

PIECE 2.0: an update for the plant gene structure comparison and evolution database

Yi Wang^{1,2,3}, Ling Xu^{2,3}, Roger Thilmony¹, Frank M. You⁴, Yong Q. Gu^{1,*} and Devin Coleman-Derr^{2,3,*}

¹USDA-ARS, Western Regional Research Center, Crop Improvement and Genetics Research Unit, Albany, CA 94710, USA, ²Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA, ³USDA-ARS, Plant Gene Expression Center, Albany, CA 94710, USA and ⁴Cereal Research Centre, Agriculture and Agri-Food Canada, Morden R6M 1Y5 MB, Canada

Received September 14, 2016; Revised October 04, 2016; Accepted October 12, 2016

ABSTRACT

PIECE (Plant Intron Exon Comparison and Evolution) is a web-accessible database that houses intron and exon information of plant genes. PIECE serves as a resource for biologists interested in comparing intron–exon organization and provides valuable insights into the evolution of gene structure in plant genomes. Recently, we updated PIECE to a new version, PIECE 2.0 (<http://probes.pw.usda.gov/piece> or <http://aegilops.wheat.ucdavis.edu/piece>). PIECE 2.0 contains annotated genes from 49 sequenced plant species as compared to 25 species in the previous version. In the current version, we also added several new features: (i) a new viewer was developed to show phylogenetic trees displayed along with the structure of individual genes; (ii) genes in the phylogenetic tree can now be also grouped according to KOG (The annotation of Eukaryotic Orthologous Groups) and KO (KEGG Orthology) in addition to Pfam domains; (iii) information on intronless genes are now included in the database; (iv) a statistical summary of global gene structure information for each species and its comparison with other species was added; and (v) an improved GSDraw tool was implemented in the web server to enhance the analysis and display of gene structure. The updated PIECE 2.0 database will be a valuable resource for the plant research community for the study of gene structure and evolution.

INTRODUCTION

Eukaryotes possess ‘genes in pieces’ in which the protein-coding exon sequences are interrupted by non-coding intron sequences (1). Most of the genes in eukaryotes contain introns and exons, thus, understanding the organization of

the intron–exon structure is important because such information reveals conserved or diverged structures of genes from different species (orthologs), and/or of the different family members (paralogs), providing insights into the process of gene evolution. Recent advances in sequencing technologies have led to an unprecedented progress in generating genome sequence data and opened a new era for comparative genomics studies (2,3). These data sets allow researchers to address many fundamental evolutionary questions at a genome-wide comparative scale. In the comparative analyses of gene structure, gene sequences from different plant genomes are often grouped by gene family members and/or ortholog clusters. Using the phylogenetic analysis tool, along with the prediction of gene structure, one can identify intron–exon patterns and intron gain or loss events in the grouped gene sequences. The reconstruction of intron gain/loss events during the evolutionary history of a gene provides valuable information for clarifying the evolutionary relationships within large gene families and facilitate a deeper understanding of the possible functional implications, including the generation or disruption of lineage-specific alternative splicing events (4).

In order to study gene structure evolution in species with sequenced genomes, user-friendly and publicly available resources are necessary. PIECE (Plant Intron Exon Comparison and Evolution) is an intron–exon database that provides a powerful platform to compare gene structure among plant species (5). It was released in 2012 and published in the 2013 Nucleic Acids Research database issue. During the past 4 years, the genomic sequence data for plant species have undergone substantial expansion. The increasing number of genes from more sequenced plant genomes has greatly enriched the gene intron–exon database, but requires the phylogenetic analysis at a much larger scale for accurate dissection of the evolution of plant intron–exon organization. Comparative analysis of intron–exon architecture is important for understanding the rules governing

