NOTE TO USERS

CO-EVOLUTIONARY RELATIONSHIP BETWEEN MOBILE DNA AND EUKARYOTES: AN INSIGHT FROM GENOME-WIDE CHARACTERIZATION OF *MUTATOR* (*Mu*)-LIKE ELEMENTS (MULEs) IN *ARABIDOPSIS THALIANA* AND *ORYZA SATIVA*

by

Zhihui Yu

Department of Biology

McGill University, Montreal

May, 2004

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 0-612-98391-9
Our file Notre référence
ISBN: 0-612-98391-9

ABSTRACT

The sequencing of eukaryotic model organisms has provided us an unprecedented opportunity for a genome-wide characterization of Transposable Elements (TEs) and the study of TE-host relationships. By developing methodologies on database mining, we explored the existence of *MUtator (Mu)*-Like Elements (MULEs) in *Arabidopsis thaliana* and *Orazy sativa*. *Mu* elements were first discovered in *Zea mays*; so far, a dozen of the elements have been identified in the genome. We identified a total of 1392 MULEs from the sequenced Arabidopsis genome. They represent one of the most abundant, diversified, yet still mobile DNA transposon families in eukaryotes. The Arabidopsis MULEs are composed of not only the elements showing the typical *Mu*-family-specific terminal structure (that is the long Terminal Inverted Repeat, TIR), but also a novel type of non-TIR MULEs. Some of this latter type of elements was found to be active both transcriptionally and transpositionally. To understand host-mediated genome-wide regulation(s) on the MULE system in *Arabidopsis*, we characterized 235 MULE mobility-specific genes (or *mudrA*-like genes) by mapping them on the sequenced Arabidopsis chromosomes and performing a genome-wide expression assay utilizing Arabidopsis METHYLTRANSFERSE1 (MET1) mutant (*met1*) plants, we showed that MET1-mediated global CpG methylation can only repress a portion of the gene family; its efficiency depends largely on the gene locations within the context of Arabidopsis chromatin remodeling: stronger in heterochromatic regions but weaker in euchromatic ones. This finding suggests that the Arabidopsis heterochromatic regions are not just a graveyard for the accumulation of defective elements; rather, they may have been playing

an important role on the repression of TE activity *via*, at least in part, exerting MET1-mediated silencing effect. Our expression analysis also suggested that a TIR structure is not necessarily required for the MET1-mediated silencing, neither is the repetition of the elements in the genome. To explore a possible role of TEs in the evolution of eukaryotes, we examined the MULE acquisitions of host DNA segments in both Arabidopsis and rice genomes. MULE-mediated amplifications of various host DNA sequences occur frequently. We identified a total of 389 Open Reading Frames (ORFs) that are not associated with a known transposase gene within the surveyed elements. Further characterization of a subset of them revealed that these MULE-contained genes were susceptible to host-mediated epigenetic regulations, show mosaic sequence organizations, and are often redundant in the two genomes respectively. MULE transposition in *Arabidopsis* has clearly facilitated the evolution of the family of Ubiquitin1-like (Ubl)-specific cytosine protease genes (*AtMULE-ULPs*), and their derived putative serine protease ones as well. Taken together, our genome-wide MULE study provides the evidence from several frontiers demonstrating that mobile DNA and eukaryotes can co-evolve.

