


TECH NOTE

Aequatus: an open-source homology browser

Anil S. Thanki ^{1,*}, Nicola Soranzo ¹, Javier Herrero ^{1,2},
Wilfried Haerty ¹ and Robert P. Davey ¹¹Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK and ²Bill Lyons Informatics Centre, UCL Cancer Institute, 72 Huntley St., London, WC1E 6DD, UK*Correspondence address. Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK. E-mail: Anil.Thanki@earlham.ac.uk  <http://orcid.org/0000-0002-8941-444X>

Abstract

Background: Phylogenetic information inferred from the study of homologous genes helps us to understand the evolution of genes and gene families, including the identification of ancestral gene duplication events as well as regions under positive or purifying selection within lineages. Gene family and orthogroup characterization enables the identification of syntenic blocks, which can then be visualized with various tools. Unfortunately, currently available tools display only an overview of syntenic regions as a whole, limited to the gene level, and none provide further details about structural changes within genes, such as the conservation of ancestral exon boundaries amongst multiple genomes. **Findings:** We present Aequatus, an open-source web-based tool that provides an in-depth view of gene structure across gene families, with various options to render and filter visualizations. It relies on precalculated alignment and gene feature information typically held in, but not limited to, the Ensembl Compara and Core databases. We also offer Aequatus.js, a reusable JavaScript module that fulfills the visualization aspects of Aequatus, available within the Galaxy web platform as a visualization plug-in, which can be used to visualize gene trees generated by the GeneSeqToFamily workflow.

Keywords: alignment, gene family, homology, phylogeny, synteny, visualization

Introduction

Sequence conservation across populations or species can be investigated at multiple levels from single nucleotides, to discrete sequences (e.g. transcription factor binding sites, exons, introns), genes, genomic blocks, and chromosomes. Analyses at each of these levels inform different evolutionary processes and time scales. While the vast majority of analyses focus on gene evolution, synteny (the conservation of genomic blocks between multiple species) can be used to trace chromosome evolutionary history [1] and infer evolutionary relationships between genes across or within species [2]. Synteny resolution and analysis typically involves carrying out multiple sequence alignments (MSAs) and phylogenetic reconstruction, comprising multiple steps that can be computationally intensive even for relatively small numbers of data points [3].

Many methods are available for the identification of genome-wide orthology (MSOAR [4], OrthoMCL [5], OMA [6], Homolo-

Gene [7], PhyOP [8], TreeFam [9], TreeBeST [10]). However, most of them do not incorporate taxonomic information (typically in the form of a species tree) while finding gene families, nor do they provide any information regarding transcript and protein structural changes across orthogroup members. The Ensembl GeneTrees pipeline [11], a computational workflow developed by the EMBL-EBI Ensembl Compara team, produces familial relationships based on clustering, MSA, and phylogenetic tree inference. The gene trees in Ensembl Compara are inferred with TreeBeST, which relies on a reference species tree to guide the process and calculates the probability of a gene tree in the context of species evolution. The data are stored in a relational database that contains information on gene families, syntenic regions, and protein families. In parallel, the Ensembl Core databases store gene feature information and other genomic annotations at the species level. The Ensembl project (release 90, August

Received: 18 June 2018; Revised: 6 September 2018; Accepted: 17 October 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

