

Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in *Arabidopsis*

Katrien M. Devos,^{1,2,3} James K.M. Brown,¹ and Jeffrey L. Bennetzen²

¹John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, United Kingdom; ²Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392, USA

Genome size varies greatly across angiosperms. It is well documented that, in addition to polyploidization, retrotransposon amplification has been a major cause of genome expansion. The lack of evidence for counterbalancing mechanisms that curtail unlimited genome growth has made many of us wonder whether angiosperms have a “one-way ticket to genomic obesity.” We have therefore investigated an angiosperm with a well-characterized and notably small genome, *Arabidopsis thaliana*, for evidence of genomic DNA loss. Our results indicate that illegitimate recombination is the driving force behind genome size decrease in *Arabidopsis*, removing at least fivefold more DNA than unequal homologous recombination. The presence of highly degraded retroelements also suggests that retrotransposon amplification has not been confined to the last 4 million years, as is indicated by the dating of intact retroelements.

Flowering plants (angiosperms) vary enormously in genome size, from <50 Mb in some members of the Cruciferae to >85,000 Mb in some Liliaceae (Bennett and Leitch 1995). The mechanisms that account for dramatic expansion of angiosperm genomes have been documented, primarily polyploidization and retrotransposon amplification (SanMiguel et al. 1996, 1998; Wendel 2000); however, counterbalancing modes of genome contraction have not been convincingly shown. In the absence of an equally comprehensive and aggressive mechanism for genome size decrease, the question remains whether angiosperms have a “one-way ticket to genome obesity” (Bennetzen and Kellogg 1997). We have addressed this fundamental issue in the genome size debate by studying the structure and evolution of long terminal repeat (LTR) retrotransposons in *Arabidopsis*.

LTR retrotransposons constitute a large part of the repetitive DNA fraction in plant species. They are characterized by LTRs that vary in size from a few 100 base pairs (bp) to several kilobases and terminate in short inverted repeats, usually 5'-TG-3' and 5'-CA-3' (Kumar and Bennetzen 1999). The well-defined structure of LTR retrotransposons, their prevalence and dispersion in the genome, their acknowledged role in genome size expansion, and the fact that individual elements have little or no selective significance make LTR retrotransposons suitable elements for studying genome evolution (Petrov 2001). The prevalence and distribution of LTR retrotransposons have been the subject of several studies, including in *Arabidopsis* (Marin and Lloréns 2000; Terol et al. 2001). These studies, however, are generally based on the analysis of intact elements of relatively recent origin and provide no information on the long-term fate of these sequences. In our study, LTR-retrotransposon families were established on the basis of homology of the LTRs rather than the open reading frames. An important advantage of this approach is that not

only complete elements but also solo LTRs and elements that have undergone a variety of deletions can be identified. It is precisely the structure of this latter group that provides the most important clues regarding plant genome evolution.

RESULTS AND DISCUSSION

We have analyzed a total of 291 LTR-retrotransposon elements belonging to 12 families (four *copia*, six *gypsy*, two unknown). The retroelements are distributed over the five *Arabidopsis* chromosomes and show the typical pericentromeric clustering previously observed for LTR retrotransposons (Lin et al. 1999; Mayer et al. 1999), indicating that the 291 elements form a representative sample. The 12 families were originally identified in two bacterial artificial chromosome (BAC) clones that were randomly chosen from a selection of annotated *Arabidopsis* BACs that contained putative LTR retroelements. The LTRs of these elements were then used as query sequences in BLAST searches against the *Arabidopsis* genomic sequence (<http://www.arabidopsis.org>). Incomplete elements were taken into account only if they retained at least one of the LTR-retrotransposon characteristics such as a primer-binding site (PBS), a polypurine tract (PPT), or a target duplication site (Kumar and Bennetzen 1999). Thus, many severely deleted LTR retrotransposons that we detected were not further studied because their highly fragmentary structure made it impossible to determine the nature of specific rearrangements that they had undergone. Of the 291 studied elements, 87 (29.9%) were found to be “complete”; that is, they contain two LTRs flanked by a 5-bp target-site duplication and separated by an internal region containing a PBS and PPT (Fig. 1A). By use of the dating strategy described by SanMiguel et al. (1998), but applying the synonymous substitution rate of 1.5×10^{-8} mutations per site per year determined for the *Chs* and *Adh* genes in the Brassicaceae (Koch et al. 2000), we estimated that these retrotransposons all inserted in the *Arabidopsis* genome during the last 4 million years, most within the last 2 million years (data not shown). These estimates are based on the assumption that LTRs evolve at approximately

³Corresponding author.

E-MAIL Katrien.devos@bbsrc.ac.uk; FAX 44 1603 450 023/24.
Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.132102>.