8-14-2018

OrthoQuery: A Tripal Database Module to Assess and Visualize Gene Family Evolution

Sumaira Zaman
sumaira.zaman@uconn.edu

OrthoQuery: A Tripal Database Module to Assess and Visualize Gene Family Evolution

Sumaira Zaman

B.S. University of Connecticut 2017

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

At the

University of Connecticut

2018

Approval Page

Master of Science Thesis

OrthoQuery: A Tripal Database Module to Assess and Visualize Gene Family Evolution

Presented by
Sumaira Zaman, B.S.

Major Advisor _____
Dr. Jill Wegrzyn

Associate Advisor _____
Dr. Yaowu Yuan

Associate Advisor _____
Dr. Dong-Guk Shin

University of Connecticut
2018

Acknowledgements

I would first and foremost like to thank my advisor Jill Wegrzyn for not only being a supportive advisor but also serving as an amazing role model. Her enthusiasm for her research can be seen through her dedication towards her work, students, and colleagues. She is truly an inspiring mentor and I cannot express enough gratitude for taking me under her wing. I am blessed to have her as my advisor and for all the knowledge she has shared with me.

I would also like to thank all the professors who have shaped me into the student and researcher I am today. I'd like to thank Jonathan Klassen for opening the world of bioinformatics to me and providing critical guidance. I would also like to thank my committee members Yaowu Yuan and Dong-Guk Shin for their time and support. I would like to express my sincere gratitude towards all the past and present members of the Plant Computational Genomics labs for being fantastic colleagues and even better friends. Thank you for the encouragement you've provided, the conversations we've had, and all the knowledge you've shared with me and others.

This section would be incomplete without mentioning my mother, Nighat Zaman, who has tirelessly fought all her life to not only provide me with a valuable education but also with a valuable life. I would not have these opportunities without her sacrifices and I cannot thank her enough for all the she has done and continues to do for me.

Table of Contents

List of Figures

List of Tables

Introduction

High throughput technologies widely accessible in genomics and proteomics have enabled scientists across the globe to assess more transcriptomes, genomes, and systems (1). Well designed bioinformatics pipelines that can efficiently connect these large datasets to analytical tools and interactive visualizations are lacking. This gap reflects the size, complexity, and diversity of these datasets, that despite standardized file formats, still present hurdles in their storage, transfer, and analysis. Biologists are tasked with gathering and filtering large datasets from multiple instruments or repositories, executing computationally intensive analysis on external High-Performance Computing (HPC) clusters, and pushing this data back out to third-party packages for visualizations. Although data derived from the scientific community are housed in public curated repositories (Genbank, EMBL), it is still heterogeneous in nature in terms of both type, source, and quality (2). General, primary repositories, such as NCBI, will collect data from a variety of experimental designs and in some sections of the database, perform only minimal automated curation. Robust software that pre-processes this data and connects researchers directly to analytical frameworks can build a foundation to accelerate discovery, particularly in organisms without a well resolved reference genome. While reference genomes remain limited when compared to the biodiversity that exists, transcriptomic studies generated from high throughput sequencing technologies are available for a much wider range of species (3). Among land plants alone, less than 200 species have a complete genome compared to over 2,000 species with at least one transcriptome study (3). Comparative genomics derived from transcriptomics, specifically comparisons across orthogroups, can help us evaluate selection

pressure, rate of gene family evolution, resolve phylogenetic relationships, identify novel gene families, and assess whole or partial genome duplication events (3,4).

Orthogroups attempt to represent a set of paralogous and orthologous genes that have descended from a single gene in the last common ancestor of all the species under consideration (5). Orthologous genes have evolved from a common ancestral gene via speciation while paralogous genes result from gene duplication events. Paralogous genes are analyzed for rate of synonymous substitution per site to infer ancient Whole Genome Duplication (WGD) events (5). Synonymous substitutions are evaluated since they are not reflective of selection pressures. The estimations are challenged by degradation of the paralogous signal over time and the impact of multiple substitution rates on a single site (5). Estimating background gene duplication and loss rates within certain orthogroups throughout the species tree can be used to calculate the probability of a WGD event (5).

The evolutionary history of land plants has been shaped by multiple whole and partial duplication events. Many angiosperm lineages have experienced multiple events of WGD genome duplication and orthogroups were informative in characterizing these events (6). WGD events have been rampant concerning angiosperm species that have been domesticated for agricultural purposes. Domestication is defined to be the breeding of wild species with specific variants that result in desirable phenotypic traits. This is done through artificial selection by cultivating variants responsible for producing favorable phenotype. It has been observed that domesticated species have distinct genotypic and phenotypic signatures. Crop species, which are angiosperms have experienced more WGD events due to domestication compared to their wild-type (7). Due to their recent evolution, detection of WGD events in angiosperms has been more detectable (4). However, Gymnosperms until recently, were thought to have few to no WGD

events. With the construction of orthogroups and inference of gene family phylogenies, it was revealed that three very ancient genome duplications may have contributed to the evolution of conifers and other gymnosperms (8). Hence, orthogroups play a significant role in the detection of WGD events, establishing phylogenetic relationships, and understanding gene and genome evolution. Importantly, they can be used with or without a reference genome.

Numerous tools have emerged for discovering and analyzing orthogroups including: OrthoFinder, OrthoMCL, and TRIBE-MCL (9,10,11). These applications conduct pairwise sequence similarity searches against proteomes available for the species of interest, followed by a clustering step that develops an orthogroup graph (9). This leads to the formation of orthogroups and inference of relationship amongst numerous gene trees. Reconciliation of gene trees then leads to species tree, depicting the phylogenetic relationship between species. OrthoFinder specifically corrects for bias imposed by gene length. Genes with reduced length may have a lower similarity search score, impacting their ability to cluster with other genes (9). OrthoFinder is ideal for transcriptomes since the de novo assembly process often generates numerous partial genes. This is also the case for early (draft) genome assemblies (12).

Such applications can also be used for the discovery of single-copy orthologs. Correct identification of single copy orthologs can be used for large phylogenetic reconstructions and can be classified for their functional relevance (13). Additionally, conserved single-copy orthologs can be used as quantitative indicators of genome completeness. Applications, such as BUSCO (Benchmarking Universal Single-Copy Orthologs), assess gene space and/or transcriptome completeness using genes expected to contain evolutionary information (14). These genes are derived from a pre-computed orthogroup resource, OrthoDB (15). This tool, and others, such as EggNOG-mapper and Clusters of Orthologous Groups (COG) allow users to interact with pre-

computed orthogroup databases that are generated from reference genomes (16,17). This is limiting since these resources depend on gene annotations derived exclusively from high quality reference genome assemblies.

We observe the limitation of OrthoDB when assessing genome annotation completeness for three non-model conifer species, *Pseudotsuga menziesii* (Douglas fir), *Pinus taeda* (loblolly pine), and *Pinus lambertiana* (sugar pine). Conifers are non-models not only due to lack of resources but also because their genomes are incredibly large and complex. With genome size ranging from 10 to 40 Gbps where much of genome is repetitive content, finding protein coding gene models is a computational challenge (18). The gene space is further convoluted with high prevalence of pseudogenes and uncharacteristic gene structure such as introns being 800 kbps long (18). Despite these complexities, we successfully annotated the three conifer species mentioned above.

