A database for orphan genes in Poaceae

CHENSONG YAO[1*], HANWEI YAN[2*], XIAODAN ZHANG[3*] and RONGFU WANG[4]

[1]Graduate School; [2]Laboratory of Modern Biotechnology; [3]School of Information and Computer Science; [4]Department of Life Sciences, Anhui Agricultural University, Hefei, Anhui 230036, P.R. China

Received October 12, 2016; Accepted June 8, 2017

DOI: 10.3892/etm.2017.4918

Abstract. Orphan genes refer to a group of protein-coding genes lacking recognizable homologs in the other organisms. Extensive studies have demonstrated that numerous newly sequenced genomes contain a significant number of orphan genes, which have important roles in plant's responses to the environment. Due to a lack of phylogenetic conservation, the origin of orphan genes and their functions are currently not well defined. In the present study, a Poaceae orphan genes database (POGD; http://bioinfo.ahau.edu.cn/pogd) was established to serve as a user-friendly web interface for entry browsing, searching and downloading orphan genes from various plants. Four Poaceae species, including *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor* and *Zea mays*, are included in the current version of POGD. The database provides gene descriptions (chromosome strands, physical location), gene product records (protein length, isoelectric point, molecular weight as well as gene and protein sequences) and functional annotations (cellular role, gene ontology category, subcellular localization prediction). Basic Local Alignment Search Tool and comparative analyses were also provided on the website. POGD will serve as a comprehensive and reliable repository, which will help uncover regulatory mechanisms of orphan genes and may assist in the development of comparative genomics in plant biology.

Introduction

Orphan genes, with coding sequences being utterly unique or genes producing previously non-existing (novel) proteins, have important roles in the function and evolution of biological networks that are conserved among various species (1). To the best of our knowledge, these genes were initially discussed when analyzing the yeast genome, and approximately one third of the identified genes were defined as orphan genes according to their genetic features (2). Previously, several studies have reported that orphan genes comprised a considerable fraction of genes in all domains of species (3-5). It was estimated that 1-71% of genes are orphan genes in various species (4,6-15), among which 5-15% are relatively typical (7,13-15). Despite the fact that these genes are abundant in quantity, their evolutionary and functional roles have remained to be elucidated. To date, several orphan genes have been reported to have imperative roles in several developmental processes. For instance, certain products of orphan genes were reported to be essential for early brain development in humans (16), as well as the regulation of tentacle formation in hydra species (17). By now, genome sequencing has been accomplished in several plant species, and several orphan genes have been reported to be closely involved in regulating responses to the environment (18-22). Therefore, identification and comparison of orphan genes on a genome-wide scale will greatly contribute to the current understanding of the molecular functions and the evolution of orphan genes.

The present study reported on a novel online database resource named Poaceae orphan genes database (POGD) that offers comprehensive information about orphan genes in four Poaceae species, including *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor* and *Zea mays*. The information was presented in a friendly web interface, which listed the orphan genes of each species and included a series of functional annotations. Compared to databases dedicated to individual organisms, the POGD database provides comparative analyses of genomic data. In conjunction with the Basic Local Alignment Search Tool (BLAST), it will efficiently facilitate the analysis and extraction of data generated from the POGD. To the best of our knowledge, databases focusing on orphan genes in plant species are currently rare. Therefore, the POGD database will be a valuable data resource, particularly for the investigation of molecular functions and the evolution of orphan genes in plant species.

Materials and methods

Dataset collection. Protein sequences and gene coordinate information of 54 plant species were downloaded from the Phytozome database (version 10.3; https://phytozome.jgi.doe.gov/pz/portal.html) (23). Furthermore, protein sequences and gene coordinate information of 69 animal species was retrieved from the Ensembl database (release 81; http://www.ensembl.org) (24).

Correspondence to: Professor Rongfu Wang, Department of Life Sciences, Anhui Agricultural University, 130 Changjiang West Road, Hefei, Anhui 230036, P.R. China
E-mail: rfwangahau@163.com

