INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
19F–046/2019/22–5–961–966
DOI: 10.17957/IJAB/15.1155
http://www.fspublishers.org

Full Length Article

TOGD: A Database of Orphan Genes in *Triticum aestivum*

Qijuan Gao[1], Hanwei Yan[2], Enhua Xia[3], Shihua Zhang[4] and Shaowen Li[1,2*]
[1]*Anhui Province Key Laboratory of Smart Agricultural Technology and Equipment, Anhui Agricultural University, Hefei 230036, China*
[2]*Key Laboratory of Crop Biology of Anhui Province, Anhui Agricultural University, Hefei 230036, China*
[3]*State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei 230036, China*
[4]*Institute of Applied Mathematics, School of Science, Anhui Agricultural University, Hefei 230036, China*
[*]For correspondence: shwli@ahau.edu.cn

Abstract

Orphan genes, genes without detectable homologous in other lineages, have important biological roles in the environmental adaptation of different species. Wheat (*Triticum aestivum* L.) is a globally consumed crop of immense economic importance, but the evolutionary genesis of wheat orphan genes remains poorly elucidated. In this study, 993 orphan genes were identified in the wheat genome by homology searching against 94 representative plant species. The features of these orphan genes (*e.g*., subcellular location, molecular weight, gene structure, and expression patterns) were characterized and integrated into a specially designed, web-accessible wheat orphan gene database (TOGD; http://togd.ahau-edu.cn/). A flexible search engine was developed with multiple options to allow users to easily extract and visualize datasets. Orphan-gene information returned by this search engine includes chromosome location, putative functions, protein length, guanine and cytosine content, isoelectric point, molecular weight, subcellular location, gene structure, and expression patterns from external databases. A BLAST tool for exploration of homologous of given sequences was also implemented. As there is no available published orphan genes database in wheat, which will help in wheat breeding and seed production through uncovering regulatory mechanisms of orphan genes and may assist in the development of comparative genomics in wheat biology. Therefore, this constructed wheat orphan gene database should serve as a comprehensive bioinformatics platform for functional and evolutionary studies of orphan genes in *T. aestivum.* © 2019 Friends Science Publishers

Keywords: Wheat; Orphan gene; Environmental adaption; Blast tool; Chromosome location

Introduction

Orphan genes, also called taxonomically restricted genes, are genes that have no detectable homologous in other lineages. Ubiquitous in plants, orphan genes account for 1–71% of genes in various species, with 5–15% being typical of whole gene contents across species (Arendsee *et al.*, 2014). The essential roles of orphan genes in developmental processes and the environmental adaptations of species are well reported (Alexandre *et al.*, 2015). For instance, *Qua-Quine Starch* (*QQS*) is an orphan gene in *Arabidopsis thaliana* that regulates carbon and nitrogen distribution between proteins and carbohydrates. The introduction of *QQS* into species lacking a *QQS* homolog, such as soybean (*Glycine max*), rice (*Oryza sativa* L.) and maize (*Zea mays* L.) can significantly increase protein levels and decrease starch contents in leaves and seeds of these species (Li *et al.*, 2009; Li and Wurtele, 2015). The *GN2*, a rice orphan gene, not only influences grain number, but also affects plant height and heading date (Chen *et al.*, 2017). Another two orphan genes (*GmHsp28.6, GmHsp28.7)* of soybean, which were found to show a unique occurrence pattern among genes correlation with nematode infection (Lopes-Caitar *et al.*, 2013).

Wheat (*Triticum aestivum* L.) is a major global cereal essential to human nutrition. In the last decade, wheat genome sequencing has progressed, and the release of a wheat reference genome has provided an unprecedented opportunity to systemically investigate the landscape of its orphan genes. Perochon *et al.* (2015) first reported that *TaFROG*, an orphan gene specific to the grass subfamily Pooideae, contributes to resistance against a wheat disease, *Fusarium* head blight. Afterwards, Ni *et al.* (2016) discovered that wheat *Ms2* is an orphan gene in grass species and regulates male sterility in wheat. These results suggest that orphan genes have a significant role in wheat sterility and disease resistance. The construction of a searchable orphan gene database would greatly aid in-depth investigation of the roles of orphan genes in species evolution and adaptation. Several orphan gene databases have been established, including yMGV (Marc *et al.*, 2001), possibly the first such

To cite this paper: Gao, Q., H. Yan, E. Xia, S. Zhang and S. Li, 2019. TOGD: a database of orphan genes in *Triticum aestivum*. *Intl. J. Agric. Biol.*, 22: 961–966

database, and POGD, related to Poaceae species (Yao *et al.*, 2017).