Annotation of these three complex conifer genomes was achieved using a novel pipeline called Braker which wraps around two programs GeneMark-ET and Augustus (19,20,21). GeneMark-ET is an iterative, self-training, machine learning algorithm developed for parameterizing exon/intron boundaries in a genome. The parameterization of exon/intron boundaries is initially dependent upon a set of heuristic parameters. The resulting *ab initio* gene predictions and those supported by raw RNA-seq alignments and then be used for parameter re-estimation. The algorithm continues to predict protein coding region and re-estimate parameters until parameters have converged between iterations (20). Upon convergence, the parameters are used to train the semi-hidden markov model in Augustus, for genome wide prediction of genes (19). This is necessary, otherwise only genes that are supported by RNA-seq data would be predicted. Since the transcriptome is only a snapshot of what is being expressed at specific points

in time, these set of genes may not fully represent the gene space (21). Therefore, Augustus uses parameters informed by the RNA-seq alignment regarding splice sites and leverages these parameters for *ab initio* gene prediction (22). However, it was observed that raw RNA-seq reads alone are not sufficient to represent the gene space entirely. The incorporation of protein evidence is necessary for training and prediction of more complex gene structures (genes with long introns) by Augustus.

Regardless of whether protein evidence is included, *ab initio* gene predictors inflate the gene space through a high number of false positive genes. A workflow was developed to further refine and reduce the number of gene models generated. This process implemented specific metrics to generate high quality gene models. These metrics include presence of start and stop codons, minimum exon length of 21 bps, minimum intron length of 9 bps, minimum CDS of 300 bps, and removal of genes with other invalid structures. Additionally, the gene models are examined for valid protein domains through functional annotation, followed by the removal of retrotransposon elements via domain association. Finally, overlapping gene models are merged through Bedtools to eliminate redundancy within the gene space (22).

Despite having achieved high quality gene models, almost all of which had functional assignments, BUSCO reported that all three conifer species were missing at least half of the conserved single-copy orthologs identified in the embryophyta lineage. However, OrthoDB delineates orthologs to an entire lineage using select species with well resolved genomes (15). Therefore, orthologs existing in the embryophyta lineage have been discovered using 26 angiosperms, one bryophyte, and one lycophyte. The estimated divergence time between gymnosperms and angiosperms is 250 million years and the divergence between early land plants and gymnosperms is even greater (23).

To utilize these applications locally, one must curate datasets from a variety of sources, install the computationally intensive applications on High Performance Computing (HPC) clusters, and interact with the results through third party packages. Moreover, communities that curate sequence resources for clade or model organism databases (CODs/MODs) do not have a mechanism for integrating computationally intensive analytics into their platforms. Tripal is a standardized framework that supports MODs/CODs with a focus on genetic/genomic data (24). This open source toolkit integrates a web content management system (Drupal) and Generic Model Organism Database schema (GMOD) known as Chado (23,24). Tripal facilitates connectivity and extensions in the form of community developed modules to extend the utility of data residing in the Chado database. The most recent release of Tripal provides an application programming interface (API) for the integration of data with Galaxy workflows. Galaxy is a platform for data analysis via documented workflows, built primarily with open-source bioinformatic command-line tools, to drive reproducibility in the scientific community (25). The Tripal project encourages development of customizable modules that can be shared throughout the scientific community to serve and analyze data.

In this study, we present a new Tripal module, OrthoQuery and demonstrate its utility in the context of TreeGenes, a Tripal powered database which houses genotypic and phenotypic data for over 1700 forest tree species (26). OrthoQuery provides a semi-automated analytical pipeline and visualization platform. The modules ease the burden of data curation, application installation, and compatibility of resulting files with visualization platforms. This robust and flexible Tripal module aims to enable researchers in conducting comparative genomics analysis for user selected species, with an emphasis on pre-processing transcriptomic resources to include non-model organisms.

Manuscript

OrthoQuery: A Tripal Database Module to Assess and Visualize Gene Family Evolution

Authors: Sumaira Zaman[1+], Emily Grau[1], Sean Buehler[1], Stephen Ficklin[2], Jill Wegrzyn[1+]

[1] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA

[2] Department of Horticulture, Washington State University, Pullman, WA, USA

[+]Corresponding Author(s):

Sumaira Zaman (sumaira.zaman@uconn.edu)

Jill L. Wegrzyn: jill.wegrzyn@uconn.edu

Journal: Formatted for BMC Bioinformatics

Abstract

Background: The abundance of transcriptomic resources for non-model organisms has enabled researchers to study comparative genomics on a larger scale. Generation of orthologous gene families facilitate the detection of genome duplication events and allows researchers to refine phylogenetic relationships and examine gene family evolution. Comparisons across orthogroups support analyzing selection pressure and novel gene families. Applications developed to study gene homology among species do not allow users to query data directly from external databases hosting resources not associated with a genome reference. In addition, real time computation of orthogroups for user selected subsets paired with interactive visualizations is lacking.

Results: OrthoQuery, a web-based Tripal module, provides a semi-automated analytical framework to enable comparisons among curated proteins and interactive visualizations in context of the resulting species tree. OrthoFinder, optimized with Diamond, is leveraged for protein level comparisons, and the Tripal database framework, coupled with Galaxy integration,

supports a variety of workflows and visualization options for the end users. OrthoQuery processes unigenes and stores a pre-computed set of orthogroups based on available species' resources in the local database. The module provides researchers with options to navigate the resulting species tree, identify ancestral/species-specific groups of genes, and associate orthogroups with functional annotations.

Conclusions: The OrthoQuery module can integrate with any of the over 30 Tripal supported databases. Tripal provides a standardized front and back-end environment for genetics/genomics focused repositories. Tripal's recent integration with Galaxy allows for functionality that extends beyond basic query operations. OrthoQuery provides the scientific community with a framework to access extensive resources for non-model systems, initiate large-scale comparative analysis, and interact with the results without leaving their web browser.

1. Background

The data derived from high throughput sequencing and housed in public repositories (Genbank, EMBL) is heterogeneous in nature, originating from a multitude of scientific communities and experimental designs. Construction of well designed bioinformatics software, that can leverage these diverse data sets, is critical for comparative genomics. Robust software that pre-processes this data and connects researchers directly to analytical frameworks can build a foundation to accelerate discovery, particularly in organisms without a well resolved reference genome. While reference genomes remain limited when compared to the biodiversity that exists, transcriptomic studies are available for a much wider range of species (1). Comparative genomics derived from transcriptomics, specifically comparisons across orthogroups, can help us evaluate selection pressure, rate of gene family evolution, resolve phylogenetic relationships, identify novel gene families, and assess whole or partial genome duplication events (2,3).

Orthogroups attempt to represent a set of paralogous and orthologous genes that have descended from a single gene in the last common ancestor of all the species under consideration (4). Orthologous genes have evolved from a common ancestral gene via speciation while paralogs result from gene duplication events. Existing tools, including: OrthoFinder, OrthoMCL, and TRIBE-MCL have emerged for discovering and analyzing orthogroups. These applications conduct pairwise sequence similarity searches against proteomes available for the species of interest, followed by a clustering step that develops an orthogroup graph (4,5,6). OrthoFinder specifically corrects for bias imposed by gene length. Genes with reduced length may have a lower similarity search score, impacting their ability to cluster with other genes (3). OrthoFinder is ideal for transcriptomes since the de novo assembly process often generates numerous partial genes. This is also the case for early (draft) genome assemblies (7).