*To whom correspondence should be addressed. Tel: +1 510 559 5732; Fax: +1 510 559 5818; Email: yong.gu@ars.usda.gov
Correspondence may also be addressed to Devin Coleman-Derr. Tel: +1 510 559 5911; Fax: +1 510 559 5678; Email: devin.coleman-derr@ars.usda.gov

gene structure organization, protein functionality and evolutionary changes among plant species. Here, we present a new version of PIECE (PIECE 2.0, <http://probes.pw.usda.gov/piece/> or <http://aegilops.wheat.ucdavis.edu/piece/>). In this updated version, we made significant modifications and improvements to the original database by increasing more genome data, enhancing web display, adding new useful features. The updated version contains 2 089 560 protein coding genes from 49 plant species; more than twice the number in the previous version (25 species). In order to view gene structure data for large gene families between multiple species, we developed a new interactive viewer that provides an easy method for displaying and analyzing intron–exon gene structures arranged via a phylogenetic analysis. Several new features have been integrated into the current version, including displaying gene structures according to KOG (The annotation of Eukaryotic Orthologous Groups) (6) and KO (KEGG Orthology) (7) information, intronless gene database. Also a global gene structure summary for each species is available, allowing species level comparisons to be made. Finally, we also updated the GSDraw tool, which can now more conveniently produce customizable, high-quality gene structure images including a phylogenetic tree using input files from users.

NEW FEATURE

Data update

In PIECE 2.0, we have updated the gene structure data from more sequenced plant species. The raw genome data sets were downloaded from Phytozome 11.0 (8) and gene structure data sets were refined by our in-house pipeline (Supplementary Figure S1). To build the PIECE database, we parsed and analyzed five types of files from the raw data to extract intron–exon information, including genomic sequence files, transcript sequence files, protein sequence files, gene annotation files and general feature format (GFF) files containing coordinate data of the gene structure in genomic sequences. Now PIECE 2 089 560 protein coding genes from 49 plant species, covering major lineages of the plant kingdom (Supplementary Table S1).

Search system improvement

PIECE 2.0 has a user-friendly system for finding target genes. Currently, the search page allows users to make queries with six types of keywords: (i) gene symbol; (ii) locus name; (iii) gene description; (iv) Pfam (9) ID; (v) KOG ID; and (vi) KO ID. For each type, there is an example link in the text box (#1 to #6) on the search page. The example link helps the user easily learn how to perform the search in PIECE. Moreover, users can choose the plant species from an interactive taxonomy tree, and the selected species are shown on the right of the search page. The taxonomy tree was generated by phyloT (<http://phylo.t.biobyte.de>) based on NCBI taxonomy (10) and manually compared with the tree on Phytozome BLAST page (<https://phytozome.jgi.doe.gov/pz/portal.html#!search?show=BLAST>). Users also can set whether the alternative splice isoforms and intronless genes are displayed on the search result page.

The main page of the search results lists all the genes meeting the search criteria and provides brief information, such as index number, species name, Pfam IDs, KOG ID, KO ID, exon number, symbol, annotation and ortholog analysis options (Figure 1). Compared with PIECE 1.0, there are two added features in the search results with PIECE 2.0. One is the exon number, which helps users quickly learn any intron and exon number variations. The other is the KOG and KO ID information for the genes. PIECE 2.0 can not only show phylogenetic trees displayed along with gene structure information by clicking on a Pfam domain, but also reconstruct the phylogenetic trees using the genes searched with KOG or KO category.

