INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Transposons and the evolutionary relationships among modern rice species

Kime Turcotte

Department of Biology, McGill University, Montreal

September 2001

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science.

© Kime Turcotte 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-75349-2

Canada

Abstract

A high-resolution computer-based survey for transposable elements performed on 910 Kb of rice genomic DNA sequences revealed the presence of both class I and class II transposable elements. Elements from most major families of plant transposable elements were identified, and new groups were reported for these families. Miniature inverted-repeat transposable elements (MITEs) are clearly the predominant type of transposable element in the rice sequences examined. Phylogenetic analysis of the putative transposases of several transposable elements indicated that *Tourist*-like miniature inverted-repeat transposable elements (MITEs) are closely related to the bacterial insertion sequence 5 (IS5) family of transposable elements, while *Emigrant*-like and *Stowaway*-like MITEs are both related to members of the IS*630*/Tc1/*mariner* superfamily of elements. Finally, the nucleotide sequences of MITEs, *Ac*-like, *Mutator*-like elements (MULE), short interspersed nuclear elements (SINEs) and other unclassified elements, as well as their insertion polymorphism data have been used to reconstruct the relationships between rice species in the AA genome. The use of a combination of transposable element data sets generated the most reliable cladograms.

Resumé

Un sondage des éléments transposables présents dans 910 Kb de séquences d'ADN génomique de riz a révélé la présence d'éléments de la classe I et de la classe II. Des éléments appartenant à la plupart des familles de transposon des plantes ont été identifiés, et de nouveaux groupes sont rapportés pour ces familles. Les éléments du type MITE («miniature inverted-repeat transposable element») sont nettement prédominants dans les séquences de riz examinées. Des analyses phylogénétiques basées sur les transposases de plusieurs transposons révèlent que les MITEs du type *Tourist* sont apparentés à la famille de transposons bactériens IS*5* («insertion sequence 5»), alors que les MITEs du type *Emigrant* et *Stowaway* sont tous deux apparentés aux membres de la superfamille de transposons IS*630*/Tc1/*mariner*. Finalement, les séquences nucléotidiques d'éléments de type MITE, *Ac*, MULE («*Mutator*-like element»), SINE («short interspersed nuclear element») et d'autres éléments non classés, ainsi que leur polymorphismes d'insertion ont été utilisés pour reconstruire les relations phylogénétiques entre les espèces de riz ayant le génome AA. Les cladogrammes les plus fiables ont été obtenus en utilisant des combinaisons de données provenant de plusieurs transposons.

Table of contents

List of tables and figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Preface

The Faculty of Graduate Studies and Research requires that the following text be reproduced in full in order to inform the reader of the Faculty regulations.

> Candidates have the option of including, as part of the thesis, the text of one or more papers submitted, or to be submitted, for publication, or the clearly duplicated text of one or more published papers. These texts must be bound as an integral part of the thesis.
>
> If this option is chosen, connecting texts that provide logical bridges between the different papers are mandatory. The thesis must be written in such a way that it is more than a mere collection of manuscript, in other words, results of a series of papers must be integrated.
>
> The thesis must still conform to all other requirements of the "Guidelines for thesis preparation". The thesis must include: a table of contents, an abstract in English and in French, an introduction which clearly states the rationale and objectives of the study, a comprehensive review of the literature, a final conclusion and summary, and a thorough bibliography or reference list.
>
> Additional material must be provided where appropriate (e.g. in appendices) and in sufficient detail to allow a clear and precise judgement to be made of the importance and originality of the research reported in the thesis.
>
> In the case of manuscript co-authored by the candidate and others, the candidate is required to make an explicit statement in the thesis as to who contributed to such work and to what extent. Supervisors must attest of the accuracy of such statements at the doctoral oral defense. Since the task of the examiners is made more difficult in these cases, it is in the candidate's interest to make perfectly clear the responsibilities of all the authors of the co-authored papers.

This thesis consists of four chapters that are prepared as individual manuscripts for publication in peer-reviewed journals. Chapter 1 consists of a manuscript published in The Plant Journal in January 2001. Chapter 2 consists of a manuscript published in Genetics in July 2001. Chapter 3 consists of a manuscript accepted for publication in Genome. Chapter 4 consists of a manuscript to be submitted soon.

Contributions of authors

To chapter 1, 3 and 4, my co-supervisors Dr. Thomas E. Bureau (McGill University) and Dr. Louise S. O'Donoughue (DNA Landmarks inc.) have provided guidance concerning the experimental protocols, the data analysis and interpretation, the structure of the manuscripts and have contributed ideas for the discussion part. They have also critiqued and edited the manuscripts.

To chapter 1, the co-author Sujatha Srinivasan (McGill University) contributed the mining of transposable elements in one of the six clones surveyed. I have used the protocol for data mining as established by other members of Dr. Bureau's laboratory. I was responsible for the data mining and analysis. I have written the manuscript.

To chapter 2, I contributed the phylogenetic analysis of selected transposable elements and its interpretation. The first author, Quang Hien Le (McGill University), was responsible for the identification of the Tc8 elements in *Caenorhabditis*, as well as the analysis and writing of the manuscript.

To chapter 3, I contributed the data manipulation, analysis, interpretation of the results, and the writing of the manuscript.

To chapter 4, I contributed the experimentation design, the laboratory work, the data manipulation and analysis, the interpretation of the results and the writing of the manuscript. I have received assistance from DNA Landmarks laboratory technicians for the sequencing part.

Acknowledgement

I want to thank my co-supervisors, Dr. Louise S. O'Donoughue (DNALandmarks inc.) and Dr. Thomas E. Bureau (McGill University), for their confidence, their availability, and their support. You gave me the latitude and the guidance that I was looking for. I also want to thank Dr. Anne Bruneau (Universite de Montreal) for advice provided.

I am grateful to FCAR (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche), McGill University, and DNA Landmarks inc. for financial support during the two years of my M. Sc.

I want to acknowledge members of DNA Landmarks laboratories and of Dr. Bureau's laboratory at McGill for their help and friendship.

Merci à ma famille !

General introduction

The first transposable element (TE) to be characterized was discovered by Barbara McClintock in 1947. The activity of the *Ac/Ds* system of TEs was found to be responsible for the unexpected phenotype (changing color of the kernels) observed in maize (McClintock 1947; McClintock 1951; McClintock 1953). Afterward, TEs were identified in virtually all organisms. The current classification of TEs into classes is based on their mechanisms of mobility. Most eukaryotic TEs identified can be classified either in Class I or Class II, although some elements remain unclassified. Class I TEs move via an RNA intermediate that is reverse transcribed prior to integration into the genome (Finnegan 1992; Grandbastien 1992; Kunze *et al.* 1997). Class II elements are often termed DNA transposons, as they move directly as DNA fragments. They can both excise and insert into a new location (Finnegan 1992; Flavell *et al.* 1994; Kunze *et al.* 1997). Elements from either class generate a duplication of the target site sequence upon insertion.

Proteins encoded by several class I and class II TEs share a specific motif, which may suggest a common evolutionary history between the two classes of elements. The DD(D/E) motif had been identified in the transposases of members of several bacterial IS families and within the integrase domain of retroviruses and retrotransposons (Fayet *et al.* 1990; Kulkosky *et al.* 1992). Later this motif was determined to be present in the eukaryotic counterparts of some transposon superfamilies, such as the IS*630*/Tc1/*mariner* superfamily (Capy *et al.* 1997; Capy *et al.* 1996; Doak *et al.* 1994; Langin *et al.* 1995; Robertson and Lampe 1995; Vos and Plasterk 1994). The three residues are essential for transposition as they define a cation binding site necessary for cleavage and strand transfer reactions (Kulkosky *et al.* 1992; Lohe *et al.* 1997; Vos and Plasterk 1994). This conserved region of the proteins has been used to clarify the phylogenetic relationships among TEs from various organisms (Capy *et al.* 1997; Capy *et al.* 1996).

In plants, the major types of class I TEs are retrotransposons that are delimited by long terminal repeat (LTR) sequences, non-LTR retrotransposons or LINEs (long interspersed

nuclear elements), and SINEs (short interspersed nuclear elements) (Kunze *et al.* 1997). LTR retrotransposons such as the tobacco *Tnt1* (Grandbastien *et al.* 1989) and *Tto1* (Hirochika 1993), the maize *Bs1* (Jin and Bennetzen 1989), the rice *Tos17* (Hirochika *et al.* 1996) are active plant *Ty1-copia*-like element, while the lily *del*1-46 (Smyth *et al.* 1989) and rice RIRE9 (Li *et al.* 2000) belong to the *Ty3-Gypsy* group of retroelements. LINE-like elements include elements such as *Cin4* in maize (Schwarz-Sommer *et al.* 1987), *del*2 in lily (Leeton and Smyth 1993) and *Tal*1-1 in Arabidopsis (Wright *et al.* 1996) Plant SINEs have first been identified in rice (Mochizuki *et al.* 1992; Umeda *et al.* 1991).

Class II TEs are very diverse elements that can be grouped into superfamilies. Generally, DNA transposons are delimited by terminal inverted repeat (TIR) sequences that are typical of the family. The *Ac*-like superfamily (McClintock 1951; McClintock 1984)) groups elements such as the maize *Activator* (McClintock, 1947 #296; McClintock, 1951 #106; McClintock, 1953 #297), the *Tam3* element from *Antirrhinum* (Hehl *et al.* 1991) and *Tag1* from the Arabidopsis genome (Tsay *et al.* 1993). The CACTA-like superfamily includes the maize *Enhancer* (Pereira *et al.* 1986), the element *Tam1* from Antirrhinum (Bonas *et al.* 1984) and the element *Tnr3* found in rice (Motohashi *et al.* 1996). MULEs include elements similar to the maize *Mutator* element (Chandler and Hardeman 1992; Robertson 1978). *Soymar1* (Jarvik and Lark 1998) is a soybean TE that belongs to the superfamily Tc1/*mariner*, which may be the class II TEs with the widest distribution among living organisms (Hartl *et al.* 1997a; Hartl *et al.* 1997b; Plasterk *et al.* 1999). Most miniature inverted-repeat transposable elements (MITEs) (Bureau *et al.* 1996; Bureau and Wessler 1992; Bureau and Wessler 1994a; Bureau and Wessler 1994b) are similar to *Stowaway* (Bureau and Wessler 1994b), *Tourist* (Bureau and Wessler 1992; Bureau and Wessler 1994a) or *Emigrant* (Casacuberta *et al.* 1998). *Tourist* and *Stowaway*-like elements have been identified in both monocots and dicots, while *Emigrant* elements have first been identified in Arabidopsis and their presence in monocots have not been reported yet. Recently, an element reported as *Basho* or *Helitron* was identified in several plant species and in *Caenorhabditis elegans* (Kapitonov and Jurka 2001; Le *et al.* 2000). This TE cannot easily be classified as either class I or class

II. This element is apparently a DNA transposon that moves via a rolling-circle replicative mechanism, which does not generate target site duplications (TSD) (Kapitonov and Jurka 2001).

TEs have been described as genetic parasites or junk DNA mostly based on the lack of evidence for positive effects on the host genome, and because they can multiply and become very abundant in those genomes (Doolittle and Sapienza 1980; Orgel and Crick 1980). Even with accumulating evidence for the contribution of TEs to the plasticity of host genomes, this conception of TEs has persisted in the literature (Charlesworth and Langley 1989; Engels 1996; Golding *et al.* 1986; Hartl 1988; Maynard Smith 1988; Maynard Smith and Szathmary 1995; Starlinger 1993). Nowadays, most authors agree that TEs do play an important role in gene and genome evolution and that they may be seen as a source of genetic variability that is used in evolution (Bennetzen 2000; Cohen and Shapiro 1980; Finnegan 1989; Flavell *et al.* 1994; Ginzburg *et al.* 1984; Kidwell and Lisch 1997; Lozovskaya *et al.* 1995; Makalowski 2000; McClintock 1951; McClintock 1978; Nevers and Saedler 1977; Schwarz-Sommer *et al.* 1985; Wessler *et al.* 1995; White *et al.* 1994). They are thought to promote evolutionary changes that may provoke speciation (Mcdonald 1995; Mcfadden and Knowles 1997).

Specifically, TE insertions may alter the structure, expression and function of genes (Britten 1996a; Britten 1996b; Coen *et al.* 1986; Flavell *et al.* 1994; Gierl *et al.* 1989; Kidwell and Lisch 1997; Mcdonald 1993; Mcdonald 1995; Schwarz-Sommer and Saedler 1987). They may contribute promoter regions or enhancer sequences, they may contribute termination or polyadenylation signals, or they may affect the splicing process (Britten 1996b; Horowitz *et al.* 1984; Purugganan and Wessler 1992). TE insertions in coding regions may result in inactivation of the protein or in the creation of an intron (Flavell *et al.* 1994; Purugganan and Wessler 1992). At the genome level, the presence of TEs obviously has an impact on genome size and may promote various chromosome rearrangements as a result of the transposition process or in facilitating non homologous recombinations (Flavell *et al.* 1994; Gray 2000; Kalendar *et al.* 2000). Although most transposition events are probably deleterious to the host genome, the insertions that are

beneficial or neutral have a good chance to survive and become a major component of the repetitive fraction of genomes (Charlesworth and Langley 1991)

With more and more TEs being characterized, it becomes obvious that these sequences are fundamental components of genomes and that TE-based strategies may be developed for various types of genome studies (Enoki and Al. 1999; Kunze *et al.* 1997). For example, *Ac* (Fedoroff *et al.* 1984; Schultes *et al.* 1996), *Tam* (Bradley *et al.* 1996; Coen *et al.* 1990), and *Mu* (Bensen *et al.* 1995; Martienssen *et al.* 1989) have been used for gene tagging experiments in plant. Polymorphisms resulting from the presence or absence of a TE in a locus, are now used in phylogenetic analysis. Class I elements were found to be very useful markers, mostly because these elements do not excise (Hillis 1999; Lum *et al.* 2000; Miyamoto 1999; Shedlock and Okada 2000). For example, SINEs and LINEs, insertion polymorphisms have provided new insights about the relationships between toothed whales and their relatives (Nikaido *et al.* 2001). Because they are not well documented in plants, SINEs insertions have mostly been used in phylogenetic analyses of animal species. However, SINEs (Hirano *et al.* 1994) and *Stowaway*-like MITEs (Kanazawa *et al.* 2000) insertions from rice have been superimposed on previously published trees, insertion polymorphisms of SINEs and other TEs have been used to classify rice strains into groups (Mochizuki *et al.* 1993) and the nucleotide sequences of homologous *Tourist-like* MITE insertions (Iwamoto *et al.* 1999) have been used in phylogenetic analysis in *Oryza*.

The objective of this project was first to characterize the TE content of the rice genome by looking for repetitive sequences and typical TE structure in rice genomic sequences and then to assay the use of transposable elements as phylogenetic markers in rice. The rice genome has been selected because of the availability of large-insert genomic clone sequences on public databases, and because very little was known about rice TEs two years ago. The rice genome has also been selected because several phylogenetic analyses of this genus, using different types of markers and different methods, have been published. Comparison of our results with the current literature was therefore possible.

Logical bridges linking the different chapters

As reported in chapter 1, a survey of transposable elements done in rice genomic sequences, allowed for identification and characterization of several TEs (Turcotte *et al.* 2001). Also, a large *Stowaway* element with an open reading frame (ORF) coding for a putative transposase was identified. This ORF show amino acid similarity to ORFs of members to the Tc1/*mariner* superfamily of transposable elements, while the TIRs and TSD are clearly reminiscent of *Stowaway* elements . In a previous survey for transposable elements in *Arabidopsis thaliana* (Le *et al.* 2000), a putative transposase for a *Tourist* element had been identified. A similar survey performed on *C. elegans* sequences also revealed large *Tourist* elements with ORFs (Le *et al.* 2001). The *Tourist* ORFs from both species show amino acid similarity to transposase of bacterial transposable elements, while the TIRs and TSD are clearly reminiscent of *Tourist*. Also, another group of researcher has identified a large *Emigrant* element in *Arabidopsis thaliana*, which also has an ORF similar to those of Tc1/*mariner*-like elements, and TIRs and TSD reminiscent to the *Emigrant* MITEs (Feschotte and Mouches 2000). *Tourist*, *Stowaway* and *Emigrant* are the three major types of MITE found in plants. Analysis of their transposases revealed that they all possess the DD(D/E) motif common to several class I and class II TEs. Therefore, in Chapter 2, I have studied the phylogenetic relationships between *Tourist* transposase and bacterial IS sequences. In Chapter 3, I have investigated the phylogenetic relationships between *Emigrant* and *Stowaway* MITEs and several members of the IS*630*/Tc1/*mariner* superfamily. Since some TEs have been shown to be useful in phylogenetic analysis, it appeared interesting to verify if elements selected from different categories would be equally good as phylogenetic markers. Therefore, in chapter 4, I have amplified and sequenced ten TEs to assay TEs as phylogenetic markers in rice. Also, the previously reported phylogenetic analyses of the Oryza species were considered as a control, to evaluate the performance of TEs in reconstructing the phylogenetic relationships.

References

Bennetzen, J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-69.

Bensen, R. J., G. S. Johal, V. C. Crane, J. T. Tossberg, P. S. Schnable, R. B. Meeley and S. P. Briggs, 1995 Cloning and characterization of the maize *An1* gene. Plant Cell 7: 75-84.

Bonas, U., H. Sommer and H. Saedler, 1984 the 17-Kb *Tam1* element of *Antirrhinum majus* induces a 3-bp duplication upon integration into the *chalcone synthase* gene. EMBO J 3: 1015-1019.

Bradley, D., R. Carpenter, L. Copsey, C. Vincent, S. Rothstein and E. Coen, 1996 Control of inflorescence architecture in *Antirrhinum*. Nature 379: 791-797.

Britten, R. J., 1996a Cases of ancient mobile element DNA insertions that now affect gene regulation. Mol Phylogenet Evol 5: 13-17.

Britten, R. J., 1996b DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A 93: 9374-9377.

Bureau, T. E., P. C. Ronald and S. R. Wessler, 1996 A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc Natl Acad Sci U S A 93: 8524-9.

Bureau, T. E. and S. R. Wessler, 1992 *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4: 1283-94.

Bureau, T. E. and S. R. Wessler, 1994a Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc Natl Acad Sci U S A 91: 1411-5.

Bureau, T. E. and S. R. Wessler, 1994b Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6: 907-16.

Capy, P., T. Langin, D. Higuet, P. Maurer and C. Bazin, 1997 Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100: 63-72.

Capy, P., R. Vitalis, T. Langin, D. Higuet and C. Bazin, 1996 Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J Mol Evol 42: 359-68.

Casacuberta, E., J. M. Casacuberta, P. Puigdomenech and A. Monfort, 1998 Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterization of the *Emigrant* family of elements. Plant J 16: 79-85.

Chandler, V. L. and K. Hardeman, 1992 The *Mu* elements of *Zea mays*. Adv. Genet. 30: 77-122.

Charlesworth, B. and C. H. Langley, 1989 the population genetics of *Drosophila* transposable elements. Annu Rev Genet 23: 251-287.

Charlesworth, B. and C. H. Langley (1991). Population genetics of transposable elements in *Drosophila*. In: Evolution at the molecular level. Selander, R. K., Clark, A. G. and Whittam, T. S. Eds. Sunderland, MA, Sinauer Associates.

Coen, E. S., R. Carpenter and C. Martin, 1986 Transposable elements generate novel spatial patterns of gene expression in *Anthirrhinum majus*. Cell 47: 285-296.

Coen, E. S., J. M. Romero, S. Doyle, R. Elliott, G. Murphy and R. Carpenter, 1990 *floricaula*: a homeotic gene required for flower development in *Antirrhinum majus*. Cell 63: 1311-1322.

Cohen, S. N. and J. A. Shapiro, 1980 Transposable genetic elements. Sci. Am. 242: 40-49.

Doak, T. G., F. P. Doerder, C. L. Jahn and G. Herrick, 1994 A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc Natl Acad Sci U S A 91: 942-6.

Doolittle, W. F. and C. Sapienza, 1980 Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603.

Engels, W. R. (1996). *P* elements in *Drosophila*. In: Transposable elements. Saedler, H. and Gierl, A. Eds. Heidelberg, Springer-Verlag: 103-123.

Enoki, H. and e. al., 1999 *Ac* as a tool for the functional genomics of rice. The Plant J 19: 605-613.

Fayet, O., P. Ramond, P. Polard, M. F. Prere and M. Chandler, 1990 Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? Mol Microbiol 4: 1771-7.

Fedoroff, N., D. B. Furtek and O. E. Nelson, 1984 Cloning of the *bronze* locus in maize by a simple and generalizable procedure using the transposable controlling element *Activator* (*Ac*). Proc Natl Acad Sci U S A 81: 3825-3829.

Feschotte, C. and C. Mouches, 2000 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: 730-7.

Finnegan, D. J., 1989 Eukaryotic transposable element and genome evolution. Trends Genet. 5: 103-107.

Finnegan, D. J., 1992 Transposable elements. Curr. Opin. Genet. Devel 2: 861-867.

Flavell, A. J., S. R. Pearce and A. Kumar, 1994 Plant transposable elements and the genome. Curr Opin Genet Dev 4: 838-44.

Gierl, A., H. Saedler and P. A. Peterson, 1989 Maize transposable elements. Annu Rev Genet 23: 71-85.

Ginzburg, L. R., P. M. Bingham and S. Yoo, 1984 On the theory of speciation induced by transposable elements. Genetics 107: 331-341.

Golding, G. B., C. F. Aquadro and C. H. Langley, 1986 Sequence evolution within populations under multiple types of mutations. Proc Natl Acad Sci U S A 83: 427-431.

Grandbastien, M.-A., 1992 Retroelements in higher plants. Trends Genet 8: 103-108.

Grandbastien, M. A., A. Spielmann and M. Caboche, 1989 *Tnt1*, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature 337: 376-380.

Gray, Y. H., 2000 It takes two transposons to tango: transposable element-mediated chromosomal rearrangements. Trends Genet 16: 461-468.

Hartl, D. H., 1988 A primer of population genetics. Sunderland, MA, Sinauer inc.

Hartl, D. L., A. R. Lohe and E. R. Lozovskaya, 1997a Modern thoughts on an ancyent marinere: function, evolution, regulation. Annu Rev Genet 31: 337-58.

Hartl, D. L., A. R. Lohe and E. R. Lozovskaya, 1997b Regulation of the transposable element *mariner*. Genetica 100: 177-84.

Hehl, R., W. K. F. Nacken, M. Krause, H. Saedler and H. Sommer, 1991 Structural analysis of *Tam3*, a transposable element from *Antirrhinum majus*, reveals homologies to the *Ac* element from maize. Plant Mol Biol 16: 369-371.

Hillis, D. M., 1999 SINEs of the perfect character. Proc Natl Acad Sci U S A 96: 9979-81.

Hirano, H. Y., K. Mochizuki, M. Umeda, H. Ohtsubo, E. Ohtsubo and Y. Sano, 1994 Retrotransposition of a plant SINE into the *wx* locus during evolution of rice. J Mol Evol 38: 132-7.

Hirochika, H., 1993 Activation of tobacco retrotransposons during tissue culture. EMBO J 12: 2521-2528.

Hirochika, H., K. Sugimoto, Y. Otsuki, H. Tsugawa and M. Kanda, 1996 Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci U S A 23: 7783-7788.

Horowitz, M., S. Luria and S. Osawa, 1984 Mechanism of activation of the mouse *c-mos* oncogene by the LTR of an intracisternal A-particule gene. EMBO J 3: 2937-2941.

Iwamoto, M., H. Nagashima, T. Nagamine, H. Higo and K. Higo, 1999 A *tourist* element in the 5'-flanking region of the catalase gene *CatA* reveals evolutionary relationships among *Oryza* species with various genome types. Mol Gen Genet 262: 493-500.

Jarvik, T. and K. G. Lark, 1998 Characterization of *Soymar1*, a *Mariner* element in soybean. Genetics 149: 1569-1574.

Jin, Y. K. and J. L. Bennetzen, 1989 Structure and coding properties of *Bs1*, a maize retrovirus-like transposon. Proc Natl Acad Sci U S A 86: 6235-6239.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo and A. H. Schulman, 2000 Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci U S A 97: 6603-6607.

Kanazawa, A., M. Akimoto, H. Morishima and Y. Shimamoto, 2000 Inter- and intra-specific distribution of *Stowaway* transposable elements in AA-genome species of wild rice. Theor Appl Genet 101: 327-335.

Kapitonov, V. V. and J. Jurka, 2001 Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. U S A 98: 8714-8719.

Kidwell, M. G. and D. Lisch, 1997 Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94: 7704-7711.

Kulkosky, J., K. S. Jones. R. A. Katz, J. P. Mack and A. M. Skalka, 1992 Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. Mol Cell Biol 12: 2331-8.

Kunze, R., H. Saedler and W. E. Lonnig, 1997 Plant transposable elements. Adv. Bot. Res. 27: 331-469.

Langin, T., P. Capy and M. J. Daboussi, 1995 The transposable element *impala*, a fungal member of the Tc1-*mariner* superfamily. Mol Gen Genet 246: 19-28.

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

Le, Q. H., S. Wright, Z. Yu and T. Bureau, 2000 Transposon diversity *in Arabidopsis thaliana*. Proc Natl Acad Sci U S A 97: 7376-81.

Leeton, P. J. and D. R. Smyth, 1993 An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Mol. Gen. Genet. 237: 97-104.

Li, Z. Y., S. Y. Chen, X. W. Zheng and L. H. Zhu, 2000 Identification and chromosomal localization of a transcriptionally active retrotransposon of *Ty3-gypsy* type in rice. Genome 43: 404-8.

Lohe, A. R., D. De Aguiar and D. L. Hartl, 1997 Mutations in the *mariner* transposase: the D,D(35)E consensus sequence is nonfunctional. Proc Natl Acad Sci U S A 94: 1293-7.

Lozovskaya, E. R., D. L. Hartl and D. A. Petrov, 1995 Genomic regulation of transposable elements in *Drosophila*. Curr Opin Genet Dev 5: 768-773.

Lum, J. K., M. Nikaido, M. Shimamura, H. Shimodaira, A. M. Shedlock, N. Okada and M. Hasegawa, 2000 Consistency of SINE insertion topology and flanking sequence tree: quantifying relationships among cetartiodactyls. Mol Biol Evol 17: 1417-24.

Makalowski, W., 2000 Genomic scrap yard: how genomes utilize all that junk. Gene 259: 61-7.

Martienssen, R. A., A. Barkan, M. Freeling and W. C. Taylor, 1989 Molecular cloning of a maize gene involved in photosynthetic membrane organization that is regulated by Robertson's *Mutator*. EMBO J 8: 1633-1639.

Maynard Smith, J., 1988 Evolutionary Genetics. Oxford, Oxford University Press.

Maynard Smith, J. and E. Szathmary, 1995 The major transitions in evolution. Oxford, Freeman, W. H.

McClintock, B., 1947 Cytogenetic studies of maize and *Neurospora*. Carnegie institution of Washington year book 46: 146-152.

McClintock, B., 1951 Chromosomal organization and genic expression. Cold Spring Harbor Symp. Quant. Biol. 16: 13-47.

McClintock, B., 1953 Introduction of instability at selected loci in maize. Genetics 38: 579-599.

McClintock, B. (1978). Mechanism that rapidly reorganize the genome. In: The discovery and characterization of transposable elements. Moore, J. A. Ed. New York, Garland publishing.

McClintock, B., 1984 The significance of responses of the genome to challenges. Science 226: 792-801.

McDonald, J. F., 1993 Transposable elements and evolution, Kluer, Dordrecht.

McDonald, J. F., 1995 Transposable elements: possible catalysts of organismic evolution. Trends Ecol Evol 10: 123-126.

McFadden, J. and G. Knowles, 1997 Escape from evolutionary stasis by transposon-mediated deleterious mutations. J Theor Biol 186: 441-7.

Miyamoto, M. M., 1999 Molecular systematics: Perfect SINEs of evolutionary history? Curr Biol 9: R816-9.

Mochizuki, K., H. Ohtsubo, H. Hirano, Y. Sano and E. Ohtsubo, 1993 Classification and relationships of rice strains with AA genome by identification of transposable elements at nine loci. Jpn J Genet 68: 205-17.

Mochizuki, K., M. Umeda, H. Ohtsubo and E. Ohtsubo, 1992 Characterization of plant SINE, p-SINE1, in rice genomes. Jap. J. Genet. 57: 155-166.

Motohashi, R., E. Ohtsubo and H. Ohtsubo, 1996 Identification of *Tnr3*, a *Supressor-Mutator/Enhancer*-like transposable element from rice. Mol. Gen. Genet. 150: 148-152.

Nevers, P. and H. Saedler, 1977 Transposable genetic elements as agents of gene instability and chromosome rearrangements. Nature 268: 109-115.

Nikaido, M., F. Matsuno, H. Hamilton, R. L. Brownell, Jr., Y. Cao, W. Ding, Z. Zuoyan, A. M. Shedlock, R. E. Fordyce, M. Hasegawa, *et al.*, 2001 Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. Proc Natl Acad Sci U S A 98: 7384-9.

Orgel, L. E. and F. H. C. Crick, 1980 Selfish DNA: the ultimate parasite. Nature 284: 604-607.

Pereira, A., H. Cuypers, A. Gierl, Z. Schwarz-Sommer and H. Saedler, 1986 Molecular analysis of the En/Spm transposable element system of *Zea mays*. EMBO J 5: 835-841.

Plasterk, R. H., Z. Izsvak and Z. Ivics, 1999 Resident aliens: the Tc1/*mariner* superfamily of transposable elements. Trends Genet 15: 326-32.

Purugganan, M. D. and S. R. Wessler, 1992 The splicing of transposable elements and its role in intron evolution. Genetica 86: 295-303.

Robertson, D. S., 1978 Characterization of a *Mutator* system in maize. Mutat. Res. 51: 21-28.

Robertson, H. M. and D. J. Lampe, 1995 Distribution of transposable elements in arthropods. Annu Rev Entomol 40: 333-57.

Schultes, N. P., T. P. Brutnell, A. Allen, S. L. Dellaporta, T. Nelson and J. Chen, 1996 *Leaf permease 1* gene of maize is required for chloroplast development. Plant Cell 8: 463-475.

Schwarz-Sommer, Z., A. Gierl, H. Cuipers, P. A. Peterson and H. Saedler, 1985 Plant transposable elements generate the DNA sequence diversity needed in evolution. EMBO J 4: 591-597.

Schwarz-Sommer, Z., L. Leclercq, E. Gobel and H. Saedler, 1987 *Cin4*, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO J 6: 3873-3883.

Schwarz-Sommer, Z. and H. Saedler, 1987 Can plant transposable elements generate novel regulatory units ? Mol Gen Genet 209: 213-221.

Shedlock, A. M. and N. Okada, 2000 SINE insertions: powerful tools for molecular systematics. Bioessays 22: 148-60.

Smyth, D. R., P. Kalitsis, J. L. Joseph and S. J. W., 1989 Plant retrotransposon from *Lilium henryi* is related to ty3 of yeast and the *gypsy* group of *Drosophila*. Proc Natl Acad Sci U S A 86: 5015-5019.

Starlinger, P., 1993 What do we still need to know about transposable element *Ac*. Gene 135: 251-255.

Tsay, Y. F., M. J. Frank, T. Page, C. Dean and N. M. Crawford, 1993 Identification of a mobile endogenous transposon in *Arabidopsis thaliana*. Science 260: 342-344.

Turcotte, K., S. Srinivasan and T. Bureau, 2001 Survey of transposable elements from rice genomic sequences. Plant J 25: 169-79.

Umeda, M., H. Ohtsubo and E. Ohtsubo, 1991 Diversification of the rice *Waxy* gene by insertion of mobile elements into introns. Jap. J. Genet. 66: 569-586.

Vos, J. C. and R. H. Plasterk, 1994 Tc1 transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. EMBO J 13: 6125-32.

Wessler, S. R., T. E. Bureau and S. E. White, 1995 LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5: 814-21.

White, S. E., L. F. Habera and S. R. Wessler, 1994 Retrotransposons in the flanking regions of normal plant genes: a role for *copia*-like element in the evolution of gene structure and expression. Proc Natl Acad Sci U S A 91: 792-796.

Wright, D. A., N. Ke, J. Smalle, B. M. Hauge, H. M. Goodman and D. F. Voytas, 1996 Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana*. Genetics 142: 569-578.

CHAPTER 1

Survey of transposable elements from rice genomic sequences

Title: **Survey of Transposable Elements from Rice Genomic Sequences**

Authors: **Kime Turcotte, Sujatha Srinivasan and Thomas Bureau**

Address: Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec, Canada, H3A 1B1

Author for correspondence: Thomas Bureau, Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec, Canada, H3A 1B1
Tel: (514) 398-6472, Fax: (514) 398-5069, Email: thomas_bureau@maclan.mcgill.ca

Author's email addresses:
Kime Turcotte: kturco2@po-box.mcgill.ca
Sujatha Srinivasan: ssrini1@po-box.mcgill.ca
Thomas Bureau: thomas_bureau@maclan.mcgill.ca

Suggested running title: Transposable elements in rice

Key words: transposon, gene annotation, MITE, *Tourist*, *Stowaway*, MULE, *Basho*

The sequences investigated in this paper were accessed from the GenBank database (GI numbers 5091496, 5042437, 5670155, 5922603, 6006355, and 6069643).