In this study, an online database (TOGD) for wheat orphan genes to allow researchers to better understand the functional and evolutionary nature of wheat orphan genes was constructed. Along with this manually curated database, a user-friendly web interface was implemented for retrieval of details of specific orphan genes, for retrieval of details of specific orphan genes, including chromosome location, putative functions, protein length, GC content, isoelectric point, molecular weight, subcellular location, gene structure, and expression patterns from external databases. The basic local alignment search tool (BLAST) can be used to efficiently facilitate the analysis and extraction of data generated from the TOGD. The current version of the database hosts 993 wheat orphan genes obtained using BLAST against 94plant genomes. TOGD is a valuable and continuously updated data resource, which is highly useful for further exploration of the molecular functions and evolution of orphan genes in wheat.

Materials and Methods

Data Collection

To construct the TOGD database, datasets were downloaded, such as datasets for annotation and protein and coding sequences, from the wheat genomics database (https://wheat-urgi.versailles.inra.fr/Seq-Repository; Alaux *et al.*, 2018). Protein sequences and gene coordinate information for 94 plant species, representing nearly all currently well-sequenced plant species, were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html; Goodstein *et al.*, 2012). Non-redundant protein sequences (NR) were obtained from the NCBI database (https://www.ncbi.nlm.nih.gov/; Wheeler *et al.*, 2005).

Identification of Wheat Orphan Genes

Identification of orphan genes in *T. aestivum* was performed according to previous studies with a few modifications (Neme and Tautz, 2013; Palmieri *et al.*, 2014). Briefly, protein BLAST (BLASTp) was used to perform a homology search of all *T. aestivum* proteins against the protein sets of the 94 plant species, with an *e*-value cutoff of 1×10^{-5}. Proteins with no detectable homologous were subjected to a further search against the genomic sequences of the same species using tBLASTn. To eliminate false positives caused by missing data in annotated protein sets, the obtained proteins were searched against the NCBI non-redundant protein database using BLASTp (Wheeler *et al.*, 2005). Finally, genes with protein lengths less than 30aa were removed. In total, 993 orphan genes were identified in the *T. aestivum* genome.

Annotation of Wheat Orphan Genes

Identified wheat orphan genes were annotated using different databases. First, the genomic features of the orphan genes, including chromosome locations, gene strand, coding and protein sequences, GC content, exon number and sequence length were retrieved from wheat annotation files. Next, isoelectric points (pIs) and molecular weights (MWs) were calculated using the EXPASY online tool (https://web.expasy.org/), and Gene Ontology functional annotations were directly collected from Phytozome. Subcellular localizations were predicted using WOLFPSORT (http://www.genscript.com/pscort.html; Horton *et al.*, 2007), and expression profiles were directly retrieved and cross-linked to the Wheat Expression Browser (http://www.wheat-expression.com/). Paralog information was extracted from the ENSEMBLE database (https://www.ensembl.org/; Shih, 2016).

Results

Characterization of Wheat Orphan Genes

Using the above-described methods, 137,052 protein sequences were selected and then searched for homologous of genes encoding 136,050 of the annotated *T. aestivum* proteins in each of the other 94 plant species. 993 orphan genes in *T. aestivum* were thereby identified, which corresponds to 0.8% of all wheat genes.

An overview of annotated orphan genes and their exon–intron structures is shown in Fig. 1 and a statistical summary of chromosomes, subcellular locations, and distributions of protein lengths, PIs, and MWs is presented in Fig. 2. Protein lengths of orphan genes were generally short, mainly ranging around 30–200aa (Fig. 2C). In addition, predicted subcellular locations were mainly on chloroplasts, nuclei and the cytoplasm. Most orphan genes were found on A-subgenome chromosomes, with nearly equal quantities present on B and D chromosomes. The chromosome with the highest number of orphan genes, 115, was chromosome 7A. The distribution of pIs of proteins encoded by orphan genes is shown in Fig. 2A and MW values of orphan genes, which ranged from 10 to 30 kDa, are given in Fig. 2B. Most orphan genes contained only a few exons; the proportion of one- and two-exon genes was 46 and 32%, respectively (Fig. 3).

