Published online 24 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D1159–D1166***
doi:10.1093/nar/gks1109

PIECE: a database for plant gene structure comparison and evolution

Yi Wang[1,2], Frank M. You[3], Gerard R. Lazo[1], Ming-Cheng Luo[2], Roger Thilmony[1], Sean Gordon[1], Shahryar F. Kianian[4] and Yong Q. Gu[1,*]

[1]USDA-Agriculture Research Service, Western Regional Research Center, Albany, CA 94710, [2]Department of Plant Sciences, University of California, Davis, CA 95616, USA, [3]Cereal Research Center, Agriculture and Agri-Food Canada, Winnipeg, Manitoba R3T 2M9, Canada and [4]Department of Plant Sciences, North Dakota State University, Fargo, ND 58105, USA

Received August 15, 2012; Revised October 18, 2012; Accepted October 19, 2012

ABSTRACT

Gene families often show degrees of differences in terms of exon–intron structures depending on their distinct evolutionary histories. Comparative analysis of gene structures is important for understanding their evolutionary and functional relationships within plant species. Here, we present a comparative genomics database named PIECE (http://wheat.pw.usda.gov/piece) for Plant Intron and Exon Comparison and Evolution studies. The database contains all the annotated genes extracted from 25 sequenced plant genomes. These genes were classified based on Pfam motifs. Phylogenetic trees were pre-constructed for each gene category. PIECE provides a user-friendly interface for different types of searches and a graphical viewer for displaying a gene structure pattern diagram linked to the resulting bootstrapped dendrogram for each gene family. The gene structure evolution of orthologous gene groups was determined using the GLOOME, Exalign and GECA software programs that can be accessed within the database. PIECE also provides a web server version of the software, GSDraw, for drawing schematic diagrams of gene structures. PIECE is a powerful tool for comparing gene sequences and provides valuable insights into the evolution of gene structure in plant genomes.

INTRODUCTION

In eukaryotes, a typical gene structure contains two elements: the exon and the intron. Exons are the DNA sequences that are transcribed and represented in the mature forms of RNA (mRNAs) that serve as template for synthesis of the encoded proteins. Introns that interrupt the exons in gene sequences are also transcribed, but they are removed from the mature RNA transcript by RNA splicing. Comparative analysis of exon–intron organization is important for understanding rules of gene structure and organization, protein functionality and evolutionary changes among species. The structural information of genes and gene families can serve as material for phylogenetic analyses to understand the gain, loss and change of gene structures (1–3), thereby elucidating mechanisms underlying the molecular evolution of genes and genomes (4–6). The increasing availability of plant genome sequences now makes it possible to conduct phylogenetic analyses of genes or gene families from a large number of plant species representing a large evolutionary distance. Typically, phylogenetic analyses of genes of interest require, first, the extraction of genes with corresponding intron and exon structure information, followed by phylogenetic analyses using available software programs. Comparing gene sequences to identify evolutionarily conserved gene structures is useful for predicting the biological function of protein-coding genes of interest. Accordingly, some plant comparative genomic databases, such as PlantGDB (7), PLAZA (8) and Phytozome (9), are well known and widely used because these databases allow users to extract gene structure data including exon–intron positions, exon and intron lengths and alternative splicing. Usually, users will still need to perform further analyses on the extracted data with available software programs to gain insight regarding the evolution and function of gene structure. Databases dealing with gene structure analyses are available, but with a primary emphasis on non-plant species. CIWOG is a plant database that displays common introns within orthologs in eight plant species (10). Furthermore, in most cases, phylogenetic trees, gene structures, protein domains and exon–intron comparisons for orthologs have not yet been integrated together and therefore the related databases do not provide a comprehensive view pertinent to evolution and function of gene structure.

*To whom correspondence should be addressed. Tel: +1 510 559 5732; Fax: +1 510 559 5818; Email: Yong.Gu@ars.usda.gov

For instance, these databases do not contain information regarding which Pfam domain in a gene family contains conserved intron sites and phases. The location of introns with exons occurs as one of three different phases; between two codons (phase 0), between the first and second nucleotides of a codon (phase 1) or between the second and third nucleotide of a codon (phase 2). Intron phases are a conservative character of eukaryotic gene structures because any phase change requires either compensatory double mutations or a more complex molecular mechanism. Therefore, the location of the introns within the same sites and phase of related genes is strong support for an evolutionary relationship. Meanwhile, it is important to understand that the evolution of gene structure is often associated with the evolutionary history and functional domains of a gene of interest.

Here, we report the development of PIECE (http://wheat.pw.usda.gov/piece), a comprehensive plant gene comparison and evolution database containing all the annotated genes described from 25 plant species with available sequenced genomes. The database includes data for 17 eudicots, 5 monocots, 2 green algae and the moss *Physcomitrella patens* (Supplementary Table S1). The annotated genes were extracted from each species and classified based on their Pfam motif (11). Phylogenetic trees were pre-constructed for each gene category by integrating exon–intron and protein motif information. The intron site data can be shown not only in the genomic sequence but also in protein alignment sequences. The sequence and gene structure information for each identified gene is available for online access within the PIECE website. The database contains orthologs in those species for comparative analysis and evolutionary studies of gene structure. Several gene structure analysis software tools including GLOOME (12), Exalign (13) and GECA (14) have been integrated into PIECE and can be executed for each orthologous group to display exon–intron gain, loss and conservation. PIECE also provides a web interface package, GSDraw (Gene Structure Draw Server), for drawing schematic diagrams of the structures of genes derived from other species in addition to the 25 sequenced plant species. Users can submit genomic coding DNA sequence (CDS) and transcript sequences. GSDraw uses this information to obtain the gene structure, protein motif and phylogenetic tree and outputs the results as diagrams. PIECE can provide valuable information for plant researchers for analyzing the evolution of gene structure and for elucidating the biological function of proteins. PIECE is a useful resource for the research community, particularly for the study of exon–intron evolution.