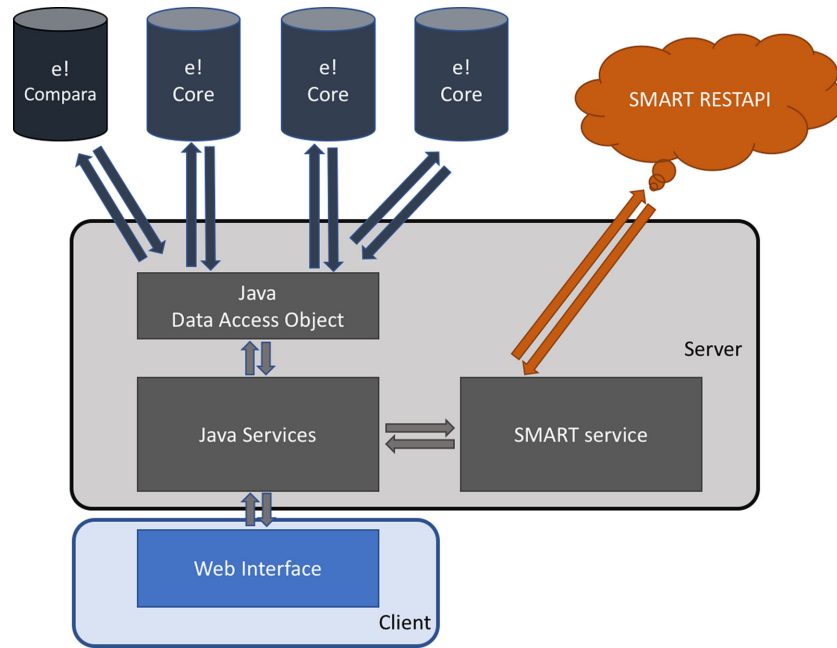


Figure 1: The Aequatus infrastructure, showing the interactions between the server-side implementation, connected to Ensembl Compara and Core database using Java Data Access Objects and Simple Modular Architecture Research Tool (SMART) server via REpresentational State Transfer (REST) application programming interface (API), and the client-side implemented using popular techniques such as JavaScript, jQuery, d3.js, and jQuery DataTables.

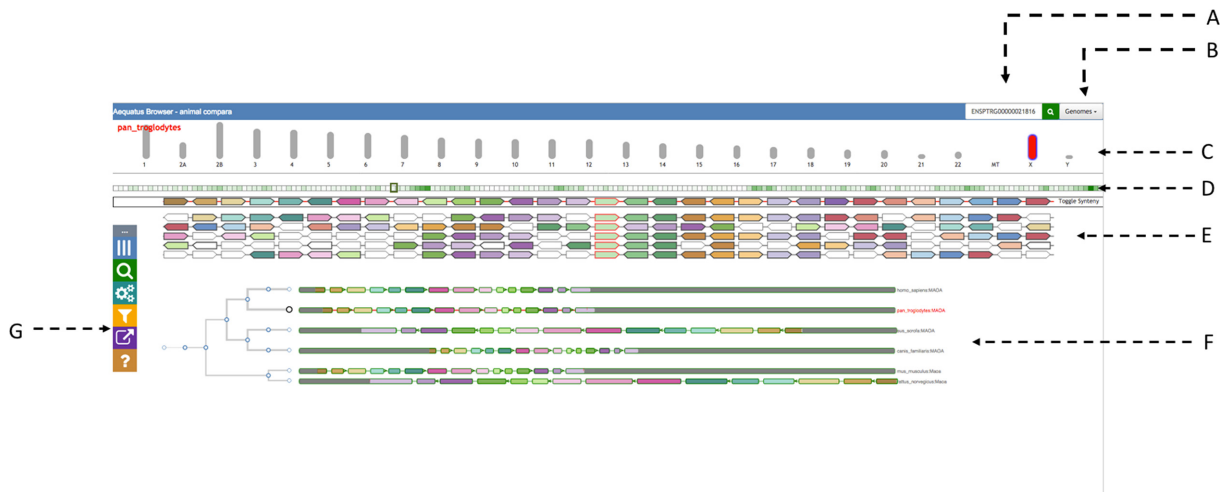


Figure 2: The main view of Aequatus. The header on top provides a search box (A) and a genome list (B). It is followed by the chromosomal view (C), where the selected chromosome is colored in red. Below there is an overview of genes (D) for the selected chromosome, followed by a zoomed area of the chromosome with genes shown in the gene order view (E) and by gene tree view (F). We are using arbitrary colors to distinguish syntenic genes (in gene order view) and matching exons (in gene tree view). The Aequatus control panel (G) is visible on the far left.

2017) at EMBL-EBI houses 100 vertebrate species [12], along with precomputed MSAs and gene family information.

Phylogenetic reconstruction is the most traditional method to represent and view comparative datasets across a given evolutionary distance, but specific tools such as Ensembl Browser [13], Genomicus [14], SyMAP [15], and MizBee [16] also exist to provide an overview of syntenic regions as a whole, with only Genomicus reaching down to the gene order and orientation level. Conversely, phylogenetic trees retain ancestral information but do not represent the underlying information regarding structural changes within genes, such as the conservation of ancestral

exon boundaries between multiple genomes or variants within genes that can be correlated to phenotypic changes. In order to build these gene-level visualizations, basic genomic feature information is required.

Therefore, we have developed Aequatus to bridge the gap between phylogenetic information and gene feature information. Here, we show that Aequatus allows the identification of exon/intron boundary changes and mutations, informing the user about underlying genetic changes.