New gene structure viewer

In PIECE 1.0, we developed a graphical viewer that produces high-quality images to display gene structure, Pfam domain and phylogenetic tree together. The elements of gene structure including exon, intron and Pfam domain are interactive that enables the user to select, click and flexibly modify colors. In PIECE 2.0, the new viewer provides users a friendly way to display and analyze intron–exon data within the phylogenetic tree (Figure 2) through HTML5 canvas visualization using the open source Javascript library named Phylo.io (11). Phylo.io is a web framework to visualize and compare phylogenetic trees side-by-side. The original Plylo.io only displays the trees. Therefore, we modified the source code so that it will also show the gene structure pattern diagram linked to the resulting bootstrapped dendrogram for each gene category. Phylo.io can improve the legibility of large trees by estimating an optimal collapsing depth using the number of leaves and the size of viewing area (11). Users can swap, collapse and expand the tree by clicking the nodes or reroot the tree by clicking the tree branches. The coordinates of gene structure elements in the diagram will be changed accordingly for each gene when the tree is adjusted. Users can also easily search for individual genes in the tree, and the gene position in the tree will be highlighted in red along with the path to that leaf from the root of the tree. By default, the highlighted gene is the one selected by the user in the search result page. The highlighted item can be removed if the user clicks the search button without any input. Unlike the previous version, users can now change the gene structure display in the viewer by directly selecting display type in the left panel. There are five types available (i) genomic sequences; (ii) aligned by the CDS start points; (iii) genomic sequences without UTR; (iv) protein sequences, and (v) aligned protein sequences. The new viewer also has several options to allow users to change the attributes of the diagram including setting color of and hiding elements of gene structure, and zoom in and zoom out function. Moreover, the user can directly download the diagram as high-quality SVG format file. As a demonstration, Figure 2 shows the gene structure of Lipoxxygenase proteins (Pfam domain PF00305) in *Arabidopsis thaliana* and *Populus trichocarpa* using the new viewer. The phylogenetic distribution suggests that *LOX* genes could be grouped into two major subclasses, Class I and Class II (Figure 2C). This result is consistent with a recent study of genome-wide analysis of *LOX* genes in Poplar (12). No sister gene

A			B				C			
Index	ID	Species	Pfam	KOG	KO	Exons	Symbol	Annotation	Ortholog	
1	AT1G73370.1	<i>Arabidopsis thaliana</i>	PF00862 PF00534 PF13692	N/A	K00695	12	ATSUS6,SUS6	sucrose synthase, putative, expressed	Show ▾	
2	Bradi1g29570.2	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	K00695	18	ATSUS6,SUS6	sucrose synthase 6	GLOOM ▾	
3	Bradi1g29570.1	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	K00695	18	ATSUS6,SUS6	sucrose synthase 6	Exalign ▾	
4	Bradi1g29570.3	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	K00695	17	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
5	Bradi1g29570.5	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	N/A	15	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
6	Bradi1g29570.4	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	N/A	17	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
7	Bradi3g60687.2	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	K00695	16	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
8	Bradi3g60687.3	<i>Brachypodium distachyon</i>	PF00862 PF00534 PF13692	KOG0853	K00695	15	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
9	Solyc02g081300.2.1	<i>Solanum lycopersicum</i>	PF00862 PF00534 PF13692	KOG0853	K00695	14	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
10	Solyc03g098290.2.1	<i>Solanum lycopersicum</i>	PF00862 PF00534 PF13692	KOG0853	K00695	15	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
11	LOC_Os02g58480.1	<i>Oryza sativa</i>	PF00862 PF00534 PF13692	KOG0853	K00695	13	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
12	LOC_Os04g17650.1	<i>Oryza sativa</i>	PF00862 PF00534 PF13692	KOG0853	N/A	16	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
13	LOC_Os04g24430.1	<i>Oryza sativa</i>	PF00862 PF00534 PF13692	KOG0853	K00695	17	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
14	Potri.002G195500.1	<i>Populus trichocarpa</i>	PF00534	N/A	N/A	2	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
15	Potri.004G081300.1	<i>Populus trichocarpa</i>	PF00862 PF00534 PF13692	KOG0853	K00695	12	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
16	Potri.004G081300.2	<i>Populus trichocarpa</i>	PF00862 PF00534 PF13692	KOG0853	K00695	13	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
17	Potri.012G037200.2	<i>Populus trichocarpa</i>	PF00862 PF00534 PF13692	KOG0853	K00695	13	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
18	Potri.012G037200.1	<i>Populus trichocarpa</i>	PF00862 PF00534 PF13692	KOG0853	K00695	16	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
19	Potri.015G029100.2	<i>Populus trichocarpa</i>	PF00862 PF00534 PF13692	KOG0853	N/A	15	ATSUS6,SUS6	sucrose synthase 6	----- ▾	
20	Potri.015G029100.3	<i>Populus trichocarpa</i>	PF00862 PF00534 PF13692	KOG0853	N/A	13	ATSUS6,SUS6	sucrose synthase 6	----- ▾	



