1056–1062 *Nucleic Acids Research, 1998, Vol. 26, No. 4*

A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome

Ralph Panstruga, Rainer Büschges, Pietro Piffanelli and Paul Schulze-Lefert*

The Sainsbury Laboratory, John Innes Centre, Norwich Research Park, Colney Lane, Norwich NR4 7UH, UK

Received September 26, 1997; Revised and Accepted December 22, 1997 DDBJ/EMBL/GenBank accession no. Y14573

ABSTRACT

The contiguous DNA sequence of a 60 kb genomic interval of barley chromosome 4HL has been assembled. The region harbours a single and novel *gypsy*-like retrotransposon, designated *BAGY-1*. Only three genes appear to reside in the genomic stretch. One predicts a plant homologue of ribophorin I, a subunit of the oligosaccharyltransferase–protein complex located in the rough endoplasmatic reticulum. The second is similar to the *Drosophila g1* gene encoding a ring finger protein involved in developmental processes. The observed gene density is ~5-fold lower than in the best characterized dicot genome of *Arabidopsis* but 6- to 10-fold higher than expected from an equidistant gene distribution in the complex barley genome. Our data suggest that the 60 kb genomic interval represents part of a gene island, a seemingly distinctive feature of grass genomes.

INTRODUCTION

Improved DNA sequence analysis facilities make large contiguous areas of plant genomes amenable for detailed investigation. This is particularly instructive for the model dicot plant *Arabidopsis thaliana*, for which a considerable amount of contiguous genomic sequence data is now publicly available. The data have provided good insights into the organization and architecture of one of the smallest plant genomes (1–3, http://genome-www.stanford.edu/Arabidopsis).

Monocotyledonous species represent the second major class of flowering plants and were separated from dicots ~120–200 million years ago (4). Grasses (*Poaceae*) represent a major family of monocotyledonous species. Grass genomes exhibit an enormous variability in DNA content, ranging in diploid species from 415 Mb for rice to 5300 Mb for the barley genome (5). However, the linear order of genes has been surprisingly conserved during speciation of grasses (6,7) and there is no evidence that the total number of genes varies substantially.

If genome size among diploid grass species can vary >10-fold but gene number and gene order are conserved, how are genes distributed in complex grass genomes such as maize, barley or wheat? Also, what is the nature and structure of the non-coding sequences? From diagnostic sequencing within a 280 kb maize genomic interval it has been concluded that >50% of the maize genome may consist of retrotransposon DNA (8) but it is not known whether this can be extrapolated to other grass genomes. There are two main scenarios as to how genes could be distributed in complex grass genomes (9). The first is a scattered distribution of genes among non-coding sequences leaving very long distances between genes. The second possibility is a clustering of genes between extended stretches of non-coding DNA. Density gradient fractionation of grass genomic DNA and analysis of the fractionated molecules with a large number of cDNA probes provided evidence for the second model indicating a clustering of genes (10,11). However, the lack of data of large contiguous genomic DNA sequences from complex grass genomes prevented detailed insights into the structure of these gene clusters.

We reported previously the map-based isolation of the barley *Mlo* gene (12,13). During the positional cloning procedure we obtained partial sequence information of a genomic 60 kb long insert of a BAC clone (BAC F15) containing *Mlo*. We sequenced the entire interval to obtain insights into the genomic organization and gene density of a complex grass genome. Here we present the analysis of the contiguous 60 kb stretch and compare our results with those obtained from similar analyses in the grass genome of maize and the dicot genome of *A.thaliana*.

MATERIALS AND METHODS

The DNA sequence of BAC F15 has been deposited at DDBJ/EMBL/GenBank under accession number Y14573.

The construction, isolation and initial sequence analysis of BAC F15 has been described before (12,13). Gaps in existing DNA sequence contigs were closed either by primer walking on plasmid subclones of BAC F15 or by applying polymerase chain reaction (PCR) using sets of specific primers. PCR products were purified by means of ‘Wizard PCR Preps’ DNA purification system (Promega). DNA sequencing was performed as described (12).

Analysis of the completed sequence of the insert of BAC F15 was done using programs of the Genetics Computer Group (GCG) or the STADEN software package for Unix users (fourth edition, 1994). Some analysis for presence of coding sequences and evaluation of deduced protein sequences has been performed using the ABIM online analysis tools (http://www-biol.univ-mrs.fr/english/logligne.html).

*To whom correspondence should be addressed. Tel: +44 1603 452571; Fax: +44 1603 250024; Email: schlef@bbsrc.ac.uk

Table 1. ESTs with high sequence similarity to insert of BAC F15

Species	DDBJ/EMBL/GenBank accession no.	High Score (BLAST)	Probability P (N)	Homology to putative genes and retrotransposons identified on BAC F15
Rice	C19562	665	3.6×10^{-80}	ribophorin gene
Rice	D24495	612	2.6×10^{-72}	ribophorin gene
Arabidopsis	T42928	555	5.3×10^{-34}	ribophorin gene
Arabidopsis	Aa395108	456	1.2×10^{-25}	3′ end of ribophorin gene
Arabidopsis	T42673	426	3.3×10^{-23}	ribophorin gene
Drosophila	Ac000747	349	1.7×10^{-16}	*BAGY-1*
Fugu rubripes	Z90168	287	1.6×10^{-11}	*BAGY-1*
Homo sapiens	Aa159170	277	4.1×10^{-10}	ribophorin gene
Drosophila	Ac00748	273	1.3×10^{-8}	*BAGY-1*
Arabidopsis	T76900	238	2.3×10^{-7}	*g1* homologue
Arabidopsis	R90681	233	5.8×10^{-7}	*g1* homologue
Rice	D41213	201	2.5×10^{-4}	5′ end of ribophorin gene

ESTs with a BLAST high score > 200 are shown.

