doi: 10.1093/nar/gkz333

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019

OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species

Ling Xu[1,†], Zhaobin Dong[2,3,†], Lu Fang[1], Yongjiang Luo[1], Zhaoyuan Wei[1], Hailong Guo[1], Guoqing Zhang[1], Yong Q. Gu[4], Devin Coleman-Derr[2,3], Qingyou Xia[1] and Yi Wang[1,*]

[1]Biological Science Research Center, Southwest University, Chongqing 400715, China, [2]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94710, USA, [3]USDA-ARS, Plant Gene Expression Center, Albany, CA 94706, USA and [4]USDA-ARS, Western Regional Research Center, Crop Improvement and Genetics Research Unit, Albany, CA 94706, USA

Received February 14, 2019; Revised April 16, 2019; Editorial Decision April 23, 2019; Accepted April 25, 2019

ABSTRACT

OrthoVenn is a powerful web platform for the comparison and analysis of whole-genome orthologous clusters. Here we present an updated version, OrthoVenn2, which provides new features that facilitate the comparative analysis of orthologous clusters among up to 12 species. Additionally, this update offers improvements to data visualization and interpretation, including an occurrence pattern table for interrogating the overlap of each orthologous group for the queried species. Within the occurrence table, the functional annotations and summaries of the disjunctions and intersections of clusters between the chosen species can be displayed through an interactive Venn diagram. To facilitate a broader range of comparisons, a larger number of species, including vertebrates, metazoa, protists, fungi, plants and bacteria, have been added in OrthoVenn2. Finally, a stand-alone version is available to perform large dataset comparisons and to visualize results locally without limitation of species number. In summary, OrthoVenn2 is an efficient and user-friendly web server freely accessible at https://orthovenn2.bioinfotoolkits.net.

INTRODUCTION

Homologous genes can be divided into two main classes: orthologs and paralogs. Orthologous genes originate from a common ancestor during specification events (1), and are usually syntenic between close-related species. Paralogs also share a common ancestor, but arise from sequence duplication events within a species, and often show limited synteny and more speciation-related divergence. If orthologous genes in multiple species show high sequence similarity, the likelihood that they may continue to perform similar biological functions (2). On the other hand, orthologs with sequences that show greater divergence from other species are more likely to perform distinct functions (2). In recent decades, identifying orthologous genes and ascertaining the degree of similarity between them are two important steps in comparative genomics studies to understand the evolution of genes and genomes (3). New innovations in sequencing technologies have rapidly generated vast amounts of genome-wide sequence data across the tree of life (2,4,5), and there is a growing need for tools that enable orthologous gene identification and the ability to explore their function and evolution across phylogenetic space.

Two fundamental methods for identifying orthologous clusters have been developed. The first is tree-based, and several tools such as software PhyloTreePruner (6) and TreeFam (7) employ this strategy. The second is graph-based, and this methodology is found in tools including PanOCT (8), OrthoFinder (9), OrthoMCL (10), COCO-CL (11), OrthoDB (12), OMA (13) and Roary (14). The tree-based and graph-based algorithms differ in many respects and each of them have their own advantage, but current tree-based algorithms are generally computationally more expensive to construct for large numbers of genes and genomes, which makes this type of approach less applicable than the graph-based approach for orthologous analysis on large datasets with large evolutionary distances (3). For these reasons, both OrthoVenn and OrthoVenn2 employed a graph-based method which has been demonstrated utility in automating and handling large datasets (3).

[*]To whom correspondence should be addressed. Yi Wang. Tel: +86 2368251683; Fax: +86 2368251128; Email: yiwang28@swu.edu.cn
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present Address: Ling Xu, Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94710, USA.

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
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.

Most applications for orthologous gene identification and comparison were developed for use on Linux-based systems, including OrthAgogue (15), Ortholog-Finder (16), Orthograph (17), PorthoMCL (18) and ProteinOrtho (19). While this feature contributes to their power, speed and versatility, it also limits the pool of potential users to researchers who have specific familiarity and knowledge with the Linux OS. To address this issue, a growing number of web-based orthology detection and comparison tools have been developed recently, such as OrthoInspector (20), Family-Companion (21), Orthonome (5), PhosphOrtholog (4), Hieranoid (22) and MorFeus (23). However, the majority of these tools lack the ability to visualize orthologous clusters, and the few that have this capability, such as ORCAN (24), contain relatively rudimentary visualization capabilities to support the comparisons of large datasets from multiple genomes.

Previously, we published a high-speed web-server based tool, OrthoVenn, which was published in the 2015 *Nucleic Acids Research* web server issue (25). This tool has been used and cited in a growing body of research (26), and thousands of users from more than sixty countries have used it to analyze their datasets. The utility of OrthoVenn has been cited more than 150 times since its release.

Here, we present an update to the OrthoVenn tool based on requests from users and the broader community of comparative genomics researchers. In this update, we have implemented new features to allow users to perform whole genome comparisons for up to twelve species of bacteria, fungi, protists and metazoa (an increase from the six in the previous version). As a result of the relatively larger genome size of plants and vertebrates as compared with the other types of organisms listed above, users can choose up to a maximum of eight species for genomic comparisons when working within these two eukaryotic groups. To improve speed, the alignment between related species within clades has already been calculated and included in the web version of OrthoVenn2. The pre-calculation and alignment between distant species would substantially increase the storage amount in our server. Due to the storage requirements and computational efficiency, the current web version of OrthoVenn compares species within clades, and comparisons between distant species (such as the species from different kingdoms) are not supported. To help facilitate cross-kingdom comparisons, and more generally for processing and visualization of larger numbers of species, users can either upload their own user-generated clustering dataset from other softwares as an input to OrthoVenn, or download the stand-alone version for use on their own server. In addition, we introduced significant enhancements to the data visualization and interpretation capabilities of OrthoVenn2. OrthoVenn2 continues to be open to all users and freely available as a web service at https://orthovenn2.bioinfotoolkits.net. The input data for OrthoVenn2 is a protein sequence in fasta format and the output is an interactive occurrence pattern table and Venn diagram with additional layers of information and downloadable content. To demonstrate its utility, we randomly chose eight *Streptomyces* species for comparative analysis of orthologous genes with OrthoVenn2. The results are discussed below in the 'sample and result analysis' section.

