Active Retrotransposons Are a Common Feature of Grass Genomes

Carlos M. Vicient, Marko J. Jääskeläinen, Ruslan Kalendar, and Alan H. Schulman²*

Plant Genomics Laboratory, Institute of Biotechnology, Viikki Biocenter, University of Helsinki, P.O. Box 56, Viikin kaari 6, FIN–00014 Helsinki, Finland

A large fraction of the genomes of grasses, members of the family Gramineae, is composed of retrotransposons. These elements resemble animal retroviruses in their structure and possess a life cycle similar to theirs that includes transcription, translation, and integration of daughter copies. We have investigated if retrotransposons are generally transcribed in the grasses and other plants, and whether the various families of elements are translationally and integrationally active in multiple grass species. A systematic search of $7.8 \times 10^5$ publicly available expressed sequence tags from plants revealed widespread retrotransposon transcripts at a frequency of one in 1,000. Monocot retrotransposons found relatively more expressed sequence tags from non-source species than those of dicots. Antibodies were raised to the capsid protein, GAG, of BARE-1, a transcribed and translated copia-like retrotransposon of barley (Hordeum vulgare). These detected immunoactive proteins of sizes identical to those of the BARE-1 GAG and polyprotein, respectively, in other species of the tribe Triticeae as well as in oats (Avena sativa) and rice (Oryza sativa). Retrotransposon-based markers showed integration polymorphisms for BARE-1 in different subfamilies of the Gramineae. The results suggest that grasses share families of transcriptionally, translationally, and integrationally active retrotransposons, enabling a comparative and integrative approach to understanding the life cycle of retrotransposons and their impact on the genome.

In most grasses (family Gramineae), genes appear to comprise less than 20% of the genome (Flavell et al., 1977; Barakat et al., 1997), with most of the rest being composed of repetitive DNA. The variation in genome size in eukaryotes, independent of differences in organismal complexity, recognized early on (Thomas, 1971) and referred to as the C-value (genome size) paradox, is particularly apparent in the grasses. Among the diploid grasses, the 1C genome size (DNA content of the unreplicated haploid set of chromosomes) varies from $2.0 \times 10^8$ bp in Oropetium thomaeum to $1.3 \times 10^{10}$ in Lygeum spartum (Bennett et al., 1998). The cereals rice (Oryza sativa, 1C = 4.3 $\times 10^8$), sorghum (Sorghum bicolor, 1C = 7.2 $\times 10^8$), maize (Zea mays, 1C = 2.6 $\times 10^9$), barley (Hordeum vulgare, 1C = 4.5 $\times 10^8$), and rye (Secale cereale, 1C = 8.0 $\times 10^8$) are arrayed in between (Kankaanpää et al., 1996; Kurata et al., 1997; Bennett et al., 1998).

Evidence is accumulating that much of this more than 50-fold variation in genome size is due to variations in the prevalence of one specific class of repetitive DNA, retrotransposons. Retrotransposons are so named because, unlike the DNA transposons such as Ac and En/Spm, they propagate not by cutting and pasting, but by a mechanism of reverse transcription followed by integration of the new cDNA copy back into the genome (Boeke and Corces, 1989; Kumar and Bennetzen, 1999). Their life cycle, encoded products, and structure (Fig. 1) resemble those of the retroviruses; the retroviruses and retrotransposons are thought to be derived from a common ancestor (Xiong and Eickbush, 1990; Doolittle and Feng, 1992; Lazcano et al., 1992). The two classes of retrotransposons, gypsy like and the copia like, differing in the order of their encoded proteins (Fig. 1), are both ubiquitous throughout the plants (Flavell et al., 1992; Voyer et al., 1992; Suonieni et al., 1998). The replicative nature of retrotransposon mobilization, combined with the large size of the elements (5 to 10 kb), indicates that active retrotransposon families have the potential to be major contributors to variation in genome size.

Mapping in the cereals showed that the genes of rice and other cereals are largely syntenic or collinear despite the large differences in genome size (Bennetzen, 2000; Keller and Feuillet, 2000). Detailed comparisons of sequenced regions in the maize and sorghum genomes (Tikhonov et al., 1999) established that their genome size difference was largely (74%) due to accumulation of retrotransposons since the divergence of these species. A major feature of cereal genomes is the localization of genes into “gene islands” interspersed by “repeat seas” (SanMiguel et al., 1996; Ananiev et al., 1998; Panstruga et al., 1998). Sequence analysis of a 66-kb contiguous region of the barley chromosome 2H (Shirasu et al., 2000) showed the three genes on that stretch to span only

¹ This work was supported by Academy of Finland grants (to C.M.V. and M.J.J.), by an Academy of Finland Senior Fellowship (to A.H.S.), and by the European Union Research Directorate (contract no. QLK5–1999–01499).

² This author also is affiliated at the present time with the Agricultural Research Centre of Finland, FIN–31600 Jokioinen, Finland.