*Contributed equally

Key words: orphan genes, database, Poaceae

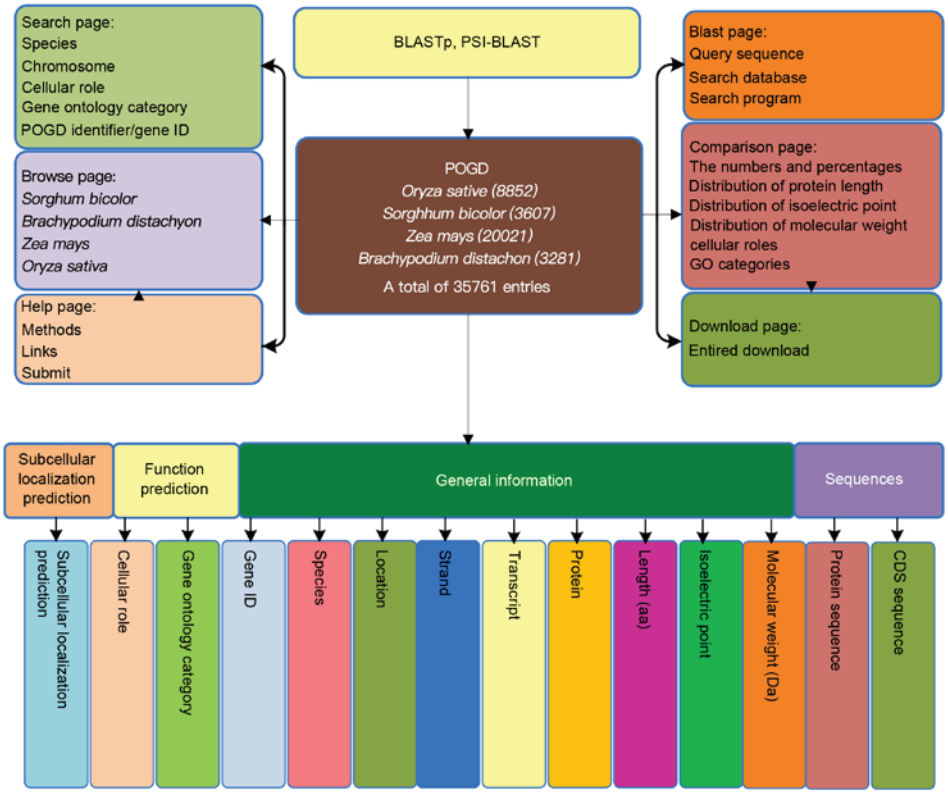

Figure 1. Overview of the architecture of POGD. The web-accessible POGD allows for four Poaceae plant orphan genes to be clearly browsed, searched, downloaded and compared, under a well-organized platform framework. POGD, Poaceae orphan genes database; GO, gene ontology; BLAST, Basic Local Alignment Search Tool; PSI, Position-Specific Iterative; BLASTp, protein BLAST; CDS, coding DNA sequence.

Identification of orphan genes. Identification of orphan genes in Poaceae species was performed according to previous studies with certain modifications (25-31). In brief, a systematic method was used based on homology search. Protein BLAST (BLASTp) was used to identify the homologs of all proteins annotated to each Poaceae species against the protein sets of other plant and animal species with an e-value cutoff at 10^{-5}. To eliminate the false positivity caused by incompleteness of the annotated protein sets, the obtained proteins were then searched against the current non-redundant protein database in of the US National Center for Biotechnology Information using BLASTp. In order to further screen for similarity between sequences, the Position-Specific Iterated BLAST (PSI-BLAST) method (32) was used to identify weaker homologous associations, which may have been missed by standard BLAST algorithms. Finally, a dataset containing 3,281 orphan genes in *Brachypodium distachyon*, 8,852 in *Oryza sativa*, 3,607 in *Sorghum bicolor* and 20,021 orphan genes in *Zea mays* were obtained.

Database construction. As a web-based platform, the POGD database was a combination of the MySQL database management system (version 5.5.8) and a dynamic web interface based on PHP (version 5.3.3) and Javascript (version 1.2). As illustrated in Fig. 1, the architecture of POGD was stratified and structured (Fig. 1).

The detailed annotations of orphan genes in Poaceae species were integrated in the POGD database. Physical locations as well as gene strand and protein sequence length were obtained from Phytozome. The isoelectric points (PI) and molecular weights (Mw) were retrieved from Expasy (http://www.expasy.org/) (33). Function prediction based on the cellular role and Gene Ontology (GO) were collected from Protfun (http://www.cbs.dtu.dk/services/ProtFun/) (34). WOLF PSORT (http://www.genscript.com/psort/wolf_psort.html) (35) was used for the prediction of subcellular localization.

Results

In total, 35,761 orphan genes from four sequenced Poaceae species were accumulated in the POGD database (Table I). The designed database web portal comprised the following components: Home, Search, BLAST, Browse, Comparison, Download and Help (Fig. 2). By clicking on the 'Browse' column, the overview and image of each species was displayed, in which the chromosomal distribution, functional categorization (cellular role and GO category) (Fig. 3), the distribution of PI, Mw and protein length were provided (Fig. 4). The option to open a new page with detailed annotations for each gene by directly clicking on the image was also available. The results were provided in the form of a hit list represented by the POGD identifier, gene ID, species, chromosome, cellular role and GO category of all associated gene entries. Access to an individual entry page containing the gene annotation was given upon clicking on the POGD identifier or Gene ID, which included chromosome strand, physical location, PI, Mw, protein length, coding DNA sequence (CDS) and protein sequence.