Word count (excluding figures and tables): 7,179

Abstract

Oryza sativa L. (domesticated rice) is a monocotyledonous plant, and its 430 Mb genome has been targeted for complete sequencing. We performed a high-resolution computer-based survey for transposable elements on 910 Kb of rice genomic DNA sequences. Both class I and II transposable elements were present, contributing 19.9% of the sequences surveyed. Class II elements greatly outnumbered class I elements (166 vs 22), although class I elements made up a greater percentage (12.2% vs 6.6%) of nucleotides surveyed. Several *Mutator*-like elements (MULEs) were identified, including rice elements that harbor truncated host cellular genes. MITEs (miniature inverted-repeat transposable elements) account for 71.6% of the mined transposable elements and are clearly the predominant type of transposable element in the sequences examined. Moreover, a putative *Stowaway* transposase has been identified based on shared sequence similarity with the mined MITEs and previously identified plant *mariner*-like elements (MLEs). Members of a group of novel rice elements resembling the structurally unusual members of the *Basho* family in Arabidopsis suggest a wide distribution of these transposons among plants. Our survey provides a preview of transposable element diversity and abundance in rice, and allows for comparison with genomes of other plant species.

Introduction

Repetitive sequences can account for a substantial portion of many eukaryotic genomes. Most of the moderately repeated sequences are mobile genetic elements, also called transposons or transposable elements (TEs) because they have the potential to insert into new genomic locations. TE insertions typically result in the duplication of a target sequence (target site duplication, TSD). The classification of TEs is based on their different modes of mobility (Charlesworth *et al.* 1994). Class I elements, or retroelements, move via an RNA intermediate that is reverse transcribed prior to integration into the genome (Finnegan 1992; Grandbastien 1992). Once inserted, class I elements cannot excise. These elements include retrotransposons, which are usually flanked by long terminal-repeats (LTRs) and may harbor the genes that encode the proteins required for their reverse transcription and integration (Finnegan 1992; Kubis *et al.* 1998). Other class I elements such as non-LTR retrotransposons or LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements) and processed pseudogenes are structurally diverse and typically terminate with a poly-A+T rich tail. Class I elements in plants are predominantly members of the Ty1-*copia* and Ty3-*gypsy* groups of LTR-retrotransposons, although both LINEs (Flavell *et al.* 1994; Leepton and Smyth 1993) and SINEs (Hirano *et al.* 1994; Mochizuki *et al.* 1992; Umeda *et al.* 1991) have been reported.

Class II elements, often termed DNA transposons, can both excise and insert into a new location (Finnegan 1992; Flavell 1994). They typically possess terminal inverted-repeats (TIRs), may encode a transposase, and are highly variable in length. Many well-described two component element systems such as the *Ac/Ds* (McClintock 1951; McClintock 1984; Scortecci *et al.* 1999), *Spm/dSpm* (Fedoroff 1999; Masson *et al.* 1989; Raina *et al.* 1998) and *MuDR/Mu* (Chandler and Hardeman 1992; Lisch *et al.* 1999; Robertson 1978; Yoshida *et al.* 1998) were first described in maize and were originally discovered by genetic analysis and subsequently cloned from mutable host alleles. *Ac/Ds*, *Spm/dSpm* and *MuDR/Mu* represent the archetypal elements defining the *Ac*-like, CACTA-like, and *Mutator*-like element (MULE) plant TE superfamilies, respectively.

Computer-based sequence similarity searches in plants have detected an abundant group of short elements (<500 bp) with conserved TIRs and a 2 or 3 bp target site duplication (TSD) (Bureau and Wessler 1992; Bureau and Wessler 1994a; Bureau and Wessler 1994b; Bureau *et al.* 1996; Wessler *et al.* 1995). These elements, called MITEs (miniature inverted-repeat transposable elements), are defined as either *Tourist*-like, having a target site preference for 5'-TAA-3', or *Stowaway*-like, having a target site preference for 5'-TA-3'. Although the vast majority of MITEs lack coding potential and are obviously non-autonomous, *Tourist*-like elements containing ORFs encoding a putative transposase have been recently described (Le *et al.* 2000).

Oryza sativa L., cultivated rice, is a model monocotyledonous plant. Its genome of ~430 Mb is small relative to other grass species (Messing and Llaca 1998). Rice has recently been targeted for complete sequencing by an international consortium (International Rice Genome Sequencing Project, http://rgp.dna.affrc.go.jp/). As a general rule, genome size is correlated to the amount of repetitive sequences (Flavell *et al.* 1994; Uozu *et al.* 1997; Wang *et al.* 1999). Clusters of genes in a conserved order suggests collinearity of segments of chromosomes between grass genomes except that intergenic regions in plants with large genomes (e.g. maize) have numerous nested retrotransposons, which are less frequent in grasses with small genomes (e.g. rice) (Bennetzen *et al.* 1998; Feuillet and Keller 1999). Copy number estimates using hybridization protocols indicate a low abundance (100 copies per haploid genome) of Ty1-*copia* retrotransposons in the *O. sativa* genome (Wang *et al.* 1999). SINEs, which predominate in mammalian genomes, are also present in the rice genome (Hirano *et al.* 1994; Mochizuki *et al.* 1992; Umeda *et al.* 1991). Their corresponding insertion polymorphisms have been used to fingerprint rice strains (Mochizuki *et al.* 1993) and hence may reflect recent mobility. Numerous MITEs have been identified in the non-coding regions of rice genes (Bureau *et al.* 1996; Bureau and Wessler 1994a; Bureau and Wessler 1994b; Ohtsubo and Sekine 1996). Likewise, *Ac*-like elements have been mined in the 3' flanking region of the rice disease resistance gene *Xa21* (Song *et al.* 1998). The presence of MULEs in rice was first suggested by the identification of cDNAs coding for a polypeptide sharing amino acid

similarity with MURA, the transposase for *MuDR* elements of maize (Eisen *et al.* 1994). Also, a CACTA-like element called Tnr3 (transposon of rice 3) is present in multiple copies in rice (Motohashi *et al.* 1996). Moreover, previous computer-based surveys of rice TEs in 105 genomic gene sequences (Bureau *et al.* 1996), a 66 Kb region containing the *Xa-21* complex gene locus (Song *et al.* 1998), a 340 Kb region containing *Adh1-Adh2* (Tarchini *et al.* 2000), and over 73,000 short genomic fragments (Sequence Tagged Connectors or STCs) representing over 50 Mb of genomic sequence (Mao *et al.* 2000) revealed that many of the previously described plant TEs have representation in the rice genome and that MITEs appear to predominate. The approaches used in these surveys were based on shared sequence similarity and subsequent examination of past mobility at a time when only a limited amount of rice sequence was available (Bureau *et al.* 1996; Song *et al.* 1998), on sequence similarity with previously identified TEs and TE-related ORFs (Mao *et al.* 2000) or on repetitiveness (Tarchini *et al.* 2000). In this report, we have taken advantage of the availability of sequences from several large-insert genomic clones and carried out a high-resolution survey of TEs in rice utilizing shared sequence similarity, structural analyses, and demonstration of past mobile histories.

Materials and Methods

Transposon mining

The sequences of six large-insert genomic clones obtained from GenBank (http://www.ncbi.nlm.nih.gov/) cover 910,705 bp of rice genomic DNA located on chromosomes 1 (GI5922603), 6 (GI5091496, 6006355, and 6069643), 10 (GI5042437) and 11 (GI5670155). The complete sequence of each clone has been initially examined by using a sliding window of 4 Kb in size as a query in sequence similarity searches against the GenBank database between September 1999 and January 2000 (BLAST 2.0; Altschul *et al.*, 1990; http://www.ncbi.nlm.nih.gov/blast/). Smaller size fragments were used for further analysis and subsequent sequence similarity searches were performed up to August 2000. Sequences that shared significant similarity to the query (BLAST score › 80) were compiled, excluding sequences annotated for non-mobility related genes, as well as simple sequence repeats. Thus, the compiled sequences are assumed to belong to the same 'group' of repetitive sequences. In addition, BLAST2 (Tatusova and Madden 1999) was used to identify putative TEs with LTRs or long TIRs by comparing each clone with its reverse complement. The compiled repetitive sequences were categorized as TEs based on nucleotide sequence similarity, structural features (TIRs, LTRs and poly-A+T regions), coding capacity and target site duplications (TSDs). TEs were sorted into previously described groups of TEs based on shared nucleotide or amino acid sequence similarity and shared structural similarity. With the exception of structurally novel TEs, element groups were given generic names based on previously established TE families or superfamilies. Groups were defined as novel if the members were, to the best of our knowledge, not cited in the literature or annotated in the sequence file (as of August 2000). Within a group individual elements were further identified by the GI number of the clones where they occurred. Amino acid sequence similarity was determined from annotated ORFs (psi-BLAST) or from predicted amino acid sequences (BLASTX). In order to provide evidence for the past mobility of newly identified TEs and to confirm the termini of the mined elements, we performed a search for sequences related to empty sites (RESites; Le *et al.* 2000). RESites were found using the TE flanking regions as queries in sequence similarity searches (BLAST) against the GenBank databases. The pairwise

comparison of the RESite and the TE-bearing sequence may reveal a gap corresponding to the position of the TE and the TSD. Detailed information regarding the elements mined from the 910 Kb of rice genomic DNA sequences may be accessed on our World-Wide Web site at http://soave.biol.mcgill.ca/rice/.

Data analysis

Nucleotide sequence alignments were performed using PILEUP from the UWGCG (University of Wisconsin Genetic Computer Group, Madison, WI, USA; version 10.0) program suite. Within-group percent similarities were obtained from pairwise comparisons (BLAST and BLAST2), and G+C contents were obtained using COMPOSITION (UWGCG). G+C contents of sequences flanking intact TEs were calculated as previously described (Le *et al.* 2000).

Results and Discussion

Using computer-based sequence similarity searches, we systematically mined TEs from six large-insert clones, totaling 910 Kb. Four clones are within euchromatic regions (GI5922603, 5091496, 6006355, and 6069643) while the other two clones have not been mapped (GI5042437 and 5670155; http://rgp.dna.affrc.go.jp/Seqcollab.html). Each clone had a similar gene density with approximately 30 genes/clone or 1 gene for every 5 Kb of sequence. Our searches revealed 204 repetitive sequences (Table 1), which belong to 69 different groups according to their shared sequence similarity (see Experimental Procedures). Most of them (134 sequences) belong to 28 previously described groups of class I and II TEs, based on shared sequence similarity, characteristic TSD lengths and sequences, and structural features such as LTRs, TIRs and poly-A+T regions. Forty-one new groups of TEs (69 repetitive sequences) were revealed in our survey. Sequences related to empty sites (RESites; Le *et al.* 2000) have been identified for some group members of the 41 new TEs. The pairwise comparison of the RESite and the TE-bearing sequence may reveal a gap corresponding to the position of the TE and the TSD. RESite analysis confirmed the termini of these elements and provided evidence for past transposition events (Figure 1).

The contribution of rice TEs to the portion of the genome surveyed is ~19.9% (Table 1), and the estimated average density is one TE for every 4.5 Kb of genomic DNA. Nested insertions (six TEs) account for 2.9% of the mined transposons. Forty-six TEs (22.5%), including the mined LINEs and *Explorer* elements, appeared to be truncated at one or both termini; these elements are unlikely to be currently mobile.

Based on sequence similarity to previously reported retroelements, we classified 22 mined sequences as class I TEs (Table 1). Ten are Ty3-*gypsy* retrotransposons and two are Ty1-*copia* retrotransposons. Four Ty3-*gypsy* solo-LTRs were identified with two elements immediately flanked by a 5 bp TSD. The solo-LTRs shared high sequence similarity to the LTRs of the corresponding full-length elements. Overall, we detected 5 times fewer Ty1-*copia* elements than Ty3-*gypsy* elements. The copy number of Ty1-

copia elements in rice has been estimated by Southern hybridization to be as low as 100 copies per haploid genome (Wang *et al.* 1999). However, our data suggests that the Ty1-*copia* copy number may be significantly higher. This discrepancy may reflect the sequence diversity among rice Ty1-*copia* retrotransposons and the likelihood that element distribution is not uniform (CSHL/WUGSC/PEB *Arabidopsis* Sequencing Consortium 2000; Dong *et al.* 1998; Presting *et al.* 1998). Five ORFs related to the reverse transcriptase of LINEs were identified. Because only a few copies were available, we could not define the termini of these putative LINEs based on sequence alignments; many plant LINEs exist as diverse 5' truncated copies (Grandbastien 1992; Okada *et al.* 1997). Four short sequences were classified as p-SINEs, an element group originally characterized from the rice *Waxy* gene (Hirano *et al.* 1994; Mochizuki *et al.* 1992; Umeda *et al.* 1991). Two of these are presumably truncated copies, because they partially align with annotated p-SINEs but lack a poly-A+T region and a TSD; RESites were not found.

Class II elements clearly outnumber class I elements in the surveyed rice clones. MULEs and MITEs are the predominant TE types. Two novel elements were mined that share TIR sequence similarity and target sequence size (i.e. 8 bp) with members of the *Ac*-like superfamily. A single CACTA-like TE with 31 bp TIRs, a perfect 3 bp TSD, but no coding capacity, was identified based on shared sequence similarity with a recently published rice CACTA-like element (Tarchini *et al.* 2000). MULEs constitute a very heterogeneous TE superfamily defined by long TIRs and a typical 9 bp TSD (Chandler and Hardeman 1992; Le *et al.* 2000; Yu *et al.* 2000). Seventeen mined elements falling into ten groups were classified as MULEs based on the presence of TIRs (83-571 bp), and a 9-10 bp TSD. Within a group, the TIRs show high sequence similarity (>80%), while the internal sequences are more variable; this result is consistent with *Mu* elements in maize (Talbert and Chandler 1988; Talbert *et al.* 1989) and a recent study of MULEs in *A. thaliana* (Yu *et al.* 2000). For example, the internal sequence of MULE-VIII:GI5091496a and MULE-VIII:GI5091496b differ by two indels (265 and 616 bp) and share 40-45% sequence similarity. MULE-VII:GI5922603 and 2 members of MULE-VIII (GI5922603 and 6006355) represent truncated versions of larger MULE-VIII members. These short copies coincide with only the TIRs, but are not flanked by direct

repeats as in the case of class I solo-LTRs. These apparent "solo" TIRs may belong to truncated or degenerated MULEs that share no internal sequence similarity with other mined members. Recombination between TIRs may be involved in generating these elements. In fact, inter-element recombination has been suggested to be a probable means for generating diversity among *A. thaliana* MULEs (Le *et al.* 2000). Alternatively, mobility of "solo" TIRs by a mechanism similar to bacterial IS (insertion sequence) movement, could explain the presence of these elements as well as the presence of MULE members sharing high TIR sequence similarity and low or no internal sequence similarity (Iida *et al.* 1983; Talbert and Chandler 1988). MULEs lacking TIRs (non-TIR MULEs) have been identified in *A. thaliana* (Le *et al.* 2000; Yu *et al.* 2000) but no definitive non-TIR MULEs were identified in the rice sequences surveyed.

Five mined MULEs contain a region corresponding to an annotated ORF or with similarity to hypothetical proteins or unknown proteins, but none have similarity to the maize *MuDR* transposase, MURA. Therefore, all the mined MULEs from the surveyed clones are presumably nonautonomous elements and require a MURA-like transposase from a corresponding autonomous MULE for mobility (Hershberger *et al.* 1995). At least one member of MULE-III, V, and VII on other recently sequenced large-insert clones have internal regions that share amino-acid sequence similarity with *mudrA* of maize (Mao *et al.* 2000; Tarchini *et al.* 2000; data not shown; Figure 2a). Six mined MULEs contain short regions sharing high sequence similarity with host cellular genes. The internal sequence of five members (one element from the surveyed clones and four additional elements on other clones) of the MULE-I group includes a 105 bp region, which shares 90-95% nucleotide similarity with the 5' half of the rice 5S rDNA gene (GI: 20162, position 58-162) (Figure 2a and b). In addition, MULE-IV:GI5042437 contains a 159 bp region that is 83% similar to a rice PCF2 binding protein gene (GI:2580459, position 333-493) (Figure 2c and d). The acquisition of truncated cellular genes, perhaps resulting from illegitimate recombination and repair events (Bennetzen and Springer 1994), has been recently reported as a common feature of MULEs in *A. thaliana* (Le *et al.* 2000; Yu *et al.* 2000). While only one complete MULE-IV element is present in the current database, the finding of several copies at various locations indicates that MULE-I

has transposed since the acquisition of the 5S rDNA gene fragment. Finally, although there were no sequence annotations for the mined MULEs, several elements overlap with one or two exons of apparently misannotated hypothetical proteins.

MITEs have been identified in several organisms, suggesting that they are ubiquitous components of eukaryotic genomes. In plants, the primary families of MITEs, *Tourist*-like and *Stowaway*-like, were originally reported by Bureau *et al.* (1992, 1994a, 1994b, 1996). Structural features such as the short length (74-490 bp), the presence of TIRs, a low G+C content, a target site preference for the dinucleotide 5'-TA-3' (*Stowaway*-like elements) or the trinucleotide 5'-TAA-3' (*Tourist*-like elements), and the potential to form stable secondary DNA structures are characteristic of the MITE superfamily (Bureau and Wessler 1992; Bureau and Wessler 1994b). These criteria were used to classify 146 repetitive sequences as MITEs (Table 1); these fall into 40 distinct MITE groups, of which 23 are first described in this paper. Within a group, members are very homogeneous in size and share 85-95% nucleotide sequence similarity.

Seventy rice *Stowaway*-like elements were further divided into 24 different groups based on sequence similarity; 16 of these are newly described. Generally, these elements have TIR lengths varying from 22-95 bp and an average G+C content of 28% consistent with previously reported MITE groups (Bureau *et al.* 1996; Bureau and Wessler 1992; Bureau and Wessler 1994a; Bureau and Wessler 1994b; Song *et al.* 1998; Wessler *et al.* 1995). Members of the rice *Stowaway*-like groups share the terminal most sequence of their TIRs that is 5'-CTCCCTCCRT-3' (consensus where R corresponds to G or A) (Figure 4) and all have a 5'-TA-3' target site preference.

Members of the *Tourist*-like family of MITEs slightly outnumber *Stowaway*-like elements with 76 elements identified falling into nine previously reported and seven newly discovered groups. These new elements have the general features of MITEs as described above, and are flanked by a 3 bp TSD and an average G+C content of 34% typical of other *Tourist*-like elements (Table 1). Based on nucleotide substitutions and indels, the rice *Tourist*-I group can be further subdivided into three distinct subgroups (data not

shown). Surprisingly, while no other MITE or other mined TE showed significant similarity (BLAST score > 80) to sequences from another plant genome, several *Tourist*-I elements shared 85% overall sequence similarity with *Tourist* elements from *Sorghum bicolor* (Figure 3). This level of conservation between the rice and Sorghum MITEs may reflect a cross-species transfer (i.e. horizontal transmission), as previously suggested (Zhang and Kochert 1998). Horizontal transmission has been well documented for class I and II TEs and is proposed to be a possible escape mechanism from host inactivation (Daniels *et al.* 1990; Jordan *et al.* 1999; Robertson 1993). In fact, the *Tourist*-I elements with high sequence similarity to elements in *S. bicolor* appear to have been active in rice recently; seven nearly identical elements were mined from the surveyed clones.

Barring horizontal transmission, the presence of several *Stowaway*-like groups and *Tourist*-like groups may reflect element diversification during the evolution of the rice genome. Some MITE groups with high sequence similarity (>90%) but few representatives, such as *Stowaway*-II, *Stowaway*-IV, *Stowaway*-V, *Stowaway*-VII, and *Tourist*-VI are most likely recently evolved groups. MITEs account for 71.6% of the mined rice TEs (Figure 5 and Table 1), or, in terms of sequence contribution, one-fifth (19.6%) of the mined TEs nucleotide sequences. Therefore, our results confirm the clear predominance of MITEs in the rice genome (Bureau *et al.* 1996; Mao *et al.* 2000; Tarchini *et al.* 2000).

A new family of transposable elements called *Basho* has been recently identified in *A. thaliana* (Le *et al.* 2000). *Basho* elements are very abundant in Arabidopsis, making up one-fourth of the overall TE content. *Basho* members in *A. thaliana* fall into seven groups and are characterized by a mononucleotide "T" TSD and a terminal sequence motif, 5'-CHH...CTAG-3' (H=A, T, or C; Le *et al.* 2000). Curiously, based on sequence and structural characteristics, *Basho* cannot be easily classified as either a class I or II element. We have identified three *Basho* elements forming one group in rice from the six rice clones surveyed. These elements are variable in size (1098-1975 bp) and do not have ORFs. Like the Arabidopsis elements, RESite analysis indicate that rice *Basho* elements have a target site preference for the mononucleotide "T" and a similar terminal sequence

signature (Figure 1, data not shown). The occurrence of these elements in rice and Arabidopsis strongly suggests that *Basho* may well be present in many plant genomes.

In addition to the rice *Basho* TEs, other difficult to categorize elements were mined in our survey. These include the previously identified elements referred to as *Crackle* and *Explorer*. The first described *Crackle* element was reported to have TIRs and a 6 bp TSD (Song *et al.* 1998). Despite being highly repetitive, we could not identify another *Crackle* element with this TIR structure. *Explorer* was also found to be repetitive, but the terminal ends could not be resolved and, as a result, no TSD could be determined (Bureau *et al.* 1996). Unfortunately, RESites that could have elucidated the identity of these elements could not be identified. In contrast, RESites corresponding to members from three other unusual element groups were found (Figure 1). The first of these unclassified group of elements (unclassified-I) is defined by a 9-11 bp TSD with no apparent TIR, LTR or poly-A+T structures and have terminal sequences that differ from non-TIR MULEs (Le *et al.* 2000; Yu *et al.* 2000). The second group (unclassified-II) has a 9 bp TSD and also lacks many of the structural features of other elements but has termini that are reminiscent of Arabidopsis non-TIR MULEs; however, no element in this group has yet been identified containing a *mudra*-like ORF (Yu *et al.* 2000). The element previously identified as *Micropon* (GI numbers 4586619, 4586620, 4586621, 4586622, and 4586623) forms the third group. Corresponding RESites are difficult to interpret but it appears that *Micropon* may have TIRs, TSDs, and/or preference for insertion into TA microsatellites (data not shown).

What are MITEs?

Recently, *Tourist*-like MITEs with ORFs related to prokaryotic IS (insertion sequences), have been identified in *A. thaliana* (Le *et al.* 2000). These elements were found based on shared sequence similarity between *Tourist*-like MITEs and longer representatives. Similarly, we have identified *Stowaway*-like MITEs with sequence similarity to a longer rice sequence. Specifically, most rice *Stowaway*-like elements share 70-100% nucleotide similarity with the termini of a previously reported rice TE (Tarchini *et al.* 2000). This

element is 5263 bp in length and was classified as a *mariner*-like element (MLE) (Tarchini *et al.* 2000) because its single ORF shares 47% amino-acid sequence identity (67% amino-acid sequence similarity) with *Soymar1*, a soybean MLE (Jarvik and Lark 1998). Strikingly, *Soymar1* also has TIRs similar to other *Stowaway*-like TEs. Taken together, these data strongly suggest that the *Stowaway* family of elements is evolutionarily related to elements belonging to the Tc1/*mariner* superfamily. Furthermore, *Stowaway* is more related to Tc1/*mariner* elements than to *Tourist*; in addition to an obvious relationship between ORFs, *Stowaway* and Tc1/*mariner* elements have a target site preference for the dinucleotide 5'-TA-3', whereas *Tourist*-like elements have a target site preference for 5'-TAA-3'. The eukaryotic class II elements making up the Tc1/*mariner* superfamily can be subdivided into three families, namely *mariner*, *pogo* and Tc1 (Plasterk *et al.* 1999). The TIRs of *Stowaway* elements, including *Soymar1* and the rice MLE, are very different from TIR sequences of members of the *pogo* and Tc1 families. Although members of the *mariner* family can have TIRs that also differ from *pogo* and Tc1 elements, the *Stowaway* TIRs are unique. Lastly, other putative ORFs, which share as much as 67% amino-acid sequence similarity with the rice MLE transposase, could be identified in rice genomic sequences, but TIRs could not be identified. These presumably represent near full-length elements that are degenerated copies or have suffered from deletion events.

Comparison with other genomes

A systematic survey of TEs has been recently performed on ~17.2 Mb of *A. thaliana* genomic sequence (Le *et al.* 2000). Class I elements (Ty1-*copia*, Ty3-*gypsy*, SINEs and LINEs), class II elements (MULEs, MITEs, *pogo*-type MLEs, *Ac*-like and CACTA-like), as well as *Basho* were mined. The transposon content is therefore qualitatively similar in the genomes of *A. thaliana* and rice. The abundance of individual groups of TEs is distinctive (Figure 5; Table 1). For example, *Basho* elements are very numerous in the Arabidopsis genome contributing over one-fourth of the elements mined in a recent survey (Le *et al.* 2000). In our survey, *Basho* elements account for <2% of the elements mined. MITEs are 26.4 times more abundant in rice than in *A. thaliana*. The latter harbors

one MITE for every 163.8 Kb of genomic DNA surveyed, which is considerably less than the estimated genomic density of one MITE for every 6.2 Kb in rice. Furthermore, MITE insertions account for 71.6% of the TEs found in rice, while the proportion of MITEs in *A. thaliana* is 16.7% (Le *et al.* 2000) (Figure 5). Also, the proportion of nucleotides contributed by rice TEs is 19.9%, which is much greater than the 4% (Meyerowitz 1994) and 5.3% (Le *et al.* 2000) previously reported for Arabidopsis.

A recent study of Arabidopsis TEs indicated that several elements (*Basho*, MITEs, MULEs, and MLEs) were preferentially found within A+T-rich regions of the genome (Le *et al.* 2000). Using the same approach (see Experimental Procedures), we determined that the G+C content of the regions flanking rice MITE insertions is significantly lower in G+C content than the average rice genomic G+C content (43%) in the sequence of the six clones surveyed. Beginning approximately 500 bp from the MITE insertion sites, there is a gradual decrease in G+C content to ≈35% immediately adjacent the MITEs. We also determined that *Stowaway*-like MITEs contribute more significantly than *Tourist*-like MITEs to this pattern and may reflect a distinction between these MITE types. With so few examples of intact members of the other TE groups, no correlation with A+T-rich regions could be determined. The patterns observed for MITE insertions within the rice and Arabidopsis genomes may indicate purifying selection against insertions within G+C or gene-rich regions or may reflect a bias for MITE insertions into the most A+T-rich regions of the genomes. Although this type of analysis has only been performed on Arabidopsis and rice elements, it is tempting to speculate that this phenomenon may be evolutionarily conserved.

Differences in the abundance of transposons within the *O. sativa* and *A. thaliana* genomes may reflect differential success of specific element types. The underlying mechanisms for success is not known but may reflect differences in the efficiency of repression mechanisms such as homology-dependant gene silencing and methylation, as well as differences in the level of ectopic recombination and gene conversion events (Kidwell and Lisch 2000). Co-adaptation between transposons and their host genomes, such as restriction of TE activity to specific tissues or insertion preferences for non-coding

regions (Kidwell and Lisch 1997) may also contribute to their differential success. In fact, some class II elements in maize preferentially insert near genes and not within retrotransposon-rich intergenic regions (Bennetzen *et al.* 1998; Cresse *et al.* 1995; Mao *et al.* 2000; Zhang *et al.* 2000). In addition, MITEs (Avramova *et al.* 1998; Tikhonov *et al.* 2000) and MULEs (Le *et al.* 2000) have been suggested to contribute sequences to matrix attachment regions (MARs), which are often located in introns and in the flanking regions of genes. In Arabidopsis, distribution analysis suggests a general negative correlation between gene and TE density and an enrichment of TEs in heterochromatic regions (Copenhaver *et al.* 1999; CSHL/WUGSC/PEB Arabidopsis Sequencing Consortium, 2000; Lin *et al.* 1999; Mayer *et al.* 1999). Unfortunately the sequence of contigs or large-insert clones from comparable regions is currently unavailable for rice. Nevertheless, understanding the diversity and distribution of elements within the rice, Arabidopsis and other plant genomes will provide the basis for further functional and regulatory characterization. Our report exemplifies that thorough TE mining is a key part of genomic cartography and, ultimately, will help determine the role of TEs in gene and genome evolution. TE mining will also provide information useful in the development of genome mapping, genome fingerprinting, gene isolation and gene functional analysis methodologies (Chang *et al.* 2000).

Acknowledgements

We would like to thank Dr. Louise O'Donoughue, Nabil Elrouby, Quang Hien Le, Stephen Wright, Zhihui Yu, and two anonymous reviewers for critical reading of the manuscript. We are also grateful to Ramm Hering and Aura Navarro-Quezada for their help in data mining and sequence analysis, as well as to Newton Agrawal, Boris-Antoine Legault and Christopher Olive for computer support. This work was supported by an FCAR (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche) fellowship to K.T., and an NSERC (Natural Sciences and Engineering Research Council of Canada) grant to T.B.

References

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

Avramova, Z., A. Tikhonov, M. Chen and J. L. Bennetzen, 1998 Matrix attachment regions and structural colinearity in the genomes of two grass species. Nucl. Acids Res. 26: 761-767.

Bennetzen, J. L., 1998 The evolution of grass genome organization and function. Symp. Soc. Exp. Biol. 51: 123-126.

Bennetzen, J. L., P. SanMiguel, M. Chen, A. Tikhonov, M. Francki and Z. Avramova, 1998 Grass genomes. Proc. Natl. Acad. Sci. USA 95: 1975-1978.

Bennetzen, J. L., and P. S. Springer, 1994 The generation of *Mutator* transposable element subfamilies in maize. Theor. Appl. Genet. 87: 657-667.

Bureau, T. E., P. C. Ronald and S. R. Wessler, 1996 A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc. Natl. Acad. Sci. USA 93: 8524-8529.

Bureau, T. E., and S. R. Wessler, 1992 *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4: 1283-1294.

Bureau, T. E., and S. R. Wessler, 1994a Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. USA 91: 1411-1415.

Bureau, T. E., and S. R. Wessler, 1994b *Stowaway*, a new family of inverted repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6: 907-916.

Chandler, V. L., and K. Hardeman, 1992 The *Mu* elements of *Zea mays*. Adv. Genet. 30: 77-122.

Chang, R.-Y., L. S. O'Donoughue and T. E. Bureau, 2000 Inter-MITE polymorphisms (IMP): a high throughput transposon-based genome mapping and fingerprinting approach. Theor. Appl. Genet. 102: 773-781.

Charlesworth, B., P. Sniegowski and W. Stephan, 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371: 215-220.

Copenhaver, G. P., K. Nickel, T. Kuromori, M.-I. Benito, S. Kaul, X. Lin, M. Bevan, G. Murphy, B. Harris, L. D. Parnell, *et al.*, 1999 Genetic definition and sequence analysis of Arabidopsis centromeres. Science 286: 2468-2474.

Cresse, A. D., S. H. Hulbert, W. E. Brown, J. R. Lucas and J. L. Bennetzen, 1995 *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. Genetics 140: 315-324.

CSHL/ WUGSC/ PEB Arabidopsis Sequencing Consortium, 2000 The complete sequence of a heterochromatic island from a higher eukaryote. Cell 100: 377-386.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell and A. Chovnick, 1990 Evidence for horizontal transmission of the *P* transposable element between Drosophila species. Genetics 124: 339-355.

Dong, F., J. T. Miller, S. A. Jackson, G.-L. Wang, P. C. Ronald and J. Jiang, 1998 Rice (*Oryza sativa*) centromeric regions consist of complex DNA. Proc. Natl. Acad. Sci. USA 95: 8135-8140.

Eisen, J. A., M. I. Benito and V. Walbot, 1994 Sequence similarity of putative transposases links the maize *Mutator* autonomous elements and a group of bacterial insertion sequences. Nucl. Acids Res. 22: 2634-2636.

Feuillet, C., and B. Keller, 1999 High gene density is conserved at syntenic loci of small and large grass genomes. Proc. Natl. Acad. Sci. USA 96: 8265-8270.

Fedoroff, N. V., 1999 The *suppressor-mutator* element and the evolutionary riddle of transposons. Genes Cells 4: 11-19.

Finnegan, D. J., 1992 Transposable elements. Curr. Opin. Genet. Devel. 2: 861-867.

Flavell, A. J., S. R. Pearce and A. Kumar, 1994 Plant transposable elements and the genome. Curr. Opin. Genet. Devel. 4: 838-844.

Grandbastien, M.-A., 1992 Retroelements in higher plants. Trends Genet. 8: 103-108.

Hershberger, R. J., M. I. Benito, K. Hardeman, C. Warren, V. L. Chandler and V. Walbot, 1995 Characterization of the major transcripts encoded by the regulatory *MuDR* transposable element of maize. Genetics 140: 1087-1098.

Hirano, H.-Y., K. Mochizuki, M. Umeda, H. Ohtsubo, E. Ohtsubo and Y. Sano, 1994 Retrotransposition of a plant SINE into the *wx* locus during evolution of rice. J. Mol. Evol. 38: 132-137.