* Corresponding author; e-mail alan.schulman@helsinki.fi; fax 358–9–191–58952.
transposons such as particles as seen for retroviruses and active retroscript is also translated, the predicted polyprotein role as contributors to growth in genome size, one rochika et al., 1996; Okamoto and Hirochika, 2000). posons investigated appear to be quiescent in somatic tissues but activated by stress including proto copia evislae Ja¨a¨skela ¨inen et al. (1999) demonstrated that the tran LTR (Suoniemi et al., 1996b; Suoniemi et al., 1997). somatic tissues from conserved promoters within the (Manninen and Schulman, 1993) is transcribed in 1284 Plant Physiol. Vol. 125, 2001 18 kb, yielding a local density only 30% greater than the average for the Arabidopsis genome and within its range (Arabidopsis Genome Initiative, 2000), most of the rest being composed of retrotransposons. If this region is typical, retroelements (intact retrotransposons and their derivatives) account for more than 60% of the barley genome, compared with approximately 5% of the Arabidopsis genome (Arabidopsis Genome Initiative, 2000). We showed earlier that a family of copia-like retrotransposons, BARE-1, alone comprises approximately 5% of the barley genome as 13.7 × 10⁹ full-length copies of 8.9 kb and >6 × 10⁶ solo LTRs of 1.8 kb (Vicient et al., 1999), dispersed along the chromosomes (Suoniemi et al., 1996a). Given 2.5 × 10⁴ to 9 × 10⁴ genes (respectively the number for Arabidopsis and the estimated human unigene set) at this density, the genic component of the barley genome would cover 1.5 × 10⁸ to 5.4 × 10⁹ bp, or 3.3% to 12% of the whole. In barley, the BARE-1 family of retrotransposons (Manninen and Schulman, 1993) is transcribed in somatic tissues from conserved promoters within the LTR (Suoniemi et al., 1996b; Suoniemi et al., 1997). Jääskeläinen et al. (1999) demonstrated that the transcript is also translated, the predicted polypeptide protein processed, and the cDNA packaged into virus-like particles as seen for retroviruses and active retrotransposons such as Ty1 of yeast (Saccharomyces cerevisiae) and copia of Drosophila melanogaster (Miyake et al., 1987; Roth, 2000). Many but not all retrotransposons investigated appear to be quiescent in somatic tissues but activated by stress including protoplast formation (Wessler, 1996; Grandbastien, 1998) and in tissue culture (Grandbastien et al., 1989; Hirochika et al., 1996; Okamoto and Hirochika, 2000). If retrotranspososons currently have a widespread role as contributors to growth in genome size, one would expect to find evidence for their activity in a broad range of species. Here we have examined whether retrotransposons are generally transcribed in the grasses and other plants, and whether the various families of elements are translationally and integrationally active in multiple grass species. We have systematically searched the expressed sequence tag (EST) databases to find transcripts and used immunoblotting to detect translational products and marker methods to reveal insertion site polymorphisms generated by integration. The results suggest that retrotransposons are widely active in the grasses and that at least some families of retrotransposons are active across multiple species.

RESULTS

Retrotransposon Transcription Is Widespread But Most Prevalent in the Grasses

The many EST sequencing projects currently under way for various plants gave us an opportunity to search for transcripts of known retrotransposons in the public EST databases. A total of 7.8 × 10⁵ ESTs were searched for homologies to the LTRs or internal domains of known retrotransposons, and 934 matches (1.2‰) of the total were found (Table I). The number of public EST sequences varies greatly by species, from almost 1.23 × 10⁵ for soybean (Glycine max) to less than 50 accession numbers each for more than 30 species. Furthermore, it should be emphasized that a taxonomically representative set of ESTs for monocots and dicots is not yet available. For plants with more than 1 × 10⁵ reported ESTs, the monocots, mostly grasses, showed the highest average fraction of retrotransposon-containing accession, 1.75‰, compared with 1.40‰ for the two conifer species and 0.92‰ for the dicots. The monocots also had both a higher maximum (3.13‰ versus 2.47‰) and a higher minimum (0.98‰ versus 0.26‰) than the dicots. In a recent in silico transcriptional profiling of EST data sets, Bortoluzzi et al. (2000) considered EST frequencies >3.6‰ to represent abundant, 3.6% to 1.25% moderate, and <1.2‰ rare transcripts. In this framework, retrotransposons are moderately expressed in only three of the nine dicots, but in four of the six monocots analyzed here.

These data suggest that retrotransposons are generally more transcriptionally active in the grasses than in other groups of plants, although transcription occurs in all groups. An important caveat in these analyses is that the ESTs in the databases are derived from a mixture of cDNA construction and sequencing methods and also display length variations. These factors, together with the inherently partial nature of EST sequences, may cause accessions for retrotransposon transcripts to be missed when searching with particular motifs in the same way it hinders annotation of ESTs generally. In addition, it is not possible to differentiate between retrotransposo-
son transcripts originating from LTRs, from elements transcriptionally active in their own right, and transcripts that are initiated in conventional cellular promoters and read through into solo LTRs or adjacent full-length retrotransposons. However, the tendency of genes in grass genomes to cluster into gene islands, discussed above, would tend to decrease the likelihood of illegitimate transcripts, at least for this group of plants.