Moving averages of observed/expected CpG dinucleotide ratios were calculated and plotted for 100 nucleotide (nt) windows across the 60 kb genomic sequence in 1 nt increments using the CpG plot program available in GCG package version 8.0. The observed/expected CpG average values of the entire 60 kb sequence and subregions were calculated by applying the equation described in Gardiner and Frommer (14).

RESULTS

Gene density

The DNA sequence of the insert of BAC F15 consists of 59 748 bp. To estimate gene density in this genomic interval we applied three criteria to search for genes: (i) homology to characterized genes or expressed sequence tags (ESTs) in the public databases; (ii) occurrence of extended high coding probabilities and (iii) application of a gene finder program (BCM gene finder). By combination of these means we identified three genes in the 60 kb insert that each matched at least two of the three criteria. These were the *Mlo* resistance gene, a plant homologue of the human ribophorin I gene and a gene encoding a ring finger protein best characterized in *Drosophila*. While *Mlo* has been described previously in detail (12), the latter two genes have not been reported in plants before.

The ribophorin I homologue is located within an ~3 kb long region which shows extensive homology to several EST sequences (DDBJ/EMBL/GenBank accession nos C19562, D24495, T42673 and T42928) from *Arabidopsis* and rice (Fig. 1A and Table 1). In addition, this interval exhibits a significantly elevated coding probability predominantly at its 5′ end. Sequence assembly and translation into the respective amino acid sequence revealed homology to ribophorin I, a type I integral membrane protein of the rough endoplasmic reticulum, characterized previously in human, rat and baker's yeast (15–17). The polypeptide represents one of the three subunits of the mammalian oligosaccharyltransferase holoprotein (18; in accordance with the gene designation in yeast we have named the barley homologue *Ost1*). We were unable to predict precisely the 3′ terminal sequence of the barley gene due to low sequence conservation to the mammalian and yeast homologues in this region and because the plant ESTs do not cover the 3′ end (Fig. 2). The predicted truncated amino acid sequence (470 amino acids) has 61.8% similar and 38.3% identical residues in comparison with the human ribophorin I polypeptide, which consists of 608 amino acids. The similarity/identity with yeast *OstI* (476 amino acids) is 48.5 and 24.0%, respectively (Fig. 2). Apart from the sequence similarity, the barley ribophorin I protein shares several characteristic features at conserved positions with the mammalian and yeast homologues such as a cleavable N-terminal signal sequence, asparagine-glycosylation sites and a single transmembrane helix (15–17).

Two ESTs, Aa395108 and D41213, align to opposite ends of the barley ribophorin I homologue (Table 1). The former could be part of the 3′ terminus of the barley ribophorin I gene since the genomic sequence exhibits also a high coding probability. In contrast, this is unlikely for D41213 exhibiting only a weak relatedness to the barley DNA sequence.

A long open reading frame (ORF) of 1047 bp is predicted at ~33 kb of BAC F15 (Fig. 1A). A region of 290 bp revealed 60% identity to two *Arabidopsis* ESTs (DDBJ/EMBL/GenBank accession nos T76900 and R90681). This ORF is located in a region of very high coding probability (Fig. 1B) and is therefore likely to be translated. Thorough analysis of the two related *Arabidopsis* ESTs and the barley coding sequence indicates a relatedness to the *Drosophila* g1 protein (19). This protein is supposedly involved in developmental processes and contains a modified ring finger motif found in several regulatory proteins. The overall homology to the g1 protein suggests that the detected ORF represents the complete coding region of the barley *g1* homologue, starting with an ATG triplet at nt 33104. The deduced amino acid sequence (295 amino acids) contains 50.5% similar and 25.7% identical residues in comparison with *g1* (284 amino acids). Each of the amino acids believed to be essential for a ring finger structure are conserved in the barley homologue (Fig. 3). Also, *g1* and the barley homologue share a single putative