DATASET

We downloaded protein sequence from the Ensembl database (release January 2019) and incorporated them into OrthoVenn2 following our previous pipeline (25). This dataset includes 142 vertebrates, 71 metazoa, 65 protists, 94 fungi, 57 plants and 111 bacteria species. The total number of protein sequences present in OrthoVenn2 is 8 858 566. In total, the protein database in OrthoVenn2 is four times the size of the previous version. The annotation of protein clusters was performed through DIAMOND analysis using the non-redundant protein database in UniProt (release January 2019) as described previously (25,27).

GENOME COMPARISON IN MULTIPLE SPECIES

The previous version of OrthoVenn only supports whole-genome comparisons for up to six species due to limitations in computational capacity and visualization methods. In OrthoVenn, we used the most popular heuristic best-match method available at the time (3) from OrthoMCL (10) to identify orthologous genes based on conservation (25). In this update, we provided new features that allows users to compare orthologous genes for more than six species (a maximum of twelve for bacteria, fungi and protists, whereas up to eight for plant and vertebrate, due to their large genome size could hinder computational efficiency of our server). First, we upgraded our web-server to have an increased computational capacity to support additional comparisons. Briefly, the prior server for Orthovenn1 had 16 core processors and 96G memory, while the new Orthovenn2 server harbors 64 core processors and 512G memory. Second, OrthoVenn2 uses DIAMOND (v0.9.24) instead of BLASTP or UBLAST to perform the all-against-all protein sequence comparison as DIAMOND has been shown to be 20 000 times faster than BLASTX and 1000 times faster than UBLAST without any significant compromise in output reliability or value (28). To test the speed improvement after update, we chose six plant and six bacteria species within our database and performed two separate orthologous clustering with both Orthovenn1 and Orthovenn2 to compare their respective speeds. Additionally, we uploaded six custom species of plant and bacteria to perform clustering with Orthovenn1 and Orthovenn2, separately. The results indicate that OrthoVenn2 is at least ten times faster than OrthoVenn1 (Figure 1). As a complement, a stand-alone version has been also developed that can process multiple species without limitations on capacity. Finally, the accuracy of OrthoVenn2 was compared with other orthology inference methods using the benchmark service from the QfO community (29). OrthoVenn2 showed accuracy similar to InParanoidCore, OMA Groups 2.0, eggNOG and SonicParanoid (fast) (Supplementary Figure S1).

TOOLS FOR VISUALIZING RESULT SETS

To display multiple comparison results more effectively, OrthoVenn2 uses an occurrence cluster table to display the orthologous cluster groups for multiple species (Figure 2A) with associated informative Venn diagrams (Figure 2B).