DATABASE CONSTRUCTION

Data collection

PIECE currently contains a total of 947 630 annotated genes from 25 sequenced plant species including low plants to Angiosperms (Supplementary Table S1). Genome sequences, transcript sequences, protein sequences and annotation GFF files were downloaded from Phytozome (9). Exon–intron site, length and intron phase data were extracted based on the genome annotation GFF files using an in-house Java program.

Plant gene family classification

Plant genes were grouped into different families based on their protein domains using the Pfam database (v26.0) (11). We applied the hmm search program in the HMMER package (15) to search against the protein sequences of each species to classify genes. An *E*-value 0.01 as a cutoff, which has been widely adopted for HMMER searches, was used for queries. Many genes have more than one Pfam domain. For example, the B3 domain (PF02362) is present in either the ABI3-VP1 family or the RAV subfamily of the AP2 family. In this case, we therefore assigned PF02362 as a gene family entry that includes genes in the ABI3-VP1 and AP2 families.

Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignment (MSA) was performed using the MUSCLE v3.831 program (16). The default parameters were used if the number of members in a gene family was ≤500, otherwise the option '*–maxiters* 2' was applied. For phylogenetic analyses, the FastTree v2.1.4 program was used (17), which implements a fast and accurate approximate maximum likelihood method. FastTree analyses were conducted with default parameters; specifically, the amino acid substitution matrix used was JTT, the number of rate categories of sites (CAT model) was 20, the local support values of each node were computed by re-sampling the site likelihoods 1000 times.

Putative ortholog annotation

To predict putative orthologous relationships of genes among these plant species, we used the BLAST score ratio (BSR) method, which has been widely adopted by ENSEMBL and other studies (18). An all-against-all BLASTP search with a strict cutoff *E*-value <1e−20 was performed, and the BSR value was calculated for each hit. After comparing results at different BSR values, we chose a BSR value ≥0.4 as the cutoff and retrieved the top sequences in the species with the largest BSR values as the putative ortholog(s).

Orthologous gene structure evolution

PIECE uses GLOOME, Exalign and GECA for the orthologous gene structure evolution analysis. GLOOME can analyze the presence and absence profiles (phyletic patterns), which are widely used in biology (12). The default parameter settings were used for GLOOME analyses. Because the required input is a phyletic pattern provided as a 0/1 MSA, we first used MUSCLE to obtain the alignment of orthologous protein sequences with default parameters, and then calculated all intron sites. For each gene, if it has an intron site in the aligned consensus sequence, we marked '1' for the site, if not, we marked '0' for the site. To obtain orthologous gene exon–intron gain and loss information, a Java pipeline was implemented to include the steps described above, i.e. gene family classification, ortholog annotation and orthologous gene structure analysis.

We performed alignments between intron/exon structures using the Exalign algorithm (13). The algorithm was run in global alignment mode, and allowed intron gain/loss detection to exclude false assignments because of intron gain/loss events between orthologs. We created the Exalign dataset for each plant species in our database, and compared the full set of plant gene structures. The length of partially coding exons was adjusted to include only the coding portions. Fully non-coding exons were excluded from the comparison.

Recently, a new tool named GECA was developed, which displays gene exon–intron organization by highlighting changes in gene structure among members of a gene family (14). In PIECE, orthologs can also be displayed using the GECA method with default settings.

UTILITY AND WEB INTERFACE

PIECE is a web-based tool combining a MySQL database management system with a dynamic web interface based on PHP and Javascript. The exon–intron data in the database are searchable and viewable.

Search system

PIECE has a user-friendly entry point for searching each gene. Users can retrieve any gene by a keyword search for gene ID, gene name or gene function or by a BLAST search using either the nucleotide or protein sequence of your gene of interest (Figure 1A). The main page of the search results lists all genes meeting the search criteria and provides brief information, such as gene accession, gene description, source organism, gene annotation, Pfam domain and gene ortholog analysis (Figure 1B). The columns in Gene ID, Pfam ID and Ortholog Gene Structures are linked to more detailed information of the analysis results. For example, clicking on a gene accession will display details on the exon–intron information for each gene, including the genomic and transcript sequences (Figure 1C). Each gene in the search result contains the location of the Pfam domain that was identified in its protein sequence. Clicking on a Pfam domain will show its phylogenetic tree along with gene structures (Figure 1D). The ortholog analysis link will display the GLOOME, Exalign and GECA results of the gene (Figure 1E and F), which link to details that include more elaborate descriptions of the orthologous gene structure evolution results. The detail of the gene structure and evolution analysis in the Pfam ID and Ortholog Gene Structures columns are also presented (see below).

Phylogenetic tree display along with gene structure and Pfam domains

The database provides a user-friendly graphical view that displays SVG-formatted output, which contains a gene structure and Pfam domain pattern diagram linked to a bootstrapped similarity dendrogram (Figure 2). Depending on the annotations present in the database, the viewer can automatically recognize elements of the gene structure, such as coding exons, introns and UTRs. Default conventions are used to render exons (thick boxes), UTRs (thin blue and green boxes) and introns (thin grey boxes), but the user can modify the display of the elements by selecting a different color or choosing to not display the element. A search function is provided to allow users to search the gene ID in he phylogenetic tree. If the ID is found, it will be highlighted in red. Controls available on the bottom of the page allow magnification of tree image (e.g. the ‘zoom in’ and ‘zoom out’ buttons) as well as movement of the magnified image with the arrow buttons. When viewing the gene structure, the exons, introns and Pfam domains for genes can be selected easily. When the user hovers the mouse over each element, the length of the element will be shown. By clicking the element, the sequence information for the selected element will be displayed. As a demonstration, the analysis results for the Lipoxygenase gene family are presented in Figure 2.