Earlier work showed that retrotransposon number and genome size is positively correlated in barley (Vicent et al., 1999; Kalendar et al., 2000). Because transcription is a prerequisite to integration of new retrotransposon copies, one might expect to find a positive association between the fraction of retrotransposon ESTs detected and genome size. The data here (Table I) display a strong and significant correlation between genome size and plant group (Pearson Product Moment, $r_p=0.616$, $P=0.025$), the grasses containing larger genomes than the others, and also a weak but not significant association between genome size and the fraction of ESTs matching retrotransposons ($r_p=0.209$, $P=0.493$). Taking the dicots alone, the correlation between genome size and EST fraction is both strong and significant ($r_p=0.895$, $P=0.006$). However, it should be kept in mind that only transcription and ultimate integration in tissues giving rise to gametes is heritable; the ESTs here are derived from many types of tissues (Table II).

### Grass Species Share Retrotransposon EST Matches

A set of 10 previously described retrotransposons from dicots, 27 from grasses, and one from a pine, including 14 which could be clearly defined as copia like and 18 as gypsy like, were used as query sequences against the EST database and the hits analyzed (Table II). Whereas the analysis for Table I looked for all expressed elements, characterized and uncharacterized, detectable in each plant, Table II reports only those matches for the specific elements listed. For the three dicotyledonous species examined encompassing $3.3 \times 10^5$ ESTs, only 12 matches to homologous retrotransposons were found, a fraction of $3.7 \times 10^{-2}$. The most active of these elements is the copia-like RetroLyc of Lycopersicon peruvianum, which found six matches in the closely related cultivated tomato. Of the six Arabidopsis elements examined, only the gypsy-like Athila matched an EST, a single-leaf accession. In parallel, only 4% of the complete retroelements in the full sequence of the Arabidopsis genome correspond to an Arabidopsis EST (Arabidopsis Genome Initiative, 2000). None of the dicot elements found inter-generic matches. In contrast, the 27 grass retrotransposons identified 259 matches within grasses corresponding to $2.6 \times 10^5$ ESTs. Furthermore, the matches from grass retrotransposons show a very different pattern from those of the dicots. Elements of both the gypsy-like and

### Table I. Plant EST databases and their share of accessions matching retrotransposons

<table>
<thead>
<tr>
<th>Genus</th>
<th>Class/Division</th>
<th>Total no.</th>
<th>No. matching retrotransposons</th>