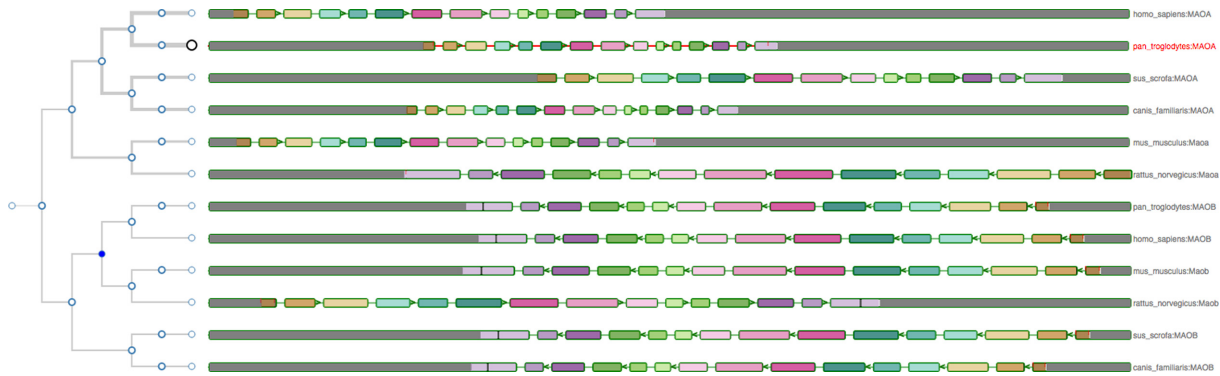


Figure 3: The gene tree for the monoamine oxidase (MAO) gene, with the Chimp gene as the reference, alongside other homologous genes in the exon-focused view. Considering the gene tree on the left, it is clear that the MAO genes are separated into two clusters, corresponding to the MAOA and MAOB gene families.

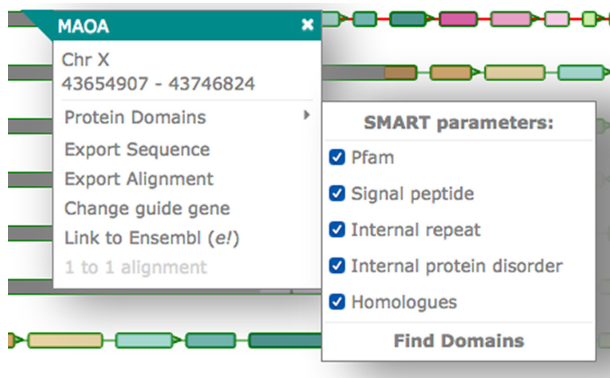


Figure 4: The pop-up in the gene tree view when clicking on a gene. The pop-up contains the chromosome name and position and options to view the protein domains, export the sequence or the alignment, change the guide gene, connect to the Ensembl page for the gene, and view the pairwise alignment.

Materials and Methods

Aequatus is built using open-source technologies and is divided into a typical server-client architecture: a web interface and a server backend (see Fig. 1).

The server-side component is implemented using the Java programming language. It retrieves and processes comparative genomics information directly from Ensembl Compara and Ensembl Core databases. Precalculated gene trees and genomic alignments, in the form of CIGAR strings [17], are held in Ensembl Compara, which are cross-referenced by Aequatus to Ensembl Core databases for each species to gather genomic feature information using the unique gene stable IDs.

The Aequatus web interface comprises well-known web technologies such as SVG, jQuery, JavaScript, and D3.js [18] to provide a fast and intuitive web-based browsing experience over complex data. Comparative and feature data are processed and rendered in an intuitive graphical interface to provide a visual representation of the phylogenetic and structural relationships among the set of chosen species.

Aequatus visualizes gene families using a phylogenetic tree generated from gene sequence conservation information, held in an Ensembl Compara database, and gene features from Ensembl Core database. Gene features are presented in the form of exon-intron boundaries and 5' and 3' untranslated regions (UTRs). In this gene tree view, users are able to select a gene from a given species as a “guide gene,” and the homologous genes

discovered through the comparative analysis are shown with respect to this guide gene. The representation of internal similarity among homologues is achieved by comparing the CIGAR strings for homologous genes with the CIGAR of the guide gene and mapping back to the homologous gene structure.

Aequatus is also able to visualize homologous genes in a customized Sankey view, using the d3.js [18] visualization library, and provides feature information in an interactive Tabular view, using the jQuery DataTable [19] library. Statistical information for each member in a set of homologues, such as percentage coverage, positivity, and identity, are fetched from *homology* and *homology_member* tables of the Ensembl Compara database.

We have integrated a Simple Modular Architecture Research Tool (SMART) [20] service to search for and visualize domain information of a protein sequence. We use the SMART Representational State Transfer (REST) application programming interface (API) to retrieve protein domains, motifs, signal, and repeats information from the SMART server using protein sequences.

