

Processed pseudogenes, processed genes and spontaneous mutations in the *Arabidopsis* genome

David Benovoy¹, Guy Drouin¹

Keywords: Processed pseudogenes, processed genes, spontaneous mutations, *Arabidopsis*

1 Introduction.

Processed pseudogenes are generated by the random integration of reversed-transcribed mature RNA (cDNA) molecules into genomes (1, 2). Given their origin, processed pseudogenes are characterized by the absence of promoters, a lack of introns, the presence of a poly(A)-tail at their 3'-end, and the presence of small direct repeats at their 5'- and 3'-termini (the result of target-site duplications). Because the sequences of processed pseudogenes are unconstrained by selective pressures, the substitutions they accumulate are considered to provide an accurate representation of the spontaneous mutations occurring in genomes (2).

2 Software and files.

We downloaded the complete *Arabidopsis thaliana* genome sequence (5 chromosomes totaling 125 Mb), and all 25,562 annotated genes present in January 2003, from the *Arabidopsis* Information Resource ftp site (TAIR web site: <ftp://ftp.arabidopsis.org/home/tair/>). We used an in-house PERL script to eliminate the genes that did not contain introns. This filter was utilized because the most important characteristic of processed pseudogenes is that they do not have the introns present in the genomic sequences of the genes from which they are derived. The introns of the 20,532 intron-containing genes were removed to mimic the molecular process that occurs in pre-mRNA molecules. These *in silico* spliced sequences were then used as queries against the entire genome of *Arabidopsis thaliana* using the program FASTA 3.4 with a Ktup of 6 (3). The alignments produced were then filtered using a PERL script that selected processed pseudogenes with a length $\geq 95\%$ that of the query. Finally, all sequences described as being retrotransposons were also discarded.

3 Results.

We identified 411 processed sequences in the *Arabidopsis* genome based on the fact that they have lost their intron(s) and have a length that is at least 95% of the length of the gene that gave rise to them. These sequences were generated by 279 different genes and clearly originated from retrotransposon events because most of them (91%) have a poly(A)-tail. They are composed of 376 sequences with frame shifts and/or premature stop codons (processed pseudogenes) and 35 sequences without disablements (processed genes). Eleven of these processed genes are likely functional retrotransposed genes because they have low Ka/Ks ratios, high Ks values and that their sequences match numerous *Arabidopsis* ETSs.

Distribution of processed sequences in the *Arabidopsis* genome

The number of processed sequences on each of the *Arabidopsis thaliana* chromosomes is proportional to their length (Table 1; Spearman's rank correlation test, $r^2 = 0.96$, $p = 0.0084$). Processed sequences are randomly distributed along the length of *Arabidopsis* chromosomes 2 to 5 ($p > 0.1$) but not on chromosome 1 because it contains four clusters with 12, 11, 13 and 11 processed sequences, respectively ($p < 0.001$).

¹ Biology Department, University of Ottawa, Ottawa, ON Canada, E-mail= DBeno620@science.Ottawa.ca

Table 1. Distribution processed sequences in the *Arabidopsis* genome.

Chromosome	Genes	length (Mb)	GC (%)	Processed sequences		
				From	On	Density
1	6517	29.6	33.4	87/10	110/8	3.72/0.27
2	4035	19.6	35.5	78/10	59/1	3.01/0.05
3	5224	23.3	35.4	83/6	87/4	3.73/0.17
4	3828	17.5	35.5	48/3	46/8	2.63/0.46
5	5956	26.3	34.5	80/6	74/14	2.81/0.53
Total (average)	25560	116.4	34.9	376/35	376/35	3.18/0.30

Genes producing processed pseudogenes

We used the TAIR web site (<http://www.arabidopsis.org/index.jsp>) to categorize the 279 genes that generated the 411 processed sequences we identified, as well as all the protein coding genes present in the *Arabidopsis* genome. The type of genes generating processed sequences is strongly correlated with the number of such genes in the *Arabidopsis* genome (results not shown; $r^2 = 0.81$, $p < 0.001$). The number of processed sequences generated by different genes is also not biased in favor of housekeeping genes highly expressed in the germ line (results not shown). In fact, the most abundant processed sequences found in the *Arabidopsis* genome are mostly hypothetical proteins and not housekeeping genes such as those coding for ribosomal proteins, actin, or tubulin (results not shown). Finally, most genes generate few processed pseudogenes (results not shown). The processed pseudogenes found in the *Arabidopsis* genome therefore originated from a random sampling of its genes.

Patterns of spontaneous mutations in *Arabidopsis*

Table 2 shows the combined pattern of substitution inferred from 66 processed pseudogenes with sequence similarities of at least 80% (2781 mutated sites / 52657 total sites).

From	To				Totals
	A	T	C	G	
A	-	6.61 ± 0.72	5.92 ± 0.74	12.72 ± 1.02	25.26
T	6.67 ± 0.87	-	6.11 ± 0.60	5.82 ± 0.65	18.60
C	9.15 ± 1.42	11.78 ± 1.48	-	7.18 ± 0.98	28.11
G	14.36 ± 1.48	8.14 ± 0.81	5.53 ± 0.82	-	28.03
Totals	30.18	26.53	17.56	25.72	

C and G are the most mutable nucleotides (with 28.11 and 28.03%, respectively) and nucleotides mutate most often to A and T nucleotides (with 30.18 and 26.53%, respectively) and least often to C nucleotides (17.56%). For individual nucleotides, we see that A changes to G more often than to other nucleotides, T changes to A more often than to other nucleotides, C changes to T more often than to other nucleotides and G changes to A more often than to other nucleotides. For transitions, we see that C to T substitutions (11.78%) are more frequent than T to C substitutions (6.11%) and that G to A substitutions (14.36%) are more frequent than A to G substitutions (12.72%).

4 Discussion.

Processed sequences are mostly randomly distributed in the *Arabidopsis* genome. In contrast with the situation observed in mammals, the processed sequences found in the *Arabidopsis* genome originate from abundant genes and not from highly expressed genes. The patterns of spontaneous mutations in *Arabidopsis* are slightly different than those of mammals but are similar to those observed in *Drosophila*. This suggests that that methylated cytosine deamination is less frequent in *Arabidopsis* than in mammals.

5 References.

- [2] Graur, D. and Li, W.-H. 1999. *Fundamentals of Molecular Evolution*. 2nd edition. Sinauer Associates, Sunderland, MA.
- [3] Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences USA* 85:2444-2448.
- [1] Weiner, A.M., Deininger, P.L. and Efstratiadis, A. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annual Review of Biochemistry* 55:631-661.