Multiple types of gene structure display

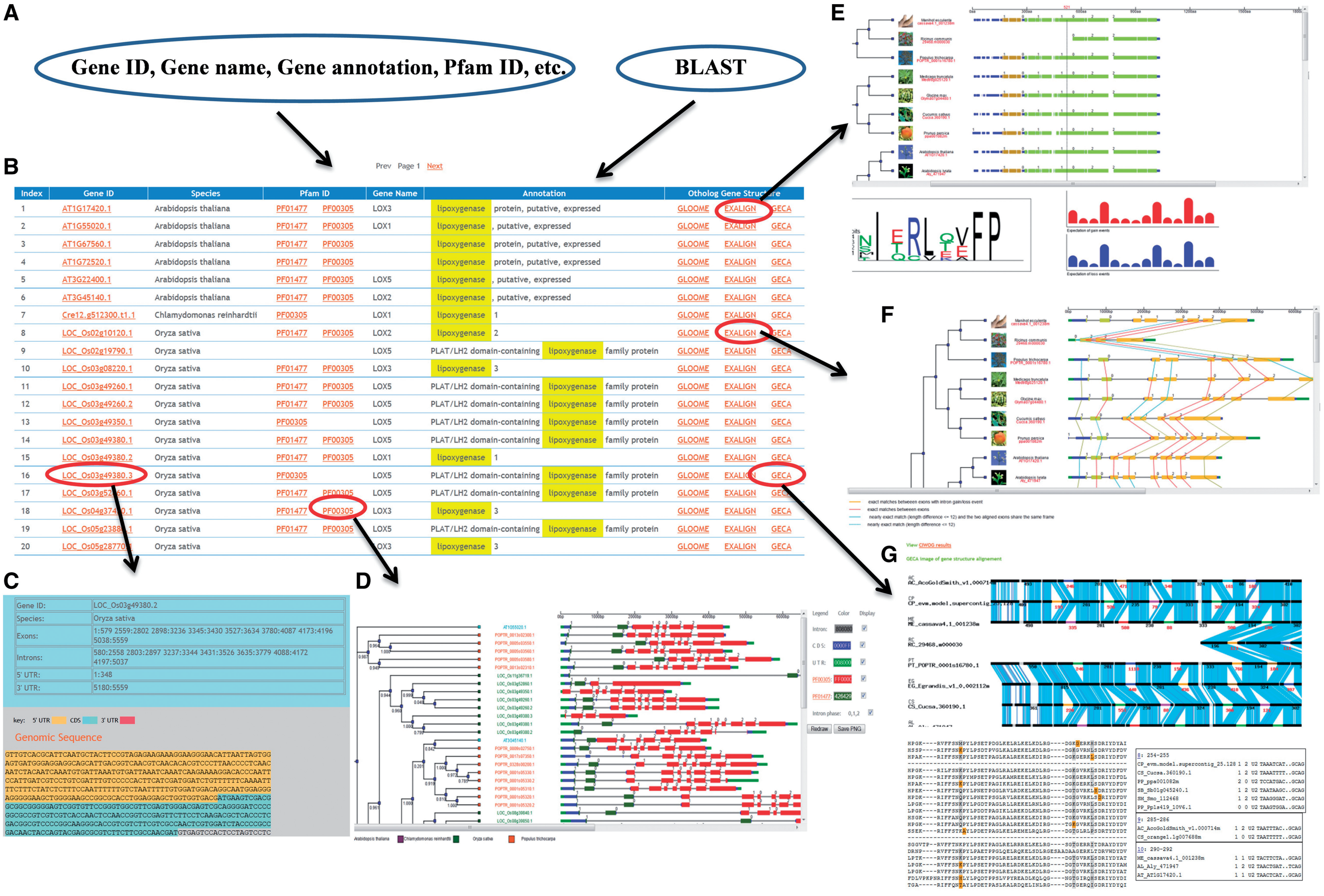
Gene structure visualization is important for analyzing exon–intron evolution. Typically, the basic components of gene structure (UTRs, intron, exons) are displayed on genomic sequence (19,20). To find relationships between exon–intron compositions in the encoded proteins, exon boundaries are also mapped onto the protein sequence (21,22). The view function in PIECE provides three types of exon–intron displays for each Pfam domain. Users can select any protein domain of interest by clicking the Pfam ID in the search results.

Analysis of gene structure evolution with groups of orthologs

On a gene family scale, global analysis is useful for dating intron changes; however, for certain genes, gene structure evolution in different species is not clear in phylogenetic trees with exon–intron pattern diagrams. Moreover, not all genes contain Pfam motifs in their encoding proteins and therefore cannot be analyzed as in Figure 2. Consequently, it is necessary to show exon–intron fluctuations for each gene in the database because intron-containing genes are spread across diverse plant phyla, whereas orthologs often have similar exon–intron organization even at large evolutionary distances (23). PIECE provides three analysis methods for each gene to infer the evolution of exon–intron structure in multiple protein-coding ortholog sets along a fixed-species phylogeny.

GLOOME

To analyze the gain and loss of introns in the ortholog group, we used GLOOME, which accurately infers branch-specific and site-specific gain and loss events with presence and absence profiles. To integrate GLOOME into PIECE, we first aligned the protein sequences within the ortholog group. We next coded intron characteristics using binary characters to denote presence (‘1’) and absence (‘0’). The 0/1 matrix, in which rows correspond to species and columns corresponds to


A
Gene ID, Gene name, Gene annotation, Pfam ID, etc.
BLAST
B
Prev Page 1 Next
C
D
E
F
G
Legend
Color
Display
Intron: 808080
CDS: 0000FF
UTR: 008000
PF00305: FF0000
PF01477: 426429
Intron phase: 0,1,2
Redraw
Save PNG
Arabidopsis thaliana
Chlamydomonas reinhardtii
Oryza sativa
Populus trichocarpa
exact matches between exons with intron gain/loss event
exact matches between exons
nearly exact match (length difference <= 12) and the two aligned exons share the same frame
nearly exact match (length difference <= 12)
View CIWOG results
GECA image of gene structure alignment
Expectation of gain events
Expectation of loss events