Finally, to complement these various visualizations for the homologous genes and their gene trees, Aequatus provides gene order information in the form of a syntenic view (see the “Gene Order” section below). For a selected gene, homologues are fetched from *homology* and *homology_member* tables of the Ensembl Compara database. The neighboring genes for these homologous genes are retrieved from the Ensembl Core databases using positional information and organized into a syntenic representation. Much like the shared conserved exon depiction in the gene tree view, syntenic genes are colored based on the shared homology.

Results

The landing page of Aequatus (see Fig. 2) contains a header with a search box (2A) and a dropdown list of species (2B), followed by a selectable chromosomal view underneath (2C).

Aequatus has a draggable control panel (2G) on the left-hand side that contains buttons to show/hide the chromosome selector on top, modify gene views and labels, access the search box, and the export options, as well as a link to the help pages.

Aequatus user interface

Aequatus provides various ways to visualize gene trees and the inferred orthology/paralogy from them.

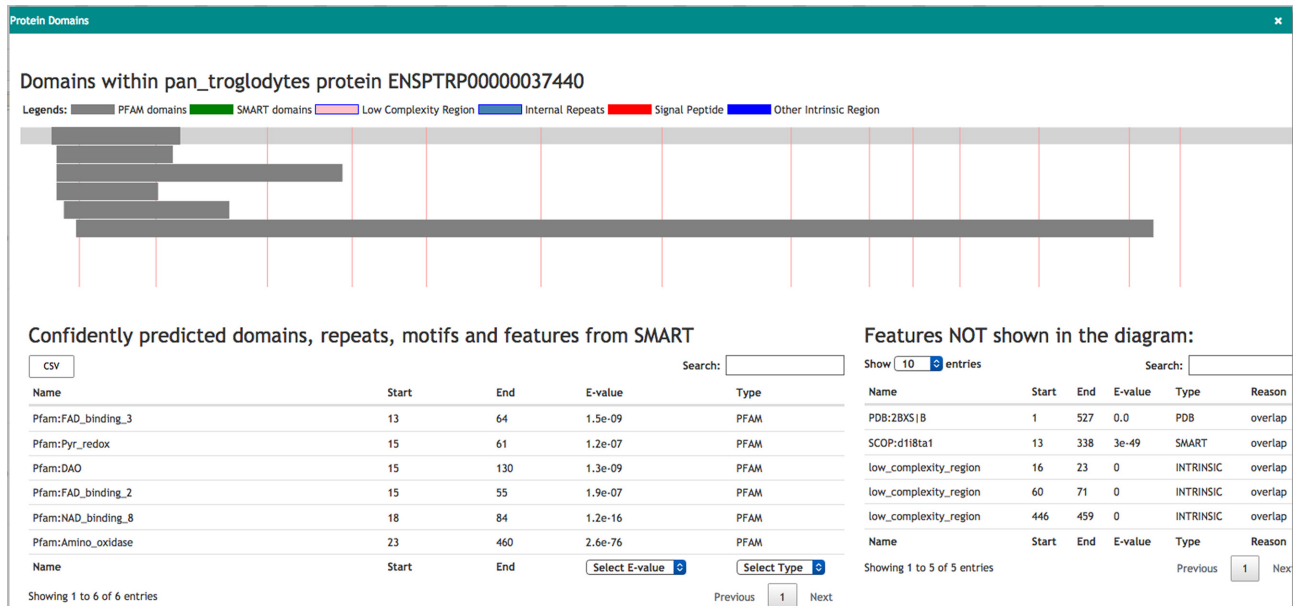


Figure 5: Visualization of the protein domain information for the protein ENSPTRP0000037440 retrieved from the SMART server. On the top, drawings of domains mapped on exons (shown with red lines). The tables below list the features shown in the diagram as well as hidden features.

Main gene trees view

The gene tree view (see Fig. 3) comprises a phylogenetic tree on the left, built from GeneTree information stored in a Ensembl Compara database [11]. Aequatus relates the genes through different events (e.g. duplication, speciation, and gene split) for the gene family and homologous genes against each respective node, which are colored based on the potential evolutionary event. Homologous genes are visualized by aligning them against a given guide gene. The selected guide gene is depicted as a larger circle black leaf node in the tree, with a red label on the right, while the other genes have a smaller circle leaf node and a gray label.