RÉSUMÉ

L'abondance d'information provenant du séquençage du génome d'organismes modèles nous offre une opportunité sans précédant pour caractériser les éléments transposables et étudier les relations entre ces derniers et leurs hôtes. Nous avons exploré les éléments de type *Mutator* (MULE) de *A. thaliana* et *O. sativa*, au moyen de sondage de bases de données, et avons identifié 1392 MULEs. Les MULEs de Arabidopsis constituent des familles de transposon d'ADN très abondantes, diversifiées et toujours fonctionnelles. Elles sont composées d'éléments ayant les TIRs typiques de *Mutator*, mais aussi d'un nouveau type d'éléments n'ayant pas ces TIRs. Certains de ces nouveaux éléments sont transcrits et transposent. Nous avons cartographié 235 gènes de mobilité de MULE (type *mudrA*) sur la séquence du génome de Arabidopsis, et nous avons étudié leur expression afin de comprendre les mécanisme génomiques de l'hôte responsable du contrôle du système MULE de Arabidopsis. Bien que la méthylation du CpG ait un effet inhibiteur sur la transcription des gènes de type *mudrA*, son efficacité dépend largement de la position des gènes sur les chromosomes : surtout réprimés dans l'hétérochromatine, mais très actifs dans l'euchromatine. Les régions hétérochromatiques jouent donc un rôle important au niveau de la répression de l'activité des transposons. Nous avons aussi démontré que les gènes de type *mudrA* ne sont pas toujours réprimés par un mécanisme impliquant un gène homologue. Pour explorer le rôle des transposons dans l'évolution des organismes eucaryotes, nous avons examiné la diversité des éléments MULEs dans le génome de Arabidopsis et du riz. Il est fréquent que des segments d'ADN de l'hôte soient amplifiés par les éléments MULEs. Nous avons identifié 389 ORFs qui ne sont pas

associés avec des gènes de mobilité. Ces gènes mosaïques sont sujets au contrôle épigénétique de l'hôte, et sont souvent redondant dans le génome. La transposition de MULE chez Arabidopsis a de toute évidence affecté l'évolution de Ubl cysteine protéases (AtMULE-Ulp) et d'une famille de serine protéase dérivé de *AtMULE-ULP*. Notre étude des MULEs d'un point de vue génomique révèle plusieurs indices démontrant que les éléments transposables et les organismes eucaryotes peuvent co-évoluer.

ACKNOWLEDGEMENTS

I wish to express my greatest gratitude to my supervisor, Dr. T. E. Bureau. I thank him for giving me such a wonderful opportunity to work on the MULE 'world' freely, for his invaluable advice, guidance, encouragement, and generous support through out these years, for his tolerance about my misunderstanding and being childish sometimes, and for his challenges I accepted, enjoyed and through which I learnt. I also would like to extend my great thanks to Drs. D. Schoen and K. Hastings for their helpful advice throughout my committee meetings.

I am greatly indebted to Steven for his kindness, encouragement, helpful suggestions, and his input in *Athb1* mutation rate analysis. Special thanks are due to Mike and Nadia for their input in MULE acquisition survey.

I also would like to thank Nikoleta for her wonderful help in sequencing-related work, to Boris, Patrick and Newton for their assistance in bioinformatics, Kime for the French translation of the abstract, and to all my fellow graduate students for ensuring my life more dynamitic during this period of time.

I am grateful to the Biology department for providing me necessary assistance. Special thanks are due to Susan, the secretary of the Graduate Study Office, and Mark and Clair of the Phytotron.

Finally, my deepest gratitude goes to my parents, my sister and her family, and my beloved friend, Louis, for their being with and loving me along all the way.

CONTRIBUTIONS TO KNOWLEDGE

As one of the pioneering studies on TE-eukaryote relationships displayed at a genome-wide scale, the data described in this thesis provide novel insights to reveal a co-evolutionary relationship between a family of Class II TEs (or DNA transposons) and eukaryotic genome. Our major contributions to knowledge are summarized as follows.

1. Hundreds of the MULEs and *mudrA*-like genes were identified from 130 and 39 Mb of the sequenced Arabidopsis and rice genome respectively by developing methodologies on the mining of large sets of the sequenced genomic data. The corresponding sequences and their positions within the genomes/sequenced BAC (YAC) clones were stored and able to be archived at www.tebureau.mcgill.ca. Such information is greatly informative for the development of a MULE-tagging system for the studies of post-genomics and further TE-host relationships.