Implementation of the Wheat Orphan Gene Database

TOGD was constructed using three major software programs: MySQL database, Windows Server 2008 R2, and PHP-based computational toolkits. The orphan gene datasets and their relevant resources are stored in a MySQL database on a Linux system. The datasets contain data on literature-curated *T. aestivum* orphan genes, correlations, and related sources (*e.g.*, NCBI data). The web services are

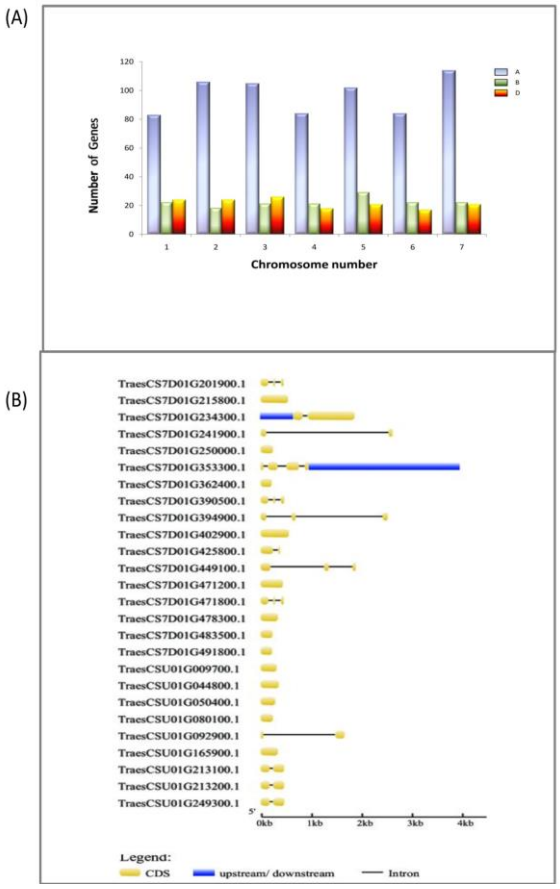

Fig. 1: Chromosome locations (**A**) shows the details: 'A' delegating A genome lineage; 'B' delegating B genome lineage; 'D' delegating D genome lineage respectively, and gene structures (**B**) of wheat orphan genes

