Genome-wide analysis of the *Emigrant* family of MITEs: amplification dynamics and evolution of genes in *Arabidopsis thaliana*

Néstor Santiago[1], Cristina Herráiz[2], J. Ramón Goñi[2], Xavier Messeguer[2], Josep M. Casacuberta[1]*
Dep. Genètica Molecular, IBMB-CSIC, Barcelona[1], Dep. Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya[2].

*Correspondence address: Dep. Genètica Molecular.
IBMB-CSIC. Jordi Girona18
08034 Barcelona, Spain
Tel. 34 93 400 61 42; Fax. 34 93 204 59 04
Email: jcsgmp@cid.csic.es

Running title: *Emigrant* amplification dynamics and gene evolution

Key words: Arabidopsis, evolution, MITE, Emigrant, master element.

ABSTRACT

MITEs are structurally similar to defective class II elements but their high copy number and the size and sequence conservation of most MITE families suggest that they can be amplified by a replicative mechanism. Here we present a genome-wide analysis of the *Emigrant* family of MITEs from *Arabidopsis thaliana*. In order to be able to detect divergent ancient copies, and low copy number subfamilies with a different internal sequence we have developed a computer program to look for *Emigrant* elements based solely on the TIR sequence. Our results show that different bursts of amplification of one or very few active, or master, elements have occurred at different times during *Arabidopsis* evolution, with an insertion dynamics similar to that of some SINEs. The analysis of the insertion sites of the *Emigrant* elements shows that, although *Emigrant* elements tend to integrate far from ORFs, the elements inserted within or close to genes are preferentially maintained during evolution.

INTRODUCTION

Transposable elements (TEs) are important components of eukaryote genomes, accounting for a high fraction of them. Most TEs can be grouped in two major classes according to their mode of transposition: class I elements transpose by a replicative mechanism invoking an RNA intermediate, while class II elements usually transpose by a non-replicative "cut-and-paste" mechanism. While most genomes contain elements of both classes, class I elements are in general much more abundant than class II elements, probably as a consequence of their replicative mode of transposition. This is the case for the human genome where the most abundant TEs are the L1 LINE (Long Interspersed Nuclear Element) and the Alu family of SINEs (Short Interspersed Nuclear Elements) (International Human Genome Sequencing Consortium, 2001), both belonging to the class I group of TEs. Plant genomes also contain retrotransposons in a high copy number, although in this case retroviral-like LTR-retrotransposons are the most abundant (Kumar and Bennetzen, 1999). In addition, plant genomes contain Miniature Inverted-repeat Transposable Elements (MITEs) in a very high copy number (Wessler et al.1995, Le et al. 2000, Turcotte et al. 2001).

MITEs are a particular class of TE first described in plants but later found to be present in other eukaryote genomes (Bureau and Wessler, 1992, Oosumi et al. 1996, Tu, 1997). They are structurally similar to defective classII elements but their high copy number and the size and sequence conservation of most MITE families suggest that they can be amplified by a replicative mechanism. It has been recently proposed that some MITEs could be a particular type of defective classII element related to the pogo subclass of Mariner transposons, or to bacterial IS elements (Feschotte and Mouchès. 2000a, Le et al. 2000, Zhang et al. 2001). Nevertheless, while it has been proposed that some MITE families could still be active in plants (Casacuberta et al. 1998, Zhang et al. 2000, Zhang et al. 2001), the characterization of a mobile MITE copy allowing the analysis of its transposition mechanisms is still lacking. In this context, the analysis of the evolution of MITE families of elements within their host genomes is probably the best approach to analyze the lifestyle of these elements and the impact of their mobility on host genomes. Here we present a genome-wide analysis of the *Emigrant* family of MITEs in *Arabidopsis thaliana*. In order to be able to detect elements with a divergent internal sequence, representing either ancient *Emigrant* elements, or previously undescribed low copy number *Emigrant* subfamilies, we have developed a computer program to detect putative MITEs in a genomic sequence based solely on their Terminal Inverted Repeat (TIR) sequences. This approach has allowed us to perform, for the first time, an evolutionary analysis of a family of MITEs within a particular genome. Our results show that different *Arabidopsis* contains different *Emigrant* subfamilies of elements that have probably been generated by the amplification of a small number of founder elements. Our results also show that, although *Emigrant* elements target very rich AT regions for insertion, elements closely linked to genes are more frequently maintained during evolution.

RESULTS

Different groups of the *Emigrant* element can be found in the genome of *Arabidopsis thaliana*

We have designed a new computer program named TRANSPO (accessible at www.lsi.upc.es/~alggen) that looks at a given sequence for the presence of a particular inverted repeat motif within a given range of distances. The independence of the search from a conserved internal sequence allows the localization of previously undetected low copy number subfamilies of a particular MITE that share the TIR sequences but differ in their internal sequence, as well as ancient MITE copies that have lost most of the sequence homology by insertion, deletions, or point mutations, within the sequence between the TIRs.

We have searched the complete available *Arabidopsis* sequence, which covers 115.4 of the 125-megabase genome, and does not include telomeres, centromeres or the rDNA repeated regions, for *Emigrant* elements by looking for *Emigrant* TIRs (tolerating up to 25% divergence) separated by more than 200 nt and less than 700 nt. We have localized 151 sequences that could represent *Emigrant* elements. In addition to the presence of *Emigrant*-like TIRs, these sequences are very A/T rich, do not have coding capacity and are flanked by the dinucleotide TA. Although all these characteristics are reminiscent of MITE-related sequences, most of these sequences are not annotated as *Emigrant* elements, MITEs or possible TEs in the databases.

The high variability of the internal sequence does not allow the correct alignement of the 151 sequences and their analysis by phylogenetic methods. We have thus used the program GROUP (also available at www.lsi.upc.es/~alggen) to put most of the 151 sequences into three main groups, that we have named *Emi*A (41 sequences), *Emi*B (26 sequences) and *Emi*C (37 sequences), based on pair-wise identity comparisons. 47 sequences were too divergent to be included in any of the defined groups and have been named as *Emi*0 elements.