Iida, S., J. Meyer and W. Arber, 1983 Prokaryotic IS elements. In Mobile genetic elements. Shapiro, J.A., Ed.. New York: Academic Press, pp. 159-221.

Jarvik, T., and K. G. Lark, 1998 Characterization of *Soymar1*, a *Mariner* element in soybean. Genetics 149: 1569-1574.

Jordan, I. K., L. V. Matyunina and J. F. McDonald, 1999 Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. Proc. Natl. Acad. Sci. USA 96: 12621-12625.

Kidwell, M. G., and D. R. Lisch. 1997 Transposable elements as sources of variation in animals and plants. Proc. Natl Acad. Sci. USA 94: 7704-7711

Kidwell, M. G., and D. R. Lisch, 2000 Transposable elements and host genome evolution. Trends Ecol. Evol. 15: 95-99.

Kubis, S., T. Schmidt and J. S. Heslop-Harrison, 1998 Repetitive DNA elements as a major component of plant genomes. Ann. Botany 82: 45-55.

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

Leepton, P. J., and D. R. Smyth, 1993 An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Mol. Gen. Genet. 237: 97-104.

Lin, X., S. Kaul, S. Rounsley, T. P. Shea, M.-I. Benito, C. D. Town, C. Y. Fujii, T. Mason, C. L. Bowman, M. Barnstead *et al.*, 1999 Sequence analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402: 769-777.

Lisch, D., L. Girard, M. Donlin and M. Freeling, 1999 Functional analysis of deletion derivatives of the maize transposon *MuDR* delineates roles for the MURA and MURB proteins. Genetics 151: 331-341.

Mao, L., T. C. Wood, Y. Yu, M. A. Budiman, J. Tomkins, S.-S. Woo, M. Sasinowsk, G. Presting, D. Frisch, D. Goff, R. A. Dean and R. A. Wing, 2000 Rice transposable elements: a survey of 73,000 sequence-tagged connectors. Genome Res. 10: 982-990.

Masson, P., G. Rutherford, J. Banks and N. Federoff, 1989 Essential large transcripts of the maize *Spm* transposable element are generated by alternative splicing. Cell 58: 755-765.

Mayer, K., C. Schuller, R. Wambutt, G. Murphy, G. Volckaert, T. Pohl, A. Dusterhoft, W. Stiekema, K.-D. Entian, N. Terryn, *et al.*, 1999 Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature 402: 769-777.

McClintock, B., 1951 Chromosomal organization and genic expression. Cold Spring Harbor Symp. Quant. Biol. 16: 13-47.

McClintock, B., 1984 The significance of responses of the genome to challenges. Science 226: 792-801.

Messing, J., and V. Llaca, 1998 Importance of anchor genomes for any plant genome project. Proc. Natl. Acad. Sci. USA 95: 2017-2020.

Meyerowitz, E. M., 1994 Structure and organization of the *Arabidopsis thaliana* nuclear genome. In Arabidopsis. Meyerowitz, E.M. and Somerville, C. Eds. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. pp. 21-36.

Mochizuki, K., H. Ohtsubo, H.-I. Hirano, Y. Sano and E. Ohtsubo. 1993 Classification and relationships of rice strains with AA genome by identification of transposable elements at nine loci. Jap. J. Genet. 68: 205-217.

Mochizuki, K., M. Umeda. H. Ohtsubo and E. Ohtsubo, 1992 Characterization of plant SINE, p-SINE1, in rice genomes. Jap. J. Genet. 57: 155-166.

Motohashi, R., E. Ohtsubo and H. Ohtsubo, 1996 Identification of Tnr3, a *Supressor-Mutator/Enhancer*-like transposable element from rice. Mol. Gen. Genet. 150: 148-152.

Ohtsubo, E. and Y. Sekine, 1996 Bacterial insertion sequences. In Transposable elements. Saedler H. and Gierld A., Eds. Berlin, Springer-Verlag, 1-26.

Okada, N., M. Hamada, I. Ogiwara and K. Ohshima, 1997 SINEs and LINEs share common 3' sequences: a review. Gene 205: 229-243.

Plasterk, R. H. A., Z. Izsvák and Z. Ivics, 1999 Resident aliens: the Tc1/*mariner* superfamily of transposable elements. Trends Genet. 15: 326-332.

Presting, G. G., L. Malysheva, J. Fuchs and I. Schubert, 1998 A *TY3/GYPSY* retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. Plant J. 16: 721-728.

Raina, R., M. Schlappi, B. Karunanandaa, A. Elhofy and N. Fedoroff, 1998 Concerted formation of macromolecular *Suppressor-mutator* transposition complexes. Proc. Natl. Acad. Sci USA 95: 8526-8531.

Robertson, D. S., 1978 Characterization of a *Mutator* system in maize. Mutat. Res. 51: 21-28.

Robertson, H. M., 1993 The *Mariner* transposable element is widespread in insects. Nature 362: 241-245.

Scortecci, K. C., R. Raina, N. V. Fedoroff and M. A. Van Sluys, 1999 Negative effect of the 5'-untranslated leader sequence on *Ac* transposon promoter expression. Plant Mol. Biol.. 40: 935-944.

Song, W.-Y., L.-Y. Pi, T. E. Bureau and P. C. Ronald, 1998 Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the *Xa21* family of disease resistance genes in rice. Mol. Gen. Genet. 258: 449-456.

Talbert, L. E., and V. L. Chandler, 1988 Characterization of a highly conserved sequence related to *Mutator* transposable element in maize. Mol. Biol. Evol. 5: 519-529.

Talbert, L. E., G. I. Patterson and V. L. Chandler. 1989 *Mu* transposable elements are structurally diverse and distributed throughout the genus *Zea*. J. Mol. Evol. 29: 28-39.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey and A. Rafalski, 2000 The complete sequence of 340 Kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell 12: 381-391.

Tatusova, T. A., and T. L. Madden, 1999 BLAST 2 sequences - a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174: 247-250.

Tenzen, T., Y. Matsuda, H. Ohtsubo and E. Ohtsubo, 1994 Transposition of Tnr1 in rice genomes to 5'-PuTAPy-3' sites, duplicating the TA sequence. Mol. Gen. Genet. 245: 441-448.

Tikhonov, A. P., J. L. Bennetzen and Z. V. Avramova, 2000 Structural domains and matrix attachment regions along colinear chromosomal segments of maize and Sorghum. Plant Cell 12: 249-264.

Umeda, M., H. Ohtsubo and E. Ohtsubo, 1991 Diversification of the rice *Waxy* gene by insertion of mobile elements into introns. Jap. J. Genet. 66: 569-586.

Uozu, S., H. Ikehashi, N. Ohmido, H. Ohtsubo, E. Ohtsubo and K. Fukui, 1997 Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. Plant Mol. Biol. 35:791-799.

Wang, S., N. Liu, K. Peng and Q. Zhang, 1999 The distribution and copy number of *copia*-like retrotransposons in rice (*Oryza sativa* L.) and their implications in the organization and evolution of the rice genome. Proc. Natl. Acad. Sci. USA 96: 6824-6828.

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

Yoshida, S., K. Tamaki, K. Watanabe, M. Fujino and C. Nakamura, 1998 A maize *MuDR*-like element expressed in rice callus subcultured with proline. Hereditas 129: 95-99.

Yu, Z., S. I. Wright. and T. E. Bureau, 2000 *Mutator*-like elements (MULEs) in *Arabidopsis thaliana*: structure. diversity and evolution. Genetics 156: 2019-2031.

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

Zhang, Q. and G. Kochert, 1998 Independent amplification of two classes of *Tourist* in some *Oryza* species. Genetica 101: 145-152.

Table 1. Occurrence of TEs in 910 Kb of *Oryza sativa* L. genomic sequences

	no. of groups	no. of TEs	sequence contribution [no. of nucleotides (%)]	contribution to TE content (%)
Class I				
Ty1-copia	2	2	23,991 (2.6)	1.0
Ty3-gypsy	5	10	65,994 (7.3)	4.9
LINEs	1	5	18,997 (2.1)	2.5
SINEs	1	4	843 (0.1)	2.0
Undetermined [a]	1	1	1,104 (0.1)	0.5
SUBTOTAL:	10	22	110,929 (12.2)	10.8
Class II				
Ac-like	2	2	1,158 (0.1)	1.0
CACTA	1	1	1,017 (0.1)	0.5
MULEs	10	17	22,458 (2.5)	8.3
Stowaway-like MITEs	24	70	16,112 (1.8)	34.3
Tourist-like MITEs	16	76	19,487 (2.1)	37.3
SUBTOTAL:	53	166	60,232 (6.6)	81.4
Other				
Basho	1	3	5,016 (0.6)	1.5
Crackle	1	1	240 (0.1)	0.5
Explorer	1	5	931 (0.1)	2.5
Micropon-like	1	2	530 (0.1)	1.0
unclassified	2	5	3,592 (0.4)	2.5
TOTAL	69	204	181,470 (19.9)	100

a, could not determine if this solo LTR belongs to Ty1-*copia* or Ty3-*gypsy*.

Figure 1. RESites (see Experimental Procedures) corresponding to sixteen of the mined rice TEs. The pairwise comparison of the RESite and the TE-bearing sequence may reveal a gap corresponding to the position of the TE and the TSD. The TSD is underlined and the TIRs are represented by arrows. GenBank GI (geninfo) numbers and nucleotide positions are indicated.

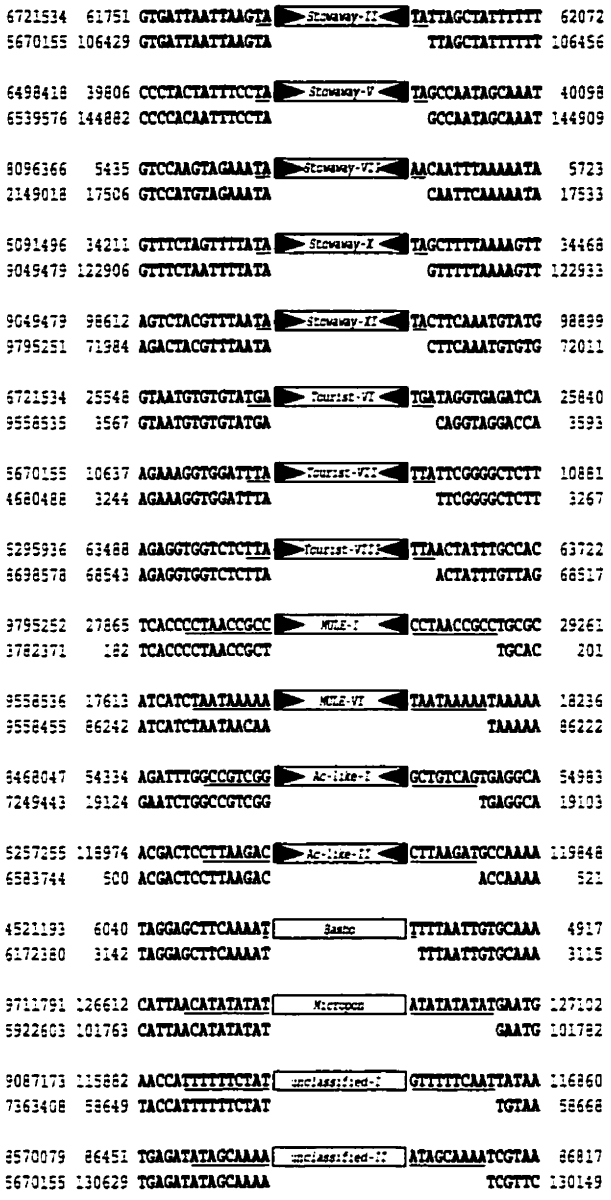

6721534 61751 GTGATTAATTAAGTA Stowaway-II TATTAGCTATTTTTT 62072
5670155 106429 GTGATTAATTAAGTA TTAGCTATTTTTT 106456
6498418 39806 CCCTACTATTTCCTA Stowaway-V TAGCCAATAGCAAAT 40098
6539576 144882 CCCCACAATTTCCTA GCCAATAGCAAAT 144909
8096366 5435 GTCCAAGTAGAAATA Stowaway-VII AACAATTTAAAAATA 5723
2149018 17506 GTCCATGTAGAAATA CAATTCAAAAATA 17533
5091496 34211 GTTTCTAGTTTTATA Stowaway-X TAGCTTTTAAAAGTT 34468
3049479 122906 GTTTCTAATTTTATA GTTTTTAAAAGTT 122933
9049479 98612 AGTCTACGTTTAATA Stowaway-XI TACTTCAAATGTATG 98899
9795251 71984 AGACTACGTTTAATA CTTCAAATGTGTG 72011
6721534 25548 GTAATGTGTGTATGA Tourist-VI TGATAGGTGAGATCA 25840
9558535 3567 GTAATGTGTGTATGA CAGGTAGGACCA 3593
5670155 10637 AGAAAGGTGGATTTA Tourist-VII TTATTCGGGGCTCTT 10881
4680488 3244 AGAAAGGTGGATTTA TTCGGGGCTCTT 3267
5295936 63488 AGAGGTGGTCTCTTA Tourist-VIII TTAACTATTTGCCAC 63722
8698578 68543 AGAGGTGGTCTCTTA ACTATTTGTTAG 68517
9795252 27865 TCACCCCTAACCGCC MULE-I CCTAACCGCCTGCGC 29261
3782371 182 TCACCCCTAACCGCT TGCAC 201
9558536 17613 ATCATCTAATAAAAA MULE-VI TAATAAAAATAAAAA 18236
9558455 86242 ATCATCTAATAACAA TAAAAA 86222
8468047 54334 AGATTTGGCCGTCGG Ac-like-I GCTGTCAGTGAGGCA 54983
7249443 19124 GAATCTGGCCGTCGG TGAGGCA 19103
5257255 118974 ACGACTCCTTAAGAC Ac-like-II CTTAAGATGCCAAAA 119648
6583744 500 ACGACTCCTTAAGAC ACCAAAA 521
4521193 6040 TAGGAGCTTCAAAAT Basho TTTTAATTGTGCAAA 4917
6172380 3142 TAGGAGCTTCAAAAT TTTAATTGTGCAAA 3115
9711791 126612 CATTAACATATATAT Micropon ATATATATATGAATG 127102
5922603 101763 CATTAACATATATAT GAATG 101782
9087173 115882 AACCATTTTTTCTAT unclassified-I GTTTTTCAATTATAA 116860
7363408 58649 TACCATTTTTTCTAT TGTAA 58668
8570079 86451 TGAGATATAGCAAAA unclassified-II ATAGCAAAATCGTAA 86817
5670155 130629 TGAGATATAGCAAAA TCGTTC 130149

Figure 2. Acquisition of truncated host cellular genes by members of MULE-I and MULE-IV. (a) and (b) Diagram and nucleotide sequence alignment of the rice 5S rDNA gene (GI:20162, position 58-162) and the MULE-I elements (GI:6006355, position 26919-27023). Five MULE-I representatives contain a 105 bp region with high nucleotide similarity (>90%) to a portion of the rice 5S rDNA gene. The black boxes represent the acquired fragment of the 5S rDNA gene, the boxes with dots represent the remaining internal sequence of the element, the boxes with an arrow represent the TIRs (common to all MULE-I elements), and the gray boxes represent the internal sequence of two representatives having no internal sequence similarity to other MULE-I members. The element on GI:8096366 contains an ORF encoding a degenerate mudrA-like gene. Conserved nucleotides are shaded in black in the alignment. (c) and (d) Diagram and nucleotide sequence alignment of the rice mRNA for PCF2 (GI:2580439, position 492-334) and the MULE-IV element (GI:5042437, position 21095-21253), which contain a 149 bp region showing high nucleotide similarity (>83%) to a region of the rice PCF2 mRNA sequence. The black boxes represent the acquired fragment corresponding to the PCF2 mRNA, the boxes with dots represent the remaining internal sequence of the element, the boxes with an arrow represent the TIRs. Conserved nucleotides are shaded in black.

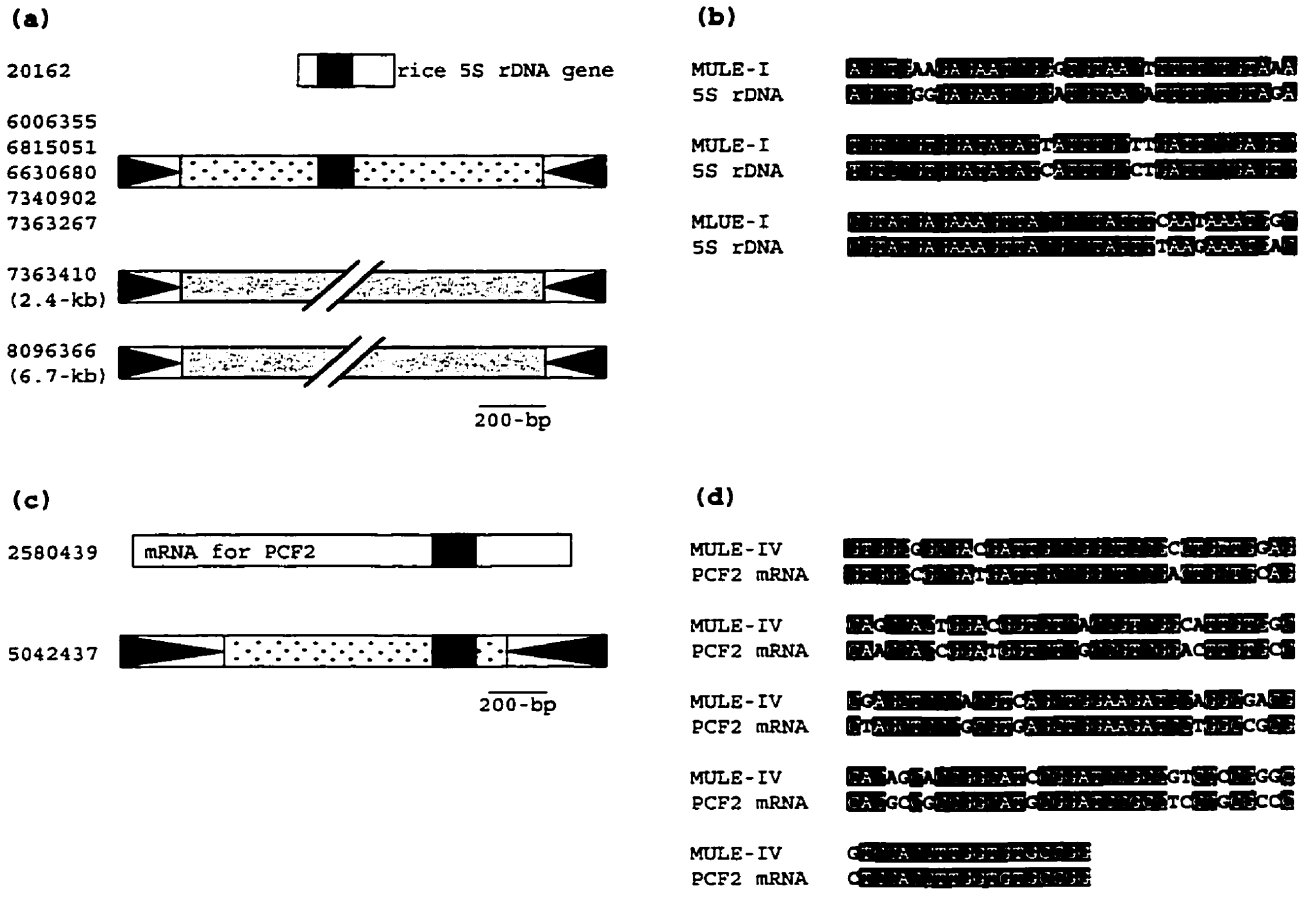

(a)
20162
rice 5S rDNA gene
6006355
6815051
6630680
7340902
7363267
7363410
(2.4-kb)
8096366
(6.7-kb)
200-bp
(b)
MULE-I
5S rDNA
MULE-I
5S rDNA
MLUE-I
5S rDNA
(c)
2580439
mRNA for PCF2
5042437
200-bp
(d)
MULE-IV
PCF2 mRNA
MULE-IV
PCF2 mRNA
MULE-IV
PCF2 mRNA
MULE-IV
PCF2 mRNA
MULE-IV
PCF2 mRNA

Figure 3. Similarity between rice and Sorghum *Tourist*-I elements. The alignment of the consensus nucleotide sequences derived from *Sorghum bicolor Tourist* elements and seven *Tourist*-I elements from *Oryza sativa*. The conserved nucleotides are shaded in black, the partially conserved nucleotides are shaded in gray, and dashes represent gaps introduced to maintain an optimal alignment.

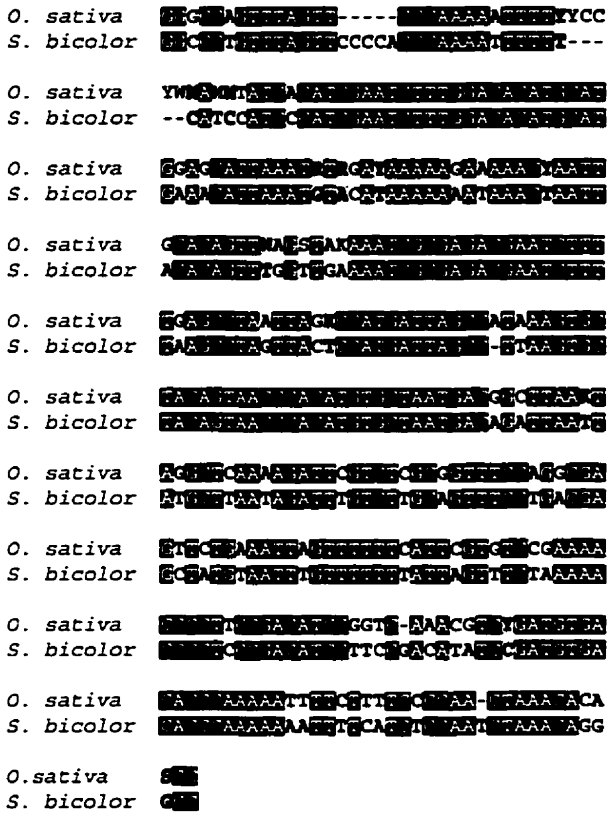

O. sativa
S. bicolor
O. sativa
S. bicolor
O. sativa
S. bicolor
O. sativa
S. bicolor
O. sativa
S. bicolor
O. sativa
S. bicolor
O. sativa
S. bicolor
O. sativa
S. bicolor
O.sativa
S. bicolor

Figure 4. Similarity between TIRs and TSD of rice *Stowaway*-like and Tc1/*mariner*-like elements. *Soymar1* is a soybean Tc1/*mariner* element (Jarvik, 1998; GI:3386533). The three terminal-most nucleotides are deduced from a smaller copy of this element (Jarvik 1998). The element on the genomic clone, GI:6979318 (position 227,524-232,786), is annotated as a Tc1/*mariner*-like transposon (Tarchini *et al.* 2000). The TSD is underlined, nucleotides sharing over 60% similarity are shaded in black, and the number in parentheses corresponds to the nucleotides not shown.


Stowaway-I
Stowaway-II
Stowaway-III
Stowaway-IV
Stowaway-V
Stowaway-VI
Stowaway-VII
Stowaway-VIII
Stowaway-IX
Stowaway-X
GI:6979318
Soymar1

Figure 5. Comparison of the TE content between *Arabidopsis thaliana* and *Oryza sativa*. The values on the x axis indicate the proportion of mined TEs corresponding to each type of TEs from the y axis. No *pogo*-like MLEs were mined in rice. SINE: short interspersed nuclear element; LINE: long interspersed nuclear element; MLE: *mariner*-like element; MULE: *Mutator*-like element; MITE: miniature inverted-repeat transposable element.

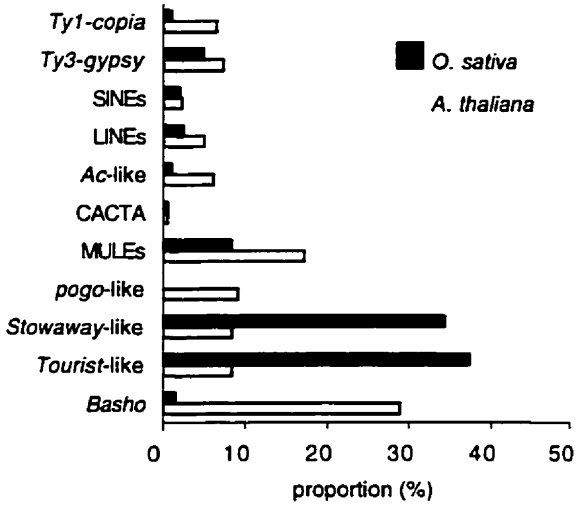
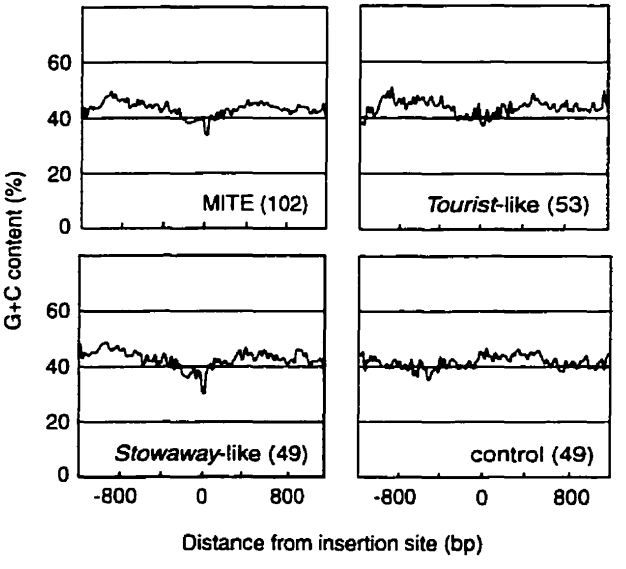

Ty1-copia
Ty3-gypsy
SINEs
LINEs
Ac-like
CACTA
MULEs
pogo-like
Stowaway-like
Tourist-like
Basho
O. sativa
A. thaliana
0
10
20
30
40
50
proportion (%)

Figure 6. G+C content of the regions flanking the insertion site of various types of mined TEs. The number in parentheses correspond to the number of TEs used in the calculations. TE types with <20 non-truncated members could not be analyzed.


G+C content (%)
60
40
20
0
MITE (102)
Tourist-like (53)
Stowaway-like (49)
control (49)
-800
0
800
Distance from insertion site (bp)

CHAPTER 2

Tc8, a *Tourist*-like transposon in *Caenorhabditis elegans*

Title: **Tc8, a *Tourist*-like Transposon in *Caenorhabditis elegans***

Authors: **Quang Hien Le, Kime Turcotte and Thomas Bureau**

Address: Department of Biology, McGill University, Montreal, Quebec, Canada, H3A 1B1.

Running head: *Tourist*-like transposons in *C. elegans*

Key words: *Tourist*, Tc8, transposon, MITE, *Caenorhabditis elegans*,

Corresponding author: Dr. Thomas Bureau, Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montreal, Quebec, Canada H3A 1B1
Tel.: (514) 398-6472; Fax.: (514) 398-5069.
E-mail: thomas_bureau@maclan.mcgill.ca

Abstract

Members of the *Tourist* family of *m*iniature *i*nverted-repeat *t*ransposable *e*lements (MITEs) are very abundant among a wide variety of plants, are frequently found associated with normal plant genes and thus, are thought to be important players in the organization and evolution of plant genomes. In Arabidopsis, the recent discovery of a *Tourist* member harboring a putative transposase has shed new light on the mobility and evolution of MITEs. Here, we analyze a family of *Tourist* transposons endogenous to the genome of the nematode *Caenorhabditis elegans* (Bristol N2). One member of this large family is 7568 bp in length, harbors an ORF similar to the putative *Tourist* transposase from Arabidopsis and is related to the IS*5* family of bacterial *i*nsertion *s*equences (IS). Using database searches, we found *e*xpressed *s*equence *t*ags (ESTs) similar to the putative *Tourist* transposases in plants, insects and vertebrates. Taken together, our data suggest that *Tourist*-like and IS*5*-like transposons form a superfamily of potentially active elements ubiquitous to prokaryotic and eukaryotic genomes.

Introduction

Transposons are mobile genetic elements found in most, if not all, prokaryotic and eukaryotic genomes. Typically, transposons are defined as either class I, members of which move *via* an RNA intermediate (*e.g.* retrotransposons), or class II, members of which move directly as DNA (*e.g.* Tc1/*mariner*, *Ac/Ds*, *En/Spm*) (Berg and Howe 1989). MITEs are transposons found in abundance among a wide variety of plant genomes. They are typically short (~100 to 500 bp), with conserved *t*erminal *i*nverted *r*epeats (TIRs), have a potential to form a stable DNA secondary structure (*i.e.* a hairpin structure), and generate a 2 or 3 bp *t*arget *s*ite *d*uplication (TSD) upon integration. Based on their TSD size and sequence, MITEs can primarily be divided into *Tourist*-like (5'-TAA-3') and *Stowaway*-like (5'-TA-3') families of elements (Bureau and Wessler 1992). MITEs are usually found in intimate association with genes and occasionally contributing *cis*-regulatory sequences (Bureau and Wessler 1994a; Bureau and Wessler 1994b). For these reasons, they are thought to play an important role in the evolution of plant genomes (Wessler *et al.* 1995). Although ubiquitous in plants, there are only few reported MITEs from other eukaryotes such as fungi, insects, *C. elegans*, *Xenopus*, teleost fish, and humans (Besansky *et al.* 1996; Feschotte and Mouches 2000; Izsvak *et al.* 1999; Oosumi *et al.* 1995; Smit and Riggs 1996; Tu 1997; Unsal and Morgan 1995; Yeadon and Catcheside 1995). However, many of these MITEs are structurally similar to Tc1/*mariner*-like elements though members with ORFs are typically not available. The reason behind this difference in MITE abundance between plants and other organisms, and how this difference impacts host genome evolution, is not clear.

Since MITEs with coding capacity were previously unknown, the mechanism underlying their transposition remained elusive. Structurally, MITEs are reminiscent of non-autonomous deletion derivatives of class II transposons, presumably mobilized *in trans* by a transposase from a related autonomous element located elsewhere in the genome. Recently however, ORFs coding for putative *Tourist* transposases from *Arabidopsis thaliana* (Columbia) and *Stowaway* transposases from Arabidopsis and *Oryza sativa*

(domesticated rice) have been found, clearly defining MITEs as class II transposons (Le *et al.* 2000; Turcotte *et al.* 2001).

Further analysis of the putative Arabidopsis *Tourist* transposases indicates that *Tourist* elements are more related to specific bacterial IS, a large and heterogeneous group of simple inverted-repeat transposons widespread among prokaryotes (Le *et al.* 2000; Mahillon and Chandler 1998), than they are to *Stowaway*. Detailed analysis of the putative transposases encoded by *Stowaway* suggests that these elements are related to members of the Tc1/*mariner* transposons (Le *et al.* 2000), Turcotte and Bureau, manuscript in preparation), a widespread superfamily of elements with representatives in vertebrates, invertebrates and fungi. Phylogenetic studies of transposases and integrases revealed that Tc1/*mariner* elements, members of diverse bacterial IS elements, retroviruses and retrotransposons share conserved catalytic residues (Doak *et al.* 1994), known as the DDE motif, suggesting a common ancestry and even more widespread distribution.

Traditionally, transposons were identified and analyzed through genetic and molecular studies. However, the availability of sequence information and bioinformatic tools has now allowed for the identification and characterization of transposons through computer-assisted searches of sequence databases (Britten 1995; Bureau and Wessler 1992; Bureau and Wessler 1994b; Le *et al.* 2000; Oosumi *et al.* 1995; Surzycki and Belknap 1999; Surzycki and Belknap 2000). In *C. elegans*, molecular, genetic and sequence database search tools have revealed the presence of diverse representatives of both classes of transposons. Among these, Tc1 is probably one of the best characterized eukaryotic class II transposon and has been effectively used as a mapping, mutagenesis and gene-tagging tool (Greenwald 1985; Plasterk and van Luenen 1997; Rushforth *et al.* 1993; Williams *et al.* 1992). Tc1 transposition and regulation has been finely dissected at the molecular and biochemical levels (Plasterk and van Luenen 1997). Most Tc1 elements have a 54 bp TIR delimited by a 5'-TA-3' TSD. Autonomous Tc1 encodes a 343 amino acid transposase, which is the only protein required for transposition (Plasterk and van Luenen 1997; Vos *et al.* 1993; Vos *et al.* 1996). Tc1 copy number can vary from 10-15

fold depending on the *C. elegans* strain and element activity is cell-type dependent (Egilmez *et al.* 1995). In Bergerac, Tc1 excision is detected in both somatic and germline cells whereas it is restricted to somatic cells in Bristol N2 (Egilmez *et al.* 1995; Eide and Anderson 1988; Plasterk and van Luenen 1997; Vos *et al.* 1993; Vos *et al.* 1996). Recently, *mut-7* was found to encode a RNaseD homolog in *C. elegans* and may be involved in the regulation of transposon activity by post-transcriptional gene silencing (Ketting *et al.* 1999).