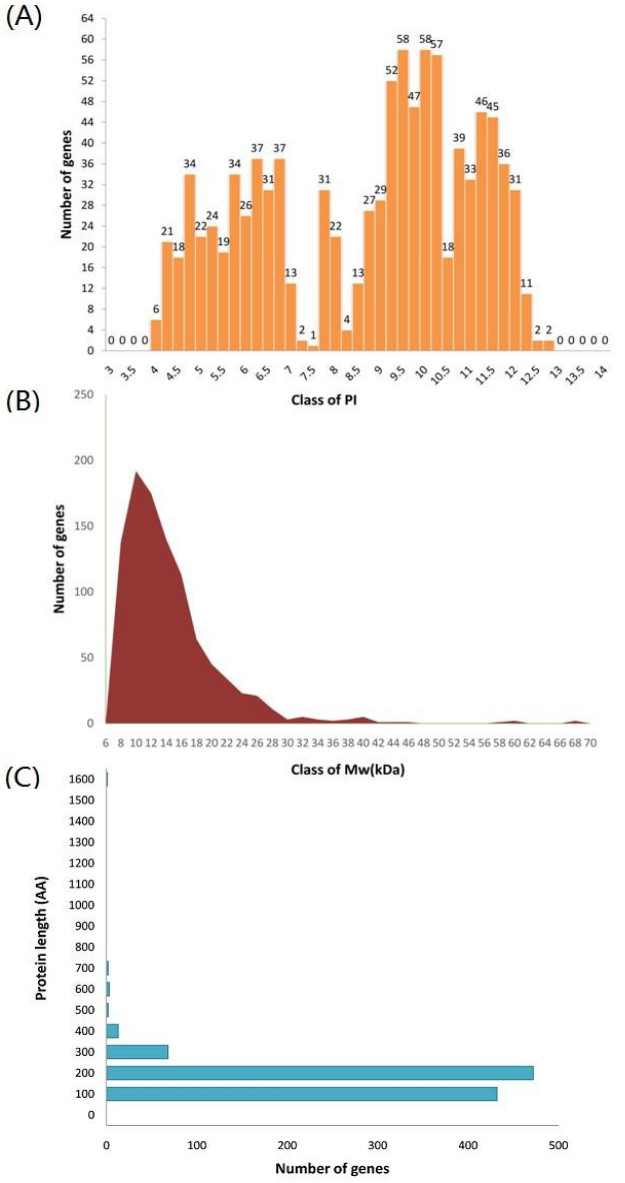

Fig. 2: Statistical summary of orphan genes identified in *Triticum aestivum.* Distributions of isoelectric points (pIs) (**A**), molecular weights (MWs) (**B**), and protein lengths (**C**)
AA = Number of amino acids

run on an Apache server, a popular, widely used application supporting multiple plug-ins that benefit server enhancement. Automatic web-page layout of data-driven documents is provided using Bootstrap, CSS, HTML, and JavaScript (version 1.2). TOGD is a web-based, cross-platform framework for rapid data visualization. To achieve the best visualization, Google Chrome and IE 9.0+ are recommended. The current release of TOGD, which is available at http://togd-ahau.edu.cn, contains 993 orphan genes from *T. aestivum*. The architecture of TOGD is shown in Fig. 4A. The web portal of the designed database comprises six main components: Home, Search, Statistics, Blast, Download, and Update (Fig. 4B).

Search Engine

A flexible search engine was developed to ease retrieval of datasets from TOGD (Fig. 5). After clicking 'Search' on the main navigation bar, users can choose one of six search options: gene ID, chromosome number, protein length, isoelectric point, molecular weight, or exon number. For instance, users can search the TOGD by gene ID, which returns a page showing the gene ID, location, gene length, and exon number. The chromosome number and exon number options can then be used to retrieve related orphan genes. Alternatively, users can enter specific information, such as protein length=100–200aa, isoelectric point= 6–8, and molecular weight=4,000–6,000, from the original search page. Detailed annotations can subsequently be easily accessed by entering a complete or truncated Gene ID at the top left of each page; this option opens a new page with detailed annotations, including gene identifier, chromosome location, strand, protein length, isoelectric

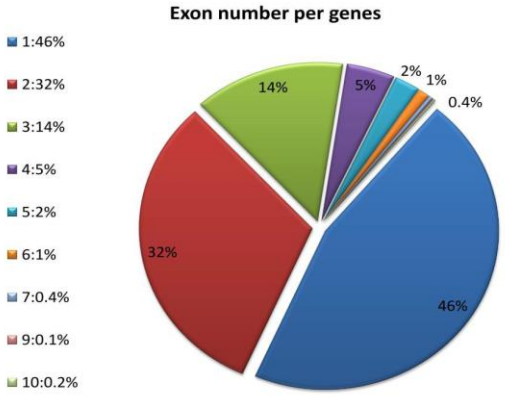

Fig. 3: Exon number distributions and the percentage of wheat orphan genes

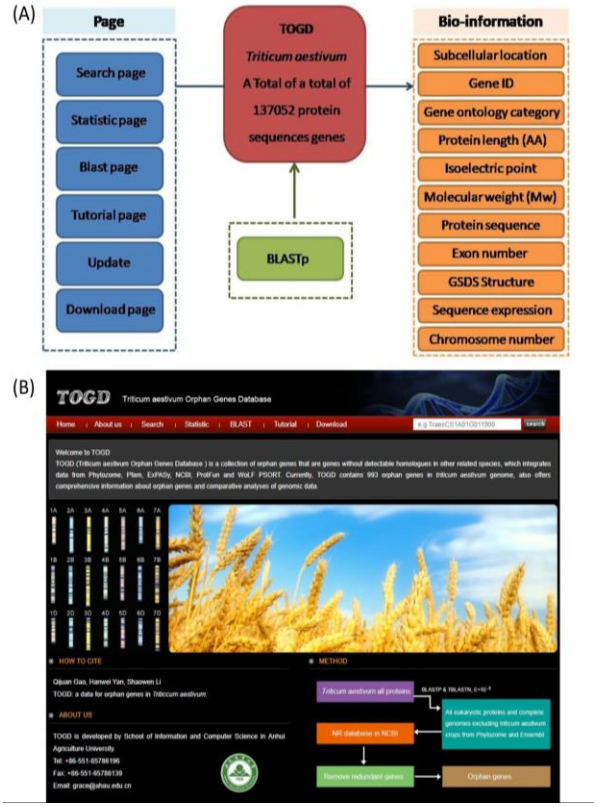

Fig. 4: Overview of the *T. aestivum* orphan gene database (TOGD): (**A**) Database architecture, (**B**) Screenshot of the homepage