Easy access to detailed annotations of interest was provided via a truncated version or the entire Gene ID at the top right of each page. In the search page, a more advanced search allowed for filtering by parameters, such as species, chromosome,

POGD identifier	Gene ID	Species	Chromosome	Cellular role	Gene ontology category
BdOG01_00001	Bradi1g00100	*Brachypodium distachyon*	1	Energy_metabolism	Hormone
BdOG01_00002	Bradi1g00205	*Brachypodium distachyon*	1	Replication_and_transcription	Transcription
BdOG01_00003	Bradi1g00253	*Brachypodium distachyon*	1	Translation	Structural_protein
BdOG01_00004	Bradi1g00291	*Brachypodium distachyon*	1	Translation	Structural_protein
BdOG01_00005	Bradi1g00316	*Brachypodium distachyon*	1	Translation	Structural_protein
BdOG01_00006	Bradi1g00341	*Brachypodium distachyon*	1	Cell_envelope	Structural_protein
BdOG01_00007	Bradi1g00362	*Brachypodium distachyon*	1	Cell_envelope	Transporter
BdOG01_00008	Bradi1g00364	*Brachypodium distachyon*	1	Energy_metabolism	Growth_factor
BdOG01_00009	Bradi1g00366	*Brachypodium distachyon*	1	Translation	Transcription_regulation
BdOG01_00010	Bradi1g00412	*Brachypodium distachyon*	1	Translation	
BdOG01_00011	Bradi1g00503	*Brachypodium distachyon*	1	Energy_metabolism	Structural_protein
BdOG01_00012	Bradi1g00625	*Brachypodium distachyon*	1	Biosynthesis_of_cofactors	Transcription_regulation
BdOG01_00013	Bradi1g00630	*Brachypodium distachyon*	1	Translation	Growth_factor
BdOG01_00014	Bradi1g00642	*Brachypodium distachyon*	1	Transport_and_binding	Cation_channel
BdOG01_00015	Bradi1g01341	*Brachypodium distachyon*	1	Translation	Transcription_regulation
BdOG01_00016	Bradi1g01362	*Brachypodium distachyon*	1	Translation	Transcription
BdOG01_00017	Bradi1g01596	*Brachypodium distachyon*	1	Translation	Structural_protein
BdOG01_00018	Bradi1g01645	*Brachypodium distachyon*	1	Energy_metabolism	Structural_protein
BdOG01_00019	Bradi1g01686	*Brachypodium distachyon*	1	Replication_and_transcription	Transcription_regulation
BdOG01_00020	Bradi1g01689	*Brachypodium distachyon*	1	Replication_and_transcription	Growth_factor

Figure 2. Overview of the website and gene annotation page. (A) Home page. (B) Search page. (C) Example of the gene annotation page. POGD, Poaceae orphan genes database.

Table I. Orphan genes in each species.

Species	Orphan genes, n (%)
Brachypodium distachyon	3,281 (10.35)
Oryza sativa	8,852 (22.78)
Sorghum bicolor	3,607 (10.92)
Zea mays	20,021 (31.54)

cellular role, GO category, POGD identifier and gene ID (Fig. 2B). Finally, the users were given free access to navigate from the search results to pages containing detailed annotations (Fig. 2C).

The POGD database provided the tool utility BLAST, where an online interface was available to input any sequence of interest in fasta format. A search against all orphan genes in this database was also available. All of the processed data contained in this database were made available on the download page.

A comparative analysis of the genomes of *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor* and *Zea mays* was available in the 'Comparison' module. The percentage of orphan genes contained in the genome of each species was 10.35% (n=3,281), 22.78% (n=8,852), 10.92% (n=3,607) and 31.54% (n=20,021), respectively (Fig. 5A; Table I). Four bar charts displaying the protein length were obtained in order to investigate the general trends in protein length distribution (Fig. 5B). The average protein length was 142 amino acids (aa) in *Brachypodium distachyon*, while that in the *Oryza sativa*, *Sorghum bicolor* and *Zea mays* was 138, 120 and 114 aa, respectively, which was in agreement with the results of previous studies, according to which the protein products of orphan genes were shorter than those of non-orphan genes (6,10,15,27). Furthermore, in all of the four Poaceae species, a bimodal distribution of PI was observed in proteins encoded by orphan genes, in which a smaller third peak between two main peaks was noticed (Fig. 5C). The Mw of each orphan gene was unimodal with a peak value ranging from 7 to 13 kDa (Fig. 6A). In the present study, GO annotations