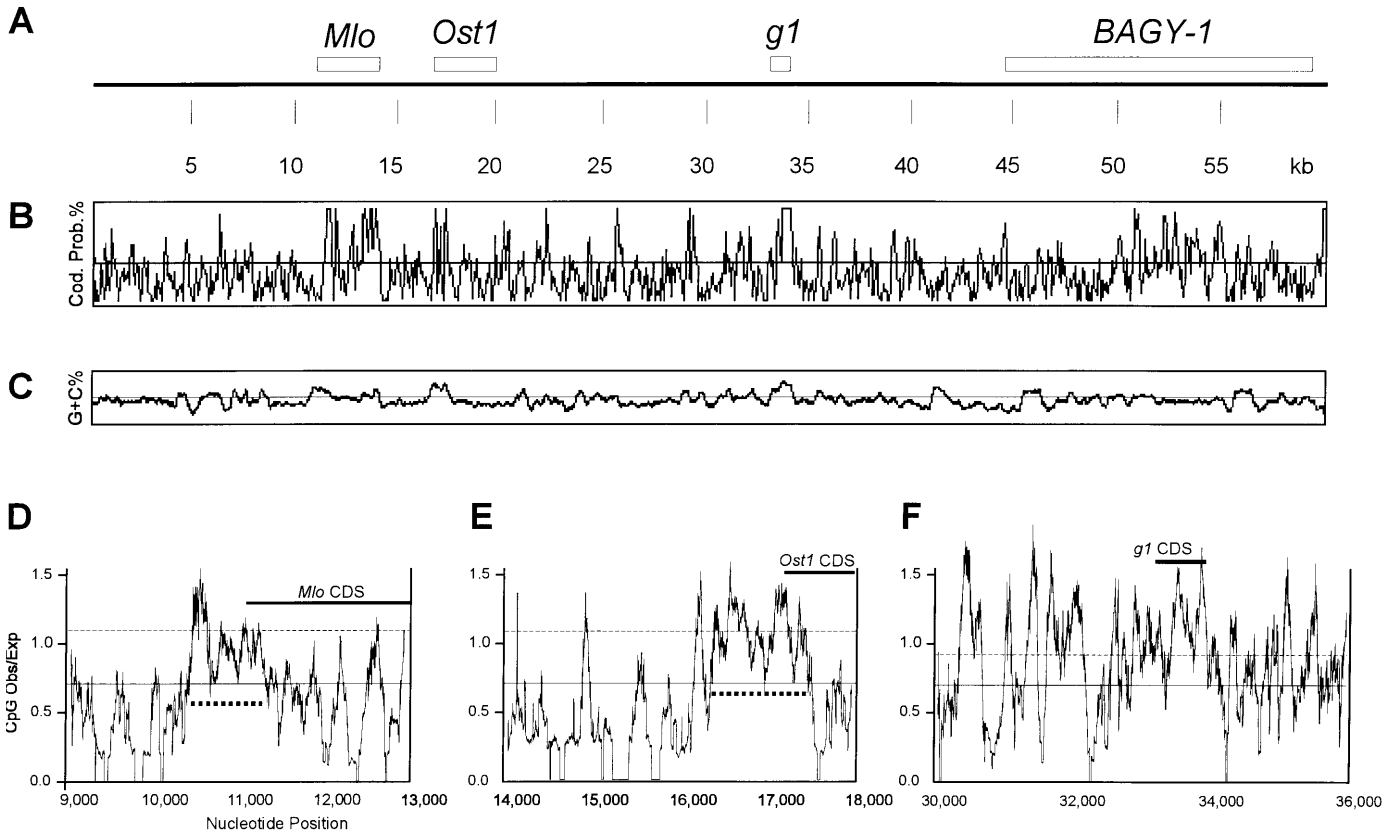

Figure 1. Graphical representation of 60 kb genomic sequences in BAC F15. (**A**) The long solid bar depicts the 60 kb insert of BAC F15. The numbers below indicate physical distances in kb. The positions of the three identified genes and the retrotransposon are depicted by open rectangles. *Ost1*: oligosacharyltransferase I homologue; *g1*: homologue of *Drosophila* g1 protein; *BAGY-1*: *gypsy*-like retrotransposon. *Mlo* has been described previously (12). (**B**) Coding probability of the insert of BAC F15. The peaks plotted in the box show the relative coding probability calculated by the uneven positional base frequencies method. The horizontal line in the box marks a threshold level indicating that peaks above this line represent in 76% coding regions. The plot was established with the following parameters: odd span length = 201, plot interval = 5. (**C**) G+C content of the insert of BAC F15. The plot was established with the following parameters: odd span length = 201, plot interval = 5. The horizontal line in the box marks a 50% G+C content. (**D**, **E** and **F**) Ratio of observed/expected CpG dinucleotides upstream of *Mlo* (D), adjacent to the ribophorin I homologue (*Ost1*, E), and adjacent to the *g1* homologue (F). Numbers given below each graph correspond to nt positions shown in (A). The average ratio of 0.79 observed/expected CpG for the 60 kb interval is marked in each graph by a hairline. Dotted hairlines indicate the average ratios of observed/expected CpG calculated for two putative CpG islands [marked by bold dotted lines in (D) and (E)] upstream of *Mlo* (nt 10 200–11 200; ratio 1.12) and the ribophorin I homologue (nt 16 200–17 200; ratio 1.26). The dotted hairline in (E) indicates the average ratio of observed/expected CpG dinucleotides for the entire 6 kb interval shown (0.93). Bold lines indicate the positions of coding sequences (CDS) for each gene.

transmembrane helix (amino acid residues 126–146) as revealed by hydropathy analysis (Fig. 3).

