W78–W84 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* *Published online 11 May 2015*
doi: 10.1093/nar/gkv487

OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species

Yi Wang[1,2,3,*], Devin Coleman-Derr[4], Guoping Chen[3] and Yong Q. Gu[1,*]

[1]USDA-ARS, Western Regional Research Center, Crop Improvement and Genetics Research Unit, Albany, CA 94710, USA, [2]Department of Plant Sciences, University of California, Davis, CA 95616, USA, [3]Bioengineering College, Campus A, Chongqing University, Chongqing 400030, China and [4]USDA-ARS, Plant Gene Expression Center, Albany, CA 94710, USA

Received February 08, 2015; Revised April 28, 2015; Accepted May 02, 2015

ABSTRACT

Genome wide analysis of orthologous clusters is an important component of comparative genomics studies. Identifying the overlap among orthologous clusters can enable us to elucidate the function and evolution of proteins across multiple species. Here, we report a web platform named OrthoVenn that is useful for genome wide comparisons and visualization of orthologous clusters. OrthoVenn provides coverage of vertebrates, metazoa, protists, fungi, plants and bacteria for the comparison of orthologous clusters and also supports uploading of customized protein sequences from user-defined species. An interactive Venn diagram, summary counts, and functional summaries of the disjunction and intersection of clusters shared between species are displayed as part of the OrthoVenn result. OrthoVenn also includes in-depth views of the clusters using various sequence analysis tools. Furthermore, OrthoVenn identifies orthologous clusters of single copy genes and allows for a customized search of clusters of specific genes through key words or BLAST. OrthoVenn is an efficient and user-friendly web server freely accessible at http://probes.pw.usda.gov/OrthoVenn or http://aegilops.wheat.ucdavis.edu/OrthoVenn.

INTRODUCTION

Orthologs or orthologous genes are clusters of genes in different species that originated by vertical descent from a single gene in the last common ancestor (1). Comparative analysis of the organization of orthologous clusters is important for understanding the rules of genome structure and gene/protein function. The information gained from comparisons of orthologous clusters can serve as raw material for taxonomic classification and phylogenetic studies of organisms, thereby shedding light on the mechanisms underlying the molecular evolution of genes and genomes (2,3). Recent advances in genome sequencing technologies has provided a wealth of genome sequence data from many organisms. The increasing availability of genome sequence data across the tree of life now makes it possible to conduct whole-genome comparative analyses of orthologous clusters across multiple species.

In order to identify orthologous genes from different genomes for classification within gene clusters, databases have employed different approaches that can be generally classified into two groups. One group is based on pairwise sequence comparisons (e.g. eggNOG (4), InParanoid (5), OrthoDB (6)), while the other uses phylogenetic methods (e.g. MetaPhOrs (7), PhylomeDB (8)). These analysis tools are well known and widely used, but their online servers are often database-oriented and focused on gene searches and analysis within specific orthologous groups, and lack the functionality to generate visualizations displaying the difference and overlapping for all orthologous clusters. Some of these databases also provide ortholog prediction software (InParanoid: http://software.sbc.su.se/cgi-bin/request.cgi?project=inparanoid, OrthoDB: http://www.orthodb.org/orthodb_software/), but generally the software has to be downloaded and run locally. The Quest for Orthologs Consortium (http://questfororthologs.org/) improves and standardizes orthology predictions and provides a list of >30 of these databases.

The tools for establishing homologies between genes or their products are becoming increasingly important to transfer knowledge from well-studied model organisms to other organisms (9). One of the simplest but most useful methods of genome wide orthologous comparison is to display the different and overlapping

[*]To whom correspondence should be addressed. Tel: +1 510 509 6146; Fax: +1 510 559 5818; Email: Yi.Wang@ars.usda.gov
Correspondence may also be addressed to Yong Q. Gu. Tel: +1 510 509 9055; Fax: +1 510 559 5818; Email: Yong.Gu@ars.usda.gov

