Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression

(retroelements/transposable elements/database searches/Hopscotch)

SHAWN E. WHITE*, LEDARE F. HABERA*, and SUSAN R. WESSLER**††

Departments of *Botany and ††Genetics, University of Georgia, Athens, GA 30602

Communicated by Charles S. Levings III, August 15, 1994 (received for review April 8, 1994)

ABSTRACT The wx-K mutation results from the insertion of a copia-like retrotransposon into exon 12 of the maize waxy gene. This retrotransposon, named Hopscotch, has one long open reading frame encoding all of the domains required for transposition. Computer-assisted database searches using Hopscotch and other plant copia-like retroelements as query sequences have revealed that ancient, degenerate retrotransposon insertions are found in close proximity to 21 previously sequenced plant genes. The data suggest that these elements may be involved in gene duplication and the regulation of gene expression. Similar searches using the Drosophila retrotransposon copia did not reveal any retrotransposon-like sequences in the flanking regions of animal genes. These results, together with the recent finding that reverse-transcriptase sequences characteristic of copia-like elements are ubiquitous and diverse in plants, suggest that copia-like retrotransposons are an ancient component of plant genomes.

The retroelement family is composed of transposable elements that move via an RNA intermediate (1). Included in this family are long-terminal-repeat (LTR) retrotransposons and retroviruses. Both LTR retrotransposons and retroviruses are flanked by LTRs that provide cis-regulatory sequences required for transcription of an RNA intermediate (2). The internal sequences of these elements encode proteins (Gag, protease, integrase, reverse transcriptase, and RNAse) necessary for reverse transcription and integration. Based on the arrangement of their protein-coding domains, LTR retrotransposons can be subdivided into two groups named after the Drosophila retrotransposons copia and gypsy (2). The integrase domain is positioned 3' of the reverse transcriptase domain in gypsy-like retrotransposons, and 5' of reverse transcriptase in copia-like retrotransposons. Both groups have been found in fungi and plants in addition to Drosophila but have not been detected in animals other than insects and fish (3). copia-like reverse transcriptases have been identified in almost every plant species surveyed (4–6) and are diverse in terms of their amino acid sequences (7). Only a few plant retrotransposons, however, are responsible for recent mutations (Tnt1 of tobacco and Bs1, Stonor, BS, and G of maize; refs. 8–10). Of these, Tnt1 is the only retrotransposon shown to be completely transcriptionally active in plants grown under normal conditions.

This paper presents the characterization of a second complete plant retrotransposon, Hopscotch. Use of this element in computer-based sequence similarity searches reveals that many normal plant genes have the remnants of copia-like retrotransposons in their upstream and downstream flanking regions. These results provide evidence that retroelements have the potential to be involved in the evolution of plant gene structure and expression by supplying genes with regulatory sequences and facilitating gene duplication. Furthermore, despite the fact that copia-like retrotransposons have been found in insects and fish, no element sequences were found in the flanking regions of normal animal genes.

MATERIALS AND METHODS

Cloning and Sequencing. Genomic DNA was isolated from maize seedlings homozygous for the wx-K mutation (11). SalI fragments of 4.5–6 kb were cloned into λZAPII phage vector (Stratagene) and the resulting plaques were screened with a waxy (wx)-specific probe, SalE (12). Both strands of a positive clone were sequenced with a Sequenase kit (United States Biochemical).

Database Searches. Computer-based amino acid similarity searches of the GenBank (version 77.0) and EMBL (version 34.0) databases were performed with the tFASTA search program of the University of Wisconsin Genetics Computer Group (GCG) software package (version 7.0) accessed through the BioScience Computing Resource at the University of Georgia. Conceptual translations of the sequences of the retrotransposons Tnt1 of tobacco (accession no. X13777) (9), TaIF-3 of Arabidopsis (X13291) (13), PDRI of pea (X66399) (14), Tst1 of potato (X52387) (15), copia of Drosophila (X02599) (16), BARE-I of barley (Z17327) (17), and Hopscotch were used as query sequences. Nucleic acid-level searches of the GenBank and European Molecular Biology Laboratory databases were performed with the BLASTN (18) search program of the National Center for Biotechnology Information (Bethesda). Pairwise DNA sequence comparisons were made using the FASTA program of GCG. The GCG PILEUP and BOXSHADE programs were used to make the alignment figures. The alignments were edited to account for frameshifts.