Retrotransposons

Three ESTs representing transcript sections of active retroelements from *Drosophila* and *F.rubripes* showed sequence similarity to a region around 52.5 kb of BAC F15 (Ac000747, Ac000748, Z90168; Table 1). They were first evidence of a single retrotransposon within the 60 kb of BAC F15, located within position 44–59 kb. It has been identified by the presence of long terminal repeats (LTRs) and by sequence relatedness of its internal region to other plant retrotransposons. The element represents a novel type of barley retrotransposon which we designated *BAGY-1* (barley *gypsy*-like retrotransposon). One further stretch with obvious homology to the barley retrotransposon *BARE-1* (20) turned out to represent a genomic sequence insertion of unknown origin in the 3′ LTR of *BARE-1* (Table 2).

BAGY-1 has a length of 14 424 bp and is flanked by two LTRs of 4202 bp (5′LTR) and 4208 bp (3′LTR) (Fig. 4). Both LTRs are 94% sequence identical. The 6% sequence differences result from single nt exchanges (transitions and transversions) and from 1-nt insertions/deletions. In contrast with many other retroelements (21) the LTRs are not bordered by short inverted repeats (IRs) but are flanked by imperfect direct repeats (DRs) consisting of 5 bp (5′-GTATG-3′ and 5′-GTATT-3′, Fig. 4). These were presumably created upon the insertion of the retroelement at this site of the barley genome. The common terminating nucleotides of many retrotransposons (5′-TG...CA-3′; 20,21) are apparently modified to 5′-TG...CC-3′ in *BAGY-1*. Next to the 5′LTR is a potential primer binding site (PBS) complementary to the 3′ end of a methionine initiator tRNA of wheat (Fig. 4). This tRNA is supposed to be virtually identical to the barley methionine initiator tRNA (20) but has not been cloned so far. A sequence of several consecutive purine bases, the so-called polypurine tract (PPT), another common structural element of retroelements, is

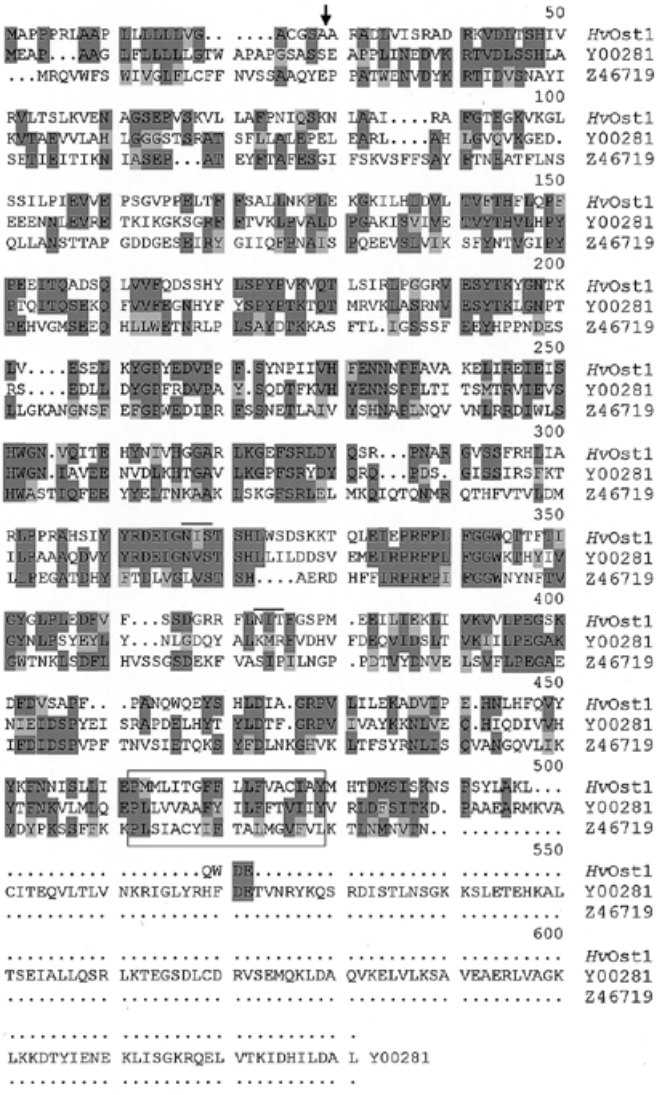

Figure 2. Amino acid sequence alignment of ribophorin I homologues. The deduced incomplete barley ribophorin I (*H.vulgare*, *Hv*Ost1), the human ribophorin I (*H.sapiens*, Y00281) and the bakers yeast Ost1 (*S.cerevisiae*, Z46719) sequence were aligned by the Pileup program of GCG. Dark grey boxes indicate identical residues, light grey boxes similar residues. A putative cleavage site for the predicted N-terminal signal peptide is indicated by an arrow, a predicted transmembrane helix is boxed and putative glycosylation sites are indicated by lines. The presence of the signal peptide, its cleavage site and the location of the membrane-spanning helix were predicted by PSORT (ABIM online analysis tools).

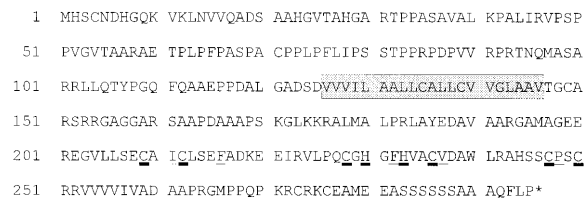

Figure 3. Deduced amino acid sequence of the barley *g1* homologue. The stop codon is marked by an asterisk, the putative transmembrane helix by a grey box. Essential C and H residues of the modified ring finger motif are underlined in bold, other conserved amino acids in this region are marked with thin lines [conserved residues according to Bouchard and Côté (19)].

located immediately upstream of the 3′LTR (11 out of 13 nt are purines; Fig. 4).