Figure 1. An example of PIECE (Plant Intron Exon Comparison and Evolution) 2.0 search results. (A) Gene ID column with link to sequence detail. (B) Pfam, KOG and KO information with link to gene structure view. (C) Ortholog analysis options include showing gene structure, GLOOM analysis and Exalign analysis.

pairs were found in the tree between *Arabidopsis thaliana* and *Populus trichocarpa*, suggesting that multiple gene duplications occurred in the course of *LOX* gene evolution in Poplar. Moreover, most *LOX* genes in the viewer contain two protein domains (Pfam domain PF01477 and Pfam domain PF00305) except Potir.001G227100.1. The number of exons in these duplicated *LOX* genes ranged from 1 to 9, indicating a complex distribution pattern of intron-exon structure in the *LOX* genes.

Intronless gene

Although many eukaryote genes carry introns, a significant portion of eukaryote genes lack introns. Intronless genes have gained increasing attention because of their implication in understanding evolutionary patterns of related genes and genomes (13). Currently, there are only two specialized database named PIGD (13) and IGDD (14) for plant intronless genes. PIGD only contains five species from the Poaceae family and IGDD only provides the data set of intronless genes from five eudicot species. In PIECE 2.0, we created a searchable database containing 353 515 predicted intronless genes from the 49 plant species (Supplementary Table S1). We annotated these intronless genes using the an-

notation information in the Phytozome database. PIECE 2.0 provides information of each intronless gene sequence, its genomic location, Pfam, isoelectric point (PI), molecular weight (MW), KOG and GO term as well as subcellular localization predicted by WoLF PSORT (15). Such information will be displayed in the result page when a search of an intronless gene is made. PIECE 2.0 also provides a function to compare intronless with intron-containing genes from the same gene family or with the orthologous genes. For example, AT1G15700.1, which encodes the gamma subunit of chloroplast ATP synthase, is intronless and contains a protein domain (Pfam domain PF00231). Supplementary Figure S2A shows that based on the search using the PF00231 domain, two related *Arabidopsis* genes are also intronless, while one related gene has 7 introns. Interestingly, most orthologs of AT1G15700.1 are intronless (Supplementary Figure S2B), but the genes from green algae all have introns. This observation could suggest that most of the introns were lost in the evolution of aquatic plants (green algae) to land plants.