Conserved single-copy orthologs can be used as quantitative indicators of genome completeness. Applications, such as BUSCO (Benchmarking Universal Single-Copy Orthologs), assess gene space and/or transcriptome completeness using genes expected to contain evolutionary information (8). These genes are derived from a pre-computed orthogroup resource, OrthoDB (9). This tool, and others, such as EggNOG-mapper and Clusters of Orthologous Groups (COG) allow users to interact with pre-computed orthogroup databases that are generated from reference genomes (10,11). This is limiting since these resources depend on gene annotations derived exclusively from reference genomes. To utilize these applications locally, one must curate datasets from a variety of sources, install the computationally intensive applications on High Performance Computing (HPC) clusters, and interact with the results through third party packages.

Communities that curate sequence resources for clade or model organism databases (CODs/MODs) do not have a mechanism for integrating computationally intensive analytics into their platforms. Tripal is a standardized framework that supports MODs/CODs with a focus on genetic/genomic data. This open source toolkit integrates a web content management system (Drupal) and Generic Model Organism Database schema (GMOD) known as Chado (12, 13). Tripal facilitates connectivity and extensions in the form of community developed modules to extend the utility of data residing in the Chado database. The most recent release of Tripal provides an application programming interface (API) for the integration of data with Galaxy workflows. Galaxy is a platform for data analysis via documented workflows, built primarily with open-source bioinformatic command-line tools, to drive reproducibility in the scientific community (14). The Tripal project encourages development of customizable modules that can be shared throughout the scientific community to serve and analyze data.

In this study, we present a new Tripal module, OrthoQuery and demonstrate its utility in the context of TreeGenes, a Tripal powered database which houses genotypic and phenotypic data for over 1700 forest tree species (15). OrthoQuery provides a semi-automated analytical pipeline and visualization platform. The modules ease the burden of data curation, application installation, and compatibility of resulting files with visualization platforms. This robust and flexible Tripal module aims to enable researchers in conducting comparative genomics analysis for user selected species, with an emphasis on pre-processing transcriptomic resources to include non-model organisms.

2. Implementation

2.1 Overview

OrthoQuery serves as the intersection between curated data from a Tripal database, executing analysis on the Galaxy backend, and delivering results along with interactive visualizations to the user at the web front-end (Figure 1). Orthoquery's pipeline begins with standardizing transcriptomic and genomic resources to ensure complete and unique protein coding genes. This provides users a centralized resource, through a Tripal database, for gathering clade specific datasets without the need for external filtering. The pipeline gathers user-specified datasets and launches one of three supported workflows through the Galaxy application server. Once the analysis has completed, OrthoQuery retrieves results from the application server and delivers them back to the Tripal web interface. Through the Tripal website profile, the user can access the analysis output as well as the interactive visualizations.

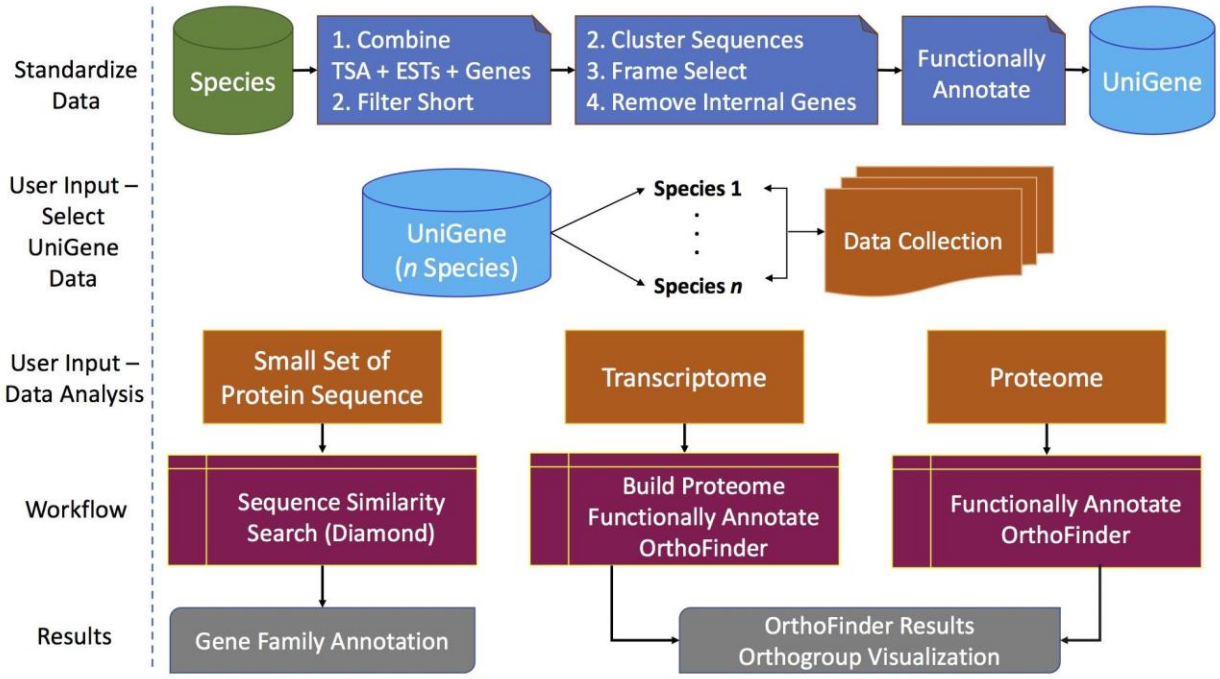

Fig 1: OrthoQuery Architecture and Workflow

2.2 Standardizing Data

OrthoQuery is responsible for creating, curating, and maintaining unigenes. The term unigene is primarily associated with NCBI's UniGene database but generally refers to sets of transcripts representing the same locus (16). Unigenes are derived from various transcriptomic and genomic resources that the database administrator can specify based upon Sequence Ontology (SO) types supported in the CHADO schema (17). Common sources of evidence may be labeled from public sources, such as Genbank: TSA (transcriptome shotgun assembly), dbEST (Expressed Sequences Tag Database), and other cDNA sources. Additionally, gene models from genome annotations can be included when available locally or through external sources. The creation of unigenes is critical since a single species may be associated with multiple transcriptomic studies representing a variety of tissue types or developmental stages.

Furthermore, the unigenes can be versioned and updated in the database when new sequences are retrieved from an external repository.

OrthoQuery is packaged with a pipeline for the creation of high quality unigene data from sequence resources available for the specified species in the CHADO database. The pipeline executes a series of filters to remove very fragmented genes and determine the coding region from the original transcripts. The remaining sequences are clustered via Vsearch at 98% identity (18). This process reduces some of the redundancy resulting from merging multiple studies. Clustered sequence sets are subsequently frame selected via GeneMarkS-T (19). GeneMarkS-T translates the transcriptome sequences using an iterative, unsupervised machine learning approach to determine the optimal frame (19). Gene models without recognized start and stop codons are removed. The final set of proteins are functionally annotated via EnTAP and loaded into CHADO via Tripal (20). Functional annotation provides information on sequence similarity, gene family assignment from pre-computed resources, Gene Ontology (GO) term assignment, protein domains, and KEGG pathway assignment terms (21,22). Pre-processing the data also includes generating a pre-computed local database of the genomic and transcriptomic resources available after unigene creation. This sets are processed when updated via OrthoFinder and the gene to orthogroup membership is stored, along with the functional annotation, in the database to support basic search operations.