Tc1, Tc2, Tc3, and Tc7 are members of the Tc1/*mariner* superfamily found in *C. elegans* (Collins *et al.* 1989; Dreyfus and Emmons 1991; Levitt and Emmons 1989; Rezsohazy *et al.* 1997). In addition, other types of transposons endogenous to *C. elegans* include LTR (*l*ong *t*erminal *r*epeat)-retrotransposons (Bowen and McDonald 1999; Britten 1995), non-LTR retrotransposons (or long interspersed nuclear elements) (Malik and Eickbush 1998) and hAT (or *Ac*)-like elements (Bigot *et al.* 1996). Elements designated as Tc4 and Tc5 were initially classified as foldback transposons (Collins and Anderson 1994; Yuan *et al.* 1991) but further analysis of their putative transposases revealed that they contain a DDE motif suggesting a distant relationship to the Tc1/*mariner* elements *pogo*, *Tigger1* and *Tigger2* (Robertson 1996; Smit and Riggs 1996). Tc6 also has a foldback structure but its terminal sequences and TSDs are reminiscent of those of Tc1/*mariner* transposons (Dreyfus and Emmons 1991). A number of MITEs have also been identified in *C. elegans* (Oosumi *et al.* 1995; Oosumi *et al.* 1996; Surzycki and Belknap 2000; The *C. elegans* Sequencing Consortium 1998) but these were not found to be related to *Tourist*. To date, *Tourist* transposons appear to be confined to plant genomes.

In this report, we analyze two predicted ORFs in *C. elegans* that share amino acid similarity to the putative Arabidopsis *Tourist* and bacterial IS transposases. The same ORFs were also identified independently by sequence similarity to an Arabidopsis element called *Harbinger* but no further attempts were made to characterize these putative *C. elegans* elements (Kapitonov and Jurka 1999). We determine that one of the *C. elegans* ORFs is part of a *Tourist*-like element which belongs to a large group of nematode transposons, referred to as Tc8. The majority of Tc8 members are short (150 to

400 bp), can potentially form DNA-hairpins and have TIRs and TSDs similar to other *Tourist*-like elements. The Tc8 transposase shares similarity to a number of plant, insect and vertebrate EST sequences, indicating that the *Tourist* family of transposons is potentially active in other organisms. In addition, several eukaryotic genomic sequences were identified with similarity to the Arabidopsis *Tourist* and *C. elegans* Tc8 transposases. Together our data confirm the presence of *Tourist*-like transposons beyond the plant kingdom and suggest that *Tourist*-like elements share a common ancestry with specific bacterial IS and are widely distributed in the prokaryote and eukaryote genomes.

Materials and Methods

Transposon mining

All sequence information and BLAST search tools (Altschul *et al.* 1990; Altschul *et al.* 1997) were accessed through GenBank (http://www.ncbi.nlm.nih.gov/), the Genome Sequencing Center (http://www.genome.wustl.edu/gsc/index.shtml) or the Sanger Centre (http://www.sanger.ac.uk/ Projects/C_elegans/). Tc8.1 was identified by using sequences corresponding to putative MITE transposases from *Arabidopsis* (Le *et al.* 2000) as queries in BLAST searches against the GenBank sequence database. Additional database searches were performed to identify other *C. elegans* elements related to Tc8.1 (BLAST score of > 80).

Identification of RESites

*R*elated *e*mpty *site*s (RESites) are sequences highly similar or nearly identical to the sequences flanking an insertion (Le *et al.* 2000). However, RESites do not harbor the transposon sequence and the target site of insertion is not duplicated. RESites may correspond to orthologous or paralogous sequences, serve to delimit the ends of the elements, determine the TSDs and suggest evidence of past mobility. The identification of RESites was performed as previously described (Le *et al.* 2000). In brief, RESites are identified using the region immediately flanking an insertion as queries in BLAST searches.

Data and phylogenetic analysis

Further sequence analysis and alignments were performed using TRANSLATE, PILEUP, BESTFIT and GAP as part of the University of Wisconsin Genetics Computing Group suite of programs (version 10.0) or additional BLAST search tools (BLASTP, BLASTX, TBLASTN, PSI-BLAST and BLAST 2 sequences) provided at NCBI (Altschul *et al.* 1997; Tatusova and Madden 1999). For phylogenetic analysis, the sequences surrounding the DDE motif of the transposase of 14 transposable elements were manually aligned, based on previous reports (Capy *et al.* 1996; Rezsohazy *et al.* 1993). Phylogenetic relationships were inferred with maximum parsimony (multiple trees heuristic search with Tree Bisection

Reconnection) and neighbor joining methods from the PAUP program, version 4.0b4a (Swofford 2000) from the amino acid sequence alignment shown in Figure 3. All amino acids and stop codons were equally weighted. Unrooted trees were displayed using the midpoint rooting method. Bootstrap analysis of the data set was done with 200 replicates.

Results and Discussion

Computer-based searches using the amino acid sequence of the putative *Tourist*-like transposases from Arabidopsis MITE X (conceptual translation of gi 4454587, positions 4650-5865) and MITE XI (gi 4585884) as queries revealed the presence of similar coding regions in several other plant, animal and eubacterial ESTs and/or genomic sequences. For the most part, no other full-length plant *Tourist*-like element could be resolved and likely reflects the presence of truncated or degenerate elements. Many of the eubacterial sequences, however, correspond to previously described mobile elements. Although, as in the case of plants, the animal sequences typically did not lead to the identification of full-length elements, a 7568 bp element was mined from the genomic sequence of *C. elegans*. This element, designated as Tc8.1, is located on chromosome 2 (clone CELF14D2, gi 2746790), has large imperfect TIRs, a putative 3 bp TSD (5'-TAA-3') and a predicted ORF coding for a hypothetical protein of 743 amino acids (gi 7499082). Overall, the presence of sequences with similarity to the *Tourist*-like element transposases in several eubacterial and eukaryotic genomes suggests a wide distribution.

Using the Tc8.1 nucleotide sequence as a query in computer-assisted database searches, we found 282 related sequences within the *C. elegans* genome. Of these sequences, 128 represent shorter versions of Tc8.1 complete with TIRs and a specific TSD sequence of 5'-TAA-3'. The remainder were truncated versions with only one discernable terminal end immediately flanked by the sequence 5'-TAA-3'. Together, Tc8.1 and the shorter intact and truncated elements comprise the Tc8 group of transposons within *C. elegans*. Even though Tc8 elements are highly abundant in *C. elegans*, the total nucleotide contribution represents only 0.05% of the genome.

Strikingly, 63 members (≈49%) of the Tc8 group are exactly 150 bp in length and share > 94% sequence similarity. Like plant *Tourist*-like members, these elements have the potential to form hairpin-like DNA secondary structures. Excluding Tc8.1, six members of the Tc8 group were > 1-kb in size but did not possess any coding capacity (Figure 1A). For four of these larger elements, the increase in size is the result of nested insertions of

other types of transposons or repetitive elements. Evidence of past mobility for some members of Tc8 was provided through the identification of *r*elated *e*mpty *site*s (RESites, Figure 2), which are sequences similar to the empty site of an insertion. Furthermore, BLASTN searches also revealed the presence of Tc8 within the genome of the related nematode *C. briggsae* (Figure 1B), which is thought to have diverged from *C. elegans* approximately 10-40 million years ago (Butler *et al.* 1981; Heschl and Baillie 1990; Kennedy *et al.* 1993). To date, eight complete and three truncated Tc8 elements were identified from *C. briggsae* genomic clones deposited in Genbank. Tc8 transposons in *C. briggsae* vary from 95 to 368 bp in length and are not as well conserved as are the Tc8 elements in *C. elegans* (data not shown). A complete list of the *C. elegans* and *C. briggsae* transposons mined in this survey is available at http://www.tebureau.mcgill.ca/. A second predicted ORF (gi 7504556) was identified in the *C. elegans* genome and shared 46.5% amino acid sequence similarity to the Arabidopsis *Tourist* transposase. However, neither TIRs nor TSD could be identified. This putative element did not share nucleotide sequence similarity to Tc8 members and was not found to be repetitive in the *C. elegans* genome.

In contrast to the 150 bp elements which appear to be recent insertions, Tc8.1 seems to be an ancient insertion since it has accumulated at least four nested insertions and a corresponding RESite appears degenerate (Figure 2). Alternatively, Tc8.1 mobilization occurred after all or some of the nested insertion events. An unusual feature of the nested insertions within Tc8.1 is the presence of another shorter member of the Tc8 group at the same position in each of the TIRs. These elements are not identical (99.3% similarity with a single nucleotide substitution and an indel). Therefore, the nested elements could be independent insertions which occurred at exactly the same location in the TIRs. However, the level of divergence between the nested insertions is similar to the level of divergence between both TIRs (data not shown). Thus, we cannot distinguish between the possibility that they represent independent insertions or that the actual structure of Tc8.1 TIRs is the result of a rearrangement, localized duplication, conversion or a combination of these events. The two other nested insertions within Tc8.1, are located at positions 26747-27109 and 29990-31095 (Figure 1) and are repetitive in the *C. elegans*

genome. The insertion at position 26747-27109 shares > 85% nucleotide similarity with members of a group of putative inverted-repeat transposons previously identified as *Cele1* (Oosumi *et al.* 1995). However, it appears to be a truncated insertion as element boundaries or TSDs could not be clearly identified. The insertion located at position 29990-31095 is annotated in ACeDB as RepQ and, upon closer examination, displays structural characteristics of *IS630*/Tc*1*/*mariner*-like elements. It is not clear whether Tc8 or any of the other elements within Tc8.1 preferentially insert into mobile elements. Such a phenomena has been suggested to reflect a mechanism to minimize deleterious insertion events in other organisms (SanMiguel *et al.* 1996; Vaury *et al.* 1989).

The short hairpin-like Tc8 elements are most likely deletion derivatives of larger elements. As in the case of other DNA-based transposons (*i.e.* class II), the mobilization and spread of deletion derivatives would be facilitated by transposase provided in *trans* by an autonomous element located elsewhere in the genome. The high copy number of very similar Tc8 elements may reflect recent activity. Tc8.1 could have been the source of transposase which mobilized the 150 bp elements. It is unlikely, however, that Tc8.1 is presently active because the putative transposase ORF is disrupted by the *Cele1* insertion (Figure 1). Alternatively, a functional transposase may have been provided by a full length Tc8 element which has been lost from the Bristol N2 strain. It is also possible that Tc8 elements, similar to the case of Tc7 which can be mobilized by a Tc1 transposase (Rezsohazy *et al.* 1997), are currently being mobilized in *trans* by the Tc4 and/or Tc5 transposases as these transposons have structural characteristics similar to Tc8 (see below). An abundance in shorter deletion derivatives is also observed for *Tourist*-like elements in Arabidopsis and other plant genomes. The selective spread of shorter elements may be a mechanism inherent in the transposition process of these elements or may reflect an active host mechanism to minimize element contribution to genome size.

Upon closer examination, the transposases from the Arabidopsis MITE X and XI elements share amino acid similarity to members of the bacterial IS*4* and IS*5* family of elements suggesting a common evolutionary history (Kapitonov and Jurka 1999; Le *et al.* 2000). Likewise, PSI-BLAST searches (Altschul *et al.* 1997) indicate that the Tc8.1

transposase also shares similarity to the same IS transposases. This similarity is restricted to the N-terminal region which contains the DDE motif found in many transposases and integrases and is known to be essential in the catalysis of DNA integration step of transposition (Rezsohazy *et al.* 1993). Members of the IS*4* and IS*5* families have additional conserved residues proximal to the DDE motif, namely, the DX(G/A)(F/Y) and $YX_2RX_3EX_6K$, or YREK, motifs neighboring the last aspartic and glutamic acid residues (Rezsohazy *et al.* 1993). These motifs are also present in the putative *Tourist* transposases (Figure 3). Although the YREK motif appears to be only partially conserved in the putative *Tourist* transposases from *C. elegans* and Arabidopsis, the invariant arginine and glutamic acid residues are conserved (Figure 3).

Phylogenetic analysis using the region corresponding to the conserved DDE motif shows that the Tc8 transposase clusters with the transposases from members of the IS*5* element family (Figure 4). According to both maximum parsimony and neighbor-joining analyses, Tc8 is more closely related to the other eukaryotic elements, Arabidopsis MITE-X and -XI, than they are to bacterial IS elements. Moreover, the eukaryotic *Tourist* elements are more related to bacterial IS5 members that generate a 3 bp TSD (typically, 5'-TAA-3'). This suggests that the Arabidopsis and *C. elegans Tourist*-like elements have evolved from a common ancestral IS sequence. We cannot rule out the possibility, however unlikely, that the results of our analysis reflect convergence of the *Tourist* and IS during evolution. With the identification of new *Tourist* transposases from other organisms, it will be interesting to test the hypothesis that eukaryotic *Tourist* emerged from a common ancestral IS*5* member. Nevertheless, these results suggest that *Tourist* and bacterial IS5 elements form an IS*5*/*Tourist* superfamily.

The terminal nucleotides of the Tc8 TIRs are not only similar to plant *Tourist*-like TIRs, but are also reminiscent to the TIRs of bacterial IS elements. The preference for insertion into the trinucleotide 5'-TAA-3' of Tc8, as confirmed by the RESite analysis, is also a feature of other *Tourist* and IS*5* elements (Figure 2, Table 1). Curiously, the 3 bp TSDs and terminal sequences of Tc8 TIRs are also reminiscent of two other well described active transposons endogenous to *C. elegans*, namely Tc4 and Tc5 (Collins and Anderson

1994; Yuan *et al.* 1991). Even though Tc4 and Tc5 putative tansposases also contain the DDE motif (Robertson 1996; Smit and Riggs 1996), no significant nucleotide or amino acid sequence similarity could be detected beyond the TSDs and terminal nucleotides. The TIRs and TSDs are important in transposase-mediated recognition of element termini and catalytic cleavage of the target sequence, respectively (Fischer *et al.* 1999; van Luenen *et al.* 1994). Consequently, sequence similarity between the TIRs and TSDs of different transposons suggests a common transposase specificity.

Although the identification of RESites provides evidence of past mobility, to date, no *Tourist* element has yet been shown to be currently active. Even though we could not identify ESTs similar to the putative *Tourist* transposases from *C. elegans* or *Arabidopsis*, we did mine ESTs from other plant species, insects and vertebrates (Table 2). Additional ESTs displayed high sequence similarity but only entries with similarities to the conserved motifs in the C-terminal end of the query (Figure 3) are shown in Table 2. These ESTs may correspond to the transposases of different *Tourist* element groups (Table 1). Alternatively, similarity to ESTs may not correspond to transposase expression but may simply represent the transcription of *Tourist* elements which have inserted into or near expressed genes. Unfortunately, genomic sequences corresponding to the EST clones were not found but TBLASTN searches using Tc8 amino acid sequences did reveal the presence of similar sequences in these organisms and in *C. briggsae* (data not shown); discernable ends of these putative elements could be identified. Despite these, it would appear that *Tourist* elements are present and possibly active in these other genomes. The lack of corresponding ESTs does not necessarily indicate lack of Tc8 activity. Transposon activity in *C. elegans* is highly strain dependent. For example, copy number of Tc1 elements can be 10-15 fold lower in Bristol N2 than in Bergerac BO (Egilmez *et al.* 1995). In Bergerac, Tc1 excision activity in somatic cells is higher than in germ line cells, indicating that element activity may be limited to specific tissues or developmental stages (Eide and Anderson 1988). In some eukaryotes (*e.g.* higher plants), host-mediated regulation such as DNA methylation may play a role in silencing transposon activity (Hirochika *et al.* 2000).

MITE insertions are often closely associated with normal plant genes (Bureau and Wessler 1992; Bureau and Wessler 1994a; Bureau and Wessler 1994b). This is in contrast to retrotransposons which are frequently found as nested insertions within heterochromatic regions (SanMiguel *et al.* 1996; Vaury *et al.* 1989). Although a more detailed examination is required to determine if Tc8 are preferentially found associated with genes or contribute *cis* regulatory sequences, there seems to be a negative correlation between Tc8 and gene distribution in *C. elegans*, where genes are clustered near the center of the chromosomes (Figure 5). However, this is not as obvious on the X chromosome where this gene clustering is less pronounced (Barnes *et al.* 1995; The *C. elegans* Sequencing Consortium 1998). Interestingly, Tc8 abundance and distribution relative to genes is similar to what has been previously observed for Tc1 elements in *C. elegans* and *Tourist*-like elements in Arabidopsis, which has a genome of approximately the same size as *C. elegans* (Korswagen *et al.* 1996; The *C. elegans* Sequencing Consortium 1998).

Eukaryotic *Tourist* and bacterial IS elements appear to have emerged from a common ancestor. This is analogous to other eukaryotic elements, namely members of the Tc1/*mariner* superfamily, which are related to IS*630* family of bacterial transposons (Capy *et al.* 1996). Furthermore, many members of the Tc1/*mariner* have been shown to be transpositionally active. Although active mobilization of a *Tourist* still needs to be demonstrated, the identification of ESTs similar to *Tourist* putative transposases suggests that this family is not merely a collection of transposon relics. Thus, *Tourist*-like transposons are potentially active elements related to the widespread bacterial IS*5* transposons, and together, members of the IS*5*/*Tourist* superfamily have successfully populated both prokaryotic and eukaryotic genomes.

Acknowledgments

We thank Julie Poupart, Dr. Rick Roy and Dr. Joseph Dent for critical comments on our manuscript. We are grateful to Boris-Antoine Legault for providing computer-programming support. This work was funded by a National Science and Engineering Research Council (NSERC) grant to T. B. and a *Formation de Chercheurs et l'Aide à la Recherche* (FCAR) fellowship to K.T.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic Local Alignment Search Tool. J. Mol. Biol. 215: 403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-33402.

Barnes, T. M., Y. Kohara, A. Coulson and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics 141: 159-179.

Berg, D. E., and M. M. Howe, 1989 *Mobile DNA*. American Society for Microbiology, Washington, D.C.

Besansky, N. J., O. Mukabayire, J. A. Bedell and H. Lusz, 1996 *Pegasus*, a small terminal inverted repeat transposable element found in the white gene of *Anopheles gambiae*. Genetica 98: 119-129.

Bigot, Y., C. Auge-Gouillou and G. Periquet, 1996 Computer analyses reveal a *hobo*-like element in the nematode *Caenorhabditis elegans*, which presents a conserved transposase domain common with the Tc1-*Mariner* transposon family. Gene 174: 265-271.

Bowen, N. J., and J. F. McDonald, 1999 Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Res. 9: 924-935.

Britten, R. J., 1995 Active *gypsy/Ty3* retrotransposons or retroviruses in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA 92: 599-601.

Bureau, T. E., and S. R. Wessler, 1992 *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4: 1283-1294.

Bureau, T. E., and S. R. Wessler, 1994a Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. U S A 91: 1411-1415.

Bureau, T. E., and S. R. Wessler, 1994b *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6: 907-916.

Butler, M. H., S. M. Wall, K. R. Luehrsen, G. E. Fox and R. M. Hecht, 1981 Molecular relationships between closely related strains and species of nematodes. J. Mol. Evol. 18: 18-23.

Capy, P., R. Vitalis, T. Langin, D. Higuet and C. Bazin, 1996 Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J. Mol. Evol. 42: 359-368.

Collins, J., E. Forbes and P. Anderson, 1989 The Tc3 family of transposable genetic elements in *Caenorhabditis elegans*. Genetics 121: 47-55.

Collins, J. J., and P. Anderson, 1994 The Tc5 family of transposable elements in *Caenorhabditis elegans*. Genetics 137: 771-781.

Doak, T. G., F. P. Doerder, C. L. Jahn and G. Herrick, 1994 A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA 91: 942-946.

Dreyfus, D. H., and S. W. Emmons, 1991 A transposon-related palindromic repetitive sequence from *C. elegans*. Nucleic Acids Res. 19: 1871-1877.

Egilmez, N. K., R. H. Ebert, 2nd and R. J. Shmookler Reis, 1995 Strain evolution in *Caenorhabditis elegans*: transposable elements as markers of interstrain evolutionary history. J. Mol. Evol. 40: 372-381.

Eide, D., and P. Anderson, 1988 Insertion and excision of *Caenorhabditis elegans* transposable element Tc1. Mol. Cell. Biol. 8: 737-746.

Feschotte, C., and C. Mouches, 2000 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.

Fischer, S. E., H. G. van Luenen and R. H. Plasterk, 1999 *Cis* requirements for transposition of Tc1-like transposons in *C. elegans*. Mol. Gen. Genet. 262: 268-274.

Greenwald, I., 1985 *lin-12*, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. Cell 43: 583-590.

Heschl, M. F., and D. L. Baillie, 1990 Functional elements and domains inferred from sequence comparisons of a heat shock gene in two nematodes. J. Mol. Evol. 31: 3-9.

Hirochika, H., H. Okamoto and T. Kakutani, 2000 Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. Plant Cell 12: 357-369.

Izsvak, Z., Z. Ivics, N. Shimoda, D. Mohn, H. Okamoto *et al.*, 1999 Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J. Mol. Evol. 48: 13-21.

Kapitonov, V. V., and J. Jurka, 1999 Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica 107: 27-37.

Kennedy, B. P., E. J. Aamodt, F. L. Allen, M. A. Chung, M. F. Heschl *et al.*, 1993 The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. J. Mol. Biol. 229: 890-908.

Ketting, R. F., T. H. Haverkamp, H. G. van Luenen and R. H. Plasterk, 1999 Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell 99: 133-141.

Korswagen, H. C., R. M. Durbin, M. T. Smits and R. H. A. Plasterk, 1996 Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as marker for gene mapping. Proc. Natl. Acad. Sci. USA 93: 14680-14685.

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

Levitt, A., and S. W. Emmons, 1989 The Tc2 transposon in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA 86: 3232-3236.

Mahillon, J., and M. Chandler, 1998 Insertion sequences. Microbiol. Mol. Biol. Rev. 62: 725-774.

Malik, H. S., and T. H. Eickbush, 1998 The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol. Biol. Evol. 15: 1123-1134.

Oosumi, T., B. Garlick and W. R. Belknap, 1995 Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U S A 92: 8886-8890.

Oosumi, T., B. Garlick and W. R. Belknap, 1996 Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. J. Mol. Evol. 43: 11-18.

Plasterk, R. H. A., and H. G. A. M. van Luenen, 1997 Transposons, pp. 97-116 in *C. elegans II.* D. L. Riddle, T. Blumenthal, B. J. Meyer and J. R. Priess Eds. CSH laboratory Press, Cold Spring Harbor.

Rezsohazy, R., B. Hallet, J. Delcour and J. Mahillon, 1993 The IS*4* family of insertion sequences: evidence for a conserved transposase motif. Mol. Microbiol. 9: 1283-1295.

Rezsohazy, R., H. G. van Luenen, R. M. Durbin and R. H. Plasterk, 1997 Tc7, a Tc1-hitch hiking transposon in *Caenorhabditis elegans*. Nucleic Acids Res. 25: 4048-4054.

Robertson, H. M., 1996 Members of the *pogo* superfamily of DNA-mediated transposons in the human genome. Mol. Gen. Genet. 252: 761-766.

Rushforth, A. M., B. Saari and P. Anderson, 1993 Site-selected insertion of the transposon Tc1 into a *Caenorhabditis elegans* myosin light chain gene. Mol. Cell. Biol. 13: 902-910.

SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Smit, A. F., and A. D. Riggs, 1996 *Tiggers* and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. U S A 93: 1443-1448.

Surzycki, S. A., and W. R. Belknap, 1999 Characterization of repetitive DNA elements in *Arabidopsis*. J. Mol. Evol. 48: 684-691.

Surzycki, S. A., and W. R. Belknap, 2000 Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc. Natl. Acad. Sci. USA 97: 245-249.

Swofford, D. L., 2000 PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), pp. Sinauer Associates, Sunderland, Massachusetts.

Tatusova, T. A., and T. L. Madden, 1999 BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174: 247-250.

The *C. elegans* Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012-2018.

Tu, Z., 1997 Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. Proc. Natl. Acad. Sci. USA 94: 7475-7480.

Turcotte, K., S. Srinivasan and T. Bureau, 2001 Survey of transposable elements from rice genomic sequences. Plant J. 25: 1-13.

Unsal, K., and G. T. Morgan, 1995 A novel group of families of short interspersed repetitive elements (SINEs) in *Xenopus*: evidence of a specific target site for DNA-mediated transposition of inverted-repeat SINEs. J. Mol. Biol. 248: 812-823.

van Luenen, H. G., S. D. Colloms and R. H. Plasterk, 1994 The mechanism of transposition of Tc3 in *C. elegans*. Cell 79: 293-301.

Vaury, C., A. Bucheton and A. Pelisson, 1989 The beta heterochromatic sequences flanking the I elements are themselves defective transposable elements. Chromosoma 98: 215-224.

Vos, J. C., I. De Baere and R. H. Plasterk, 1996 Transposase is the only nematode protein required for in vitro transposition of Tc1. Genes Dev. 10: 755-761.

Vos, J. C., H. G. van Luenen and R. H. Plasterk, 1993 Characterization of the *Caenorhabditis elegans* Tc1 transposase *in vivo* and *in vitro*. Genes Dev. 7: 1244-1253.

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

Williams, B. D., B. Schrank, C. Huynh, R. Shownkeen and R. H. Waterston, 1992 A genetic mapping system in *Caenorhabditis elegans* based on polymorphic sequence-tagged sites. Genetics 131: 609-624.

Yeadon, P. J., and D. E. Catcheside, 1995 *Guest*: a 98 bp inverted repeat transposable element in *Neurospora crassa*. Mol. Gen. Genet. 247: 105-109.

Yuan, J. Y., M. Finney, N. Tsung and H. R. Horvitz, 1991 Tc4, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. Proc. Natl. Acad. Sci. USA 88: 3334-3338.

Table 1. TSD and TIR sequence similarity between insertion transposons

Element	Host Organism	TSD	TIR*	Nucleotide gi Nb.	Protein gi Nb.†
MITE X	*Arabidopsis thaliana*	TAA	GGGGGTGTTATTGGT	4454587	N/A‡
MITE XI	*Arabidopsis thaliana*	TAA	GGTCCTGTTTGTTTG	4235150	4585884
Tc8	*Caenorhabditis elegans*	TAA	TGGGGTTATTCAAGT	2746790	7499082
IS*5*Sa	*Synechocystis* sp.	TAA	GAGCGTGTTTGAAAA	1256581§	1256580
IS*1031*	*Acetobacter xylinum*	TAA	GAGCCTGATCCGAAA	349283	349284
IS*12528*	*Gluconobacter suboxydans*	TCA	GAGCCCTTTTGGAAA	2055292	2055293
IS*903*	*Escherichia coli*	9-bp¶	GGCTTTGTTGAATAA	43025	581236
ISH*11*	*Halobacterium halobium*	7-bp¶	GAGGGTGTCATAGAA	43507	43508
IS*6501*	*Brucella ovis*	TA¶	CTAGAGCTTGTCTGC	295859	454221
IS*5*	*E. coli* (lambda KH 100)	CWAR‖	GGAAGGTGCGAATAA	49080	455278
IS*112*	*Streptomyces albus*	TA‖	AGGGCTGTCCCGTAA	46594	581565
IS*493*	*Streptomyces lividans*	ANT‖	GAGCGTTTTTCAACC	1707869	1196467
IS*702*	*Calothrix* sp. PCC 7601	CGT‖	AGAACTCTTGCAAAA	40657	581004
ISL*2*	*Lactobacillus helveticus*	AAT‖	AGAGTTACTGCGAAA	763132	763133
Tc4	*Caenorhabditis elegans*	TNA	CTAGGGAATGACCAG	156456	**
Tc5	*Caenorhabditis elegans*	TNA	CAAGGGAAGGTTCTG	529006	**

* 15 bases of terminal end sequences are shown (left TIR, from 5' to 3'). The actual TIR may be longer or shorter. † Amino acid sequence used for phylogenetic analysis (Fig. 4). ‡ conceptual translation of gi 4454587 from position 4650-5865. § Sequence shown in table is the reverse-complement of the sequence found on clone in GenBank. ¶ No TSD consensus sequence. ‖ TSD sequence information extracted from the IS database (http://pc4.sisc.ucl.ac.be/is.html). W = A or T, R = A or G and N = any nucleotide. ** Not included in the phylogenetic analysis.

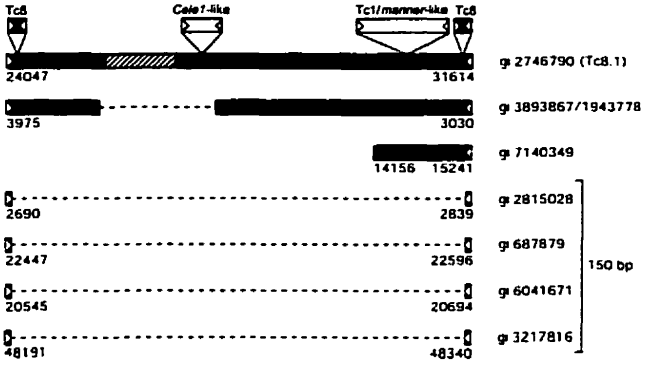
Table 2. dbEST entries with TBLASTN similarity to the putative *Tourist* transposases from *C. elegans* and Arabidopsis*

Query†	Organism	Genbank Accession Number‡
C. elegans	*Bombyx mori*	AV404936
(Tc8)	*Anopheles gambiae*	AJ281674, AJ284723
	Danio rerio	AI882971
	Zea mays	AW165620, AI783120, AW165636, BE345915, AW172080
	Lycopersicon hirsutum	AW616734, AW616736
	Sorghum bicolor	AW746190
	Oryza sativa	AU068684
	Glycine max	AW759312
	Bos taurus	AW353649
	Gallus gallus	AI981097
	Oryza sativa	AU068684
	Glycine max	AW759312
Arabidopsis	*Lycopersicon hirsutum*	AW616734, AW616736
(MITE-X	*Lycopersicon esculentum*	AW032471
and MITE-XI)	*Zea mays*	AW438057, AW438058, AW000374, AW172080, AI677510, AW165620, AI1783120, AW165636, BE345915, AI941809, AI947769, BE510064, BE509925, AW052790, AW352677, AW000051
	Triticum monococcum	BE492211
	Triticum aestivum	BE489993
	Sorghum bicolor	BE365933, BE593965, AW746190
	Oryza sativa	AU075345, AU068684, C25943
	Danio rerio	AI882971
	Gallus gallus	AI981097
	Bos taurus	AW353649
	Xenopus laevis	BE191894
	Anopheles gambiae	AJ281674
	Mus musculus	AA003110

* Entries listed have similarity to the DDE motif of the query sequence. † Amino acid sequence used as query for *C. elegans* were Tc8.1 (gi 7499082) and the second putative *C. elegans Tourist* ORF (gi 7504556); for Arabidopsis MITE-X from gi 4585884 and MITE-XI from the conceptual translation of gi 4454587, nucleotide position 4650-5865. ‡ Alignments can be viewed at http://www.tebureau.mcgill.ca/.

Figure 1. *Tourist*-like elements in the *C. elegans* genome. (A) Diagram of Tc8 members. The majority of the elements are 150 bp and share high sequence similarity. Additional longer members were found but were truncated. Dashes represent gaps and the hatched box indicates the region (~300 amino acids) of the annotated ORF which is similar to the C-terminal domain of other transposases. Boxes above Tc8.1 with inverted arrowheads represent nested insertions. The TIR and TSD sequences of the nested *Tourist*-like element are degenerate. GenBank gi numbers and positions of elements on corresponding clones are indicated to the right and below, respectively. (B) A Tc8-like group of elements in the genome of *C. briggsae*. Nucleotide alignment of *C. elegans* Tc8 from gi 2746790, position 24108-24275 and *C. briggsae Cb*-Tc8 from gi 11095049, position 34282-34399.

A



B

```
Tc8-Ce   1 tggggttattcaagtagtgtcggaaaattaaaaagtgtagaaaaattacg 50
           ||||||||||||| |||| |  |  |  | || |  |       |
Tc8-Cb   1 tggggttattcacgtagcgaaaatatgtattgaaatgcatttccgcgcct 50

Tc8-Ce  51 tcacaactgtatttaaaatacataaaaacatgtgttttaatacatttgtga 100
           || || | | || |  |  | |  | |||| ||| ||||||| |
Tc8-Cb  51 tcgaaattcaatgcatgagcgtcgcgaaatgtatttcaatacatat.... 96

Tc8-Ce 101 cgtcacaaatgtatttaaatacatttgctacattacttgaataacccca 150
                                    | |  ||| ||||||||||||
Tc8-Cb  97 .... ............ .......tttcgctacatgaataacccca 118
```

Figure 2. RESites corresponding to Tc8 element insertions. The identification of RESites suggests evidence of past mobility. Black boxes with inverted arrowheads represent Tc8 insertions. GenBank gi numbers and position on clones are indicated. Target sites are underlined and TSDs are in bold.

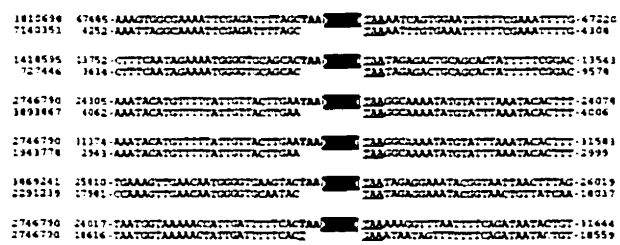


Figure 3. Similarity between putative MITE and bacterial IS transposases. GenBank gi numbers for amino acid sequences used in the alignment are indicated in Table 1. Amino acid coordinates are indicated to the left. The conserved DDE residues are indicated below the alignment. The DX(G/A)(Y/F) and YREK motifs found in other IS families (Capy *et al.* 1996) are boxed.

