilies, functions and pathways**. *Nucleic Acids Res* 2005, **33**(Database issue):D284–8.