2.3 Use Cases

OrthoQuery provides three specific use cases to researchers. The first use case is the simplest: the researcher has a single (or small set) of protein sequences and the goal is to determine the best orthogroup assignment for those sequences. The sequences are compared to one (or all) orthogroup sets housed in the database, through a rapid sequence similarity search

conducted by Diamond (23). Diamond is a faster alternative to BLAST for protein searches and provides comparable sensitivity. Since the functional annotation information for all unigenes are available in the Tripal database, information regarding gene families can be easily extrapolated through the protein level comparison.

The latter two use cases require a more comprehensive input such that a comparative genomic analysis can be executed in real time between the user provided species of interest and those available in the database. This use case requires the user to provide a transcriptome or set of genes that they have independently assembled. Given this set of transcripts, OrthoQuery will build a proteome through a series of steps. This is similar to the process by which the unigenes are created with the exception of gathering multiple transcriptomic resources. The final use case allows the user to provide their own proteome, which will generally result from their own downstream processing of a *de novo* transcriptome or a set of predicted gene models that have been translated regardless of whether the user provides a transcriptome or a proteome, both will be functionally annotated via EnTAP. The pairing of functional annotation to the proteins is imperative to answering biologically meaningful questions. Both of these use cases require the researcher to select the species they would like to compare with from those available as unigene. Following preparation of the input sets, the OrthoFinder run will commence.

2.4 Workflow Development & Execution

All analytical workflows, supported through Tripal modules, must be executed in a Galaxy instance. While Galaxy provides a Graphical User Interface (GUI), OrthoQuery leverages the Tripal Galaxy API to avoid redirecting users to the local Galaxy instance. Galaxy currently supports two APIs to support databases, BioBlend written in Python and blend4php written in PHP (24,25). The Tripal Galaxy module uses the blend4php API to transfer data from

one or more Tripal databases to Galaxy. This also invokes the appropriate workflow so that analysis can be performed on the Galaxy sever. By utilizing the API, the module can retrieve results from the local Galaxy application server and deliver them back to Tripal.

Independent workflows were developed for each of the three use cases since each requires different tools and parameters. In the first use case, the workflow simply confirms the appropriate input(s) from the user and executes Diamond on the application server. The second use case, involving a user provided transcriptome, is first processed via Galaxy and the resulting proteome, in addition to the selected unigenes, are compiled into a *data collection* and sent for execution via OrthoFinder (with Diamond support). The final use case can take the proteome and compile all selections (user provided and database stored) into a *data collection* and launch OrthoFinder in Galaxy. All stages of the runs are logged, including the summary outputs. OrthoFinder's processing includes formation of orthogroups, multiple sequence alignment of genes, generation of gene trees, and construction of a final species tree (4). Despite the shorter run times associated with Diamond in OrthoFinder, one can expect a few hours of processing time depending on the number of proteomes compared and the resources available on the Galaxy server. Implementation in Tripal allows OrthoQuery to provide results within a profile accessible only to the user associated with that run. The profile connects the researcher to the output files as well as the interactive visualizations.

2.5 Visualizations

OrthoFinder provides detailed logs and useful summaries for the end users. Depending on the number of species represented in the analysis, these summary files can be unwieldy for biologists to parse. In addition, there is no efficient method for connecting the resulting orthogroups with functional annotation information. OrthoQuery's visualization bridges this gap

by analyzing how different orthogroups are evolving in a species tree and providing connection to putative functional data. OrthoQuery leverages three outputs from OrthoFinder: (1) species tree in Newick format, (2) gene counts for each orthogroup, and (3) species represented in each orthogroup. Sequence source information for the unigenes and the functional data is retrieved from the database.

The visualization is presented via the website and executed with D3 to support interactivity (26). OrthoQuery displays an interactive and labeled species tree where the user can select any node and the resulting subtree will be highlighted. Summary information for the entire analysis is provided as well as the ability to download the summary files that are generated by OrthoFinder. Summary statistics presented to the user, include: percentage of genes assigned to orthogroups, total number of orthogroups, statistics regarding size and membership of orthogroups, and the number of single-copy orthologs discovered. Upon selection of a specific node on the tree, a panel depicting a histogram is displayed that quantifies five different relationships within the tree at that position (Table 1).

Table 1 - Description of orthogroup sets that can be parsed in the interactive species tree generated by OrthoQuery

Orthogroup category	Description of orthogroup
Absent	An orthogroup that is not present in the selected subtree.
Species-Specific	An orthogroup that is strictly found in one only species within the subtree and is absent elsewhere in the tree.
Clade-Specific	An orthogroup that is strictly found in all species present in the subtree and is absent elsewhere in the tree.
Ancestral	An orthogroups that is present in all descendant species of the subtree resulting from the most recent common ancestor of the selected node.
Present elsewhere in the tree	An orthogroup that is present elsewhere in the tree and in the selected subree (excluding ancestral orthogroups). These orthogroups might have been present earlier in time, had been lost, and evolved again later in time.

Once the user selects any of the five sets listed above, a second panel appears listing all the orthogroups which represent that specific relationship. Users can see the size of the orthogroups and the number of species present in that orthogroup. The size of the orthogroup can be used to evaluate gain/loss events within that gene family. Furthermore, comparing the orthogroups size and the number of species can also aid in discovering single copy orthologs. Finally, a user can choose to study a specific orthogroup by selecting it. This selection will highlight which species are present in the orthogroup and will also prompt the third panel which displays functional annotation information. The number of genes constituting an orthogroup may range from two to hundreds, however, OrthoQuery only displays functional annotation information for the most informative sequence present in the orthogroup. The functional annotation information for all the sequences in that group is available through a downloadable file. This file also contains information on the source of each gene in terms of species and unigene composition.

2.6 OrthoQuery Implementation

The front end of the user interface is developed for Drupal v.7 integrated in Tripal v.3.0. The Galaxy v18.05 instance, used in the development of OrthoQuery and supported by blend4php v0.1a, is hosted locally by TreeGenes. Processing scripts were written with Python v2.7. The local Galaxy instance is running Diamond v0.9.19 and OrthoFinder v2.1.2. Dependencies for OrthoFinder are installed and managed by the Conda environment through the Bioconda channel. The OrthoQuery visualization is developed with D3 v5.5.

3. Results & Discussion

3.1 Application in a Tripal Database

The TreeGenes database is one of the over 30 Tripal supported websites. This curated, web-based relational database houses a wide range of genetic data describing just over 1700 forest tree species representing 16 orders and 124 genera. Despite this diversity, genomes are only available for 40 species while transcriptomic resources are available for 370. Transcriptomics resources in TreeGenes are sourced primarily from Genbank submissions, and include: TSA, ESTs, cDNAs, as well as gene annotations derived from sequenced genomes.

