

# Computational Identification of 69 Retroposons in *Arabidopsis*<sup>1[w]</sup>

Yujun Zhang<sup>2</sup>, Yongrui Wu<sup>2</sup>, Yilei Liu, and Bin Han\*

National Center for Gene Research (Y.Z., Y.W., Y.L., B.H.), and Shanghai Institute of Plant Physiology and Ecology (Y.Z., B.H.), Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China

Retroposition is a shot-gun strategy of the genome to achieve evolutionary diversities by mixing and matching coding sequences with novel regulatory elements. We have identified 69 retroposons in the *Arabidopsis* (*Arabidopsis thaliana*) genome by a computational approach. Most of them were derivatives of mature mRNAs, and 20 genes contained relics of the reverse transcription process, such as truncations, deletions, and extra sequence additions. Of them, 22 are processed pseudogenes, and 52 genes are likely to be actively transcribed, especially in tissues from apical meristems (roots and flowers). Functional compositions of these retroposon parental genes imply that not the mRNA itself but its expression in gamete cells defines a suitable template for retroposition. The presence/absence patterns of retroposons can be used as cladistic markers for biogeographic research. Effects of human and the Mediterranean Pleistocene refugia in *Arabidopsis* biogeographic distributions were revealed based on two recent retroposons (At1g61410 and At5g52090). An evolutionary rate of new gene creation by retroposition was calculated as 0.6 genes per million years. Retroposons can also be used as molecular fossils of the parental gene expressions in ancient time. Extensions of 3' untranslated regions for those expressed parental genes are revealed as a possible trend of plant transcriptome evolution. In addition, we reported the first plant functional chimeric gene that adapts to intercompartmental transport by capturing two additional exons after retroposition.

The scientific community has witnessed an accelerated advance toward understanding genomes and their evolutionary dynamics in recent years. Gene numbers among these organisms vary greatly, indicating the existence of a general process of new gene creation, a central theme for evolutionary diversity and speciation. Several molecular mechanisms have been known for this process (Long et al., 2003) such as: (1) Gene duplication; (2) exon shuffling; (3) horizontal gene transfer; (4) DNA translocation; and (5) retropositions, or a combination of these mechanisms.

Retroposition describes a process by which an RNA species was reverse transcribed, and the resultant double strand cDNA was then incorporated back into the genome (Brosius, 1991). This process yields "retroposons," a term coined by Rogers (1983) to indicate those genomic sequences generated from reverse transcription (RT) of mRNAs. In fact, all retroelements that have RNA intermediates can be called retroposons by definition. In this article, we focused only on those genes that derived from mRNA molecules rather

than mobile elements. Since promoters of eukaryotic mRNAs are located upstream of the transcripts, such retroposons are most likely to be transcriptionally silenced once integrated into the chromosome, then accumulate mutations such as frameshifts, in-frame stop codons, fragment deletions or interspersed repeat element in coding regions, and finally are eliminated during evolution. That is why they are usually called processed pseudogenes. In this point, all retroposons can be considered as insertional mutagens, since they cause loss of function of both the retroposon itself and the targeted active gene as well. However, if the integration happens just downstream of an active gene, it can use the resident promoter to resurrect transcription, and a changed expression pattern will be achieved by mixing and matching already existent coding sequence with novel promoters or enhancers (for review, see Brosius, 2003).

With the availability of whole genome sequences, extensive analysis has been done in human, mouse, *Drosophila*, and *Caenorhabditis elegans*, and many retroposons have been identified by experimental or computational methods in recent years (Gonçalves et al., 2000; Harrison et al., 2001; Betrán et al., 2002; Harrison et al., 2003; Zhang et al., 2003). However, little is known about plant retroposons, being poorly annotated in plant genome databases except for the first reported potato processed actin gene (Drouin and Dover, 1987). We adopted a computational method to investigate retroposons in *Arabidopsis* (*Arabidopsis thaliana*) genome, their numbers, organizations, and transcriptional activities. Functional roles of retroposition in *Arabidopsis* are also discussed.

<sup>1</sup> This work was supported by the Ministry of Sciences and Technology (grant nos. 2002AA2Z1003 and 2003AA222091), by the Chinese Academy of Sciences, by the Shanghai Municipal Commission of Sciences and Technology (grant no. 038019315), and by the National Natural Science Foundation of China (grant no. 30325014).

<sup>2</sup> These authors contributed equally to the paper.

\* Corresponding author; e-mail bhan@ncgr.ac.cn; fax 86-21-64825775.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at [www.plantphysiol.org/cgi/doi/10.1104/pp.105.060244](http://www.plantphysiol.org/cgi/doi/10.1104/pp.105.060244).

## RESULTS

### Annotation and Functional Classification

All *Arabidopsis* genes were clustered according to protein sequence similarities. Those with fewer exons in the same cluster were selected as candidates to inspect the presence of three retropositional hallmarks: (1) lack of introns in the coding region compared with their paralogs; (2) remnant of the poly-(A) tail at the 3' end; and (3) target site duplications (TSDs; 5 bp < length < 15 bp) generated by retroposon integration. Finally, 59 retroposons from multiple exonic gene families as well as 4 genes (At3g11810, At4g03290, At5g49490, and At5g66830) from clusters of single exonic genes were identified. In addition, six single exonic singletons were confirmed by comparison with their rice (*Oryza sativa*) orthologs that have more than four exons, giving rise to a total of 69 retroposons in current *Arabidopsis* genome (Columbia ecotype [Col-0]; Table I; the detailed sequence annotation information can be found in Supplemental File 1). Fifty-one parental genes were also determined (with some parental genes created more than one retrogenes) according to phylogenetic tree topology, conservation in untranslated regions (UTRs), or synonymous substitution rate (Ks) of each retroposon with its paralogs as described in "Materials and Methods." Figure 1 shows the gene structure of a recent retroposon (At1g63760) with regard to its parental gene (At1g05890).

As evidence of the retropositional event, 16 retroposons have at least one conserved UTR longer than 50 bp. For example, At3g60610 has a 178-bp 3' UTR that shows 90.302 nucleotide identity with its parental gene as well as a 61-bp conserved 5' UTR. We identified that both At1g03300 and At1g45100 are retroposons. Of these retroposons, 67 are intronless, and only two genes (At1g63210 and At5g56720) kept one intron each that were not spliced out from pre-mRNAs. In most cases, the remnants of poly-(A) tails and TSDs were hardly detected; only 15 retroposons have both predicted poly-(A)s and TSDs, while 27 retroposons (39%) have neither of them.

Truncation of retroposons is a striking feature as a result of the low processivity of the reverse transcriptases, especially toward the 5' direction. Thirteen retroposons were truncated in coding regions in 5' direction with clear retropositional endpoints (Fig. 2A). We also identified six 3' truncated retroposons (Fig. 2B). For At1g63210, truncations toward both 5' and 3' directions were observed. Three of these 18 truncated genes kept TSDs, suggesting that the truncations might be the results of RT processes rather than genomic rearrangements subsequent to retropositions.

Another structural defect as internal sequence deletion was found. The seven internal exons of At4g38030 (from the fourth to the 10th in part, totally 1,130 bp) were completely deleted in its retroposed copy, At1g65210 (Fig. 2C).

In addition, one case of extra sequence addition was observed in At3g58390, a retroposon from the *Arabidopsis* cell division gene *pelota* (PEL1, At4g27650). It was found that a 63-bp fragment in the coding sequence (CDS) of this retroposon was derived from mitochondrial ORF159 (AtMg01050; Fig. 2D). In total, there are 20 genes that have structural traits of the RT process (e.g. truncations, deletions, and/or extra sequence additions).

Among these genes, 22 are processed pseudogenes (32%) that lost coding potential after integration, and subsequent mutations had accumulated in their ORFs that preclude their functionality. For example, sequence mutation from A to T introduced premature stop codon in At2g25500, while an insertion of a 150-bp AG rich (93%) fragment caused loss of function of At3g27720.

Most of these parental genes that gave rise to retroposons can be classified into four functional categories: 10 are involved in cell division, chromosome partition, or DNA repair; 15 are related to transcription or translation; 11 have DNA or protein binding activities; and the remaining 15 genes encoded enzymes that are involved in miscellaneous cellular processes. The average length and GC content of these retroposons in retroposed regions are 1,225 bp and 43.7%, respectively, close to the average values for the entire gene set of *Arabidopsis* (1,301 bp and 40.302, respectively; *Arabidopsis* Genome Initiative, 2000), which is in striking contrast to human retrogenes that are usually short in size with lower GC percentage (Gonçalves et al., 2000).

### Retroposon Expression

We identified potentially expressed retroposons through BLASTN search against expressed sequence tag (EST) databases that exhibited  $\geq 99\%$  similarity. Twenty-nine expressed retroposons had been detected finally. In addition, expressions of 23 retroposons were confirmed by RT-PCRs with gene specific primers, giving rise to a total of 52 expressed retroposons. Five of them (At2g01180, At2g03270, At2g36500, At3g52950, and At5g63610) even evolved into two exonic gene structures by recruiting the 5' TSD upstream sequences and obtaining canonical intron splicing signals.