All the previously described *Emigrant* elements (Casacuberta el al. 1998) belong to the *Emi*A group. We have previously demonstrated that *Emi*A elements were mobile in the recent past, as some of them were found to be polymorphic among *Arabidopsis* ecotypes (Casacuberta et al. 1998). In order to obtain data on the possible mobility of the other groups of elements we searched the *Arabidopsis* genome for related empty site (RESites), representing genome duplication events occurring prior to the transposition of these newly described elements. The presence of (RESites) within a genome has been successfully used as an indication of mobility when analyzing possible TEs within a single genome (Le et al. 2000, Tu, 2001). We found more than twenty well conserved RESites corresponding to the different groups of *Emigrant* elements, although we found more RESites corresponding to the *Emi*A, *Emi*B and *Emi*C classes than the *Emi*0. Figure 1 shows an example of such RESites. These data, and the presence in each case of a TA duplication accompanying the insertion of the element, strongly suggests that the different elements here described are indeed mobile elements related to the *Emigrant* family of MITEs.

Distribution of *Emigrant* elements in Arabidopsis chromosomes

Figure 2 shows the distribution of *Emigrant* elements in *Arabidopsis* chromosomes. The 5 chromosomes of Arabidopsis contain *Emigrant* elements although the density of

insertions varies slightly among them. Chromosome 4 has the highest concentration of *Emigrant* elements (1.8 Emi/Mb) while chromosome 5 the lowest (1 Emi/Mb), and chromosomes 1, 2 and 3 intermediate *Emigrant* concentrations. Although the concentration of the different *Emigrant* groups varies slightly between the 5 chromosomes, they all contain representatives of each Emigrant group. The concentration of *Emi*A elements in chromosome 4 is 6 times higher than in chromosome 5, and 2x the concentration of *Emi*C elements is present in chromosome 2 and 4 compared to chromosome 3, while there are more *Emi*0 elements (3x) in chromosome 4 than in chromosome 1.
Within each chromosome there is a slight tendency for *Emigrant* elements to be found close to peri-centromeric regions, but these concentrations seem less pronounced than that found for the overall TE population in Arabidopsis (The Arabidopsis Genome Initiative 2000).

Analysis of *Emigrant* insertion sites.

Most of the previously described *Emigrant* elements are flanked by a TA dinucleotide probably duplicated upon insertion (Casacuberta et al. 1998). Here we show that this is also the case for most of the *Emigrant* elements present in the genome of Arabidopsis, regardless of the subfamily they belong to. The analysis of 60 nt around the insertion site of the 151 *Emigrant* insertions shows that, in addition to insertion within TA sequences, *Emigrant* elements, like other MITEs and MLEs (Le et al. 2000), target regions of very high AT content. The average AT content of the Arabidopsis chromosomes ranges from 64.5% to 66.6%, rising to 67.6% in non-coding regions (The Arabidopsis Genome Initiative 2000), while sequences flanking *Emigrant* elements have 74.3% AT. This preference for AT-rich regions, their strict target specificity for the TA dinucleotide, and the fact that TA repetitions form the most frequent microsatellite in *Arabidopsis* (Casacuberta et al. 2000) indicates a preference of *Emigrant* to integrate within microsatellites. It has been recently reported that the rice poly(AT)n microsatellites are frequently associated with the *Micropon* family of MITEs (Temnykh et al, 2001). Nevertheless, although 25% of the *Emigrant* elements here analyzed lie in a TATA sequence, only 3% are found in a sequence containing a repetition of more than 4 TAs, suggesting that *Emigrant* elements do not target microsatellites for integration.

Analysis of the sequence and size variability of the different *Emigrant* subfamilies.

The relatively high sequence identity within each of the *Emi*A, *Emi*B and *Emi*C groups of elements has allowed us to perform conventional phylogenetic analysis. Sequences belonging to each group were aligned using the ClustalW program, and the alignments were used to obtain neighbor-joining trees. Figure 3 presents the trees obtained. Different monophyletic groups supported by high bootstrap values can be defined within each tree. Whitin each *Emigrant* group most of the sequences can be sub-divided in three different subfamilies (A1, A2, and A3, B1, B2, and B3, C1, C2, and C3). By performing new alignments with the sequences belonging to each subfamily, we have deduced a consensus sequence for each of them, and compared these consensus sequences in order to obtain information about the phylogenetic relationships among the different *Emigrant* subfamilies. A neighbor-joining tree, obtained comparing the consensus sequences of each subfamily, is also shown in Figure 3. The three *Emi*A subfamilies, the three *Emi*B subfamilies and the three *Emi*C subfamilies seem

phylogenetically related, as the three different groups cluster together with high bootstrap values.
The alignments of the sequences belonging to each subfamily were also used to calculate the nucleotide diversity, π (Nei, 1957), and the size variability for each *Emigrant* subfamily. These results show that each *Emigrant* subfamily displays a different degree of variability (Table 1). While some subfamilies such as the *Emi*A2 are highly homogeneous both in size and sequence, other subfamilies like the *Emi*B3 subfamily are highly variable. Each *Emigrant* group contains subfamilies of different variability. Within the *Emi*A group the A2 subfamily is more homogeneous than the A1 subfamily; within the *Emi*B group the most homogeneous is the B2 subfamily and the most variable is the B3, the B1 displaying an intermediate degree of variability. Within the *Emi*C group the C1 subfamily is the most homogeneous in sequence although is relatively variable in size, and C3 seems to be the most variable group.