Table 1. Total numbers of categorized protein sequences in OrthoVenn

Category	Number of species	Number of protein sequences
Vertebrates	69	1 276 453
Metazoa	55	1 020 988
Protists	32	458 802
Fungi	52	567 086
Plants	38	1 321 298
Bacteria	26	56 830
Total	272	4 701 457

orthologous clusters in a Venn diagram, which in our case provides circles or other shapes representing each species with overlapping regions that illustrate the genes or gene clusters that are unique to or shared between each species. As an example, a recent analysis of the banana genome (http://www.promusa.org/blogpost174-The-best-genomics-Venn-diagram-ever-deconstructed), presented such a Venn diagrams for orthologous clusters comparison among six plant genomes. Venn diagrams allow for quick visualization of relationships by revealing intersections (overlaps) and disjunctions (non-overlaps) for large biological datasets obtained from different species, and are often used in the whole-genome analysis across species (10–12). Currently, a number of online Venn diagram applications have been developed to provide simultaneous visual interpretation of large amounts of biological data. The Pangloss Venn diagram generator (http://www.pangloss.com/seidel/Protocols/venn4.cgi) and Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html) are web applications that can create Venn diagrams from user-provided ID lists. BioVenn (13) provides a comparison and visualization of biological lists using area-proportional Venn diagrams. GeneVenn (14) and VennMaster (15) possess the additional feature of linking genes within each group to related information in the NCBI Entrez Nucleotide database or the Gene Ontology database. NetVenn (16) compares and analyzes gene lists by combining a Venn diagram visualization with an interactome network and biological annotation data. A database named EDGAR provides Venn diagrams of the common gene pools for the comparative analysis of prokaryotic genomes (17).

To our knowledge, a web application that offers genome wide comparison and analysis of orthologous clusters across multiple species is not available. SPOCS (18) provides a web server for prediction of orthologs and paralogs among closely related genomes, however, it's simple visualization function does not allow for comparison of overlapping orthologous clusters among species. Here, we present a web platform named OrthoVenn (http://probes.pw.usda.gov/OrthoVenn or http://aegilops.wheat.ucdavis.edu/OrthoVenn) for the comparison and analysis of genome wide orthologous clusters across multiple species. In OrthoVenn, users can select protein sequence data for genome wide comparisons from 272 species in the database, including vertebrates, metazoa, protists, fungi, plants and bacteria. OrthoVenn also allows user-defined species to be uploaded as customized protein sequences. An efficient and interactive graphics tool is employed to provide a Venn diagram view of the genome wide comparison of orthologous clusters based on the protein sequence data selected from up to six species. The intersection of orthologous clusters is analyzed by GO Slim annotation and UniProt search. In the output of OrthoVenn, each orthologous cluster provide sequence analysis data, single copy gene cluster identification, protein similarity comparisons, and the phylogenetic relationships among clustered genes. OrthoVenn also provides key word search and BLAST functions for finding clusters of specific interest to the user. In addition, OrthoVenn allows the user to create Venn diagrams from orthologous cluster files generated by other software.

