J Mol Evol (2009) 68:80–89
DOI 10.1007/s00239-008-9190-5

The Contribution of Transposable Elements to Expressed Coding Sequence in *Arabidopsis thaliana*

Steven Lockton · Brandon S. Gaut

Received: 29 June 2008 / Accepted: 2 December 2008 / Published online: 3 January 2009

Abstract The goal of this study was to assess the extent to which transposable elements (TEs) have contributed to protein-coding regions in *Arabidopsis thaliana*. To do this, we first characterized the extent of chimeric TE-gene constructs. We compared a genome-wide TE database to genomic sequences, annotated coding regions, and EST data. The comparison revealed that 7.8% of expressed genes contained a region with close similarity to a known TE sequence. Some groups of TEs, such as *helitrons*, were underrepresented in exons relative to their genome-wide distribution; in contrast, *Copia*-like and En/Spm-like sequences were overrepresented in exons. These 7.8% percent of genes were enriched for some GO-based functions, particularly kinase activity, and lacking in other functions, notably structural molecule activity. We also examined gene family evolution for these genes. Gene family information helped clarify whether the sequence similarity between TE and gene was due to a TE contributing to the gene or, instead, the TE co-opting a portion of the gene. Most (66%) of these genes were not easily assigned to a gene family, and for these we could not infer the direction of the relationship between TE and gene. For the remainder, where appropriate, we built phylogenetic trees to infer the direction of the TE-gene relationship by parsimony. By this method, we verified examples where TEs contributed to expressed proteins. Our results are undoubtedly conservative but suggest that TEs may have contributed small protein segments to as many as 1.2% of all expressed, annotated *A. thaliana* genes.

Keywords *Arabidopsis thaliana* · Transposable element · Chimera · Expressed · Paralogues · Gene family

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9190-5) contains supplementary material, which is available to authorized users.

S. Lockton · B. S. Gaut
Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA

S. Lockton (✉)
Department of Integrative Biology, University of California, 3060 Valley Life Sciences Building, Berkeley, CA 94720-3140, USA
e-mail: slockton@gmail.com

Introduction

Transposable elements (TEs) are a ubiquitous feature of plant genomes. In maize, for example, TEs comprise 60–80% of the genome (SanMiguel et al. 1996; Messing et al. 2004). The proportion is lower, but still substantial, in compact genomes like those of rice and *Arabidopsis thaliana*. TEs represent 29% of the rice genome (Messing et al. 2004) and 10% of the 125-Mb *Arabidopsis* genome (*Arabidopsis* Genome Initiative 2000). TEs are traditionally categorized into two groups based on their mode of transposition. Class I elements, or retrotransposons, copy and paste to a new location via an RNA intermediate, which then reintegrates into the genome at a new location after reverse transcription. Class II elements are DNA transposons. DNA transposons excise out of their chromosomal location as DNA and reinsert elsewhere in the genome. Maize and other grasses contain predominantly class I elements. In contrast, class I TE activity is apparently suppressed in *Arabidopsis* (Wright and Voytas 1998), with DNA transposons approximately equaling retrotransposons in copy number (Wright and Voytas 1998; *Arabidopsis* Genome Initiative 2000).

The roles of TEs in genome evolution are varied (Le Rouzic et al. 2007) but many are harmful to genome

function. Common examples include insertional inactivation of genes (Greene et al. 1994) and DNA rearrangement via ectopic recombination (Kazazian 2004; Bennetzen 2005). Nonetheless, a subset of TE-mediated events is adaptive: in *Drosophila*, for example, TE insertions have contributed to enhanced insecticide resistance, either by affecting gene expression or by changing gene structure (Schlenke and Begun 2004; Aminetzach et al. 2005). Similarly, a "domesticated" TE-derived transposase domain contributed directly to two vertebrate proteins, RAG1 and RAG2, that are central to the immune system of jawed vertebrates (Kapitonov and Jurka 2005). The insertion of TE sequence fragments into open reading frames (ORFs) of vertebrate genes may be a general phenomenon (Nekrutenko and Li 2001). Consistent with this conjecture, TEs share sequence similarity with thousands of human protein-coding sequences (Britten 2006), many of which remain functional (Wu et al. 2007).

The contribution of TEs to plant genes is not yet clear, but some TE-based phenomena have been well documented. For example, reverse transcription of mRNA transcripts by class I transposons has generated more than 1000 retroposed genes in rice, many of which have recruited exons from flanking regions to produce functional genes (Wang et al. 2006). TEs also capture and shuffle gene fragments (Jiang et al. 2004; Brunner et al. 2005; Lai et al. 2005). Maize *helitrons*, for example, capture and move gene fragments to the extent that ~20% of genes (or gene fragments) differ in location between two maize lines (Lai et al. 2005; Morgante et al. 2005). Additionally, in *A. thaliana*, *helitrons* proliferated after the acquisition of exon fragments (Hollister and Gaut 2007). Many of the gene fragments captured by TEs are expressed (Jiang et al. 2004; Brunner et al. 2005; Lai et al. 2005), fueling speculation that TE-mediated gene shuffling can lead to novel genes. While it is clear that TEs can capture gene fragments, there are few direct examples that TE sequences have contributed to functional plant genes. One exception is the domestication of a *hAT*-like transposase by the *DAYSLEEPER* gene in *Arabidopsis* (Bundock and Hooykaas 2005), but the genome-wide extent of TE incorporation into functional genes remains unknown.