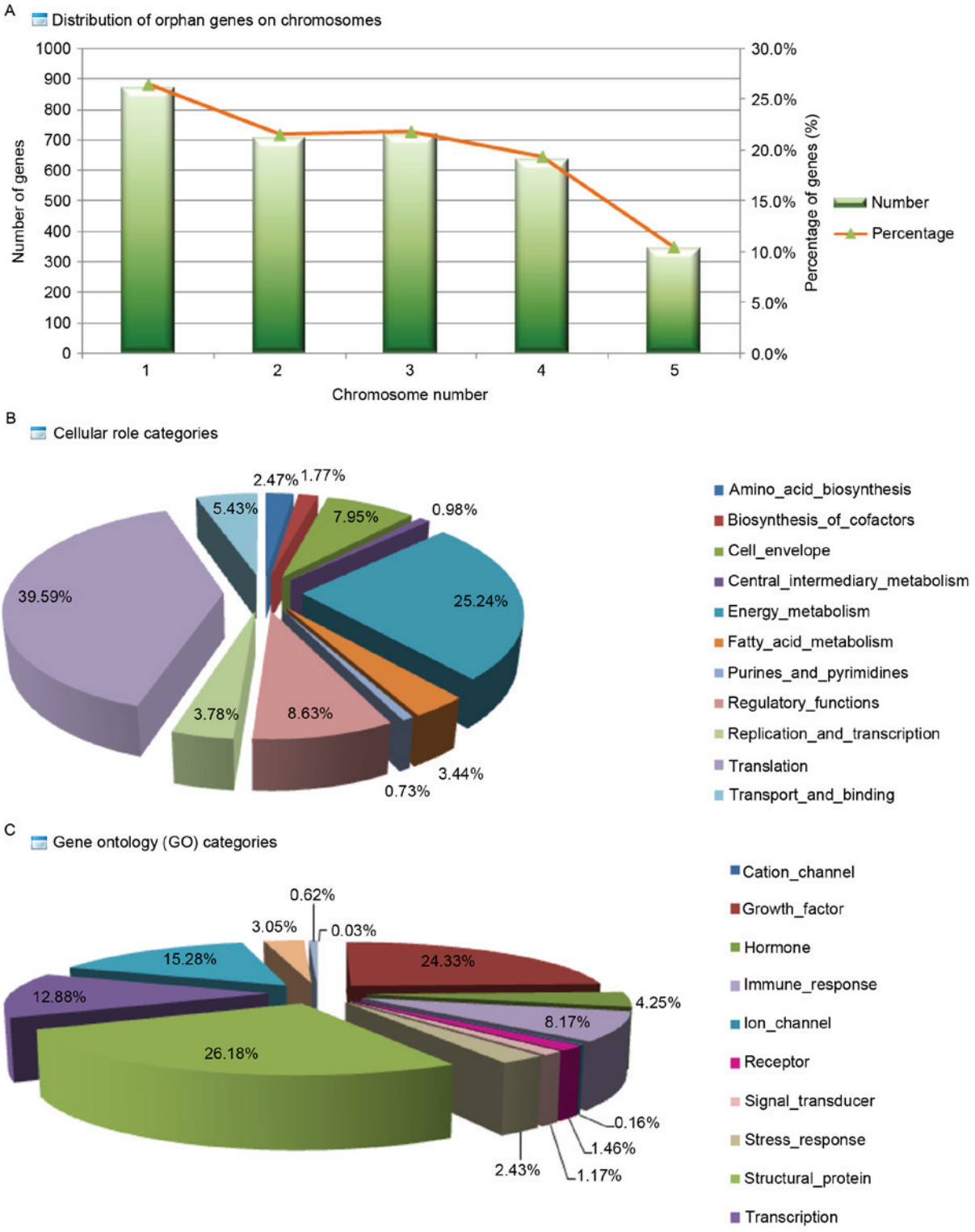

Figure 3. Summary data for orphan genes identified in *Brachypodium distachyon*. (A) Distribution of orphan genes on chromosomes. (B) Cellular role categories. (C) GO categories. GO, gene ontology.

were predicted together with a comprehensive comparison to trace the potential functional significance of orphan genes. The results indicated a similarity across species in the percentages of orphan genes that fell under the different functional categories. Of note, translation was the most abundant functional category among the cellular roles and structural protein was the most common GO category (Fig. 6B and C). Comparative analysis of orphan genes indicated the presence of several common functional and evolutionary characteristics, which may be present in other major eukaryote kingdoms.