Index	Gene ID	Species	Pfam ID		Gene Name	Annotation	Otholog Gene Structure		
1	AT1G17420.1	Arabidopsis thaliana	PF01477	PF00305	LOX3	lipoxygenase protein, putative, expressed	GLOOME	EXALIGN	GECA
2	AT1G55020.1	Arabidopsis thaliana	PF01477	PF00305	LOX1	lipoxygenase , putative, expressed	GLOOME	EXALIGN	GECA
3	AT1G67560.1	Arabidopsis thaliana	PF01477	PF00305		lipoxygenase protein, putative, expressed	GLOOME	EXALIGN	GECA
4	AT1G72520.1	Arabidopsis thaliana	PF01477	PF00305		lipoxygenase protein, putative, expressed	GLOOME	EXALIGN	GECA
5	AT3G22400.1	Arabidopsis thaliana	PF01477	PF00305	LOX5	lipoxygenase , putative, expressed	GLOOME	EXALIGN	GECA
6	AT3G45140.1	Arabidopsis thaliana	PF01477	PF00305	LOX2	lipoxygenase , putative, expressed	GLOOME	EXALIGN	GECA
7	Cre12.g512300.t1.1	Chlamydomonas reinhardtii	PF00305		LOX1	lipoxygenase 1	GLOOME	EXALIGN	GECA
8	LOC_Os02g10120.1	Oryza sativa	PF01477	PF00305	LOX2	lipoxygenase 2	GLOOME	EXALIGN	GECA
9	LOC_Os02g19790.1	Oryza sativa			LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
10	LOC_Os03g08220.1	Oryza sativa	PF01477	PF00305	LOX3	lipoxygenase 3	GLOOME	EXALIGN	GECA
11	LOC_Os03g49260.1	Oryza sativa	PF01477	PF00305	LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
12	LOC_Os03g49260.2	Oryza sativa	PF01477	PF00305	LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
13	LOC_Os03g49350.1	Oryza sativa	PF00305		LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
14	LOC_Os03g49380.1	Oryza sativa	PF01477	PF00305	LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
15	LOC_Os03g49380.2	Oryza sativa	PF01477	PF00305	LOX1	lipoxygenase 1	GLOOME	EXALIGN	GECA
16	LOC_Os03g49380.3	Oryza sativa	PF00305		LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
17	LOC_Os03g52860.1	Oryza sativa	PF01477	PF00305	LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
18	LOC_Os04g37430.1	Oryza sativa	PF01477	PF00305	LOX3	lipoxygenase 3	GLOOME	EXALIGN	GECA
19	LOC_Os05g23880.1	Oryza sativa	PF01477	PF00305	LOX5	PLAT/LH2 domain-containing lipoxygenase family protein	GLOOME	EXALIGN	GECA
20	LOC_Os05g28770.1	Oryza sativa			LOX3	lipoxygenase 3	GLOOME	EXALIGN	GECA

Gene ID:	LOC_Os03g49380.2
Species:	Oryza sativa
Exons:	1:579 2559:2802 2898:3236 3345:3430 3527:3634 3780:4087 4173:4196 5038:5559
Introns:	580:2558 2803:2897 3237:3344 3431:3526 3635:3779 4088:4172 4197:5037
5' UTR:	1:348
3' UTR:	5180:5559

key: 5' UTR 🟧 CDS 🟦 3' UTR 🟥

Genomic Sequence

```
GTTGTCACGCATTCAATGCTACTTCCGTAGAGAAGAAAGGAAGGGAACATTAATTAGTGG
AGTGATGGGAGGAGGCAGCATTGACGGTCAACGTCAACACACGTCCCTTAACCCCTCAAC
AATCTACAATCAAATGTGATTAAATGTGATTAAATCAAATCAAGAAAAGGACACCCAATT
CCATTGATCGTCCTGTCGATTTGTCCCCCACTTCATCGATGGATCTGTTTTTTCAAAATT
TTCTTTCTATCTCTTTCCAATTTTTTGTCTAATTTTTGTGGATGGACAGGCAATGGAGGG
AGGGGGAAGCTGGGGAAGCCGGCGCACCTGGAGGAGCTGGTGGTGACGATGAAGTCGACG
GCGGCGGGGGAGTCGGTGTTCCGGGTGGCGTTCGAGTGGGACGAGTCGCAGGGGATCCCC
GGCGCCGTCGTCGTCACCAACTCCAACCGGTCCGAGTTCTTCCTCAAGACGCTCACCCTC
GACGGCGTCCCCGGCAAGGGCACCGTCGTCTTCGTCGCCAACTCGTGGATCTACCCCGCC
GACAACTACCAGTACGAGCGCGTCTTCTTCGCCAACGATGTGAGTCCACTCCTAGTCCTC
```

```
HPGK----RVFFSNMPYLPSETPDGLKELRLKELKDLRG----DGKGERKHSDRIYDYAI
HSSP----RIFFSNKPYLPSDTPVGLKALREKELKDLRG----DGKGVRNLSDRVYDFDV
HPAK----RIFFSNKPYLPSETPAGLKALREKELKDIRG----DGKGERKLSDRIYDFDV
------------------------------------------------------------
HPGK----RILFSNKPYLPSETPAGLRALREKELRDLRG----DGKGVRKLSDRIYDFDV
HPGK----RIFFSNKPYMPSETPPGLKAMREEELKYLRG----DGKGERKLSDRIYDFDL
HPAK----RIFFSNKPYLPGETPAGIKKLREIELKDIRG----DGKGERKLSDRVYDFDV
HPEK----RIFFTNQPYLPNETPGGLRVLRDKELKNLRG----DGTGVRKLSDRIYDFDV
HPDK----RIFFTNQPYLPNETPSGLRVLREKELKNLRG----DGSGVRKLSDRIYDFDV
LPGK----RIFFSNKPYLPGDTPAGLRLLREKELRNLRG----DGKGVRNLSDRIYDYDI
LPGK----RIFFSNKPYLPDDTPAGIKLLREKDLKNLRG----DGKGVRKLSDRIYDYDT
HSGK----RIFFANQPYLPSETPAGLRALREKELKDIRG----TGKGVRKLSDRIYDYDV
HPGK----RIFFANQPYLPSETPAGLRALREKELKDIRG----TGKGVRKLSDRIYDYDV
SSEK----RIFFSTKAYLPDETPAGLKELREMELRQLRG----DGTGLRLPSDRTYDYDT
------------------------------------------------------------
SGGVTP--RVFFTNKPYLPSKTPAGLRELRRRELKELRG----SGTGERRITDRAYDYDV
DRNP----RVFFTNKPYLPAETPPGLQELRQKELSDLRGESAADAAGERKLTDRVWDYDV
LPTK----RVFFSNKPYLPSETPPGLRELREKELKDLRG----DGTGVRKLSDRIYDYAT
LPGK----RVFFSNKPYLPSETPPGLRELRDKELKDLRG----DGTGVRKLSDRIYDYAM
LPGK----RVFFSNKPYLPSETPPGLRELRDKELKDLRG----DGTGVRKLSDRIYDYAT
FDLVPKPNRIFFSNHLYLPQDTPPSLVPYREADLKQLQG----NGTGIRQHSERIYDYDI
TGA-----RIFFQNTAYLPDETPASLKALRDQELINLRG----DGTGERQIGDRIYDYQI
```

```
8: 254-255
CP_evm.model.supercontig_25.128 1 2 U2 TAAATCAT..GCAG
CS_Cucsa.360190.1               1 2 U2 TAAATTTT..GCAG
PP_ppa001082m                   2 0 U2 TCCATGAC..GCAG
SB_Sb01g045240.1                1 2 U2 TAATAAGC..GCAG
SM_Smo_112468                   1 2 U2 TAAGGGAT..GCAG
PP_Pp1s419_10V6.1               0 0 U2 TAAGTGGA..GCAG
9: 285-286
AC_AcoGoldSmith_v1.000714m     1 2 U2 TAATTTAC..GCAG
CS_orange1.1g007688m           1 0 U2 TAATTTTT..GCAG
10: 290-292
ME_cassava4.1_001238m          1 1 U2 TACTTCTA..GCAG
AL_Aly_471947                  1 2 U2 TAACTGAT..TCAG
AT_AT1G17420.1                 1 1 U2 TAACTCAT..GCAG
```