The internal region of *BAGY-1*, 6014 bp in size between the LTRs, exhibits at its 3′ end high sequence similarity to retrotransposon-derived sequences from other species (Table 2) encoding the putative integrase/endonuclease of these elements. Close inspection of ORFs in this region revealed all conserved motifs of retroelements, the RNA binding (*gag*), protease (*prot*), reverse transcriptase (*RT*), RNAse H (*RNAse H*) and integrase (*int*) domains. The linear order of the domains in *BAGY-1* (*gag*/*prot*/*RT*/*RNase H*/*int*) is characteristic for members of the Ty3/*gypsy* class of retroelements (21–25). However, to uncover the characteristic assembly of these retrotransposon domains in a single contiguous ORF, we had to presume two frame shifts at the nt level. In addition, the coding region of the predicted polyprotein is interrupted by eight stop codons and therefore not functional (Fig. 5).

Intergenic sequences

Intergenic sequences (stretches of DNA which predict no genes or retrotransposons) account for 64.3% of the analyzed interval. This includes regulatory 5′- and 3′-flanking sequences of the genes as well as ‘stuffer’ DNA between genes and the retrotransposon. With the exception of a complex repeat structure, the intergenic sequences reveal no obvious structure. The repeat structure, located at ~27 kb, is composed basically of two sets of repeats (Fig. 6). The first set consists of two 206 bp blocks with 100% sequence identity in tandem orientation and 1 bp overlap. The second tandem repeat set is juxtaposed to the former, and consists of 152 bp units with 100% sequence identity which are separated by 98 bp. These 98 bp represent a third truncated version (3′ half) of the first repeat structure. The 152 bp repeat unit is in turn part of a 442 bp segment which is highly homologous to a region upstream of the barley aleurain gene (Table 2).

BLAST analysis (Table 2) identified one further sequence block in the upstream region of the barley lipid transferase protein, LTP, exhibiting clear sequence relatedness to an intergenic sequence of BAC F15 (nt 711–1386). While the homology to the above mentioned aleurain upstream sequence is contiguous, the homology to the LTP upstream region is non-contiguous due to several short interspersed stretches.

Nucleotide composition and CpG dinucleotide frequencies

The average G+C content of the 60 kb region is 46.0%. The distribution of G+C in the interval is not homogenous but some short sections display remarkable high or low fluctuations from this mean value (Fig. 1C). The G+C content of the three genes embedded in the 60 kb interval varies considerably from 42.4–45.7% (complete genomic sequence/coding sequence only) for the ribophorin I gene and 53.9–59.6% for *Mlo* to 70.8% for the *Drosophila* g1 homologue (no intron). The identified retroelement consists of 46.4% G+C, essentially the same in both LTRs (46.7 and 46.3%) and the internal region (46.2%). This compares with a G+C content of 44.9% for intergenic sequences. Thus, both the retroelement and intergenic sequences have a G+C content very similar to the average value of the 60 kb interval.

Table 2. Related plant sequences of DDBJ/EMBL/Genbank database entries to insert of BAC F15

Entry name	DDBJ/EMBL/GenBank accession no.	High score (BLAST)	Probability P (N)	Area of homology
Barley gene for thiol protease aleurain	X05167	2138	2.6×10^{-190}	5′ upstream region
Hordeum vulgare mRNA for Mlo protein	Z83834	2131	0.0	coding region
Thinopyrum bessarabicum RAPD marker	U43516	1583	1.7×10^{-167}	(not applicable)
Hordeum vulgare DNA for BARE-1 copia-like retroelement	Z17327	1472	0.0	insertion in 3′LTR
Lilium henryi del transposon	X13886	1135	2.2×10^{-154}	integrase motif
Nicotiana alata retrotransposon	Z35426	932	4.6×10^{-65}	integrase motif
H.vulgare gene for LTP 1	X60292	610	7.5×10^{-137}	5′ upstream region

Sequences with a BLAST high score of >600 are shown.

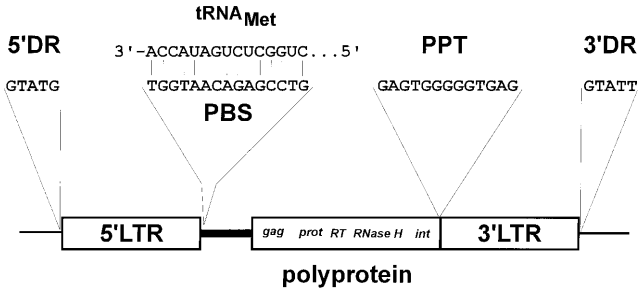

Figure 4. Structure of the barley retroelement *BAGY-1*. The boxes at the left and right side indicate LTRs, the box in the middle the region of the polyprotein encoding the RNA binding (*gag*), protease (*prot*), reverse transcriptase (*RT*), RNase H and integrase (*int*) domains. Positions of a putative primer binding site (PBS) complementary to the methionyl initiator tRNA ($tRNA_{Met}$) of wheat and the polypurine tract (PPT) are indicated. The 5′ and 3′ DRs are also indicated.