2. A novel type of MULEs, or non-TIR MULEs, were discovered and characterized. Our phylogenic study suggested that the TIR- and non-TIR elements evolved independently in the Arabidopsis genome. We revealed that in the Arabidopsis genome the *mudrA*-like genes are associated mostly with non-TIR MULEs. As some of them were active both transcriptionally and transpositionally, this novel type of elements are not just the defective form of the TIR-MULEs. This finding of functional non-TIR MULEs in Arabidopsis (1) challenges the common notion of the imperativeness of a TIR structure in the regulation of the mobility of *Mu* elements or DNA transposons in general, and (2) reveals the limitation of simply utilizing the presence (or lack) of a TIR alone to categorize families of Class II TEs.

3. The expression of a total of 92 Arabidopsis *mudrA* homologues was systematically examined for their response to the MET1-mediated silencing effect. This survey represents the first study on host-mediated genome-wide regulations of a family of indigenous transposase genes. Our two major discoveries include (1) MET1-mediated CpG methylation can differentially regulate individual *mudrA*-like genes simultaneously and (2) eukaryotic heterochromatic regions are not merely a graveyard for the accumulating of defective TEs; instead, the formation of heterochromatin may promote MET1-mediated silencing on TEs.

4. We identified a total of 389 non-transposase genes within the surveyed MULEs. We showed that the MULE mobility may be important in the evolution of the family of Arabidopsis *ULP* genes and the *MULE-ULP*-derived ORFs encoding a group of putative serine proteases. Taken together, we provide the novel evidence demonstrating a positive role of DNA transposon mobility in eukaryotic gene evolution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

AtPME: Pectin MethylEsterase gene in *Arabidopsis*.

AtMULE-PME: PME gene within an Arabidopsis MULE.

At-ULP: Ubiquitin1-Like (Ubl)-specific cytosine Protease gene in *Arabidopsis*.

AtMULE-ULP: Ubiquitin1-Like (Ubl)-specific cytosine Protease gene within an Arabidopsis MULE.

AtMULE-dULP: *AtMULE-ULP* derived ORF.

CMT3: CHROMOMETHYLASE3.

cmt3: CMT3 mutant.

ddm1: Decrease in DNA Methylation1.

DNMTs: DNA Methyltransferases.

DSB: Double-Strand DNA Break.

EST: Expressed Sequence Tag.

LINE: Long Interspersed Nuclear Elements.

LTR: Long Terminal Repeat.

MET1: METHYLTRANSFERSE1.

met1: MET1 mutant.

Mu: *Mutator*.

MULE: *Mutator*-Like Element.

ORF: open reading frame.

PEV: Position-Effect Variegation.

PTGS: Post Transcriptional Gene Silencing

SINE: Short Interspersed Nuclear Element.

SUMO: Small Ubiquitin Modifier.

TE: Trandposable Element.

TGS: Transcriptional Gene Silencing.

TIR: Terminal Inverted Repeat.

TSD:Target Site Duplication .

CONTRIBUTIONS OF CO-AUTHORS

TO MANUSCRIPTS FOR PUBLICATIONS

The major data described in this thesis were arranged into 3 manuscripts that correspond respectively to Chapter 3, 4, and 5, and two appendix papers. Manuscript 1 (Chapter 3) was published in *Genetics* (2000). Manuscript 2 (Chapter 4) will be submitted to *Genetics*. Manuscript 3 (Chapter 5) will be submitted to *Nature*. The appendix papers were published in *PNAS* (2000) and *Nature* (2000) respectively. Professor Bureau is co-author for all the manuscripts. As my supervisor, Dr. Bureau was responsible for overall design and supervision of the research project, and for correcting the writing. S. Wright is the co-author of manuscript 1. He made a major contribution to the statistical analysis of the *Athb1* mutation rate. M. Huynh and N. Mohabir are the co-authors of manuscript 3. M. Huynh made a major contribution to rice MULE analyses; N. Mohabir worked with me in organizing MULE data and identifying MULE acquisitions in Arabidopsis. My contribution to the *PNAS* paper includes providing most of the MULE data, more than one third of the other TE-mining data and experimental results for the demonstration of past TE-mobility, and to the *Nature* paper includes providing the MULE data.