Figure 1. Graphical user interface overview. The user interface provides multiple ways to access information stored in the database. (**A**) Multiple inputs for PIECE. (**B**) Search results interface. (**C**) Sequence detail including exon–intron information. (**D**) The graphical viewer for PIECE. (**E**) GLOOME analysis results. (**F**) Exalign analysis results. (**G**) GECA (CIWOG) analysis results.

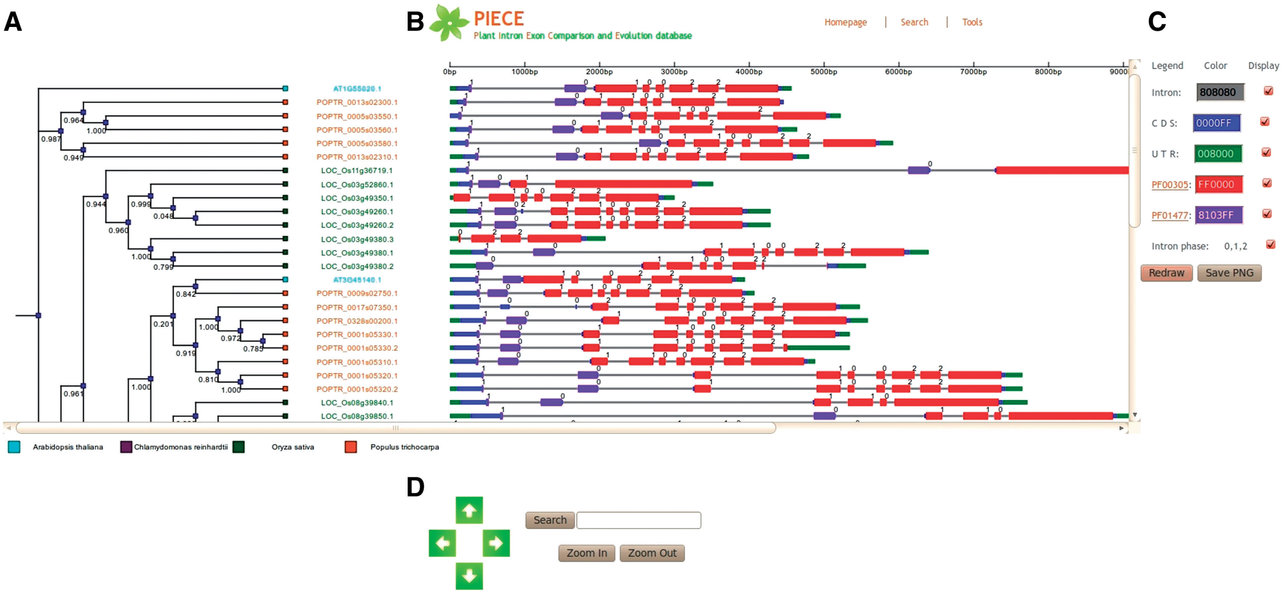


Figure 2. The PIECE viewer. Data for the *LOX* gene family (PF00305) in *Arabidopsis thaliana*, rice, poplar and *Chlamydomonas reinhardtii*. (**A**) Dendrogram of sequences clustered according to the presence and similarity of extracted Pfam motifs. (**B**) Diagram that displays positional information of the gene structure in each sequence. (**C**) Color selector and check boxes for displaying introns, CDS, UTRs, Pfam domains and intron phases, and save button to save the output as a PNG file. (**D**) Color for the plant species and operation panel for manipulating the output.

binary characters, is termed a phylogenetic profile of presence–absence or phyletic pattern and is equivalent to a MSA. In PIECE, the output of GLOOME includes plant species trees, gene structure displays, intron site sequence logos and the expected number of gains and losses for each intron site (Figure 1E). When users click each box in the histogram, the viewer will show the intron site in the aligned protein sequence. The alignments were generated using MUSCLE, and the sequence alignment graphical display was implemented in the Jalview (24) Java applet. GLOOME provides useful analytical facilities for exploring the degree of conservation of intron evolution across proteins in the ortholog group and also for analyzing the distribution of exonic sequences within the aligned coding sequences of domains.