Next we tested the frequency and distribution of CpG dinucleotides in the 60 kb interval (Fig. 1D–F). CpG dinucleotides are 4-fold underrepresented in bulk vertebrate genomic DNA but clusters of 1–2 kb CpG islands are regarded as gene markers in vertebrates (26). The average observed/expected CpG frequency of the 60 kb interval is 0.79, significantly higher than in vertebrate genomes but similar to 0.75 reported for a sample of 53 monocot and dicot genes (14). A close inspection of the 60 kb barley interval revealed only two stretches longer than 1 kb with an unusually high observed/expected CpG ratio of 1.12 and 1.26. These are located immediately upstream of the coding sequences of the *Mlo* and the ribophorin I homologue genes, respectively (Fig. 1D and E). The two stretches include the first exon of both genes, a characteristic feature of CpG islands in vertebrates (27). In contrast, no comparable extended CpG cluster was found adjacent to the *g1* homologue (Fig. 1E). The region flanking *g1* is characterized by extreme fluctuations of observed/expected CpG frequencies, reflected by a higher ratio in a 6 kb interval (0.93) compared with the average of 0.79 for the entire 60 kb.

DISCUSSION

The 60 kb stretch of barley chromosome 4HL provides the largest consecutive DNA sequence from a complex grass genome reported so far, enabling us to gain insights in gene density and patterns of intergenic sequences. In maize, a region of 280 kb around the *Adh1-F* and *u22* genes was analyzed by diagnostic sequencing and did not result in a single sequence contig (8). It revealed that this region of the maize genome is mainly composed of retrotransposons inserted within each other accounting for >60% of the investigated interval and for at least 50% of the entire maize genome. In contrast, in barley, we identified a single retroelement in the 60 kb stretch which, due to its unusually large size of 14.4 kb, accounts for 24% of the entire interval. Thus, the occurrence of retrotransposons in complex grass genomes is either not homogenous, leading to clusters and retroelement-poor regions, or they account for much less of the barley genome than in maize.

Our data indicate that apart from retrotransposons other short repeat structures are likely to contribute to the complexity of the barley genome. Since only very few genomic barley sequences are available in the databases, it is remarkable that the 60 kb interval harbours two different sequence repeats sharing very high sequence relatedness to upstream regions of two barley genes at other chromosomal sites. It is unclear how these short sequence repeats have spread across the barley genome.

Based on the utilization of coding probability, database searches and a gene finder program, we have located only three genes in the investigated genomic interval resulting in an average gene density of one gene per 20 kb. This ratio is remarkably different from two chromosomal regions of the dicot *A. thaliana* in which genes are separated on average by 4–5 kb (1,2). The findings are corroborated by current DNA sequence analysis of large genomic contigs of the *A. thaliana* genome (http://genome-www.stanford.edu/Arabidopsis). Our gene density estimate needs to be interpreted with restrictions. Firstly, it is unclear to what extent the tested genomic interval is representative of the rest of the barley genome. Secondly, limitations of computer-aided gene identification are well known (28).

Evidence suggests that all evolutionary closely related grass genomes contain a comparable number of genes in the same order despite >10-fold differences in their genome size (6,7). Current estimates of gene number in higher plants vary between 25 000 and 43 000 (29). Since barley has a genome size of 5300 Mb/haploid genome (5), a gene density of one gene per 123–212 kb can be expected if genes are distributed at equal distance. However, the observed gene density in the 60 kb interval exceeds this calculation by a factor of 6–10 suggesting a clustering of genes. Experimental evidence obtained by a different approach supports the idea of gene clusters in grass genomes (10,11,30). By testing analytical CsCl profiles of genomic DNA in the 50–100 kb size

Figure 5. Hypothetical polyprotein encoded by *BAGY-1*. Stop codons are marked by asterisks and two reading frame shifts are indicated by arrows above the sequence. Conserved protein motifs according to Smyth *et al.* (23) are underlined. Abbreviations of protein domains given on the right are described in Figure 4.

range from barley, maize and rice with a large number of gene probes only a small percentage of each genome (12, 17 and 24%), in each case characterized by a well defined G+C content, was found to harbour coding regions (11). It was concluded that grass genomes are characterized by a compositional compartmentalization with gene islands (also termed 'gene space') of 100–200 kb. In contrast, genes in the similarly complex human genome are scattered among G+C-rich and G+C-poor regions (31). Unfortunately, comparable data are not available for the *Arabidopsis* genome.

Transformation of the buoyant densities corresponding to the barley gene islands into G+C values (1.7017–1.7025 g/cm^3; 11) result in a range of 45.8–46.7% G+C. Strikingly, the mean G+C value of the analyzed 60 kb interval is 46.0% and matches precisely the predicted narrow 0.8% G+C 'window' of barley gene islands. This is noteworthy since each of the three identified genes in the 60 kb interval shows clearly different G+C values (42.4, 53.9 and 70.8%, respectively), a surprising fluctuation which is also documented for genes in other grass species (11). In conclusion, the observed gene density and the G+C value of the 60 kb interval suggest it to represent a stretch of a gene island.