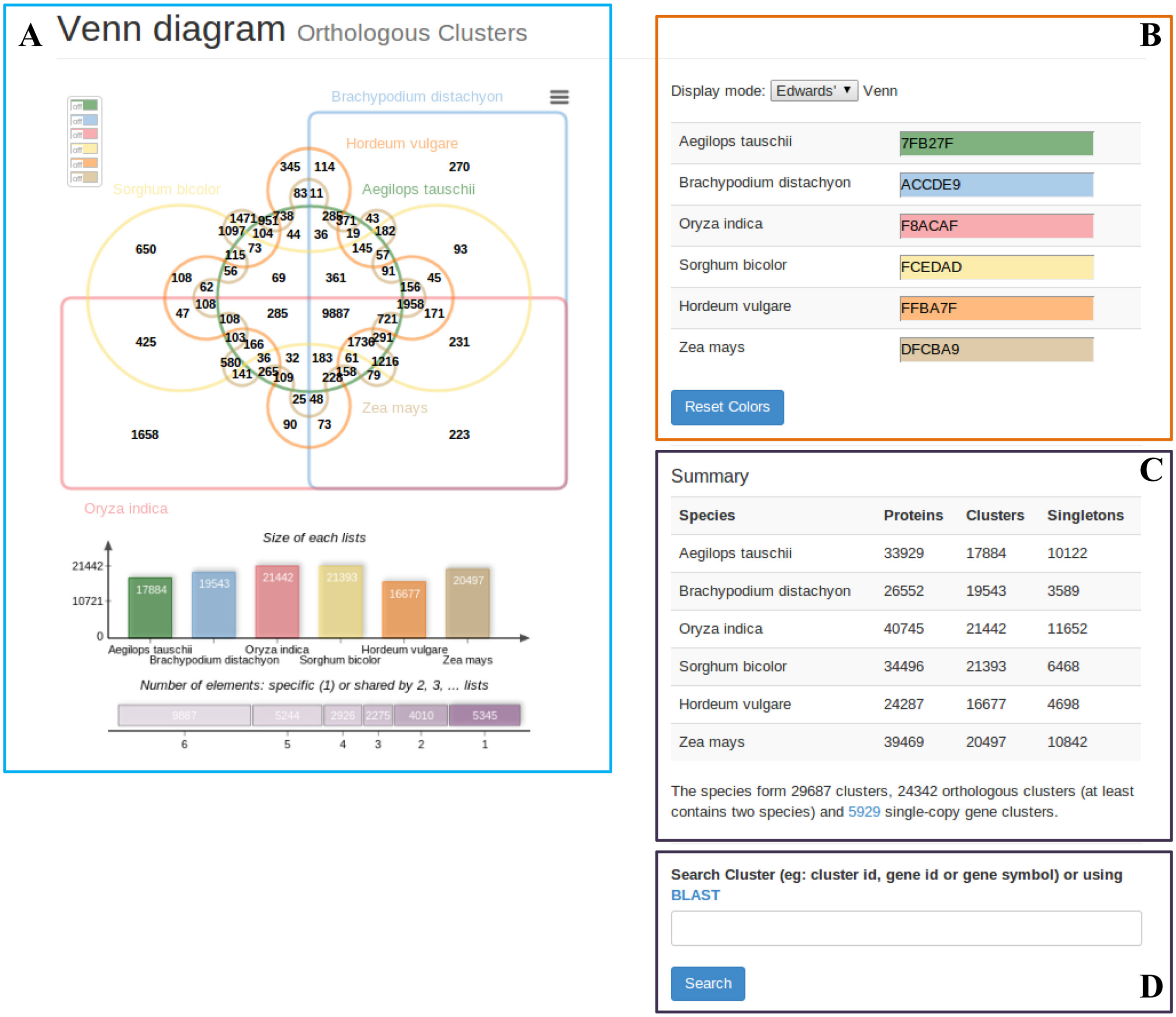
WEB SERVER CONSTRUCTION

Protein sequences collection

We downloaded sequence data from the Ensembl database, which is a genomic interpretation system providing the most up-to-date annotations of whole-genome protein sequences for many species (19). After quality analysis (sequences contains illegal characters or $<$20 amino acids were removed) of the amino acid fasta file, we retrieved the protein sequences of all gene coding sequences without alternative splice variants. We grouped the protein sequences into six categories: vertebrates, metazoa, protists, fungi, plants and bacteria. A summary of the collected protein sequences in each category is presented in Table 1. OrthoVenn also allows user-defined protein sequences in fasta or compressed fasta (.tar.gz and .zip) format for any non-represented species. Each of these novel species can be given their own name so that the user can immediately see which part of the output in the Venn diagram corresponds to which species.

Identification of orthologous cluster

There are multiple methods for orthology prediction. We used the popular heuristic approach named OrthoMCL (20) to identify ortholog groups. The OrthoMCL performs an all-against-all BLASTP alignment, identifies putative orthology and inparalogy relationships with the Inparanoid algorithm (21) and generates disjoint clusters of closely related proteins with the Markov Clustering Algorithm (MCL) (22). The steps in the OrthoMCL are time-consuming for large numbers of protein sequences from multiple species. In order to make the approach suitable for the web application, we modified portions of the data processing steps in the OrthoMCL algorithm. First, we used UBLAST (v7.0.1090) (23) to do the all-against-all similarity search. UBLAST is ∼350× faster than BLASTP and achieves very similar results for ortholog searches (24,25). Second, each of the protein sequences downloaded from Ensembl have been aligned with one another and the result