Discussion

Orphan genes have been widely identified across all domains of life with the advance of next-generation sequencing technology (3,4,18). At present, research focuses on the traits

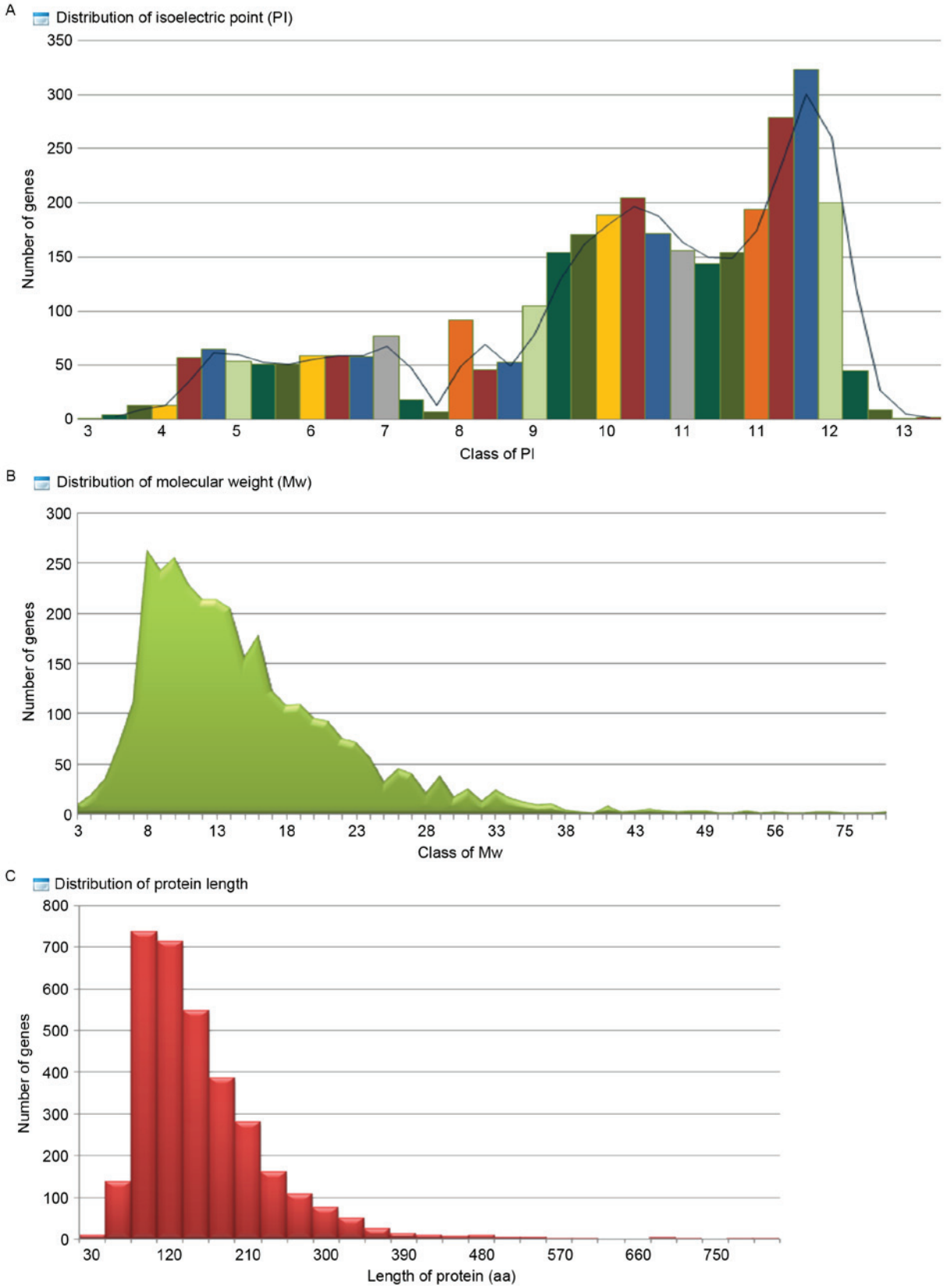

Figure 4. Summary data for orphan genes identified in *Brachypodium distachyon*. (A) Distribution of PI. (B) Distribution of Mw. (C) Distribution of protein length. PI, isoelectric point; Mw, molecular weight; aa, number of amino acids.

of orphan genes, and numerous attempts have been made to explore their evolutionary and functional roles in different species. In particular, accumulating evidence has demonstrated that orphan genes are closely involved in the response to environmental stress and species-specific traits or regulatory patterns in plants. Therefore, it is necessary to understand the basis for orphan gene evolution in synthetic biology. Establishing a database for plant orphan genes contributes to