The web-accessible TOGD facilitates searching, downloading, and analysis of orphan genes of *T. aestivum*. BLAST, basic local alignment search tool; PSI, position-specific iterative; BLASTp, protein BLAST

point (PI), molecular weight (MW), homolog groups, and expression profiles for each gene. Those protein sequence features display protein identities that can hyperlinked to external databases to access more information. 401 paralogous pairs of orphan genes were identified in *T. aestivum* and have provided their *Ka* and *Ks* values to allow determination of their evolutionary distances. Expression profiles of genes from different tissues, developmental stages and experimental treatments can be retrieved, and a visualized line diagram showing expression changes across different tissues and developmental stages is also available. In addition, subcellular localization predictions are shown because they are important clues to protein roles.

BLAST-based Homology Search

Users can submit their own sequences to search for homologous in TOGD. Four databases are available: CDS and gene sequences of 993 orphan genes of *T. aestivum* and CDS and protein sequences of all *T. aestivum* genes. Multiple output formats are supported (*e.g.*, flat, XML, and tabular), and output files can be bulk-downloaded as FASTA or Excel files for further analysis.

Data Download

Details of wheat orphan genes, including gene sequence (FASTA), protein sequence (FASTA), coding sequence (FASTA), chromosome location (TXT), exon number (TXT), molecular weight (TXT), isoelectric point (TXT), protein length (TXT), and other related data, can be downloaded from http://togd.ahau-edu.cn/download-data/.

A Case Application of TOGD

TaFROG, originally identified as a gene responsive to a mycotoxic virulence factor in wheat, is located on chromosome 4A of the wheat genome. Researchers can search the TOGD database to find details of other orphan genes on chromosome 4A available on the website, which facilitates further study of the contribution of these orphan genes in wheat. Moreover, the adopted chromosome-based strategy will support detailed analysis of a region of interest.

Discussion

Orphan genes have been widely identified in variety of species. In order to aid in-depth investigation of the roles of orphan genes in species, the construction of database would provide a platform for questing (Yao *et al*., 2017). As there was no database of orphan gene currently available in *T. aestivum*; therefore, in this study the database of TOGD focusing on the identification and characterization of orphan genes in *T. aestivum* was constructed. Several analysis tools were applied to extract the 993 orphan genes with extensive annotations. This database will provide a centralized platform for functional genomics studies of wheat orphan genes in future. The method of identification of orphan genes in *T. aestivum* was reliable, which can be further applied to investigate the orphan genes of other plants (Yao *et al.*, 2017).