Position of *Emigrant* elements relative to open reading frames

Although MITEs seem to target very high AT-rich regions, they have often been found close to transcribed sequences (Wessler et al 1995, Yang et al. 2001). We analyzed the regions flanking the 151 *Emigrant* insertions and calculated the distance from the closest ORF. 10% of the elements lie within an ORF (7% in introns and 3% in exons), 24% lie at less than 500 nt from an ORF, 23% at more than 500 nt and less than 1000 nt from an ORF. 29% are located at more than 1000 nt from any ORF and 13% are inserted within a repetitive region. Nevertheless, the position of the *Emigrant* elements with respect to the ORFs greatly varies among the different subfamilies analyzed (see Table 2). While 55.5% of the *Emi*B3 elements are found at less than 500 nt from an ORF, and 42% of *Emi*0 are located within or close to an ORF, the vast majority of the *Emi*A2 (85%) are located at more than 500 nt from any ORF.
Among the 53 *Emigrant* elements located at less than 500 nt from the closest ORF, 46.5% are located downstream, 27.5% are located upstream, and 26% are located within an ORF region. These elements can affect promoter activity, splicing, transcriptional termination or RNA stability, as well as the coding capacity of the ORF. We have thus analyzed these insertions in some detail, and Figure 4 shows examples of such close-gene insertions. Figure 4A shows an *Emi*0 element, found within the transcribed downstream region of the *Det1* gene, as an example of an element lying downstream of an ORF. The availability of the genomic and the cDNA sequence for the *Det1* gene has allowed us to determine that the transcription of the *Det1* gene stops within the *Emigrant* element, probably using polyadenylation sequences provided by *Emi*116. Figure 4B shows an example of an *Emi*0 element located within a predicted ORF coding for a GATA-like transcription factor. The insertion of the element has provided a new putative ATG and 48 new aminoacids within the C-terminal region of the protein. We also found 5 Emigrant elements lying at less than 500 nt from two different ORFs. The insertion of those elements could potentially affect to the expression of both the upstream and the downstream gene. Alternatively, the insertion of an Emigrant element in these extremely short intergenic regions could help to avoid transcriptional interference between both genes. Related to this, it is interesting to note that it has been proposed that some MITEs could act as matrix attachment regions isolating their neighboring genes (Tikhonov et al, 2000). This possible effect of MITE insertion could be particularly useful in Arabidopsis, which is a very compact genome and genes are sometimes found extremely close to one another.

DISCUSSION

A new approach to study MITE evolution

Most MITE families described to date are characterized by a high degree of sequence and size conservation (Bureau and Wessler, 1992, Casacuberta et al, 1998, Feschotte and Mouchès, 2000b, Oosumi et al. 1996, Tu, 1997, Yang et al, 2001). However, after insertion MITEs are subjected to random mutation and the sequence and size homogeneity of a particular MITE family will decrease with time. Ancient divergent copies, and low copy number families of MITEs, are difficult to detect by sequence similarity-based search methods, and have probably been missed in searches performed to date. In order to get access to *Emigrant* divergent elements and analyze the evolution of this particular family of MITEs we developed a computer program based solely on the presence of relatively conserved TIR sequences. The program TRANSPO has allowed us to detect all the sequences present within the genome of *Arabidopsis* that contain TIRs 75% identical to the previously defined Emigrant TIR (Casacuberta et al. 1998), separated by more than 200 nt and less than 700 nt. The presence of target site duplications and RESites for the different *Emigrant* groups here described has allowed us to confirm the mobile nature of these sequences.

Different amplification bursts of *Emigrant* elements have occurred during *Arabidopsis* evolution.

The 47 *Emi*0 sequences are too divergent to be included in any of the 9 *Emigrant* subfamilies here defined. The high divergence of these elements suggests that they represent old *Emigrant* insertions that have accumulated a high number of mutations. The phylogenetic analysis of the other *Emigrant* elements shows that they belong to different subfamilies with different degrees of variability. While the 20 *Emi*A2 elements are highly homogeneous, the *Emi*B3 subfamily is highly variable. This suggests that different amplification bursts have occurred at different times during the evolution of *Arabidopsis*, giving rise to these different subfamilies, the more variable a subfamily is the more ancient the amplification burst that has generated it should be. The start-type topology of the *Emigrant* subfamilies in the different trees suggests that each subfamily has been generated from the amplification of a single *Emigrant* active element. These insertion dynamics fits the predictions of the master gene model that has been developed to explain the evolutionary dynamics of some SINEs (Denninger and Batzer, 1995). In this model only one, or very few, active loci are capable of amplification giving rise to non-propagating offspring copies. MITEs are assumed to be defective elements transposed and amplified by a trans acting transposase that recognizes their TIRs. Most *Emigrant* elements belonging to a recently amplified subfamily such as A2 show high conservation of both the internal and the TIR sequences, which suggests that the chromosomal region in which each element lies influences the ability of a particular element to be amplified and act as a master element. We have not found any particular distribution of *Emigrant* copies within the 5 different chromosomes (see Fig.2), suggesting that *Emigrant* elements do not target particular chromosomal regions for integration. This lack of insertion preference with respect to chromosomal locations, together with the influence that these regions may have in determining the ability of a particular copy to act as a founder amplifying element, could explain the lack of activity of most newly inserted copies.

The different *Emigrant* subfamilies cluster into different groups suggesting that they have been generated by different but related founder elements. This could be explained by a low frequency acquisition of the ability to be amplified and act as founder element by the newly inserted copies. Alternatively, as the conservation of the internal *Emigrant* sequence does not seem to be essential for amplification, a single master element could maintain its potential activity during a long period of time in spite of the accumulation of mutations between the TIRs, giving rise in each amplification burst to a different but related subfamily.

Elements close to genes have been preferentially maintained during *Arabidopsis* evolution

Although MITEs seem to target very high AT-rich regions, they have often been found close to transcribed sequences (Wessler et al 1995, Yang et al. 2001). Nevertheless, a recent survey failed to detect transposon insertions in *Arabidopsis thaliana* coding regions, suggesting a purifying selection against deleterious mutation (Le et al. 2000). The presence of mobile elements at particular locations within a genome is the result of their transpositional activity, but also of the selection of the best fit genomes. Thus, elements transposing randomly within a genome can be found at particular locations as the result of a positive selection of their insertion within those sites or the negative selection of insertion in other locations. The effect of target site specificity should be more easily detected for recently inserted elements, while the effect of selection will be more apparent for ancient insertions. The comparison of the distribution of ancient versus recent elements should reveal the effect of selection and thus the impact of transposon insertions. So, we compared the relative distribution of the different subfamilies of *Emigrant* elements, which represent amplification burst occurred at different times of the evolution of *Arabidopsis*, with respect to predicted genes in order to determine their insertion specificity as well as the effect of selection and the impact of *Emigrant* insertions.

*Emi*A2 is the most homogeneous subfamily here described, both in sequence and size, and probably represents the most recent burst of amplification of *Emigrant* elements. 85% of the 20 *Emi*A2 elements lie at more than 500 nt from the closest ORF (see Table2). The genome of *Arabidopsis* is extremely compact and the intergenic regions are very short. The mean size of *Arabidopsis* genes is 2Kb and there is one gene every 5 Kb, which implies that the mean distance between two genes is only 3 Kb (The Arabidopsis Genome Initiative 2000). Thus genic regions occupy 40% of the genome space, and the regions closely linked to the genes, that most probably contain gene regulatory regions (arbitrarily taken here as 500 nt) occupy 20% of the genome space, which means that 60% of the genome is occupied by genes and their potentially regulatory regions. The regions not linked to genes occupy only 40 % of the genome space (20% the region arbitrarily defined here as between 500 and 1000 nt, and 20 % the region arbitrarily defined here as >1000 nt). The distribution of *Emi*A2 elements is thus far from random, with *Emigrant* elements inserting preferentially far from ORFs. The strict specificity of *Emigrant* and other MITEs for the TA dinucleotide as insertion site, as well as the preference for very high AT containing regions (74.3% AT in the case of *Emigrant*) probably helps these elements to avoid genes even in extremely compact genomes such as *Arabidopsis*.