Expression profiling of those retroposons and their parental genes were carried out in four tissues (root, shoot, leaf, and flower) by use of the specific primers. It turned out that most of the parental genes were actively transcribed (45 out of 51 parental genes; Table I). Interestingly, no expressions of three parental genes (At5g58340, At1g18310, and At1g69090) were detected, but all of their corresponding retroposons (At1g15720, At5g15870, and At5g66830) have evidence of expression. In addition, an obvious tendency was observed that most of the expressed retroposons were active in both roots and flowers, tissues from apical

**Table 1.** List of all identified retroposons in *Arabidopsis* genome (*Columbia* ecotype *Col-0*)

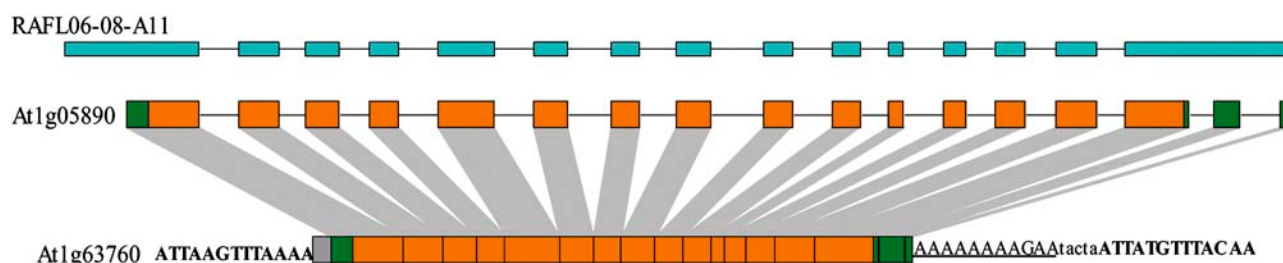
Retroposons					Parental Genes			Hallmarks of Retrogenes		
TAIR Accession	Description	EST <sup>a</sup>	Paralog <sup>b</sup>	Tree <sup>c</sup>	TAIR Accession	EST <sup>a</sup>	Identity/Ks <sup>d</sup>	Intron Loss <sup>e</sup>	TSDs <sup>e</sup>	Poly-(A) Tail <sup>e</sup>
At1g03300 <sup>f</sup>	Pseudo Agenet domain-containing protein	Z17450	2	$\beta$	At2g47230	AU237015	74%/0.302	+	-	-
At1g08250 <sup>g</sup>	Putative chloroplast prephenate dehydratase	AV826414	5	$\beta$	At3g07630	AV826197	62%/1.764	+	-	+
At1g14680	Hypothetical protein	+	1	$\alpha$	At4g09060	BX826843	78%/0.414	+	-	-
At1g15040 <sup>g</sup>	Putative glutamine aminotransferase <sup>h</sup>	T42473	2	$\beta$	At1g66860	Al618755	71%/0.943	+	-	-
At1g15720	Myb family transcription factor TRFL5 <sup>h</sup>	AU235363	1	$\alpha$	At5g58340 <sup>f</sup>	-	75%/0.397	+	-	-
At1g19810 <sup>f</sup>	Pseudo cell division control protein <sup>h</sup>	-	1	$\alpha$	At3g01610	AV523483	63%/0.679	+	-	-
At1g35440 <sup>g</sup>	Putative cyclin	-	4	$\alpha$	At1g27630	AU238537	56%/1.020	+	-	-
At1g45100 <sup>f</sup>	Pseudo poly(A)-binding protein PABP	-	2	$\alpha$	At3g10845 <sup>f</sup>	-	61%/0.297	+	+	+
At1g53345	Putative phosphohydrolase domain containing protein	AV544926	0	$\gamma$	-	-	60%/1.109	+	-	+
At1g55035	Pseudo importin $\alpha$ -subunit	-	7	$\alpha$	At3g06720	AY054164	82%/0.329	+	+	+
At1g61410 <sup>f</sup>	Pseudo double strand break repair protein (XRCC4)	Al998918	1	$\alpha$	At3g23100	AF233528	94%/0.160	+	+	-
<sup>g</sup> At1g63210 <sup>g</sup>	Pseudo chromatin structure regulator GTB1 homolog	-	1	$\alpha$	At1g65440 <sup>f</sup>	CB518248	80%/0.427	-	+	-
At1g63760	Pseudo ARIADNE-like protein ARI6 <sup>h</sup>	-	3	$\alpha$	At1g05890	AV823727	93%/0.091	+	+	+
At1g65210 <sup>f</sup>	Pseudo LG127/30-like protein	-	1	$\alpha$	At4g38030	-	78%/0.229	+	-	-
At1g66770	Putative nodulin MtN3 family protein	-	3	$\alpha$	At4g10850	AV558521	80%/0.357	+	+	-
At1g68610	Putative DUF614 domain-containing protein	+	7	$\beta$	At1g14880	AU239746	56%/0.671	+	+	+
At2g01180 <sup>f</sup>	Lipid phosphate phosphatase AtLPP1	AV824728	3	$\beta$	At3g02600	AY084915	65%/0.970	+	-	-
At2g03270	Repressor of transcription <sup>h</sup>	AV784668	0	$\gamma$	-	-	69%/1.081	+	-	+
At2g03410	Putative calcium binding protein MO25 <sup>h</sup>	+	2	$\beta$	At4g17270	CF652338	64%/0.873	+	+	+
At2g19550	Putative esterase <sup>h</sup>	+	4	$\beta$	At1g29840	BT012359	76%/0.468	+	-	-
<sup>g</sup> At2g24410 <sup>f</sup>	Pseudo Smad nuclear interacting protein DDL	+	1	$\alpha$	At3g20550	AV782440	83%/0.196	+	+	+
At2g25500 <sup>f</sup>	Pseudo Maf-like protein	+	1	$\alpha$	At5g66550	AV786815	87%/0.279	+	+	+
At2g25630 <sup>f</sup>	Pseudo $\beta$ -glucosidase	+	14	$\alpha$	At2g44450	AV554511	85%/0.401	+	+	+
<sup>g</sup> At2g27820	Putative chloroplast prephenate dehydratase	BX836869	5	$\beta$	At3g07630	AV826197	61%/1.484	+	-	-
At2g32050 <sup>g</sup>	Pseudo cell cycle control related protein <sup>h</sup>	+	2	$\beta$	At1g17130	AU237365	70%/0.864	+	-	-
At2g36500	CBS domain containing protein	AU236945	4	$\beta$	At5g50530	CK119272	64%/0.796	+	-	+
At2g45530	C3HC4-type RING finger family protein <sup>h</sup>	BP561635	0	$\gamma$	-	-	68%/0.906	+	-	+
At3g03160	Putative Erwinia induced protein	AV824551	3	$\beta$	At3g17780	AV824916	51%/1.035	+	-	-
At3g09850	D111/G-patch domain-containing protein <sup>h</sup>	Al996551	0	$\gamma$	-	-	64%/0.906	+	-	+
At3g11810	Expressed membrane protein <sup>h</sup>	BU636041	1	$\alpha$	At2g03330 <sup>f</sup>	BX838116	95%/0.037	+	+	+
<sup>g</sup> At3g22910	Ca <sup>2+</sup> -transporting ATPase, plasma membrane-type (ACA13)	+	11	$\beta$	At5g57110	AV828770	59%/1.239	+	-	-
At3g24500	Putative ethylene-responsive transcriptional coactivator	AU235871	2	$\beta$	At3g58680	Al996308	60%/0.974	+	-	+
At3g27710	ARIADNE-like protein ARI3	AV830663	3	$\beta$	At4g34370	AV825264	76%/0.427	+	-	+
At3g27720 <sup>f</sup>	Pseudo ARIADNE-like protein ARI4 <sup>h</sup>	-	3	$\beta$	At4g34370	AV825264	68%/0.414	+	-	-
At3g43250 <sup>g</sup>	Pseudo cell cycle control related protein	+	2	$\beta$	At1g17130	AU237365	70%/0.672	+	-	+
<sup>g</sup> At3g44720	Putative chloroplast prephenate dehydratase	AV517927	5	$\beta$	At3g07630	AV826197	58%/1.119	+	-	+
At3g45950 <sup>g</sup>	Step II splicing factor like protein	+	2	$\beta$	At4g37120	AV828772	83%/0.305	+	-	-
At3g52950	CBS domain containing protein	AV792536	4	$\beta$	At5g50530	CK119272	64%/0.777	+	+	+
At3g53640	Putative pre-mRNA processing factor	-	2	$\alpha$	At1g13350	AV792311	87%/0.138	+	+	+

(Table continues on following page.)