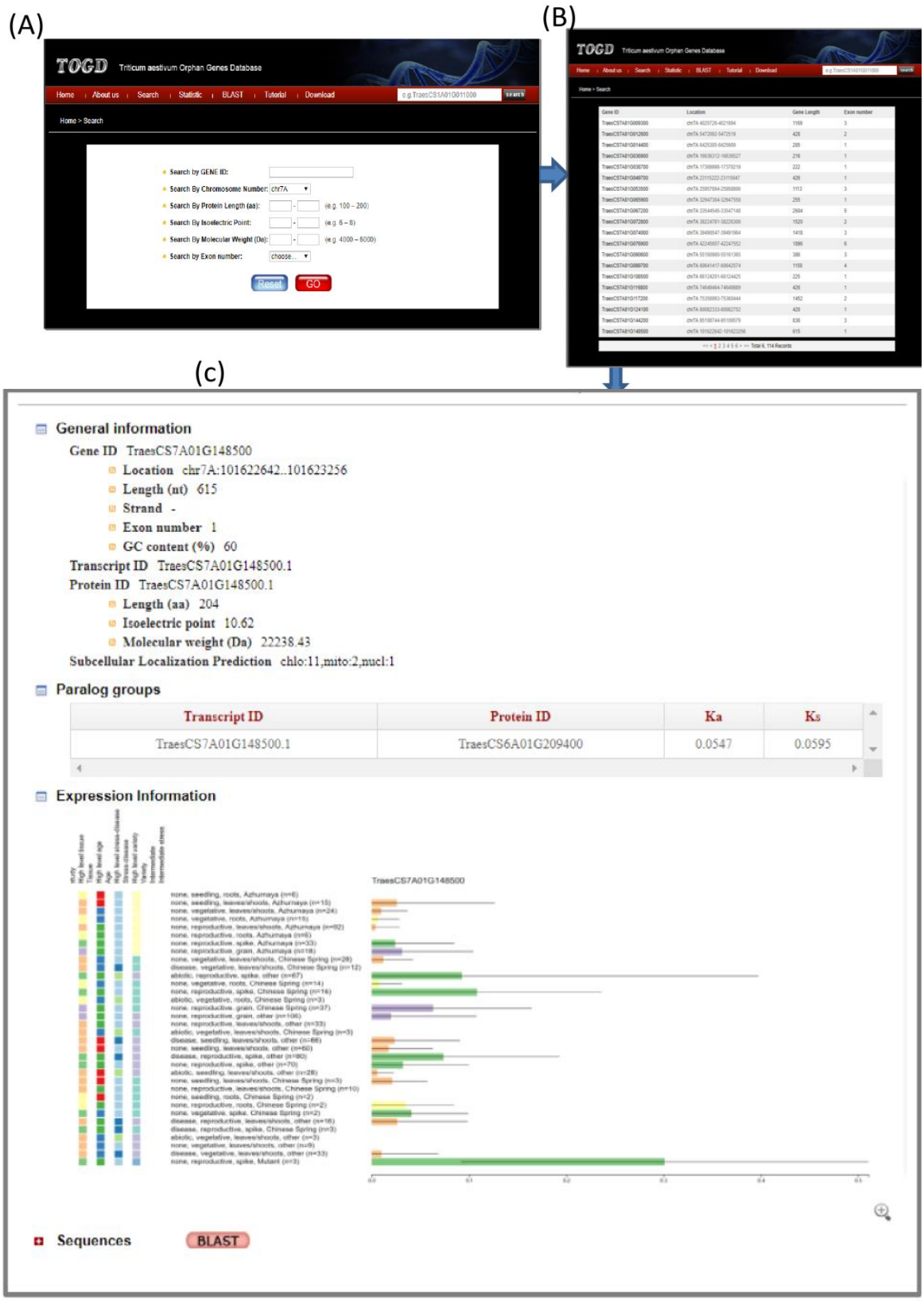

Fig. 5: An overview of the website and gene annotation page in TOGD: (**A**) The Search page, (**B**) Searching list of one kind method by chromosome method and (**C**) An example of the orphan gene annotation page

TOGA provided detailed information for each wheat orphan gene, including gene ID, subcellular location, gene ontology category, protein length, isoelectric point, molecular weight, protein sequence, exon number, gene structure, gene expression as well as chromosome distribution. Besides, the database provides online BLAST to help predict the putative homologous groups of the orphan genes. Associated external databases are accessible *via* the web links provided on the TOGD database platform. Though the current version of TOGD was restricted in collection of *T. aestivum* orphan genes; however, it is still a useful resource for the research community, and particularly an important method for studying about molecular function and evolution of orphan genes. Like the orphan gene of *QQS* in *arabidopsis*, using the same method we identified orphan genes, it has a specific role for plant adaption to low-carbon, low-energy and a noxic stress conditions (Li and Wurtele, 2015).

Overall, TOGD will serve as an important platform for research community to investigate the wheat orphan genes in the future. This database will help researchers to attain some valuable genomic resources, which will assist to explore some valuable trait for wheat breeding. In the near future, the database will be updated according to the latest released genome of *T. aestivum*, and several related species will be added with annotation for further study on orphan genes.

Conclusion

In summary, 993 orphan genes were identified in the wheat genome and incorporated them into a specifically designed TOGD. A flexible search engine and BLAST-based homology search tool were developed to facilitate the extraction and visualization of datasets in TOGD.

Acknowledgements

This study was supported by the Anhui Provincial Natural Science Foundation, China (Grant No.2016-X34).