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019

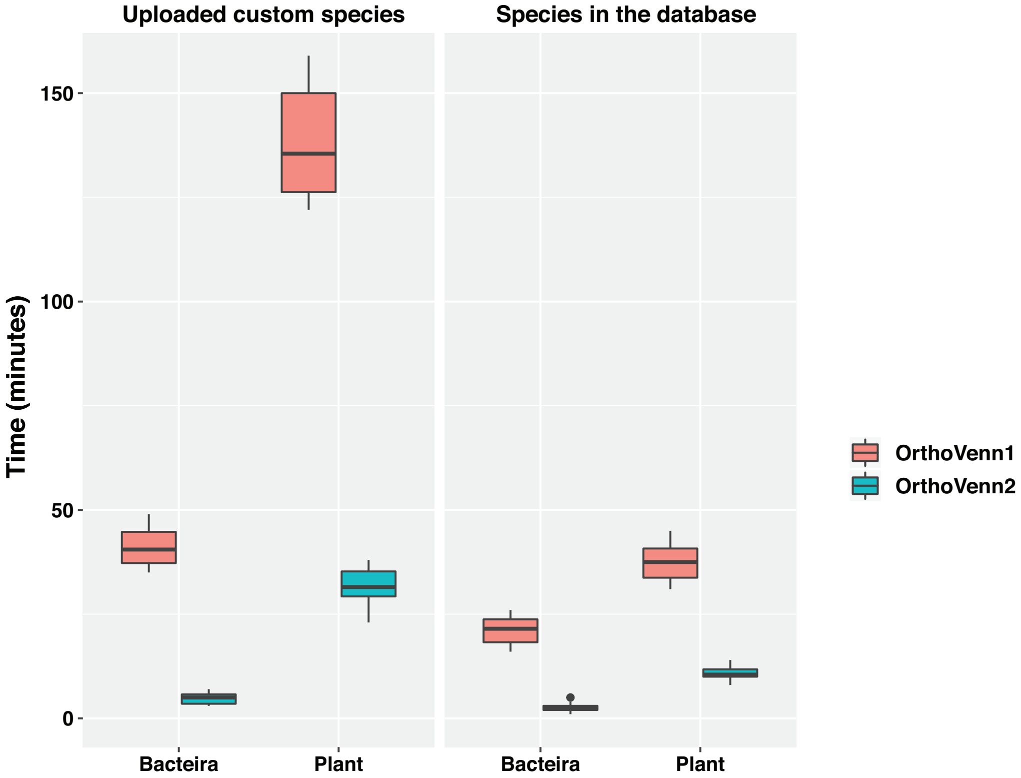


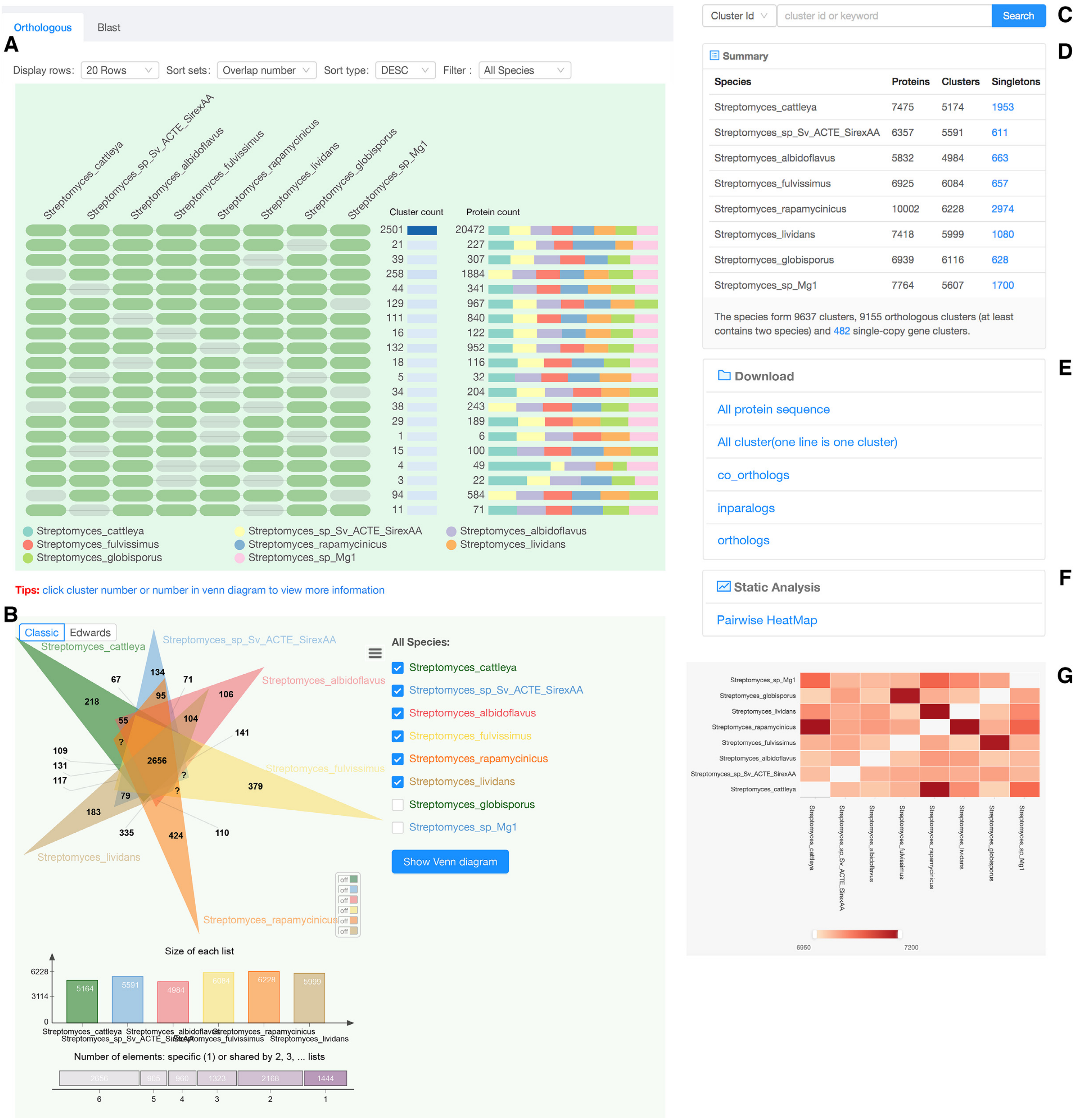
Figure 1. Speed comparison of OrthoVenn1 and OrthoVenn2. Boxplots of median time lapse per job for OrthoVenn1 (red) or OrthoVenn2 (Green). The boxplots indicate the median (central line), the first and third quartiles (upper and lower box bounds) and the minimum and maximum value (lower and upper whiskers).

This clustering result is summarized as a cell graph in which each row represents an ortholog cluster group and each column indicates a species. A green cell indicates the presence of a cluster group in the corresponding species, and a gray bar represents the absence of a cluster group in that species (Figure 2A). There are other visualization options that users can select before downloading the figure, including changes to cell color, height, width, and font size by clicking the icon in the upper right corner. The table can be redrawn with several user-defined sorting schemes, including sorting by protein number, overlap count, or cluster count in descending ('DESC') or ascending ('ASC') order (Figure 2A). Overlaying the cursor on each cell will display the species identity associated with the current cell (Figure 2A). A stacked bar plot at right displays the cumulative number of protein sequences present in the cluster group for each species (Figure 2A).