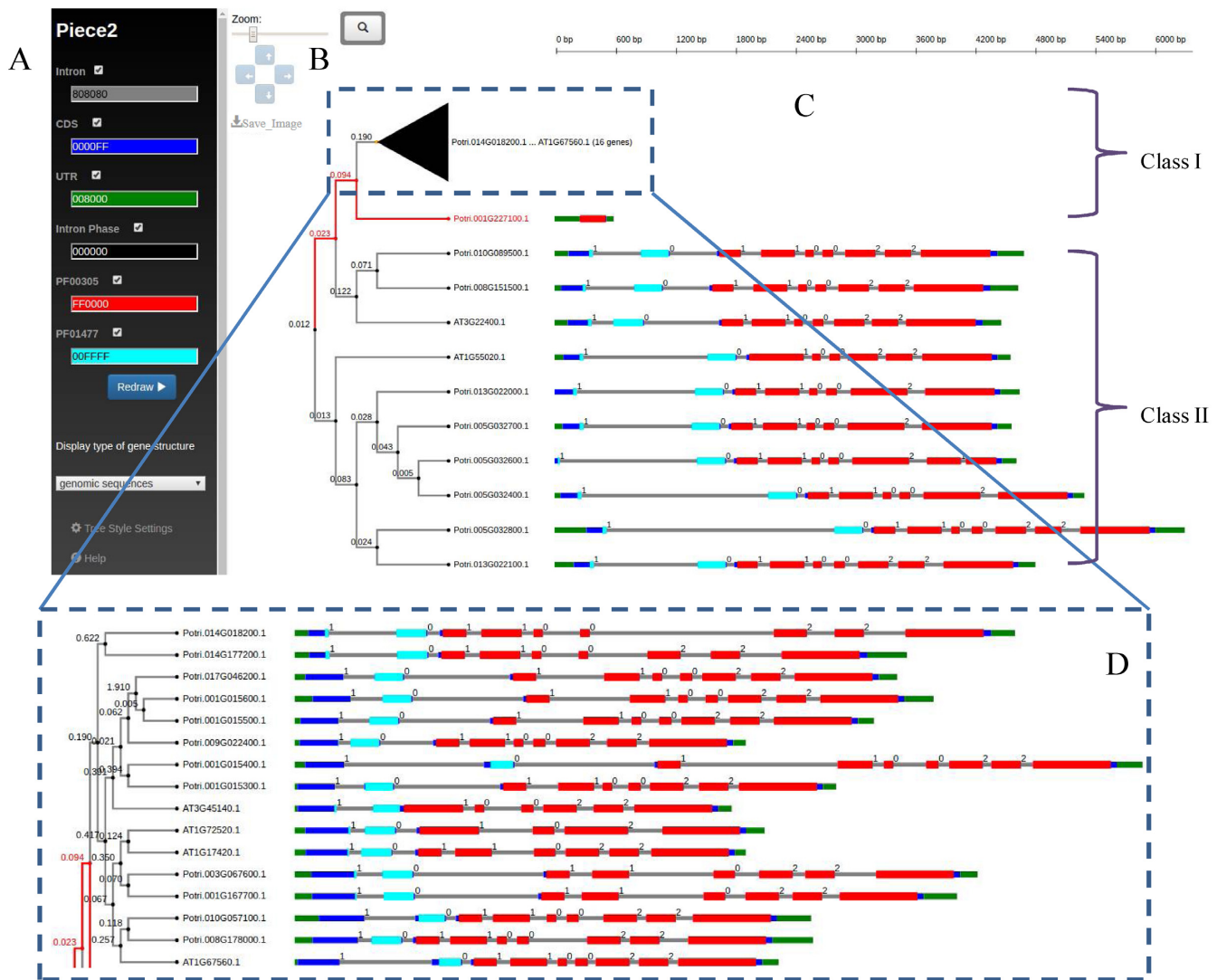


Figure 2. The new PIECE viewer showing the gene structures of *LOX* proteins (Pfam domain PF00305) in *Arabidopsis thaliana* and *Populus trichocarpa*. (A) Control panel for user selection of parameters in phylogenetic tree reconstruction and gene structure display. (B) Operation panel for search ID, manipulating the tree and download the picture. (C) Gene structure pattern diagram linked to the resulting bootstrapped dendrogram. (D) A example showing the expanding of a subgroup from the phylogenetic tree.

Statistics of gene structure characterization

In PIECE 2.0, clicking on 'Browse' from the home page will provide the statistical summary of the global gene structure information for each species in the current database, including total number, total length, median length, average length of genes as shown in a table format. A box plot of gene, exon, intron, 5' UTR and 3' UTR are also displayed in the page, allowing the user easily compare the median, maximum and minimum length of each element (Supplementary Figure S3). PIECE 2.0 also provides the utility to compare the attribute of gene structure among plant species using the 'Compare' page. The result is shown in a dot plot, which has zoom in and zoom out feature by clicking and dragging. Users can choose an element of a gene structure for x-axis or y-axis to redraw the plot. Figure 3A is a dot plot showing the comparison of average intron length (x-axis) with average exon length (y-axis) among plant species.

The result shows that the ratio of the average intron length and average exon length is in the range of 1.5–3.5 for most plant species, meaning that they have a similar average intron and exon length ratio (Figure 3B). Some exceptions include *Ostreococcus lucimarinus*, *Micromonas sp. RCC299* and *Micromonas pusilla*, where the intron average length is shorter than the exon average length (ratio < 1). In addition, the average intron length of *Vitis vinifera* and *Amborella trichopoda* is much longer than the exon average length (ratio > 5) (Figure 3A).