References

Alaux, M., J. Rogers, T. Letellier, R. Flores, F. Alfama, C. Pommier, N. Mohellibi, S. Durand, E. Kimmel, C. Michotey, C. Guerche, M. Loaec, M. Laine, D. Steinbach, F. Choulet, H. Rimbert, P. Leroy, N. Guilhot, J. Salse, C. Feuillet, International Wheat Genome Sequencing Consortium, E. Paux, K. Eversole, A.F. Adam-Blondon and H. Quesneville, 2018. Linking the international wheat genome sequencing consortium bread wheat reference genome sequence to wheat genetic and phenomic data. *Genomics Biol*., 19: 111

Alexandre, P., J.G. Jia, A. Kahla and C. Arunachalam, 2015. TaFROG Encodes a pooideae orphan protein that interacts with SnRK1 and enhances resistance to the mycotoxigenic fungus *Fusarium graminearum*. *Plant Physiol*., 169: 2895–2906

Arendsee, Z.W., L. Li and E.S. Wurtele, 2014. Coming of age: orphan genes in plants. *Trends Plant Sci*., 19: 698–708

Chen, H., Y. Tang, J. Liu, L. Tan, J. Jiang and M. Wang, 2017. Emergence of a novel chimeric gene underlying grain number in rice. *Genetics*, 205: 993–1002

Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes and J. Fazo, 2012. Phytozome: a comparative platform for green plant genomics. *Nucl. Acids Res.*, 40: 1178–1186

Horton, P., K.J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier and K. Nakai, 2007. WoLF PSORT: protein localization predictor. *Nucl. Acids Res*., 35: 585–587

Li, L. and E.S. Wurtele, 2015. The *QQS* orphan gene of *Arabidopsis* modulates carbon and nitrogen allocation in soybean. *Plant Biotechnol. J*., 13: 177–187

Li, L., C.M. Foster, Q. Gan, D. Nettleton, M.G. James and A.M. Myers, 2009. Identification of the novel protein *QQS* as a component of the starch metabolic network in *Arabidopsis* leaves. *Plant J*., 58: 485–498

Lopes-Caitar, V.S., M.C.D. Carvalho, L.M. Darben, M.K. Kuwahara, A.L. Nepomuceno, W.P. Dias, R.V. Abdelnoor and F.C. Marcelino-Guimaraes, 2013. Genome-wide analysis of the *Hsp20* gene family in soybean: comprehensive sequence, genomic organization and expression profile analysis under abiotic and biotic stresses. *BMC Genom.*, 14: 577–594

Marc, P., F. Devau and C. Jacq, 2001. Ymgv: a database for visualization and data mining of published genome-wide yeast expression data. *Nucl. Acids Res*., 29: 13–19

Neme, R. and D. Tautz, 2013. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. *BMC Genomics*, 14: 117

Ni, F., J. Qi, Q. Hao, B. Lyu, M.C. Luo, Y. Wang, F. Chen, S. Wang, C. Zhang, L. Epstein, X. Zhao, H. Wang, X. Zhang, C. Chen, L. Suan and D. Fu, 2016. Wheat Ms2 Encodes for an orphan protein that confers male sterility in grass species. *Nat. Commun*., 8: 1–12

Palmieri, N., C. Kosiol and C. Schlotterer, 2014. The life cycle of *Drosophila* orphan genes. *Elife*, 3: 1–21

Perochon, A., J. Jianguang, A. Kahla, C. Arunachalam, S.R. Scofield and S. Bowden, 2015. Tafrog encodes a pooideae orphan protein that interacts with snrk1 and enhances resistance to the mycotoxigenic fungus *Fusarium graminearum*. *Plant Physiol*., 169: 2895–2906

Shih, W.I.C., 2016. Ensemble Based Estimators of a Latent Variable: Application in Aging Research. *Doctoral Dissertation*. University of California, Los Angeles, USA

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J.U. Pontius, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner and E. Yaschenko, 2005. Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res*., 33: 39–45

Yao, C., H. Yan, X. Zhang and R. Wang, 2017. A database for orphan genes in Poaceae. *Exp. Ther. Med*., 14: 2917–2924

[Received 09 Apr 2019; Accepted 30 Apr 2019 Published (online) 10 Nov 2019]