RESULTS

The wx-K mutation of maize results from an ≈4.5-kb insertion in the wx gene (12). Analysis of the DNA sequence of this insertion revealed that it has the structure of an LTR retrotransposon. We have named this element Hopscotch. Hopscotch has identical 231-bp LTRs, is 4828 bp long, and has a single open reading frame of 4320 nt (1440 aa). A potential primer binding site with similarity (18/19 nt) to the 3' end of wheat initiator methionine tRNA (GenBank accession no. V01383) is found adjacent to the 5' LTR, and a polypurine tract lies next to the 3' LTR.

Comparison of Hopscotch with Known Retrotransposons. The nucleotide and derived amino acid sequences of Hop-
scotch were compared with other transposable elements by computer searches of the GenBank and European Molecular Biology Laboratory databases. The searches identified significant nucleic acid (35–51%) and amino acid (23–32%) similarities between Hopscotch and several known copia-like retrotransposons. These comparisons also revealed that Hopscotch contains all of the amino acid domains (nucleic acid binding, protease, integrase, reverse transcriptase, and RNase H) that are found to be conserved among autonomously active retropulement (Fig. 1A). Both the amino acid conservation and the domain order serve to identify Hopscotch as a copia-like retrotransposon (Fig. 1B).

Retrotransposon-Like Sequences Flank Many Plant Genes. Surprisingly, these searches also revealed that 16 previously described plant genes have amino acid similarity to the conserved domains of copia-like retrotransposons (Figs. 2 and 3). Additional searches using the derived amino acid sequence of the retrotransposon Tnt1 of tobacco as a query sequence detected 3 more plant genes with flanking regions similar to these domains (cotlea4a, cotdgala, and whtgerma) as well as 13 of the retrotransposon-like sequences identified in the Hopscotch searches. No additional plant genes were detected by using the amino acid sequences of other plant copia-like elements (Tal1-3 of Arabidopsis, PDR1 of pea, BARE-1 of barley, and Tstl of potato) as query sequences.

Many of the retrotransposon-like sequences in the flanking regions of the genes probably represent ancient insertions. In several of the genes, retroelement similarity is degenerate, ends abruptly (cotmat5a, cucacc1, gmchsl, mzeq3pd, ricmtnad3a, phvarcla, pschs1, zmpgalac, and zmpmslg2), or contains internal deletions (cotmat5a, cucacc1, phvarcla, and pschs1). In three cases, retrotransposon-like sequences are found in the same position in several members of a gene family, indicating that insertion preceded gene duplication. Comparison of the 5' end of the pea ribulose-bisphosphate carboxylase gene rbcS-E9 with other members of the rbcS gene family revealed that two other rbcS genes (rbcS-8.0 and rbcS-3.6) (37) have insertions at the same site in their upstream flanking regions. Similarly, several members of the maize 19-kDa zein gene family have elements inserted at the same position, as do members of the maize polygalacturonase gene (PG) family (zmpgalac, zmpgnsg, zmpgg14) (38).

Nucleic acid-level searches using the plant retrotransposon sequences revealed two more genes with flanking regions similar to copia-like retrotransposons. The 3' flanking region of the pea glyceraldehyde-3-phosphate dehydrogenase gene (Gpbl) (39) has 92% similarity to the PDR1 LTR and probably represents part of an LTR from another copy of the PDR1 retrotransposon. Likewise, the 5' flanking sequence of a tomato gene expressed during pollen development (LAT59) (40) has 65% similarity to the LTR of Tnt1. Since Tnt1 has been detected in the tomato genome by Southern blot analysis (9), the retrotransposon-like sequence in LAT59 is probably the LTR of a Tnt1-related retrotransposon in tomato.