GSDraw update

The comparison and visualization of gene structure for gene family or homologous genes offers an important method for biologists to analyze gene evolution. A number of web server applications for gene structure analysis have been developed, such as FancyGene (16), GSDraw (5), Gene-

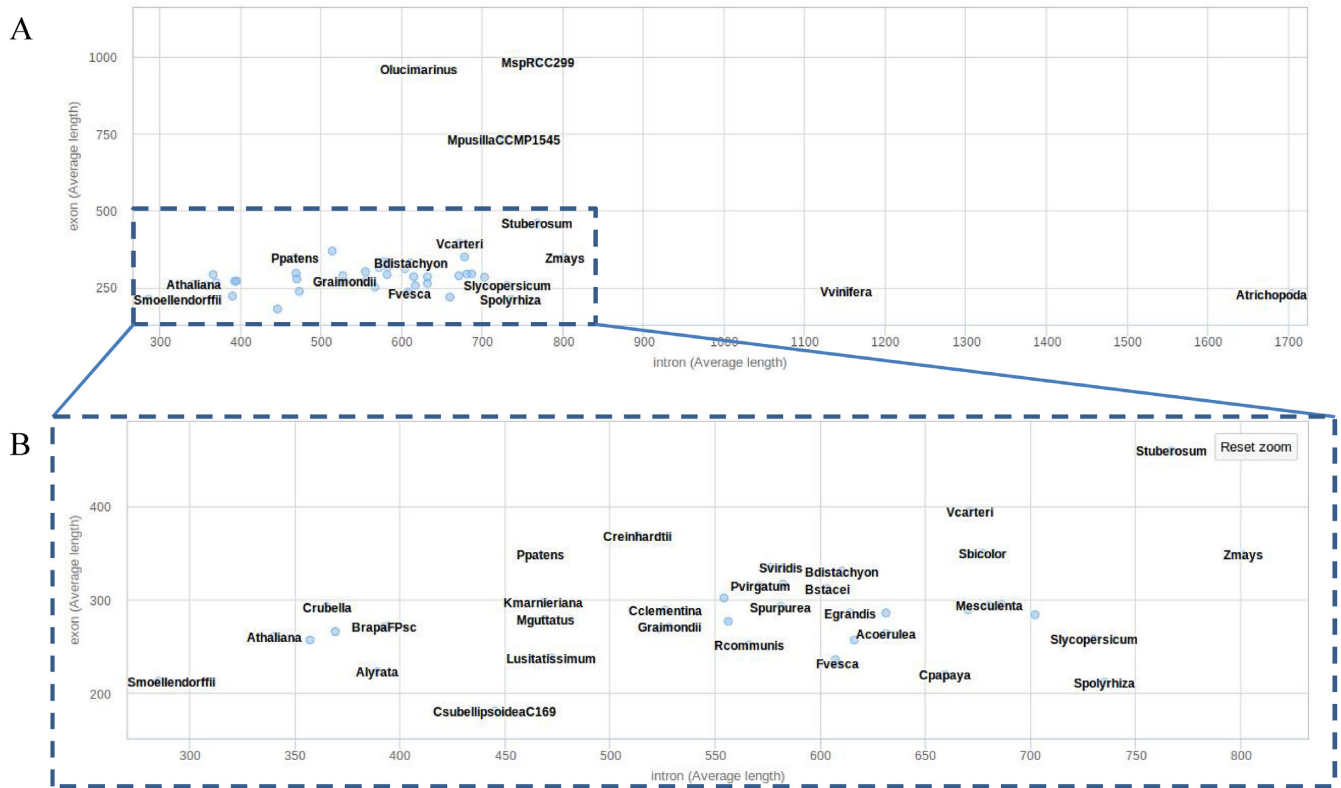


Figure 3. Dot plot for intron and exon average length comparison in the 49 species in PIECE 2.0. (A) Display of all 49 species. (B) A zoom-in picture of the analysis result of selected species.