Plant genomes possess not only TEs but also an abundance of gene duplications. Duplicated genes provide functional redundancy, a potential template for evolutionary innovation and a comparative context to infer the incorporation of TE-like sequence in individual genes (Gotea and Makalowski 2006). Plants are a particularly rich system in this respect. All plant genomes studied to date exhibit evidence of ancient whole-genome duplication events in their evolutionary past, including relatively small genomes like that of *A. thaliana* (Adams and Wendel 2005). In *Arabidopsis*, duplicated chromosomal regions retain ~25% of their genes as duplicates (Blanc et al. 2003), and a similar proportion of *Arabidopsis* genes (~16%) have been duplicated as a result of local, tandem duplication events (Zhang and Gaut 2003). An important consequence of this extensive duplication is large gene families. Plants possess more gene families—with more members per gene family, on average—than other eukaryotes (Lockton and Gaut 2005).

In this study we exploit gene family data from *Arabidopsis* to assess the possibility that TEs have contributed to expressed peptides. To achieve this, we search for TE-related sequences in expressed sequence tags (ESTs) of *Arabidopsis* and verify that the TE-related sequence had a genomic counterpart in an annotated protein-coding region. However, there is an inherent difficulty with this approach: when there is clear evidence that a TE is homologous to a portion of a coding sequence, it is difficult to discriminate whether the TE *contributed* to the coding region or *acquired* the coding fragment, as commonly occurs with *helitrons* and other TEs. To address this uncertainty, we examine gene family data. In the comparative and phylogenetic context of gene families, one can use parsimony arguments to infer whether a subset of gene family members contains a unique insertion consistent with contribution from a TE. We find evidence for TE homology to expressed regions in more than 2000 genes and demonstrate that TE insertion events have led to the formation of TE-gene chimeras.

Materials and Methods

Genomic, EST, and TE Sequences

We downloaded three types of *Arabidopsis* sequences from TAIR (The *Arabidopsis* Information Resource; http://www.arabidopsis.org). All three sequence types were based on *Arabidopsis thaliana* genome release 8. The first type was genomic "Seq" gene sequences, which consist of 5′ and 3′ untranslated regions (UTRs), introns, and exons; the second was coding sequence (CDS, or exon-only sequences); and the third was peptide sequences. The Seq data contained 30,271 annotated sequences; the CDS and protein data each contained 29,161 sequences (Fig. 1). In addition, we downloaded *A. thaliana* UniGenes. UniGene Build No. 49 was downloaded via NCBI's Entrez Web site. Our database consisted of 25,693 UniGene sequences.

Our TE database was comprised of sequences derived from a BLASTn query (1e-20 cutoff, no repeat filtering) against the *A. thaliana* release 5 genome. TE queries were tabulated from three sources: (i) TEs described in a previous survey of 17 Mb of the *A. thaliana* genome (Le et al. 2000); (ii) *A. thaliana* TEs found in TIGR's repeat