The amino acid- and nucleic acid-level searches combined identified 21 genes with flanking regions similar to copia-like retrotransposons. Retrotransposon similarity in 20 of these 21 genes had gone undetected until this study. Only the retrotransposon-like sequence at the 3' end of the cotton 2S albumin storage-protein gene (Mat5-A) had been reported (22). The Tnt1 amino acid searches detected another retrotransposon-like sequence in a 5' flanking region of this gene previously described as repetitive (22).

Identifying Element LTRs. The flanking sequences of the genes identified in this study do not encompass complete retrotransposons, so the limits of the elements cannot be resolved by comparing LTRs. Although tRNA binding sites or purine tracts characteristically found immediately internal to the 5' and 3' LTRs, respectively, can be used to determine one end of an LTR, the other end is often unidentifiable. By comparing related elements or gene family members with and without insertions, however, we have been able to approximate the LTRs of several retrotransposon-like sequences. Alignment of the pea rbcS-E9 gene with a closely-related pea rbcS gene (rbcS-3A) lacking the retrotransposon insertion allows the limits of the retrotransposon to be defined. The LTRs of the related elements adjacent to the cotton Mat5a and Lea44a genes and the pea Chsl and phenylalanine ammonia-lyase (PAL2) genes were defined by pairwise alignment of their nucleic acid sequences. The point at which retrotransposon sequence similarity ends was used to approximate one end of an LTR and the position of tRNA binding sites was used to define the other end. The ends of the retrotransposons in the 19-kDa zein genes were found by comparing the related upstream and downstream retrotransposons which share 90% sequence identity. The upstream element's LTR found in this manner corresponds precisely with the element end as determined by comparison with another member of the 19-kDa zein gene family lacking the insertions (ze19).

Retrotransposon-Like Sequences Are Not Found in Drosophila Genes. To identify genes harboring nearby retroelement sequences in organisms other than plants, the searches were repeated with the Drosophila retrotransposon copia as a query sequence. These searches failed to identify a single normal insect or other animal gene with copia-like retrotransposon sequences. They did, however, detect 15 of the plant sequences shown in Fig. 2.
DISCUSSION

Computer-assisted searches of the GenBank and European Molecular Biology Laboratory databases using three plant retrotransposons as query sequences revealed unexpected similarity to the flanking regions of 21 plant genes. The identity of these sequences as retrotransposon-like had been recognized previously in only 1 of these 21 genes. Although most regions of similarity appear to be the remnants of ancient insertions, we are confident of their retrotransposon origin for the following reasons: (i) the extent of amino acid similarity is striking (e.g., the ricmtna3 element has 25 matches to either TntI or Hopscotch over 33 aa, and the element in the zmmps2g gene has 136 matches over 475 aa); (ii) the regions of amino acid similarity include conserved retrotransposon domains (i.e., Gag, protease, integrase, reverse transcriptase, or RNase H); (iii) several flanking sequences contain similarity to more than a single domain; and (iv) in many instances, other distinguishing structural features of retrotransposons such as LTRs, tRNA binding sites, and polypurine tracts can be identified. Despite the fact that retroelement similarity usually lies within 1 kb of the coding regions, 20 of the 21 published genes discussed in this paper are normal rather than mutant (the exception being the cotmat5a gene). Over 30 normal plant genes have been previously found to contain the inverted repeat transposable elements Tourist or Stowaway (41–43). Thus, the total number of plant genes harboring mobile elements or their remnants is >100.

The Retrotransposon-Like Sequences Contain Previously Identified Cis-Regulatory Elements. Four examples of ancient retroviral insertions that provide regulatory sequences to adjacent genes have been previously described. The mouse sex-limited protein gene is expressed in the presence of androgen due to a hormone-responsive enhancer in the LTR of an endogenous provirus (44). In humans, an upstream endogenous retroviral insertion has been found to be responsible for parotid-gland tissue specificity of the salivary amylase genes (45). Finally, the rat oncomodulin gene and the mouse IAP-promoted placental gene are under the control of promoters in solo LTRs of rodent intracisternal A particles (IAPs) (46, 47).