Painter (17) and GSDS (18). Among them, GSDraw is the one that can use sequence data (genomic, CDS and transcript sequences) to extract the gene structure, protein motif, and phylogenetic information, and then process the information together to draw the display diagram. In PIECE 2.0, we integrated GSDraw with the new gene structure viewer discussed above to provide an efficient and interactive graphical interface for analyzing gene structure. Users can set their own parameters in GSDraw for protein motif identification and phylogenetic tree reconstruction. GSDraw can now also draw a gene structure diagram by uploading a GFF file, a feature that was added based on the feedback from PIECE 1.0 users.

DISCUSSION

A feature of eukaryotic gene structure is that the genomic sequences of protein-coding genes are frequently interrupted by introns, its generation are relics of primordial genes (19–21). The evolutionary dynamics of intron–exon organization can be used to infer the evolutionary history of a gene family. One of the principle methods of performing this type of analysis is to reconstruct a phylogenetic tree that best represents the evolutionary histories of gene families. In such analysis, sequence alignments are used to create phylogenetic trees and then this information is combined with the intron–exon organization of each member in the gene family. In plants, very few databases provide comparative information on gene structure evolution. Phy-

tozome (8) and PLAZA (22) are well known plant genome database where intron–exon sequences of plant genes can be obtained. However, these resources lack comparative analytical capabilities to investigate intron–exon structural evolution through comparing gene families and orthologs from different species (5). A recent released database, named JuncDB (23), is a database for comparing exonic architectures. However, it only uses transcripts of orthologous genes in data process and does not focus on plant species. Unlike other genomic databases, PIECE is a plant intron–exon database that allows comparative analysis at a genome-wide level. Users can easily integrate, visualize and analyze phylogenetic trees, intron–exon structure, protein domains and intron phase. PIECE has been widely used by plant biologists for functional and evolutionary studies of gene families and orthologs (24–29). Here, we reported the new PIECE 2.0 version with updated gene structure data sets by including 24 newly sequenced genomes since the release of PIECE 1.0 and with a significant improvement of the view function to display phylogenetic trees along with gene structure and protein domains. In addition, PIECE 2.0 can now compare gene structures with four categories, including Pfam, KOG, KO and ortholog. This can help users better understand the evolution of gene structure. PIECE 2.0 also provides a web resource for the collection and analysis of plant intronless genes. In conclusion, it is anticipated that PIECE 2.0 will prove a useful resource for addressing important biology questions regarding the organization and evolution of gene structure in the evolutionary history of plant species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

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

FUNDING

United State Department of Agriculture, Agriculture Research Service CRIS projects [5325-21000-021D and 2030-21430-00D]. Funding for open access charge: Agriculture Research Service CRIS projects [5325-21000-021D and 2030-21430-00D].

Conflict of interest statement. None declared.