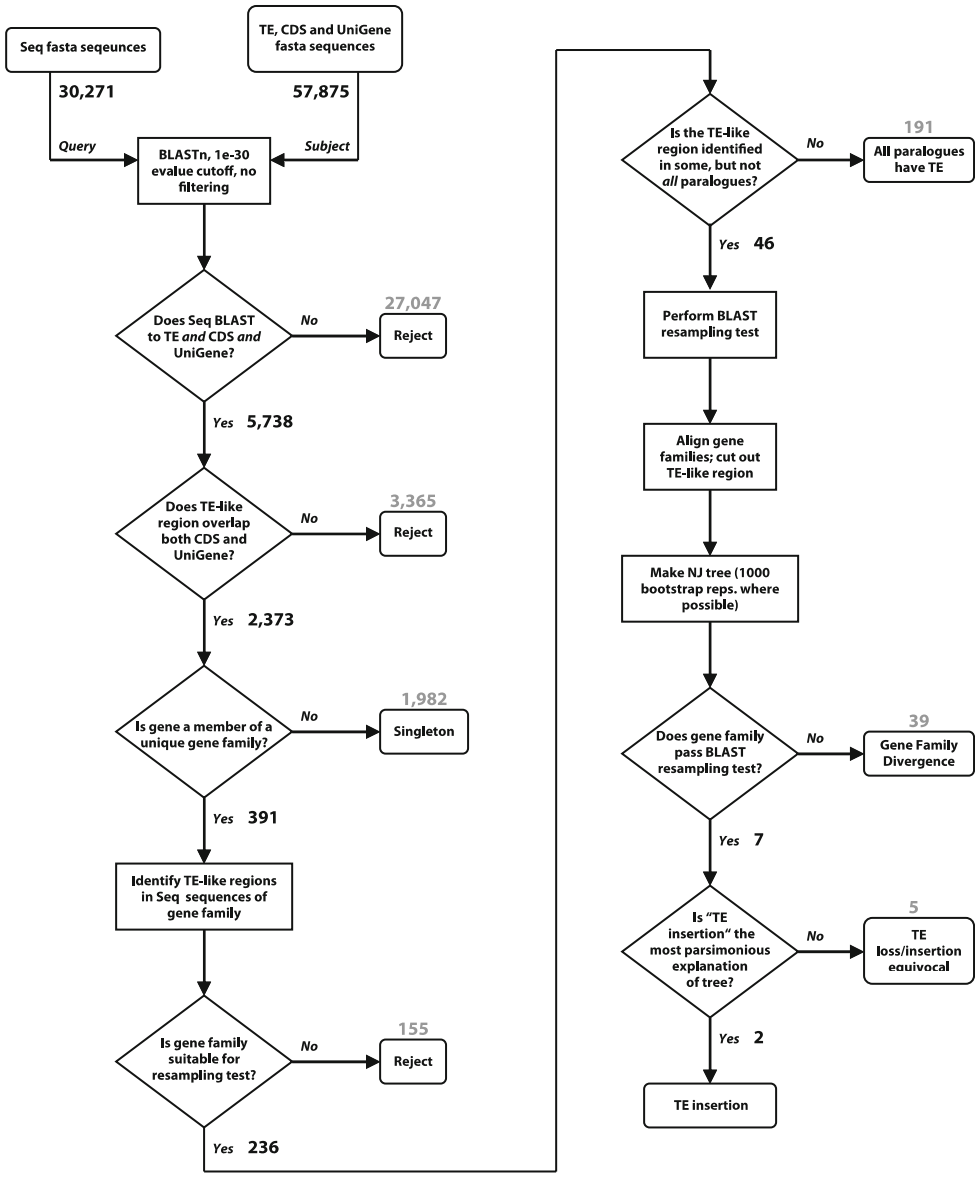

Fig. 1 Flowchart giving an overview of the methods used in this study: 30,271 Seq sequences were queried against a BLAST subject database of 57,875 TE, UniGene, and CDS sequences. Subsequently, the number of individual BLAST alignments was whittled down based on different criteria: gray numbers represent the number of BLAST alignments rejected from further analysis at each step

database; and (iii) all GenBank ORFs annotated as transposase-related in the *Arabidopsis* genome. Our final TE database consisted of 3079 nonredundant TE sequences in the *Arabidopsis thaliana* genome. The TE sequences ranged in length from a 65-base *mariner*-like TE fragment to a 15.8-kb MULE. The mean length of our TE sequences was 1134 bases.

Candidate Identification

To identify genes that consist in part of TE-like sequence, we implemented a decision tree based on a BLAST search among TE, UniGene, CDS, and genomic sequences (Fig. 1). In this initial BLAST, the TE, CDS, and UniGene FASTA sequences were combined into a single database. The Seq file was used as the query in a tBLASTx (Altschul et al. 1997), with repeat filtering off, against the TE, CDS, and UniGene subject database.

BLAST results were parsed to find Seq sequences that hit, at an e-value <1e-10, all three types of sequences (TE, CDS, and UniGene) in the subject database. These data were further parsed to find BLAST alignments in which all three types of subject sequences aligned to a common region of the Seq sequence, so that a TE was found to overlap expressed (UniGene), exonic (CDS) sequence. Seq sequences that did not meet this criterion were not studied further. Each comparison among sequence types provided some information. For example, the BLAST alignment between Seq and Unigene confirmed that the UniGene was not an EST cloning artifact and confirmed gene expression; the alignment among Seq, UniGene, and CDS confirmed exon/intron boundaries; the alignment of TE with

UniGenes confirmed expression of a TE-like sequence; and the alignment of TE with annotated Seq data confirmed that the TE-like region is found in genomic sequence.

Gene Ontology Analysis

For each genomic Seq query that successfully hit a TE, CDS, and UniGene sequence, we assessed its classification according to Gene Ontology (GO) (Ashburner et al. 2000) to determine if any biases in molecular function existed. The *A. thaliana* GO Slim database (Berardini et al. 2004) was downloaded from TAIR, and only entries that corresponded to the function of our genes of interest were parsed. These data were compared to the distribution of GO Slim functional categories for the whole genome using 2×2 chi-square contingency tables.

Gene Family Identification and Evolution

When there is homology between a TE and a coding region, one cannot infer the direction of the TE event. Did the TE contribute to the gene or, conversely, did the TE acquire a copy of the gene fragment? To address this question, we used gene family data. The phylogenetic distribution of the TE on a gene family should allow one to distinguish TE insertion from acquisition, using parsimony arguments. For these analyses, we relied on the *Arabidopsis* high-stringency gene family data set of Rizzon et al. (2006). These gene families were defined by a homology criterion of pairwise identities $\geq$50% over $\geq$90% of the peptide sequence; paralogues were grouped using the single-linkage criterion, resulting in 10,542 genes clustered into 3544 gene families.

For each gene family, we took the following steps. We first assembled each gene family as a FASTA file of Seq genomic sequences and then identified the location of the TE-like region in the originally identified "TE gene" from the initial BLAST. Each Seq sequence represents an unspliced genic region (including introns, exons, and UTRs), as found on the chromosome. Because TE activity takes place at the chromosomal level, we aimed to identify TE-like regions in Seq sequences. We used tBLASTx to compare the region of strong TE homology to the Seq sequence of all other paralogues in an attempt to identify further TE-like regions. Each resulting e-value was recorded. As sequence divergence among paralogues is a confounding factor in the identification of TE insertions in gene families, we devised a BLAST resampling procedure to determine if TE-like regions were atypical relative to the other genic regions. To do this, 100 coding sequence fragments were randomly chosen from the gene family member that was originally identified as containing a TE-like region. These random fragments were the same length as the TE-like region but did not overlap with it. Each fragment was used as a tBLASTx query against the entire gene family, using identical criteria as in the initial BLAST. Coding sequence fragments that hit a paralogue at an e-value less than the previous TE-region tBLASTx value for that same paralogue were considered a successful hit and recorded. Paralogues with more successful hits had higher BLAST resampling scores. In summary, high blast resampling values indicated that the non-TE regions were more similar among paralogues than the TE region, suggesting that the putative TE insertion into protein-coding regions represented a region of aberrant sequence evolution.

We subsequently aligned all peptides within gene families using T-Coffee (Notredame et al. 2000) with default parameters. The TE-like regions of the genes were excluded from the aligned peptides, as these could bias both phylogenetic analyses and our inferences. Alignments were visually inspected and hand-adjusted, then employed to construct Poisson-corrected neighbor-joining trees with 1000 bootstrap replications, using MEGA v3.1 (Kumar et al. 2004).