On the right, Aequatus depicts the internal gene structure, using a shared color scheme for coding regions, to represent similarity across homologues. Homologous genes are visualized by aligning them against a given guide gene. Aequatus is also able to indicate insertions and deletions in homologous genes with respect to shared ancestors. Black bars within exons represent insertions, while red lines represent deletions specific to a given gene compared with the guide.

Aequatus provides two view types for gene families. The first (default) view is exon focused (as in Fig. 3), where all introns are set to a fixed width, since long introns can adversely affect the visibility of surrounding exons. This provides easier browsing of the actual gene structure, especially when less screen real estate is available. Conversely, in the second view, all homologous genes are resized to the maximum available width in the web browser, showing introns and exons proportional to the real gene size. Users can switch between these views from the “Introns” settings in the control panel.

In gene tree view, gray blocks at the start and end of each gene represent UTRs, black bars within exons indicate insertions, red lines represent deletions specific to a given gene compared with the guide, and tiny arrows denote the coding strand of the gene.

Pop-ups Aequatus provides a contextual menu system via interactive pop-up menus, which are displayed when a user clicks on

a gene (see Fig. 4). Each pop-up shows the gene name and its position; a link to find protein domain information using SMART; links to export the protein sequence or the CIGAR alignment; an option to set the current gene as the guide in order to see insertions and deletions in homologous genes relative to the selected guide gene; a link out to the Ensembl page for the gene; and an option to view the pairwise alignment.

Protein domain Aequatus can provide an interactive visualization of the protein domains for the selected gene. Aequatus finds the protein domains by connecting to the SMART web server via its REST API and querying the protein sequence for domains, motifs, internal repeats, and similar information. In this view (see Fig. 5), a user can filter and sort domains based on type, E-value, position, and source of domain. The features shown in the diagram can be exported in comma-separated value (CSV) or Excel file format.

Homologous genes

The underlying information describing homologous genes contained within the Compara database schema can be visualized using either a tabular view or Sankey plot.

Tabular view The tabular view (see Fig. 6) contains statistical information for the homologous relationships. This view is dynamic, allowing the user to search for any homolog using a search box (6A) as well as filter results for the type of homology (6E) (1-to-1 orthologs, 1-to-many orthologs, and paralogs) or one or more specified species (6D). Homologous genes can be exported from the tabular view as Excel, CSV, or PDF.

Extra details for the pairwise alignment between homologues can be shown by using the “+” button for the homologue entry. The first button (6B) will show statistical comparisons for identity, coverage and similarity, while the second button (6C) will visualize the pairwise alignment with the gene structures as detailed in the “1-to-1 alignments” subsection below.