REFERENCES

- Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.
- Chi, K.R. (2008) The year of sequencing. *Nat. Methods*, **5**, 11–14.
- Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
- Coulombe-Huntington, J. and Majewski, J. (2007) Characterization of intron loss events in mammals. *Genome Res.*, **17**, 23–32.
- Wang, Y., You, F.M., Lazo, G.R., Luo, M.C., Thilmoney, R., Gordon, S., Kianian, S.F. and Gu, Y.Q. (2013) PIECE: A database for plant gene structure comparison and evolution. *Nucleic Acids Res.*, **41**, D1159–D1166.
- Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
- Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. *et al.* (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic acids Res.*, **40**, D1178–D1186.
- Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. *et al.* (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
- Federhen, S. (2015) Type material in the NCBI Taxonomy Database. *Nucleic Acids Res.*, **43**, D1086–D1098.
- Robinson, O., Dylus, D. and Dessimoz, C. (2016) Phylo.io: Interactive viewing and comparison of large phylogenetic trees on the web. *Mol. Biol. Evol.*, **33**, 2163–2166.
- Chen, Z., Chen, X., Yan, H., Li, W., Li, Y., Cai, R. and Xiang, Y. (2015) The Lipoxygenase Gene Family in Poplar: Identification, classification, and expression in response to MeJA treatment. *PLoS One*, **10**, e0125526.
- Yan, H., Jiang, C., Li, X., Sheng, L., Dong, Q., Peng, X., Li, Q., Zhao, Y., Jiang, H. and Cheng, B. (2014) PIGD: A database for intronless genes in the Poaceae. *BMC Genomics*, **15**, 832.
- Yan, H., Dai, X., Feng, K., Ma, Q. and Yin, T. (2016) IGDD: A database of intronless genes in dicots. *BMC Bioinformatics*, **17**, 289.
- Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K. (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
- Rambaldi, D. and Ciccarelli, F.D. (2009) FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics*, **25**, 2281–2282.
- Muhlhausen, S., Hellkamp, M. and Kollmar, M. (2015) GenePainter v. 2.0 resolves the taxonomic distribution of intron positions. *Bioinformatics*, **31**, 1302–1304.
- Hu, B., Jin, J., Guo, A.Y., Zhang, H., Luo, J. and Gao, G. (2015) GSDD 2.0: an upgraded gene feature visualization server. *Bioinformatics*, **31**, 1296–1297.
- Rogozin, I.B., Sverdlov, A.V., Babenko, V.N. and Koonin, E.V. (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief. Bioinform.*, **6**, 118–134.
- de Roos, A.D. (2007) Conserved intron positions in ancient protein modules. *Biol. Direct*, **2**, 7.
- Csuros, M., Holey, J.A. and Rogozin, I.B. (2007) In search of lost introns. *Bioinformatics*, **23**, i87–i96.
- Proost, S., Van Bel, M., Vanechoutte, D., Van de Peer, Y., Inze, D., Mueller-Roeber, B. and Vandepoel, K. (2015) PLAZA 3.0: An access point for plant comparative genomics. *Nucleic Acids Res.*, **43**, D974–D981.
- Chorev, M., Guy, L. and Carmel, L. (2016) JuncDB: An exon-exon junction database. *Nucleic Acids Res.*, **44**, D101–D109.
- Akkuratov, E.E., Walters, L., Saha-Mandal, A., Khandekar, S., Crawford, E., Zirbel, C.L., Leisner, S., Prakash, A., Fedorova, L. and Fedorov, A. (2014) Bioinformatics analysis of plant orthologous introns: identification of an intronic tRNA-like sequence. *Gene*, **548**, 81–90.
- Qiao, X., Li, M., Li, L., Yin, H., Wu, J. and Zhang, S. (2015) Genome-wide identification and comparative analysis of the heat shock transcription factor family in Chinese white pear (*Pyrus bretschneideri*) and five other Rosaceae species. *BMC Plant Biol.*, **15**, 12.
- Hawkins, L.K., Mylroie, J.E., Oliveira, D.A., Smith, J.S., Ozkan, S., Windham, G.L., Williams, W.P. and Warburton, M.L. (2015) Characterization of the maize chitinase genes and their effect on *Aspergillus flavus* and Aflatoxin Accumulation resistance. *PLoS One*, **10**, e0126185.
- Ding, M., Chen, J., Jiang, Y., Lin, L., Cao, Y., Wang, M., Zhang, Y., Rong, J. and Ye, W. (2015) Genome-wide investigation and transcriptome analysis of the WRKY gene family in *Gossypium*. *Mol. Genet. Genomics*, **290**, 151–171.
- Kaur, S., Dhugga, K.S., Gill, K. and Singh, J. (2016) Novel structural and functional motifs in cellulose synthase (CesA) genes of bread wheat (*Triticum aestivum*, L.). *PLoS One*, **11**, e0147046.
- Korbes, A.P., Kulcheski, F.R., Margis, R., Margis-Pinheiro, M. and Turchetto-Zolet, A.C. (2016) Molecular evolution of the lysophosphatidic acid acyltransferase (LPAAT) gene family. *Mol. Phylogenet. Evol.*, **96**, 55–69.