Exalign

During evolution, one exon may split into multiple exons or multiple exons may fuse into one; such events have stringent constraints in exon length, and this characteristic can be used to determine cases of exon fusion or division. To analyze the evolution of gene structure of orthologs, we use another tool named Exalign (13). The Exalign viewer of PIECE can show the relationship of exons in orthologous genes from different plant species. This viewer provides exon–intron display for orthologs of gene structure data sets linked to the species phylogeny (Figure 1F). The gene-exon comparison between species is shown as colored lines. Different colors indicate different exon comparison results. In PIECE, any gene data with its set of orthologs can be put into the Exalign viewer at the user's request to easily find the evolution history of genes and, particularly, to detect exon relationships and fusion events.

GECA

Aligning exon–intron structures accompanied with similarities between sequences is helpful for annotating gene structure information. GECA can analyze gene exon–intron organization and highlight changes in gene structure (14). GECA relies on protein alignments, completed with the identification of common introns in corresponding genes using CIWOG (Common Introns Within Orthologous Genes) (10). In PIECE, each gene has a GECA link to view the orthologs that are aligned using their common introns detected by CIWOG. The similarities between orthologous sequences in the alignment are represented at the level of amino acids in the translated exons. A blue line links two amino acids if they are identical, a purple line indicates conservative substitutions, and intron type is detected by CIWOG (Figure 1G).

GSDraw web server

A number of web tools have been developed for gene structure annotation, such as GSDS (19), FancyGene (20) and GECA (14). The purpose of these programs is to represent the exon–intron structure of several genes in a single image to perform global gene structure comparisons (14). However, these resources do not display sequences with phylogenetic relationships and automatically detected protein motifs. Therefore, we developed GSDraw as part of PIECE. GSDraw is a convenient and easy-to-use interface for gene structure annotation that integrates Sim4 (25), MEME (26), MUSCLE (16) and FastTree (17) into a single web-based tool. The procedures for designing and implementing the GSDraw server are illustrated in Figure 3. Users submit a query sequence set (in multi-FASTA format) consisting of genomic, CDS or transcript sequences to GSDraw

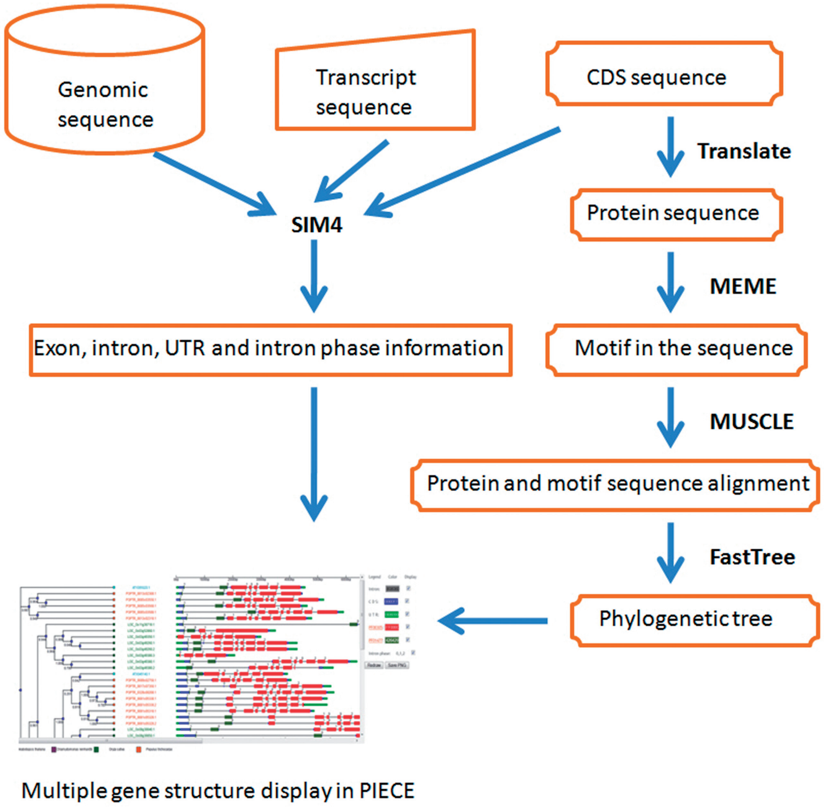


Figure 3. Workflow chart of GSDraw.

(http://wheat.pw.usda.gov/piece/GSDraw.php) and obtain schematic diagrams of their gene structures with annotated Pfam protein motifs and a phylogenetic tree. This capability allows users to view a PIECE database-style display for any selected gene family group (of three or more genes) from any species with available data. The GSDraw output for three rice LRR-Kinase genes is shown in Supplementary Figure S1. The user can modify the gene structure display to their own preferences by selecting different colors for the annotated sequences and/or choose whether or not to display each of the Pfam motifs, similar to what is allowed in the PIECE viewer.

DISCUSSION

Simple sequence alignment and comparison usually is unable to provide a clear picture of the structural evolution of genes, e.g. how their intron–exon structures, intron lengths, alternative splicing and untranslated regions change over time. Although there has been a rapid growth in the number of plant genome databases, such as PlantGDB (7), PLAZA (8), Phytozome (9) and GreenPhylDB (27), these resources lack comparative analytical capabilities for integrating protein domains from multiple species to investigate exon–intron structural evolution. ExDom (28) contains an extensive collection of exon–intron gene structures mapped to protein domains, but it primarily focuses on non-plant species. Furthermore, most related databases do not display the phylogenetic tree of gene families and orthologous gene evolution histories. To address these limitations, we developed PIECE, which characterizes the number, position and length of introns and exons from 25 individual sequenced plant genomes. The PIECE database provides a panoramic perspective from which to investigate the evolution of gene structures on a broad evolutionary time scale. Furthermore, PIECE provides an easy entry point for researchers to immediately access gene structure evolution information without having to install any software.