Display mode: Edwards' ▾ Venn

Aegilops tauschii	7FB27F
Brachypodium distachyon	ACCDE9
Oryza indica	F8ACAF
Sorghum bicolor	FCEDAD
Hordeum vulgare	FFBA7F
Zea mays	DFCBA9

Reset Colors

Summary

Species	Proteins	Clusters	Singletons
Aegilops tauschii	33929	17884	10122
Brachypodium distachyon	26552	19543	3589
Oryza indica	40745	21442	11652
Sorghum bicolor	34496	21393	6468
Hordeum vulgare	24287	16677	4698
Zea mays	39469	20497	10842

The species form 29687 clusters, 24342 orthologous clusters (at least contains two species) and 5929 single-copy gene clusters.

Search Cluster (eg: cluster id, gene id or gene symbol) or using BLAST

Search

Figure 1. A results page in OrthoVenn. (**A**) Venn diagram showing the distribution of shared gene families (orthologous clusters) among *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Hordeum vulgare* and *Zea mays*. The cluster number in each component is listed. (**B**) Color selector and display mode setting for the Venn diagram. (**C**) Counts of clusters in each genome. (**D**) Key word search and BLAST links for finding specific clusters in the result.

is stored in the server as BLAST tab format data (∼1.6TB). If the input species for a new OrthoVenn analysis is contained in the previously sorted alignment table, the step for sequence comparison will be omitted in the process. Finally, we used orthAgogue (v1.0.3) for identification of putative orthology and inparalogy relations. The orthAgogue is a multithreaded C application for high-speed estimation of homology relations in massive datasets (26). We have compared the orthologous clustering results between OrthoMCL and OrthoVenn, and they have similar clustering behavior (Supplemental Material).

Orthologous cluster annotation

To deduce the putative function of each ortholog, the first protein sequence in each cluster are subjected to BLASTP analysis against the non-redundant protein database in UniProt (27). We only selected one sequence in each cluster because the proteins are very similar and performing BLASTP on all sequences in each cluster against UniProt is computationally too expensive, especially for large genomes which contain many protein sequences. Any hit with an e-value <1e−5 is considered and the top hit is defined as the putative function of each cluster. The GOSlim terms for biological process, molecular function, and cellular component categories are assigned to the corresponding orthologous cluster. The GOSlimViewer (28) is used to provide a high-level summary of functions for orthologous clusters in the Venn diagram overlapping regions using pre-computed GO Slim sets. An alignment of the protein sequences in each cluster is performed using the multiple sequence alignment tool MUSCLE (29). The conserved motifs in the protein sequences are identified by Multiple Expectation Maximization for Motif Elicitation program (MEME) (30). Phylogenetic trees for the orthologous clusters are built in FastTree (31).

Implementation

OrthoVenn was implemented with a Java code as the back end, configured on a Ubuntu Linux with an Apache server (http://www.apache.org/). PHP 5.5, JavaScript and HTML were used to develop the front-end user interface. The user input is sent to a PHP script through an interactive form and calls a core program written by Java. After the program is finished running, the PHP script retrieves and displays the results with hyperlinks to additional information. The Venn diagram is presented through jvenn, which is a new JavaScript library that allows the developer to embed Venn diagram viewers within HTML documents (32). Since we expect that the protein sequences in the Ensembl database will be updated regularly, we will update the OrthoVenn dataset every half year with our semi-automatic sequence identification pipeline.

SAMPLE AND RESULT ANALYSIS

We applied OrthoVenn clustering to identify gene clusters enriched in six grass genomes including *Aegilops tauschii* (Tausch's goatgrass), *Brachypodium distachyon* (false brome), *Oryza sativa* (rice), *Sorghum bicolor* (sorghum), *Hordeum vulgare* (Barley) and *Zea mays* (maize). Pairwise sequence similarities between all input protein sequences were calculated with an e-value cut-off of 1e−5. An inflation value (*−I*) of 1.5 was used to define orthologous cluster structure. The result is present at http://probes.pw.usda.gov/OrthoVenn/result.php?ID=g5.