This preference for regions far from genes is less pronounced for other *Emigrant* subfamilies. More than 50% of the *Emi*B3 elements, and 43 % of the *Emi*0 group of elements, are less than 500 nt from the closest ORF. Interestingly, the *Emi*0 group

contains the most divergent *Emigrant* elements, and the *Emi*B3 subfamily is one of the most variable subfamilies, suggesting that both the *Emi*0 group and the *Emi*B3 subfamily represent the most ancient insertions and have been subjected to selection for a relatively long period of time. The other *Emigrant* subfamilies show different distribution patterns with respect to ORFs but, although the low number of elements makes it difficult to draw conclusions in some cases, the more variable a subfamily the more closely it is associated to genes.

These results suggest that while *Emigrant* elements preferentially insert far from ORFs, the elements closely linked to genes are more frequently maintained during evolution. This is reminiscent of what has been shown for the Alu family of SINEs in the human genome. Alus tend to insert in AT rich regions, and recently transposed Alu subfamilies are found in poor-gene regions, while ancient Alu subfamilies are found preferentially in GC-rich regions closely associated to genes (International Human Genome Sequencing Consortium, 2001). It has been proposed that a positive selection in favor of the minority of Alus in GC-rich DNA, rather than against the majority that lie in AT-rich regions, could explain the difference in distribution between old and new Alu subfamilies (International Human Genome Sequencing Consortium, 2001). *Emigrant* and other MITEs resemble SINEs in their short size and their high copy number, and here we have shown that their amplification dynamics is also very similar. It is thus tempting to hypothesize that, as is the case for Alu elements within the human genome, there has been a positive selection for *Emigrant* elements lying within or close to genes during *Arabidopsis* evolution.

A role for *Emigrant* elements in the evolution of *Arabidopsis* genes

Over the last 10 years a growing body of evidence has pointed towards a modular nature for the regulation of gene expression. Promoters, and probably terminators, are constituted by a complex array of regulatory elements. Most of these elements are found in many different promoters or terminators, although each promoter/terminator contains a particular combination of them. With the completion of genome sequencing projects it has become more and more clear that coding regions of eukaryote genes are also often composed of domains or modules that have been reshuffled during evolution. There are many different mechanisms that can account for the amplification and distribution of particularly successful coding or regulatory modules, but short replicative elements such as SINEs and MITEs, would be particularly suitable candidates for such a function. SINEs are frequently found within or close to genes in Arabidopsis (Lenoir et al, 2001) and other organisms (Makalowski, 1995), and it has been recently found that some of them can play an important biological role as coding or transcriptional regulatory regions (Shimamura et al, 1998, Ferrigno et al, 2001). Moreover, it has been proposed that B2 SINEs may have the potential to distribute a functional pol II promoter throughout the genome (Ferrigno et al, 2001). Here we showthat a high number of *Emigrant* elements within potential promoters, terminators, introns, and coding sequences which may affect gene coding capacity or regulation, have been conserved during evolution. Although molecular experiments to unambiguously determine the impact of these insertions have yet to be performed, our results suggest that the insertion of *Emigrant* elements has played an important role in the evolution of *Arabidopsis* genes. MITEs, as has been proposed for SINEs, could have been recruited by genomes as an evolutionary mechanism to generate novel coding or regulatory sequences. The fact that MITEs can probably be excised (Yang et al 2001), makes them even more suitable for such a function.

METHODS

Construction of the TRANSPO program

TRANSPO implements the Fast Bit-Vector algorithm (Myers, 1998) that finds all locations at which a query (the *Emigrant* TIR sequence in this case) approximately matches a sequence (the sequence of 5 *Arabidopsis* chromosomes). The expected time is linear on the length of the sequence. Although this algorithm is based on dynamic programming with quadratic cost on the lengths of the sequence and the query, the linearity can be accomplished if the length of the query is shorter than the length of the computer word. In this case the query is implemented as a vector of bits and at each step all the values of a column (or a row) can be computed.
MATDIS computes the matrix of similarities between all the sequences. In this case we have taken as measure of similarity between each pair of sequences the quotient obtained when the number of matches of the best global pair wise alignment is divided by the length of the shortest sequence. The SPAT clustering program groups the sequences into a hierarchical classification, i.e. a nested sequence of partition (Gordon, 1999). The similarity matrix is viewed as an undirected, weighted graph where the nodes are the sequences and the weight of each edge is the similarity measure between the pair of nodes connected by this edge. Then the maximum spanning binary tree is found. Note that this tree ‘‘spans’’ the graph in the following sense: it connects all the nodes in such away that the sum of the weights of the edges is maximized. Finally this tree is partitioned by removing the edges with minimum weight (Zahn, 1971, Delattre and Hansen, 1980). It could be that these edges are terminal edges, then its extraction only separates one node from the tree, but in our case, the extraction of edges defined the groups A, B, C (group 0 is composed of those nodes connected by terminal edges). We consider that a group of sequences is cohesive when a consecutive extraction of a significant number of edges does not change the composition of the group.

Emigrant element mining

The TRANSPO program was used to look at the entire available *Arabidopsis* genome sequence (www.arabidopsis.org) for inverted repeated sequences 75% identical to the first 20 nt of the previously defined *Emigrant* TIR (CAGTAAAACCTCTATAAATT) located within a range of 200-700 nt. Overlapping elements generated from subterminal inverted repeat sequences were eliminated. The sequences obtained were grouped using the GROUP program and a graphical distribution of the different elements in *Arabidopsis* chromosomes was obtained using the … program.
To obtain information about the ORF located close to the *Emigrant* elements, the 30nt flanking each *Emigrant* element upstream and downstream were used as probes in sequence similarity searches (BLAST 2.0; Altschul et al, 1990; http://www.ncbi.nlm.nih.gov/blast/). A table containing the BAC accession number and the nucleotide position of each *Emigrant* element, as well as the name and the distance of the elements to the closest upstream and downstream ORF can be obtained as additional information.
Sequence similarity searches (BLAST) with the sequences flanking *Emigrant* elements were also used to look for RESites.