For example, heat shock response in eukaryotes is transcriptionally regulated by conserved heat shock transcription factors (Hsfs). *Hsf* genes are represented by a large multigene family in plants. To illustrate the possible mechanisms of structural evolution of *Hsf* homologs, we used PIECE to compare the exon–intron structures of individual *Hsf* genes in 10 plant lineages. Supplementary Figure S2 provides a detailed illustration of the relative length of introns and the conservation of the corresponding exon sequences within each of the *Hsf* genes. Notably, although the members of the *Hsf* gene family exhibited differences in intron number and intron length, the intron positions and intron phases were remarkably well conserved, with conserved splicing sites between adjacent exons.

To further investigate the structural evolution of *Hsf* genes in different lineage species, we also used PIECE to create images that contain gene structure information in

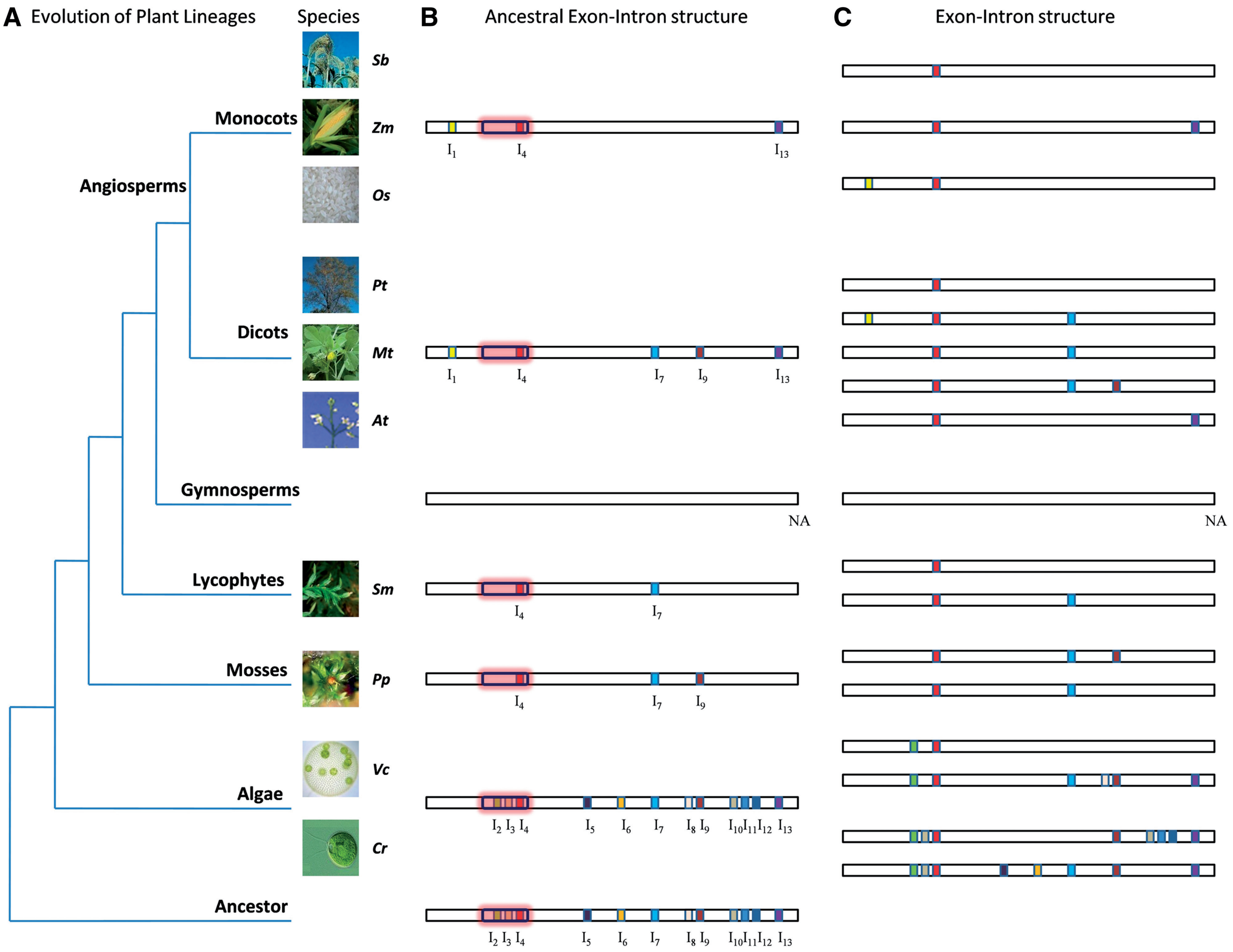

A Evolution of Plant Lineages
Species
B Ancestral Exon-Intron structure
C Exon-Intron structure
Sb
Monocots
Zm
Angiosperms
Os
Pt
Dicots
Mt
At
Gymnosperms
NA
NA
Lycophytes
Sm
Mosses
Pp
Vc
Algae
Cr
Ancestor
I_1 I_4 I_{13}
I_1 I_4 I_7 I_9 I_{13}
I_4 I_7
I_4 I_7 I_9
I_2 I_3 I_4 I_5 I_6 I_7 I_8 I_9 I_{10} I_{11} I_{12} I_{13}
I_2 I_3 I_4 I_5 I_6 I_7 I_8 I_9 I_{10} I_{11} I_{12} I_{13}

Figure 4. An evolutionary model for the structural evolution of the *Hsf* gene family in plants. (**A**) Dendogram representing the evolutionary relationship of all plant lineages. (**B**) Proposed exon–intron structure of the ancestral *Hsf* gene in each plant lineage. (**C**) Current exon–intron structure of *Hsf* genes. The exon–intron structure of the *Hsf* genes in gymnosperms is represented with a empty bar because genomic sequences are unavailable.