**Table 1.** (Continued from previous page.)

Retroposons					Parental Genes			Hallmarks of Retrogenes		
TAIR Accession	Description	EST <sup>a</sup>	Paralog <sup>b</sup>	Tree <sup>c</sup>	TAIR Accession	EST <sup>a</sup>	Identity/Ks <sup>d</sup>	Intron Loss <sup>e</sup>	TSDs <sup>e</sup>	Poly-(A) Tail <sup>e</sup>
At3g58390	Pseudo cell division protein pelota	+	1	$\alpha$	At4g27650	AV544568	91%/0.269	+	+	+
At3g60610 <sup>f</sup>	Pseudo snRNP-associated emb1220 protein	-	1	$\alpha$	At1g60170	AV561075	94%/0.089	+	+	+
<sup>§</sup> At3g63380	Ca <sup>2+</sup> -transporting ATPase, plasma membrane-type (ACA12)	AU238075	11	$\beta$	At5g57110	AV828770	57%/1.667	+	-	+
At4g02630	Putative protein kinase	AU236868	8	$\beta$	At4g34500	AV787745	60%/1.069	+	-	+
At4g03290	Putative calmodulin	AU227798	5	$\alpha$	At1g05990	T46471	80%/0.601	+	-	+
<sup>§</sup> At4g16580	Putative protein phosphatase 2C <sup>h</sup>	AU238779	1	$\alpha$	At5g66720	AU237910	72%/0.982	+	-	-
<sup>§</sup> At4g30300 <sup>f</sup>	Pseudo RNase L inhibitor AtNAP15	-	2	$\beta$	At4g19210	AV830986	72%/0.838	+	-	-
<sup>§</sup> At5g14900	Pseudo pre-mRNA splicing factor RNA helicase, C terminal	-	2	$\beta$	At3g62310	CK117525	81%/0.438	+	+	+
At5g15870	Putative $\beta$ -glucan elicitor receptor	+	1	$\alpha$	At1g18310 <sup>f</sup>	-	75%/0.602	+	-	-
At5g17190	Putative Erwinia induced protein	AV521320	3	$\beta$	At3g17780	AV824916	51%/0.702	+	-	-
At5g17630	Glc-6-P/phosphate translocator XPT <sup>h</sup>	AV827364	4	$\beta$	At1g61800	AV821996	56%/0.849	+	-	-
At5g18070	DNA-damage-repair/toleration protein DRT101	AV829614	0	$\gamma$	-	-	58%/1.062	+	-	-
<sup>§</sup> At5g22630	Putative chloroplast prephenate dehydratase	BP617975	5	$\beta$	At3g07630	AV826197	59%/1.509	+	-	+
<sup>§</sup> At5g23600	Pseudo tRNA splicing 2' phosphotransferase	+	1	$\alpha$	At2g45330	AA585768	95%/0.082	+	-	+
At5g26900	WD-40 repeat family CDC20 homolog	+	4	$\beta$	At4g33270	BE520538	83%/0.301	+	-	+
At5g27080 <sup>f</sup>	WD-40 repeat family CDC20 homolog	-	4	$\beta$	At4g33270	BE520538	83%/0.332	+	-	+
At5g27570	WD-40 repeat family CDC20 homolog	+	4	$\beta$	At4g33270	BE520538	84%/0.335	+	-	+
At5g28210	mRNA capping enzyme <sup>h</sup>	-	2	$\beta$	At5g01290	AV822021	73%/0.505	+	-	-
At5g37445	Pseudo spliceosome associated protein	+	1	$\alpha$	At2g32600	AV543938	93%/0.259	+	+	-
At5g42260	$\beta$ -Glucosidase	+	14	$\alpha$	At2g44450	AV554511	84%/0.395	+	-	-
At5g44640	$\beta$ -Glucosidase	+	14	$\alpha$	At2g44450	AV554511	85%/0.396	+	-	+
At5g49200	Putative zfwf protein	+	1	$\alpha$	At5g40880	CB263571	86%/0.154	+	+	+
At5g49490 <sup>f</sup>	MADS-box family protein	-	1	$\alpha$	At5g38620	-	84%/0.179	+	-	+
<sup>§</sup> At5g52090	Pseudo nonsense-mediated mRNA decay trans-acting factor Sen1	-	4	$\alpha$	At5g37150 <sup>f</sup>	-	97%/0.043	+	-	-
At5g54960	Pyruvate decarboxylase PDC2	AV824559	3	$\beta$	At5g01320	AU236252	76%/0.969	+	-	-
At5g56720	Pseudo cytosolic malate dehydrogenase <sup>h</sup>	+	2	$\beta$	At1g04410	CF652159	57%/0.911	-	-	+
At5g63070	40S ribosomal protein S15-like RPS15F	+	5	$\beta$	At5g09490	AY086958	66%/0.531	+	-	+
At5g63370	CDC2-like protein kinase	AV828884	1	$\alpha$	At1g67580	AV826526	62%/1.194	+	-	-
At5g63610	Cyclin-dependent kinase E HEN3	AU237126	0	$\gamma$	-	-	66%/1.436	+	-	-
At5g66830	Putative F-box protein <sup>h</sup>	+	1	$\alpha$	At1g69090	-	71%/0.287	+	-	+

<sup>a</sup>Accession numbers of the ESTs matched. Minus (-) indicates no EST can be matched to this gene in GenBank, and plus (+) represents the expression signal of this gene had been detected by RT-PCRs in this work. <sup>b</sup>Paralog numbers of this retroposon in Arabidopsis genome, including their parental genes. <sup>c</sup>Tree topology of this retroposon with its paralog. " $\alpha$ " tree for those parental genes in the same clade with retroposons, otherwise " $\beta$ " tree. For those orphan retroposons, " $\gamma$ " tree would be constructed. <sup>d</sup>DNA level identity and synonymous substitution rate (Ks) of each retroposon with the parental gene. For those orphan retroposons, the two values were calculated with their respective rice homologs. <sup>e</sup>Plus (+) indicates this kind of structural hallmark can be found in the retroposon, otherwise minus (-). <sup>f</sup>The original annotation for this gene was wrong and had been corrected. <sup>§</sup>Indicates 5' truncated (when in the left shoulder of retroposon) or 3' truncated (when in the right shoulder) retroposons. <sup>h</sup>Indicates those retrogenes integrated into the 3' UTRs of their neighboring genes in reverse orientation (in a tail-to-tail configurations).



**Figure 1.** Structural relationship of the retroposon At1g63760 with its parental gene At1g05890. Coding exons are shown as orange boxes, conserved UTRs are shown as green boxes, and gray box represents a 59-bp filler DNA of unknown origin between 5' TSD and 5' homology endpoint. EST match of the parental gene is shown as blue boxes and is named after the Arabidopsis full-length EST clone name. Lines between boxes indicate intron positions, and homologous regions are linked by gray lines. The putative poly(A) tail is underlined, and the TSDs are in bold letters. Not to scale.

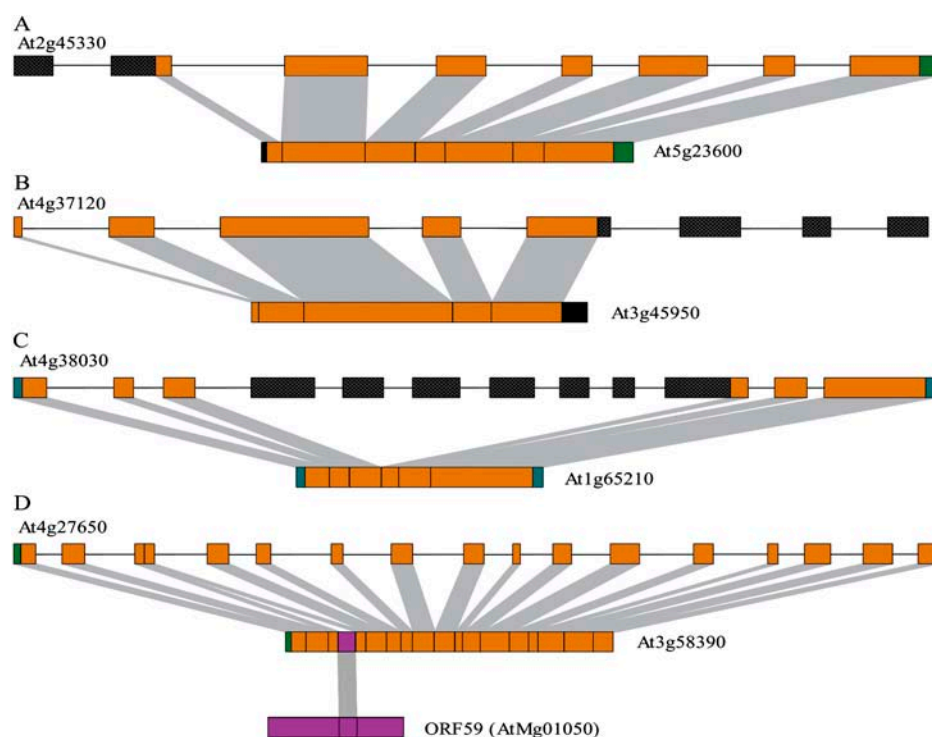
meristems. RT-PCR results of four representative retroposons were shown in Figure 3.

### Gene and Genome Organization

To illustrate the relationship between these retroposons and their parental genes, we mapped all of them on the Arabidopsis genome (Fig. 4). It seems that these retrogenes have randomly incorporated into the chromosomes, and no hotspots for insertion were detected. No genes were presented in the pericentromeric regions. A high density of retroposons in chromosome 5 (23 retroposons, 33%) and a relatively lower density of the retroposons in chromosome 4 (only four genes present) were also observed. Eleven retrogenes are located on the short arm of chromosome 5, representing one of the highest density regions. It is striking that

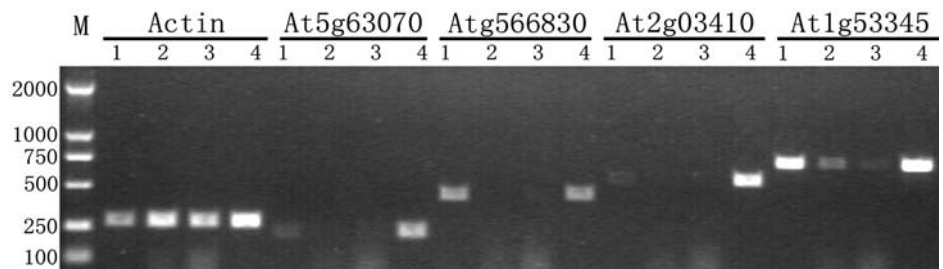
there is a paucity of genes out of chromosome 2 that has only five parental genes. An exodus scenario was also remarkable for chromosome 4 (especially the long arm), which has the highest parental gene density but lowest retroposon density. No neighboring retroposons, as those in retroposons of human high-mobility group, were identified (Strichman-Almashau et al., 2003). The distances between two nearest retroposons (At2g03270 and At2g03410) and parental genes (At5g01290 and At5g01320) are 40 kb and 8 kb, respectively.