Phylogenetic Analysis.

Sequences were aligned using the CLUSTAL W multiple-alignment program (version 1.5; Thompson, Higgins, and Gibson 1994) with some minor refinements. DNADIST in Felsestein's PHYLIP package (Felsenstein 1989) was used to generate a distance matrix based on the Jukes-Cantor algorithm (Jukes, and Cantor 1969). This was used to generate neighbor-joining trees (Saitou, and Nei 1987). Bootstrap analyses were performed using the Seqboot and Consense programs from Felsestein's PHYLIP package (Felsenstein 1989). Sequence variability, as measured by Nei's measure of nucleotide diversity, π (Nei, 1957) was calculated using the DnaSP program (Rozas, and Rozas 1997).

ACKNOWLEDGEMENTS

We would like to thank E. Casacuberta, M.L. Espinás, J. Martinez-García, and P. Puigdomènech for critical reading of the manuscript. This work was supported by a grant from the Ministerio de Ciencia y Tecnología to JMC (grant BIO2000-0953).

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol*. **215:** 403-410.

Casacuberta, E., Casacuberta, J.M., Puigdomènech, P., and Monfort, A. 1998. Presence of Miniature Inverted-repeat Transposable Elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. *Plant J.* **16:** 79-85

Casacuberta, E., Puigdomènech, P., Monfort, A. 2000. Distribution of microsatellites in relation to coding sequences within the Arabidopsis thaliana genome. *Plant Sci.* **157:** 97-104.

Deininger, P.L., and Batzer, M.A. 1995. SINE master genes and population biology. In *The impact of short interspersed elements (SINEs) on the host genome.* (ed. Maraia, R.J.) pp. 43-60. RG Landes Company, Springer, Austin, Texas.

Delattre, M., and Hansen, P. 1980. Bicriterion cluster analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2,277-291.

Felsenstein, J. 1989. PHYLIP-phylogeny inference package (version 3.56). *Cladistics* **5:** 164-166

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.P., White, R.J., and Aberdam, D. 2001. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat Genet.* **28:**77-81.

Feschotte, C.., and Mouches, C. 2000a. Evidence that a family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol.* **17:** 730-737

Feschotte, C., and Mouches, C. 2000b. Recent amplification of miniature inverted-repeat transposable elements in the vector mosquito Culex pipiens: characterization of the Mimo family. *Gene.* **250:** 109-116.

Gordon, A.D. 1999. Classification. Chapman&Hall/CRC

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-922

Kumar, A., and Bennetzen, J. 1999. Plant Retrotransposons. *Annu. Rev. Genet.* **33:** 479-532

Le, Q.H., Wright, S., and Bureau, T. 2000 Transposon diversity in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **97:** 7376-7381

Le, Q.H., Turcotte, K., and Bureau, T. 2001 Tc8, a Tourist-like Transposon in Caenorhabditis elegans. *Genetics.* **158:** 1081-1088.

Lenoir. A, Lavie, L., Prieto, J.L., Goubely, C., Cote, J.C., Pelissier, T., and Deragon, J.M. 2001. The evolutionary origin and genomic organization of SINEs in Arabidopsis thaliana. *Mol. Biol. Evol.* **18:**2315-2322.

Makalowski, W. 1995. SINEs as a genomic scrap yard: an essay on genomic evolution. In *The impact of short interspersed elements (SINEs) on the host genome* (ed. Maraia, R.J.) pp. 81-104. RG Landes Company, Springer, Austin, Texas.

Myers, G. (1998) A Fast Bit-Vector Algorithm for Approximate String Matching Based on Dynamic Progamming. *Proc. Ninth Combinatorial Pattern Matching Conference. Springer-Verlag LNCS series* **1448:** 1-13.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York

Oosumi, T., Garlick, B., and Belknap, W.R. 1996. Identification of putative nonautonomous transposable elements associated with several transposon families in Caenorhabditis elegans. *J. Mol. Evol.*.**43:** 11-8

Rozas, J., and Rozas, R.. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15:** 174-175.

Saitou, N., and Nei, N. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425

Shimamura, M., Nikaido, M., Ohshima, K., and Okada, N. 1998. A SINE that acquired a role in signal transduction during evolution. *Mol. Biol. Evol.* **15:** 923-925

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*. **11:** 1441-1452

Thompson, J.D., D.G. Higgins, and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignement through sequence weighting, population-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673-4680

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815

Tikhonov, A.P., Bennetzen, J.L., and Avramova, Z.V. 2000. Structural domains and matrix attachment regions along colinear chromosomal segments of maize and sorghum. *Plant Cell* **12:** 249-64.

Tu, Z. 2000. Molecular and evolutionary analysis of two divergent subfamilies of a novel miniature inverted repeat transposable element in the yellow fever mosquito, Aedes aegypti. *Mol. Biol. Evol.* **17:** 313-25.

Tu, Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the african malaria mosquito, Anopheles gambiae. *Proc. Natl. Acad. Sci. USA* **98:** 1699-1704

Turcotte, K., Srinivasan, S., and Bureau, T. 2001. Survey of transposable elements from rice genomic sequences. *Plant J.* **25:** 169-179

Wessler, S., Bureau, T., and White, S.E. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Devel.* **5:** 814-821

Yang, G., Dong, J., Chandrasekharan, M.B., and Hall, T.C. 2001. *Kiddo*, a new transposable element family closely associated with rice genes. *Mol. Genet. Genomics* **266:** 417-424

Zhang, Q., Arbuckle, J., and Wessler, S.R. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker* into genic regions in maize. *Proc. Natl. Acad. Sci. USA* **97:** 1160-1165

Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W.R., and Wessler, S.R. 2001. P instability factor: An active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA.* **98:** 12572-12577.

Zahn, C.T. (1971) Graph-theoretical methods for detecting anddescribing gestalt clusters. IEEE Transactions on Computers, C-20,68-86.