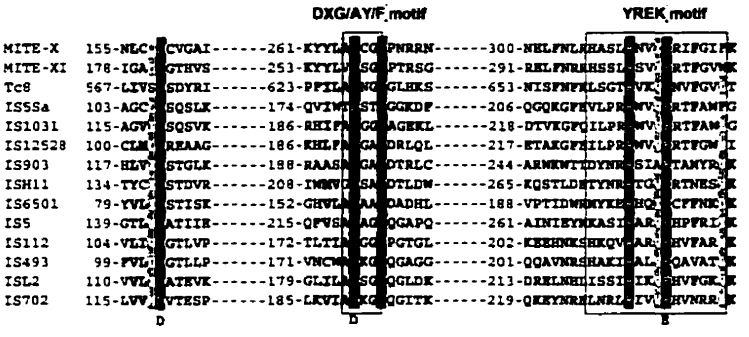

DXG/AY/F motif
YREK motif
MITE-X
MITE-XI
Tc8
IS5Sa
IS1031
IS12528
IS903
ISH11
IS6501
IS5
IS112
IS493
ISL2
IS702
D
D
E

Figure 4. Evolutionary relationship between *Tourist* and bacterial IS*5* elements. Phylogenetic analysis showing the relationship between the putative *Tourist* transposases from Arabidopsis (MITE-X and -XI) and from *C. elegans* to transposases from bacterial IS*5* elements. GenBank gi numbers for amino acid sequences used in the alignment are indicated in Table 1. Parsimony and the neighbor joining trees (not shown) are all concordant with the placement of Tc8 transposases with representatives of the bacterial IS*5* family. The unique and most parsimonious tree, based on the alignment in Figure 3, is shown (length = 335, CI = 0.818, RI = 0.564). Numbers at the base of nodes correspond to boostrap values. Typical size of TSDs are indicated to the right.

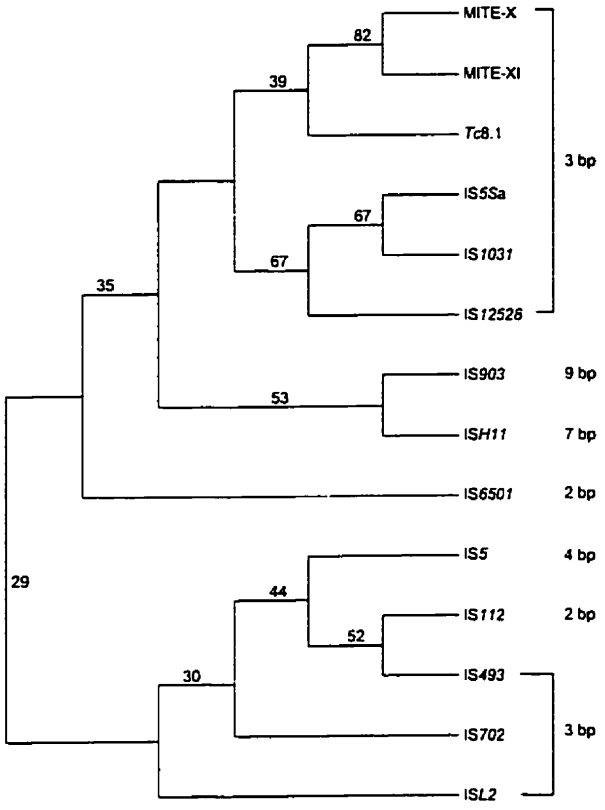
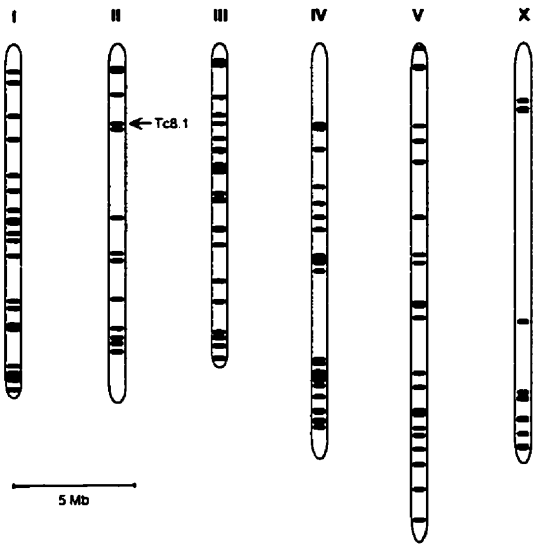

MITE-X
MITE-XI
Tc8.1
IS5Sa
IS1031
IS12526
IS903
ISH11
IS6501
IS5
IS112
IS493
IS702
ISL2
82
39
67
67
35
53
29
44
52
30
3 bp
3 bp
9 bp
7 bp
2 bp
4 bp
2 bp

Figure 5. Distribution of Tc8 elements in the *C. elegans* genome. Each line represents a Tc8 insertion. Location of Tc8.1 is indicated. Scale is indicated at the bottom. The first nucleotide position on chromosome is located at the top.


I
II
III
IV
V
X
Tc8.1
5 Mb

CHAPTER 3

Stowaway and *Emigrant* miniature inverted-repeat transposable elements share a common evolutionary history with IS*630*/Tc1/*mariner*-like transposons

Title: ***Stowaway* and *Emigrant* Miniature Inverted-repeat Transposable Elements Share a Common Evolutionary History with IS*630*/Tc1/*mariner*-like Transposons**

Authors: **Kime Turcotte and Thomas Bureau**

Address: Department of Biology, McGill University, Montreal, Quebec, Canada H3A 1B1

Running head: Evolutionary history of MITEs

Keywords: *Stowaway*, *Emigrant*, MITE, *mariner*, transposon

Corresponding author: Thomas Bureau, Department of Biology, McGill University, 1205 Docteur Penfield, Montreal, Quebec, Canada H3A 1B1
Tel.: (514) 398-6472; Fax.: (514) 398-5069;
E-mail: thomas_bureau@maclan.mcgill.ca

Abbreviations: MITE: Miniature Inverted-repeat Transposable Element; TE: Transposable element; LTR: Long Terminal Repeat; TIR: Terminal Inverted-Repeat; TSD: Target Site Duplication; IS: Insertion Sequence; ORF: Open Reading Frame; EST: Expressed Sequence Tag; MLE: *mariner*-like element;.

Abstract

The genomes of plants, like virtually all other eukaryotic organisms, harbor a diverse array of mobile elements or transposons. In terms of numbers, the predominant type of transposons in many plants is Miniature Inverted-repeat Transposable Elements or MITEs. These elements are typically short and have terminal inverted-repeats and target site sequence preference. There are three archetypal MITEs known as *Tourist*, *Stowaway* and *Emigrant*, each of which can be defined by a specific terminal inverted-repeat sequence signature. Although their presence was known for over a decade, only recently have open reading frames been identified that correspond to putative transposases for each of the archetypes. *Tourist*-like MITEs along with prokaryotic IS*5*-like elements form a unique transposon superfamily. *Stowaway* and *Emigrant*, on the other hand, are more similar to members of the previously characterized IS*630*/Tc1/*mariner* superfamily. In this report we provide a high resolution phylogenetic analysis of the evolutionary relationship between *Stowaway*, *Emigrant*, and members of the IS*630*/Tc1/*mariner* superfamily. We show that while *Emigrant* is closely related to the *pogo*-like family of elements, *Stowaway* may represent a novel family.

Introduction

Transposable elements (TEs) or transposons are fundamental components of most, if not all, prokaryotic and eukaryotic genomes. Though many elements move through a RNA intermediate and rely on the action of reverse transcriptase (Class I), still others move in a DNA form through a "cut and paste" mechanism (Class II). Class I and II elements can be divided into superfamilies based on shared sequence similarity, terminal motifs, coding capacity, and target site size and sequence. For example, the IS*630*/Tc1/*mariner* superfamily contains members from several prokaryotic and eukaryotic organisms. Some elements, such as the recently identified *Basho* elements (Le *et al.* 2000) cannot be easily classified into any known Class I or II superfamily and remain enigmatic.

In plants, a diverse number of transposons have been identified. Plants with very large genomes are heavily populated with *copia*- and *gypsy*-like LTR (Long Terminal Repeat)-retrotransposons (Class I). Plants with smaller genomes have fewer Class I elements but harbor numerous types of Class II and other unclassified elements. The diversity of transposons within plant genomes is also plant family dependent. For instance, *Arabidopsis thaliana*, a dicotyledon and a member of the Brassicaceae, and *Oryza sativa* (domesticated rice), a monocotyledon and a member of the Poaceae, both have small genomes but have different complements of transposons (Le *et al.* 2000; Turcotte *et al.* 2001).

Regardless of the plant species, MITEs (Miniature Inverted-repeat Transposable Elements) appear to be highly abundant and in rice clearly predominate. Discovered almost a decade ago through computer-based sequence similarity searches, MITEs were initially divided into two types based on sequence and structural features and are defined by the archetypal elements, *Tourist* and *Stowaway* (Bureau and Wessler 1992; Bureau and Wessler 1994a; Bureau and Wessler 1994b). Recently, a third archetypal element has emerged called *Emigrant* (Casacuberta *et al.* 1998). All three MITE types were originally identified and characterised as non-autonomous elements due to a lack of coding capacity and small size (<500 bp) and were assumed to be Class II elements since they possessed

terminal inverted-repeats (TIRs). Members of the three MITE types appear to have a target site preference for either 5'-TAA-3' (*Tourist*) or 5'-TA-3' (*Stowaway* and *Emigrant*). This feature appears to be unique among elements in plants. The location of elements in AT-rich sequences extends beyond the target site as many MITEs are located within AT-rich domains ranging from 500-1000 bp (Le *et al.* 2000; Turcotte *et al.* 2001). It is unclear whether this is the result of purifying selection from GC (gene-rich) regions over evolutionary time or the consequence of targeted insertions. Furthermore, some studies have indicated that MITEs may preferentially insert near transcription units and may, therefore, play an important role in plant gene evolution (Bennetzen 2000; Bureau and Wessler 1992; Bureau and Wessler 1994a; Bureau and Wessler 1994b; Wessler 1998; Zhang *et al.* 2000). MITEs are not restricted to plants, as elements have been reported in insects (Braquart *et al.* 1999; Feschotte and Mouches 2000b; Tu 1997; Tu 2000), fishes (Izsvak *et al.* 1999), nematodes (Le *et al.* 2001; Surzycki and Belknap 2000) and mammals (Smit and Riggs 1996). Many of the MITEs appear to be structurally similar to the members of the IS*630*/Tc1/*mariner* superfamily of transposons.

For many years, the relationship between the MITE types has remained unresolved since no larger elements with coding capacity could be identified. However, analysis of information generated from the Arabidopsis and rice genome sequence initiatives has recently uncovered the first glimpses into the *Tourist*, *Stowaway* and *Emigrant* transposases. Members of two Arabidopsis *Tourist*-like MITE groups were found with coding capacity (Le *et al.* 2000). Although quite different at the nucleotide level, the open reading frames (ORFs) from both MITE groups share a striking level of similarity in their amino acid sequences. The Arabidopsis *Tourist* transposases share similarity with bacterial Insertion Sequences (IS), specifically with members of the IS*5* family. In addition, ORFs and expressed sequence tags (ESTs) with significant similarity to the Arabidopsis *Tourist* transposases were identified in other plant and non-plant species. In fact, this information led to the identification of a repetitive *Tourist*-like element in the *C. elegans* (Bristol N2) genome, designated Tc8 (Le *et al.* 2001). A larger *Stowaway* element was identified in rice that possessed coding capacity for a putative transposase (Turcotte *et al.* 2001). This ORF shares high sequence similarity with another ORF

harbored by *Soymar1*, a *mariner*-like element (MLE) from soybean (*Glycine max*) (Jarvik and Lark 1998; Tarchini *et al.* 2000). The rice and soybean elements share 70-100% nucleotide identity with the TIRs of shorter previously identified *Stowaway* elements and have a characteristic 5'-TA-3' target site duplication (TSD) (Turcotte *et al.* 2001). Despite their possible relationship with members of the IS*630*/Tc1/*mariner* superfamily, the *Stowaway* TIR sequence signature is unique. In contrast, the recently identified transposase for *Emigrant* confirms its relationship with the IS*630*/Tc1/*mariner* superfamily (Feschotte and Mouches 2000a). The TIR signature sequence for *Emigrant* and many members of the IS*630*/Tc1/*mariner* superfamily share similarity with the four terminal-most basepairs, namely the catalytically important 5'-CAGT-3'. Furthermore, *Emigrant* also has a preference for insertion into the dinucleotide 5'-TA-3' (Casacuberta *et al.* 1998).

The *Stowaway*, *Tourist*, and *Emigrant*, transposases harbor a conserved motif called the DD(D/E) signature (Feschotte and Mouches 2000a). The DDE motif had been identified in the transposases of members of several bacterial IS families and within the integrase domain of retroviruses and retrotransposons (Fayet *et al.* 1990; Kulkosky *et al.* 1992). Later this motif was determined to be present in the eukaryotic counterparts of some transposon superfamilies, such as the IS*630*/Tc1/*mariner* superfamily (Capy *et al.* 1997; Capy *et al.* 1996; Doak *et al.* 1994; Vos and Plasterk 1994). The DDD signature found in *pogo*-like and *mariner*-like transposases is assumed to be homologous to the DDE motif (Doak *et al.* 1994; Langin *et al.* 1995; Robertson and Lampe 1995a). The three residues are essential for transposition as they define a cation binding site necessary for cleavage and strand transfer reactions. Mutation in any of these residues abolishes activity (Kulkosky *et al.* 1992; Lohe *et al.* 1997; Vos and Plasterk 1994). This motif is present in many transposases and integrases, suggesting a common evolutionary history between the two classes of TEs.

We will provide in this report details of an additional *Stowaway* element harboring ORFs. This second complete element in rice confirms earlier results suggesting the similarity of the *Stowaway* and IS*630*/Tc1/*mariner* transposases. Moreover, we will show how we

took advantage of the virtual landslide of genome EST sequence information to unravel the relationship between *Stowaway* and *Emigrant*-type elements and discuss the implications of this relationship in the evolution of MITEs.

Materials and Methods

Nucleotide and protein sequences used in this study (Table 1) were retrieved from the GenBank public database. Identification of the residues corresponding to the DD(D/E) motif and the alignment of the amino acid sequences of this region are based on previous publications (Capy *et al.* 1996; Doak *et al.* 1994; Feschotte and Mouches 2000a; Garcia-Fernandez *et al.* 1995; Jarvik and Lark 1998; Smit and Riggs 1996). Sequences of the transposases for *Stowaway*-like elements were found using, BLAST tools from NCBI (http://www.ncbi.nlm.nih.gov/blast/) (Altschul and Koonin 1998; Altschul *et al.* 1997). The amino acid sequence of Osa-*Stow1* was obtained using ORF FINDER (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) with GI:6979318 (227524 to 232786). The 576 amino acids long sequence in the +2 frame corresponds to the putative transposase including the DDD residues (Table1). Alignments of the 32 amino acid sequences was done manually. A few gaps were allowed in the alignment in order to maintain a maximum amino acid identity. The sequences used for phylogenetic analysis are shown in Figure 1. Phylogenetic analysis was carried out using PAUP* version 4.0b8 for MacIntosh (Swofford 2000), with maximum parsimony method (multiple trees heuristic search option with tree bisection-reconnection) and the neighbour joining distance method. Bootstrap analysis was done with 200 replicates, resampling 53 characters and including groups compatible with the 50% majority rule consensus tree. Based on previous studies (Capy *et al.* 1997; Capy *et al.* 1996; Plasterk *et al.* 1999), IS*630* was specified as an outgroup to the monophyletic Tc1/*mariner* superfamily of TEs. Distances were calculated as mean pairwise differences adjusted for missing data, obtained from PAUP* version 4.0b8 (Swofford 2000).

Results

Residues flanking the conserved DD(D/E) signature were extremely variable, difficult to align, and too ambiguous to provide reliable phylogenetic characters. Therefore, we decided to restrict the alignment to the most conserved region between the transposases (53 characters). These sequences provide sufficient resolution of the transposases with 49 parsimony informative sites.

The sequences analysed, except the MITE transposases, have been previously shown to form a monophyletic superfamily of TEs (Capy *et al.* 1997; Capy *et al.* 1996; Plasterk *et al.* 1999). These sequences clearly define three main monophyletic groups corresponding to families of TEs (Figure 2). The elements harboring the DDD signature comprise the *mariner* and *pogo* families. The third main family consists of the Tc1-like sequences, characterised by a DDE signature. These three lineages are robust since applying the maximum parsimony (Figure 2) and the neighbour joining (Figure 3) methods yielded nearly-identical clades. The maximum parsimony method yielded 30 equally parsimonious trees with variable branching order for these three major clades, resulting in an absence of resolution of these clades in the consensus tree (Figure 2). However, the bootstrap tree (data not shown) reflects the topology obtained with the neighbour joining method (Figure 3) and conforms with the general scheme established by Capy et al. (1997). Other than the conserved DD(D/E) signature, conserved residues were identified within and between family members. For example, the histidine (H) at position 27 and the glutamic acid (E) at position 6 of the alignment shown in Figure 1 are almost perfect symplesiomorphies (shared ancestral character states) that only Tc4 and Tc5 do not share. The proline (P) at position 43 is also shared by the majority of the sequences. The residues at position 3 and 42 are very useful characters corresponding to almost non-homoplasic apomorphies (a unique derived character state) for each of the families. Also, the histidine (H) at position 33 and the tyrosine (Y) at position 46 are unique to members of the *mariner* family. Finally, at position 3, 42, 45, and 49, Tc1-like elements have a plesiomorphic residue (ancestral character state), responsible for their basal position relative to the *mariner* and *pogo* clades in many trees obtained.

The three main clades (*mariner*, *pogo* and Tc1) also correspond to the topology obtained by other authors, using similar methods (Capy *et al.* 1997; Capy *et al.* 1996; Plasterk *et al.* 1999). In those studies, the IS*630* family of bacterial elements was determined to be ancestral to the Tc1/*mariner* superfamily. We therefore specified IS*630* as an outgroup for our phylogenetic analysis. In order to reduce the chances of an erroneous polarization of characters, we have selected two IS*630* sequences from different organisms. In fact, these two sequences, although not identical, cluster together in all trees.

The rice *Stowaway* Osa-*Stow1* was identified based on TIR sequence similarity between *Stowaway* MITEs and the flanking region of a sequence annotated as a Tc1/*mariner*-like element on a large insert clone (GI:6979318) (Tarchini *et al.* 2000). Similarity searches (BLASTX) using the nucleotide sequence of Osa-*Stow1* transposase as a query, led to the identification of another *Stowaway* transposase in rice (Osa-*Stow2*). The two rice transposase genes are flanked by *Stowaway*-like TIRs and TSDs. These elements are respectively 5,3 kb and 3,6 kb long. The *Stowaway* transposases found show high amino acid similarity to the transposase for *Soymar1* from *Glycine max* (Jarvik and Lark 1998). Similarity searches (TBLASTN) using the amino acid sequence of the rice *Stowaway* transposases revealed one more sequence from *Oryza sativa* and eight sequences from *Arabidopsis thaliana* genomic clones, which share high similarity with the query (data not shown). However, TIRs could not be identified. Also, two EST sequences from *Triticum aestivum* and three from *Zea mays* were found with high similarity to the *Stowaway* transposases around the DD(D/E) signature. Three sequences are missing the region corresponding to the first or the last aspartic acid residue (D) of the DD(D/E) signature, because the EST sequences were too short. Phylogenetic analysis indicated that the ESTs are more closely related to the transposase of members of the *Stowaway* family than they are to Tc1, *mariner* or *pogo* (data not shown). Using a reduced data set of transposase sequences, we have investigated the relationships between these ESTs and members of the *Stowaway* family (Figure 4). The topology of the neighbour joining tree (data not shown) and the maximum parsimony tree (Figure 4) are similar. *Soymar1* has apparently recently emerged within the family and is closely related to the three EST

sequences from *Zea mays*. The two EST sequences from *Triticum aestivum* appear to be ancestral to the rest of the family, while Osa-*Stow1* and Osa-*Stow2* have an intermediate position in the tree. Most of the trees group one of the rice *Stowaway* sequences with *Soymar1*, while the other rice *Stowaway* sequence shows earlier divergence.

The three *Stowaway*-like elements cluster together as a sister group to the pogo family in our analysis. The presence of an aspartic acid residue (D) instead of a glutamic acid residue (E) at position 45 support the closer relationship with *mariner* and *pogo*, than with Tc1, although the *Stowaway*-like elements share conserved residues, other than the DD(D/E) signature, with members of the three families. For example, the presence of two glutamines (Q) at position 20 and 21 and an asparagine (N) at position 42 (Figure 1) is characteristic of several members of the Tc1 family, but is also present in the *Stowaway*-like elements. Also, the presence of a lysine (K) and a tryptophan (W) at position 7 and 8 associates these sequences with those of the *mariner* family, while the leucine (L) at position 44 is reminiscent of members of the *pogo* family. Overall, the sequences of *Stowaway*-like elements are significantly different from sequences of all three main families.

The *Emigrant* MITE transposase sequence clusters within the *pogo* family in close association with the human *Tigger* and Drosophila *pogo* elements. This is consistent with a previous report suggesting that *Emigrant* is a *pogo*-like element, based on TIR, TSD and transposase sequence similarity with other *pogo* elements (Feschotte and Mouches 2000a). The residues at position 49 and 52 are unique to the *pogo-Tigger-Emigrant* cluster, while the asparagine (N) and glutamine (Q) residues at position 3 and 42, respectively, support their position within the *pogo* family.

On average, elements from different families differ by 70 to 79% amino acids, while elements from the same family differ by less than 61% amino acids (Table 2). We observe that the most conserved region of *Stowaway*-like transposases differ by 28% amino acids, while this group of elements show 66 to 75% amino acid difference with the three main families, according to the alignment in Figure 1. In comparison, the level of

sequence similarity between the *Emigrant* transposase and the other members of the *pogo* family is high (60% amino acid difference, Table 2). We also observe that Tc4 and Tc5 differ significantly from the rest of the *pogo* family (70% amino acid difference). Similarly, Tec1 and Tec2 differ from the other Tc1 members (68% amino acid difference). However, *Emigrant*, Tc4, Tc5, Tec1 and Tec2 harbor the conserved residues typical of their family. In addition to a difference in amino acid sequence, Tc4 and Tc5 have a 3 bp target site and have TIRs different from other *pogo*-like elements (Table 1).

Discussion

The phylogeny of the elements analysed in this report does not necessarily reflect the phylogeny of their hosts. Some elements may have been present in the genome of a common ancestor of related taxa, and may have been vertically transmitted to successive generations. However, it is likely that horizontal transfer has played an important role in the history of these elements. In fact, there have been many reports of horizontal transmission events involving members of the IS*630*/Tc1/*mariner* superfamily (Capy *et al.* 1994; Garcia-Fernandez *et al.* 1995; Hartl *et al.* 1997; Lohe *et al.* 1995; Robertson and Lampe 1995b). Horizontal transfer has been suggested to be a possible mechanism for element survival and spread. It may also explain the widespread distribution of these class II elements in diverse eukaryotes (Robertson and Lampe 1995a). With this in mind, we have randomly chosen TEs from different hosts and loci, to determine the evolutionary history of element sequences rather than the relationship of their hosts.

Through the analysis of the most conserved region of the transposase from each element, we observe that the *Emigrant* transposase clearly clusters within the *pogo* family. These MITEs have TIRs reminiscent of the typical terminal-most sequences of *pogo* and *mariner*-like elements, that is 5'-CAGT-3', and harbor residues that are unique to the *pogo* family. Therefore, it is likely that *Emigrant* and the *pogo*-like elements originated from a common ancestor relatively recently. Although the transposases corresponding to the *Stowaway*-like elements do not share such obvious similarity with any of the sequences analysed, being equally distant to Tc1, *mariner*, and *pogo* families (Table 2), they nonetheless form a sister group to the *pogo* family. It is therefore likely that the *Stowaway*-like elements may represent a fourth family of IS*630*/Tc1/*mariner*-like elements.

Capy et al. have proposed that an ancestral transposase gene would have acquired TIRs, giving rise to the ancestral DD(D/E) transposon. This element would have given rise to several bacterial IS groups. More specifically, an IS*630*-like element would have evolved into Tc1/*mariner* elements (Capy *et al.* 1997; Capy *et al.* 1996). Certain positions within

the TIR sequence that act as binding sites for the transposase, would have been more conserved, while the remainder of the element sequence evolved defining the different groups. Many of the IS*630*/Tc1/*mariner* elements share a conserved terminal-most sequence motif, that is 5'-CAGT-3'. Within the *pogo* family, the members of the *pogo-Tigger-Emigrant* group also possess the 5'-CAGT-3' motif, however, there are exceptions. For examples, the termini of Tan1, Fot1, Pot2, Tc4, and Tc5 all differ from the 5'-CAGT-3' motif. Within the Tc1 family, Tec1, Tec2, *Minos2*, and Tes1 have diverse terminal sequences. Within *mariner*, most elements start with 5'-TTAG-3' or 5'-CCAG-3'. The *Stowaway*-like elements appear to be the most aberrant with the terminal motif 5'-CTCC-3'. Therefore, terminal motif signatures may not be conclusive evidence for determining phylogenetic relationships among elements in the IS*630*/Tc1/*mariner* superfamily. Target site size and, in some cases ,sequence appear to be much more consistent traits than TIR motifs within known TE superfamilies. Within the IS*630*/Tc1/*mariner* superfamily all of the elements have a target site preference for the dinucleotide 5'-TA-3' with the exception of Tc4 and Tc5 (5'-TAA-3').

It is clear from our analysis that the three types of MITEs have different evolutionary histories despite similarity in structure. *Tourist* has been shown to be more related to prokaryotic IS*5*-like elements than to *Stowaway* or *Emigrant* (Le *et al.* 2000; Le *et al.* 2001). Recently computer-based searches using the *Tourist* transposase as a query has led to the identification of *Tourist*-like elements, designated Tc8 group, within the *C. elegans* genome (Le *et al.* 2001). *Emigrant* and *Stowaway* are closely related based on our study but have very different TIR sequences. With these observations, it is no longer possible to consider MITEs as defining a superfamily of elements. Clearly, *Tourist* may define a unique superfamily of elements but *Stowaway* and *Emigrant* belong to the previously established IS*630*/Tc1/*mariner* superfamily. Many non-plant eukaryotic MITEs have been reported though, for the most part, ORFs encoding putative transposases have yet to be identified. We would suggest that the term MITE be used only to reflect a mobile element that has TIRs and target site sequence preference for 5'-TA-3' or 5'-TAA-3'. As transposase sequences become available, only then can the classification in terms of superfamily designation be determined.

Phylogenetic analysis has yielded revealing insights into the evolutionary history of several TE superfamilies. In plants, the class II elements *Tourist*-like, *Stowaway*-like, and MULEs predominate. MULEs have been shown to be part of a superfamily of elements that include prokaryotic IS*256*-like elements although non-plant eukaryotic members have yet to be identified. *Ac*-like and CACTA-like elements are also represented in many plant genomes but are not as numerous. These two element types have also been suggested to form distinct superfamilies though prokaryotic counterparts have not been reported. Interestingly, members of the prokaryotic ISA*s1* family have a terminal sequence (5'-CAGGG-3') and target site sequence size (i.e. 8 bp) consistent with many members of the *Ac*-like superfamily. However, the putative transposases for these elements do not share sequence similarity so the actual evolutionary relationship, if any, remains unclear. Nevertheless, these relationships have provided evidence that many eukaryotic elements have prokaryotic counterparts suggesting a long evolutionary history.

Acknowledgements

We would like to thank Dr. Louise O'Donoughue, Dr. Anne Bruneau, and Quang Hien Le for critical reading of the manuscript. This work was supported by an FCAR (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche) fellowship to K.T., and an NSERC (Natural Sciences and Engineering Research Council of Canada) grant to T.B.

References

Altschul, S. F., and E. V. Koonin, 1998 Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem. Sci. 23: 444-447.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman,1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Amutan, M., E. Nyyssonen, J. Stubbs, M. R. Diaz-Torres and N. Dunn-Coleman,1996. Identification and cloning of a mobile transposon from *Aspergillus niger* var. awamori. Curr. Genet. 29: 468-473.

Bennetzen, J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 42: 251-269.

Braquart, C., V. Royer and H. Bouhin, 1999 DEC: a new miniature inverted-repeat transposable element from the genome of the beetle *Tenebrio molitor*. Insect Mol. Biol. 8: 571-474.

Bureau, T. E., and S. R. Wessler, 1992 *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4: 1283-1294.

Bureau, T. E., and S. R. Wessler, 1994a Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. U S A 91: 1411-1415.

Bureau, T. E., and S. R. Wessler, 1994b *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6: 907-916.

Caizzi, R., C. Caggese and S. Pimpinelli, 1993 *Bari-1*, a new transposon-like family in *Drosophila melanogaster* with a unique heterochromatic organization. Genetics 133: 335-345.

Capy, P., T. Langin, Y. Bigot, F. Brunet, M. J. Daboussi, G. Periquet, J. R. David and D. L. Hartl, 1994 Horizontal transmission versus ancient origin: *mariner* in the witness box. Genetica 93: 161-170.

Capy, P., T. Langin, D. Higuet, P. Maurer and C. Bazin, 1997 Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100: 63-72.

Capy, P., R. Vitalis, T. Langin, D. Higuet and C. Bazin, 1996 Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J. Mol. Evol. 42: 359-368.

Casacuberta, E., J. M. Casacuberta, P. Puigdomenech and A. Monfort, 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.

Collins, J. J., and P. Anderson, 1994 The Tc5 family of transposable elements in *Caenorhabditis elegans*. Genetics 137: 771-781.

Daboussi, M. J., T. Langin and Y. Brygoo, 1992 Fot1, a new family of fungal transposable elements. Mol. Gen. Genet. 232: 12-16.

Doak, T. G., F. P. Doerder, C. L. Jahn and G. Herrick, 1994 A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. U S A 91: 942-946.

Fayet, O., P. Ramond, P. Polard, M. F. Prere and M. Chandler, 1990 Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? Mol. Microbiol. 4: 1771-1777.

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

Franz, G., and C. Savakis, 1991 *Minos*, a new transposable element from *Drosophila hydei*, is a member of the Tc1-like family of transposons. Nucleic Acids Res. 19: 6646.

Garcia-Fernandez, J., J. R. Bayascas-Ramirez, G. Marfany, A. M. Munoz-Marmol, A. Casali, J. Baguna and E. Salo, 1995 High copy number of highly similar *mariner*-like transposons in planarian (Platyhelminthe): evidence for a trans-phyla horizontal transfer. Mol. Biol. Evol. 12: 421-431.

Harris, L. J., D. L. Baillie and A. M. Rose, 1988 Sequence identity between an inverted repeat family of transposable elements in Drosophila and Caenorhabditis. Nucleic Acids Res. 16: 5991-5998.

Hartl, D. L., A. R. Lohe and E. R. Lozovskaya, 1997 Modern thoughts on an ancyent *marinere*: function, evolution, regulation. Annu. Rev. Genet. 31: 337-358.

Heierhorst, J., K. Lederis and D. Richter, 1992 Presence of a member of the Tc1-like transposon family from nematodes and Drosophila within the vasotocin gene of a primitive vertebrate, the Pacific hagfish *Eptatretus stouti*. Proc. Natl. Acad. Sci. U S A 89: 6798-6802.

Izsvak, Z., Z. Ivics, N. Shimoda, D. Mohn, H. Okamoto and P. B. Hackett, 1999 Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J. Mol. Evol. 48: 13-21.

Jacobson, J. W., M. M. Medhora and D. L. Hartl, 1986 Molecular structure of a somatically unstable transposable element in Drosophila. Proc. Natl. Acad. Sci. U S A 83: 8684-8688.

Jahn, C. L., S. Z. Doktor, J. S. Frels, J. W. Jaraczewski and M. F. Krikau, 1993 Structures of the *Euplotes crassus* Tec1 and Tec2 elements: identification of putative transposase coding regions. Gene 133: 71-78.

Jarvik, T., and K. G. Lark, 1998 Characterization of *Soymar1*, a *Mariner* element in soybean. Genetics 149: 1569-1574.

Kachroo, P., S. A. Leong and B. B. Chattoo, 1994 Pot2, an inverted repeat transposon from the rice blast fungus *Magnaporthe grisea*. Mol. Gen. Genet. 245: 339-348.

Krause, M., J. Harwood, J. Fierer and D. Guiney, 1991 Genetic analysis of homology between the virulence plasmids of *Salmonella dublin* and *Yersinia pseudotuberculosis*. Infect. Immun. 59: 1860-1863.

Kulkosky, J., K. S. Jones, R. A. Katz, J. P. Mack and A. M. Skalka, 1992 Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. Mol. Cell. Biol. 12: 2331-2338.

Langin, T., P. Capy and M. J. Daboussi, 1995 The transposable element *impala*, a fungal member of the Tc1-*mariner* superfamily. Mol. Gen. Genet. 246: 19-28.

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

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

Li, W., and J. E. Shaw, 1993 A variant Tc4 transposable element in the nematode *C. elegans* could encode a novel protein. Nucleic Acids Res. 21: 59-67.