To analyze the genomic context of integration sites, we calculated base compositions of the 250-bp flanking sequences on both sides of each retroposon. The average GC contents of the 5' and 3' flanking sequence were 37% and 33%, respectively, close to the average value of the overall GC content of the genome (34.7%;



**Figure 2.** Comparative structures between defective retroposons and their respective parental genes. A, A 5' truncation. One and one-half exons in the 5' coding region of At2g45330 were missed in its retroposon copy. B, A 3' truncation. Three and one-half exons (516 bp in length) in the 3' coding region of At4g37120 were absent in its retroposon, At3g45950. C, Internal deletion. The exons of At4g38030 from the fourth to 10th (in part) were completely deleted in At1g65210. D, Extra sequence addition. A 63-bp fragment from mitochondrial ORF59 (AtMg01050) has been incorporated between CDS position 160 and 161 of At3g58390. Black boxes with grids represent exons that were not presented in the retroposon copies, while magenta box indicates extra sequence added. Not to scale.

**Figure 3.** RT-PCR analysis of expressional patterns of 4 retroposons (At5g63070, At5g66830, At2g03410, and At1g53345) in different Arabidopsis tissues. Total RNA samples were prepared from roots (lanes marked with 1), shoots (2), leaves (3), and flowers (4). M, DL2000 molecular marker. The RT-PCR products corresponding to Actin transcripts were indicated.



Arabidopsis Genome Initiative, 2000), and no repeat sequence was detected. Only 19 TSDs were identified, ranging from 8 bp to 13 bp with an obvious compositional bias for ATs (78.0.302). At3g11810 inserted into an A-rich region, being flanked by 18 and 20 As on either side.

Even though a majority of these retroposons were in intergenic spaces, some were located in proximity of genes in a tail-to-tail configuration. Seventeen retrogenes had incorporated into the 3' UTRs of their neighboring genes in opposite orientations within 500-bp distance, resulting in overlapped transcriptions in complementary directions. For example, At1g15720 has a 79-bp complementary 3' UTR with the downstream gene (At1g15730) in opposite orientations. For some of the retroposons, portions of 3' coding regions can only be transcribed as 3' UTRs of their downstream genes in a reverse direction. No expression of At5g28210 was detected, but the expression of the 3' UTRs of its downstream neighboring gene (At5g28220) in the reverse orientation (GenBank accession no. BT004260) was identified to have a 104-bp overlap with this retroposon.

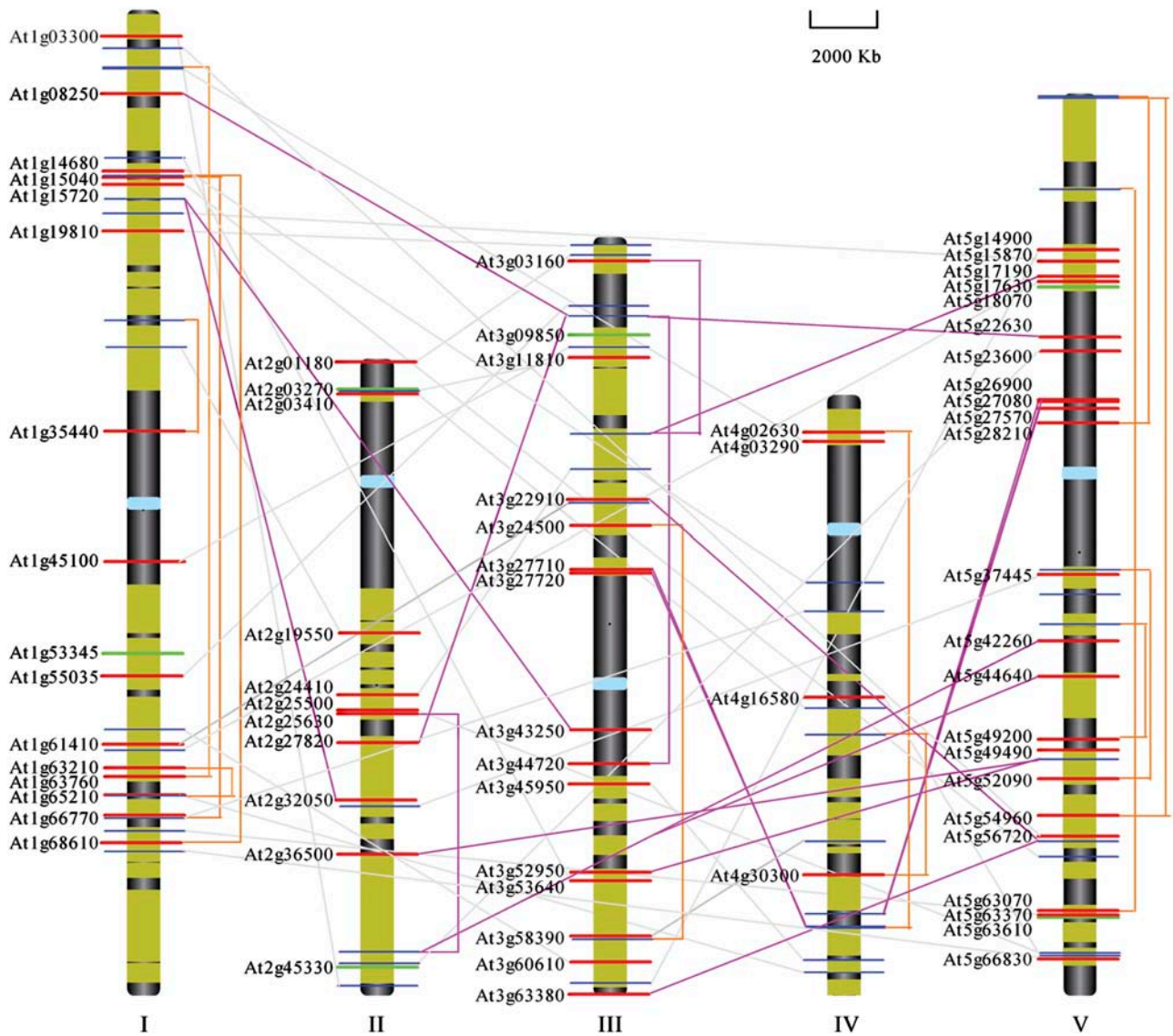
Arabidopsis has been characterized as the most extensively duplicated and reorganized genome that has three rounds of genome-wide duplications (Bowers et al., 2003; Blanc and Wolfe, 2004a, 2004b). For the 69 retroposons, 52 (75%) are located in segmental duplication blocks, among which 10 were flanked by clusters of tandemly repeated genes. Eight parental genes produced more than one retrogene, suggesting that these retroposons were subjected to tandem or segmental duplications after initial integrations. For example, At2g36500 is a segmental duplication copy of At3g52950, while At5g26900, At5g27080, and At5g27570 are tandem duplicates after retroposition from At4g33270. Retroposons of the  $\beta$ -glucosidase family (three out of 15 members) witnessed a combination of these mechanisms; a retroposed copy of At2g44450 was fixed as At2g25630, and subsequent genome-wide duplication gave birth to At5g42260, then a tandem duplicate was formed as At5g44640. Most of these retroposons (44) were from families of multiple genes. We even noticed that three retroposons might be derived from two different parental genes in the same functional gene cluster. That is that At3g27710 (and its pseudo tandem copy At3g27720) was from At4g34370 (the subgroup A of Arabidopsis

ARIADNE class of RING-finger proteins), and that At1g63760 was from At1g05890 (the subgroup B ARIADNE protein), representing the most extensive involvement of retropositions for gene family expansion (Mladek et al., 2003). No segmental or tandem duplication relationship can be found between At3g22910 and At3g63380, two retroposons from the same parental gene (At5g57110), raising the possibility that several reverse transcribed copies of the same parental gene might have been inserted independently several times through evolution.

#### Evolutionary Dynamics of Retroposition in Arabidopsis

Elucidation of the emergence pattern of retroposons will provide insight into the evolutionary dynamics of Arabidopsis. The calculated  $K_s$  of these retroposons ranged evenly from 0.037 to 1.764, implying a continuous retropositional process through evolution. Of these retrogenes, some have intronless counterparts in the corresponding genome segmental duplication blocks (Fig. 4), suggesting that the formation of them is earlier than the latest genome duplication event of about 14 to 83 million years ago (MYA; Bowers et al., 2003). At3g03160 and At5g17190 are in such case.