LEGENDS

Figure 1 RESites corresponding to elements of the different *Emigrant* groups

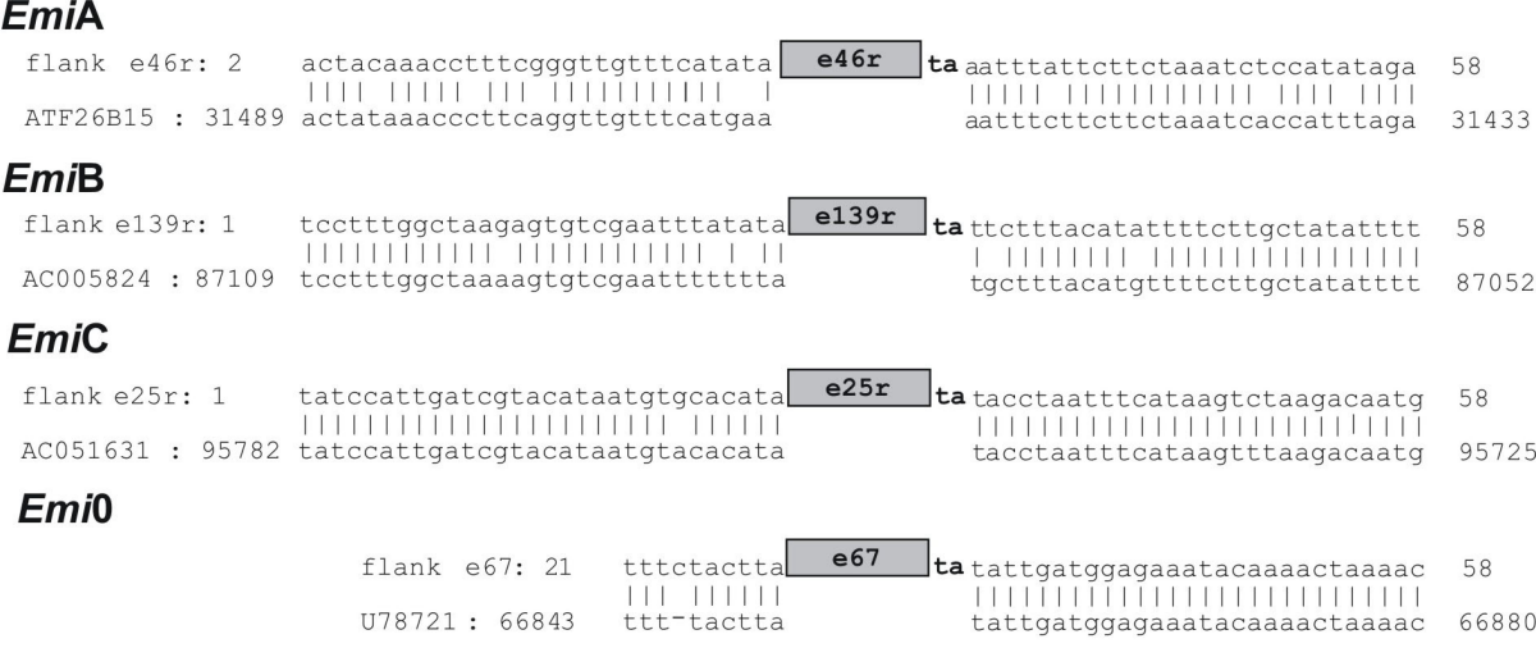
Figure 2 Distribution of Emigrant elements of the different groups in the 5 chromosomes of Arabidopsis. The number of *Emigrant* elements per Mb is indicated for each subfamily.

Figure 3 Phylogenetic analysis of the *Emigrant* elements. Neighbor-joining trees obtained comparing the *Emi*A **(A)**, *Emi*B **(B)**, and the *Emi*C **(C)** elements. The subfamilies defined within each *Emigrant* group are shown as A#, B# or C#. (**D)** Neighbor-joining tree obtained comparing the consensus sequences of the different *Emigrant* subfamilies. Bootstrap values above 60% supporting major clusters are shown. Distances are proportional to evolutionary divergence expressed in substitutions per hundred sites.

Figure 4. *Emigrant* elements inserted within ORFs. Schematic representation of the insertions of Emi116 (A) and Emi130 (B) within ORFs. Open boxes represent coding sequences, and filled boxes represent Emigrant elements. The name of the gene or the accession number of the ORF in which the *Emigrant* elements are inserted is indicated. The nucleotide and deduced aminoacidic sequences are shown under the scheme. Sequences corresponding to the *Emigrant* element are in bold and the TIR sequence is boxed. The poly A site is also indicated and the polyadenylation signal is underlined.

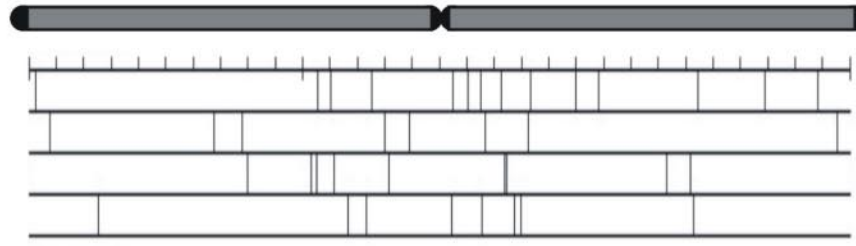
Table1 Nucleotide diversity, π (Nei, 1957), and size variability of the different *Emigrant* subfamilies. Size variability has been calculated as the standard deviation from the mean size. Nucleotide and size variability was not determined (n.d.) for *Emi*0 as these elements do not constitute a homogeneous group and can not be aligned.

Table 2. Position of the *Emigrant* elements of the different subfamilies with respect to the closest ORF. The number of sequences is shown in brackets. We did not determine (n.d.) the distance to the closest ORF for 7 *Emigrant* elements as we have not obtained the annotation of the corresponding genomic region. *The e79 element (belonging to the *Emi*A group and lying <500nt of an ORF), the e104 element (belonging to the *Emi*B group and lying at <500 nt of an ORF) and the e111 element (belonging to the *Emi*C group and lying within an ORF) have not been included in any of the *Emigrant* subfamilies but have been included in the global analysis of elements.