OrthoVenn2 also provides a search function to retrieve specific clusters from the archived cluster results with a keyword or cluster ID (Figure 2C). The total number of shared clusters for each pair of species are summarized in a table (Figure 2D). Links are also provided for users to download the computed datasets (Figure 2E), and a 'Pairwise HeatMap' button will render a heatmap to visualize the overlapping cluster numbers for the working species in a pairwise fashion (Figure 2F). The overlapping cluster numbers are indicated through a color gradient with user-defined minimum and maximum thresholds. Overlapping clusters less than the minimum threshold are ignored and indicated by blank cells within the figure (Figure 2G). Additionally, we offer a BLAST tool on a second tab to allow users to input their own FASTA-format sequences to compare against the output clusters (Figure 2A).

Finally, to render a Venn diagram in 'Classic' or 'Edwards' format for up to six species from the occurrence table, the user may select the checkboxes for individual species names located below the occurrence table. By default, a Venn diagram is automatically drawn for two to six of the species present in the analysis (Figure 2B). The color of each species in the Venn diagram can be adjusted by selecting the species name (Figure 2B). Additionally, a bar plot below the Venn diagram shows the number of clusters found in each species (Figure 2B, bottom). One can download these figures by selecting the icon above and to the right of the Venn diagram. Selecting the number found within each cell of the Venn diagram or the cluster number in the occurrence table will produce a new page showing the functional information associated with the chosen cluster group (Figure 3). It should be noted that the functional annotation can only be performed for up to eight species. Three pie plots are used to visualize the proportion of GO terms for the three main functional categories: biological processes, molecular functions and cellular components within the chosen orthologous group. The corresponding numeric information can be found on the 'Cluster list' page. The network for each selected cluster ID can also be generated (Figure 4A). To download the fasta file for the current cluster, the user can select the 'fasta' link. An alignment figure can be gener-

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019



Display rows: 20 Rows Sort sets: Overlap number Sort type: DESC Filter : All Species

Cluster count	Protein count
2501	20472
21	227
39	307
258	1884
44	341
129	967
111	840
16	122
132	952
18	116
5	32
34	204
38	243
29	189
1	6
15	100
4	49
3	22
94	584
11	71

Tips: click cluster number or number in venn diagram to view more information

All Species:

- ☑ Streptomyces_cattleya
- ☑ Streptomyces_sp_Sv_ACTE_SirexAA
- ☑ Streptomyces_albidoflavus
- ☑ Streptomyces_fulvissimus
- ☑ Streptomyces_rapamycinicus
- ☑ Streptomyces_lividans
- □ Streptomyces_globisporus
- □ Streptomyces_sp_Mg1

Show Venn diagram

Cluster Id | cluster id or keyword | Search

Summary

Species	Proteins	Clusters	Singletons
Streptomyces_cattleya	7475	5174	1953
Streptomyces_sp_Sv_ACTE_SirexAA	6357	5591	611
Streptomyces_albidoflavus	5832	4984	663
Streptomyces_fulvissimus	6925	6084	657
Streptomyces_rapamycinicus	10002	6228	2974
Streptomyces_lividans	7418	5999	1080
Streptomyces_globisporus	6939	6116	628
Streptomyces_sp_Mg1	7764	5607	1700

The species form 9637 clusters, 9155 orthologous clusters (at least contains two species) and 482 single-copy gene clusters.

Download

All protein sequence

All cluster(one line is one cluster)