We also identified 13 retrogenes that have intronless counterparts in rice genome. There are two possible origins for these genes: they have inserted in the common ancestor before differentiation of the two species, or they have evolved independently during the evolution of the two species (for their high transcription levels that are prone to RTs). To discriminate the two possibilities, we searched against genomes of another monocot, maize (*Zea mays*; Palmer et al., 2003; Whitelaw et al., 2003), and another eudicot, *Medicago truncatula* (Lamblin et al., 2003). A total of 9 intronless counterparts were identified in both maize and medicago, suggesting that most of these 13 retrogenes should have been inserted before the divergence of monocots and eudicots. It is therefore believed that active promoters should be needed for them to maintain their functions through millions of years of evolution. In fact, nearly all of these genes are still actively transcribed in both species contemporarily. Rice chimeric esterase, the rice counterpart of At2g19550, should have jumped into the second intron region of its target gene, and then was fortuitously rescued by an alternative splicing mechanism

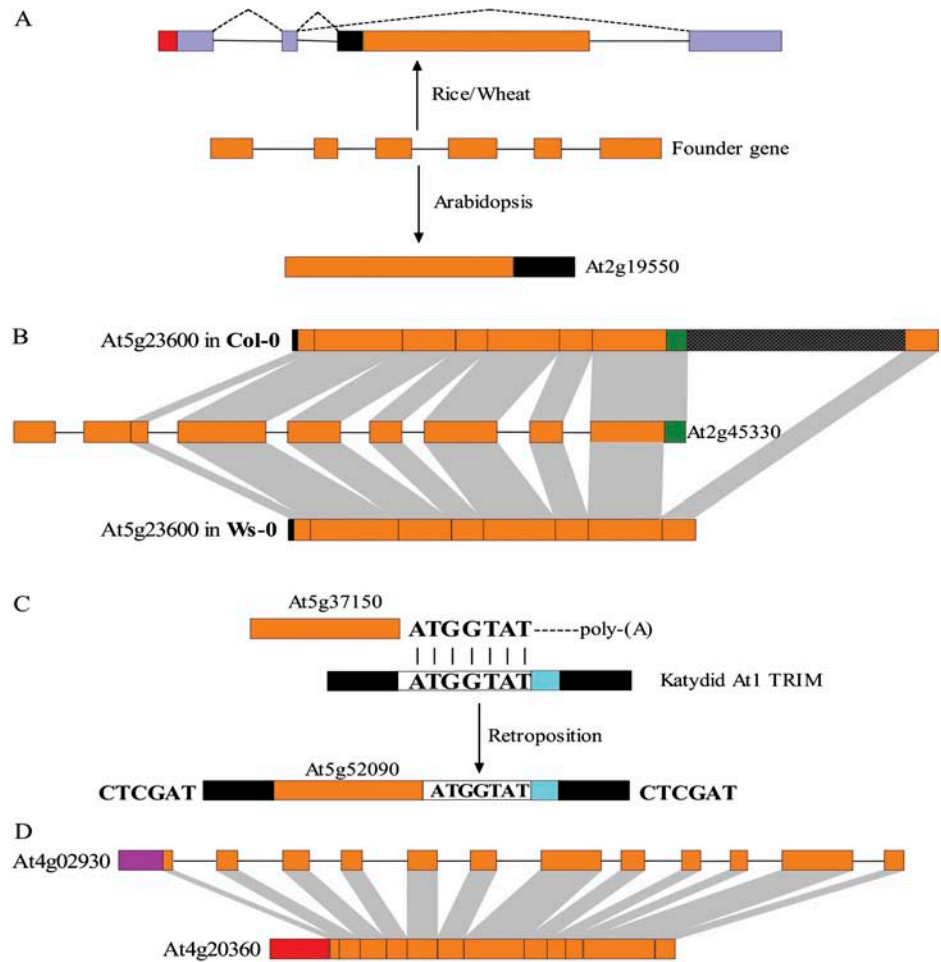


**Figure 4.** Chromosomal positions of all identified retroposons and their parental genes in the Arabidopsis genome. Retroposons are shown in bold red horizontal bars and the corresponding parental genes in blue bars. Orphan retroposons are shown in bold dark green bars. Retroposons resident on different chromosomes with their parental genes are linked by light gray lines or by orange lines if they locate within the same chromosome. Retroposons that belong to tandem or segmental duplications are linked by magenta lines. Centromeres of each chromosome are shown in light blue rectangles. The Arabidopsis whole genome duplication blocks are highlighted in dark yellow on each chromosome. Five chromosomes are indicated from number I to V. TAIR accession numbers of the parental genes were not shown for display convenience.

(Fig. 5A). Since the target gene in rice is a single copy gene encoding chloroplast 30S ribosomal protein S10 that contains a plastid transit signal in the first exon, the retroposon captured the two N-terminal exons of the target gene, resulting in chloroplast localization of fusion protein. Thus, the cytoplasm localized parental gene obtained an ectopical expression in chloroplast. The same retropositional result was observed in wheat as revealed by EST data (GenBank accession no. BE488550). This may be a functional chimeric retrogene that adapts to intercompartmental transport. Further experimental validation of functions of this chimeric gene in grasses is now underway.

Of these retroposons, eight showed higher than 90% DNA sequence similarities with their founder genes (among which five retrogenes have  $K_s$  values less than 0.1), suggesting very recent retropositional events. These genes may not be presented in other Arabidopsis ecotype strains, if only the target site sequences remain at the corresponding loci. We screened 16 additional Arabidopsis ecotypes by PCRs. It turned out that four retroposons (At1g63760, At3g11810, At3g58390, and At3g60610) were presented in all of these ecotypes, and four Arabidopsis ecotypes (Ba-1, Br-0, Ler-0, and Mt-0) contained all of these recent retroposons as Col-0, while the other 12 ecotypes lacked

**Figure 5.** A, A possible retropositional event occurred before the divergence of monocots and eudicots that had a distinct outcome. Light purple boxes represent exons of the target gene where retroposon integrated, and dashed lines linking exons indicate alternative splicing isoforms. B, Different performances across *Arabidopsis* ecotypes on one recent retroposon (At5g23600) integration. Gridded black boxes represent a 369-bp sequence identified in Col-0 that was not included in the corresponding regions of the remaining 12 ecotypes such as Wassilewskija (Ws-0), where only a 27-bp unknown sequence exists. C, An exception to the general retroposition process. At5g52090 is coreverse transcribed with a Katydid-At1 type TRIM element by a 7-bp microhomologous region between mRNA of At5g37150 and this TRIM sequence as anchor site. CTCGAT is the identified TSDs generated upon integration. D, An ambiguous intronless gene that was not retroposon. The intronless translation elongation factor EF-Tu (At4g20360) was a nuclear transfer product of the chloroplast *tufA* gene, rather than a retroposed copy of the multiple exonic paralog (At4g02930, 12 exons) that encoded the mitochondrial precursor. Red and purple boxes represent the chloroplast and mitochondrial transit peptides of genes, respectively.



some of the remaining four retrogenes (Table II). Surprisingly, we could not detect the band corresponding to At5g23600, an expressed tRNA 2' phosphotransferase pseudogene, in up to 12 ecotypes checked. Subsequent cloning and sequencing of fragments encompassing the retroposed region revealed polymorphisms in the 3' end where the reverse PCR primer was unfortunately located; in Col-0 as well as Ba-1, Br-0, Ler-0, and Mt-0, the retroposed region extended 52 bp downstream of the stop codon of the parental gene, but in the other 12 ecotypes, it ended 15 bp before the stop codon and had lost a 369-bp downstream fragment (Fig. 5B), implying different

retroposon integration performances across ecotypes. With 5.2 MYA as an upper time limit of divergence between different ecotypes of *Arabidopsis* (Koch et al., 2000), an evolutionary rate of new gene creation in *Arabidopsis* was calculated as 0.6 genes per million years (original genes from RTs rather than duplication products subsequent to retropositions) since there were three novel retroposons (At1g61410, At5g37445, and At5g52090) created after radiation of *Arabidopsis*.

At5g52090, the most recent retroposon, however, is an exception to the general retropositional process. It is not directly reverse transcribed by RTs, but coreverse transcribed with a helper sequence: a 7-bp

**Table II.** Presence of eight recent retroposons in different *Arabidopsis* ecotypes

“√” indicates that the retroposon was identified in the corresponding ecotype, while “×” indicates that no such retroposon was identified in the corresponding ecotype. Each of the 12 ecotypes has a shortened At5g23600 towards the 3' direction compared with Col-0, but it is marked with “-” for all of them.

	No-0	Est-0	Ws-0	N1	Da(1)-12	Kas-1	Gr-1	Di-0	Ge-0	En-1	Uod-1	Gel-1
At1g61410	√	√	√	√	√	×	×	√	√	√	√	√
At5g23600	-	-	-	-	-	-	-	-	-	-	-	-
At5g37445	√	√	√	√	√	√	√	×	√	√	√	√
At5g52090	√	√	√	√	√	√	√	×	×	×	×	×

microhomologous fragment between mRNA of At5g37150 (parent of At5g52090) and this helper sequence provides an anchor site for co-RT, then this sequence is integrated into the chromosome. It has been revealed that the helper sequence was a *Katydid-At1* type terminal-repeat retrotransposons in miniature (TRIM) element (Witte et al., 2001). Since this TRIM sequence lacked mobility-related coding sequences such as gag and pol domains, its retrotransposition activity was achieved by borrowing from those autonomous elements, and At5g37150 hitchhiked this TRIM sequence again (Fig. 5C). An interesting finding is that this retrogene is absent only in ecotypes from continental Europe along the Rhine river region, implying a genetic isolation of these geographical populations from the outside world in the recent geological periods.