Several lines of evidence suggest that some of the retrotransposon-like sequences identified in this study may influence the expression of adjacent genes. The retrotransposon-like sequences in the maize polygalacturonase (PG) genes contain sequence motifs that are common among genes expressed during pollen development (35). In addition, a 501-bp fragment containing a positive regulatory region of a tomato gene expressed during pollen development (LAT59) (40) is composed entirely of a retrotransposon-like sequence.

The region upstream of nt −250 of the pea rbcS-E9, −8.0, and −3.6 genes corresponds to a retrotransposon insertion that occurred prior to gene duplication. Another family member, rbcS-3A, lacks the insertion and, for this reason, has distinct sequences from nt −250 upstream to at least −410. Interestingly, the rbcS-E9, −8.0, and −3.6 genes are coordinately expressed in a manner different from rbcS-3A (48). The combined expression of the rbcS-E9, −8.0, and −3.6 genes in leaves is 30–50% lower than rbcS-3A gene expression, and their transcripts are underrepresented in pea petals and seeds when compared with rbcS-3A transcripts. On the basis of promoter domain-swapping experiments, the region upstream of −170 of the rbcS-E9 gene (∼90% of which corresponds to the retrotransposon-like sequence) has been hypothesized to harbor a negative regulator of transcription (49).
The retrotransposon-like sequences flanking the 19-kDa zein genes of maize may have influenced both the expression and structure of this gene family. At least five of the seven sequenced 19-kDa zein gene family members (pms1, pms2, ze19ba, ze1l9, and ze25) have retrotransposon-like sequences at the same site in upstream flanking regions (Fig. 4). These sequences have 90% nucleic acid sequence identity to another element found in the downstream flanking regions of at least two members of this family (pms1 and pms2). Many of the 19-kDa zein genes have two promoters, P1 and P2, with P1 accounting for ~0.1% of zein gene transcripts (53). Our analysis indicates that P1 and the nearby start site of transcription are composed entirely of retrotransposon LTR sequences. The P1 promoter sequence has also been identified in the downstream flanking DNA of the pms2 gene (36) and lies within the LTR of the downstream retrotransposon.

In addition to providing a zein promoter, the retrotransposon sequences may have facilitated the amplification of this gene family. Since the 19-kDa zein genes have been found clustered on the short chromosome 7 (54), there is a possibility that they are tandemly arranged. In fact, ze19ba and ze25 have been found in such a tandem arrangement (52). This organization suggests that the 19-kDa zein genes were duplicated by homologous, unequal crossing-over between retrotransposons inserted on either side of a progenitor gene.

**Involvement of copia-Like Elements in Plant vs. Insect Evolution.** Only plant gene sequences were identified as having significant similarity to either plant (TntI, Hopscotch) or Drosophila (copia) retrotransposons. This is surprising, since copia-like elements are highly expressed in Drosophila and have been shown to be the causative agent of many spontaneous mutations (55). In contrast, plant copia-like retrotransposons are transcribed at low levels under normal conditions and have been found to be responsible for only a few mutations.

The disparity between plant and Drosophila genes may reflect a lack of selection against retrotransposon insertions near plant genes. Alternatively, copia-like elements may be an older component of plant genomes and may have had a longer time frame for insertions into the flanking regions of genes to occur and become fixed. This hypothesis is consistent with the results of several recent surveys of reverse transcriptase domains in plant genomes. These studies have revealed that reverse transcriptase sequences characteristic of copia-like retrotransposons are heterogeneous and ubiquitous among plant species and were probably inherited by vertical transmission from a common ancestor (4–7). In contrast, analysis of both the codon usage of copia and its phylogenetic relationship to other retrotransposons has led to the hypothesis that copia-like elements were horizontally transmitted to Drosophila or one of its ancestors.