Results and Discussion

Exon Sequences Containing TE-Related Fragments

We queried a database of 57,875 TE, CDS, and UniGene sequences with 30,271 genomic Seq sequences. Of the 30,271 BLAST queries, 5738 hit to all three types of sequence in the subject database (Fig. 1). These 5738 results were further parsed to find BLAST alignments in which TE, CDS and UniGene sequences overlap, aligning to the same region of the Seq sequence—2373 alignments passed this criterion, leaving 3365 to be rejected. None of the 2373 alignments involved genes functionally annotated as a TE or a pseudogene. Of the rejected alignments, in 2472 cases only the TE hit the Seq query, with no evidence of exon overlap or gene expression, suggesting that the TE-like sequence was found in an intron. In 835 rejected alignments, the TE sequence hit the Seq query and overlapped with CDS, but not its corresponding UniGene, perhaps indicating either a match to a pseudogene or an erroneous structural gene call. For a further 58 rejected results, the TE aligned to the Seq query and a UniGene, but not its CDS, likely indicating a match to a UTR.

The remaining 2373 BLAST hits possessed the expected gene structure and a TE in at least one expressed exon (Fig. 1, Table 1); they comprised our dataset for further analysis (see Supplementary Table S1 for a comprehensive list). For the 2373 alignments, the Seq genomic sequence matched the TE sequence, with a mean alignment length of

Table 1 Summary of the gene families for which the presence of TE-like regions varies among paralogues

Gene[a]	TE family	TE- region size	Inference	Gene family size	Genes with TE	Gene family annotation[b]
At1g74290	*Ac*-like	463	Insertion	9	1	Esterase/lipase/thioesterase family protein
At3g20950	*Copia*-like	363	Insertion	22	1	Cytochrome P450 family protein
At3g27150	EnSpm	573	Equivocal	2	1	Kelch repeat-containing F-box family protein
At3g44540	*Ac*-like	371	Equivocal	2	1	Acyl CoA reductase, putative
At4g02810	EnSpm	134	Equivocal	2	1	Expressed protein
At4g33390	EnSpm	150	Equivocal	2	1	Hypothetical protein
At5g39030	EnSpm	491	Equivocal	4	2	Protein kinase family protein

[a] The gene originally identified in the original Seq vs. UniGene-CDS-TE BLAST

[b] Gene annotations retrieved from TAIR (The *Arabidopsis* Information Resource; http://www.arabidopsis.org)

833 bases and a mean BLAST e-value score of 2.65e-12. A subset of 201 alignments showed strong TE sequence homology across 90% or more of the length of the gene. This suggests that, rather than contributing a small segment to the gene as with the majority of alignments, these 201 TEs may have been involved in TE domestication events.

Of the 2373 genes, 162 had stop codon overlapping the TE-related sequence, suggesting that TEs may have either contributed additional 3′ exon sequence or truncated the gene product by contributing a stop codon. These 162 genes remain expressed, and at least several are functionally well characterized: for example, DET3 (At1g12840) (Schumacher et al. 1999), PGP4 (At2g47000) (Terasaka et al. 2005), and HYD1 (At1g20050) (Topping et al. 1997). Of the 2373 "TE genes," 125 were annotated as alternatively spliced. In 43 cases, the TE was found to overlap the gene's splice junction, raising the possibility that, of 2373 putative TE contributions to genes, 43 contributed both protein-coding sequence and an alternative splice site. We also examined the chromosomal locations of the 2373 genes and compared their distribution across chromosomes to that of our TE database. We found no bias toward any chromosome for these putative TE-gene chimaeras (data not shown).

Figure 2 shows the proportions of TEs involved in the 2373 putative TE-gene chimeras compared to all TEs in our TE database. Perhaps the most striking observation is the statistically significant bias against *helitron*-like sequences within exons; *helitrons* represent 18.9% of the TEs in our database and only 2.4% of exon hits. *Helitrons* have been shown to capture gene fragments (Lai et al. 2005; Hollister and Gaut 2007). If exon capture commonly leads to novel gene formation, the signal of remnant *helitrons* is not discernible in our data. However, *helitron* TE sequence is similar only at the 5′ and 3′ ends, varying considerably internally (Kapitonov and Jurka 2001). Thus, an intrinsic bias against finding *helitrons* may exist in our BLAST analysis. *Mariner*-like class II transposon sequences are also significantly underrepresented in exons. Members of the *mariner* TE family have a 5′-TA-3′ target site, so it may not be surprising that these TEs are not often found in GC-rich, gene-rich regions of the genome.

En/Spm elements are significantly overrepresented in exons. TEs in the *En/Spm* superfamily are known to preferentially insert into hypomethylated gene-rich regions of plant genomes (Kunze and Weil 2002), to the extent that they are used as plant mutagens (T-DNA) (Wisman et al. 1998; Krysan et al. 1999). Another surprising result is the significant overrepresentation of *copia*-like LTR retrotransposons in the putative chimeric gene dataset compared to our whole TE database. In contrast, there is a significant bias against *gypsy*-like LTR retrotransposons. While Wright et al. (2003) estimated roughly equal numbers of *copia*- and *gypsy*-like retroelements in the *A. thaliana* genome, our TE database contains considerably more