## DISCUSSION

Even with the availability of Arabidopsis genome sequence, our knowledge about plant retropositional process is still scant because of the poor annotations of pseudogenes and the difficulties to detect highly divergent sequences. In this article, we report the identification of 69 retroposons in Arabidopsis genome by a computational approach with stringent parameters. There is perhaps an underestimation of the total number of retroposons in the genome. However, the goal of this work was not to be exhaustive but to minimize the false positive rate (retroposons confirmed by paralogs less than three exons that may result from an adjacent-exon-merge process and those ambiguous clusters of single exonic genes with no obvious poly-A relics) and to set up a blueprint of plant retroposons. In fact, most of the 69 genes are likely to be derivatives of mature mRNAs, and only two genes kept one intron each that were not spliced out from pre-mRNAs. Records of the RT process such as truncations, deletions, and extra sequence additions can be found in 20 of these genes, validating the authenticity of this identification. Characterization of retroposons provides us a unique perspective to understand genome and gene dynamics over millions of years of evolution. Four genes (At1g14680, At4g02630, At5g17630, and At5g63370) that were once assigned as segmental duplication products from multiple exonic paralogs were finally elucidated to be retroposons here. In addition, retroposon identification helped to improve sequence annotations, since original annotations for 14 retroposons and five parental genes had been corrected by correlating with retropositions.

In this work, an evolutionary rate of new gene creation by retroposition (0.6 retroposons per million years) in Arabidopsis was reached based on the presence/absence patterns of four recent retrogenes in different ecotypes. Assuming the evolutionary distance between Arabidopsis and rice as 170 to 235 MYA (Yang et al., 1999), the number of retroposons gener-

ated after that should be 98 to 136, much higher than retroposons identified in this work (only 49, 56 minus 7 duplication products). This discrepancy may have three possible origins: the massive gene loss after the latest duplication event of Arabidopsis; missed annotations of pseudogenes from retropositions; or those unsuccessful retroposons have diverged to such an extent that they are no longer recognizable.

Retroposons were in fact jumping genes like retrotransposons (Roos et al., 2004) since they were reverse transcribed. Considering the high self-fertilizing rate of Arabidopsis (approximately 99% in the natural populations; Abbott and Gomes, 1989), the recent retropositional event had recorded the evolutionary history of different populations and can be used as a cladistic marker for biogeographic research with great fidelity (Deininger and Batzer, 2002). It is well accepted that Arabidopsis across Europe has two major origins: one from Asia, another from the Mediterranean refugia as a result of the Pleistocene glacial period about 1.64 MYA (Sharbel et al., 2000; Minorsky, 2001). The Mediterranean Pleistocene refugia effect was highlighted by the revealed genetic isolation of those ecotypes of the Rhine river region from the outside world with their paucity of the most recent retroposon (At5g52090). Interestingly, we noticed from Table II that Kashmir (Kas-1), the only Indian representative ecotype in the Arabidopsis Resource Centers of both Europe and North America, has the same retroposon inventory as that of Gr-1 from Austria; one recent retrogene, At1g61410, was absent in both of the two ecotypes that also had a shortened At5g23600, suggesting that Kas-1 may have originated in Europe, not India. This substantiated a previous speculation that Kashmir is the descendant of a Europe ecotype that hitchhiked to India in a shipment of grain 150 years ago according to polymorphic patterns based on 15 microsatellite loci (Vander Zwan et al., 2000). Further investigations of the presence/absence patterns of these recent retroposons on more ecotypes will help to elucidate the effect of human and Refugia in Arabidopsis biogeographic distribution across the world.

### Criteria for Retroposon Identifications

Characterization of retroposons can be regarded as a genomic archaeological process; what you can excavate depends on what have successfully deposited in the genome through millions of years of evolution. We can no longer discern any retroposons unless structural hallmarks and reference genes (intron-containing paralogs or exactly the founder genes) were identified in the genome. However, of the 69 retroposons identified, only 38 had predicted poly-(A) tails, and 27 had neither recognizable poly-(A) tails nor TSDs. This posed the biggest obstacle to identify the authentic retrogenes. Generally, a nascent poly-(A) tail will tend to decrease in length with time due to slippage during

DNA replication, and long homopolymeric runs of As will also lead to genetic instability (Symers et al., 2002). Thus, an inverse correlation of the length of poly-(A) tail with the elements' age is obvious. This is the same with TSDs because of accumulation of random mutations through evolution. In addition, large TSDs also introduced instability for the genome by illegitimate recombination, resulting in gradual shortening of TSDs (Gilbert et al., 2002). Then, only "intronless," the most conspicuous trait, survived. This is underlined by the fact that nearly all of these retroposons have kept intronless. A bold hypothesis tells that almost the entire yeast genome, with its paucity of introns, consists of retroposons that have replaced their founder genes by homologous recombination (Fink, 1987). So it's plausible that those intronless Arabidopsis genes with no multiple exonic paralogs, at least in part, are ancient retroposons that had gradually lost their parental genes as well as structural hallmarks and cannot be recognizable any more. In fact, it has been suggested that the single exonic SET-domain containing SU(var)3-9 gene family that has eight members (Baumbusch et al., 2001), and the  $\beta$ 2 expansin family with three genes (Li et al., 2002) are all ancient retroposons.

With rice genome as a reference, six single exonic singletons were characterized as retroposons by comparing with those rice orthologs more than four exons, suggesting the existence of orphan retroposons (" $\gamma$ " type of tree topology). The assumption of this approach is that the possibility of more than three introns gained (in rice lineage) or lost (in Arabidopsis lineage) in the same orthologous gene of one lineage is nil after divergence of rice and Arabidopsis, and such situation can only be explained as a result of the RT process (Roy et al., 2003). It seemed to imply that such retrogenes had some adaptive advantages over their parental genes that were gradually eliminated from the genome, leaving those orphan genes alone. As an indication of functionality, all of these orphan retrogenes can be actively transcribed. It is conceivable that with more genomic sequences available, especially from a cofamilial Brassicaceae species, more recent functional orphan retrogenes will be identified.

However, not all intronless genes were retroposons. For example, the Arabidopsis genome has two nuclear genes that encoded the translation elongation factor EF-Tu (Fig. 5D): one multiple exonic isoform (At4g02930, 12 exons) encoded the mitochondrial precursor, and one intronless isoform (At4g20360) encoded the chloroplast precursor. Even though the overall DNA sequence similarity between them was 59%, the chloroplast precursor was not a retropositional copy of the multiple exonic gene. Molecular phylogenetic evidence has revealed that At4g20360 was in fact a nuclear transfer product of the chloroplast tufA gene in the green algal ancestor of land plants (Baldauf and Palmer, 1990), which was then retargeted specifically to the organelle from which it was donated with acquisition of a proper organelle-targeting transit pep-

ptide (Martin and Herrmann, 1998). Considering the fact that about 18% of the total Arabidopsis protein-coding genes were acquired from the cyanobacterial ancestor of plastids (Martin et al., 2002), some of the ambiguous intronless genes with multiple exonic paralogs can be traced back to the organelle transfer process rather than retropositions. In addition, some intronless genes may just represent an ancient gene structure in the "intron late" scenario. The best example was observed in the intronless chloroplast NAD-dependent malate dehydrogenase (At3g47520) that has two seven exonic paralogs (At1g53240 and At3g15020) that targeted to mitochondrion as well as two nine exonic paralogs (At2g22780 and At5g09660) that targeted to glyoxysome. Extensive analysis revealed that all of these three isoforms were derived from duplications of an ancient intronless mitochondrial isoform, and then only the chloroplast gene has kept intronless with a moderate divergence, occupying a different clade in the phylogenetic tree (Berkemeyer et al., 1998). Several tens of intronless genes were identified as ancient gene structures rather than retroposons and were not included in this work, such as 20S proteasome  $\alpha$ -subunit C (At3g22110) and the histone H3-like protein family with seven members.

### Functional Compositions of Retroposons

Logically, a gene that can generate processed copy must meet three prerequisites: (1) be highly expressed, for a higher probability to be transported into the nucleus and reverse transcribed; (2) can be expressed, at least in some stage, in the germ line cells (or apical meristem cells that finally differentiate into gamete cells in plants), so as to be fixed in the genome and successfully transmitted into the next generation; and (3) recruits an active promoter to avoid transcriptional silencing, being maintained through millions of years of evolution. This stipulates that retroposon parental genes should be widely expressed, highly conserved house-keeping genes as have been shown in this report. In fact, 45 out of 51 parental genes are highly expressed in examined tissues. These functional components reflect the nature of the retropositional process: not the sequence per se of the RNA molecule, but its expression level in gamete cells determines the probability of an RNA to be reverse transcribed. Virtually all types of mRNA are capable of retroposition (Brosius, 1999), but genes involved in transcription and translation as well as cell division are the most efficient (Venter et al., 2001). Twenty-five genes (40.302) identified in this work belong to this functional groups.