Lohe, A. R., D. De Aguiar and D. L. Hartl, 1997 Mutations in the *mariner* transposase: the D,D(35)E consensus sequence is nonfunctional. Proc. Natl. Acad. Sci. U S A 94: 1293-1297.

Lohe, A. R., E. N. Moriyama, D. A. Lidholm and D. L. Hartl, 1995 Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. Mol. Biol. Evol. 12: 62-72.

Matsutani, S., H. Ohtsubo, Y. Maeda and E. Ohtsubo, 1987 Isolation and characterization of IS elements repeated in the bacterial chromosome. J. Mol. Biol. 196: 445-455.

Nyyssonen, E., M. Amutan, L. Enfield, J. Stubbs and N. S. Dunn-Coleman, 1996 The transposable element Tan1 of *Aspergillus niger* var. awamori, a new member of the Fot1 family. Mol. Gen. Genet. 253: 50-56.

Plasterk, R. H., Z. Izsvak and Z. Ivics, 1999 Resident aliens: the Tc1/*mariner* superfamily of transposable elements. Trends Genet. 15: 326-332.

Prasad, S. S., L. J. Harris, D. L. Baillie and A. M. Rose, 1991 Evolutionarily conserved regions in Caenorhabditis transposable elements deduced by sequence comparison. Genome 34: 6-12.

Robertson, H. M., 1996 Members of the *pogo* superfamily of DNA-mediated transposons in the human genome. Mol. Gen. Genet. 252: 761-766.

Robertson, H. M., and D. J. Lampe, 1995a Distribution of transposable elements in arthropods. Annu. Rev. Entomol. 40: 333-357.

Robertson, H. M., and D. J. Lampe, 1995b Recent horizontal transfer of a *mariner* transposable element among and between Diptera and Neuroptera. Mol. Biol. Evol. 12: 850-862.

Robertson, H. M., D. J. Lampe and E. G. Macleod, 1992 A *mariner* transposable element from a lacewing. Nucleic Acids Res. 20: 6409.

Robertson, H. M., and K. L. Zumpano, 1997 Molecular evolution of an ancient *mariner* transposon, *Hsmar1*, in the human genome. Gene 205: 203-217.

Rosenzweig, B., L. W. Liao and D. Hirsh, 1983 Sequence of the *C. elegans* transposable element Tc1. Nucleic Acids Res. 11: 4201-4209.

Smit, A. F., and A. D. Riggs, 1996 *Tiggers* and DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A 93: 1443-1448.

Surzycki, S. A., and W. R. Belknap, 2000 Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc. Natl. Acad. Sci. U S A 97: 245-249.

Swofford, D. L., 2000 PAUP*. Phylogenetic Analysis Using Parcimony (*and Other Methods)., pp. Sinauer Associates, Sunderland, Massachusetts.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey and A. Rafalski, 2000 The complete sequence of 340 kb of DNA around the rice *Adh1-adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell 12: 381-391.

Tu, Z., 1997 Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. Proc. Natl. Acad. Sci. U S A 94: 7475-7480.

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: 1313-1325.

Tudor, M., M. Lobocka, M. Goodell, J. Pettitt and K. O'hare, 1992 The *pogo* transposable element family of *Drosophila melanogaster*. Mol. Gen. Genet. 232: 126-134.

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

Vos, J. C., and R. H. Plasterk, 1994 Tc1 transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. EMBO J. 13: 6125-6132.

Wessler, S. R., 1998 Transposable elements and the evolution of gene expression. Symp. Soc. Exp. Biol. 51: 115-122.

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

Table 1. Sequences used for phylogenetic analysis

Name	Organism	Protein GI number	Position D-D-(D/E)	terminal nucleotide sequence of TIR	Reference
Gti-*mariner1*	*Girardia tigrina*	887424	151-243-278	TTAGGTTGTTCG	(Garcia-Fernandez *et al.* 1995)
Hsa-mar1	*Homo sapiens*	1263081	155-247-282	TTAGGTTGGTGC	(Robertson and Zumpano 1997)
Bmo-tpase	*Bombix mori*	2627179	151-244-278	TTAGGTCCTTAC	[b]
Cel-*mariner1*	*Caenorhabiditis elegans*	7331821	125-214-249	TTAGGCTGGTCA	[b]
Cca-*mariner*	*Ceratitis capitata*	1399036	149-241-276	TTGGATGAGTGC	[b]
Dma-*mariner*	*Drosophila mauritana*	157832	156-249-284	CCAGGTGTACAA	(Jacobson *et al.* 1986)
Cpl-*mariner*	*Chrysoperla plorabunda*	156618	158-252-287	TTAGGTTGGCTG	(Robertson *et al.* 1992)
Dme-*pogoR11*	*Drosophila melanogaster*	2133672	169-272-303	CAGTATAATTCG	(Tudor *et al.* 1992)
Hsa-*Tigger1*	*Homo sapiens*	2226004	187-302-335	CAGGCATACCTC	(Robertson 1996, Smit and Riggs 1996)
Hsa-*Tigger2*	*Homo sapiens*	7513081	11-124-157	CAGTTGACCCTT	
Ath-*Emigrant*	*Arabidopsis thaliana*	4262216	81-184-206	CAGTAAAACCTC	(Casacuberta *et al.* 1998, Feschotte and Mouches 2000a)
Cel-Tc4	*Caenorhabiditis elegans*	156453	261-374-412	CTAGGGAATGAC	(Li and Shaw 1993)
Cel-Tc5	*Caenorhabiditis elegans*	515796	257-368-406	CAAGGGAAGGTT	(Collins and Anderson 1994)
Ani-Tan1	*aspergillus niger*	1805251	175-289-325	ACGTAATCAACG	(Amutan *et al.* 1996, Nyyssonen *et al.* 1996)
Fox-Fot1	*Fusarium oxysporum*	2723	158-275-311	AGTCAAGCACCC	(Daboussi *et al.* 1992)
Pgr-Pot2	*Pyricularia grisea*	496854	160-273-309	TAACGTTGCGTA	(Kachroo *et al.* 1994)

Gma-*Soymar1*	*Glycine max*	7488706	163-287-327	CTCCCTNGTTTC	(Jarvik and Lark 1998)
Osa-*Stow1*	*Oryza sativa*	6979318[a]		CTCCCTCCGTTT	(Tarchini *et al.* 2000)
Osa-*Stow2*	*Oryza sativa*	6539555	160-283-318	CTTCCTCMGTCC	[b]
Cbr-Tcb1	*Caenorhabiditis briggsae*	84417	87-177-212	CAGTACTGGCCA	(Harris *et al.* 1988)
Cbr-Tcb2	*Caenorhabiditis briggsae*	156455	87-177-212	CAGTACTGGCCA	(Prasad *et al.* 1991)
Cel-Tc1	*Caenorhabiditis elegans*	6888	87-177-212	CAGTGCTGGCCA	(Rosenzweig *et al.* 1983)
Aal-*Quetzal*	*Anopheles albimanus*	1196520	154-245-280	CACTTCTCACAA	[b]
Est-Tes1	*Eptatretus stouti*	422537	139-269-308	(TA)CTCTACCGGTCG	(Heierhorst *et al.* 1992)
Dme-*Bari1*	*Drosophila melanogaster*	7641	154-243-278	CAGTCATGGTCA	(Caizzi *et al.* 1993)
Dhi-*Minos2*	*Drosophila hidei*	436464	175-265-300	CGCTTAACTTAA	(Franz and Savakis 1991)
Fox-*Impala*	*Fusarium oxysporum*	10764493	149-234-270	CAGTGGGGTGCA	(Langin *et al.* 1995)
Cel-Tc3	*Caenorhabiditis elegans*	283537	144-231-266	TACAGTGTGGGA	(Langin *et al.* 1995)
Ecr-Tec1	*Euplotes crassus*	397762	230-316-351	TAGAGGGAGTTT	(Jahn *et al.* 1993)
Ecr-Tec2	*Euplotes crassus*	1216295	231-317-352	TAGAGGGATAAA	(Jahn *et al.* 1993)
Sso-IS*630*	*Shigella sonnei*	47542	181-261-297	AATAGCTGCGCG	(Matsutani *et al.* 1987)
Sth-IS*630*	*Salmonella thyphimurium*	154216	181-261-297	(A)AATAGCTTCGC	(Krause *et al.* 1991)
Zma-EST1	*Zea mays*	6102034[a]			
Zma-EST2	*Zea mays*	6227252[a]			
Zma-EST3	*Zea mays*	5922112[a]			
Tae-EST1	*Triticum aestivum*	9424806[a]			
Tae-EST2	*Triticum aestivum*	9424809[a]			

[a] This GI number refers to a nucleotide sequence. The amino acid sequence have been obtained using ORF FINDER or TBLASTN, as described in Material and Methods.

[b] The terminal nucleotide sequence of these elements are deduced from our sequence analysis.

Table 2. Percent differences between DD(D/E) domains [a]

	mariner	*pogo*	Tc1	*Emigrant*	*Stowaway*-like [b]
mariner	41	75	70	74	66
	(11)	(6)	(4)	(5)	(4)
pogo		61	79	60	75
		(12)	(4)	(9)	(6)
Tc1			54	79	69
			(15)	(2)	(3)
Emigrant				0	79
				(0)	(2)
Stowaway-like					28
					(4)

[a] Within group average and standard errors (in parentheses) are calculated from the mean pairwise distances adjusted for missing data (PAUP).

[b] *Stowaway*-like includes the sequences of Gma-*Soymar1*, Osa-*Stow1*, and Osa-*Stow2*.

Figure 1. Amino acid sequence alignment corresponding to the DD(D/E) motif of the transposase of members of the IS*630*/Tc1/*mariner* superfamily of transposable elements. The last five sequences correspond to EST sequences. GI numbers and amino acid positions are given in Table 1. The conserved D, D and D/E residues are indicated with an arrow.

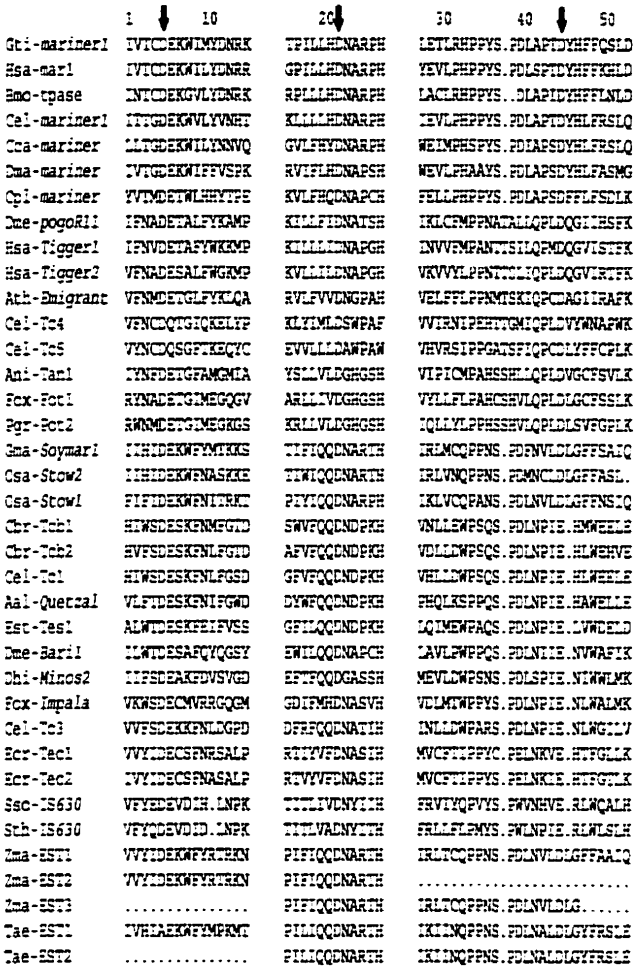

1 10 20 30 40 50
Gti-mariner1 IVTCDEKWIMYDNRK TPILLHDNARPH LETLRHPPYS.PDLAPTDYHFFQSLD
Hsa-mar1 IVTCDEKWILYDNRR GPILLHDNARPH YEVLPHPPYS.PDLSPTDYHFFKGLD
Bmo-tpase INTCDEKGVLYDNRK RPLLLHDNARPH LACLRHPPYS..DLAPIDYHFFLNLD
Cel-mariner1 ITTGDEKGVLYVNHT KLLLLHDNARPH IEVLPHPPYS.PDLAPTDYHLFRSLQ
Cca-mariner LITGDEKWILYNNVQ GVLFHYDNARPH WEIMPHSPYS.PDIAPSDYHLFRSLQ
Dma-mariner IVTGDEKWIFFVSPK RVIFLHDNAPSH WEVLPHAAYS.PDLAPSDYHLFASMG
Cpl-mariner YVTMDETWLHHYTPE KVLFHQDNAPCH FELLPHPPYS.PDLAPSDFFLFSDLK
Dme-pogoRI1 IFNADETALFYKAMP KILLFIDNATSH IKLCPMPPNATALLQPLDQGIIHSFK
Hsa-Tigger1 IFNVDETAFYWKKMP KILLLIDNAPGH INVVPMPANTTSILQPMDQGVISTFK
Hsa-Tigger2 VFNADESALFWGKMP KVLLILDNAPGH VKVVYLPPNTTCLIQPLDQGVIRTFK
Ath-Emigrant VFNMDETGLFYKLQA RVLFVVDNGPAH VELFFLPPNMTSKIQPCDAGIIRAFK
Cel-Tc4 VFNCDQTGIQRELYP KLYIMLDSWPAF VVIRNIPEHTTGMIQPLDVYWNAPWK
Cel-Tc5 VYNCDQSGFTKEQYC EVVLLLDAWPAW VHVRSIPPGATSFIQPCDLYFFCPLK
Ani-Tan1 IYNFDETGFAMGMIA YSLLVLDGHGSH VIPICMPAHSSHLLQPLDVGCFSVLK
Fcx-Fot1 KYNADETGIMEGQGV ARLLIVDGHGSH VYLLFLPAHCSHVLQPLDLGCFSSLK
Pgr-Pot2 KWNMDETGIMEGKGS KRLLVLDGHGSH IQLLYLPPHSSHVLQPLDLSVFGPLK
Sma-Soymar1 IIHIDEKWFYMTKKS TIFIQQDNARTH IRLMCQPPNS.PDFNVLDLGFFSAIQ
Osa-Stow2 IIHIDEKWFNASKKE TIWIQQDNARTH IRLVNQPPNS.PDMNCLDLGFFASL.
Osa-Stow1 FIFIDEKWFNITRKT PIYIQQDNARPH IKLVCQPANS.PDLNVLDLGFFNSIQ
Cbr-Tcb1 HIWSDESKFNMFGTD SWVFQQDNDPKH VNLLEWPSQS.PDLNPIE.EMWEELE
Cbr-Tcb2 EVFSDESKFNLFGTD AFVFQQDNDPKH VDLLDWPSQS.PDLNPIE.HLWEHVE
Cel-Tc1 HIWSDESKFNLFGSD GFVFQQDNDPKH VELLDWPSQS.PDLNPIE.HLWEELE
Aal-Quetzal VLFTDESKFNIFGWD DYWFQQDNDPKH PHQLKSPPQS.PDLNPIE.HAWELLE
Est-Tes1 ALWTDESKFEIFVSS GFILQQDNDPKH LQIMEWPAQS.PDLNPIE.LVWDELD
Dme-Bari1 ILWTDESAFQYQGSY EWILQQDNAPCH LAVLPWPPQS.PDLNIIE.NVWAFIK
Dhi-Minos2 IIFSDEAKFDVSVGD EFTFQQDGASSH MEVLDWPSNS.PDLSPIE.NIWWLMK
Fcx-Impala VKWSDECMVRRGQGM GDIFMHDNASVH VDLMTWPPYS.PDLNPIE.NLWALMK
Cel-Tc3 VVFSDEKKFNLDGPD DFRFQQDNATIH INLLDWPARS.PDLNPIE.NLWGILV
Ecr-Tec1 VVYIDECSFNHSALP RTIVVFDNASIH MVCFTIPPYC.PELNKVE.HIFGLLK
Ecr-Tec2 IVYIDECSFNASALP RTVVVFDNASIH MVCFTIPPYS.PELNKIE.HIFGTLK
Ssc-IS630 VFYEDEVDIH.LNPK TITLIVDNYIIH FRVIYQPVYS.PWVNHVE.RLWQALH
Sth-IS630 VFYQDEVDID.LNPK TITLVADNYITH FRLLFLPMYS.PWLNPIE.RLWLSLH
Zma-EST1 VVYIDEKWFYRTRKN PIFIQQDNARTH IRLTCQPPNS.PDLNVLDLGFFAAIQ
Zma-EST2 VVYIDEKWFYRTRKN PIFIQQDNARTH .....
Zma-EST3 ..... PIFIQQDNARTH IRLTCQPPNS.PDLNVLDLG.....
Tae-EST1 IVHIAEKWFYMPKMT PILIQQDNARTH IKIINQPPNS.PDLNALDLGYFRSLE
Tae-EST2 ..... PILIQQDNARTH IKIINQPPNS.PDLNALDLGYFRSLE

Figure 2. Phylogenetic tree of members of the IS*630*/Tc1/*mariner* superfamily, inferred with maximum parsimony method with the alignment shown in Figure 1 (excluding ESTs). The majority rule consensus tree is shown, derived from 30 equally parsimonious trees of 525 steps (CI=0.636; RI=0.671). Bootstrap values are indicated below the corresponding nodes. The clades corresponding to families of transposable elements are indicated. Dashed lines represent clades that collapse in the strict consensus tree. In the bootstrap tree, the DDD sequences form a monophyletic group.

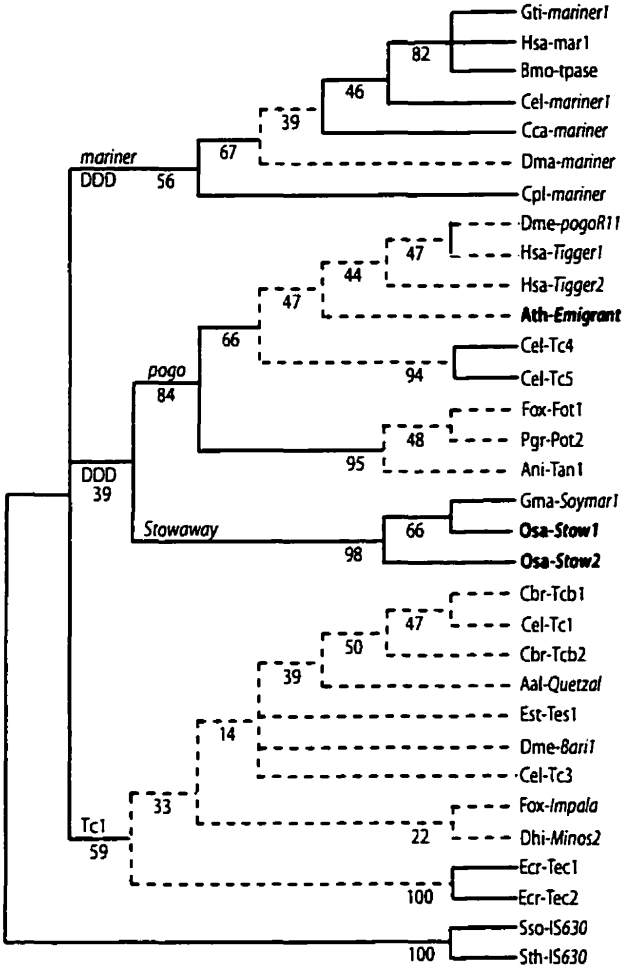

Gti-mariner1
Hsa-mar1
Bmo-tpase
Cel-mariner1
Cca-mariner
Dma-mariner
Cpl-mariner
Dme-pogoR11
Hsa-Tigger1
Hsa-Tigger2
Ath-Emigrant
Cel-Tc4
Cel-Tc5
Fox-Fot1
Pgr-Pot2
Ani-Tan1
Gma-Soymar1
Osa-Stow1
Osa-Stow2
Cbr-Tcb1
Cel-Tc1
Cbr-Tcb2
Aal-Quetzal
Est-Tes1
Dme-Bari1
Cel-Tc3
Fox-Impala
Dhi-Minos2
Ecr-Tec1
Ecr-Tec2
Sso-IS630
Sth-IS630
82
46
39
67
mariner
DDD
56
47
44
47
66
pogo
84
94
48
95
DDD
39
Stowaway
66
98
47
50
39
14
33
Tc1
59
22
100
100

Figure 3. Phylogenetic tree of members of the IS*630*/Tc1/*mariner* superfamily, inferred with neighbour joining method with the alignment shown in Figure 1 (excluding ESTs). The tree score is 7,15. Branch lengths are proportional to the estimated number of substitutions per site. Bootstrap values are indicated below the corresponding nodes. The clades corresponding to families of transposable elements are indicated.

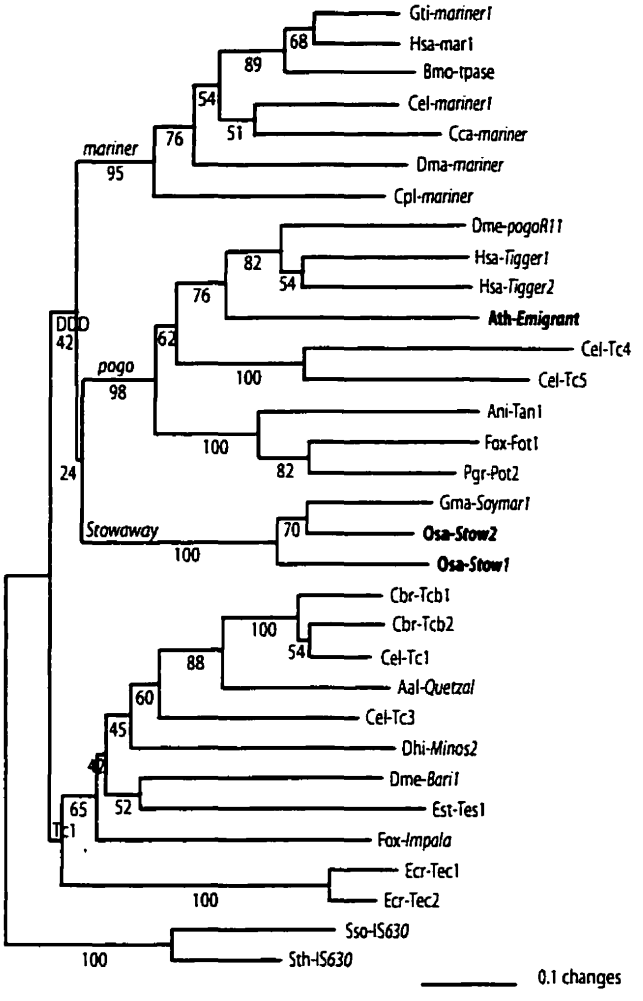
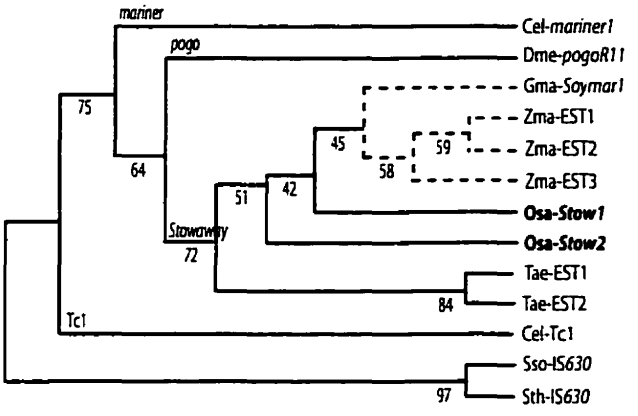

Gti-mariner1
Hsa-mar1
Bmo-tpase
Cel-mariner1
Cca-mariner
Dma-mariner
Cpl-mariner
mariner
Dme-pogoR11
Hsa-Tigger1
Hsa-Tigger2
Ath-Emigrant
Cel-Tc4
Cel-Tc5
pogo
Ani-Tan1
Fox-Fot1
Pgr-Pot2
Gma-Soymar1
Osa-Stow2
Osa-Stow1
Stowaway
Cbr-Tcb1
Cbr-Tcb2
Cel-Tc1
Aal-Quetzal
Cel-Tc3
Dhi-Minos2
Dme-Bari1
Est-Tes1
Fox-Impala
Tc1
Ecr-Tec1
Ecr-Tec2
Sso-IS630
Sth-IS630
0.1 changes

Figure 4. Phylogenetic tree of members of the *Stowaway* family, including EST sequences and one representative of the Tc1, *mariner* and *pogo* families. The tree was inferred with maximum parsimony method. The majority rule consensus tree is shown, derived from 5 equally parsimonious trees of 167 steps (CI=0.910; RI=0.810). Bootstrap values are indicated below the corresponding nodes. The clades corresponding to families of transposable elements are indicated. Dashed lines represent clades that collapse in the strict consensus tree.


mariner
pogo
Stowaway
Tc1
75
64
72
51
42
45
58
59
84
97
Cel-mariner1
Dme-pogoR11
Gma-Soymar1
Zma-EST1
Zma-EST2
Zma-EST3
Osa-Stow1
Osa-Stow2
Tae-EST1
Tae-EST2
Cel-Tc1
Sso-IS630
Sth-IS630

CHAPTER 4

Transposon mobility on evolutionary time scales: utility of transposable elements in phylogenetic analysis of the AA genome species of rice

Title: **Transposon Mobility on Evolutionary Time Scales: Utility of Transposable Elements in Phylogenetic Analysis of the AA Genome Species of Rice**

Authors: **Kime Turcotte, Thomas Bureau, and Louise S. O'Donoughue†**

Address: Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec, Canada, H3A 1B1

†DNA Landmarks Inc., 84 Richelieu, St-Jean-sur-Richelieu, Quebec, Canada, J3B 6X3

Author for correspondence: Thomas Bureau, Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec, Canada, H3A 1B1

Tel: (514) 398-6472, Fax: (514) 398-5069, Email: thomas_bureau@maclan.mcgill.ca

Author's email addresses:

Kime Turcotte: kturco2@po-box.mcgill.ca

Louise O'Donoughue: odonoughuel@dnalandmarks.ca

Thomas Bureau: thomas_bureau@maclan.mcgill.ca

Abstract

Asian and African cultivated rice (*Oryza sativa* and *O. glaberrima*) belong to the *O.sativa* complex that groups diploid rice species with the AA genome type. The genome of *O. sativa* is known to contain various types of transposable elements. In this study, we examined transposable elements at ten *O. sativa* loci, and determined their presence/ absence in eight species of the *O. sativa* complex. We also used transposon nucleotide sequences for phylogenetic analyses of the rice species. Miniature inverted-repeat transposable elements (MITEs), *Ac*-like elements, *Mutator*-like elements (MULEs), short interspersed nuclear elements (SINEs) and three unclassified elements were investigated. We found that generally, the sequences of a single element do not provide sufficient phylogenetic signal to reconstruct the relationships with good resolution. However, the use of a combination of transposable elements, and the corresponding insertion polymorphism information results in robust cladograms of the rice species. These trees agree with previous analyses of the relationships among rice species, and support the hypothesis of domestication of *O. sativa* from *O. nivara*, and *O. glaberrima* from *O. barthii*. In addition, the insertion of a MITE and two MULEs shed light on the position of *O. longistaminata* at a basal position relative to both types of cultivated rice.

Introduction

Virtually all eukaryotic and prokaryotic genomes have been found to harbor transposable elements (TEs). These ubiquitous genetic elements differ from the gene complement of genomes in that they are mobile. Movement can be mediated by an RNA intermediate and be essentially replicative (e.g. class I TEs) or directly through a DNA form (e.g., class II TEs). Class I elements include endogenous retroviruses, LTR-retrotransposons, non-LTR-retrotransposons (also referred to as long interspersed nuclear elements or LINEs), short interspersed nuclear elements (e.g., SINEs), and processed pseudogenes. In contrast to class I elements, class II elements can excise generating a so-called footprint, which may include terminal element sequences and/or the target site duplication (TSD). Structurally, class II elements are delimited by terminal-inverted repeats (TIRs). Although historically referred to as junk or parasitic DNA as they commonly populate non-genic regions of the genome, a new view of TEs spurred by the wealth of genome sequencing information has emerged suggesting that they are involved in adaptation and evolution at the gene, genome, population and species levels (Capy *et al.* 2000; Kidwell and Lisch 1997; Makalowski 2000). Transposon copy number may be very high, as in the maize genome where at least 50% of the genome is composed of class I elements (Sanmiguel *et al.* 1996). TEs are less abundant in other genomes, as in *Arabidopsis thaliana*, where all TEs together account for only approximately 5% (Le *et al.* 2000; Meyerowitz 1994) to 10% (Initiative 2000) of the genome. Generally, eukaryotes with smaller genomes contain less repetitive sequences and/or fewer large-TE insertions, such as retroelements.

Cultivated rice (*Oryza sativa*) is a model genome for monocotyledonous plants. Its small genome is currently being completely sequenced, and the availability of genomic sequences has facilitated computer-based surveys of TEs (Bureau *et al.* 1996; Mao *et al.* 2000; Song *et al.* 1998; Tarchini *et al.* 2000; Turcotte *et al.* 2001). These surveys have revealed a diverse number of TEs including many novel elements. In a high-resolution survey, miniature inverted-repeat transposable elements (MITEs) were found to make up 72% of the rice TE insertions mined (Turcotte *et al.* 2001). Other class II TEs such as *Ac*-like, CACTA-like and *Mutator*-like elements (MULEs) and class I TEs such as Ty1-

copia and Ty3-*gypsy*-like retrotransposons, LINEs and SINEs were identified. The element referred to as *Basho* (Le *et al.* 2000) was also identified in rice. Overall, TEs mined in *O. sativa* accounted for approximately 20% of the nucleotide sequences surveyed.

The genus Oryza belongs to the family Gramineae and has a worldwide distribution. Two Oryza species, *O. sativa* (Asian) and *O. glaberrima* (African) have been domesticated. In addition, there are 21 known wild species of Oryza (Khush 1997). These species are often grouped into four complexes based on interspecific hybridization aptitude (Khush 1997). Species in the *O. sativa* complex are diploid (2n=24) and correspond to the genome type AA. The *O. officinalis* complex is made up of diploid and tetraploid species, with genome types BB, BBCC, CC, CCDD, EE, and FF. The *O. meyeriana* complex includes diploid species of genome type GG, while species from the *O. ridleyi* complex are tetraploid of the genome type HHJJ. Genome types were chiefly defined by genomic DNA hybridization and chromosome pairing behavior of hybrids (Aggarwal *et al.* 1997; Khush 1997).

Cultivated rice was most likely domesticated as two separate events. *O. sativa* was probably domesticated from an annual type of *O. rufipogon* called *O. nivara*. Whereas, *O. glaberrima* was probably domesticated from *O. barthii* (=*O. breviligulata*), which originated from the African species *O. longistaminata*. Relationships between cultivated and wild rice species were originally studied based on their morphology, physiology and crossing ability (Morishima *et al.* 1963; Morishima and Oka 1960; Oka 1974; Oka 1988). More recently, relationships between cultivated and wild rice species have been investigated using many different modern methods. For example, isozymes (Second 1982; Second 1985), RFLP from chloroplast DNA (Dally and Second 1990; Ishii *et al.* 1988) and nuclear DNA (Wang *et al.* 1992), RAPD (Ishii *et al.* 1996; Yi *et al.* 1995), ribosomal DNA spacer length polymorphism (Cordesse *et al.* 1993; Sano and Sano 1990), and nuclear and chloroplastic gene sequences (Ge *et al.* 1999) have been exploited. Transposable elements have also been used in phylogenetic studies in rice. The polymorphisms generated by the presence or absence of SINEs (Hirano *et al.* 1994) and

Stowaway-like MITEs (Kanazawa *et al.* 2000) at a given locus have been superimposed on previously published trees, insertion polymorphisms of SINEs and other TEs have been used to classify rice strains into groups (Mochizuki *et al.* 1993) and the nucleotide sequences of *Tourist-like* MITE insertions (Iwamoto *et al.* 1999) have been used in phylogenetic analysis in Oryza. Here we compare different types of TEs as phylogenetic markers in rice. Specifically, we have examined *Tourist*-like MITEs, an *Ac*-like element, MULEs, SINEs, and unclassified TEs. The objective of our study is first to evaluate various TEs as phylogenetic markers, and then to determine if such markers can contribute to the resolution of phylogenetic relationships in Oryza.

Materials and Methods

Plant material

Germplasm of eight diploid, AA genome Oryza species was obtained from the International Rice Research Institute (IRRI, Los Banos, Philippines). One to three accessions per species for a total of 22 accessions were analyzed. The species accession numbers are listed in Table 1. Genomic DNA was extracted from frozen leaves according to Dellaporta (Dellaporta *et al.* 1983; Ishii *et al.* 1996), with the exception that precipitation was done with isopropanol instead of ethanol.