OrthoQuery exists as an analytical tool utilizing the unigene data that resides in the TreeGenes database. The landing page of OrthoQuery asks for the type of input the user will provide, a small set of protein sequences, transcriptome, or a proteome. The user must also specify whether to select the entire unigene dataset or whether to select specific targets and the number of target species (Fig. 2A). If sub-setting the dataset, users will be redirected to select species and submit the job (Fig. 2B). The default parameters for each analysis are exposed to the user for reproducibility and troubleshooting analysis if needed.

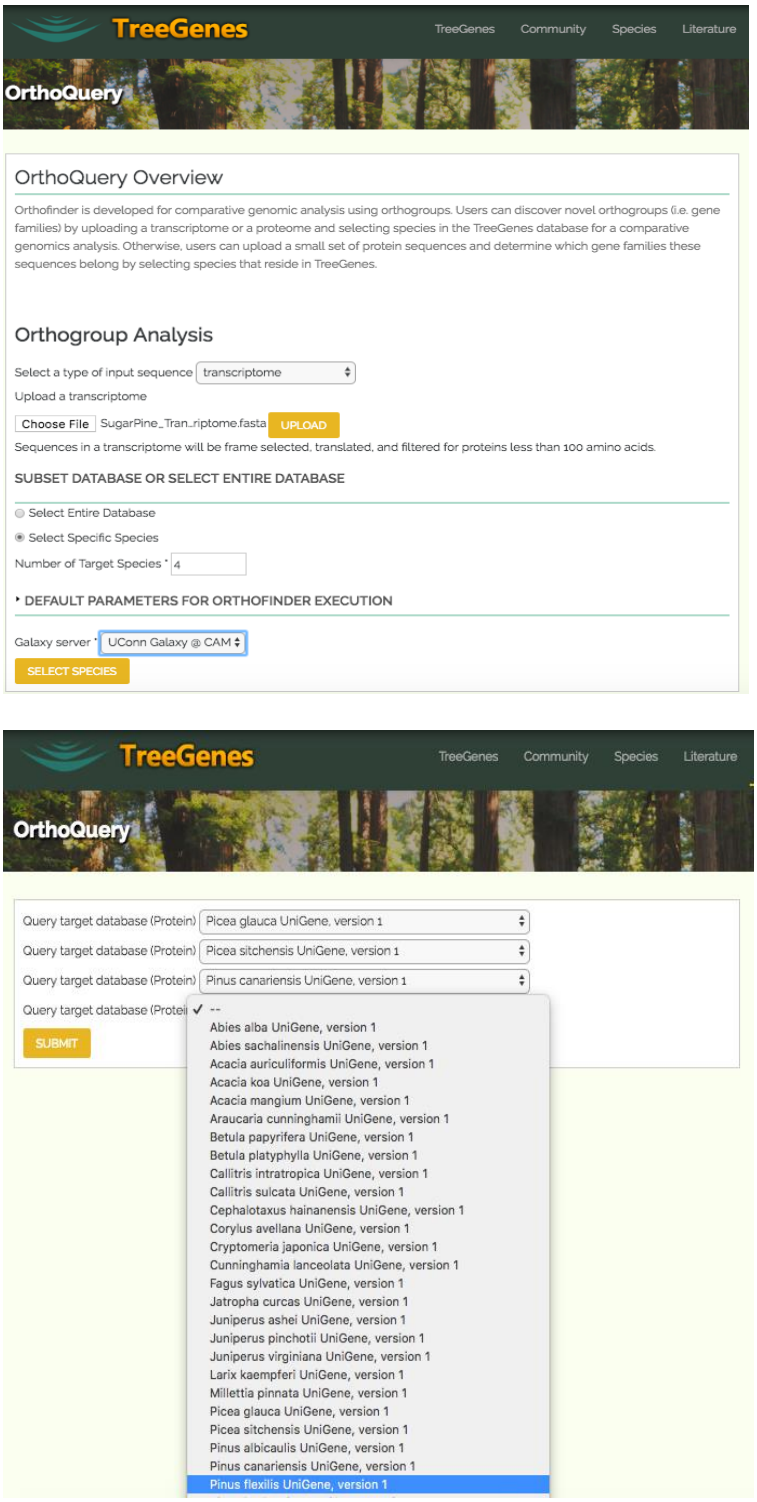

Figure 2: (A) The OrthoQuery analysis user interface landing page allows the user to select from one of three supported workflows. (B) The user can select any subset of sequences to customize their OrthoQuery run.

3.2 Analyzing Non-Model Organisms: Application in Gymnosperms

From TreeGenes, a total of 40 species representing 25 genera had transcriptomic support and were used as input to the unigene pipeline. These sequences were frame selected, clustered, and length filtered to generate a total of 21 unigene sets (Table S1). These 21 species, local to TreeGenes were selected in addition to 132 non-tree angiosperms sourced from 1KP (27). OrthoQuery is installed as a module on TreeGenes to examine species-specific families associated with gymnosperms and understand their phylogenetic relationships.

Gymnosperms appeared between 250 and 300 million years ago, are characterized as naked seed plants, and have only recently been assessed due to their large and complex genomes that range from 10 to 40 Gbp in size (28). These genomes are difficult to assemble and characterizing them through genome annotation is even more elusive. Existing reference assemblies and their associated annotations remain incomplete despite the availability of six gymnosperm reference genomes. Pre-computed orthogroup databases, such as OrthoDB, do not currently contain representation from this group. The challenges associated with the reference genome annotations in poorly characterized species may be assisted by the inclusion of transcriptomic resources from species within the same phylum, order, or genus.

OrthoQuery was executed using two different datasets. The first data set consisted only of species with an available reference genome while the second data set consisted of both species with genome annotation and these same species combined with those with a unigene set. The first dataset included two early land plants (*Physcomitrella patens* and *Selaginella moellendorffii),* five gymnosperms (*Ginkgo biloba, Picea abies, Pseudotsuga menziesii, Pinus taeda, Pinus lambertiana*) and five angiosperms (*Amborella trichopoda, Arabidopsis thaliana, Vitis vinifera, Sorghum bicolor, Oryza sativa*). The second analysis included five new species

with transcriptomic data (unigenes): *Picea sitchensis*, *Pinus patula*, *Pinus canariensis*, *Jatropha curcas* and *Quercus suber*. All putative gene models were filtered for representation of correct gene structure by custom in-house scripts.

3.2.1 Results

The dataset consisting only of species with reference genomes, resulted in an incorrect placement of *Picea abies* with *Pinus taeda* (Figure 3A). The species tree was resolved by adding unigene data from TreeGenes to support specific clades, represented in the second dataset. The additional unigene data corrected the species tree and preserved correct phylogeny (Figure 3B). The resulting species tree has specific subtrees for *Pinus* and *Picea*. Furthermore, within genus *Pinus* the sub-genus *Pinus* (*Pinus patula*, *Pinus canariensis*, and *Pinus taeda*) is also present a subtree while the sub-genus *Strobus* (*Pinus lambertiana*) is distinct. This demonstrates the need for a comprehensive, well curated data set to improve comparative genomics analysis across non-model species.