Four main components are shown in the results page: a Venn diagram showing the shared orthologous gene clusters among six grass genomes (Figure 1A), a control panel for setting the colors and Venn diagram display mode (Figure 1B), a table of the number of clusters in each genome (Figure 1C), and the key words input and BLAST link for performing user-defined searches (Figure 1D). The analysis showed that 24 342 orthologous clusters were formed based on the protein sequences from the 6 species. These observations are consistent with that reported in a recent study which found 23 202 orthologous groups in five grass genomes (33). The numbers in the Venn diagram represent the number of orthologous clusters that *A. tauschii* shares with the five other species. The diagram shows that 9887 gene clusters are shared by all six species, suggesting their conservation in the lineage after speciation. Additionally, it shows that there were 951 clusters specific to *A. tauschii*. These clusters are likely gene clusters within multiple genes or in-paralog clusters. The presence of the in-paralog clusters suggests that there might be a lineage specific gene expansion in these gene families in *A. tauschii*. Based on annotation of these clusters, some of these lineage specific clusters could be potentially involved in important biological processes (e.g. nitrogen compound metabolic process or cellular aromatic compound metabolic process) (Figure 2).

The identification of single copy orthologs in any group of species is important for phylogenetic studies (34). Based on the orthologous clustering result, we have identified a set of 5929 clusters representing single copy genes shared in all six genomes, suggesting they maintained single copy status through evolutionary time after the divergence of these species. Each orthologous cluster in the OrthoVenn result provides an in-depth view of the genes and their sequences (Figure 3). Users can search specific clusters using key words, however, a BLAST search function is also provided. In the case of the key word search, input of 'Agamous-like MADS-box' will result in an orthologous cluster search result. Selection of cluster602 from the search result indicates that this orthologous group contains 12 protein sequences from five species (Figure 3A). Multiple sequence alignment and motif analysis indicate that the cluster represents a typical MADS-box protein (Figure 3B and C). Phylogenetic tree data shows that in cluster602, *A. tauschii* contains more paralogous genes, suggesting there is lineage-specific expansion of this specific type of MADS-box protein (Figure 3D). Furthermore, OrthoVenn includes the Cytoscape Web application (35) for visualizing and manipulating the graphs of the cluster and the relationship between clusters. In cluster602, proteins from the same species had higher similarity to one another than proteins between different species (Figure 3E), suggesting that duplication events after speciation resulted in closely related paralogous genes. In addition, cluster602 is closely related to cluster20273, cluster1325 and cluster23528, as revealed by the width of the lines connecting the cluster nodes (Figure 3F). Related clusters are defined by sequence similarity of proteins in these clusters as determined by the orthAgogue analysis result. The edge weight means the amount of similar sequences by counting similar sequence pairs between the clusters. Annotation data indicates that all of these clusters are Agamous-like MADS-box proteins.

VENN DIAGRAM GENERATION FROM CLUSTER FILE

In order to display orthologous cluster files directly in a Venn diagram, we developed a tool named ClusterVenn in the web site of OrthoVenn for converting clusters file to Venn diagrams. This allows users to submit their own orthologous cluster files generated by other software such as OrthoMCL. ClusterVenn calculates the number of species in the file, and provides a convenient way for users to choose which species should be compared. ClusterVenn then identifies the shared and unshared orthologous clusters based on identifiers in the orthologous cluster file and displays the Venn diagram as the result.

DISCUSSION

With the rapid increase of genome sequence data from high-throughput sequencing technologies, effective bioinformatics tools for genome wide comparisons among species has become an important component of gene and genome analysis. The identification of orthologs among species is a fundamental step in many comparative genomics studies. Visualization of disjointed or intersecting orthologous clusters provides insights into genome evolution across multiple