DNA amplification and sequencing

The transposable elements and the corresponding loci in *O. sativa* are listed in Table 2. Primers were designed using the program OLIGO6 (Rychlik and Rhoades 1989) from genomic sequences flanking the TEs (Table 2, GenBank: http://www.ncbi.nlm.nih.gov/). Primer pairs were designed to amplify fragments (≈1 kb) suitable for bi-directional direct sequencing. PCR amplification was performed in 50 μL reactions, with 100 ng total DNA, 2.0mM $MgCl_2$, 0.3 to 0.4 μM primers, 0.8mM dNTP, and 2 units AMPLITAQ DNA polymerase (Applied Biosystems). Amplification was carried out in GENEAMP PCR system 9700 (Applied Biosystems) with the following program: initial denaturation at 94 for 3 minutes, then 30 to 40 cycles of denaturation at 94°C for 1 minute, annealing for 1 minute, elongation at 72°C for 1 minute, and a final elongation step of 7 minutes at 72°C. The annealing temperature was 60°C for T1A-1 and T8R-2 (30 seconds), 65°C for AC1-1, 51°C for MU9-2, 50°C for MU9-3, UNC2-2, and SINE1-1, 54°C for SINE1-4 and UNC1-1, and 57°C for CRA-1. Amplification was verified by agarose (2% gels) electrophoresis. Purification of PCR products was done using a QIAQUICK PCR purification kit (Qiagen). In some cases, the amplification was not specific enough, and more than one band was observed. The bands corresponding to the size of a fragment containing the TE insertion were gel purified using a QIAQUICK gel extraction kit (Qiagen). Purified PCR products were directly sequenced (both strands) using Dye-terminator chemistry (BIG DYE) on an ABI prism DNA analyzer model 3700 (Applied Biosystems). In some cases, PCR products were ligated in a PMO vector, transformed

into competent cells (XL1-blue) by heat shock, purified using QUIAPREP miniprep plasmid purification kit (Qiagen), and sequenced on a LI-COR DNA sequencer model 4200 using Dye-primer chemistry using universal M13 primers.

Sequence and phylogenetic analyses

Sequences were edited and assembled using the STADEN software package on a Linux workstation (Staden *et al.* 2000). Sequence alignment of the resulting consensus sequences was performed using CLUSTALW (Thompson *et al.* 1994) implemented with BIOEDIT version 4.8.10 (Hall 1999), with a gap creation penalty of 5 to 10 and a gap extension penalty of 0.2. Alignments were corrected manually when necessary. Because of their phylogenetic information, indels were considered as characters, and manually added to the data matrix for each data set. When an indel of variable length was encountered, different character states were employed to describe the situation. In individual TE matrices, all characters were unordered and of equal weight, without regards to the size of the indel events. A few sequences showed nucleotide heterozygozity; the corresponding sites and any other position for which homology was uncertain were eliminated from phylogenetic analyses. Each variable site has been verified with the original sequences. Phylogenetic reconstructions were performed using PAUP* version 4.0 beta 8 for the Macintosh (Swofford 2000), with maximum parsimony (MP; multiple trees heuristic search option with tree bisection-reconnection) and the neighbor-joining (NJ) distance methods, using Jukes-Cantor distance measure. Unrooted trees were displayed using the mid-point rooting option. Analyses were performed on each data set separately, using the TE sequence of the accessions that possess the insertion. Matrices containing variable characters from several loci combined with insertion polymorphisms information have also been used for phylogenetic reconstructions. Predictions of the secondary structure were obtained from DNA MFOLD server (http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi) with the free energy values calculated following SantaLucia (Santalucia 1998).

Results

We have amplified and sequenced TEs from ten loci in eight rice species (22 accessions) of the *O. sativa* complex. Occasionally, no PCR product was obtained (Table 1). Failure to amplify was most likely due to mutations that made primer annealing impossible. The absence of an amplicon was interpreted as missing information, rather than absence of the TE. Most amplifications resulted in a single band on agarose gels corresponding to the expected size of either a TE insertion (+) or an empty site (-) (Table1). Some species or accessions were clearly heterozygous as two bands of the expected sizes were observed. Generally, the TE insertions investigated defined monomorphic traits for most species, as all accessions gave the same banding pattern within a species. Nevertheless, accessions of *O. meridionalis* showed intra species polymorphisms at several loci, where accession 105283 invariably differed from accessions 101146 and 105281 (Table 1). Accession polymorphism has also been obtained in several species with SINE1-4 primers.

Nucleotide sequences were easily aligned due to a generally high level of sequence identity (74% to 96%). In fact, most positions of the alignments were constant through the 22 accessions. A few gaps were introduced to maintain alignment of homologous positions, and the resulting indels were treated as characters. The number of parsimony informative sites has been calculated for each data set (Table 3) because phylogenetic analyses are based on variable nucleotides while constant sites are uninformative. The trees reconstructed using maximum parsimony (MP) and neighbor joining (NJ) methods using data sets from individual TEs are shown in figure 1 and 2. Figure 3 shows the cladogram obtained with combined matrices.

MITEs

Two MITEs were included in our experiment. These elements belong to the *Tourist*-like family. T1A-1 and T8R-2 have a 3 bp TSD, are located within introns, and are delimited by 12 and 14 bp TIR sequences (Table 2). Previous studies of *A. thaliana* and *C. elegans* have associated *Tourist* elements with members of the IS*5* family of bacterial TEs, based on amino acid similarity in their transposases (Le *et al.* 2001; Le *et al.* 2000). In our

study, T1A-1 (Bureau and Wessler 1994), a member of the *Tourist* I group of rice TEs (Turcotte *et al.* 2001), could be amplified in most species. In *O. meridionalis*, accessions 101146 and 105281 have an empty site (-), while accession 105283 does have the insertion (+). Also, a product of unexpected size was obtained in *O. longistaminata*. Amplification with the three accessions gave a fragment shorter than the size expected for a *Tourist* insertion in the T1A-1 locus, but larger than the corresponding empty site. Nucleotide sequence alignment of T1A-1 sequences shed light on this unexpected polymorphism. This *Tourist*, amplified from intron 5 of the *phy18* gene, is absent from the three *O. longistaminata* accessions. These lineages however have a 149 bp insertion located 283 bp downstream of the *Tourist* insertion site (position 8682 in *O. sativa* GI:20287). This insertion corresponds to a *Stowaway* element with typical *Stowaway* TIRs and the dinucleotide TA as a TSD (data not shown). The locus of T1A-1 contained four other indels, of which one is located within the *Tourist* sequence. This indel appears to be a 20 bp deletion in the 3'TIR of T1A-1, in five of the six *O. sativa* accessions. Although other species have presumably complete TIR sequences, these are still imperfect inverted repeats. T1A-1 has very few parsimony informative variable sites because the sequences of the different species are highly similar. The T1A-1 sequences comprise 338 characters, of which 323 are invariable. Only 9 characters (including the indel character) were parsimony informative (Table 3). The sequences of this element were actually less variable than the sequences flanking the insertion (intron 5 from *phy 18*), which had 23 parsimony informative characters among the 399 characters observed (data not shown). Consequently the MP strict consensus tree (Figure 1a) derived from T1A-1 is poorly resolved. All accessions of *O. nivara* form a branch, and *O. barthii* and *O. glaberrima* accessions combined form another branch. Surprisingly, *O. meridionalis* accession 105283 appeared more closely related to *O. sativa* accession 611. NJ method yielded a tree with a slightly better resolution (Figure 2a), where *O. sativa* is polyphyletic. A cluster formed by *O. sativa* accession 611, *O. rufipogon* and *O. meridionalis* accession 105283 is more closely related to the African rice species than they are to the other *O. sativa* accessions.

The other MITE, T8R-2, was absent from *O. rufipogon*, *O. glumaepatula*, and *O. longistaminata*. Slippage suring sequencing within a short poly A track in the 5' flanking sequence of T8R-2 was observed. Therefore, although an amplification product was obtained, accurate sequences of *O. sativa* accessions 636 and 611, *O. nivara* accession 104684, *O. glaberrima* accession 100139, and *O. barthii* accession 104290 could not be obtained. Nucleotide sequence alignment of the T8R-2 sequences revealed two indels and only six parsimony informative characters. Consequently, the MP and NJ trees (figure 1b and 2b), although they are congruent with each other, show poor resolution. Out of the six parsimony informative characters available, four differentiate *O. nivara* from all other species, resulting in an unusual rooting.

Ac-like

AC1-1 belongs to the group I of *Activator*-like elements identified in rice (Turcotte *et al.* 2001). AC1-1 is delimited by 24 bp TIR sequences and is flanked by a 8 bp TSD. AC1-1 is also located in an intron. This TE could not be amplified from accessions 101146 and 105281 of *O. meridionalis*. Four accessions were not included in the phylogenetic analyses because sequencing could not be achieved with sufficient accuracy. All accessions for which sequence information could be obtained, contained the *Ac*-like element. Seven indels were revealed by the alignment of the remaining 16 AC1-1 sequences. Two equally parsimonious trees of 82 steps were obtained with MP method. The strict consensus tree (Figure 1c) is well resolved and perfectly congruent with the NJ tree (Figure 2c), although the latter has better resolution. Again, one *O. sativa* accession forms a cluster with *O. rufipogon* and accession 105283 of *O. meridionalis*, while the other *O. sativa* accessions form a distinct branch. In contrast to the T1A-1 NJ tree, the *O. sativa/ O. rufipogon/ O. meridionalis* cluster appears more closely related to the other *O. sativa* accessions than to the *O. glaberrima/ O. barthii* group.

Mutator-like (MULE)

MU9-2 and MU9-3 are two members of the MULE IX group of rice TEs (Turcotte *et al.* 2001). MULEs are class II transposons that are structurally similar to the maize *Mutator* element. The two copies selected are short non-autonomous members from different

intergenic loci. These elements have ≈250 bp TIRs and a 9 bp TSD, typical to this family. Their TIRs are very long compared to the size of the element (561 bp), such that a hairpin structure can potentially be formed. In fact, a free energy value of dG = –195.8 kcal/mole have been predicted for MU9-2 and MU9-3 DNA sequences, using the MFOLD program (http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi). This secondary structure may be responsible for the difficulties encountered with direct sequencing of the amplified products, where sequences had half the expected length. It appears that one TIR was normally sequenced and the sequencing reactions prematurely terminated after that segment. The same results were obtained while sequencing either strand. Cloning of these elements and sequencing using an alternative chemistry (see Materials and Methods) were also unsuccessful. Therefore, we could not derive trees from the nucleotide sequence alignments of these MULEs, but the insertion polymorphisms have been taken into account in the combined data matrices (Figure 3). These elements could be amplified from most accessions, and empty sites were found in *O. longistaminata* and *O. meridionalis* accessions 101146 and 105281.

p-SINE

SINE1-1 and SINE1-4 are two p-SINE1 members, previously identified and characterized (Hirano *et al.* 1994; Mochizuki *et al.* 1993; Mochizuki *et al.* 1992). These two very short elements (124 bp and 128 bp) are terminated with a poly-A tail, and are flanked by a 15 bp TSD. Previous analysis of their distribution among rice species has revealed their presence in all AA genome species (Mochizuki *et al.* 1993). Consequently, we could amplify SINE1-1 from all species, but no amplification product was obtained from accessions 101146 and 105281 of *O. meridionalis*. The sequences of SINE1-1 contained two indels. A unique most parsimonious tree of 30 steps was derived from these sequences (Figure 1d). The tree poorly resolves the species relationships, but performed well on grouping the accessions into species, with the exception of the *O. sativa/ O. rufipogon/ O. meridionalis* clustering. Again, the *O. sativa* accessions were polyphyletic. The NJ method (Figure 2d) grouped the same taxa into branches, but the branching order is a little clearer than in the MP tree. Again, *O. barthii and O. glaberrima* form a distinct

branch in both trees. Surprisingly, four accessions of *O. sativa* appear at a basal position in the NJ tree, but the length of this branch is very short (>0.1 substitution/site).

SINE1-4 could be amplified from all accessions. *O. longistaminata* was polymorphic with accessions 101198 and 101205 appearing as heterozygotes while accession 81968 had only one amplicon. Amplification with the SINE1-4 primers yielded additional polymorphisms due to a 166 bp insertion in several accessions. As previously reported (Mochizuki *et al.* 1993), the insertion corresponds to a MITE (e.g. Tnr2) within the targeted p-SINE element. The nested MITE was found in two *O. nivara* accessions, *O. glumaepatula*, *O. glaberrima*, two *O. longistaminata* accessions, *O. barthii*, and two *O. meridionalis* accessions. Tnr2 was also found in one *O. sativa* accession, while absent in the five other accessions. A total of five indels (other than the Tnr2 insertion) were found in the amplified sequences. Four were associated with Tnr2 and one was found in SINE1-4. The sequences of SINE1-4 (excluding the sequence of Tnr2) yielded three equally most parsimonious trees with 39 steps. The poor resolution of the MP consensus tree (Figure 1e) is most likely due to the uneven distribution of the characters among the taxa. Ten characters support the basal dichotomy, and six characters support the monophyletic group of sequences lacking Tnr2. The remaining characters cannot provide much more resolution. Although the character corresponding to presence/absence of Tnr2 carries a weight of 1, the trees obtained with SINE1-4 clearly divided the accessions of *O. sativa*, *O. meridionalis* and *O. nivara* into two groups, according to the Tnr2 insertion, resulting in an unexpected topology where most *O.sativa* accessions are close relatives of *O.longistaminata* and *O.meridionalis*. A similar topology was generated using NJ method (Figure 2e).

Although the sequences of SINE1-1 and SINE1-4 have a number of parsimony informative sites similar to the sequences of AC1-1, CRA-1, UNC1-1, and UNC2-2 (Table 3), the length of the SINEs is considerably shorter than any other TE sequences investigated. SINE1-1 and SINE1-4 are therefore the most variable markers in our study, with respectively 15.9% and 20.2% of their characters being parsimony informative.

Unclassified rice transposons

CRA-1 is similar to the previously reported rice *crackle* element (Song *et al.* 1998). This element has no apparent TIRs, does contain subterminal inverted repeats (SIRs) of 91 bp, and has a 9 bp TSD. These characteristics are reminiscent of non-TIR MULEs but the actual classification of *crackle* is still unknown. CRA-1 could not be amplified from accession 101198 and 101205 of *O. longistaminata*, and accessions 101146 and 105281 of *O. meridionalis*. Although a product has been amplified in accession 81968 of *O.longistaminata*, it could not be accurately sequenced. The elements were variable in length because of 5 to 48 bp indels shared by diverse species. Four indels were added to our matrix as new characters. Thirteen equally parsimonious trees of 39 steps were obtained. Although *O. sativa* accessions were grouped together, accession 105283 of *O. meridionalis* and *O. rufipogon* still form a cluster. In both the MP and NJ trees, the *O. nivara* accessions form a branch sister to the *O. sativa* group. The other species were clustered diversely in the 13 most parsimonious trees, so that the strict consensus MP tree (figure 1f) shows much less resolution than the NJ tree (Figure 2f)

UNC1-1 has been described as unclassified-I (Turcotte *et al.* 2001) with no apparent TIR, LTR or poly A structure, and a TSD of 9-11 bp. We could determine that UNC1-1 contains SIRs that are very conserved among members of this group, while the termini of these elements are more variable. Like *crackle*, these features are reminiscent of non-TIR MULEs. Amplification of UNC1-1 from all 22 accessions investigated was possible. Seven indels present within the sequence of this element were added to the matrix. 23 parsimony informative characters were determined. A unique most parsimonious tree of 40 steps was obtained. Except for the position of *O.meridionalis* accession 105283, the MP tree (Figure 1g) is concordant with the NJ tree (Figure 2g) although less resolved. Using UNC1-1 as a marker resulted in an unusual topology, where *O. sativa* is polyphyletic and four accessions appear closely related to *O.meridionalis*.

UNC2-2 belongs to the group unclassified-II (Turcotte *et al.* 2001). This group of element is also reminiscent of non-TIR MULEs. We have amplified UNC2-2 from an intergenic region in most AA genome rice species. An empty site was found in *O. meridionalis*

accessions 101146 and 105281, and revealed a perfect 8 bp TSD. Other than the empty sites, sequences of UNC2-2 revealed three indels. These sequences generated a unique most parsimonious tree of 45 steps (Figure 1h). This tree groups all *O. sativa* accessions in a cluster with *O. rufipogon*, *O. nivara* and accessions 105283 of *O. meridionalis*. *O. barthii* and *O. glaberrima* again form a distinct branch, and the three *O. longistaminata* accessions appear at a basal position. Again, MP (Figure 1h) and NJ (Figure 2h) methods yielded very similar trees.

Combined matrices

All variable characters (parsimony informative or not) from TE sequences of all data sets were used to form a combined matrix. Information about insertion polymorphisms was scored as presence or absence of TE and added to this matrix for a total of 277 characters. Characters corresponding to TE insertions were given more weight than nucleotide characters. Different values have been tested. A weight of 1:1 (all characters equally weighted) yielded a MP strict consensus tree with almost no resolution and a highly homoplasic NJ tree where most insertion/excision events occured independantly in different lineages (not shown). Increasing the weight of insertion polymorphisms reduced the homoplasy, but *O. nivara* accession 101508 no longer cluster with the two other *O. nivara* accessions because it lacks Tnr2. *O. longistaminata* accession 81968 was closer to the other accessions of this species, at the base of the tree, with *O. meridionalis* accessions 101146 and 105281. Weighting the insertion polymorphisms with increasing values (6:1 to 50:1) yielded identical trees using MP method. These trees are perfectly concordant with a NJ tree obtained when insertion polymorphisms carry a weight of 15:1 (Figure 3a). Increasing the weight ratio over 15:1 results in *O. sativa* accession 636 and *O. glumaepatula* becoming more basal and loss of resolution. In any case, *O. sativa* and *O.nivara* are polyphyletic because of the Tnr2 insertion. It would be possible to root these trees with *O. meridionalis* accessions 101146 and 105281 because these sequences lack UNC2-2, T1A-1 and both MULEs, or with *O. longistaminata* which lacks T8R-2, T1A-1, and both MULEs.

Because of the unusual distribution of Tnr2 and other SINE1-4 characters, a combined matrix excluding these variables was also used to generate cladograms (Figure 3b). A topology much more in line with established phylogenetic relationships was obtained using MP (not shown) and NJ (Figure 3b) methods. All *O. sativa* accession form a monophyletic clade, and all *O. nivara* accessions form a second monophyletic clade sister to *O. sativa*. Accessions of *O. barthii* and *O. glaberrima* still cluster together, but their position as sister to the *O. sativa*/ *O. nivara* clade is new. Again, *O. rufipogon* and *O. meridionalis* accession 105283 cluster together, as well as accessions of *O.longistaminata* and the remaining accessions of *O. meridionalis*, which appear as the most distantly related species because they lack MU9-2, MU9-3, and T1A-1. As for the previous tree, it is possible to root this tree with *O. longistaminata* or *O. meridionalis* accessions 101146 and 105281.

Discussion

Phylogeny of rice species

Except for the clustering of *O. barthii* with *O. glaberrima*, trees generated with the individual data sets have variable topologies. This incongruance probably indicates that the nucleotide sequences of a single TE do not provide sufficient phylogenetic signal to clearly resolve relationships between species of the *O.sativa* complex. A similar analysis based on a rice *Tourist* element yielded a tree with a poor resolution as well (Iwamoto *et al.* 1999). The phylogenetic trees shown in figures 1 and 2, although variable, reflect some trends found among previously published trees. For example, *O. longistaminata* and *O. meridionalis* appear at variable basal positions in many of the TE trees, in agreement with the trees derived from *Tourist* sequences (Iwamoto *et al.* 1999), RAPD (Ishii *et al.* 1996), nuclear and chloroplastic genes (Ge *et al.* 1999), and RFLP (Dally and Second 1990; Wang *et al.* 1992).

Generally, trees derived from individual data sets more clearly reflect the hypothesis of domestication of the cultivated African rice (*O. glaberrima*) from an African progenitor (*O. barthii*) than the domestication of the Asian cultivated rice (*O. sativa*) from *O. nivara*. *O. longistaminata*, which is the presumed perennial ancestor to the annual African species, actually appears ancestral in the UNC1-1 and UNC2-2 trees, but several other trees do not support this relationship. *O. rufipogon*, which is the presumed perrenial ancestor of the annual Asian species, appears at variable positions in the individual TE trees. Also, most data sets did not resolve the relationships between *O. sativa* and *O. nivara* ,or these data sets did not cluster all accessions of a species together, resulting in obscure relationships among the Asian species. Obviously, sequences of SINE1-4 divided the six accessions of *O. sativa* and the three accessions of *O. nivara* according to Tnr2 insertion, but also according to 5 other point mutations. Similarly, other data sets caused a division of these species into different branches based on the sequence variation. In a previous study, rice strains were classified based on several TE insertions (Mochizuki *et al.* 1993). The results also divided accessions of several Oryza species into different

groups. For example, accessions of *O. sativa* and *O. rufipogon* lacking Tnr2 were found in group A, while accessions having the Tnr2 insertion belonged to another group.

The frequent clustering of *O. rufipogon* and accession 105283 of *O. meridionalis* was quite unexpected. The hypothesis of DNA contamination can be eliminated because none of the accessions investigated are identical in sequence to *O. meridionalis* accession 105283. Among the ten loci investigated, at least three revealed an amplification polymorphism between accession 105283 and the other two *O. meridionalis* accessions. For example, the Tnr2 insertion in SINE1-4 sequence is present in accessions 101146 and 105281, but not in accession 105283. This is a good indication that these three accessions have different evolutionary backgrounds. Before 1981, *O. meridionalis* was not recognized as a species, and was defined as the Australian annual type of *O. rufipogon* (Khush 1997). According to our results, *O. meridionalis* 105283 is more accuratly classified as *O. rufipogon*, while the other two accessions are distinct from *O. rufipogon*.

Despite the uncertain individual TE trees, we could obtain robust trees with good resolution using combined matrices made of variable characters from several data sets added to the corresponding insertion polymorphisms information. These trees are robust since different character weighting and different methods (MP, NJ) resulted in very similar topologies. As long as insertion polymorphisms carry more weight than other characters, the trees were robust and less homoplasic for TE insertion/excision events. The tree inferred from the combined matrix made of all variable characters from all TEs is obviously the less homoplasic because most accessions lacking Tnr2 are clustered together. However, this tree cluster accessions of a species in different groups and is incongruent with most trees in the current literature.

The best topology was actually obtained by excluding the Tnr2 insertion from the matrix as well as the SINE1-4 sequence, which carries Tnr2 (Figure 3b). Tnr2 was excluded because of its very unusual distribution among the accessions, and SINE1-4 was excluded because the distribution of several SINE1-4 characters follows the distribution of Tnr2. In this new tree (figure 3b), *O. sativa* accessions form a monophyletic group, and all *O.*

nivara accessions form a monophyletic group sister to *O. sativa*. Accessions of *O. barthii* and *O. glaberrima* cluster together, as a sister group to the *O. sativa/ O. nivara* group. Again, *O. rufipogon* and *O. meridionalis* accession 105283 cluster together. Accessions of *O. longistaminata* and the remaining accessions of *O. meridionalis* appear distantly related to the other species because they lack MU9-2, MU9-3, and T1A-1. The position of *O. glumaepatula* is not clearly resolved in the current literature. Our data strongly support its inclusion in the group defined by the insertion of MU9-2, MU9-3, and T1A-1, but its position relative to other species within this group is not defined by a shared apomorphic TE insertion. The tree shown in figure 3b is very consistent with the hypothesis of domestication of the two cultivated rice species from the annual *O. nivara* and *O. barthii* as previously described. However, the relationships with putative ancestral perennial species are more obscure. The position of *O. rufipogon* is not strongly supported in the NJ tree and is not resolved in the MP tree (not shown). On the other hand, the distribution of T1A-1, MU9-2 and MU9-3 strongly support the dichotomy between the *O. meridionalis/ O. longistaminata* cluster and the cluster formed by all others. This result supports those based on RFLP (Wang *et al.* 1992), RAPD (Ishii *et al.* 1996), and *Stowaway* insertions (Kanazawa *et al.* 2000), but strongly disagree with topologies obtained from *Tourist* sequences (Iwamoto *et al.* 1999), and nuclear and chloroplastic gene sequences (Ge *et al.* 1999). Therefore, our data is in contradiction with the hypothesis that *O. longistaminata* is the ancestor to the cultivated African rice (Khush 1997), in which case the ancestor to O. barthii and *O. glaberrima*, the ancestor to O. *nivara* and *O. sativa* , and the ancestor to *O. glumaepatula* would have acquired independently the same T1A-1, MU9-2 and MU9-3 insertions in exact homologous loci. Insertion polymorphisms rather suggest an earlier divergence of *O. longistaminata*, before insertion of these three TEs in their locus.

The distribution of MU9-2, MU9-3, T1A-1, and UNC2-2, which are absent from *O. meridionalis* accessions 101146 and 105281, indicates that the root may be placed at the node leading to these two *O. meridionalis* accessions. Independent excision events seem more likely to occur than independent insertions of the same element in the same locus, between the same two nucleotides. Consequently, it is likely that T8R-2 inserted in the

genome of an ancestor to the AA genome rice species, before the speciation events occurred. Multiple parallel excisions of this TE must be assumed to reflect its distribution among the accessions investigated. Indeed, *O. longistaminata*, *O.glumaepatula*, and *O.rufipogon* all lack T8R-2, but do not form a monophyletic clade. Similarly, it is likely that the presence of Tnr2 was the ancestral state and that it has independently excised from the accessions lacking this element. The tree could alternatively be rooted with *O. longistaminata*. In this situation, the common ancestor probably contained the UNC2-2 element, which afterward excised from the ancestor to *O. meridionalis* accessions 101146 and 105281, and T8R-2 would have inserted in its locus after *O. longistaminata* diverge. Previous studies of rice species from different genome types place *O. meridionalis* at the base of the AA-genome species (Ge *et al.* 1999; Wang *et al.* 1992), and an analysis of a *Tourist* insertion even suggests that this node may represent the common ancestor to all Oryza species (Iwamoto *et al.* 1999). We therefore may hypothesize that the AA genome originated in Australia where O. meridionalis evolved, and spread to Africa *(O. longistaminata*, *O. barthii*. *O. glaberrima*), Asia (*O. nivara*, *O. sativa*), and America (*O. glumaepatula*).

Although SINE1-4 gives an unusual phylogenetic signal, SINEs are the most informative markers tested in this study, according to their high number of parsimony informative sites. However, the identification of SINEs in genomic sequences is not as simple as the identification of TEs that are delimited by TIRs. MITEs, *Ac*-like elements and MULEs for examples can easily be identified by characterisation of their TIRs and TSD. However, the nucleotide sequences of the MITEs and *Ac*-like elements investigated in this analysis are not variable enough to resolve phylogenetic relationships among species when used individually. The sequences of AC1-1 could generate a tree with good resolution (Figure 1c and 2c), but a tree based on a single marker may not be reliable. Therefore, we combined the information relative to T1A-1, T8R-2, and AC1-1 and constructed a cladogram (not shown) based on this third combined matrix, which contained 98 characters. Again, a weight ratio of 15:1 was given to insertion polymorphism characters. In the NJ tree, *O. sativa* is a polyphyletic group that includes *O. nivara* and *O. meridionalis* accession 105283. The *O. barthii/ O. glaberrima* group is

sister to the *O. sativa* group. *O. meridionalis* accessions 101146 and 105281 appear at an intermediate position rather than at the base of the tree. The topology obtained therefore contrasts with most trees previously published. Moreover, the fact that 2798 equally parsimonious topologies have been obtained using MP method, and that the resulting strict consensus tree is almost not resolved, indicates that this matrix may not be reliable.

TEs as phylogenetic markers

Different aspects of transposable elements can influence their use as phylogenetic markers. The absence of a TE from its locus may correspond to a site that has never received the TE insertion (e.g. empty site) or to a site that has once contained the TE, but the element has excised (e.g. excision site). An empty site would most likely be ancestral to other sequences having the corresponding insertion. However, in the case of a perfect excision (e.g. no footprints), the excision site is identical to an empty site. In the present study, we have reconstructed phylogenetic hypotheses from nucleotide sequences of TEs. Therefore, accessions having an empty site were excluded from individual data sets. However, the presence and absence of the insertion was scored as unordered character states, because no footprints were found to indicate an excision event, and this valuable information was added to combined matrices. Though the presence or absence of TEs are very informative phylogenetic markers, the exemple of the possible multiple parallel excisions of Tnr2 and T8R-2 indicate that care must be taken in the intrepretation of cladograms resulting from such markers and that the inclusion of more active transposable elements may complicate phylogenetic interpretation.

The topology of the trees inferred in this study imply excision of at least two TEs. Tnr2 and T8R-2 are absent from accessions that cannot be grouped in monophyletic groups. For example, if we force *O. longistaminata* accession 81968 to cluster with other accessions lacking Tnr2, we consequently introduce an accession lacking T1A-1, MU9-2 and MU9-3 into a group of accessions that possess these TE insertions. The resulting topology is more homoplasic. Tnr2 has been reported as putative class II transposon of 157 bp, with 56 bp TIRs and a TSD of 8 bp. To the best of our knowledge, the mobility of this element is not documented. T8R-2 is a *Tourist* MITE. Excision of MITEs is thought

to be a rare event (Wessler 1998; Wessler *et al.* 1995) and has not been extensively documented. Evidence for excision of a *Stowaway* MITE in Triticeae have recently been reported (Petersen and Seberg 2000). The authors found that the species having a putative excision footprint also share a mononucleotide that is absent form sequences bearing the *Stowaway* element and from sequences having a perfect empty site. They suggest that the mononucleotide C is inserted during excision of the element. Analysis of our data did not reveal any footprint. However, accessions lacking T8R-2 share 10 nucleotide substitutions in the sequences flanking the TE insertion site. Likewise, the accessions lacking Tnr2 share 5 nucleotide substitutions in the sequences flanking the Tnr2 insertion site. Based on the other data sets, these accessions do not cluster in a monophyletic group, meaning that the excisions have probably occured independantly, rather than in a unique common ancestor. As suggested for the insertion of a C in the the excision event reported for the Triticeae (Petersen and Seberg 2000), we may hypothesize that the shared nucleotide substitutions in the flanking sequences are generated by a mechanism of excision that is very specific.

The time of insertion of a TE in its locus would also impact phylogenetic reconstructions. Very recent insertions would be present in a limited number of closely related taxa, and would most likely have very similar sequences. Ancient insertions may allow comparison of more distantly related taxa, whose sequences may also have accumulated more mutations. The occurrence of these mutations at a constant rate during evolution of the taxa is also important. In this study, characters were often dispersed on most internal branches (e.g. the tree constructed from AC1-1 sequences). However, the sequences of SINE1-4 provided many characters supporting the early divergence of the taxa examined, but very few characters for further diversification.

A good marker should be easy to amplify and to sequence. The presence of a poly A track in TEs can be problematic. For example, the sequences of T8R-2 contain a short poly A track that resulted in slippage during sequencing. However, other sequences in this analysis contained longer poly A tracks and were still easily sequenced. Certain structural features typical of TEs may also complicate their use as phylogenetic markers. Although

T8R-2 has short TIRs, the presence of long TIRs may result in the formation of a hairpin structure, that could hinder the amplification or sequencing process. The elements MU9-2 and MU9-3 were selected for their short length compared to other MULE members. Their TIRs are approximately 250 bp long, so that only ~50 bp separate the 5' TIR from the 3' TIR sequences. The resulting potential to form secondary structure (dG=-195.8 kcal/mole) may therefore be responsible for the difficulties encountered with direct sequencing of the amplification products. Attemps were made to solve this problem using an alternative chemistry on a LI-COR system, but as cloning of MU9-2 and MU9-3 was problematic, the situation was not resolved.

The elements used in this study were selected primarily for their small size such that amplification and direct sequencing could be easily and rapidly performed. TEs from certain families, such as LTR retroelements, were definitely too large to be used as markers in our study. In contrast, the size of the elements was not a concern for non-autonomous MITEs and SINEs because they are all of small size. Moreover, for the purpose of primer design, it is interesting to have intronic TEs that can be amplified from conserved flanking exons, and MITEs are often found in the flanking sequences of genes or in their introns. Other families, such as *Ac*-like and MULEs, have members variable in size, and smaller representatives could be selected.

Use of TEs in phylogenetics

The variability obtained from nucleotide sequences of transposable elements appears to be appropriate for studying relationships at the species level. Because phylogenetic hypotheses rely on sequence alignments, it is of primary importance that nucleotide sequences be aligned with confidence, as they were in this analysis. Loci offering more variable sites, such as SINE1-1 and SINE1-4, may be more informative than those more conserved like the loci of T1A-1 and T8R-2, although a reasonable tree could be obtained with NJ method for the T1A-1 data set. Most data sets in this study were not variable enough to allow discrimination of accessions within a species or in certain cases accessions of closely related species such as *O. barthii* and *O. glaberrima*. Resolution of such lower level relationships would require more variable markers. Although the

nucleotide sequence of TEs did not provide such variability in our study, it is possible to take advantage of the abundance and genome-wide distribution of TEs and obtain the required variability for discriminating cultivars. For example, inter-MITE polymorphisms (IMP) were successful in fingerprinting barley lines (Chang *et al.* 2001). *Alu*-based amplification has also been found useful in human genome analyses (Nelson *et al.* 1991).