CHAPTER 1

INTRODUCTION

Mobility, or transposition, distinguishes transposable elements (TEs) from host DNA segments; but how the mobility is carried out determines TE classes. Class I elements, or retrotransposons, require an RNA intermediate to accomplish their mobility; whereas Class II elements, or DNA transposons, transpose themselves directly (Plasterk, 1995). Since the first element was identified in *Z. mays* 50 years ago (McClintock, 1946), TEs have been found ubiquitously in all the surveyed eukaryotes where they make up a conspicuous fraction of the host genomes (recent reviews see Kidwell, 2002, Feschotte *et al.*, 2002; and Eline *et al.*, 2000). However, a long-term TE-host relationship remains largely unknown. Understanding it may facilitate to unveil the mechanisms driving eukaryotic gene and genome evolution.

Many eukaryotic TEs are capable of moving around. There is ample evidence showing that *de novo* TE motility can cause host gene mutations and chromosomal rearrangements, which, in many cases, can subsequently create visible mutant phenotypes (Lönnig and Saedler, 2002). As such, TE mobility has been perceived as an exceptional force for mutagenesis. But how can a eukaryote carry functional TEs while maintaining its stability? An obvious strategy would be *via* silencing. Several pioneering studies have shown that TEs can be suppressed by DNA methylation (review see Yoder, 1997). However, many questions were left unanswered regarding this mechanism. For example, as only one or two elements were examined at each time of these studies, it is currently

not known how an entire TE family is regulated simultaneously. It is also not clear whether or not there are other silencing systems working on TEs. If it is the case, can different mechanisms interact with each other to boost their silencing effect?

Previous studies have indicated some long-term beneficial influence of TE mobility on the evolution of eukaryotes (Kidwell and Lisch, 1997 and 2001). For example, insertions of *Alu* retrotransposons upstream of the human IgE receptor gene alter its expression pattern (Makalowski and Labuda, 1995). TEs may also directly participate in a host function. In *Drosophila*, a class of retrotransposases performs telomerase function during chromosomal duplications (Eickbush, 1997). However, compared to the great diversity and abundance of eukaryotic TEs, what was discovered may well be 'the tip of iceberg'.

Our ultimate goal is to examine TE-host relationships in eukaryotes. In this study, we chose to systematic study of the *Mutator* (*Mu*)-like elements (MULEs) in *Arabidopsis thaliana* mainly. Arabidopsis is one of higher plant species selected for a genome-wide sequencing project through international co-operation (Meinke *et al.* 1998). The vast sequence information generated by such an effort provides us an excellent opportunity to examine the molecular organization of the genome and explore systematically TEs and TE-host relationships. The *Mu* system was first identified in maize in 1987 (Robertson, 1987). One important merit of this system is to produce high forward mutation rate *per* generation, which makes it potentially the best candidate for mutagenesis applications among the known TEs discovered in higher plants (Walbot, 1992). A large number of maize genes have been characterized *via Mu* activity (Walbot, 2000). However, unlike *Ac/Ds* and several other eukaryotic TE systems that can be delivered into a heterogeneous

species while maintaining their mobility, all the attempts of introducing the functional *Mu* system into several higher plant organisms were not successful (Walbot, 1992). This could be because of unknown host-specific factor(s) that regulate *Mu* transposition (Walbot and Rudenko, 2002). Therefore, choosing to characterize the MULE system in Arabidopsis not only allow us to examine TE-host relationships, the information gathered will also be useful for a better understanding of the *Mu*/MULE family, consequently facilitating the development of a new *Mu*/MULE-tagging tool for.

We approached this study by testing three hypotheses (see the connecting statements of the thesis). We demonstrated that the MULEs in Arabidopsis belong to one of the largest, most diversified, yet still mobile TE families in eukaryotes. We showed that Arabidopsis MET1-mediated global CpG methylation can only suppress a portion of the *mudrA*-like gene family; its silencing efficiency is correlated with the formation of Arabidopsis heterochromatin. We revealed that MULE mobility in Arabidopsis and rice can facilitate to create eukaryotic genes and multiplegene families. In conclusion, TEs and eukaryotes are capable of co-evolving.