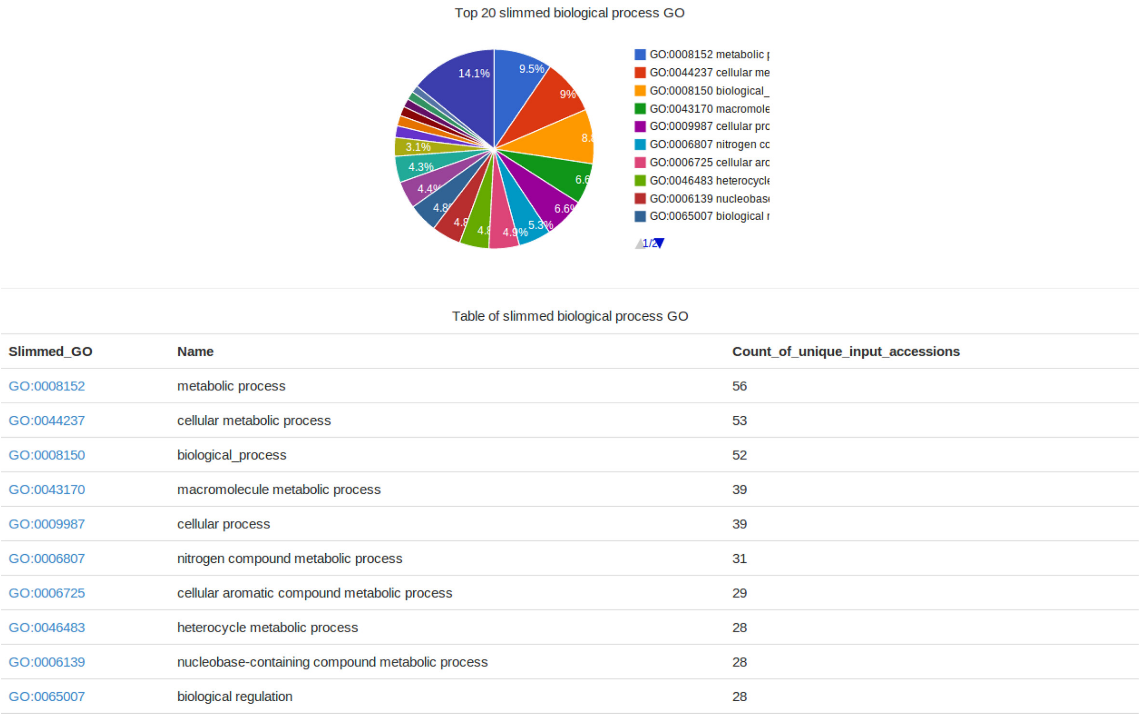


Table of slimmed biological process GO

Slimmed_GO	Name	Count_of_unique_input_accessions
GO:0008152	metabolic process	56
GO:0044237	cellular metabolic process	53
GO:0008150	biological_process	52
GO:0043170	macromolecule metabolic process	39
GO:0009987	cellular process	39
GO:0006807	nitrogen compound metabolic process	31
GO:0006725	cellular aromatic compound metabolic process	29
GO:0046483	heterocycle metabolic process	28
GO:0006139	nucleobase-containing compound metabolic process	28
GO:0065007	biological regulation	28

Figure 2. Distributions of *Aegilops tauschii* specific gene sets in biological process GO slim terms.