<th>% Total</th>
<th>Genome Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>Dicot</td>
<td>122,843</td>
<td>153</td>
<td>1.25</td>
<td>2.25</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Dicot</td>
<td>112,467</td>
<td>59</td>
<td>0.52</td>
<td>0.35</td>
</tr>
<tr>
<td>Lycopersicon</td>
<td>Dicot</td>
<td>93,115</td>
<td>112</td>
<td>1.20</td>
<td>2.05</td>
</tr>
<tr>
<td>Medicago</td>
<td>Dicot</td>
<td>82,320</td>
<td>21</td>
<td>0.26</td>
<td>1.75</td>
</tr>
<tr>
<td>Zea</td>
<td>Monocot</td>
<td>73,965</td>
<td>184</td>
<td>2.49</td>
<td>5.45</td>
</tr>
<tr>
<td>Oryza</td>
<td>Monocot</td>
<td>62,420</td>
<td>61</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>Triticum</td>
<td>Monocot</td>
<td>47,435</td>
<td>60</td>
<td>1.26</td>
<td>11.55</td>
</tr>
<tr>
<td>Sorghum</td>
<td>Monocot</td>
<td>47,098</td>
<td>75</td>
<td>1.59</td>
<td>1.5</td>
</tr>
<tr>
<td>Lotus</td>
<td>Dicot</td>
<td>26,844</td>
<td>15</td>
<td>0.56</td>
<td>0.95</td>
</tr>
<tr>
<td>Hordeum</td>
<td>Monocot</td>
<td>23,947</td>
<td>75</td>
<td>3.13</td>
<td>11.10</td>
</tr>
<tr>
<td>Gossypium</td>
<td>Dicot</td>
<td>23,100</td>
<td>57</td>
<td>2.47</td>
<td>3.23</td>
</tr>
<tr>
<td>Pinus</td>
<td>Coniferales</td>
<td>21,866</td>
<td>23</td>
<td>1.05</td>
<td>–</td>
</tr>
<tr>
<td>Solanum</td>
<td>Dicot</td>
<td>14,083</td>
<td>12</td>
<td>0.85</td>
<td>1.75</td>
</tr>
<tr>
<td>Mesembryanthemum</td>
<td>Dicot</td>
<td>11,115</td>
<td>7</td>
<td>0.63</td>
<td>–</td>
</tr>
<tr>
<td>Secale</td>
<td>Monocot</td>
<td>6,574</td>
<td>3</td>
<td>0.46</td>
<td>16.55</td>
</tr>
<tr>
<td>Brassica</td>
<td>Dicot</td>
<td>3,627</td>
<td>2</td>
<td>0.55</td>
<td>1.60</td>
</tr>
<tr>
<td>Cryptomeria</td>
<td>Coniferales</td>
<td>2,293</td>
<td>4</td>
<td>1.74</td>
<td>–</td>
</tr>
<tr>
<td>Suaeda</td>
<td>Dicot</td>
<td>742</td>
<td>2</td>
<td>2.70</td>
<td>–</td>
</tr>
<tr>
<td>Saccharum</td>
<td>Monocot</td>
<td>495</td>
<td>1</td>
<td>2.02</td>
<td>–</td>
</tr>
<tr>
<td>Avena</td>
<td>Monocot</td>
<td>492</td>
<td>1</td>
<td>2.03</td>
<td>8.82</td>
</tr>
<tr>
<td>Capsicum</td>
<td>Dicot</td>
<td>252</td>
<td>1</td>
<td>3.97</td>
<td>8.00</td>
</tr>
<tr>
<td>Nicotiana</td>
<td>Dicot</td>
<td>130</td>
<td>3</td>
<td>23.08</td>
<td>5.85</td>
</tr>
<tr>
<td>Allium</td>
<td>Monocot</td>
<td>79</td>
<td>2</td>
<td>25.32</td>
<td>33.50</td>
</tr>
<tr>
<td>Vigna</td>
<td>Dicot</td>
<td>69</td>
<td>1</td>
<td>14.49</td>
<td>–</td>
</tr>
</tbody>
</table>
Published retrotransposon elements were used as query sequences in searches of plant EST accessions in the EST database. The matching ESTs for each retrotransposon are listed according to their source. The databases are not necessarily nonredundant.