EmiA
flank e46r: 2 actacaaacctttcgggttgtttcatata e46r ta aatttattcttctaaatctccatataga 58
ATF26B15 : 31489 actataaaccttcaggttgtttcatgaa aatttcttcttctaaatcaccatttaga 31433
EmiB
flank e139r: 1 tcctttggctaagagtgtcgaatttatata e139r ta ttctttacatattttcttgctatatttt 58
AC005824 : 87109 tcctttggctaaaagtgtcgaattttttta tgctttacatgttttcttgctatatttt 87052
EmiC
flank e25r: 1 tatccattgatcgtacataatgtgcacata e25r ta tacctaatttcataagtctaagacaatg 58
AC051631 : 95782 tatccattgatcgtacataatgtacacata tacctaatttcataagtttaagacaatg 95725
Emi0
flank e67: 21 tttctactta e67 ta tattgatggagaaatacaaaactaaaac 58
U78721 : 66843 ttt-tactta tattgatggagaaatacaaaactaaaac 66880

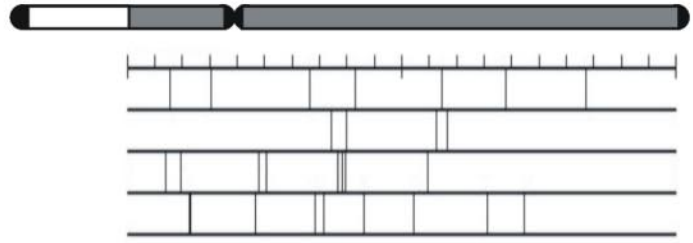
Chr. 1 29.1 Mb

A	14 (0.41 Emi/Mb)
B	8 (0.34 Emi/Mb)
C	9 (0.34 Emi/Mb)
0	8 (0.24 Emi/Mb)
T	39 (1.3 Emi/Mb)



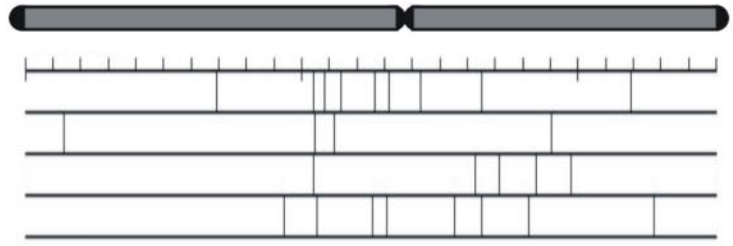
Chr. 2 19.6 Mb

A	7 (0.36 Emi/Mb)
B	4 (0.20 Emi/Mb)
C	9 (0.46 Emi/Mb)
0	7 (0.36 Emi/Mb)
T	27 (1.4 Emi/Mb)



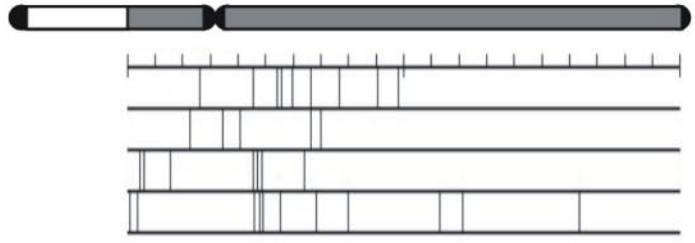
Chr. 3 23.2 Mb

A	7 (0.30 Emi/Mb)
B	6 (0.26 Emi/Mb)
C	5 (0.22 Emi/Mb)
0	9 (0.39 Emi/Mb)
T	27 (1.2 Emi/Mb)



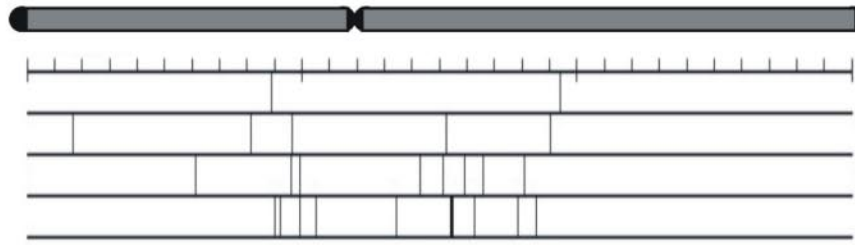
Chr. 4 17.5 Mb

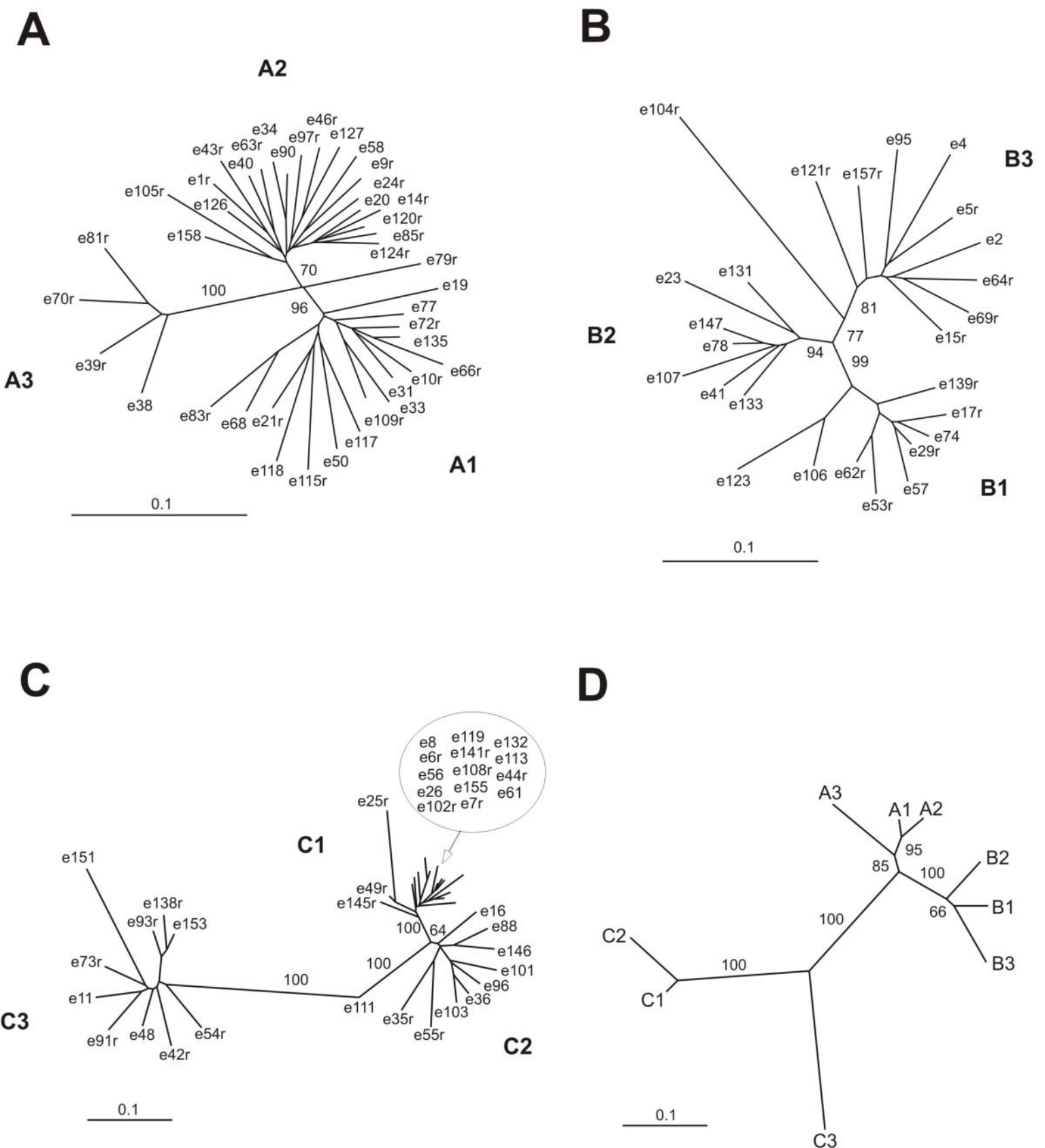
A	9 (0.51 Emi/Mb)
B	5 (0.28 Emi/Mb)
C	7 (0.40 Emi/Mb)
0	11 (0.63 Emi/Mb)
T	32 (1.8 Emi/Mb)