co_orthologs

inparalogs

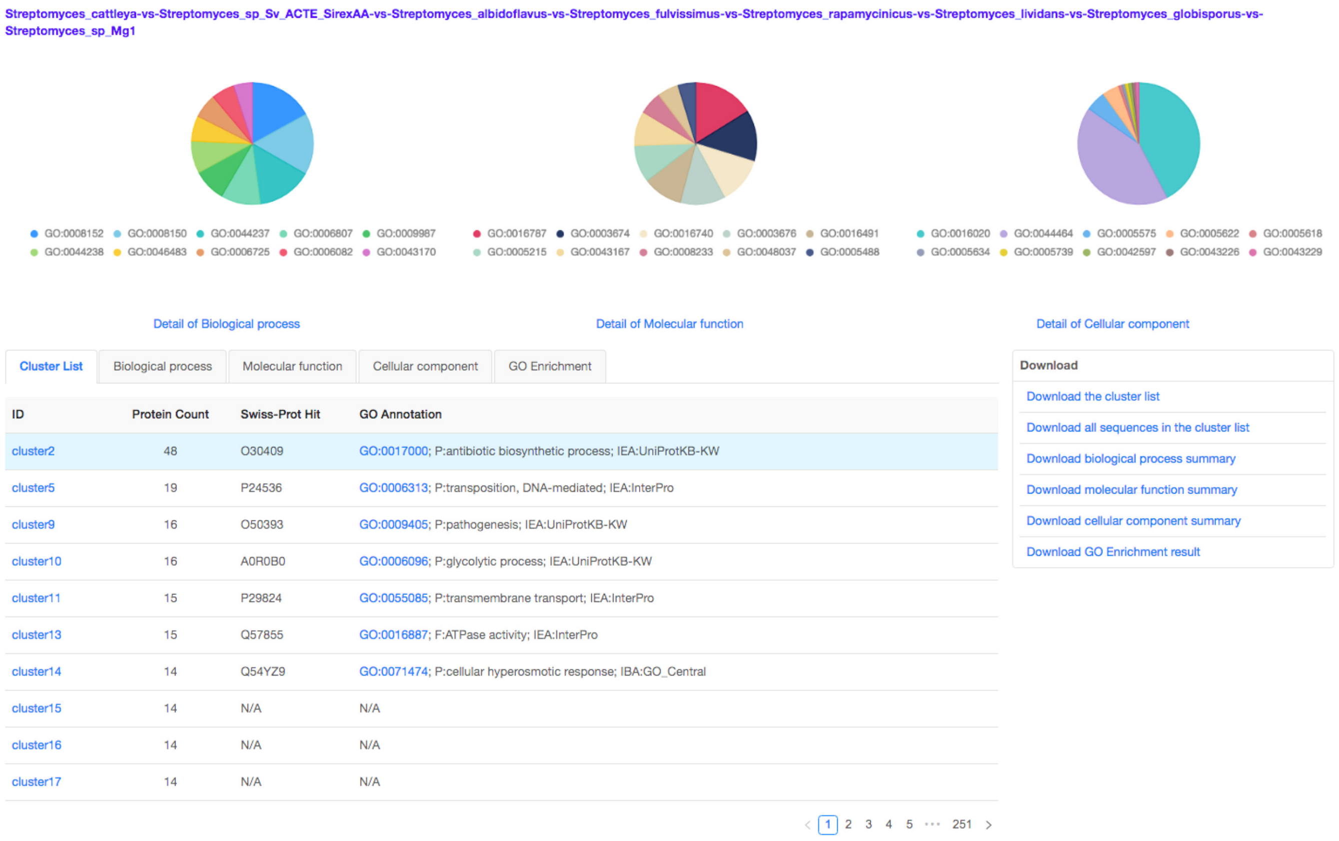
orthologs

Static Analysis

Pairwise HeatMap

Figure 2. A result page of Orthovenn2. (**A**) The occurrence table shows the occurrence pattern of shared orthologous groups among *Streptomyces cattleya*, *Streptomyces sp. Sv ACTE SirexAA*, *Streptomyces albidoflavus*, *Streptomyces fulvissimus*, *Streptomyces rapamycinicus*, *Streptomyces lividans*, *Streptomyces globisporus* and *Streptomyces sp. Mg1*. The pattern to the left indicates which species are in the clusters, cluster count is the number of clusters shared between species, and protein count is the number of protein members in the shared clusters. (**B**) Venn diagram displays the distribution of shared orthologous clusters among the first six species. (**C**) Keyword and cluster ID search for specific clusters in the results. (**D**) Counts of clusters in each genome. The singleton at the top right describes the genes for which no orthologs could be found in other species; single copy gene clusters at the bottom indicate the clusters that contain single copy gene in each species. (**E**) Download links for computed datasets. (**F**) A link to the pairwise heatmap of overlapping cluster numbers between pair-wise genomes. (**G**) The pairwise heatmap of overlapping cluster numbers appears in the pop-up. The heatmap of overlapping cluster numbers between each pair of genomes. Each cell indicates the overlap cluster numbers between each pair of species. The overlapping cluster numbers refers to the cluster numbers that were shared between species. The overlapping cluster numbers would appear when users hovering cursor over each heatmap cell.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019