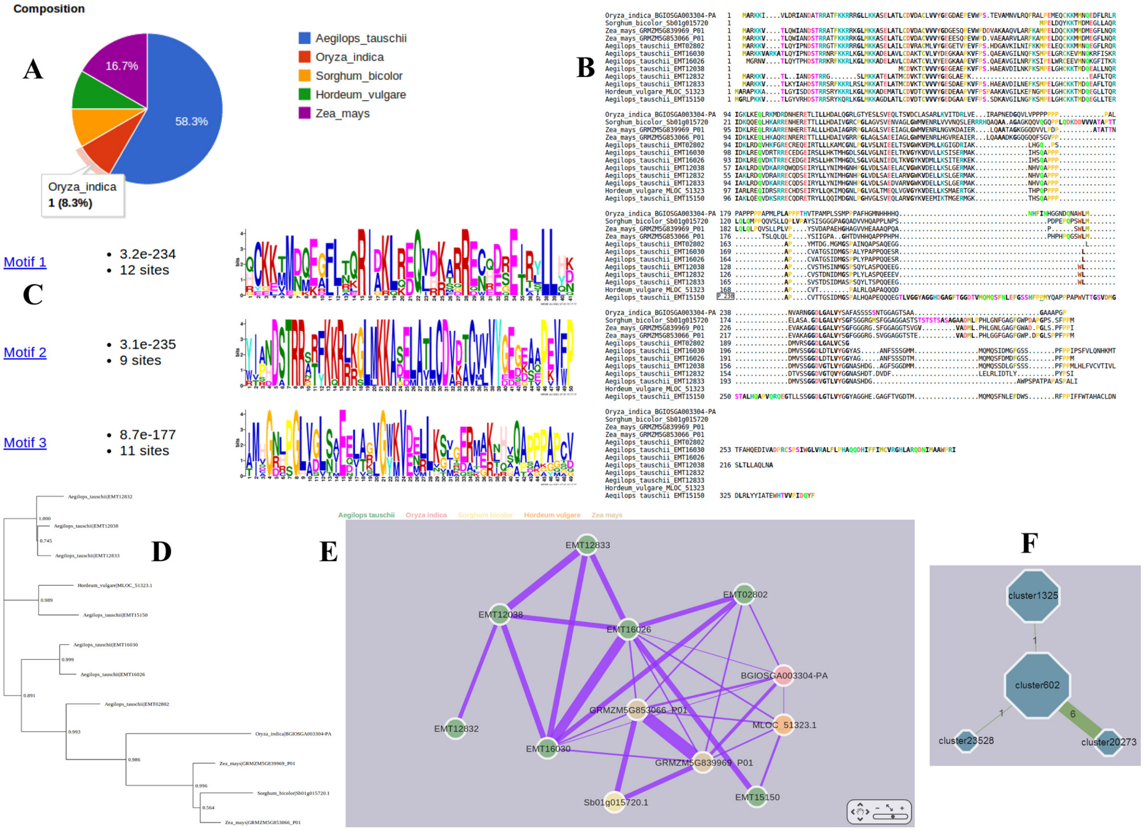


Figure 3. The annotation of cluster602 with different methods. (**A**) Composition of the cluster. (**B**) Multiple sequence alignment viewer. (**C**) Motifs in the proteins of the cluster. (**D**) Phylogenetic tree showing the inferred evolutionary relationships among the sequences in cluster602. (**E**) Network layout of the cluster. Nodes represent proteins and the edge width indicates the similarity between protein nodes. (**F**) The relationships between cluster602 and other clusters. Each node is a cluster and node size represents the number of proteins in the cluster. The edge weight means the amount of similar sequences by counting similar sequence pairs between the clusters.

species. OrthoVenn is a unique web tool for the visualization of genome wide comparisons of orthologous clusters with an interactive Venn diagram view and provides a high-level summary of functions for overlapping and non-overlapping orthologous gene sets. The integration of several sequence analysis methods in OrthoVenn provides further information related to the sequence conservation of orthologous genes of interest to the user. Furthermore, OrthoVenn is an easy-to-use web server that only requires users to choose species from the available list of 272 organisms or to submit a list of protein sequences for their species of interest, and allows the ortholog analysis to be performed without the need to install programs locally. We think OrthoVenn provides a powerful platform for identifying, analyzing and assigning the biological meaning to orthologous genes and facilitates a better understanding of gene and genome evolution across diverse taxa.

AVAILABILITY

http://probes.pw.usda.gov/OrthoVenn or http://aegilops.wheat.ucdavis.edu/OrthoVenn.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Xiaohua He for the critical reading of the manuscript. We are very grateful to many anonymous reviewers for testing the server and offering valuable comments.