Chr. 5 26 Mb

A	2 (0.08 Emi/Mb)
B	5 (0.19 Emi/Mb)
C	8 (0.31 Emi/Mb)
0	11 (0.42 Emi/Mb)
T	26 (1 Emi/Mb)




A
A2
e46r
e34
e97r
e127
e43r
e63r
e90
e58
e40
e9r
e1r
e24r
e105r
e20
e14r
e126
e120r
e158
e85r
e124r
e81r
e79r
70
e19
e70r
100
96
e77
e72r
e135
e39r
e66r
A3
e10r
e31
e38
e33
e83r
e68
e21r
e109r
e117
e50
A1
e118
e115r
0.1
B
e104r
e95
e4
B3
e121r
e157r
e5r
e2
e64r
e23
e131
81
e69r
e147
B2
77
e15r
e78
94
99
e107
e139r
e41
e133
e17r
e74
e29r
e123
e106
e62r
e57
B1
e53r
0.1
C
e8
e119
e132
e6r
e141r
e113
e56
e108r
e44r
e26
e155
e61
e102r
e7r
e25r
C1
e151
e49r
e145r
e138r
e16
e93r
e153
100
64
e88
e146
e73r
100
100
e101
e11
e96
e36
e111
e35r
e103
C3
e91r
e48
e54r
e55r
C2
e42r
0.1
D
A3
A1
A2
95
B2
85
100
B1
66
100
C2
100
B3
C1
0.1
C3

A

tctgtcgtca atatccattt taggcgatga ctctttttgg tttctatcag tctagtagtc

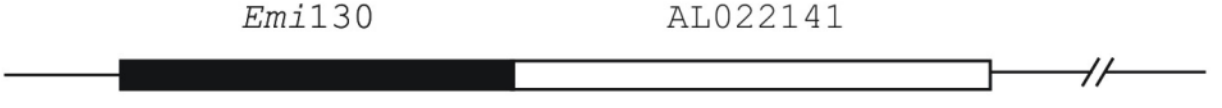
S V V N I H F R R

ttcctacatg ttgtaatctat tgtttgtaa caaaagtt**ca gtaaaacctct ataaattcc**

gagttgggcc gatgtaaaaat taacaaact tcgataaaat aataatataat aatcttttg

polyA

B



agacaaaaat ggagtgtgta gaagcatttc tcggtgattt ttccgtcgac gatcttctcg

M E C V E A F L G D F S V D D L L D

acctctctaa cgccgacact tctttagagt cgtcttcgtc acaaagaaaa gaagacgaac

L S N A D T S L E S S S S Q R K E D E Q

aagaacgtga gaaatttaag agcttttctg accaaagtac ccgtctttca ccgccagagg

E R E K F K S F S D Q S T R L S P P E D

	#	Nucleotide variability	Size variability
***Emi* A1**	18	0.108	5.6%
***Emi* A2**	20	0.095	2.7%
***Emi* A3**	4	0.081	5.2%
***Emi* B1**	9	0.102	3.9%
***Emi* B2**	7	0.077	3.8%
***Emi* B3**	6	0.115	8.2%
***Emi* C1**	15	0.068	7.1%
***Emi* C2**	9	0.112	4.2%
***Emi* C3**	9	0.131	5.9%
***Emi* 0**	47	n.d	n.d

	Distance to the closest ORF (nt)					
	0 (within)	**<500**	**500-1000**	**>1000**	**rep.**	**n.d.**
***Emi* A1 (16)**	12.5% (2)	19% (3)	19% (3)	44% (7)	6% (1)	
***Emi* A2 (20)**		15% (3)	20% (4)	50% (10)	15% (3)	
***Emi* A3 (4)**		25% (1)	50% (2)		25% (1)	
***Emi* B1 (9)**	12.5% (1)	12.5% (1)	12.5% (1)	38% (3)	25% (2)	1
***Emi* B2 (7)**		43% (3)	43% (3)		14% (1)	
***Emi* B3 (9)**		55.5% (5)	33% (3)	11% (1)		
***Emi* C1 (17)**	20% (3)	20% (3)	7% (1)	40% (6)	13% (2)	2
***Emi* C2 (9)**	11% (1)	22% (2)	44% (4)	22% (2)		
***Emi* C3 (10)**	10% (1)	20% (2)	30% (3)	30% (3)	10% (1)	
***Emi* 0 (47)**	14% (6)	28% (12)	21% (9)	18.5% (8)	18.5% (8)	4
Total (151)*	10% (15)	25.5% (37)	23% (33)	27.5% (40)	13% (19)	7