Gene 609 (2017) 1–8

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

The fate of retrotransposed processed genes in *Arabidopsis thaliana*

Basma T.M. Abdelkarim, Vincent Maranda, Guy Drouin *

Département de biologie et Centre de recherche avancée en génomique environnementale, Université d'Ottawa, Ottawa, Ontario K1N 6N5, Canada

ARTICLE INFO

Article history:
Received 22 November 2016
Received in revised form 20 January 2017
Accepted 24 January 2017
Available online 26 January 2017

Keywords:
Processed genes
Plants
Retrotranscription
Retrotransposition
Selective pressures
Synonymous substitutions
Nonsynonymous substitutions

ABSTRACT

Processed genes are functional genes that have arisen as a result of the retrotransposition of mRNA molecules. We found 6 genes that generated processed genes in the common ancestor of five Brassicaceae species (*Arabidopsis thaliana*, *Arabidopsis lyrata*, *Capsella rubella*, *Brassica rapa* and *Thellungiella parvula*). These processed genes have therefore been kept for at least 30 million years. Analyses of the Ka/Ks ratio of these genes, and of those having given rise to them, show that they evolve relatively slowly and suggest that the processed genes maintained the same function as that of their parental gene. There is a significant negative correlation between the number of ESTs and transcripts produced and the Ka/Ks ratios of the parental genes but not of the processed genes. This suggests that selection has not yet adapted the selective pressure the processed genes experience to their expression level. However, the *A. thaliana* processed genes tend to be expressed in the same tissues as that of their parental genes. Furthermore, most have a CAATT-box, a TATA-box and are located about 1 kb from another protein-coding gene. Altogether, our results suggest that the processed genes found in the *A. thaliana* genome have been kept to produce more of the same product, and in the same tissues, as that encoded by their parental gene.

© 2017 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Processed sequences are generated through the reverse transcription of mature mRNA molecules which are subsequently randomly integrated into the genome (Fink, 1987; Brosius, 1999; Kaessmann et al., 2009; Carelli et al., 2016). These processed sequences are therefore characterized by the lack of a promoter, the absence of introns and the presence of a 3′poly-A tail. Since their random insertion site usually lacks regulatory elements, these processed sequences are most often not expressed and they quickly become processed pseudogenes (Drouin and Dover, 1987; Zou et al., 2009). However, this is not always the case. Occasionally, the processed sequences are inserted in a locus where promoter elements support transcription. In such cases they can be transcriptionally active and become processed genes (Brosius, 1999; Ding et al., 2006; Vinckenbosch et al., 2006; Wang et al., 2006; Sakai et al., 2008; Pei et al., 2012; Carelli et al., 2016). These promoter elements can have at least five sources: they can be derived from the parental gene if the parental mRNA was transcribed from an upstream alternative promoter, they can be recruited from another gene by inserting themselves downstream of the promoter of another gene, they can be recruited from another gene by inserting themselves next to another gene which has a bidirectional promoter, they can be recruited from another gene by inserting themselves into the intron of a gene or sequences able to act as promoters can simply be present at a given insertion site by chance (Carelli et al., 2016).

Although functional processed genes are much less abundant than processed pseudogenes, many of them have been characterized in mammalian genomes (Brosius, 1999; Wang, 2004; Vinckenbosch et al., 2006; Sakai et al., 2008; Carelli et al., 2016). However, there is much less information regarding the evolutionary fate of processed genes in plants (Benovoy and Drouin, 2006; Wang et al., 2006; Zou et al., 2009). Here, we wanted to learn more about the functional relevance of the expressed *A. thaliana* processed genes we identified in a previous study (Benovoy and Drouin, 2006). In particular, we were interested in finding out whether these processed genes were specific to *A. thaliana* or whether they were also present in related species, whether they evolved under purifying or positive selection, whether they experienced similar selective pressures in different species, whether their expression levels were similar to that of their parental genes, whether they were expressed in the same tissues as their parental gene and to elucidate how they acquired the promoters necessary to their expression.

2. Materials and methods

2.1. Gene sequences

We used the gene names of the genes which gave rise to processed genes in *A. thaliana* to retrieve the sequences of these parental genes

Abbreviations: kb, kilo bases; bp, base pairs; Ka, number of nonsynonymous substitutions per nonsynonymous site; Ks, number of synonymous substitutions per synonymous site; EST, expressed sequence tag.

* Corresponding author.
E-mail address: gdrouin@science.uottawa.ca (G. Drouin).

http://dx.doi.org/10.1016/j.gene.2017.01.029
0378-1119/© 2017 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 1
Distribution of processed genes in five Brassicaceae species.

Gene	Function	Number of introns[a]	Expressed in[b]	*A. thaliana*	*A. lyrata*	*C. rubella*	*B. rapa*	*T. parvula*	*A. lyrata*
At2g44450	Beta glucosidase 15	11	23 structures	Present	Present	Present	Present	Present	Present
At3g14510	Geranylgeranyl pyrophosphate synthase 3	1	Roots	Present	Present	Present	Present	Present	Present
At4g27130	Trans initiation factor SUI1 family protein	3	24 structures	Present	Present	Present	Present	Present	Present
At4g40030	Histone H3.3	2	25 structures	Present	Present	Present	Present	Present	Present
At5g01320	Pyruvate decarboxylase	4	Unknown	Present	Present	Present	Present	Present	Present
At5g10980	Histone H3.3	2	25 structures	Present	Present	Present	Present	Present	Present

[a] The number of introns are those found in the coding region of *A. thaliana* genes.
[b] Function and tissue expression data were obtained from the TAIR web site (https://www.arabidopsis.org/).