<table>
<thead>
<tr>
<th>Element^a</th>
<th>Source species</th>
<th>Type</th>
<th>Total ESTs</th>
<th>Matching species</th>
<th>EST source tissue^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRE-1</td>
<td>Soybean</td>
<td>Copia</td>
<td>2</td>
<td>Soybean</td>
<td></td>
</tr>
<tr>
<td>Tgmr</td>
<td>Soybean</td>
<td>Copia</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>diaspora</td>
<td>Soybean</td>
<td>Gypsy</td>
<td>3</td>
<td>Soybean</td>
<td>– 1 2</td>
</tr>
<tr>
<td>Athila</td>
<td>Arabidopsis</td>
<td>Gypsy</td>
<td>1</td>
<td>Arabidopsis</td>
<td>1</td>
</tr>
<tr>
<td>AtRE1</td>
<td>Arabidopsis</td>
<td>Copia</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AtRE2</td>
<td>Arabidopsis</td>
<td>Copia</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ta1</td>
<td>Arabidopsis</td>
<td>Copia</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tat1</td>
<td>Arabidopsis</td>
<td>Gypsy</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Evelknievel</td>
<td>Arabidopsis</td>
<td>Copia</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>RetroLyc</td>
<td>L. peruvianum</td>
<td>Copia</td>
<td>6</td>
<td>Cultivated tomato</td>
<td>3 2 1</td>
</tr>
<tr>
<td>B1</td>
<td>Maize</td>
<td>?</td>
<td>1^c</td>
<td>Maize</td>
<td></td>
</tr>
<tr>
<td>Hopscotch</td>
<td>Maize</td>
<td>Copia</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Opie-2</td>
<td>Maize</td>
<td>Copia</td>
<td>21</td>
<td>Maize</td>
<td>2 9 3 5 2</td>
</tr>
<tr>
<td>PREM-2</td>
<td>Maize</td>
<td>Copia</td>
<td>17</td>
<td>Maize</td>
<td>– 8 1 7 1</td>
</tr>
<tr>
<td>Stonor</td>
<td>Maize</td>
<td>Copia</td>
<td>3</td>
<td>Maize</td>
<td>1 – 2</td>
</tr>
<tr>
<td>CentA</td>
<td>Maize</td>
<td>Gypsy</td>
<td>2</td>
<td>Rice</td>
<td>– 1</td>
</tr>
<tr>
<td>Cinful1</td>
<td>Maize</td>
<td>Gypsy</td>
<td>23</td>
<td>Maize</td>
<td>3 7 3 10</td>
</tr>
<tr>
<td>Cinful2</td>
<td>Maize</td>
<td>?</td>
<td>22</td>
<td>Maize</td>
<td>4 – 4 2 12</td>
</tr>
<tr>
<td>Grande1</td>
<td>Zea diploperennis</td>
<td>Gypsy</td>
<td>2</td>
<td>Maize</td>
<td>2 – – – –</td>
</tr>
<tr>
<td>Reina</td>
<td>Maize</td>
<td>Gypsy</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tekay</td>
<td>Maize</td>
<td>Gypsy</td>
<td>2</td>
<td>Maize</td>
<td>2 – – – –</td>
</tr>
<tr>
<td>Zeon-1</td>
<td>Maize</td>
<td>Gypsy</td>
<td>10</td>
<td>Maize</td>
<td>4 – 5 1 – –</td>
</tr>
<tr>
<td>RIRE1</td>
<td>Oryza australiensis</td>
<td>Copia</td>
<td>2</td>
<td>Maize</td>
<td>1 – – – –</td>
</tr>
<tr>
<td>RIRE2</td>
<td>Rice</td>
<td>Gypsy</td>
<td>6</td>
<td>Rice</td>
<td>– – 1 2 3</td>
</tr>
<tr>
<td>RIRE3</td>
<td>Rice</td>
<td>Gypsy</td>
<td>8</td>
<td>Gossypium arboresum</td>
<td>– – 1 – –</td>
</tr>
<tr>
<td>RIRE7</td>
<td>Rice</td>
<td>Gypsy</td>
<td>7</td>
<td>Rice</td>
<td>– – – – 1 1 –</td>
</tr>
<tr>
<td>RIRE8</td>
<td>Rice</td>
<td>Gypsy</td>
<td>5</td>
<td>Maize</td>
<td>– – 1 1 2 –</td>
</tr>
<tr>
<td>Wis2-1</td>
<td>Bread wheat</td>
<td>Copia</td>
<td>19</td>
<td>Bread wheat</td>
<td>– – 1 4 2 2 –</td>
</tr>
<tr>
<td>RetroSor1</td>
<td>Sorghum</td>
<td>Gypsy</td>
<td>4</td>
<td>Sorghum</td>
<td>– – – – 3 1 –</td>
</tr>
<tr>
<td>Levithan</td>
<td>Sorghum</td>
<td>LTR</td>
<td>1</td>
<td>Sorghum</td>
<td>– – – – 1 –</td>
</tr>
<tr>
<td>BARE-1</td>
<td>barley</td>
<td>Copia</td>
<td>23</td>
<td>barley</td>
<td>7 – 6 – – –</td>
</tr>
<tr>
<td>Bagy-1</td>
<td>barley</td>
<td>Gypsy</td>
<td>6</td>
<td>barley</td>
<td>– – 1 5 1 2 –</td>
</tr>
<tr>
<td>Bagy-2</td>
<td>barley</td>
<td>Gypsy</td>
<td>7</td>
<td>barley</td>
<td>– – 1 1 – –</td>
</tr>
<tr>
<td>Cereba</td>
<td>barley</td>
<td>Gypsy</td>
<td>5</td>
<td>barley</td>
<td>– – – – 2 –</td>
</tr>
<tr>
<td>Nikita</td>
<td>barley</td>
<td>LTR</td>
<td>1</td>
<td>barley</td>
<td>– – – – 1 –</td>
</tr>
<tr>
<td>Sabrina</td>
<td>barley</td>
<td>?</td>
<td>29</td>
<td>barley</td>
<td>22 – 3 – – –</td>
</tr>
</tbody>
</table>