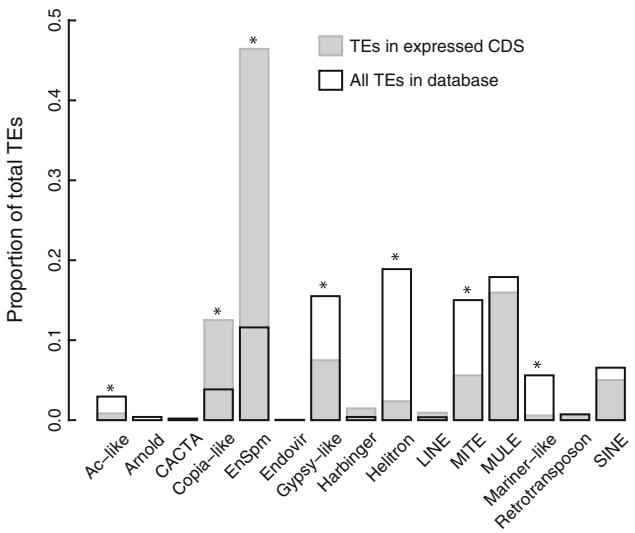

Fig. 2 Bar chart comparing the distributions of TEs involved in the 2373 putative TE-gene chimeras (gray) against all 3079 genomic TEs in our starting database (open bars with *black* borders). Each bar represents the percentage of each TE family within the 2373 TE-gene chimeras and the 3079 TEs, respectively. $*p < 0.05$, 2×2 χ^2 contingency test, after Bonferroni correction

gypsy-like than *copia*-like TEs (468 and 116, respectively). Both of these observations may suggest that the bias toward *copia*-like and against *gypsy*-like elements in coding regions may not be a biological phenomenon, but the result of a deficiency of *copia* sequences in our original TE database and an excess of *gypsy* sequences. Lending support to the veracity of this result, however, *copia*-like elements have been identified previously as having an insertion preference near genes in maize, while *gypsy*-like elements have been observed to preferentially insert into other repetitive elements (Bennetzen 1996).

This discussion of *copia* and *gypsy* make the important point that these comparisons could be sensitive to the method of genome-wide TE identification that was used to compile the original TE database. Our initial compilation of a genome-wide TE query database was conservative with respect to method (using BLASTn as opposed to tBLASTn or other repeat-finding criteria) and stringency (using BLASTn hit e-values <1e-20). As a result, our TE database used in the Seq-Unigene-CDS-TE blast comparison was smaller, in terms of the number of identified TE sequences, than previous estimates of the genome-wide complement of TEs in *A. thaliana* (*Arabidopsis* Genome Initiative 2000; Wright et al. 2003). However, our use of this database also ensures that our results are conservative, with respect both to the number of genes found to have TE homologies and to the believability of results. Even so, our trends are comparable. Qualitatively, the trends in Fig. 2 remained unaffected using the genome-wide percentage TE estimates based on Wright et al. (2003), except for the aforementioned *copia* and *gypsy* result. For example, Wright et al. (2003) estimated that *helitrons* and SINE elements comprise ~23% and ~3% of genomic TEs, respectively, whereas we estimate ~20% and ~7%, respectively.

Functional Biases of Genes with Homology to TEs

The ORFs of functional TEs encode a narrow range of functions. For example, in order to transpose successfully, a class II DNA transposon only needs to bind and cut both its terminal inverted repeats and its target site using a single transposase enzyme. One might expect, therefore, that chimeras between TEs and genes would also encompass limited function. An example is the human SETMAR protein, which is a chimera between a *mariner* class II TE and a previously existing protein (Cordaux et al. 2006). The function of SETMAR is unknown, but it appears that the TE contributed a transposase domain and, consequently, a new DNA-binding function to SETMAR. Following this example, one could predict that exons with TE-like sequences may be enriched for binding functions.

Accordingly, we assessed GO functions for genes containing TE-like sequences (Fig. 3). Of all 15 GO Slim functional categories, 10 categories were significantly over- or underrepresented for genes with TE-like sequences compared to all genes in the *Arabidopsis* genome. Meeting our expectations, the "transcription factor activity" GO Slim category was significantly overrepresented for putative TE-gene chimeras. Contrary to our prediction, however, neither "nucleic acid binding" nor "DNA or RNA binding" functions were significantly over- or underrepresented. Most significantly overrepresented were both the "kinase activity" and the "transferase activity" functions. Both kinase and transferase genes are known to form large gene families in *Arabidopsis* (Meyers et al. 1998; Frova 2003). Perhaps this functional redundancy permits the acquisition of TE sequence with few detrimental consequences. Also, "TE genes" were significantly overrepresented in the "nucleotide binding" and "receptor binding/activity" functional categories.

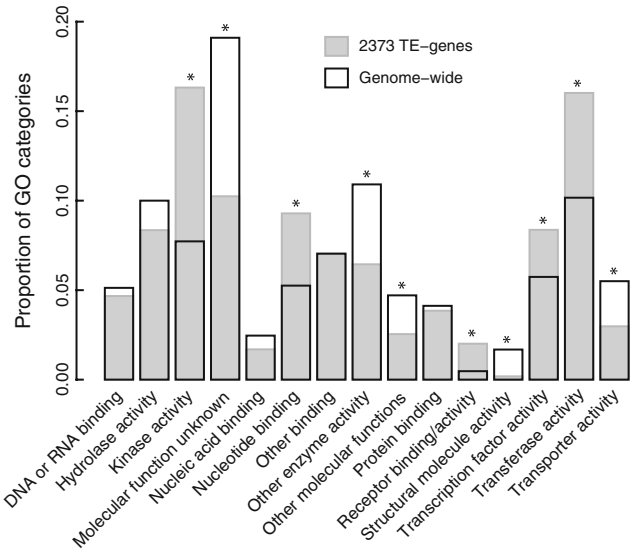

Fig. 3 Comparison of the distributions of GO Slim annotations for the genes involved in the 2373 putative TE-gene chimeras (*gray*) against all genes in the *Arabidopsis* genome (open bars with *black* borders). Each bar represents the percentage of genes that fall into a particular GO annotation, for both types of genes in question. $*p < 0.05$, 2×2 χ^2 contingency test, after Bonferroni correction

"TE genes" were significantly underrepresented in the "transporter activity" GO Slim functional class. This group of functions encompasses proteins which facilitate transmembrane transport—a function not commonly associated with TEs. Also underrepresented were the three "catch-all" categories of "molecular function unknown," "other enzyme activity," and "other molecular functions." Putative TE-gene chimeras were also poorly represented in the "structural molecule activity" GO Slim category, with only 1 of the 907 genes in this category demonstrating homology to TE-related sequences. If the "structural molecule activity" category consists primarily of conserved

housekeeping genes, these genes could be more sensitive to perturbation by TE insertion than genes in other functional groups.