(listed in Table 2 of Benovoy and Drouin, 2006). Since all these genes contain introns (Table 1), we then removed the introns of these parental genes and used the resulting sequences to perform BLASTn searches against the genome sequence of *A. thaliana*, *A. lyrata*, *C. rubella*, *B. rapa* and *T. parvula* in order to recover the processed genes these parental sequences had produced (Altschul et al., 1997). As in our 2006 study, sequences were regarded as processed sequences when they were intronless (and derived from intron containing genes), that their sequence did not have any premature stop codons or frame-shifts and that their length was at least 95% of the length of their parental gene (Benovoy and Drouin, 2006).

2.2. Sequence analyses

Sequences were manipulated using the BioEdit program (Hall, 1999). They were aligned using the ClustalW function in BioEdit (Thompson et al., 1994). The number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number synonymous substitutions per synonymous site (Ks) between sequences were calculated using the Kumar method implemented in MEGA6 (Tamura et al., 2013). As described in the MEGA6 help files, this method is a modification of the Pamilo and Bianchi (1993), Li (1993) and Comeron (1995) methods that is able to handle the problematic degeneracy class assignments of arginine and isoleucine codons. As a rule, neutral evolution is defined by a Ka/Ks value of one, a value higher than one indicates positive selection, while a value lower than one indicates purifying (negative) selection (Graur and Li, 2000).

Phylogenetic trees were constructed using MEGA 6 with aligned DNA sequences, the maximum likelihood method with the general time reversible model with 5 discrete gamma categories plus invariant sites and 100 bootstraps (Tamura et al., 2013).

2.3. Gene expression data

Expressed sequence tags (ESTs) are often used to detect if a gene is being expressed (Dong et al., 2004). The number of ESTs produced by each parental and processed genes in each species were counted using BLAST searches against the EST database for each of *A. thaliana* and *B. rapa* at NCBI. We did not consider EST data from the other three species either because too few ESTs are yet known (*A. lyrata*) or because EST data are not yet available (*C. rubella* and *T. parvula*). Only identity matches larger or equal to 98% over at least 300 bp were considered positive matches. The plant tissues that generated the ESTs were obtained from the description of the sequence in the GenBank link for these EST sequences.

The *A. thaliana* RNA-Seq abundance data was obtained from the Arabidopsis Next-Gen Sequences DBs (https://mpss.danforthcenter.org/index.php; Nakano et al., 2006). The average number of distinct transcripts was calculated based on the average of seven different experiments. These average numbers of transcripts are all normalized assuming a total number of 35,000,000 transcripts. We could not find RNA-Seq data for the other four species.

2.4. Regulatory motifs

Potential regulatory motifs were identified using the Arabidopsis Information Resource (TAIR) website using the Seqviewer tool (https://seqviewer.arabidopsis.org/). Information on surrounding genes, CCAAT-Box and sequences were also obtained from the TAIR website (Swarbreck et al., 2008). Potential GC-Islands were identified using the EMBOSS Cpgplot tool (http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/). TATA-Box sequences were identified using the Softberry TSSP Prediction of PLANT Promoters tool (http://www.softberry.com/berry.phtml?topic=tssp&group=programs&subgroup=promoter).

3. Results

3.1. Processed genes

Only 6 of the 22 parental sequences we had identified in 2006 have corresponding processed genes in the most recent (2016) version of the *A. thaliana* genome (Benovoy and Drouin, 2006; Table 1). This is partially due to the fact that spurious sequences have been removed from the most recent version of the *A. thaliana* genome. This is also due to the fact that we only considered sequences which had lost all of their introns. For example, we did not consider the partially processed gene sequences generated by the At5g43330 gene (a malate dehydrogenase gene with 6 introns) because these sequences still have the first intron of this gene (i.e., they only lost 5 of the 6 introns; results not shown).

Interestingly, these 6 parental genes have processed genes in the five Brassicaceae species we studied (Fig. 1; Tables 1 and 2). Moreover, genes At2g44450, At3g14510, and At5g10980 generated more than one processed sequence for a total of 64 processed genes, 12 of which are found in the *A. thaliana* genome (Table 2). It is not unusual for a functional gene to produce several processed sequences (Benovoy and Drouin, 2006; McDonell and Drouin, 2012). Note that BLASTn searches with these six in silico processed parental sequences did not reveal any corresponding processed gene sequences in the genome of *Carica papaya*, the species that is most closely related to the above five Brassicaceae species and for which a genome sequence is available (results not shown). This is likely due to the absence of these processed genes in this species because BLASTn searches should be able to detect these

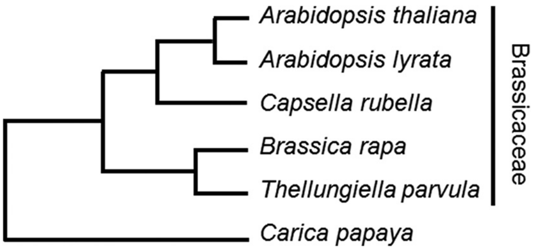

Fig. 1. Phylogenetic relationships of the Brassicaceae species studied in this paper. Source: Dicots Plaza 3.0, http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/.

Download English Version:

https://daneshyari.com/en/article/5589335

Download Persian Version:

https://daneshyari.com/article/5589335

Daneshyari.com