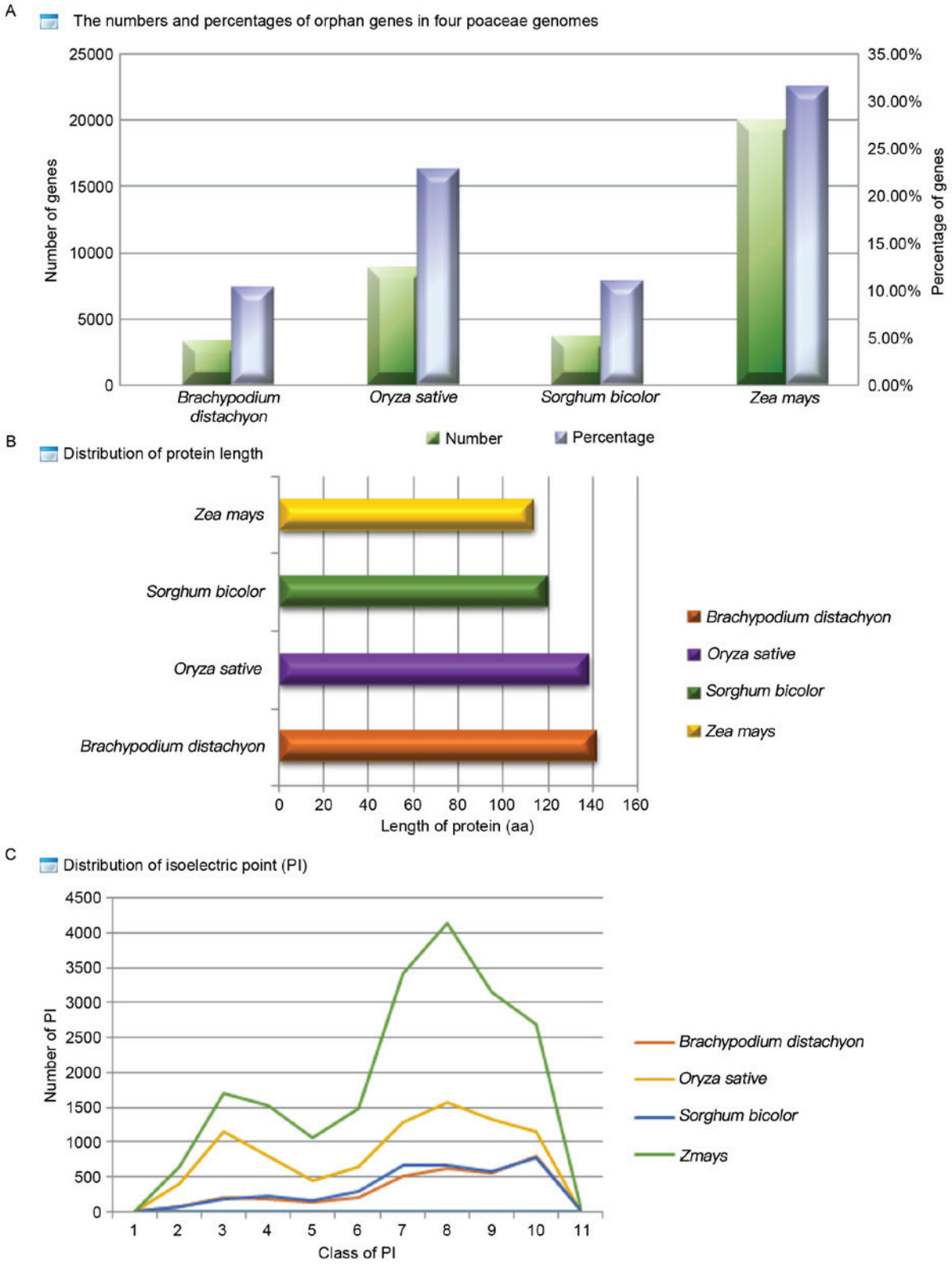

Figure 5. Comparative analysis of orphan genes in four different species in the POGD. (A) Number and percentage of orphan genes. (B) Length of protein. (C) PI. PI, isoelectric point; aa, number of amino acids.

advances in the field by providing valuable genomic resources for data mining. To the best of our knowledge, no databases of orphan genes in plant species are currently available. In the present study, the POGD database providing a platform for obtaining detailed information on these orphan genes was established. The database web portal offers a comprehensive repository of four Poaceae species. A total of 35,761 orphan genes with extensive annotations and a user-friendly web interface are contained in the database.

A detailed bioinformatics analysis was provided for each gene, including chromosome strand, physical location, protein length, PI, Mw, CDS, protein sequence, cellular role,