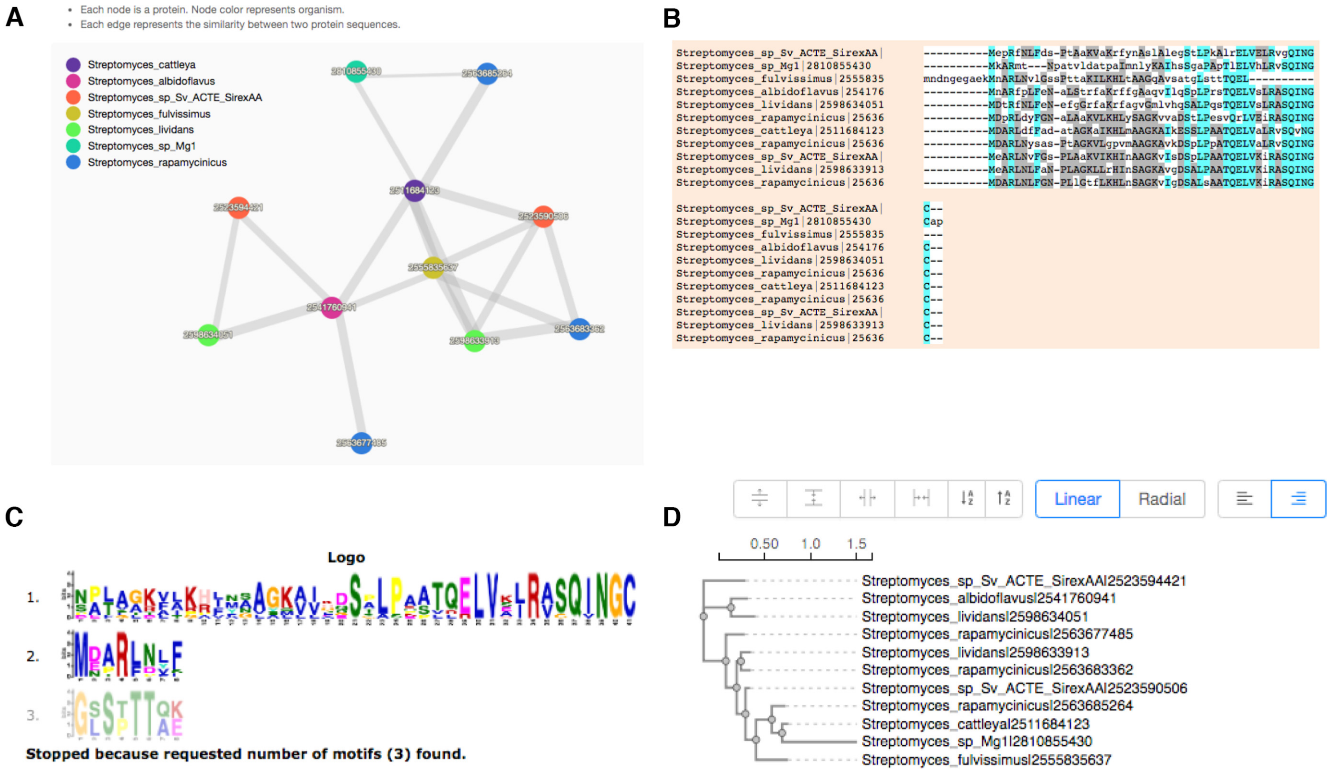

Streptomyces_cattleya-vs-Streptomyces_sp_Sv_ACTE_SirexAA-vs-Streptomyces_albidoflavus-vs-Streptomyces_fulvissimus-vs-Streptomyces_rapamycinicus-vs-Streptomyces_lividans-vs-Streptomyces_globisporus-vs-Streptomyces_sp_Mg1
● GO:0008152 ● GO:0008150 ● GO:0044237 ● GO:0006807 ● GO:0009987
● GO:0044238 ● GO:0046483 ● GO:0006725 ● GO:0006082 ● GO:0043170
● GO:0016787 ● GO:0003674 ● GO:0016740 ● GO:0003676 ● GO:0016491
● GO:0005215 ● GO:0043167 ● GO:0008233 ● GO:0048037 ● GO:0005488
● GO:0016020 ● GO:0044464 ● GO:0005575 ● GO:0005622 ● GO:0005618
● GO:0005634 ● GO:0005739 ● GO:0042597 ● GO:0043226 ● GO:0043229
Detail of Biological process
Detail of Molecular function
Detail of Cellular component
Cluster List
Biological process
Molecular function
Cellular component
GO Enrichment
1 2 3 4 5 ••• 251

ID	Protein Count	Swiss-Prot Hit	GO Annotation
cluster2	48	O30409	GO:0017000; P:antibiotic biosynthetic process; IEA:UniProtKB-KW
cluster5	19	P24536	GO:0006313; P:transposition, DNA-mediated; IEA:InterPro
cluster9	16	O50393	GO:0009405; P:pathogenesis; IEA:UniProtKB-KW
cluster10	16	A0R0B0	GO:0006096; P:glycolytic process; IEA:UniProtKB-KW
cluster11	15	P29824	GO:0055085; P:transmembrane transport; IEA:InterPro
cluster13	15	Q57855	GO:0016887; F:ATPase activity; IEA:InterPro
cluster14	14	Q54YZ9	GO:0071474; P:cellular hyperosmotic response; IBA:GO_Central
cluster15	14	N/A	N/A
cluster16	14	N/A	N/A
cluster17	14	N/A	N/A

Download

- Download the cluster list
- Download all sequences in the cluster list
- Download biological process summary
- Download molecular function summary
- Download cellular component summary
- Download GO Enrichment result

Figure 3. Distribution of GO terms for core orthologous gene clusters of eight *Streptomyces* species.


A
B
C
D
Streptomyces_cattleya
Streptomyces_albidoflavus
Streptomyces_sp_Sv_ACTE_SirexAA
Streptomyces_fulvissimus
Streptomyces_lividans
Streptomyces_sp_Mg1
Streptomyces_rapamycinicus
2810855430
2563685264
2511684123
2523594421
2523590506
2555835637
2541760941
2598634051
2598633913
2563683362
2563677485
Logo
1.
2.
3.
Linear
Radial
0.50
1.0
1.5
Streptomyces_sp_Sv_ACTE_SirexAAl2523594421
Streptomyces_albidoflavusl2541760941
Streptomyces_lividansl2598634051
Streptomyces_rapamycinicusl2563677485
Streptomyces_lividansl2598633913
Streptomyces_rapamycinicusl2563683362
Streptomyces_sp_Sv_ACTE_SirexAAl2523590506
Streptomyces_rapamycinicusl2563685264
Streptomyces_cattleyal2511684123
Streptomyces_sp_Mg1l2810855430
Streptomyces_fulvissimusl2555835637

- Each node is a protein. Node color represents organism.
- Each edge represents the similarity between two protein sequences.

```
Streptomyces_sp_Sv_ACTE_SirexAA|          ----------MepRfNLFds-PtAaKVaKrfynAslAlegStLPkAlrELVELRvgQING
Streptomyces_sp_Mg1|2810855430            ----------MkARmt---NpatvldatpaImnlyKAIhsSgaPApTlELVhLRvSQING
Streptomyces_fulvissimus|2555835          mndngegaekMnARLNvlGssPttaKILKHLtAAGqAvsatgLsttTQEL----------
Streptomyces_albidoflavus|254176          ----------MnARfpLFeN-aLStrfaKrffgAaqvIlqSpLPrsTQELVsLRASQING
Streptomyces_lividans|2598634051          ----------MDtRfNLFeN-efgGrfaKrfagvGmlvhqSALPqsTQELVsLRASQING
Streptomyces_rapamycinicus|25636          ----------MDpRLdyFGN-aLAaKVLKHLySAGKvvaDStLPesvQrLVEiRASQING
Streptomyces_cattleya|2511684123          ----------MDARLdfFad-atAGKaIKHLmAAGKAIkESSLPAATQELVaLRvSQvNG
Streptomyces_rapamycinicus|25636          ----------MDARLNysas-PtAGKVLgpvmAAGKAvkDSpLPpATQELVaLRvSQING
Streptomyces_sp_Sv_ACTE_SirexAA|          ----------MeARLNvFGs-PLAaKVIKHInAAGKvIsDSpLPAATQELVKiRASQING
Streptomyces_lividans|2598633913          ----------MeARLNLFaN-PLAGKLLrHInSAGKAvgDSALPAATQELVKiRASQING
Streptomyces_rapamycinicus|25636          ----------MDARLNLFGN-PLlGtfLKHLnSAGKvIgDSALsAATQELVKiRASQING

Streptomyces_sp_Sv_ACTE_SirexAA|          C--
Streptomyces_sp_Mg1|2810855430            Cap
Streptomyces_fulvissimus|2555835          ---
Streptomyces_albidoflavus|254176          C--
Streptomyces_lividans|2598634051          C--
Streptomyces_rapamycinicus|25636          C--
Streptomyces_cattleya|2511684123          C--
Streptomyces_rapamycinicus|25636          C--
Streptomyces_sp_Sv_ACTE_SirexAA|          C--
Streptomyces_lividans|2598633913          C--
Streptomyces_rapamycinicus|25636          C--
```