### Retroposon: A Snapshot of Gene Expression in Ancient Time

Since retroposons are derivatives of mature mRNA emerged in various evolutionary periods, they can serve as molecular fossils to reflect their parental

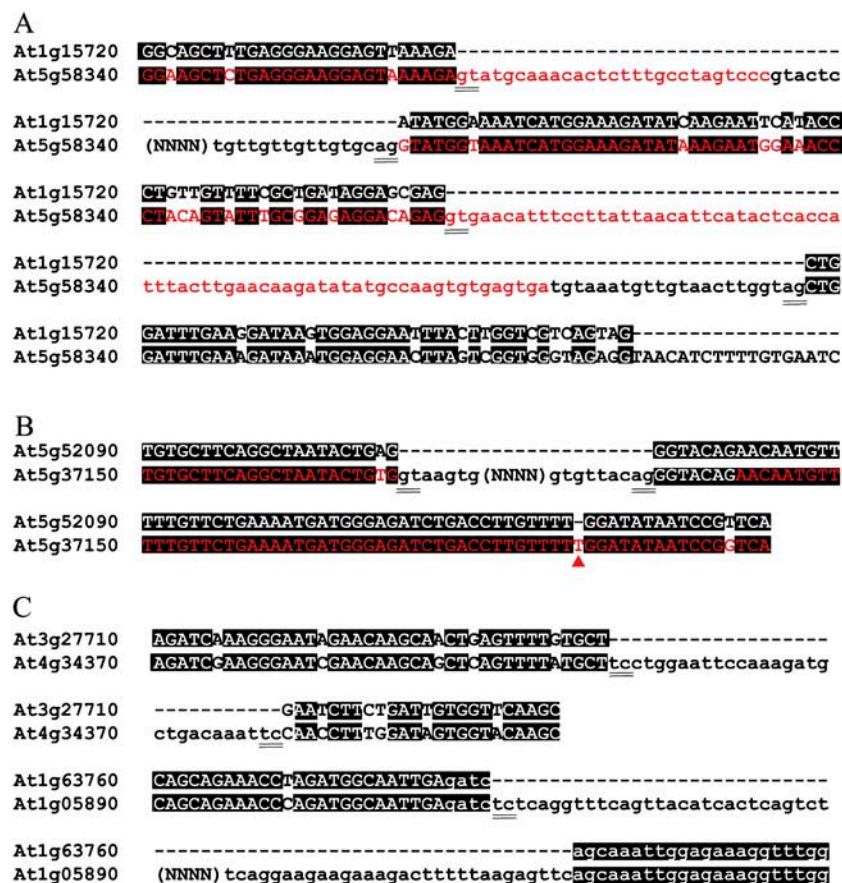
genes' expressions at the time when they are formed, and sequence comparisons will help to identify some variant splicing forms that are expressed only under specific conditions or are extinct (Strichman-Almashau et al., 2003).

For At1g45100, sequence alignment indicated the existence of a 24-bp mini-exon in the parental gene. The identification of At2g45530, a putative C3HC4-type RING-finger family protein, revealed a 63-bp exon in its rice ortholog that is missed by all gene prediction programs. This exon is confirmed by RT-PCR in rice as well as sugarcane EST data (GenBank accession no. CA202140). Comparisons between At1g15720 and At5g58340 suggested different gene structures toward the 3' direction (Fig. 6A). Since only the retroposed copy can be actively transcribed, we changed the parental gene prediction according to its fossil record of expression.

Such comparisons can also help to identify three parental genes that have been disabled after producing retroposed copies. For example, the splicing acceptor site of the second intron of At5g37150 was wrongly predicted because of one T insertion in the third exon (Fig. 6B), while the last intron of At1g18310 was indeed a defective exon as confirmed by retroposon sequence as well as rape paralog data (GenBank

accession no. AAP42646). In addition, alignments for two retroposons with their respective parental genes seemed to imply the existence of an aberrant splicing signal as TC-TC (Fig. 6C). We even noticed an extensive exon reshuffling in At1g13350 toward the 5' direction after retroposition.

Such comparisons also revealed changes in 3' UTRs through evolution. Of the 18 retroposons that showed higher than 80% sequence similarities in retroposed regions with their respective expressed founder genes, there are seven genes whose retropositional 3' end points matched well with the parental genes' transcriptional stop sites, suggesting the transcriptional patterns of those parental genes have not been changed after retroposon creations. But for the remaining nine genes (except for two genes that were truncated toward 3'), extensions of the 3' UTRs of the parental genes were striking, from 90 to 300 bp according to EST data. These changes may imply possible alternative poly-(A) signals, but efforts to clone these parental mRNAs similar to retroposons failed. When all retroposons were checked, 28 extensions on 3' UTRs out of 51 founder genes were observed, suggesting that extension of 3' UTR, which plays a crucial role in posttranscriptional regulation of gene expression by modulating nucleocytoplasmic



**Figure 6.** Retroposons as molecular fossils of the parental genes' expression in ancient time. A, Alignment helps to change the parental gene prediction according to its expressed retroposon. B, Alignment indicates the parental gene has been disabled by one base insertion after retroposition. Red triangle indicated the inserted nucleotide. C, Alignments imply the existence of an aberrant intron splicing signal as TC-TC. Identical nucleotides are shaded, and original exon predictions of parental genes are in red letters. Intron sequences are in lowercase, the splicing donor and acceptor sites are underlined, and (NNNN) are used for display convenience.

mRNA transport, translation efficiency, subcellular localization, or mRNA stability, may be a common mechanism of plant transcriptome evolution.

### Retroposition as a Novel Force of Genome Evolution

Retroposition represents a reverse flow of genetic information via RNA intermediates. It is a shot-gun approach of the genome to achieve functional innovation and thus evolutionary diversities by mixing and matching coding sequences with novel regulatory elements. Generally, the low proportion of coding region in a genome minimizes the chance of a successful retroposition event. In fact, considering the insertional mutagenesis nature of retroposons, those genes inserted into intragenic regions (especially inside exons) were prone to purification through evolution, leaving those intergenic retroposons (especially to a close proximity of genes) aside. As a result of the retropositional process, retroposons usually have traits of structural defects (such as truncations, deletions, additions, and premature stop codons) that preclude their functionality. Nevertheless, decisive advantages are also clear for retroposition over segmental duplication for its nature of juxtaposing already existent coding sequence with a different regulatory element. For segmental duplications, subsequent changes in the corresponding regulatory elements are required so as to generate different control regions and alter temporal or spatial expression patterns. In this point, retroposition has been viewed as sowing seeds of evolution for new gene origination, rather than just representing an evolutionary dead end (Brosius, 1991; Long et al., 2003).

In this work, we noticed some expressed retroposons that had changed their target signals through retroposition, and such intersignal exchange may have considerable impacts. For example, At4g16580 had lost the original chloroplast transit peptide because of 5' truncation after retroposition, while both retroposon and the parental gene have kept coding potential and transcriptional activities. We even identified one rice chimeric gene that had acquired a chloroplast transit peptide by capturing two additional exons; thus, the cytoplasmic localized protein got an innovated function in chloroplast.

We calculated the ratio of nonsynonymous substitution rate ( $K_a$ ) to  $K_s$  ( $K_a/K_s$ ) of each retroposon with its parental gene as an indicator of selective constraints on the new retroposed gene. It is generally believed that pseudogenes evolved neutrally with no selective pressures, giving a  $K_a/K_s$  value of 1 ( $K_a = K_s$ ; Torrents et al., 2003). However, it turned out that the ratio is significantly smaller than 1 for most of these retroposons identified here with no frame disruptions in coding regions, suggesting that some of these retrogenes may still be functional. In fact, 75% of these retroposons had kept transcriptional activities, especially in tissues from apical meristems (roots and flowers), after incorporation back into the chromo-

some, implying that they may functionally active after retropositions.

In addition, 11 out of 22 processed pseudogenes have evidence of transcriptions. Recently, a new functional role of such expressed pseudogenes was revealed as ncRNA by regulating mRNA stability of its homologous parental gene in coding region (Hirotsume et al., 2003), implying a possible function of those genes. Except for protein-coding genes, there may also be many functionally important noncoding RNA genes that can be traced back to retropositional origins, such as BC1 RNA gene in rodents and BC200 RNA gene in primates (Martignetti and Brosius, 1993). Then scrutiny on those noncoding RNAs can also provide valuable information about functional impacts of retropositions on plant genomes. However, it is hard to achieve significant statistical powers with only 69 retroposons present in the Arabidopsis genome, about 0.2% of the total protein coding genes. To get a further understanding of plant retroposons, an in vitro system that mimics the retropositional process is needed, as has been done in human (Gilbert et al., 2002; Symers et al., 2002).