(continues on facing page.)
copia-like classes find related and expressed elements in other grass species. Although all of the matches are to cereals, many of the matches cross tribe lines. This is particularly the case for the elements of barley and rice, whereas with the exception of CentA, the maize elements found matches only among maize ESTs.

Expression of BARE-Like Capsid Protein GAG Is Common in the Triticeae

Retroviruses and retrotransposons express their encoded proteins as a single polypeptide or as two (GAG and POL) separated by a frameshift (Fig. 1); these are processed into functional units by the aspartic proteinase present as part of the polypeptide itself (Garfinkel et al., 1991; Katz and Skalka, 1994). Earlier we demonstrated that GAG encoded by BARE-1 is translated and processed to the predicted mature size of 32.0 kD in dry and germinating embryos, leaves, and cell cultures of barley (Jääskeläinen et al., 1999). Oat (Avena sativa), bread wheat, and rye contain elements similar to BARE-1 (Pearce et al., 1997). The only reported full-length retrotransposon from which the EST was obtained: A, Leaf; B, shoot and root; C, flower and other reproductive organs; D, embryo and seed; E, seedling; F, stress-induced plants; and G, cultured cells.  

The antibodies recognize proteins of 150 kD and 90 kD on immunoblots of virtually all tested samples, including those of species outside the Triticeae (Fig. 2A). The 150-kD band is weak in lime grass (Leymus arenarius L. Hochst.) and bread wheat. The 150-kD protein corresponds to the 146.9 kD predicted for the unprocessed polypeptide of BARE-1. The 90-kD band corresponds to the mass predicted for the BARE-1 polypeptide following endoproteolytic cleavage of the reverse transcriptase-ribonuclease H domain. In addition, these antibodies recognized a 31.5-kD protein, matching the predicted size of the mature, proteolytically processed GAG (Jääskeläinen et al., 1999), as well as a 34-kD band from rice which may be its equivalent. The approximately 53- and 54-kD bands and weak 36-kD band in barley on the immunoblots also reacted with the pre-immune serum (shown on the right); otherwise, all reactions seen were specific.

In other experiments (Fig. 2B), immunoblots were reacted with the antibodies earlier generated to the N-terminal portion of GAG (Jääskeläinen et al., 1999). These antibodies recognized a 31-kD protein, a mass virtually identical to that of BARE-1 GAG, in all species tested; the lyme grass reaction was weak (lane 5) but detectable by eye. The oat protein was slightly smaller, 29 kD. Maize extracts contain a reacting protein of this size as well (not shown). The additional bands seen in the oat and rice lanes may represent intermediates in the proteolytic processing of the polypeptide. As is seen from the negative response to the pre-immune serum on the parallel blot, the recognition of the 29- to 31-kD proteins was specific. The antibody used in Figure 2B, earlier raised to only the amino-terminal half of BARE-1 (Jääskeläinen et al., 1999), visualizes the mature GAG but not the polypeptide. The two antisera were raised to nonidentical BARE-1 variants (85% identity in the overlapping, expressed GAG region) from the genomic population; their differing but specific protein recognition patterns may reflect variations in the processing kinetics of distinct BARE-1 subfamilies. The observed immunoresponse detected for rice is consistent with the RIRE1 retrotransposon of O. australiensis (accession no. D85597) and other Oryza spp. having been reported to be most similar to BARE-1 (Noma et al., 1997). An alignment of the RIRE1 polypeptide (GenPept accession no. BAA22288) to the GAG of BARE-1a shows 48.9% similarity, suggesting
that an active RIRE-1-like element may be responsible for the immunoreaction detected.

**BARE-1 Appears Integrationally Active in Many Grasses**

The creation by LTR retrotransposons of new joints with the flanking genomic DNA upon integration allows molecular fingerprints of the insertion pattern to be detected. Marker bands are generated by PCR using outward facing primers matching retroelement LTRs in combination with primers corresponding to dispersed genomic components, alternatively another retroelement (the inter-retrotransposon amplified polymorphism [IRAP] method; Kalendar et al., 1999), a microsatellite (the retrotransposon-microsatellite amplified polymorphism method; Kalendar et al., 1999), or a restriction site adapter (the sequence-specific amplified polymorphism [SSAP] method; Waugh et al., 1997). Because LTRs do not excise as part of retrotransposition, marker band polymorphisms are generally due to integration of new retrotransposon copies. We established earlier (Gribbin et al., 1999) that the SSAP method displays BARE-1 insertional polymorphisms throughout the Triticeae as well as in oat cultivars. Because SSAP polymorphisms may also be due to variability in the occurrence of restriction sites, we could not attribute them solely to retroelement mobility. Here, we have used a method relying only on LTR primers (IRAP) as well as looked for BARE-1 integrational polymorphisms in grasses distant from the Triticeae.

The BARE-1 element appeared among the EST database matches for three species in the tribe Triticeae (Table II), and we performed IRAP for BARE-1 on these and a range of other grass species (Fig. 3). The primers produced bands not only for all the Triticeae accessions and for timothy and oat, like the Triticeae in the subfamily Pooideae, but also for S. maritima and cordgrass of the subfamily Chloridoideae. Polymorphisms were observed between the pairs of, respectively, wheat cultivars, both Nordic spring types, rye lines, timothy lines, and cordgrass species. This suggests that BARE-1 has been integrationally active within these groups of cultivars or accessions since their divergence from their last common ancestor.
Rice, in the subfamily Erhartoideae, produced a single band, and maize, in the subfamily Panicoideae, produced only very weak bands (data not shown), although other barley retroelements are useful for IRAP in these species (our unpublished results).

BARE-like elements are translated in rice (Fig. 2), so the data suggest either that few are clustered or that the LTRs are dissimilar from those of BARE-1 of barley. Rye produces many bands with the BARE-1 primers; these are better resolved on sequencing gels than on agarose gels as in Figure 3.

A. tauschii represents the D-genome donor to bread wheat. A set of IRAP bands that are both monomorphic and shared between the A. tauschii and bread wheat accessions (Fig. 3) may represent integration events arising in the D genome before the advent of bread wheat.