Fluctuations in the G+C content of individual genes must somehow be compensated by intergenic sequences to explain the observed uniform base composition in 100–200 kb gene space regions. In maize, it has been proposed that this balancing is due to the presence of multiple interspersed retroelements (11). This hypothesis presupposes that retrotransposons exhibit a G+C content close to the G+C range of gene islands in each grass species. The observed 46.4% G+C value of *BAGY-1* matches precisely the narrow 0.8% G+C 'window' of the barley gene space. Interestingly, the previously described *BARE-1* element has a G+C value of 47% (20), very close to the predicted 0.8% G+C 'window', further supporting this relationship.

Vertebrate genomes are characterized by the existence of CpG islands, GC-rich (60–70%) 1–2 kb short segments usually located immediately upstream of genes (27). CpG dinucleotides in these islands are unmethylated, in contrast with the bulk genomic DNA which is methylated as 5mCpG. CpG islands are regarded as signposts within vertebrate genomes, possibly providing an 'open' chromatin stucture allowing access of transcription factors to gene promoters. In addition, CpG dinucleotides are ~4-fold underrepresented in the bulk of vertebrate genomic DNA (27). In plants, evidence for the existence of unmethylated CpG islands has been obtained in maize, wheat and barley (32–34). Our analysis of the 60 kb contiguous barley genomic DNA revealed an observed/expected ratio of CpG content of 0.79, markedly larger compared with 0.2 for human genomic DNA (27). However, the observed/expected ratio of CpG in the 60 kb barley interval is strikingly similar to the 0.77 ratio reported for bulk wheat genomic DNA (35). Our findings corroborate data derived from the analysis of a number of monocot and dicot genomic sequences, indicating that higher plant genomes generally contain CpG dinucleotides at a frequency much closer to the expected level than do vertebrate genomes (0.75 and 0.23, respectively; 14,27). Thus, in contrast with vertebrate genomes there is little suppression of CpG dinucleotides in plants. The absence of a CpG deficiency makes the detection of CpG islands by DNA sequence analysis alone more difficult.

Despite this difficulty, we found extended DNA stretches exhibiting a noticeably elevated observed/expected CpG dinucleotide ratio (>1.0) immediately upstream of *Mlo* and the ribophorin I homologue but not adjacent to the *g1* homologue (Fig. 1D–F). If these sites are functionally relevant, it would be expected that the clustered CpG in these islands are non-methylated. It is interesting to note that among 375 tested human genes all housekeeping and widely expressed genes were found to have a CpG island covering the transcription start, whereas only 40% of genes with a tissue-restricted expression are associated with CpG islands (26). Preliminary data suggest that *Mlo* is constitutively expressed at least in leaves and roots and a housekeeping function is known for the ribophorin I gene, encoding one of the three subunits of the oligosaccharyltransferase holo-protein (Panstruga *et al*., unpublished data; 18). In contrast, the *Drosophila g1* gene is known to be transiently expressed in the mesoderm of the embryo (19). Thus, the observed putative CpG islands adjacent to genes in the 60 kb BAC does not contradict their distribution among vertebrate genes.

The order of functional elements in *BAGY-1* allows us to assign the element to the Ty3/*gypsy* group of retroelements (21–25); it represents the first element of this class described in *H.vulgare*. We have not tested copy numbers of *BAGY-1* in the barley genome, but generally members of the Ty3/*gypsy* class are found

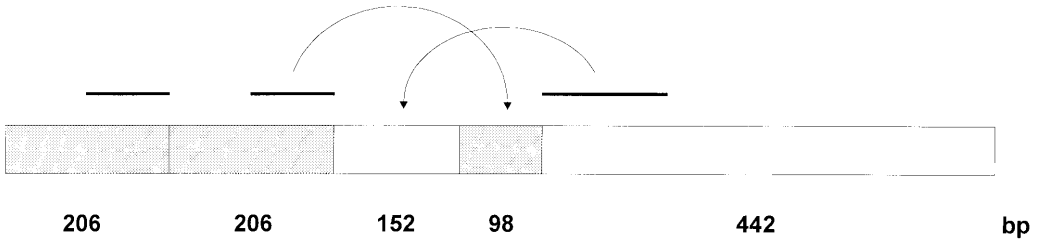

Figure 6. Structure of a complex direct repeat cluster. Two sets of direct repeats nested within each other are depicted. One set is represented by two copies of 206 bp and a 3′ subfragment (98 bp) of these (grey boxes). The 98 bp segment separates the second set (white boxes) represented by copies of 442 bp and a 152 bp subfragment. The 442 bp fragment is highly homologous (98% identity) to an upstream sequence of the barley aleurain gene (Table 2).

at high copy numbers in contrast with low copy numbers detected for representatives of the Ty1/*copia* type (22–23). The direct repeat of 5 bp flanking the insertion site of *BAGY-1* is imperfect by 1 bp which has also been observed for the retrotransposon *victim* of maize (8). The predicted *BAGY-1* polyprotein is not functional due to frame shifts and stop codons in the DNA sequence (Fig. 5). Likewise, one frame shift in the nt sequence of barley *BARE-1* had to be assumed to derive the polyprotein sequence which is interrupted by four stop codons (20). An extreme case of 21 stop codons in the polyprotein reading frame of a plant retroelement has been reported for the *del* retroelement of *Lilium henryi* (23). It is therefore very likely that each of these defective copies of retroelements, including the copy of *BAGY-1*, are inactive. Hence, the retrotransposition of *BAGY-1*, if it occurs, must be mediated by *trans*-acting factors.