FUNDING

U.S. National Science Foundation [IOS 0822100]; United State Department of Agriculture, Agriculture Research Service CRIS project [5325-21000-021]. Funding for open access charge: U.S. National Science Foundation [IOS 0822100].
Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.
3. Mushegian,A.R., Garey,J.R., Martin,J. and Liu,L.X. (1998) Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.*, **8**, 590–598.
4. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldon,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
5. Sonnhammer,E.L. and Ostlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.
6. Kriventseva,E.V., Tegenfeldt,F., Petty,T.J., Waterhouse,R.M., Simao,F.A., Pozdnyakov,I.A., Ioannidis,P. and Zdobnov,E.M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, **43**, D250–D256.
7. Pryszcz,L.P., Huerta-Cepas,J. and Gabaldon,T. (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, **39**, e32.
8. Huerta-Cepas,J., Capella-Gutierrez,S., Pryszcz,L.P., Marcet-Houben,M. and Gabaldon,T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.
9. Altenhoff,A.M. and Dessimoz,C. (2012) Inferring orthology and paralogy. *Methods Mol. Biol.*, **855**, 259–279.
10. Argout,X., Salse,J., Aury,J.M., Guiltinan,M.J., Droc,G., Gouzy,J., Allegre,M., Chaparro,C., Legavre,T., Maximova,S.N. *et al.* (2011) The genome of Theobroma cacao. *Nat. Genet.*, **43**, 101–108.
11. D'Hont,A., Denoeud,F., Aury,J.M., Baurens,F.C., Carreel,F., Garsmeur,O., Noel,B., Bocs,S., Droc,G., Rouard,M. *et al.* (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217.
12. Cho,Y.S., Hu,L., Hou,H., Lee,H., Xu,J., Kwon,S., Oh,S., Kim,H.M., Jho,S., Kim,S. *et al.* (2013) The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat. Commun.*, **4**, 2433.
13. Hulsen,T., de Vlieg,J. and Alkema,W. (2008) BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, **9**, 488.
14. Pirooznia,M., Nagarajan,V. and Deng,Y. (2007) GeneVenn – a web application for comparing gene lists using Venn diagrams. *Bioinformation*, **1**, 420–422.
15. Kestler,H.A., Muller,A., Kraus,J.M., Buchholz,M., Gress,T.M., Liu,H., Kane,D.W., Zeeberg,B.R. and Weinstein,J.N. (2008) VennMaster: area-proportional Euler diagrams for functional GO analysis of microarrays. *BMC Bioinformatics*, **9**, 67.
16. Wang,Y., Thilmony,R. and Gu,Y.Q. (2014) NetVenn: an integrated network analysis web platform for gene lists. *Nucleic Acids Res.*, **42**, W161–W166.
17. Blom,J., Albaum,S.P., Doppmeier,D., Puhler,A., Vorholter,F.J., Zakrzewski,M. and Goesmann,A. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, **10**, 154.
18. Curtis,D.S., Phillips,A.R., Callister,S.J., Conlan,S. and McCue,L.A. (2013) SPOCS: software for predicting and visualizing orthology/paralogy relationships among genomes. *Bioinformatics*, **29**, 2641–2642.
19. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
20. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
21. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
22. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
23. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
24. Moreno-Hagelsieb,G. and Hudy-Yuffa,B. (2014) Estimating overannotation across prokaryotic genomes using BLAST+, UBLAST, LAST and BLAT. *BMC Res. Notes*, **7**, 651.
25. Ward,N. and Moreno-Hagelsieb,G. (2014) Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PLoS One*, **9**, e101850.
26. Ekseth,O.K., Kuiper,M. and Mironov,V. (2014) orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics*, **30**, 734–736.
27. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
28. McCarthy,F.M., Wang,N., Magee,G.B., Nanduri,B., Lawrence,M.L., Camon,E.B., Barrell,D.G., Hill,D.P., Dolan,M.E., Williams,W.P. *et al.* (2006) AgBase: a functional genomics resource for agriculture. *BMC Genomics*, **7**, 229.
29. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

30. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
31. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
32. Bardou,P., Mariette,J., Escudie,F., Djemiel,C. and Klopp,C. (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 293.
33. Jia,J., Zhao,S., Kong,X., Li,Y., Zhao,G., He,W., Appels,R., Pfeifer,M., Tao,Y., Zhang,X. *et al.* (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91–95.
34. Creevey,C.J., Muller,J., Doerks,T., Thompson,J.D., Arendt,D. and Bork,P. (2011) Identifying single copy orthologs in Metazoa. *PLoS Comput. Biol.*, **7**, e1002269.
35. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.