CHAPTER 2

LITERATURE REVIEW

Transposable elements (TEs), or mobile DNAs, were first discovered in Z. *mays* by Barbara McClintock in 1946. Since then, they have been identified in all examined eukaryotes, where they represent a conspicuous fraction of eukaryote genome (Kidwell, 2002). It was estimated that nearly 70% of the maize, 45% of the human, 15% of the Drosophila and 14% of the Arabidopsis genome are made up of various transposons (Kidwell, 2002, Eline *et al.*, 2002). The embracement of such abundant foreign DNA segments by a living eukaryotic organism raises several fundamental questions regarding their relationship(s). For example, how is TE activity regulated within a eukaryote? Can TEs offer any benefits for host development and evolution? Answers to these questions are fundamentally important for our understanding of the principles governing the evolution of a eukaryote.

The aims of this Chapter are to review recent development on the studies of TE-eukaryote relationships and to summarize features of a particular family of TEs, namely the *Mutator* (*Mu*)/*Mu*-like elements (MULEs) in higher plants. I address them by (1) introducing the major TE families discovered in eukaryotes, (2) discussing the main mechanisms of the regulation of TE activity, (3) reviewing the major discoveries of a positive role of TE activity in eukaryote evolution, (4) summarizing *Mu*/MULE studies and (5) drawing conclusions on our current understanding of TE-eukaryote relationships.

2.1 Classes of eukaryote TEs

Eukaryote TEs are mainly classified into two classes (Shapiro, 1995). Class I TEs include various families of retrotransposons, whereas Class II TEs refer to different families of DNA transposons. The key difference between the two is whether or not an RNA intermediate is required during transposition: the mobility of the former class members requires the step where TEs are reverse transcribed; in contrast, the latter class of elements transpose themselves directly (Plasterk, 1995). In addition, different classes also exhibit different terminal structural features (Shapiro, 1995). For example, Class I TEs usually contain either a Long Terminal Repeat (LTR) structure or show a polyA tail at the 3'-end, whereas Class II TEs are usually associated with a Terminal Inverted Repeat (TIR) structure. As these terminal features are generally conserved within intra-family members, they have been the hallmarks for the identification and classification purposes. Within each TE family, elements behave as either autonomous or non-autonomous. Autonomous elements carry a family-specific mobility-essential gene (or transposase gene) for both themselves *in vivo* and for the non-autonomous members (those don't carry the gene) *in vitro* (Plasterk, 1995). During transposition, a transposase conducts multiple functions, including DNA binding, cleavage, and strand transfer. One common outcome of a TE insertion is the creation of variable sizes of Target Site Duplications (TSDs). As TSD lengths within each family are usually conserved, they have been used as a guide for TE identifications and categorizations (Le *et al.*, 2000).

2.1.1 Retrotransposons

Based on features of terminal structures, eukaryotic retrotransposons can be further classified into three major subclasses: LTR-retrotransposons, Long Interspersed

Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) (Prak and Kazazian, 2000, Kumar and Bennetzen, 1999).

LTR retrotransposons The prominent features of this subclass involve an LTR structure and 5-bp TSDs (Kumar and Bennetzen, 1999). In addition, an autonomous element contains a polyprotein gene encoding *gag*, *prot*, *endo* and *RT/RNaseH* domains (Kumar and Bennetzen, 1999). The *gag* genes encode a class of viral structural proteins, which are not mobility-related. The pro genes encode proteinases for the cleavage of primary polyproteins into separate units. The endo genes encode integrases required for cDNA insertion into a host genome. The RT/RNaseH genes encode for reverse transcripteases essential for reverse transcription of the elements. The RT domain is also the most conserved region among eukaryotic Class I TEs. Within this subclass, different style of domain arrangements further divides the elements into two groups: *copia*- or *gypsy*-like elements (Prak and Kazazian, 2000, Kumar and Bennetzen, 1999). A *copia*-like element shows the domain organization of *gag-prot-endo-RT/RNaseH*. *Copia*-like element families include *Ty1* in yeast (Jordan and Macdonald, 1999) and *Ta1* in Arabidopsis (Bennetzen, 1999). In a *gypsy*-like element, the *endo* domain is positioned after *RT/RNaseH*. Examples of *gypsy*-like elements include yeast *Ty3* (Jordan and Macdonald, 1999) and tobacco *Tna1* (Bennetzen, 1999).