The OrthoQuery run produces a summary of the OrthoFinder analysis. From the OrthoQuery visualization summary, we learn that 81.5% of the genes were assigned to 20,783 orthogroups. Fifty percent of all the genes were in orthogroups with 37 or more genes and were contained in the largest 2,252 orthogroups. There are 1,356 species specific orthologs in the entire tree. Selecting the ancestral node that gave rise to the gymnosperms (Fig. 4) summarizes the following relationship between gymnosperms and the remaining tree (Table 2).

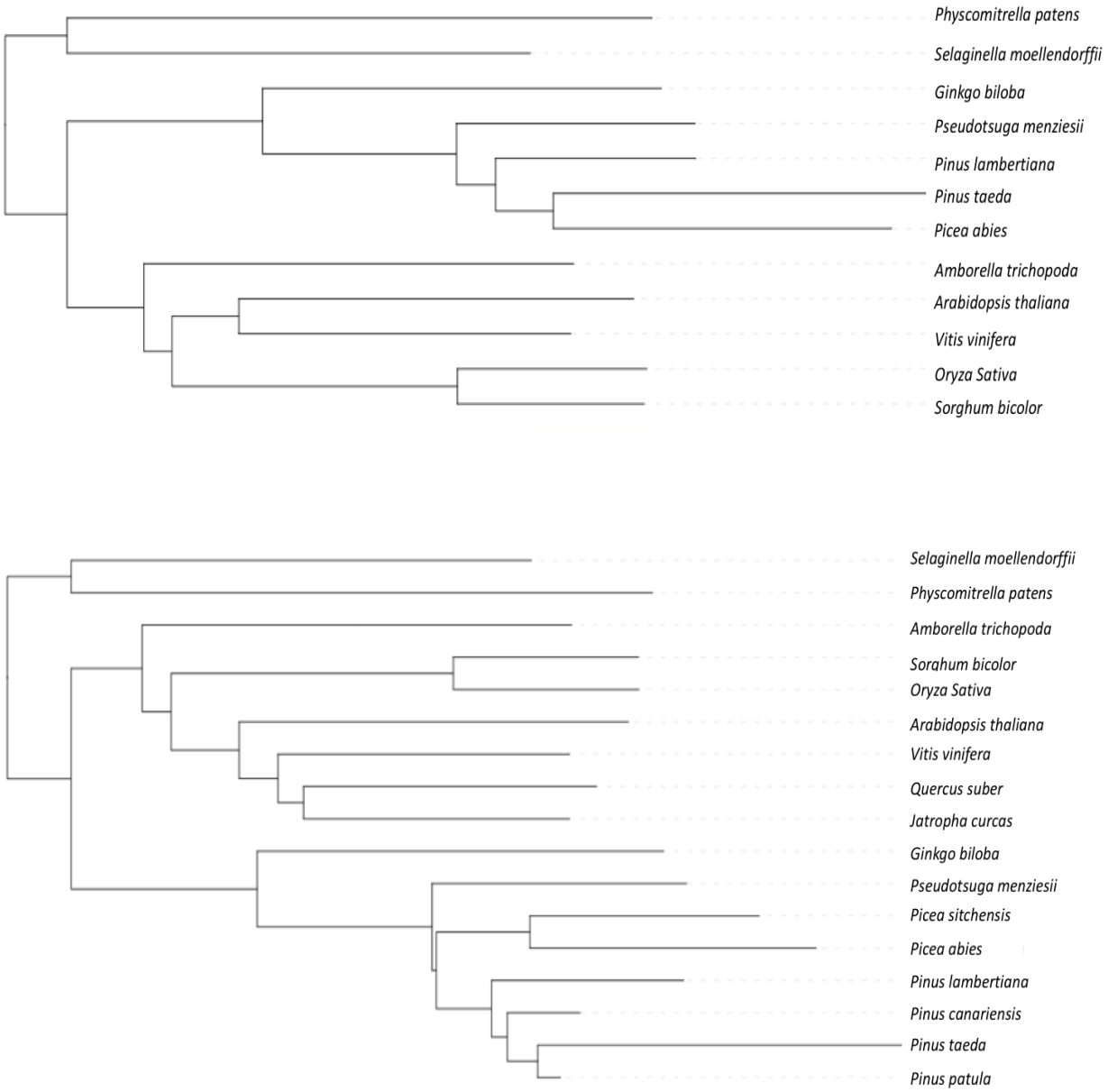

Fig 3: (A) Resulting species tree when only comparing species with reference genomes available. (B) Resulting species tree when including additional species from unigene, possessing high quality transcriptomic data while lacking a reference genome.

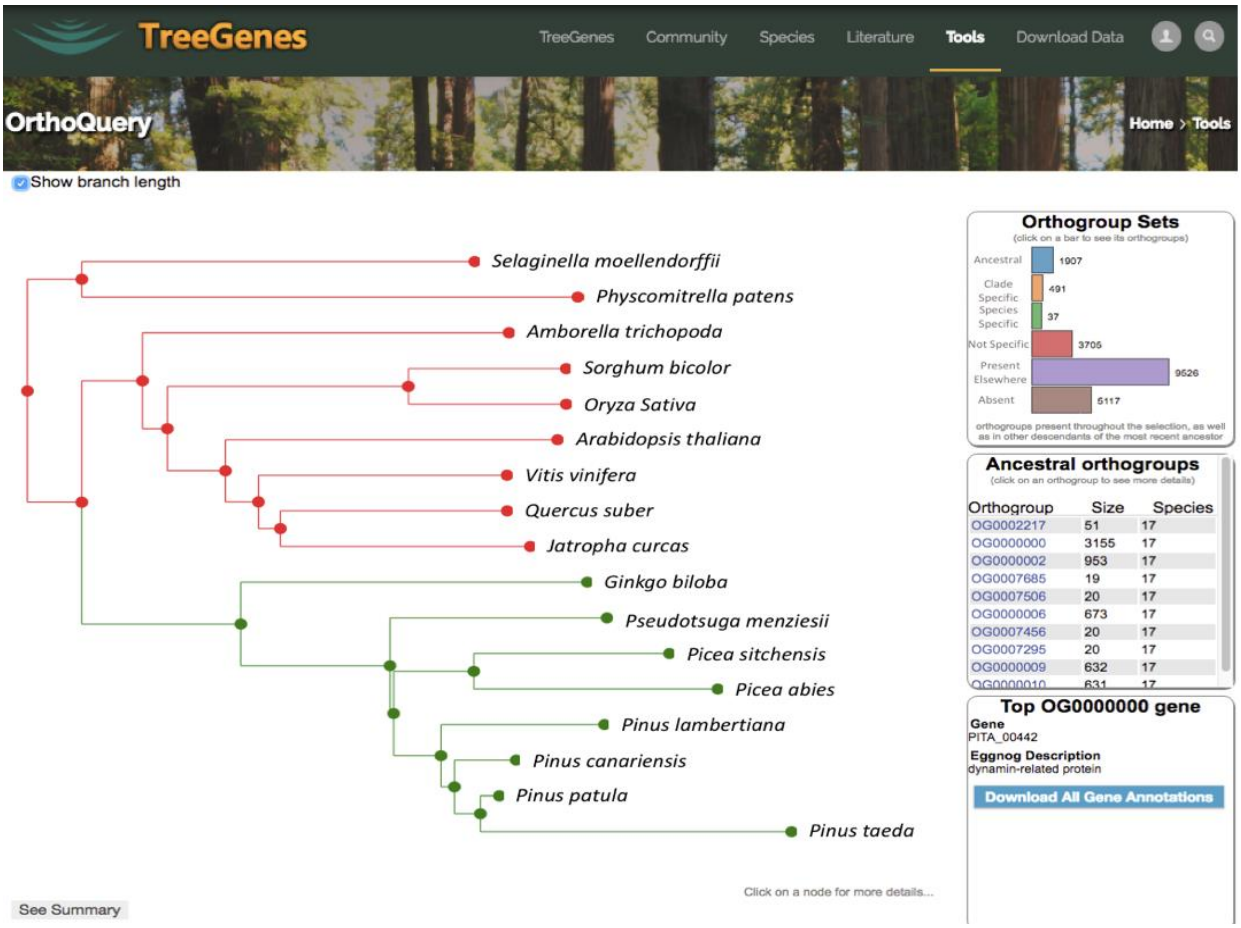

Fig 4: Interactive visualization supported by OrthoQuery via D3.