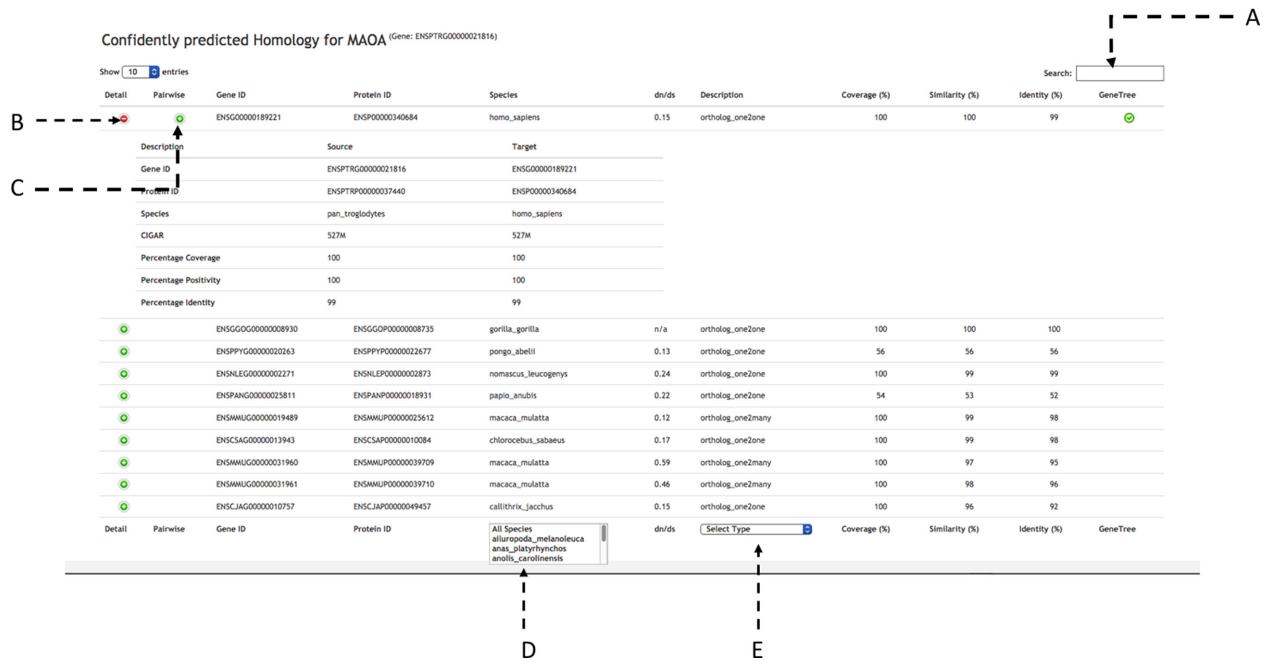


Figure 6: Homologous for the gene MAOA (ENSPTRG00000021816) in tabular view with statistical comparison about homologues. The tabular view contains a search box on top (A). There are two buttons to visualize statistical comparisons (C) and pairwise alignment (D) for each homolog. At the bottom it is possible to select from a list of species (D) and the type of homology (E).

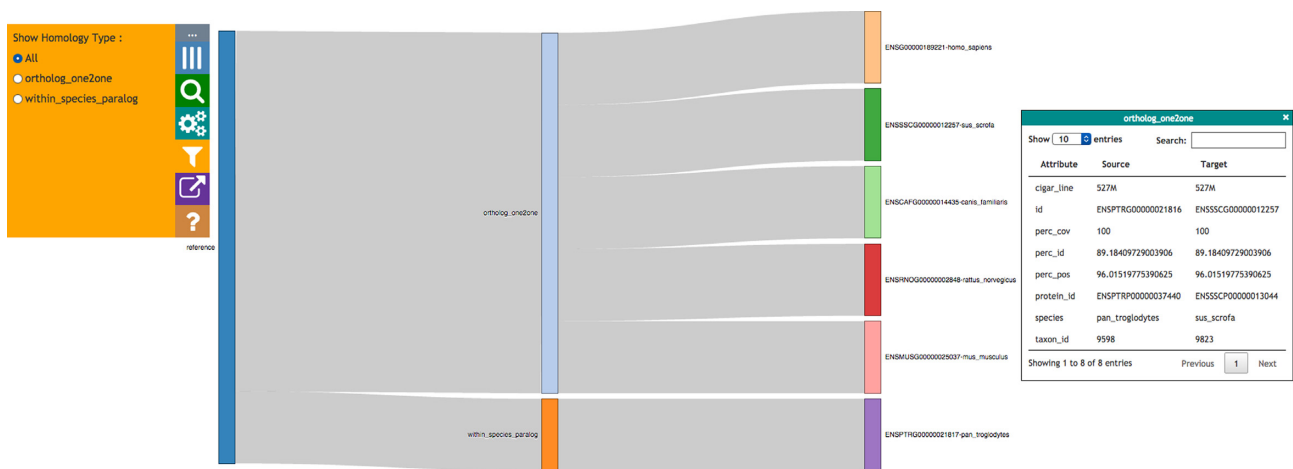


Figure 7: Homologues for a gene in Sankey format, grouped together by type of homology. The control panel on the left shows filters for the view. Additional information for any homologue can be retrieved by clicking on it; the information is then shown in a box on the right.

Sankey view The Sankey view (see Fig. 7) visualizes homology as an interactive diagram, where the homologues of a selected gene are distinguished by homology type, i.e. paralogs, 1-to-1 orthologs, or 1-to-many orthologs. The nodes for homologous genes are colored by species, which helps when finding genes from the same species in the case of 1-to-many and many-to-many orthologs.

When clicking on a homologous gene, additional details for the homologous pair are displayed in the info panel on the right-hand side.

1-to-1 alignments

Aequatus provides 1-to-1 alignments between homologous genes to facilitate pairwise comparisons. These 1-to-1 align-

ments (Fig. 8) can be seen by clicking on the corresponding option either in the pop-up for the gene tree view or in the homologous genes tabular view. This will fetch the relevant alignment from the homology table of the Ensembl Compara database and visualize it based on the gene structure (8A) together with the pairwise protein sequence alignments (8B).