Gene Family Data Help Discriminate Between TE Insertion and Co-option

Thus far we have described 2373 examples of homology between TEs and expressed exonic sequence. But with homology data alone, we cannot infer the direction of the relationship. That is, did TEs *contribute* sequence to exons, thus providing potentially adaptive material, as has been widely argued (Britten 2006; Cordaux et al. 2006), or did TEs *co-opt* genic sequence, as has been demonstrated previously (Jiang et al. 2004; Lai et al. 2005)? Although the direction of sequence relationship can be difficult to decipher, gene family phylogenies can provide insight (Gotea and Makalowski 2006). With gene family data and a phylogenetic context, there is the possibility to infer directionality using parsimony arguments.

We compared each of the 2373 Seq to TE-UniGene-CDS homologues to determine if they belonged to gene families. Of the 2373 genes, 1928 were single-copy (Fig. 1), which provided no information as to directionality, as they are not present in a gene family, and were thus not considered further. For each of the remaining genes, the gene families in which they belonged were assembled into 391 unique gene families and aligned. For each of these 391 multigene families, we characterized the distribution of the TE-like region and, after determining the length of the TE-like region, performed our BLAST resampling test. This resampling test compares randomly chosen, non-TE fragments of the same length as the TE-like region. In 155 cases, the BLAST resampling test could not be performed because the TE region was longer than the flanking gene regions.

BLAST-based searches for regions of TE homology in the remaining 236 genes revealed that every paralogue of 191 gene families contained the same TE-like sequence; these gene families were discarded from further consideration as no phylogenetic inference regarding TE insertion or cooption could be made, leaving 46 gene families under consideration. In 39 cases low BLAST resampling scores suggested that the TE-like region was not out of the ordinary with regard to divergence among gene family members. Thus, in these cases we could not clearly identify the TE region as a unique insertion. This led us to conclude that sequence evolution among the gene family members was responsible for the result, and not a TE insertion event (see Fig. 4a for an example). Seven gene families remained on which to perform further analyses.

Given the seven gene families that met our strict requirements, we employed two additional criteria to discriminate between TE insertion or gene sequence acquisition. The first employed parsimony arguments: a clade of "TE-gene" chimeras within a gene family with many paralogues that lack the TE-like sequence argues strongly for a TE insertion event. Second, we examined each alignment by eye for either an insertion or an unusually divergent region at the location of the TE-like sequence (as identified by the original BLAST). For example, the large, 22-paralogue cytochrome P450 gene family contained only a single paralogue (At3g20950) in which 363 base pairs (bp) of a *Copia*-like element perfectly matched the start of the gene, contributing an intron, and extending into the second exon (Supplemental Fig. 1). This TE-like region exists as an insertion only in paralogue At3g20950. The other 21 paralogues in this gene family do not possess this same sequence. Moreover, the BLAST resampling results also indicated that the TE-like region is atypical with regard to sequence divergence. Thus one can infer that At3g20950 is an example of a single TE insertion into a coding region of an expressed gene (Fig. 4b).

We applied our additional criteria to all seven of the remaining gene families. A second example of TE insertion was found in one paralogue (At1g74290) of an "esterase/lipase/thioesterase" gene family (Fig. 4c), where a single paralogue in the nine-member gene family was found to contain a high sequence similarity to an *Ac*-like TE. For the remaining five gene families, we were unable to conclude that either "TE contribution" or "TE co-option" was the cause of the pattern of TE-like regions on the phylogenies. Most (four or five) of these gene families were two-member gene families, in which parmisony arguments are impossible to apply (Table 1). One of these five additional gene families was four paralogues in size, in which two putative TE-gene chimeras formed a single clade. Two other paralogues formed their own clade, thus making it unclear whether an ancient TE insertion or co-option was responsible for the pattern of TE-like sequence on the tree (Fig. 4d). Although our parsimony arguments cannot differentiate between TE acquisition and TE insertion in these five examples, these may still represent true TE contributions to coding sequence. In these cases, the availability of an outgroup sequence, such as *A. thaliana*'s sister species *Arabidopsis lyrata*, would facilitate this distinction. Additionally, although parsimony arguments based solely on distributions of TE-regions on phylogenetic trees cannot be made in these five examples, one may argue that, since exon capture by TEs is a relatively rare event in comparison with TE insertion via transposition, TE insertion may be the most parsimonious conclusion.

Implications of the Methods and Results

Overall, we found 2373 genes where coding sequence, ESTs, and TEs showed strong BLAST identity to the same