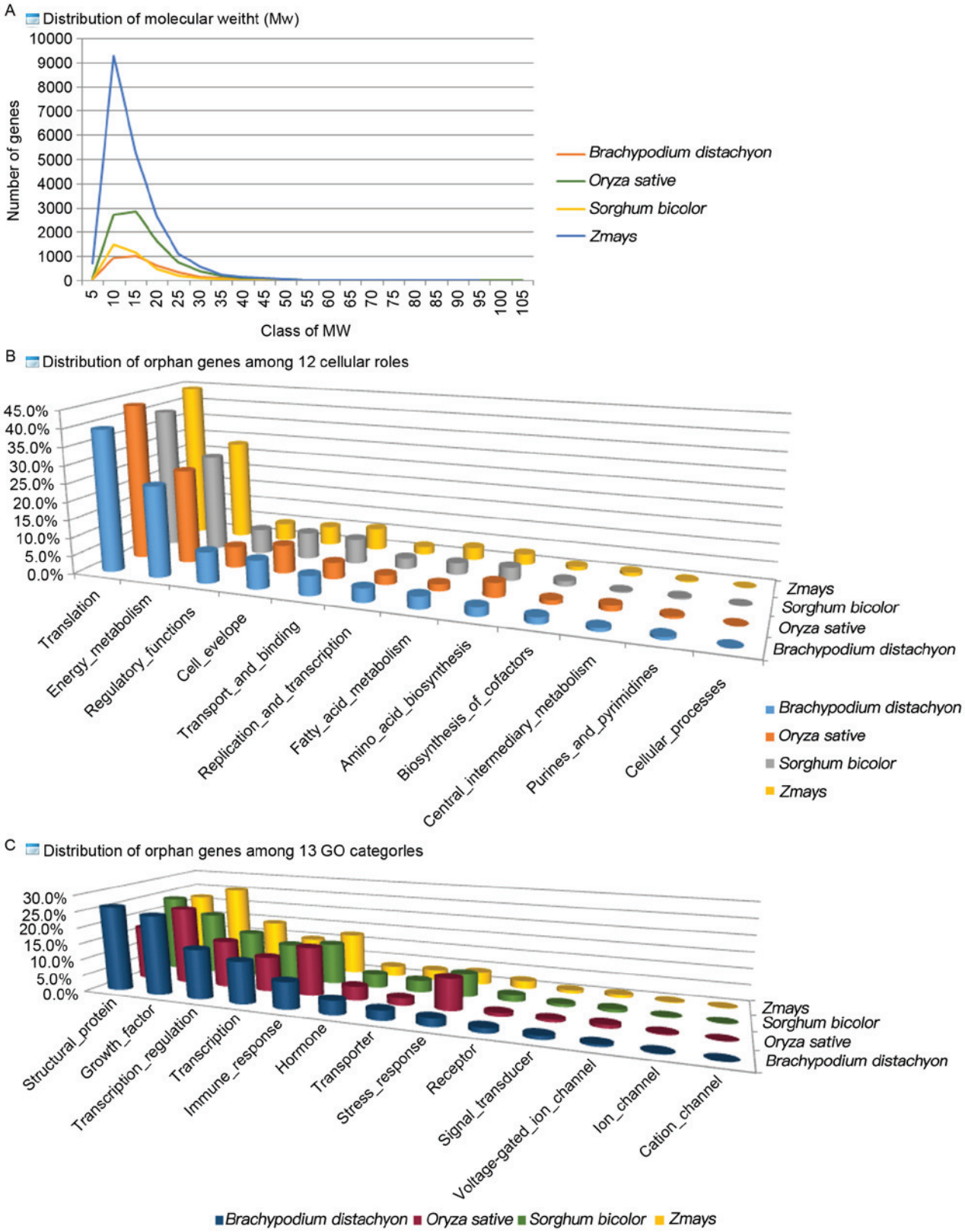

Figure 6. Comparative analysis of orphan genes in four different species in the POGD. (A) Molecular weight. (B) Cellular role. (C) GO. GO, gene ontology; Mw, molecular weight.

GO category, as well as subcellular localization prediction. Furthermore, the database provides online tools for sequence similarity searches such as BLAST. Furthermore, a convenient submission interface was also provided, through which independent researchers may upload information on novel orphan genes. For scientific communication and data sharing, an interactive platform was established in the database, where researchers may download freely available data. In addition, associated external databases were accessible via the web links provided on the POGD database platform.

POGD was developed whilst keeping an eye on scalable aspects of the datasets in order to expand this project in the near future. On this basis, several available plant species will be added with more annotations for comparative analysis. Furthermore, powerful comparative analysis tools will be developed for further in-depth study on orphan genes. The POGD database will be updated with the progression in the field. Indeed, integration and analysis of orphan gene data may provide crucial clues to understand differences in the differentiation, morphology and chemical composition between species.

Acknowledgements

This study was supported by the Anhui Provincial Natural Science Foundation (grant no. 1608085QC65), the China Postdoctoral Science Foundation (grant no. 2015M581806) and the Cultivating Academic Backbone Foundation of Anhui Agricultural University (grant no. 2014XKPY-04).