unaligned and aligned protein sequences (Supplementary Figures S3 and S4). We next constructed an evolutionary model that could predict the current *Hsf* genes in plant species of different lineages (Figure 4). Under the assumption that introns, which were located at identical positions and given identical phase, should be present in the common ancestor, we reconstructed the ancestral exon–intron structure of *Hsf* for all plant lineages (Figures 4). The results obtained from the *Hsf* intron analysis suggested that the ancestral *Hsf* contained ≥12 introns, symmetrically distributed throughout its coding sequence (Figure 4). The aquatic plants (green algae) have a large number of introns. Most introns were lost in the evolution of aquatic plants (green algae) to lower land plants (mosses and lycophytes), including I_2, I_3, I_5, I_6, I_8 and I_{10}–I_{12}. Moreover, single intron losses also occurred during the expansion and divergence of the *Hsf* gene family in each plant lineage. For example, the ancestral *Hsf* in monocots contained at least 3 introns, whereas all *Hsf* genes in monocots contained only 1 or 2 introns (Figure 4B and C). It appears that I_7 and I_9 are not present in monocots, but are present in the dicot ancestor. Besides the intron loss, gain of an intron is also observed. I_1 is only present in angiosperms. Furthermore, the analysis revealed that I_4 is present in the *Hsf* gene of the common ancestor of all plant lineages (Figure 4B), and its position is in the DNA binding domain (DBD) (Figure 4C). This observation indicates that the Hsf family in plants not only has a conserved DBD motif but also contains a conserved intron in the DBD domain.

As the examples demonstrate, the capabilities of PIECE will provide researchers with many hypotheses for designing molecular biology studies and will help to elucidate the evolutionary history of plant genes. Future efforts will extend the number of available plant species and enhance the analytical capabilities of PIECE. Newly published plant genomes will enable efficient phylogenetic analyses of exon-domain relationships in plants and in-depth analysis of the evolutionary history of protein domains. Alternative splicing is an important biological process that greatly increases the biodiversity of proteins that can be encoded by the genome. One of the future directions will focus on the integration of alternative splice data into PIECE for gene evolution and structure analyses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

The authors thank Eric Yan for his assistance on the database development and Xiaohao He for the critical reading of the manuscript. The authors are very grateful to Farhad Ghavami, Ajay Kumar, Vijay Tiwari, Jiajie Wu, Filippo Bassi and the anonymous reviewers for testing the server and offering valuable comments.

FUNDING

U.S. National Science Foundation [IOS 0701916 and IOS 0822100]; United State Department of Agriculture, Agriculture Research Service CRIS project [5325-21000-014]. Funding for open access charge: US Department of Agriculture, Agriculture Research Service.

Conflict of interest statement. None declared.

REFERENCES

1. Javelle,M., Klein-Cosson,C., Vernoud,V., Boltz,V., Maher,C., Timmermans,M., Depege-Fargeix,N. and Rogowsky,P.M. (2011) Genome-wide characterization of the HD-ZIP IV transcription factor family in maize: preferential expression in the epidermis. *Plant Physiol.*, **157**, 790–803.
2. Turchetto-Zolet,A.C., Maraschin,F.S., de Morais,G.L., Cagliari,A., Andrade,C.M., Margis-Pinheiro,M. and Margis,R. (2011) Evolutionary view of acyl-CoA diacylglycerol acyltransferase (DGAT), a key enzyme in neutral lipid biosynthesis. *BMC Evol. Biol.*, **11**, 263.
3. Strommer,J. (2011) The plant ADH gene family. *Plant J.*, **66**, 128–142.
4. Zhu,Z., Zhang,Y. and Long,M. (2009) Extensive structural renovation of retrogenes in the evolution of the Populus genome. *Plant Physiol.*, **151**, 1943–1951.
5. Garcia-Espana,A., Mares,R., Sun,T.T. and Desalle,R. (2009) Intron evolution: testing hypotheses of intron evolution using the phylogenomics of tetraspanins. *PLoS One*, **4**, e4680.
6. Lin,H., Zhu,W., Silva,J.C., Gu,X. and Buell,C.R. (2006) Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol.*, **7**, R41.
7. Duvick,J., Fu,A., Muppirala,U., Sabharwal,M., Wilkerson,M.D., Lawrence,C.J., Lushbough,C. and Brendel,V. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **36**, D959–D965.
8. Proost,S., Van Bel,M., Sterck,L., Billiau,K., Van Parys,T., Van de Peer,Y. and Vandepoele,K. (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.
9. 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.
10. Wilkerson,M.D., Ru,Y. and Brendel,V.P. (2009) Common introns within orthologous genes: software and application to plants. *Brief Bioinform.*, **10**, 631–644.
11. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
12. Cohen,O., Ashkenazy,H., Belinky,F., Huchon,D. and Pupko,T. (2010) GLOOME: gain loss mapping engine. *Bioinformatics*, **26**, 2914–2915.
13. Pavesi,G., Zambelli,F., Caggese,C. and Pesole,G. (2008) Exalign: a new method for comparative analysis of exon-intron gene structures. *Nucleic Acids Res.*, **36**, e47.
14. Fawal,N., Savelli,B., Dunand,C. and Mathe,C. (2012) GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families. *Bioinformatics*, **28**, 1398–1399.
15. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
16. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
17. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
18. Rasko,D.A., Myers,G.S. and Ravel,J. (2005) Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics*, **6**, 2.
19. Guo,A.Y., Zhu,Q.H., Chen,X. and Luo,J.C. (2007) [GSDS: a gene structure display server]. *Yi Chuan*, **29**, 1023–1026.
20. Rambaldi,D. and Ciccarelli,F.D. (2009) FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics*, **25**, 2281–2282.
21. Leslin,C.M., Abyzov,A. and Ilyin,V.A. (2004) Structural exon database, SEDB, mapping exon boundaries on multiple protein structures. *Bioinformatics*, **20**, 1801–1803.
22. Csuros,M. (2008) Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics*, **24**, 1538–1539.
23. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.*, **13**, 1512–1517.
24. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
25. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
26. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
27. Rouard,M., Guignon,V., Aluome,C., Laporte,M.A., Droc,G., Walde,C., Zmasek,C.M., Perin,C. and Conte,M.G. (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, **39**, D1095–D1102.
28. Bhasi,A., Philip,P., Manikandan,V. and Senapathy,P. (2009) ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.*, **37**, D703–D711.