**DISCUSSION**

Retrotransposons are well established as ubiquitous and highly prevalent components of plant genomes in general. These elements are nevertheless often referred to as “junk DNA,” implying that they are inert, in contrast to the genes required for cellular function. Their prevalence might be explained by ancient retrotransposition without requiring activity in the present. However, for some individual retroelements, evidence exists for transcription (Suoniemi et al., 1996b; Hirochika, 1997; Vernhettes et al., 1997; Okamoto and Hirochika, 2000), stress activation (Wessler, 1996; Grandbastien, 1998; Kalendar et al., 2000), translation (Hu et al., 1995; Jääskeläinen et al., 1999), and integration at specific loci (Johns et al., 1985; Grandbastien et al., 1989; Hirochika et al., 1996).

Here, we have undertaken to demonstrate that three components of the life cycle of active retrotransposons (transcription, translation, and integration) are widespread. First, we have made a systematic search for transcribed retrotransposon sequences in EST databases. Among the $7.8 \times 10^5$ ESTs searched, retrotransposons represent about 1.2% of transcripts, the frequency being somewhat higher among monocots than dicots. The frequency would be several-fold higher were only the identifiable ESTs (“hits”) to be considered. Retrotransposons from the grasses tend to match ESTs across multiple genera, whereas elements from the dicots tend to find matches only in their host species. These analyses cannot be regarded as exhaustive because even large EST databases may be 30% to 40% incomplete (Arabidopsis Genome Initiative, 2000; Penn et al., 2000).

Antibodies raised to BARE-1 GAG expressed in Escherichia coli recognized proteins in the seeds of virtually all species tested, not only barley and other grasses of the Triticeae, but also in species in different tribes and subfamilies of the Gramineae. The sizes of the bands detected are almost identical to those predicted for the BARE-1 polyprotein and the proteolytically processed, mature GAG, consistent with expression of polyproteins of size similar to those in the other species. These results are the first evidences for pools of retrotransposon polyproteins in plant cells. They show that BARE-like retrotransposons are translationally active and sufficiently well conserved for immunological cross reaction in a wide range of species in the Gramineae. The translation of BARE-like elements in other grasses is consistent with the evidence from the EST database searches that transcriptionally active retrotransposon families are shared among the grasses.

The third line of evidence that grass genomes share families of active retrotransposons is the demonstration that a retrotransposon originally identified in barley can generate polymorphic marker bands in distant species. The IRAP products result from two retrotransposons near enough to each other to permit amplification of a PCR fragment between them. Their
prevalence is consistent with the observed retroelement clusters in grass genomes (SanMiguel et al., 1996; Ananiev et al., 1998; Panstruga et al., 1998; Manninen et al., 2000). The IRAP polymorphisms reported here indicate integration events subsequent to the last common ancestors of lines or cultivars outside barley and the tribe Triticeae. Similar findings for other retrotransposons, including gypsy-like Bagy-1, Sukkula, and others (our unpublished observations), together with the EST and translation data here, suggest that the broad activity of retrotransposon families across the grasses may be a general phenomenon.

In recent years, the large-scale syntenic nature of the grass genomes has been recognized and applied (Ahn and Tanksley, 1993; Kilian et al., 1997). Exceptions to the microcolinearity of genes can be due to insertion of transposable elements (Benetzen, 2000); recombinational loss of sequences intervening between the LTRs of retrotransposons (Shirasu et al., 2000) may help explain other exceptions to microcolinearity caused by deletions. Given synteny and active, shared families of retrotransposons, the grasses may be well suited for a comparative approach to the understanding of the impact of retrotransposons on the genome. Examination of the changes in genome organization among the grasses wrought by specific retrotransposon families may help untangle shared mechanisms of propagation and regulation from the contingent history of a these families in any given species.

MATERIALS AND METHODS

Plant Materials

Barley (Hordeum vulgare var. Himalaya) was obtained from Washington State University (Pullman). Rye (Secale cereale var. Riihi), spring bread wheat (Triticum aestivum vars. Tjälve and Mahti), and durum wheat (Triticum durum) were gifts of the Department of Food Technology, University of Helsinki (Finland). Rye lines P105 and P87 were gifts of Tamara S. Schulko (Academy of Sciences, Minsk, Belarus). Oat (Avena sativa var. Veli) and timothy grass (Phleum pratense) lines 22 and 16 were gifts of Boreal Plant Breeding Ltd. (Jokioinen, Finland). Aegilops tauschii lines 1691 and 1704 were gifts of Bikram Gill (Kansas State University, Manhattan). Spartina maritima (Curtis) Fernald and cordgrass (Spartina alterniflora Loisel) are from Malika Ainouche (University of Rennes, France). Inbred maize (Zea mays) lines Oh43 and Mo17 were gifts of the U.S. Department of Agriculture, Agricultural Research Service, North Central Regional Plant Introduction Station, Iowa State University, Regional Plant Introduction Station (Ames). Lyme grass (Leptus arenarius L. Hochst.) with a provenance in Eyyarbakkí, Iceland was a gift of Kesara Anamthawat-Janssens (University of Reykjavik, Iceland). Rice (Oryza sativa) line IRRI 52886 was from the International Rice Research Institute (Los Baños, Laguna, The Philippines).