Table 2: OrthoQuery results from TreeGenes.

Type of Orthogroup	Number of Orthogroups
Ancestral	1907
Species specific	491
Clade specific	37
Absent	5117
Present elsewhere in the tree	9526

4. Conclusion

OrthoQuery sits at the intersection of the data repository and the analytic software. The OrthoQuery module identifies orthologous genes via a Tripal database, standardizes the data for comparative analysis, performs analysis through the Tripal Galaxy API with OrthoFinder, sends the data to the user's database profile, and provides interactive visualizations. Visualization features focus on facilitating the interrogation of large gene families, examining relationships among families, and allowing direct query of the stored orthogroups. OrthoQuery was demonstrated in the TreeGenes database in order to assess orthogroups in gymnosperms when compared to other land plants. The module is extensible to any Tripal genomics databases running with Galaxy integration.

5. Appendix

5.1 Source Code

https://gitlab.com/TreeGenes/orthoquery

5.2 Documentation

Installation of OrthoQuery can be found here: http://tripal.info/extensions/modules/orthoquery

5.3 Supplementary Table

Table S1: Availability of genome and transcriptomic data in TreeGenes.

Species	Genome Annotation	TSA	EST	Unigene
1. *Acacia koa*	x	91069	x	24196
2. Salix integra	x	79977	x	20871
3. Pseudotsuga menziesii	52,865	331725	3755	67214
4. Betula papyrifera	x	275545	x	60044
5. Pinus lambertiana	39,443	33112	x	32579
6. Cryptomeria japonica	x	9966	19994	12241
7. Pinus monticola	x	65191	x	23796
8. Wollemia nobilis	x	41289	x	12173
9. Picea sitchensis	x	18688	19999	14522
10. Pinus patula	x	105454	23	40400
11. Araucaria cunninghamii	x	80474	x	19665
12. Pinus albicaulis	x	357872	x	73958
13. Quercus suber	x	87826	6698	40440
14. Picea glauca	567	455504	39999	64490
15. Tectona grandis	x	237418	9	60276
16. Fagus sylvatica	x	151667	10000	14559
17. Millettia pinnata	x	53586	x	22438
18. Pinus canariensis	x	92641	x	37016
19. Pinus massoniana	x	274404	124	64540
20. Cephalotaxus hainanensis	x	49355	x	23140
21. Jatropha curcas	57437	91954	9967	42306

6. Introduction References

1. “High-Throughput Sequencing Technologies.” Egyptian Journal of Medical Human Genetics, Elsevier, 21 May 2015, www.sciencedirect.com/science/article/pii/S1097276515003408?via=ihub.
2. Clark, et al. “GenBank | Nucleic Acids Research | Oxford Academic.” OUP Academic, Oxford University Press, 20 Nov. 2015, academic.oup.com/nar/article/44/D1/D67/2503088.
3. Bolger, Marie E., et al. “Plant Genome and Transcriptome Annotations: from Misconceptions to Simple Solutions”, OUP Academic, Oxford University Press, 5 Jan. 2017, academic.oup.com/bib/article/2843630.
4. McKain, Michael R., et al. “Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales | Genome Biology and Evolution | Oxford Academic.” OUP Academic, Oxford University Press, 17 Mar. 2016, academic.oup.com/gbe/article/8/4/1150/2574085/.
5. Tiley, et al. “Evaluating and Characterizing Ancient Whole-Genome Duplications in Plants with Gene Count Data | Genome Biology and Evolution | Oxford Academic.” OUP Academic, Oxford University Press, 17 Mar. 2016, academic.oup.com/gbe/article/8/4/1023/2574077.
6. McKain, Michael R., et al. “Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales | Genome Biology and Evolution | Oxford Academic.” OUP Academic, Oxford University Press, 17 Mar. 2016, academic.oup.com/gbe/article/8/4/1150/2574085/A-Phylogenomic-Assessment-of-Ancient-Polyploidy.
7. Salman-Minkov, Ayelet, et al. “Whole-Genome Duplication as a Key Factor in Crop Domestication.” Nature News, Nature Publishing Group, 1 Aug. 2016, www.nature.com/articles/nplants2016115.
8. Li, Zheng, et al. “Early Genome Duplications in Conifers and Other Seed Plants.” Science Advances, American Association for the Advancement of Science, 1 Nov. 2015, advances.sciencemag.org/content/1/10/e1501084.
9. Emms, David M., and Steven Kelly. “OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy.” Genome Biology, BioMed Central, 6 Aug. 2015, genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0721-2
10. Li, Li, et al. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” Genome Research, Cold Spring Harbor Lab, 1 Jan. 1970, genome.cshlp.org/content/13/9/2178.full.
11. Enright, A J, et al. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” Nucleic Acid Research., U.S. National Library of Medicine, 1 Apr. 2002, www.ncbi.nlm.nih.gov/pubmed/11917018.