ACKNOWLEDGEMENTS

We thank David Baker, Patrick Bovill and Leony Chamberlain for their technical assistance in sequencing. This work was funded by a Research Grant from the Gatsby foundation to P.S.-L.

REFERENCES

1 Le Guen, L., Thomas, M. and Kreis, M. (1994) *Mol. Gen. Genet.*, **245**, 390–396.
2 Quigley, F., Dao, P., Cottet, A. and Mache, R. (1996) *Nucleic Acids Res.*, **24**, 4313–4318.
3 Tremousaygue, D., Bardet, C., Dabos, P., Regad, F., Pelese, F., Nazer, R., Gander, E. and Lescure, B. (1997) *Genome Res.*, **7**, 198–209.
4 Martin, W., Gierl, A. and Saedler, H. (1989) *Nature*, **339**, 46–48.
5 Bennett, M. D. and Smith, J. B. (1991) *Philos. Trans. R. Soc. Lond., Ser. B.*, **334**, 309–345.
6 Moore, G., Devos, K. M., Wang, Z. and Gale, M. D. (1995) *Curr. Biol.*, **5**, 737–739.
7 Devos, K. M. and Gale, M. D. (1997) *Plant Mol. Biol.*, in press.
8 SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. and Bennetzen, J. L. (1996) *Science*, **274**, 765–768.
9 Moore, G., Gale, M., Kurata, N. and Flavell, R. (1993) *Bio/Technology*, **11**, 584–589.
10 Carels, N., Barakat, A. and Bernardi, G. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 11057–11060.
11 Barakat, A., Carels, N. and Bernardi, G. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 6857–6861.
12 Büschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., van Daelen, R., van der Lee, T., Diergaarde, P., Groenendijk, J., Töpsch, S., Vos, P., Salamini, F. and Schulze-Lefert, P. (1997) *Cell*, **88**, 695–705.
13 Simons, G., van der Lee, T., Diergaarde, P., van Daelen, R., Groenendijk, J., Frijters, A., Büschges, R., Hollricher, K., Töpsch, S., Schulze-Lefert, P., Salamini, F., Zabeau, F. and Vos, P. (1997) *Genomics*, in press.
14 Gardiner-Garden, M. and Frommer, M. (1992) *J. Mol. Evol.*, **34**, 231–245.
15 Crimaudo, C., Hortsch, M., Gausepohl, H. and Meyer, D. I. (1987) *EMBO J.*, **6**, 75–82.
16 Harnik-Ort, V., Prakash, K., Marcantonio, E., Colman, D. R., Rosenfeld, M. G., Adesnik, M., Sabatini, D. D. and Kreibich, G. (1987) *J. Cell Biol.*, **104**, 855–863.
17 Silberstein, S., Collins, P. G., Kelleher, D. J., Rapiejko, P. J. and Gilmore, R. (1995) *J. Cell Biol.*, **128**, 525–536.
18 Kelleher, D. J., Kreibich, G. and Gilmore, R. (1992) *Cell*, **69**, 55–65.
19 Bouchard, M. L. and Côté, S. (1993) *Gene*, **125**, 205–209.
20 Manninen, I. and Schulman, A. H. (1993) *Plant Mol. Biol.*, **22**, 829–846.
21 Grandbastien, M.-A. (1992) *Trends Genet.*, **8**, 103–108.
22 Bennetzen, J. L. (1996) *Trends Microbiol.*, **4**, 347–353.
23 Smyth, D. R., Kalitsis, P., Joseph, J. L. and Sentry, J. W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5015–5019.
24 Purugganan, M. and Wessler, S. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11674–11678.
25 Knoop, V., Unseld, M., Marienfeld, J., Brandt, P., Sünkel, S., Ullrich, H. and Brennicke, A. (1996) *Genetics*, **142**, 579–585.
26 Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) *Genomics*, **3**, 1095–1107.
27 Cross, S. H. and Bird, A. P. (1995) *Curr. Opin. Genet. Dev.*, **5**, 309–314.
28 Fickett, J. W. (1996) *Trends Genet.*, **12**, 316–320.
29 Miklos, G. L. G. and Rubin, G. M. (1996) *Cell*, **86**, 521–529.
30 Montero, L. M., Salinas, J., Matassi, G. and Bernardi, G. (1990) *Nucleic Acids Res.*, **18**, 1859–1867.
31 Bernardi, G. (1995) *Annu. Rev. Genetics*, **29**, 445–476.
32 Antequera, B. and Bird, A. (1988) *EMBO J.*, **7**, 2295–2299.
33 Langdale, J. A., Taylor, W. C. and Nelson, T. (1991) *Mol. Gen. Genet.*, **225**, 49–55.
34 Sørenson, M. B., Müller, M., Skerritt, J. and Simpson, D. (1996) *Mol. Gen. Genet.*, **250**, 750–760.
35 Swartz, M. N., Trautner, T. A. and Kornberg, A. (1962) *J. Biol. Chem.*, **237**, 1961–1967.