References

1. Arendsee ZW, Li L and Wurtele ES: Coming of age: Orphan genes in plants. Trends Plant Sci 19: 698-708, 2014.
2. Dujon B: The yeast genome project: What did we learn? Trends Genet 12: 263-270, 1996.
3. Khalturin K, Hemmrich G, Fraune S, Augustin R and Bosch TC: More than just orphans: Are taxonomically-restricted genes important in evolution? Trends Genet 25: 404-413, 2009.
4. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ and Field D: Orphans as taxonomically restricted and ecologically important genes. Microbiology 151: 2499-2501, 2005.
5. Yin Y and Fischer D: Identification and investigation of ORFans in the viral world. Bmc Genomics 9: 24, 2008.
6. Donoghue MT, Keshavaiah C, Swamidatta SH and Spillane C: Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. BMC Evol Biol 11: 47, 2011.
7. Wissler L, Gadau J, Simola DF, Helmkampf M and Bornberg-Bauer E: Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol 5: 439-455, 2013.
8. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, *et al*: Proto-genes and de novo gene birth. Nature 487: 370-374, 2012.
9. Ekman D and Elofsson A: Identifying and quantifying orphan protein sequences in fungi. J Mol Biol 396: 396-405, 2010.
10. Yang L, Zou M, Fu B and He S: Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. Bmc Genomics 14: 65, 2013.
11. Hahn MW, Han MV and Han SG: Gene family evolution across 12 Drosophila genomes. PLoS Genet 3: e197, 2007.
12. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, *et al*: The ecoresponsive genome of *Daphnia pulex*. Science 331: 555-561, 2011.
13. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F, *et al*: Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen *Dothideomycetes* fungi. PLoS Pathog 8: e1003037, 2012.
14. Gibson AK, Smith Z, Fuqua C, Clay K and Colbourne JK: Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick *Amblyomma americanum*. Bmc Genomics 14: 135, 2013.
15. Kuo CH and Kissinger JC: Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. BMC Evol Biol 8: 108,2008.
16. Zhang YE, Landback P, Vibranovski MD and Long M: Accelerated recruitment of new brain development genes into the human genome. PLoS Biol 9: e1001179, 2011.
17. Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G and Bosch TC: A novel gene family controls species-specific morphological traits in *Hydra*. PLoS Biol 6: e278, 2008.
18. Gollery M, Harper J, Cushman J, Mittler T, Girke T, Zhu JK, Bailey-Serres J and Mittler R: What makes species unique? The contribution of proteins with obscure features. Genome Biol 7: R57, 2006.
19. Chen WH, Trachana K, Lercher MJ and Bork P: Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol 29: 1703-1706, 2012.
20. Luhua S, Ciftci-Yilmaz S, Harper J, Cushman J and Mittler R: Enhanced tolerance to oxidative stress in transgenic Arabidopsis plants expressing proteins of unknown function. Plant Physiol 148: 280-292, 2008.
21. Gollery M, Harper J, Cushman J, Mittler T and Mittler R: POFs: What we don't know can hurt us. Trends Plant Sci 12: 492-496, 2007.
22. Lacombe S, Rougon-Cardoso A, Sherwood E, Peeters N, Dahlbeck D, van Esse HP, Smoker M, Rallapalli G, Thomma BP, Staskawicz B, *et al*: Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. Nat Biotechnol 28: 365-369, 2010.
23. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N and Rokhsar DS: Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res 40 (Database issue): D1178-D1186, 2012.
24. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, *et al*: Ensembl 2015. Nucleic Acids Res 43 (Database Issue): D662-669, 2015.
25. Tautz D and Domazet-Lošo T: The evolutionary origin of orphan genes. Nat Rev Genet 12: 692-702, 2011.
26. Domazet-Loso T and Tautz D: An evolutionary analysis of orphan genes in *Drosophila*. Genome Res 13: 2213-2219, 2003.
27. Tollriera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X and Albà MM: Origin of primate orphan genes: A comparative genomics approach. Mol Biol Evol 26: 603-612, 2009.
28. Lin H, Moghe G, Ouyang S, Iezzoni A, Shiu SH, Gu X and Buell CR: Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. BMC Evol Biol 10: 41, 2010.
29. Zhang G, Wang H, Shi J, Wang X, Zheng H, Wong GK, Clark T, Wang W, Wang J and Kang L: Identification and characterization of insect-specific proteins by genome data analysis. BMC Genomics 8: 93, 2007.
30. Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, Haas BJ, Hamilton JP and Buell CR: Identification and characterization of lineage-specific genes within the Poaceae. Plant Physiol 145: 1311-1322, 2007.
31. Yang X, Jawdy S, Tschaplinski TJ and Tuskan GA: Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*. Genomics 93: 473-480, 2009.
32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997.
33. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, *et al*: ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40: W597-W603, 2012.
34. Jensen LJ, Gupta R, Staerfeldt HH and Brunak S: Prediction of human protein function according to gene ontology. Bioinformatics 19: 635-642, 2003.
35. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ and Nakai K: WoLF PSORT: Protein localization predictor. Nucleic Acids Res 35: W585-W587, 2007.