12. https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/
13. Creevey, Christopher J., et al. “Identifying Single Copy Orthologs in Metazoa.” PLOS Medicine, Public Library of Science, 1 Dec. 2011, journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002269.
14. Sim, et al. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs | Bioinformatics | Oxford Academic.” OUP Academic, Oxford University Press, 9 June 2015, academic.oup.com/bioinformatics/article/31/19/3210/211866.
15. Zdobnov, Evgeny M., et al. “OrthoDB v9.1: Cataloging Evolutionary and Functional Annotations for Animal, Fungal, Plant, Archaeal, Bacterial and Viral Orthologs.” Nucleic Acid Research., U.S. National Library of Medicine, 4 Jan. 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5210582/.
16. Huerta-Cepas, J, et al. “Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper.” Nucleic Acid Research., U.S. National Library of Medicine, 1 Aug. 2017, www.ncbi.nlm.nih.gov/pubmed/28460117.
17. Tatusov, R L, et al. “The COG Database: an Updated Version Includes Eukaryotes.” Nucleic Acid Research., U.S. National Library of Medicine, 11 Sept. 2003, www.ncbi.nlm.nih.gov/pubmed/12969510.
18. Torre, Amanda R. De La, et al. “Insights into Conifer Giga-Genomes.” Plant Physiology, American Society of Plant Biologists, 1 Dec. 2014, www.plantphysiol.org/content/166/4/1724.
19. Hoff, Katrina J, et al,. "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS". Bioinformatics, 2015, Nov 11.
20. Alexandre, et al. “Integration of Mapped RNA-Seq Reads into Automatic Training of Eukaryotic Gene Finding Algorithm | Nucleic Acids Research | Oxford Academic.” OUP Academic, Oxford University Press, 2 July 2014, nar.oxfordjournals.org/content/early/2014/06/30/nar.gku557.long.
21. Mario, et al. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts | Nucleic Acids Research | Oxford Academic.” OUP Academic, Oxford University Press, 1 July 2006, academic.oup.com/nar/article/34/suppl_2/W435/2505582.
22. Quinlan, et al. “BEDTools: a Flexible Suite of Utilities for Comparing Genomic Features | Bioinformatics | Oxford Academic.” PLOS Biology, Public Library of Science, 28 Jan. 2010, doi.org/10.1093/bioinformatics/btq033.
23. Ying Lu,Jin-Hua Ran,Dong-Mei Guo,Zu-Yu Yang,Xiao-Quan Wang. “Phylogeny and Divergence Times of Gymnosperms Inferred from Single-Copy Nuclear Genes.” PLOS Medicine, Public Library of Science, 2014, journals.plos.org/plosone/article?id=10.1371/journal.pone.0107679.
24. Ficklin, Stephen P., et al. “Tripal: a Construction Toolkit for Online Genome Databases.” Database: The Journal of Biological Databases and Curation, Oxford University Press, 29 Sept. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3263599/.

25. Jung, Sook, et al. “Chado Use Case: Storing Genomic, Genetic and Breeding Data of Rosaceae and Gossypium Crops in Chado.” Nucleic Acid Research., U.S. National Library of Medicine, 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4795932/.
26. Afgan, Enis, et al. “The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update.” Nucleic Acid Research., U.S. National Library of Medicine, 8 July 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4987906/.
27. re3data.org: TreeGenes; editing status 2017-04-26; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3NQ19 last accessed: 2018-08-06

7. Manuscript References

1. Bolger, Marie E., et al. “Plant Genome and Transcriptome Annotations: from Misconceptions to Simple Solutions”, OUP Academic, Oxford University Press, 5 Jan. 2017, academic.oup.com/bib/article/2843630.
2. McKain, Michael R., et al. “Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales | Genome Biology and Evolution | Oxford Academic.” OUP Academic, Oxford University Press, 17 Mar. 2016, academic.oup.com/gbe/article/8/4/1150/2574085/.
3. Bolger, Marie E., et al. “Plant Genome and Transcriptome Annotations: from Misconceptions to Simple Solutions”, OUP Academic, Oxford University Press, 5 Jan. 2017, academic.oup.com/bib/article/2843630.
4. Emms, David M., and Steven Kelly. “OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy.” Genome Biology, BioMed Central, 6 Aug. 2015, genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0721-2
5. Li, Li, et al. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” Genome Research, Cold Spring Harbor Lab, 1 Jan. 1970, genome.cshlp.org/content/13/9/2178.full.
6. Enright, A J, et al. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” Nucleic Acid Research., U.S. National Library of Medicine, 1 Apr. 2002, www.ncbi.nlm.nih.gov/pubmed/11917018.
7. https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/
8. Sim, et al. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs | Bioinformatics | Oxford Academic.” OUP Academic, Oxford University Press, 9 June 2015, academic.oup.com/bioinformatics/article/31/19/3210/211866.
9. Zdobnov, Evgeny M., et al. “OrthoDB v9.1: Cataloging Evolutionary and Functional Annotations for Animal, Fungal, Plant, Archaeal, Bacterial and Viral Orthologs.” Nucleic Acid Research., U.S. National Library of Medicine, 4 Jan. 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5210582/.
10. Huerta-Cepas, J, et al. “Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper.” Nucleic Acid Research., U.S. National Library of Medicine, 1 Aug. 2017, www.ncbi.nlm.nih.gov/pubmed/28460117.
11. Tatusov, R L, et al. “The COG Database: an Updated Version Includes Eukaryotes.” Nucleic Acid Research., U.S. National Library of Medicine, 11 Sept. 2003, www.ncbi.nlm.nih.gov/pubmed/12969510.

12. Ficklin, Stephen P., et al. “Tripal: a Construction Toolkit for Online Genome Databases.” Database: The Journal of Biological Databases and Curation, Oxford University Press, 29 Sept. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3263599/.
13. Jung, Sook, et al. “Chado Use Case: Storing Genomic, Genetic and Breeding Data of Rosaceae and Gossypium Crops in Chado.” Nucleic Acid Research., U.S. National Library of Medicine, 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4795932/.
14. Afgan, Enis, et al. “The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update.” Nucleic Acid Research., U.S. National Library of Medicine, 8 July 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4987906/.
15. re3data.org: TreeGenes; editing status 2017-04-26; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3NQ19 last accessed: 2018-08-06
16. UniGene. 2018. https://www.ncbi.nlm.nih.gov/unigene
17. Karen Eilbeck, et al. “The Sequence Ontology: a Tool for the Unification of Genome Annotations.” Genome Biology, BioMed Central, 29 Apr. 2005, genomebiology.biomedcentral.com/articles/10.1186/gb-2005-6-5-r44.
18. Rognes T, et al. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584 https://doi.org/10.7717/peerj.2584
19. Tang S, et al. “Identification of protein coding regions in RNA transcripts.” Nucleic Acids Research. 13 July 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4499116/.
20. Hart, Alexander J, et al. EnTAP: Bringing Faster and Smarter Functional Annotation to Non-Model Eukaryotic Transcriptomes. BioRxiv, 24 Apr. 2018, www.biorxiv.org/content/biorxiv/early/2018/04/24/307868.full.pdf.
21. The Gene Ontology Consortium. “Expansion of the Gene Ontology Knowledgebase and Resources.” Nucleic Acid Research., U.S. National Library of Medicine, 4 Jan. 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5210579/.
22. Afgan, Enis, et al. “The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update.” Nucleic Acid Research., U.S. National Library of Medicine, 8 July 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4987906/.
23. Buchfink, Benjamin, et al. “Fast and Sensitive Protein Alignment Using DIAMOND.” Nature News, Nature Publishing Group, 17 Nov. 2014, www.nature.com/articles/nmeth.3176.
24. Leo, Simone, et al. BioBlend.objects: Metacomputing with Galaxy. Bioinformatics, Oct. 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4173020/.
25. Wytko, Connor, et al. “blend4php: a PHP API for Galaxy | Database | Oxford Academic.” OUP Academic, Oxford University Press, 10 Jan. 2017, academic.oup.com/database/article/doi/10.1093/database/baw154/2884891.
26. D3. 2018. https://d3js.org/
27. Van, M, et al. “PLAZA 4.0: an Integrative Resource for Functional, Evolutionary and Comparative Plant Genomics.” Advances in Pediatrics., U.S. National Library of Medicine, 4 Jan. 2018, www.ncbi.nlm.nih.gov/pubmed/29069403.

28. Nystedt, Björn, et al. “The Norway Spruce Genome Sequence and Conifer Genome Evolution.” Nature News, Nature Publishing Group, 22 May 2013, www.nature.com/articles/nature12211.