Gene order

Genes that share a common ancestor and are part of a consecutive block of genes are likely to have a transcriptional and/or functional relationship [21]. Hence, inferred homologues that are present in all species and in the same order are more likely to be real homologues. In the Gene Order view, neighboring genes are displayed for the selected gene and its homologues (shown

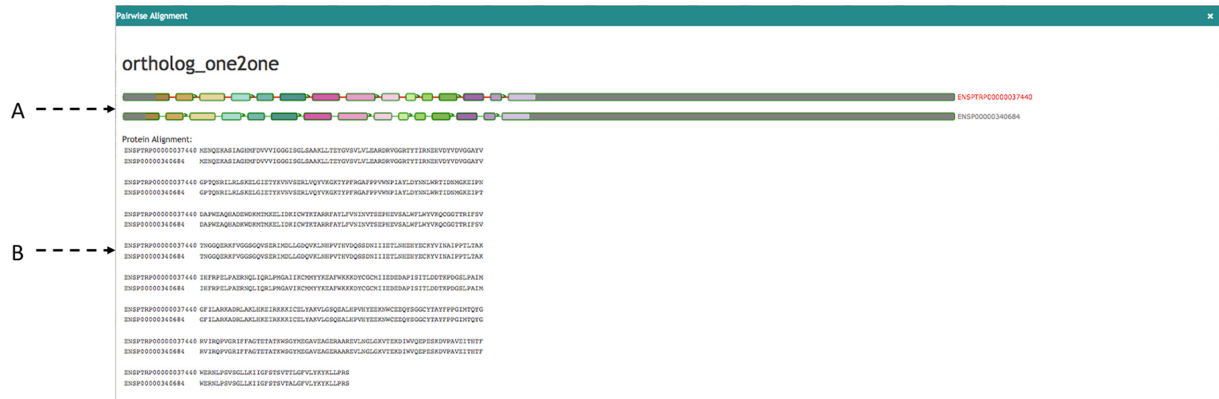


Figure 8: The 1-to-1 alignments between homologous genes. (A) Visualizing alignment on gene structure and (B) visualizing pairwise sequence alignments.

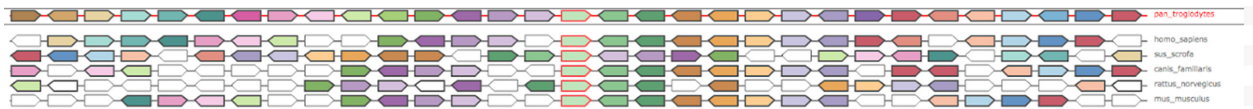


Figure 9: Gene Order for the MAOA gene in Pan troglodytes, where they are colored by homologous genes. The selected gene and its homologous have a red border. White genes are the ones that do not have any homologous in the current visible region.

in Fig. 9). Homologues of the genes in neighboring species are colored based on the matching genes from the reference species. Clicking on a gene feature will open a search panel with various viewing options, and mousing over a given gene will highlight all homologous genes within the same region. The syntenic view complements the main functionality of Aequatus by providing evidence for the conservation information for the genes of interest.

Search

Aequatus has keyword-based search functionality, whereby the user can provide search terms and a list of all the relevant genes is returned. Aequatus can query for matching gene symbols, Ensembl stable IDs (unique identifiers in the Ensembl project for each genomic annotation), common names for genes and proteins, or any keyword in the description. Search results then allow the user to visualize the corresponding gene tree view or homologous genes in the tabular or Sankey views.

Export

Users can export data at different points in the visualization. In the gene tree view, the underlying genomic data for the gene families can be exported in various forms, such as a list of gene IDs, the sequence alignments, or the gene trees in Newick [22] or JavaScript Object Notation (JSON) [23] format, for use in downstream tools. The tabular view can be exported in CSV, XLS, and PDF format.

Persistent uniform resource locators

Aequatus provides persistent unique uniform resource locators (URLs) to enable consistent access to genes of interest, making it easy to go back to the results of a previous search, to share information with collaborators, or for use in publications. Users can share the link for the visualization of a specific gene, the results of a search for a term, or a specific reference to a given species and chromosome.

Discussion

The ultimate goal of Aequatus is to provide a unique and informative way to render and explore complex relationships between genes from various species at a level of detail that has so far been unrealized in a single platform. Supplementary Table S1 shows a detailed comparison of Aequatus with various phylogenetic visualization tools, which highlights the signature feature of Aequatus, i.e. genetic structural comparison. Supplementary Figs. S1–S3 allow a comparison of the visualizations of monoamine oxidase B (MAOB) genes from the tools offering a gene tree-focused view.