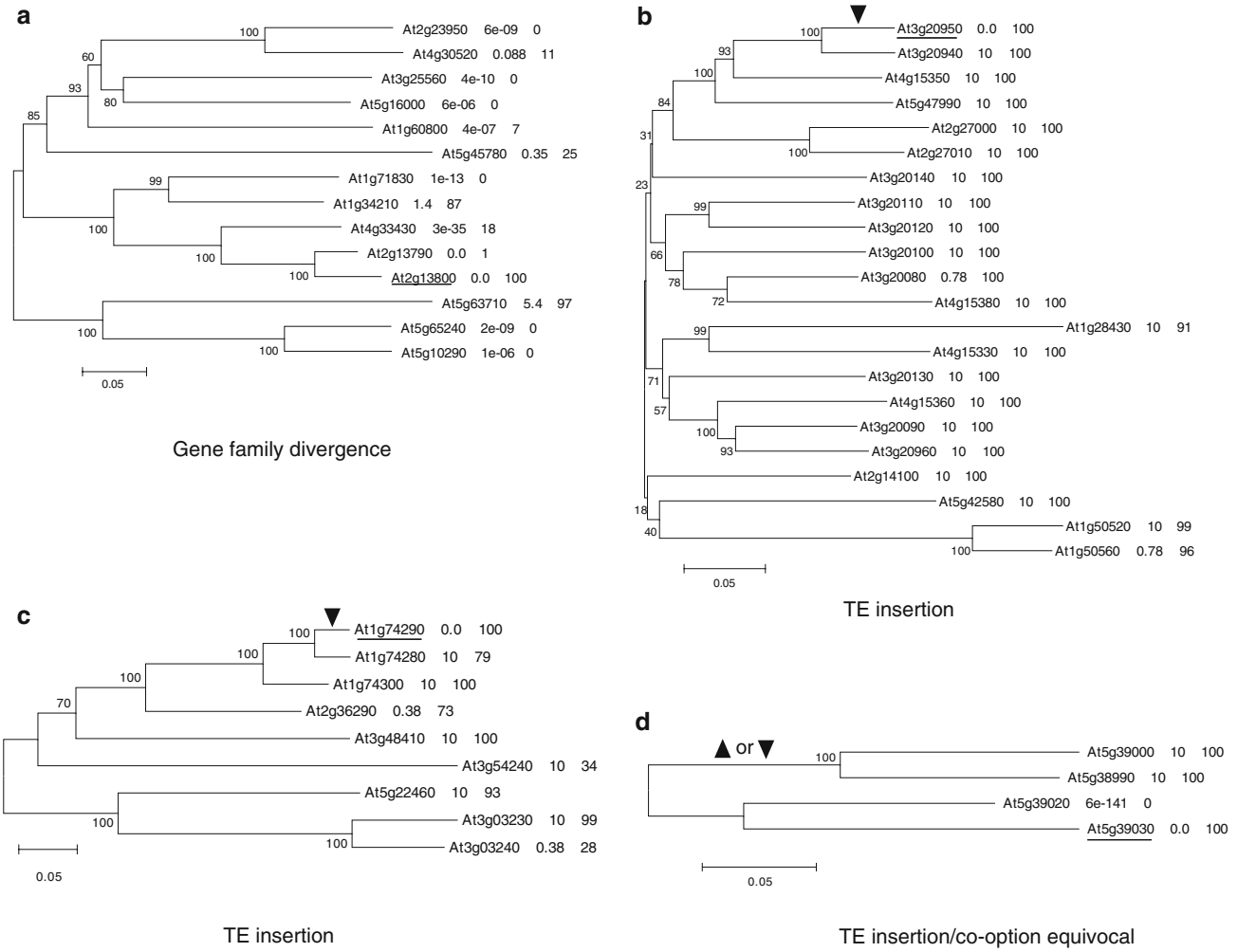

Fig. 4 Gene family phylogenetic trees (**a**–**d**). Phylogenies that have downward-pointing black triangles are examples of trees in which the inference of a TE event was possible. Conclusions are based on parsimonious events, assuming equal probability of excision and insertion, as well as consideration of the mode of TE replication. The underlined gene was originally identified in the initial BLAST as possessing a putative expressed TE in CDS. To the right of each locus tag are two numbers: first, the TE vs. gene family tBLASTx e-value; second, the result of the BLAST resampling (as a percentage). The text below each figure describes the inference drawn. Gene family names and biological functions are given in Table 1

genomic region. At the level of sequence homology, then, we provide evidence that TE-like sequences are present in expressed protein-coding sequences in 7.8% of *Arabidopsis* annotated genes. We caution that some of these could be expressed transcripts that do not contribute to the proteome, but at present the number of confirmed proteins is not sufficient to provide an unbiased genome-wide analysis at the protein level (Gotea and Makalowski 2006). We then addressed the question of directionality and found only a handful of gene families with compelling evidence for a TE insertion event. These few cases likely do not represent the full extent of TE contribution to exons in *A. thaliana* and, as such, likely underestimate the true picture.

What factors led us to believe that we underestimated the number of TE-gene chimeras? First, as mentioned above, we began with a TE query database that was compiled using conservative methods. Second, our phylogenetic analyses are biased against older gene families, as only young gene families tended to pass our BLAST resampling test, which rejected overly divergent paralogues. Third, our phylogenetic methods were amenable only to *Arabidopsis* genes within gene families. Further inferences about the direction of TE-gene homologies for singleton genes may be possible from multispecies analysis (e.g., Gotea and Makalowski, 2006); to this end, ongoing genome sequencing of additional *Brassica* taxa will provide a valuable resource for deciphering the contributions of TEs to annotated protein-coding regions. Finally, we used gene family data based on stringent parameters (≥50% BLASTp identity over ≥90% of the sequence), and only 34.8% of *Arabidopsis* genes were included in gene families under this definition (Rizzon

et al. 2006). TE-gene chimeras are likely to be assigned more often to the ~65% of genes that are not in a gene family. The reason is that a TE insertion changes the sequence of its peptide, making it less similar to its homologues and resulting in its exclusion from a gene family.

To examine the effect of gene family definitions on our study, we repeated the phylogenetic analyses using lower-stringency gene family definitions (paralogues with ≥30% identity over ≥70% of the peptide) (Rizzon et al. 2006). BLAST resampling of the originally-identified TE genes, however, often showed very low match proportions, making it difficult to interpret whether the TE-like region was unique. After repeating the phylogenetic analysis with low-stringency gene families, it became clear that our use of high-stringency paralogues led to fewer inferences of TE insertion events, but limited false positives.