Stopped because requested number of motifs (3) found.

Figure 4. The annotation of cluster65 using multiple methods. (**A**) The network of proteins within cluster65. (**B**) Multiple sequence alignment for proteins in cluster65. (**C**) Motifs in the protein sequences in cluster65. (**D**) Phylogenetic tree for the proteins within cluster65.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019

ated by clicking on the 'Multiple Sequence Alignment' button (Figure 4B). Finally, OrthoVenn2 also incorporates the software 'Multiple Em for Motif Elicitation' to perform a motif analysis (Figure 4C) and a phylogenetic tree for users (Figure 4D).

SAMPLE AND RESULTS ANALYSIS

To demonstrate the utility of OrthoVenn2, we applied OrthoVenn2 clustering to eight randomly picked *Streptomyces* species. Pairwise genome comparisons were performed with the following parameters: e-value 1e-5 and an inflation value 1.5. The results can be found at https://orthovenn2.bioinfotoolkits.net/task/result/8576ec1a59ad3d78b2680ab78b181191. Briefly, the analysis identified 9,637 orthologous clusters, which includes 2501 core genome orthologs (Figure 2B). The number of core genome orthologs is less than the number (3,096) reported in a study in which five *Streptomyces* genomes were compared (30). To address the discrepancy, we hypothesized that different *Streptomyces* species sets harbor different core genome orthologs. To test our hypothesis, we downloaded these five genomes of *Streptomyces* species from NCBI http://www.ncbi.nlm.nih.gov/genome/browse/, predicted genes with prodigal v2.6.3 (31) and performed the genome comparison using OrthoVenn2 with threshold e-value 1e-5 and inflation 1.5. We identified 8694 orthologous clusters in total with 3188 core orthologous gene set https://orthovenn2.bioinfotoolkits.net/task/result/ae808cda97a1384ca03d200fdc4571c0. These results are consistent with the previous study. The number of predicted orthologous groups ranges from 8,341–9,209 using *E*-values of 10, 1, 0.1, 0.01, 10–3, 10–4, 10–5, 10–6, 10–7, 10–8, 10–9 and 10–10 and inflation indexes of 1.0, 1.5 and 2.0 (30). Our observations demonstrated that OrthoVenn2 is a reliable and user-friendly tool to perform genome orthologous comparison.

CLUSTERVENN IMPROVEMENT

As there are many methods available for ortholog clustering, for example, OrthoMCL can generate cluster file. In order to provide users a function of viewing the shared clusters between species, we developed a tool named ClusterVenn to visualize this cluster file in OrthoVenn1. This early version was limited to analysis of orthologous clusters for a maximum of six species due to known visualization constraints of Venn diagrams with more than six categories (32). In this update, we used the occurrence table to display the occurrence of cluster groups between species, allowing users to upload and compare clusters without limitations of species number. Simultaneously, users can choose up to six species in the occurrence table to display the intersection and disjunction relationship between species with a Venn diagram. Cluster files from most available orthologous identification tools are compatible and visualizable with OrthoVenn2.

STAND-ALONE VERSION

In response to users' requests, we generated a stand-alone version of OrthoVenn2 for large datasets comparisons. Our tool is merged with Docker technology to build reproducible and convenient types of workflows. Docker is an open source project and platform for building, shipping and running any app, enabling the widespread distribution of applications (https://docs.docker.com). The release of OrthoVenn2 as a Docker provides an isolated and self-contained package without the need to install dependencies and change environmental settings. This feature increases its reusability and reproducibility while simplifying its ease of use. The installation and usage instructions are available at https://orthovenn2.bioinfotoolkits.net/download. Users can input their own fasta file without limitation of species number. In almost all other respects, the usage is the same as that for the web server, including data analysis and visualization.

FUTURE PLAN

OrthoVenn2 is an open-source web server that identifies and compares genome orthologs from different species. We upgraded our server capacity to process larger datasets and offered improvements to data visualization and interpretation. However, OrthoVenn2 might have some backlog in term of data analysis and interpretation. Currently, OrthoVenn2 only takes protein sequence data as input. We are working to improve our computational capacity to allow users to upload both protein and the genome sequences. Input of genome sequences could allow our tool to predict genes and perform protein translation. The protein sequence will be clustered and annotated in our current version. We also aim to continue to improve the visualization and annotation of orthologous groups.

Some of the orthologous regions show collinearity characteristic (33,34). Comparative analysis of the collinearity for these segments is important to understand the genome rearrangements and evolution. We wish to add the collinearity comparison for our orthologous in our future version.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

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

FUNDING

National Natural Science Foundation of China [31871330]. Funding for open access charge: National Natural Science Foundation of China [31871330].

Conflict of interest statement. None declared.