EST Database Searches

The plant accessions found in the dbEST division of the combined GenBank (release 120.0), EMBL, and DNA Data Bank of Japan databases were searched for similarities to retrotransposons in two ways. First, we selected all entries in the overall nonredundant, combined nucleic acid database containing the words “retrotransposon,” “copia-like,” or “gypsy-like” in the descriptor, as well as all entries of previously published plant retrotransposons. We then selected the parts of these sequences corresponding to retrotransposons. These were used as the query sequences against the dbEST database with the Advanced BLAST program using a cutoff value of 0.0001. Second, for those general database entries having a putative translation, we queried the EST databases using the TBLASTN program applying a cutoff value of 1.0. All searches were done using the online service of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/blast/blast.cgi).

Immunoblotting

Two anti-GAG antisera were used: one (Fig. 2B) was identical to that used earlier (Jääskeläinen et al., 1999), and the other (Fig. 2A) was prepared as follows. The region corresponding to the full-length BARE-1 GAG was isolated from a gag clone (accession no. AJ295226) by amplification with the forward primer 5′ GTG TGT AGA CAT ATG GCT CGC GGA GTA GC 3′ (start codon underlined) and the reverse primer 5′ GAC ATG TGG ATC CAA TAT ACC TCA TTT TTC 3′ (stop codon underlined). The forward primer introduced a NdeI site (CATATG) and the reverse primer a BamHI site (GGATCC). The reaction product was digested with BamHI and NdeI and cloned into the vector pET14b (Novagen, Madison, WI). This vector tags the N terminus of the protein with a poly-His stretch. The construct was expressed in Escherichia coli and purified on a TALON metal affinity resin (CLONTECH, Palo Alto, CA) column under denaturing conditions according to the manufacturer’s instructions. The expressed protein was further purified by SDS-PAGE electrophoresis prior to injection into the rabbit. The antisera was raised and antibodies purified as previously described (Jääskeläinen et al., 1999). Pre-immune antisera were from the same rabbits and purified as for the anti-GAG antisera.

For immunoblotting, embryo halves of the various seeds were isolated and pulverized under liquid Nz2, then extracted in 50 mM Tris-HCl (pH 7.5), 10 mM EDTA, 10 mM CHAPS [3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate], 15 mM KCl, 5 mM MgCl2, 3 mM dithiothreitol, 10 μM trans-epoxysuccinyl-L-leucylamido-(4-guanidino) butane, 4 μM Pepstatin, and 2 μM Leupeptin. The protein contents of the extracts were determined (Bio-Rad Protein Assay, Bio-Rad, Hercules, CA) and an equivalent of 20 μg protein for each sample separated by SDS-PAGE electrophoresis. Electrophoresis, blotting, and immunoreactions were carried out as previously described (Jääskeläinen et al., 1999).
IRAP Polymorphism Detection

The IRAP markers were generated as before (Kalander et al., 1999; Manninen et al., 2000) in a thermocycler (Master Cycler Gradient, Eppendorf AG, Hamburg, Germany) in 0.2-mL tubes (AB-0337, ABgene, Epsom, Surrey, UK). The 20-μL reactions contained 75 mm Tris-HCl (pH 8.8), 20 mm (NH₄)₂SO₄, 1.5 mm MgCl₂, 0.01% (v/v) Tween 20 (polyoxyethylene-sorbitan), 20 ng DNA, 200 nm LTR primers, 200 μM dNTPs, and 1.2 units thermostable DNA polymerase (FIREPol, Solis Biodyne, Tartu, Estonia). The BARE-1 LTR primer consisted of 5’ TCC CAT GCG ACG TTC CCC 3’, matching nt 2,116 to 2,133 of accession number Z17327 at 4 nt from the 3’ end of the LTR. The template DNA was isolated as earlier described (Kalander et al., 1999). The reaction mixture was heated to 94°C for 2 min, then 30 cycles were carried out, which were as follows: 94°C, 20 s; 60°C, 20 s; and 72°C, 2 min. The reaction was terminated by a final extension at 72°C for 10 min followed by maintenance at 4°C. One-fifth of the reaction mixture was analyzed by gel electrophoresis, carried out in 2% (w/v) agarose (RESolute LE agarose, BIOzym, Landgraaf, The Netherlands) at 80 V for 7 h and visualized by staining with ethidium bromide.

ACKNOWLEDGMENTS

The expert technical assistance of Anne-Mari Narvanto (University of Helsinki) is deeply appreciated. The involvement of Alex Baumel and Malika Ainouche (University of Rennes, France) in developing the IRAP markers for Spatina spp. is gratefully acknowledged.

Received November 13, 2000; returned for revision December 8, 2000; accepted December 20, 2000.

LITERATURE CITED

Kilian A, Chen J, Han F, Steffenson B, Kleinhofs A (1997) Towards map-based cloning of the barley stem rust re-


Vicient et al.