There is no doubt that TEs are major contributors to the evolution of plant genomes (Jiang et al. 2004; Bennetzen 2005; Brunner et al. 2005; Lai et al. 2005). It is also clear that chimeric constructs are relatively common in plants, particularly when TEs acquire portions of coding regions (Jiang et al. 2004; Wang et al. 2006). Thus far, however, the extent of TE contribution to expressed and putatively functional proteins has not been assessed. Despite the conservative nature of our analysis, we found compelling evidence for TE insertion into expressed protein-coding sequences. Our results reside between two extremes, represented by human studies, which claim that more than 1000 proteins contain TE sequence (Nekrutenko and Li 2001; Britten 2006), and *Drosophila melanogaster*, which seems to possess very few expressed TE-gene chimeras (Lipatov et al. 2005). With very few exceptions (e.g., Gotea and Makalowski, 2006; Bundock and Hooykaas 2005), the directionality of these relationships (contribution or co-option of genic regions by TEs) has not been determined. We have found only a handful of cases for which the evidence of TE contribution to a coding region is strong but expect that larger plant genomes, with correspondingly larger TE complements, contain more evidence for TE contributions to coding regions. Even so, the contribution of TEs to TE-gene chimeras may not be small in *Arabidopsis*. We found that 15% (7 of 46) of the examined multigene families provided compelling evidence for incorporation of TE sequence into coding regions. If this proportion is representative, then ~361 of our initial set of 2373 "TE genes" represent TE contributions to coding regions, representing ~1.2% of all annotated *A. thaliana* proteins.

Acknowledgments We thank J. Hollister, L. DeRose-Wilson, and J. Ross-Ibarra for discussion. This research was supported by an NSF grant to B. S. Gaut.

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135–141

Altschul SF, Madden TL, Schaffer 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

Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. Science 309(5735):764–767

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Bennetzen JL (1996) The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol 4:347–353

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–627

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G, Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY (2004) Functional annotation of the *Arabidopsis* genome using controlled vocabularies. Plant Physiol 135:745–755

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res 13:137–144

Britten R (2006) Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci USA 103:1798–1803

Brunner S, Pea G, Rafalski A (2005) Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. Plant J 43:799–810

Bundock P, Hooykaas P (2005) An *Arabidopsis hAT*-like transposase is essential for plant development. Nature 436:282–284

Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA 103:8101–8106

Frova C (2003) The plant glutathione transferase gene family: genomic structure, functions, expression and evolution. Physiol Plantarum 119:469–479

Gotea V, Makalowski W (2006) Do transposable elements really contribute to proteomes? Trends Genet 22:260–267

Greene B, Walko R, Hake S (1994) Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. Genetics 138:1275–1285

Hollister JD, Gaut BS (2007) Population and evolutionary dynamics of helitron transposable elements in *Arabidopsis thaliana*. Mol Biol Evol 24:2515–2524

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569–573

Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714–8719

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3:e181

Kazazian HH Jr (2004) Mobile elements:drivers of genome evolution. Science 303:1626–1632

Krysan PJ, Young JC, Sussman MR (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell 11:2283–2290

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5:150–163

Kunze R, Weil C (2002) The hAT and CACTA superfamilies of plant transposons. In: Craig NL, Gragie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, pp 565–610

Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA 102:9068–9073

Le QH, Wright S, Yu Z et al (2000) Transposon diversity in *Arabidopsis* thaliana. Proc Natl Acad Sci USA 97:7376–7381

Le Rouzic A, Boutin TS, Capy P (2007) Long-term evolution of transposable elements. Proc Natl Acad Sci USA 104:19375–19380

Lipatov M, Lenkov K, Petrov DA, Bergman CM (2005) Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. BMC Biol 3:24

Lockton S, Gaut BS (2005) Plant conserved non-coding sequences and paralogue evolution. Trends Genet 21:60–65

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA 101:14349–14354

Meyers BC, Shen KA, Rohani P, Gaut BS, Michelmore RW (1998) Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell 10:1833–1846

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002

Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619–621

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol 2:e115

SanMiguel P, Tickhonov A, Jin Y-K, Melake-Berhan A, Springer PS, Edwards KJ, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon insertion in Drosophila simulans. Proc Natl Acad Sci USA 101:1626–1631

Schumacher K, Vafeados D, McCarthy M, Sze H, Wilkins T, Chory J (1999) The *Arabidopsis det3* mutant reveals a central role for the vacuolar H(+)-ATPase in plant growth and development. Genes Dev 13:3259–3270

Terasaka K, Blakeslee JJ, Titapiwatanakun B, Peer WA, Bandyopadhyay A, Makam SN, Lee OR, Richards EL, Murphy AS, Sato F, Yazaki K (2005) PGP4, an ATP binding cassette P–glycoprotein, catalyzes auxin transport in *Arabidopsis thaliana* roots. Plant Cell 17:2922–2939

Topping JF, May VJ, Muskett PR, Lindsey K (1997) Mutations in the HYDRA1 gene of Arabidopsis perturb cell shape and disrupt embryonic and seedling morphogenesis. Development 124:4415–4424

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK, Long M, Wang J (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791–1802

Wisman E, Cardon GH, Fransz P, Saedler H (1998) The behaviour of the autonomous maize transposable element En/Spm in *Arabidopsis thaliana* allows efficient mutagenesis. Plant Mol Biol 37:989–999

Wright DA, Voytas DF (1998) Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/gypsy retrotransposons that encode envelope-like proteins. Genetics 149:703–715

Wright SI, Agrawal N, Bureau TE (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res 13:1897–1903

Wu M, Li L, Sun Z (2007) Transposable element fragments in protein-coding regions and their contributions to human functional proteins. Gene 401:165–171

Zhang L, Gaut BS (2003) Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome Res 13:2533–2540