REFERENCES

1. Jensen,R.A. (2001) Orthologs and paralogs - we need to get it right. *Genome Biol.*, **2**, interactions1002.1–interactions1002.3.
2. Fang,G., Bhardwaj,N., Robilotto,R. and Gerstein,M.B. (2010) Getting started in gene orthology and functional analysis. *PLoS Comput. Biol.*, **6**, e1000703.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019

3. Kristensen,D.M., Wolf,Y.I., Mushegian,A.R. and Koonin,E.V. (2011) Computational methods for Gene Orthology inference. *Brief. Bioinform.*, **12**, 379–391.
4. Chaudhuri,R., Sadrieh,A., Hoffman,N.J., Parker,B.L., Humphrey,S.J., Stöckli,J., Hill,A.P., James,D.E. and Yang,J.Y.H. (2015) PhosphOrtholog: a web-based tool for cross-species mapping of orthologous protein post-translational modifications. *BMC Genomics*, **16**, 617.
5. Rane,R.V., Oakeshott,J.G., Nguyen,T., Hoffmann,A.A. and Lee,S.F. (2017) Orthonome - a new pipeline for predicting high quality orthologue gene sets applicable to complete and draft genomes. *BMC Genomics*, **18**, 673.
6. Kocot,K.M., Citarella,M.R., Moroz,L.L. and Halanych,K.M. (2013) PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinform. Online*, **9**, 429–435.
7. Schreiber,F., Patricio,M., Muffato,M., Pignatelli,M. and Bateman,A. (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.*, **42**, D922–D925.
8. Fouts,D.E., Brinkac,L., Beck,E., Inman,J. and Sutton,G. (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res.*, **40**, e172.
9. Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**, 157.
10. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
11. Jothi,R., Zotenko,E., Tasneem,A. and Przytycka,T.M. (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, **22**, 779–788.
12. Zdobnov,E.M., Tegenfeldt,F., Kuznetsov,D., Waterhouse,R.M., Simão,F.A., Ioannidis,P., Seppey,M., Loetscher,A. and Kriventseva,E.V. (2017) OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.*, **45**, D744–D749.
13. Altenhoff,A.M., Glover,N.M., Train,C.-M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
14. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T.G., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.
15. Ekseth,O.K., Kuiper,M. and Mironov,V. (2014) orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics*, **30**, 734–736.
16. Horiike,T., Minai,R., Miyata,D., Nakamura,Y. and Tateno,Y. (2016) Ortholog-Finder: a tool for constructing an ortholog data set. *Genome Biol. Evol.*, **8**, 446–457.
17. Petersen,M., Meusemann,K., Donath,A., Dowling,D., Liu,S., Peters,R.S., Podsiadlowski,L., Vasilikopoulos,A., Zhou,X., Misof,B. *et al.* (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, **18**, 111.
18. Tabari,E. and Su,Z. (2017) PorthoMCL: parallel orthology prediction using MCL for the realm of massive genome availability. *Big Data Analytics*, **2**, 4.
19. Lechner,M., Findeiss,S., Steiner,L., Marz,M., Stadler,P.F. and Prohaska,S.J. (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, **12**, 124.
20. Linard,B., Thompson,J.D., Poch,O. and Lecompte,O. (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, **12**, 11.
21. Cottret,L., Rancurel,C., Briand,M. and Carrere,S. (2018) Family-Companion: analyse, visualise, browse, query and share your homology clusters. bioRxiv doi: https://doi.org/10.1101/266742, 18 February 2018, pre-print: not peer-reviewed.
22. Schreiber,F. and Sonnhammer,E.L.L. (2013) Hieranoid: hierarchical orthology inference. *J. Mol. Biol.*, **425**, 2072–2081.
23. Wagner,I., Volkmer,M., Sharan,M., Villaveces,J.M., Oswald,F., Surendranath,V. and Habermann,B.H. (2014) morFeus: a web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. *BMC Bioinformatics*, **15**, 263.
24. Zielezinski,A., Dziubek,M., Sliski,J. and Karlowski,W.M. (2017) ORCAN-a web-based meta-server for real-time detection and functional annotation of orthologs. *Bioinformatics*, **33**, 1224–1226.
25. Wang,Y., Coleman-Derr,D., Chen,G. and Gu,Y.Q. (2015) OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.*, **43**, W78–W84.
26. Nichio,B.T.L., Marchaukoski,J.N. and Raittz,R.T. (2017) New tools in orthology analysis: a brief review of promising perspectives. *Front. Genet.*, **8**, 165.
27. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
28. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
29. Altenhoff,A.M., Boeckmann,B., Capella-Gutierrez,S., Dalquen,D.A., DeLuca,T., Forslund,K., Huerta-Cepas,J., Linard,B., Pereira,C., Pryszcz,L.P. *et al.* (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.
30. Zhou,Z., Gu,J., Li,Y.-Q. and Wang,Y. (2012) Genome plasticity and systems evolution in Streptomyces. *BMC Bioinformatics*, **13**(Suppl. 10), S8.
31. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
32. Mamakani,K. and Ruskey,F. (2012) A New Rose: The First Simple Symmetric 11-Venn Diagram. arXiv doi: https://arxiv.org/abs/1207.6452, 27 July 2012, pre-print: not peer-reviewed.
33. Li,W. and Gill,B.S. (2002) The colinearity of the Sh2/A1 orthologous region in rice, sorghum and maize is interrupted and accompanied by genome expansion in the triticeae. *Genetics*, **160**, 1153–1162.
34. Ilic,K., SanMiguel,P.J. and Bennetzen,J.L. (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 12265–12270.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz333/5485531 by Xishuangbanna Tropical Botanical Garden (XTBG) user on 12 May 2019