## MATERIALS AND METHODS

### Identification Strategies and Bioinformatics Methods

The Arabidopsis (*Arabidopsis thaliana*) genomic sequence and the total gene set were retrieved from Munich Information Center for Protein Sequences (MIPS; <http://ftpmips.gsf.de/cress>), and exon/intron structure of each gene was derived from alignment of each CDS with genomic sequence using AAT package (Huang et al., 1997). Then each protein was used as a query to search against the whole protein set with BLAST (Altschul et al., 1997). All of these proteins were then classified into different groups with a threshold of 66% length match and  $E = 10^{-20}$  groups of mobile elements were discarded. For genes of the same group, exon/intron structures in coding regions were compared, and those genes with less if any exons were selected as candidates for further annotation; occurrence of introns outside the coding regions was also included. For each candidate gene, genomic sequence was inspected for presence of retropositional hallmarks. Candidates that located on the same segmental duplication blocks (data from MIPS at [http://mips.gsf.de/proj/thal/db/gv/rv/rv\\_frame.html](http://mips.gsf.de/proj/thal/db/gv/rv/rv_frame.html)) or tandem duplication regions with their parental genes were filtered. Intronless genes with paralogs of less than three exons were discarded, and UTRs of a retroposon were aligned with its candidate parental gene for confirmation. For those single exonic gene groups, presence of a poly-(A) (longer than 20 bp) in the vicinity of the 3' terminal within 500 bp was taken as retroposons. Considering the propensity of retroposons to be pseudogenes, those intronless pseudogenes were checked manually case by case. For those single exonic singletons, gene structures were compared with their orthologs in rice genome (sequence from RGP at <ftp://ftp.dna.affrc.go.jp/>). Phylogenetic trees were constructed based on CDS with ClustalW (Higgins et al., 1994) for confirmation of parental genes. We can easily figure out the parental gene as homolog in the same clade with retroposon ("a" type tree topology); otherwise, the presence of a conserved UTR was taken as an indicator of paternity ("β" type tree topology), and when no UTR conservation is found, parental gene is arbitrarily determined as the one with smaller synonymous substitution rate, or  $K_s$  (Betrán et al., 2002). Ratios of  $K_a/K_s$  were calculated for retroposons with their parental genes in retroposed region with K-estimator (Comeron, 1999), and subcellular locations of them were predicted by TargetP (Emanuelsson et al., 2000). The intronless counterparts of those 13 conserved Arabidopsis-Rice retroposons in maize (*Zea mays*) and medicago (*Medicago truncatula*) were searched with TBLASTN against TIGR maize ([http://tigrblast.tigr.org/tgi\\_maize/index.cgi](http://tigrblast.tigr.org/tgi_maize/index.cgi)) and medicago (<http://tigrblast.tigr.org/er-blast/index.cgi?project=mtbe>) genomic sequence database, respectively.

For each retroposon, exact homologous endpoints of matched region were derived from sequence alignment, and the presence of a poly-(A) tail was checked in a 500-bp region downstream of the 3' endpoint. A poly-(A) tract was defined as  $\geq 8$  As out of a moving window of 10-bp. When a poly-(A) was found, 50-bp downstream of it was aligned with the 250-bp sequence upstream of the 5' endpoint, and TSDs were defined as  $\geq 90\%$  identical matches in a 7-bp or longer tract (Strichman-Almashau et al., 2003). When no poly-(A) is present, two 250-bp sequence segments outside each endpoint were aligned in a 10-bp moving window to find TSDs. To reduce ambiguity, TSDs were only identified for those retrogenes with predicted poly-(A) tails or those that showed higher than 80% sequence identities with their parental genes. The integration sites were characterized by calculating the base-pair composition of flanking 250-bp sequences on either side of a retroposon, and repetitive elements were filtered by RepeatMasker (Jurka, 2000). Truncation was defined when more than one exon was missed in either direction.

## Polymerase Chain Reactions

For DNA PCRs on different *Arabidopsis* ecotypes, primers on conserved regions of the eight recent retroposons and their parental genes were designed. Total genomic DNA was prepared from light-grown *Arabidopsis* seedlings with DNeasy Plant Mini kit (Qiagen, Valencia, CA). For RT-PCRs, gene specific primers were designed for each retroposon and their parental genes. RNAs were extracted from tissue samples from roots, shoots, leaves, and flowers with the RNeasy Plant Mini kit (Qiagen). First strand cDNA was synthesized by SuperScript II RT (Invitrogen, Carlsbad, CA) at 42°C for 1 h, and PCR was carried out by Taq DNA polymerase (TaKaRa Biotechnology, Dalian, China) with the following program: an initial 95°C for 1 min followed by 30 cycles of 95°C for 30 s, 56°C for 30 s, and 72°C for 1 min. Considering the intronless nature of retroposons, total RNA was digested by RNase-free DNase to avoid any possible DNA contamination and ran controls including mRNA without being reverse transcribed. Information about PCR primers and *Arabidopsis* ecotypes used can be found in the supplemental tables.

## ACKNOWLEDGMENTS

We thank Dr. Hairong Wang for help in data collection and Yanlei Fu for taking care of *Arabidopsis* plants. We also thank Professor Daoxiu Zhou (Université Paris-sud XI) and two anonymous reviewers for their constructive suggestions on manuscript revision. We are greatly indebted to the *Arabidopsis* Biological Resource Center (ABRC) for providing seed stocks of different *Arabidopsis* ecotypes.

Received January 26, 2005; revised March 14, 2005; accepted March 17, 2005; published May 27, 2005.

## LITERATURE CITED

- Abbott RJ, Gomes MF (1989) Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity* **62**: 411–418
- Altschul SE, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402
- Baldauf SL, Palmer JD (1990) Evolutionary transfer of the chloroplast tufA gene to the nucleus. *Nature* **344**: 262–265
- Baumbusch LO, Thorstensen T, Krauss V, Fischer A, Naumann K, Assalkhou R, Schulz I, Reuter G, Aalen RB (2001) The *Arabidopsis thaliana* genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes. *Nucleic Acids Res* **29**: 4319–4333
- Berkemeyer M, Scheibe R, Ocheretina O (1998) A novel, non-redox-regulated NAD-dependent malate dehydrogenase from chloroplasts of *Arabidopsis thaliana* L. *J Biol Chem* **273**: 27927–27933
- Betrán E, Thornton K, Long M (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Res* **12**: 1854–1859
- Blanc G, Wolfe KH (2004a) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**: 1667–1678
- Blanc G, Wolfe KH (2004b) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**: 1679–1691

- Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438
- Brosius J (1991) Retroposons: seeds of evolution. *Science* **251**: 753
- Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**: 115–134
- Brosius J (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118**: 99–116
- Cameron JM (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15**: 763–764
- Deininger PL, Batzer MA (2002) Mammalian retroelements. *Genome Res* **12**: 1455–1465
- Drouin G, Dover GA (1987) A plant processed pseudogene. *Nature* **328**: 557–558
- Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**: 1005–1016
- Fink GR (1987) Pseudogenes in yeast? *Cell* **49**: 5–6
- Gilbert N, Lutz-Prigge S, Moran JV (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**: 315–325
- Gonçalves I, Duret L, Mouchiroud D (2000) Natural and structure of human genes that generate retrospseudogenes. *Genome Res* **10**: 672–678
- Harrison PM, Echols N, Gerstein MB (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res* **29**: 818–830
- Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M (2003) Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res* **31**: 1033–1037
- Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680
- Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**: 91–96
- Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. *Genomics* **46**: 37–45
- Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**: 418–420
- Koch M, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of the chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis* and related genera. *Mol Biol Evol* **17**: 1483–1498
- Lamblin AF, Crow JA, Johnson JE, Silverstein KA, Kunau TM, Kilian A, Benz D, Stromvik M, Endre G, VandenBosch KA, et al (2003) MtDB: a database for personalized data mining of the model legume *Medicago truncatula* transcriptome. *Nucleic Acids Res* **31**: 196–201
- Li Y, Darley CP, Ongaro V, Fleming A, Schipper O, Baldauf SL, McQueen-Mason SJ (2002) Plant expansins are a complex multigene family with an ancient evolutionary origin. *Plant Physiol* **128**: 854–864
- Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875
- Martignetti JA, Brosius J (1993) BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. *Proc Natl Acad Sci USA* **90**: 11563–11567
- Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol* **118**: 9–17
- Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* **99**: 12246–12251
- Minorsky PV (2001) The hot and the classic. *Plant Physiol* **126**: 471–472
- Mladek C, Guger K, Hauser M-T (2003) Identification and characterization of the ARIADNE gene family in *Arabidopsis*. A group of putative E3 ligases. *Plant Physiol* **131**: 27–40
- Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. *Science* **302**: 2115–2117

- Rogers J** (1983) Retroposons defined. *Nature* **301**: 460
- Roos C, Schmitz J, Zischler H** (2004) Primate jumping genes elucidate strepsirrhine phylogeny. *Proc Natl Acad Sci USA* **101**: 10650–10654
- Roy SW, Fedorov A, Gilbert W** (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci USA* **100**: 7158–7162
- Sharbel TF, Haubold B, Mitchell-Olds T** (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol Ecol* **9**: 2109–2118
- Strichman-Almashau LZ, Bustin M, Landsman D** (2003) Retroposed copies of the HMG genes: a window to genome dynamics. *Genome Res* **13**: 800–812
- Symers DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke J** (2002) Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**: 327–338
- The Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815
- Torrents D, Suyama M, Zdobnov E, Bork P** (2003) A genome-wide survey of human pseudogenes. *Genome Res* **13**: 2559–2567
- Vander Zwan C, Brodie SA, Campanella JJ** (2000) The intraspecific phylogenetics of *Arabidopsis thaliana* in worldwide populations. *Syst Bot* **25**: 47–59
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al** (2001) The sequence of the human genome. *Science* **291**: 1304–1351
- Whitelaw CA, Barbazuk WB, Perlea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, et al** (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302**: 2118–2120
- Witte C-P, Le QH, Bureau T, Kumar A** (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc Natl Acad Sci USA* **98**: 13778–13783
- Yang Y-W, Lai K-N, Tai P-Y, Li W-H** (1999) Rate of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J Mol Evol* **48**: 597–604
- Zhang Z, Harrison PM, Liu Y, Gerstein M** (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13**: 2541–2558