Higher level relationships can still be studied with TEs. In our study, most species in AA genome do possess the selected TE insertions. Some of these insertions may be shared by more distantly related species or even closely related genera. Other than sequence comparison, insertion polymorphism appears to be a very powerful tool for phylogenetic analyses at this level. Because class I elements apparently do not excise, class I insertion polymorphisms are unequivocal characters that are very useful for phylogenetic studies (Hillis 1999; Miyamoto 1999; Shedlock and Okada 2000). Studies of primates (Schmitz *et al.* 2001), cetartiodactyls (Lum *et al.* 2000), and cetacean lineages (Nikaido *et al.* 2001) for example, took advantage of the SINE and LINE unequivocal insertion polymorphisms to shed light on the relationships among these organisms. The usefulness of these retroelements in phylogenetic analyses is now broadly recognized (Hillis 1999; Miyamoto 1999; Shedlock and Okada 2000).

In our study, insertion polymorphism and sequence alignements of transposable element have confirmed the close relationship between the two cultivated rice species and their wild annual progenitor. Information from MITE T1A-1 as well as two MULEs, MU9-2 and MU9-3, have shed light on the position of *O. longistaminata* relative to other AA genome species. Though the presence or absence of TEs are very informative phylogenetic markers, possible multiple parallel excisions of TEs indicate that care must be taken in the intrepretation of cladograms resulting from such markers and that the inclusion of more active TEs may complicate phylogenetic interpretation. Although the structure and sequence of transposons can sometimes complicate the amplification or sequencing procedures, they still provide valuable phylogenetic signals that should be seen as a complement to analyses based on other types of markers.

Acknowledgments

M. Jackson (IRRI, Los Banos, Philippines) is acknowledged for kindly providing seeds. We are grateful to Richard Langlois, Julie Landry and Nancy St-Gelais for valuable technical assistance in sequencing. This work was supported by DNA Landmarks inc., an FCAR (Fonds pour la Formation de Chercheurs et l'Aide a la Recherche) scholarship to K.T., and an NSERC (Natural Sciences and Engineering Research Council of Canada) grant to T.B.

References

Aggarwal, R. K., D. S. Brar and G. S. Khush, 1997 Two new genomes in the Oryza complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. Mol. Gen. Genet. 254: 1-12.

Bureau, T. E., P. C. Ronald and S. R. Wessler, 1996 A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc. Natl. Acad. Sci. U S A 93: 8524-8529.

Bureau, T. E. and S. R. Wessler, 1994 Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. U S A 91: 1411-1415.

Capy, P., G. Gasperi, C. Biemont and C. Bazin, 2000 Stress and transposable elements: co-evolution or useful parasites? Heredity 85: 101-106.

Chang, R.-Y., L. S. O'Donoughue and T. E. Bureau, 2001 Inter-MITE polymorphisms (IMP): a high throughput transposon-based genome mapping and fingerprinting approach. Theor. Appl. Genet. 102: 773-781.

Cordesse, F., R. Cooke, D. Tremousaygue, F. Grellet and M. Delseny, 1993 Fine structure and evolution of the rDNA intergenic spacer in rice and other cereals. J. Mol. Evol. 36: 369-379.

Dally, A. M. and G. Second, 1990 chloroplast DNA diversity in wild and cultivated species of rice (Genus Oryza, section Oryza). Clasdistic-mutation and genetic-distance analysis. Theor. Appl. Genet. 80: 209-222.

Dellaporta, S. L., J. Wood and J. B. Hicks, 1983 A plant DNA minipreparation: version II. Plant Mol. Biol. 1: 19-21.

Ge, S., T. Sang, B. R. Lu and D. Y. Hong, 1999 Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc. Natl. Acad. Sci. U S A 96: 14400-14405.

Hall, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41: 95-98.

Hillis, D. M., 1999 SINEs of the perfect character. Proc. Natl. Acad. Sci. U S A 96: 9979-9981.

Hirano, H. Y., K. Mochizuki, M. Umeda, H. Ohtsubo, E. Ohtsubo and Y. Sano, 1994 Retrotransposition of a plant SINE into the *wx* locus during evolution of rice. J. Mol. Evol. 38: 132-137.

Initiative, T. A. G., 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

Ishii, T., T. Nakano, H. Maeda and O. Kamijima, 1996 Phylogenetic relationships in A-genome species of rice as revealed by RAPD analysis. Genes Genet. Syst. 71: 195-201.

Ishii, T., T. Terachi and K. Tsunewaki, 1988 Restriction endonuclease analysis of chloroplast DNA from A-genome diploid species of rice. Jpn. J. Genet. 63: 523-536.

Iwamoto, M., H. Nagashima, T. Nagamine, H. Higo and K. Higo, 1999 A *Tourist* element in the 5'-flanking region of the catalase gene CatA reveals evolutionary relationships among Oryza species with various genome types. Mol. Gen. Genet. 262: 493-500.

Kanazawa, A., M. Akimoto, H. Morishima and Y. Shimamoto, 2000 Inter- and intra-specific distribution of *Stowaway* transposable elements in AA-genome species of wild rice. Theor. Appl. Genet. 101: 327-335.

Khush, G. S., 1997 Origin, dispersal, cultivation and variation of rice. Plant Mol Biol 35: 25-34.

Kidwell, M. G. and D. Lisch, 1997 Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. U S A 94: 7704-7711.

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

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

Lum, J. K., M. Nikaido, M. Shimamura, H. Shimodaira, A. M. Shedlock, N. Okada and M. Hasegawa, 2000 Consistency of SINE insertion topology and flanking sequence tree: quantifying relationships among cetartiodactyls. Mol. Biol. Evol. 17: 1417-1424.

Makalowski, W., 2000 Genomic scrap yard: how genomes utilize all that junk. Gene 259: 61-67.

Mao, L., T. C. Wood, Y. Yu, M. A. Budiman, J. Tomkins, S. Woo, M. Sasinowski, G. Presting, D. Frisch, S. Goff, *et al.*, 2000 Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res. 10: 982-990.

Meyerowitz, E. M. (1994). Structure and organisation of the *Arabidopsis thaliana* nuclear genome. In: Arabidopsis. Meyerowitz, E. M., and Sommerville, C. Eds. Cold Spring Harbor, Cold Spring Harbor Laboratory Press: 21-36.

Miyamoto, M. M., 1999 Molecular systematics: Perfect SINEs of evolutionary history? Curr. Biol. 9: R816-819.

Mochizuki, K., H. Ohtsubo, H. Hirano, Y. Sano and E. Ohtsubo, 1993 Classification and relationships of rice strains with AA genome by identification of transposable elements at nine loci. Jpn. J. Genet. 68: 205-217.

Mochizuki, K., M. Umeda, H. Ohtsubo and E. Ohtsubo, 1992 Characterization of plant SINE, p-SINE1, in rice genomes. Jap. J. Genet. 57: 155-166.

Morishima, H., K. Hinata and H. I. Oka, 1963 Comparison of modes of evolution of cultivated forms from two wild rice species, *Oryza breviligulata* and *O. perennis*. Evolution 17: 170-181.

Morishima, H. and H.-I. Oka, 1960 The pattern of interspecfic variation in the genus *Oryza* : its quantitative representation by statistical methods. Evolution 14: 153-165.

Nelson, D. L., A. Ballabio, M. F. Victoria, M. Pieretti, R. D. Bies, R. A. Gibbs, J. A. Maley, A. C. Chinault, T. D. Webster and C. T. Caskey, 1991 *Alu*-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: identification of clones associated with a disease locus. Proc. Natl. Acad. Sci. U S A 88: 6157-6161.

Nikaido, M., F. Matsuno, H. Hamilton, R. L. Brownell, Jr., Y. Cao, W. Ding, Z. Zuoyan, A. M. Shedlock, R. E. Fordyce, M. Hasegawa, *et al.*, 2001 Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. Proc. Natl. Acad. Sci. U S A 98: 7384-7389.

Oka, H. I., 1974 Experimental studies on the origin of cultivated rice. Genetics 78: 475-486.

Oka, H. I., 1988 Origin of cultivated rice. Tokyo, Japan Scientific Societies Press.

Petersen, G. and O. Seberg, 2000 Phylogenetic evidence for excision of *Stowaway* miniature inverted-repeat transposable elements in Triticeae (Poaceae). Mol. Biol. Evol. 17: 1589-1596.

Rychlik, W. and R. E. Rhoades, 1989 A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res. 17: 8543-8551.

SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P. S. Springer, K. J. Edwards, M. Lee, Z. Avramova *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Sano, Y. and R. Sano, 1990 Variation in the intergenic spacer region of ribosomal DNA in cultivated and wild rice species. Genome 33: 209-218.

SantaLucia, J. J., 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA 95: 1460-1465.

Schmitz, J., M. Ohme and H. Zischler, 2001 SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. Genetics 157: 777-784.

Second, G., 1982 Origin of the genic diversity of cultivated rice (Oryza spp.); study of the polymorphism scored at 40 isozyme loci. Jpn. J. Genet. 57: 25-57.

Second, G., 1985 Evolutionary relationships in the sativa group of Oryza based on isozyme data. Genet. Sel. Evol. 17: 89-114.

Shedlock, A. M. and N. Okada, 2000 SINE insertions: powerful tools for molecular systematics. Bioessays 22: 148-160.

Song, W. Y., L. Y. Pi, T. E. Bureau and P. C. Ronald, 1998 Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the Xa21 family of disease resistance genes in rice. Mol. Gen. Genet. 258: 449-456.

Staden, R., K. F. Beal and J. K. Bonfield, 2000 The Staden package, 1998. Methods Mol. Biol. 132: 115-130.

Swofford, D. L. (2000). PAUP*. Phylogenetic Analysis Using Parcimony (*and Other Methods). Sunderland, Massachusetts, Sinauer Associates.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey and A. Rafalski, 2000 The complete sequence of 340 kb of DNA around the rice *Adh1-adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell 12: 381-391.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

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

Wang, Z. Y., G. Second and S. D. Tanksley, 1992 Polymorphism and phylogenetic relationships among species in the genus Oryza as determined by analysis of nuclear RFLPs. Theor. Appl. Genet. 83: 565-581.

Wessler, S. R., 1998 Transposable elements and the evolution of gene expression. Symp. Soc. Exp. Biol. 51: 115-122.

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

Yi, Q. M., W. G. Deng, Z. P. Xia and H. H. Pang, 1995 Polymorphism and genetic relatedness among wild and cultivated rice species determined by AP-PCR analysis. Hereditas 122: 135-141.

Table 1 Distribution of TEs among all accessions tested

species	accession	TE distribution									
		T1A-1	T8R-2	Ac1-1	MU9-2	MU9-3	SINE1-1	SINE1-4	CRA-1	UNC1-1	UNC2-2
O. sativa L. (Indica)	636	+	+	+	+	+	+	>	+	+	+
O. sativa L. (Japonica)	7288	+	+	+	+	+	+	+	+	+	+
O. sativa L. (Japonica)	244	+	+	+	+	+	+	+	+	+	+
O. sativa L. (Indica)	754	+	+	+	+	+	+	+	+	+	+
O. sativa L. (Indica)	611	+	+	+	+	+	+	+	+	+	+
O. sativa L. (Japonica)	42576	+	+	+	+	+	+	+	+	+	+
O. rufipogon Griff.	104642	+	-	+	+	+	+	+	+	+	+
O. nivara Sharma.	101508	+	+	+	+	+	+	+	+	+	+
O. nivara Sharma.	104684	+	+	+	+	+	+	>	+	+	+
O. nivara Sharma.	105703	+	+	+	+	+	+	>	+	+	+
O. glumaepatula Steud.	100971	+	-	+	+	+	+	>	+	+	+
O. glaberrima Steud.	100139	+	+	+	+	+	+	>	+	+	+
O. glaberrima Steud.	104562	+	+	+	?	+	+	>	+	+	+
O. longistaminata A. Chev. et Roehr.	81968	<	-	+	-	-	+	+	+	+	+
O. longistaminata A. Chev. et Roehr.	101198	<	-	+	-	-	+	> +	?	+	+
O. longistaminata A. Chev. et Roehr.	101205	<	-	+	-	-	+	> +	?	+	+
O. barthii A. Chev.	101252	+	+	+	+	+	+	>	+	+	+
O. barthii A. Chev.	104286	+	+	+	+	+	+	>	+	+	+
O. barthii A. Chev.	104290	+	+	+	+	+	+	>	+	+	+
O. meridionalis Ng	101146	-	+	?	-	-	+	>	?	+	-
O. meridionalis Ng	105281	-	+	?	-	-	+	>	?	+	-
O. meridionalis Ng	105283	+	+	+	?	+	+	+	+	+	+

+ identifies a fragment corresponding to the expected length of TE insertion

- identifies a fragment corresponding to the expected length an empty site

< and > identifies fragments respectively shorter and longer than the expected size of a TE insertion

? identifies locus that could not be amplified in the corresponding accession.

Table 2 Selected TEs and expected size of the amplification products

TE	TE type	Position of primers [a]	Length PCR product [b]	Locus
T1A-1	Tourist MITE	20287 F (8565-8583) R (9490-9470)	926 (606)	phy18 (intron 5)
T8R-2	Tourist MITE	8468049 F (25812-25829) R (26631-26609)	820 (624)	Putative eukaryotic translation initiation factor 3 large subunit (intron 1)
AC1-1	*Ac*-like	10179050 F (39138-39155) R (40015-39994)	878 (178)	Hypothetical protein (intron 1)
MU9-2	MULE	6006355 F (137696-137714) R (138570-138549)	897 (336)	Intergenic region
MU9-3	MULE	11875148 F (59672-59692) R (60544-60523)	894 (333)	Intergenic region
SINE1-1	p-SINE	(Hirano *et al.* 1994)	283 (144)	*wx*+ (untranslated 5' region)
SINE1-4	p-SINE	(Mochizuki *et al.* 1993)	433 (125)	n/a
CRA-1	undeter-mined	5042437 F (32330-32351) R (32862-32843)	552 (312)	Hypothetical protein (intron 8)
UNC1-1	undeter-mined	5670155 F (44997-44977) R (45844-25821)	866 (297)	n/a
UNC2-2	undeter-mined	5091496 F (93758-93779) R (94446-94430)	705 (367)	Intergenic region

[a] The first number refers to the Genbank GI number of the sequence from which primers were designed. Number in parenthesis refers to the position of the selected primers. F and R indicate respectively the forward and reverse primers.

[b] The number in parentheses corresponds to the expected size of an empty site.

n/a Not available

Table 3 Number of parsimony informative sites.

	# sites/ length	%
T1A-1	9/338	2,7
T8R-2	6/198	3,0
AC1-1	29/635	4,5
MU9-2	n/a	n/a
MU9-3	n/a	n/a
SINE1-1	22/138	15,9
SINE1-4	26/129	20,2
CRA-1	20/387	5,2
UNC1-1	23/431	5,3
UNC2-2	25/363	6,9

Figure 1. Maximum parsimony strict consensus trees derived from individual TE data sets.

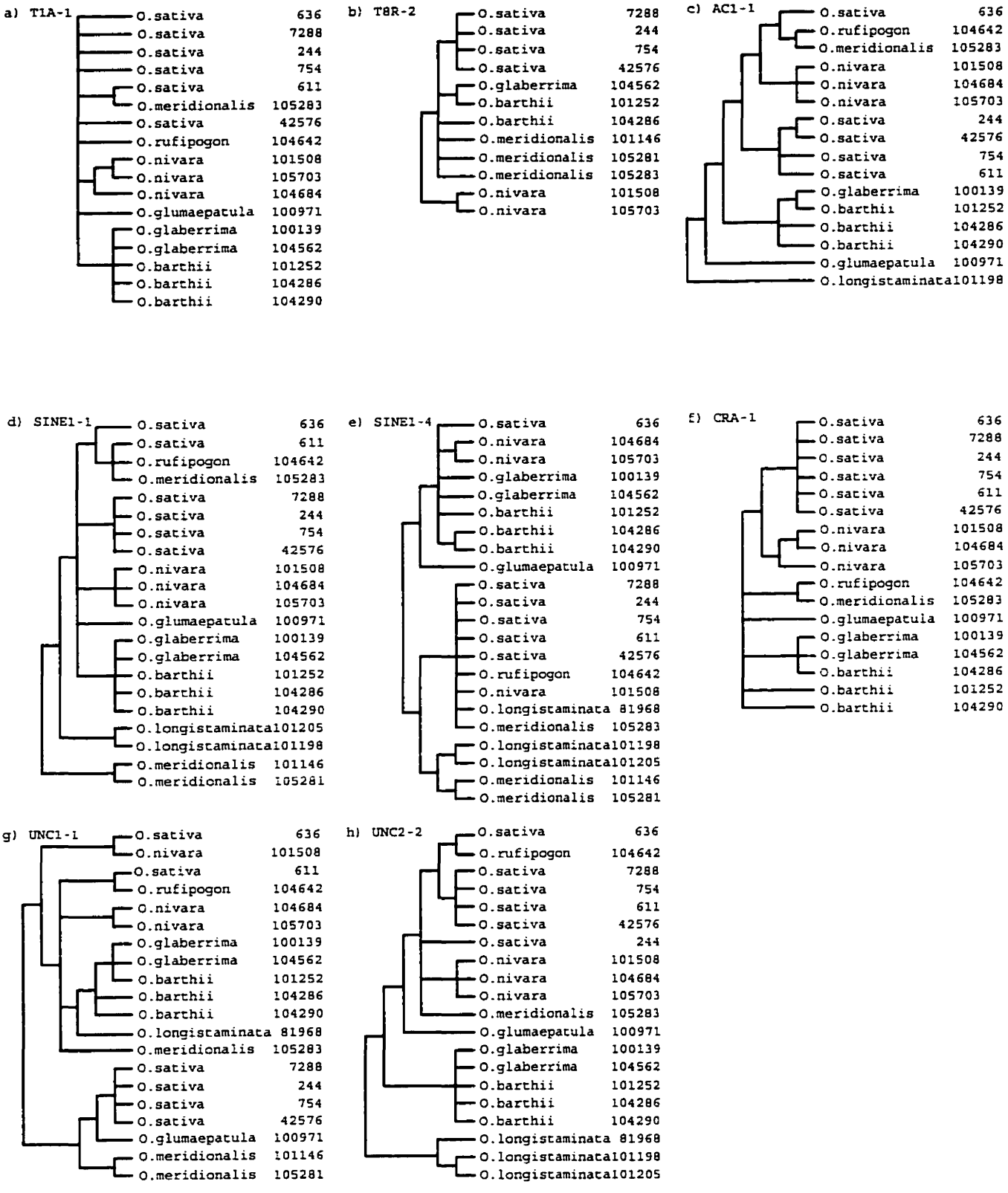

a) T1A-1
O.sativa 636
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 611
O.meridionalis 105283
O.sativa 42576
O.rufipogon 104642
O.nivara 101508
O.nivara 105703
O.nivara 104684
O.glumaepatula 100971
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
b) T8R-2
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.meridionalis 101146
O.meridionalis 105281
O.meridionalis 105283
O.nivara 101508
O.nivara 105703
c) AC1-1
O.sativa 636
O.rufipogon 104642
O.meridionalis 105283
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.sativa 244
O.sativa 42576
O.sativa 754
O.sativa 611
O.glaberrima 100139
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.glumaepatula 100971
O.longistaminata 101198
d) SINE1-1
O.sativa 636
O.sativa 611
O.rufipogon 104642
O.meridionalis 105283
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.glumaepatula 100971
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.longistaminata 101205
O.longistaminata 101198
O.meridionalis 101146
O.meridionalis 105281
e) SINE1-4
O.sativa 636
O.nivara 104684
O.nivara 105703
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.glumaepatula 100971
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 611
O.sativa 42576
O.rufipogon 104642
O.nivara 101508
O.longistaminata 81968
O.meridionalis 105283
O.longistaminata 101198
O.longistaminata 101205
O.meridionalis 101146
O.meridionalis 105281
f) CRA-1
O.sativa 636
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 611
O.sativa 42576
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.rufipogon 104642
O.meridionalis 105283
O.glumaepatula 100971
O.glaberrima 100139
O.glaberrima 104562
O.barthii 104286
O.barthii 101252
O.barthii 104290
g) UNC1-1
O.sativa 636
O.nivara 101508
O.sativa 611
O.rufipogon 104642
O.nivara 104684
O.nivara 105703
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.longistaminata 81968
O.meridionalis 105283
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.glumaepatula 100971
O.meridionalis 101146
O.meridionalis 105281
h) UNC2-2
O.sativa 636
O.rufipogon 104642
O.sativa 7288
O.sativa 754
O.sativa 611
O.sativa 42576
O.sativa 244
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.meridionalis 105283
O.glumaepatula 100971
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.longistaminata 81968
O.longistaminata 101198
O.longistaminata 101205

Figure 2. Neighbor joining trees derived from individual TE data sets. Dashed lines identify branches of length < 0.1 (total character difference, Jukes Cantor).

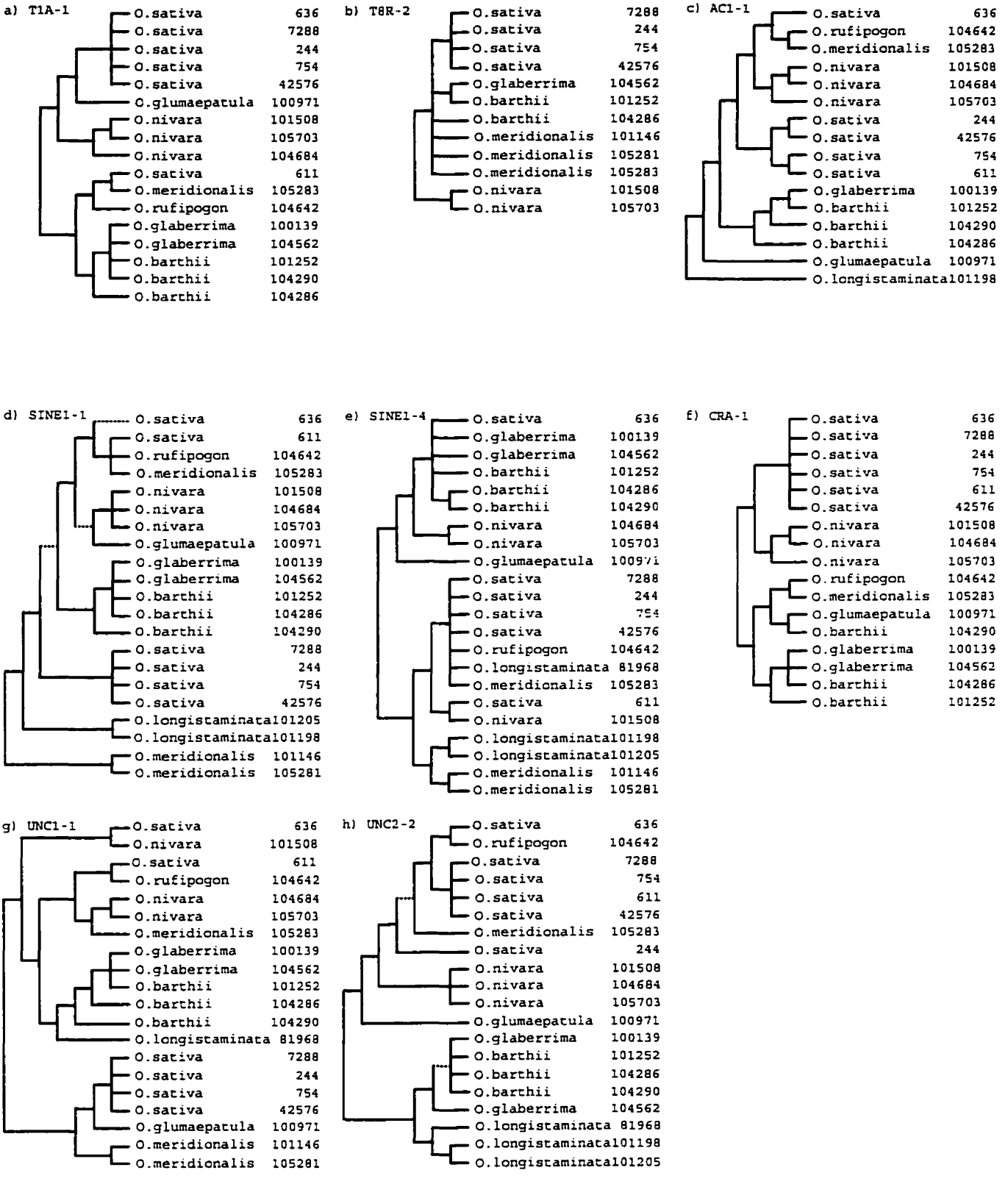
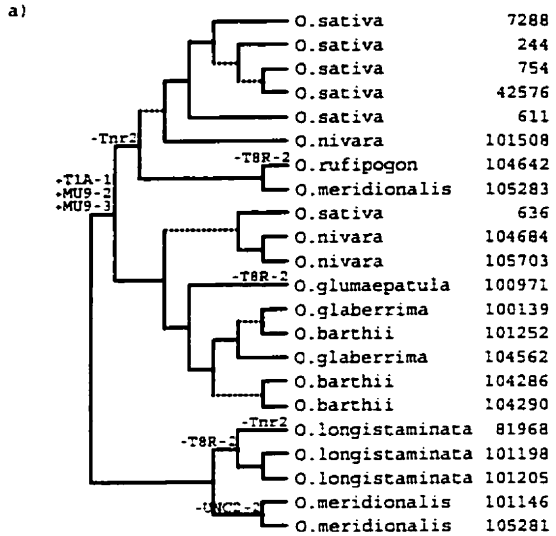

a) T1A-1
O.sativa 636
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.glumaepatula 100971
O.nivara 101508
O.nivara 105703
O.nivara 104684
O.sativa 611
O.meridionalis 105283
O.rufipogon 104642
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104290
O.barthii 104286
b) T8R-2
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.meridionalis 101146
O.meridionalis 105281
O.meridionalis 105283
O.nivara 101508
O.nivara 105703
c) AC1-1
O.sativa 636
O.rufipogon 104642
O.meridionalis 105283
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.sativa 244
O.sativa 42576
O.sativa 754
O.sativa 611
O.glaberrima 100139
O.barthii 101252
O.barthii 104290
O.barthii 104286
O.glumaepatula 100971
O.longistaminata 101198
d) SINE1-1
O.sativa 636
O.sativa 611
O.rufipogon 104642
O.meridionalis 105283
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.glumaepatula 100971
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.longistaminata 101205
O.longistaminata 101198
O.meridionalis 101146
O.meridionalis 105281
e) SINE1-4
O.sativa 636
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.nivara 104684
O.nivara 105703
O.glumaepatula 100971
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.rufipogon 104642
O.longistaminata 81968
O.meridionalis 105283
O.sativa 611
O.nivara 101508
O.longistaminata 101198
O.longistaminata 101205
O.meridionalis 101146
O.meridionalis 105281
f) CRA-1
O.sativa 636
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 611
O.sativa 42576
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.rufipogon 104642
O.meridionalis 105283
O.glumaepatula 100971
O.barthii 104290
O.glaberrima 100139
O.glaberrima 104562
O.barthii 104286
O.barthii 101252
g) UNC1-1
O.sativa 636
O.nivara 101508
O.sativa 611
O.rufipogon 104642
O.nivara 104684
O.nivara 105703
O.meridionalis 105283
O.glaberrima 100139
O.glaberrima 104562
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.longistaminata 81968
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.glumaepatula 100971
O.meridionalis 101146
O.meridionalis 105281
h) UNC2-2
O.sativa 636
O.rufipogon 104642
O.sativa 7288
O.sativa 754
O.sativa 611
O.sativa 42576
O.meridionalis 105283
O.sativa 244
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.glumaepatula 100971
O.glaberrima 100139
O.barthii 101252
O.barthii 104286
O.barthii 104290
O.glaberrima 104562
O.longistaminata 81968
O.longistaminata 101198
O.longistaminata 101205

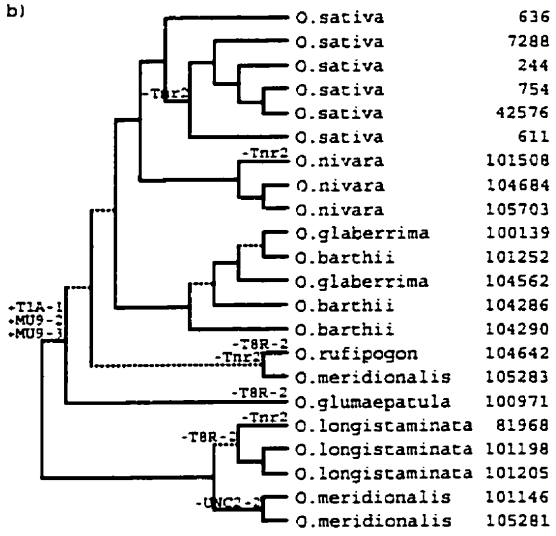
Figure 3. a) Neighbor joining tree derived from a combined matrix of all variable characters from all eight data sets and the corresponding insertion polymorphisms information. Insertion polymorphisms carry a weight of 15:1 relative to other characters. The MP strict consensus tree derived from the same matrix is perfectly congruent with this tree, although less resolved. Putative insertion and excision events are indicated with symbol + and – respectively. TEs that are not indicated were most likely present in the common ancestor. Dashed lines identify branches that collapse in the MP strict consensus tree.

b) Neighbor joining tree derived from a combined matrix of variable characters from several data sets and the corresponding insertion polymorphisms information, excluding Tnr2 and SINE1-4. Insertion polymorphisms carry a weight of 15:1 relative to other characters. The MP strict consensus tree derived from the same matrix is perfectly congruent with this tree, although less resolved. Putative insertion and excision events are indicated with symbol + and – respectively. TEs that are not indicated were most likely present in the common ancestor. Dashed lines identify branches that collapse in the MP strict consensus tree.

a)


O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.sativa 611
-Tnr2
O.nivara 101508
-T8R-2
O.rufipogon 104642
+T1A-1
+MU9-2
+MU9-3
O.meridionalis 105283
O.sativa 636
O.nivara 104684
O.nivara 105703
-T8R-2
O.glumaepatula 100971
O.glaberrima 100139
O.barthii 101252
O.glaberrima 104562
O.barthii 104286
O.barthii 104290
-Tnr2
O.longistaminata 81968
-T8R-2
O.longistaminata 101198
O.longistaminata 101205
O.meridionalis 101146
O.meridionalis 105281

b)


O.sativa 636
O.sativa 7288
O.sativa 244
O.sativa 754
O.sativa 42576
O.sativa 611
-Tnr2
O.nivara 101508
O.nivara 104684
O.nivara 105703
O.glaberrima 100139
O.barthii 101252
O.glaberrima 104562
O.barthii 104286
+T1A-1
+MU9-2
+MU9-3
O.barthii 104290
-T8R-2
O.rufipogon 104642
-Tnr2
O.meridionalis 105283
-T8R-2
O.glumaepatula 100971
-Tnr2
O.longistaminata 81968
-T8R-2
O.longistaminata 101198
O.longistaminata 101205
O.meridionalis 101146
O.meridionalis 105281

Final conclusion

The genome of *Oryza sativa* L. (domesticated rice) contains diverse transposable elements from both class I and II. The TEs identified contributed approximately 20% of the sequences surveyed. Elements from most major families of plant transposable elements were identified, and new groups were reported for these families. MITEs (miniature inverted-repeat transposable elements) account for >70% of the mined TEs and are clearly the predominant type of element in the sequences examined. There are three archetypal plant MITEs known as *Tourist*, *Stowaway* and *Emigrant*, each of which can be defined by a specific terminal inverted-repeat sequence signature. In rice, only *Tourist*-like and *Stowaway*-like MITEs have been identified, and a putative *Stowaway* transposase has been identified based on shared sequence similarity with the mined MITEs and previously identified transposases. A high resolution phylogenetic analysis of the putative transposases of several transposable elements revealed that *Stowaway* and *Emigrant* are both related to members of the previously characterized IS*630*/Tc1/*mariner* superfamily. More specifically *Emigrant* is closely related to the *pogo*-like family of elements, as reported by Feschotte and Mouches (2000), while *Stowaway* may represent a novel family. On the other hand, *Tourist*-like transposons along with prokaryotic IS*5*-like elements form a unique transposon superfamily.

Nucleotide sequences of MITEs, *Ac*-like, *Mutator*-like elements (MULE), short interspersed nuclear elements (SINEs) and other unclassified elements, as well as their insertion polymorphism data have been used to reconstruct the relationships between rice species in the AA genome. These analyses have confirmed the relationships between the two cultivated rice species and their wild annual progenitor. Informations from the MITE T1A-1 as well as two MULEs, MU9-2 and MU9-3, have shed light on the position of *O. longistaminata* relative to other AA genome species. The use of information from a combination of TEs generated robust cladograms that are in line with the current literature, and therefore confirm the usefulness of TEs as phylogenetic markers.

The progress in sequencing technologies and the increasing interest in genome sequencing will most likely result in the increasing availability from public databases of genomic sequences information from different organisms. Therefore, analyses such as high-resolution surveys of TEs will be facilitated and will continue to provide valuable information. Understanding the diversity and distribution of TEs within the rice, Arabidopsis and other plant genomes will provide the basis for further functional and regulatory characterization. Our report exemplifies that thorough TE mining is a key part of genomic cartography and, ultimately, will help determine the role of TEs in gene and genome evolution. TE mining will also provide information useful in the development of genome mapping, genome fingerprinting, gene isolation and gene functional analysis methodologies. TE-related informations obviously offer new opportunities for phylogenetic analysis, as a complement to traditional markers.