While applicable to species with high-quality gold-standard reference genomes present in core database resources such as human or mouse, Aequatus has been designed to accommodate users who need to explore large, fragmented, nonmodel genome references that are held in institutional databases. Comparing nonmodel organism genes with gold standard genomes allows the identification of exon/intron boundary changes and mutations, informing the user about underlying genetic changes, but can also highlight mis-annotations, pseudogenes [24], or polyploidization (see Fig. 10). We are currently testing Aequatus with a range of nonmodel organisms, such as koala, polyploid crops, and spiny mouse. As Aequatus can visualize relationships using simple CIGAR strings, any tool that outputs this format can use Aequatus to view them. We produce input for Aequatus using the GeneSeqToFamily pipeline, a freely available Galaxy workflow [25] for finding and visualizing gene families for genomes that are not available from Ensembl databases.

In order to make Aequatus more accessible and reusable, the gene tree visualization module from the stand-alone Aequatus browser is available as Aequatus.js [26], an open-source JavaScript library. In this way, it preserves the interactive functionality of the Aequatus browser tool but can be integrated with other third-party web applications. We have demonstrated this by integrating the Aequatus.js library into Galaxy [27], where gene families generated by running the GeneSeqToFamily workflow can be visualized using the Aequatus plug-in within Galaxy.

5. Li L, Stoeckert CJ, Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13:2178–89.
6. Altenhoff AM, Dessimoz C. Inferring orthology and paralogy. *Methods in Molecular Biology.* 2012, p. 259–79.
7. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2008;36:D13–21.
8. Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol.* 2006;2:e133.
9. Li H, Coghlan A, Ruan J, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 2006;34:D572–80.
10. TreeSoft: TreeBeST. <http://treesoft.sourceforge.net/treebest.shtml>. Accessed 9 June 2018.
11. Clamp M, Andrews D, Barker D, et al. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* 2003;31:38–42.
12. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Res.* 2018;46:D754–61.
13. Stalker J, Gibbins B, Meidl P, et al. The Ensembl web site: mechanics of a genome browser. *Genome Res.* 2004;14:951–5.
14. Muffato M, Louis A, Poinsel CE, et al. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics.* 2010;26:1119–21.
15. Soderlund C, Nelson W, Shoemaker A, et al. SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res.* 2006;16:1159–68.
16. Meyer M, Munzner T, Pfister H. MizBee: a multiscale synteny browser. *IEEE Trans Vis Comput Graph.* 2009;15:897–904.
17. Sequence Alignment/Map Format Specification. <http://samtools.github.io/hts-specs/SAMv1.pdf>. Accessed 9 June 2018.
18. Bostock M. D3.js - Data-Driven Documents. <http://d3js.org/>. Accessed 9 June 2018.
19. DataTables | Table plug-in for jQuery. <https://datatables.net>. Accessed 9 June 2018.
20. Schultz J, Milpetz F, Bork P. et al. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci.* 1998;95:5857–64.
21. Dávila López M, Martínez Guerra JJ, Samuelsson T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One.* 2010;5(5):e10654. doi: 10.1371/journal.pone.0010654.
22. Newick Format. <http://evolution.genetics.washington.edu/phylip/newick.doc.html>. Accessed 8 Apr 2018.
23. JSON format. <http://www.json.org>. Accessed 9 June 2018.
24. Vanin EF. Processed pseudogenes: characteristics and evolution. *Annu Rev Genet.* 1985;19:253–72.
25. Thanki AS, Soranzo N, Haerty W, et al. GeneSeqToFamily: a Galaxy workflow to find gene families based on the Ensembl Compara GeneTrees pipeline. *GigaScience.* 2018;7(3):1–10.
26. Thanki AS, Davey RP. TGAC/aequatus.js GitHub Repository. <https://github.com/TGAC/aequatus.js>. Accessed 9 June 2018.
27. Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 2016;44:W3–10.
28. UseGalaxy.eu <https://usegalaxy.eu>. Accessed 8 June 2018.
29. Yates A, Beal K, Keenan S, et al. The Ensembl REST API: Ensembl data for any language. *Bioinformatics.* 2015;31:143–5.
30. Goff SA, Vaughn M, McKay S, et al. The iPlant Collaborative: cyberinfrastructure for plant biology. *Front Plant Sci.* 2011;2:34.
31. Grüning B, Dale R, Sjödin A, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods.* 2018;15(7):475–6.
32. Thanki AS, Bian X, Davey RP. TGAC Browser: visualisation solutions for big data in the genomic era. <http://browser.earlham.ac.uk/> Accessed 8 June 2018.
33. Thanki AS, Soranzo N, Herrero J, et al. Supporting data for “Aequatus: An open-source homology browser.” *GigaScience Database.* 2018. <http://dx.doi.org/10.5524/100509>.