LINE/LINE-like elements In contrast to LTR-retrotransposons, LINE/LINE-like elements don't have the LTR structure, but show a poly(A) stretch at their 3'-ends (Kumar and Bennetzen, 1999; Eline *et al.*, 2002). Besides, they typically create 14-bp TSDs. Some LINE/LINE-like elements encode an RT homologue and may function as autonomous elements (Feng *et al.*, 1996). The frequent occurrence of incomplete reverse

transcription creates many 5'-end truncated elements. Large numbers of LINE/LINE-like elements were identified in mammals. The best known family is human LINE1 (L1) (Moran *et al.*, 1999).

SINE/SINE-like elements Despite that they are also non-LTR retrotransposons, SINE/SINE-like elements are much shorter than the former subclass, don't encode an RT and are relatively rare in eukaryotes (Kunze *et al.*, 1997). As they usually share sequence homology with an RNA gene, SINE/SINE-like elements were thought to be evolved from an ancestral RNA gene. The best known families of this subclass include human SINE and *Alu*.

2.1.2 DNA transposons

A common feature of Class II TEs is the TIR structure (review see Shapiro, 1995). Unlike Class I TEs that don't excise themselves during transposition, Class II elements can perform both excision and insertion activity. Excision of an element from a donor site leaves a double-strand gap, which is subsequently repaired by host-mediated DNA repair system (review see Plasterk, 1995). Categories of Class II TE families were typically carried out based on TIRs, transposase domains, or both (Kunze *et al.*, 1997). The best-known Class II TEs include *hAT*, CACTA and MITE subclasses.

hAT The prominent features of this subclass are the conserved short TIR sequence, 8-bp TSDs, and DDE motif within the transposases (Kunze *et al.*, 1997). Since first discovered in maize, *hAT* members have been identified brodaly in eukaryotes (review sees Bennetzen, 2000; Kunze *et al.*, 1997; Prak and Kazazian, 2000). Of the best-known *hAT* family, maize *Ac/Ds* family, *Ac* elements contain the gene encoding the *Ac* transposase, thereby regarded as autonomous TEs. In contrast, *Ds* elements behave as

non-autonomous TEs and were mainly derived from a defective *Ac*. An important feature of *Ac/Ds* system is that *Ac* transposase alone is sufficient for the regulation of *Ac* and *Ds* transposition. As such, *Ac/Ds* system has been introduced successfully into different heterogeneous organisms (Kunze *et al.*, 1997).

CACTA The elements are characterized by the TIR ending with the sequence CACTA (review see Bennetzen, 2000; Prak and Kazazian, 2000). The other two diagnostic features are 3-bp TSDs and the conserved family-specific transposase domain. TEs belonging to this subclass have been identified in both animals and plants. The most thoroughly studied CACTA family is maize *En/Spam* (Gierls, 1996). Members of this family contain a perfect 13-bp TIR. The autonomous *En/Spm* elements produce 2.4 kb and 6 kb mRNAs that encode two proteins termed TNPA and TNPD respectively. Although the two are both required for transposition, TNPA doesn't bind to the TIRs. As such, the TNPD was suggested to be the transposase of the family (Gierls, 1996).

MITEs There are two distinguishing features of this superfamily: small size (around 200 bp) and TA or TAA target-site preference (Bureau and Wessler, 1992, 1994; Bureau *et al.*, 1996, Le *et al.*, 2000). MITEs often maintain high copy numbers in a genome and preferentially inserted near host genes (review see Feschotte *et al.*, 2002). Since the identification of the first MITE a decade ago (Bureau and Wessler, 1992), MITEs have been found in a number of eukaryotes. Like other DNA transposons, all MITEs display a TIR structure. Based on their TIRs and TSDs, MITEs were further classified as either *Stowaway* or *Tourist* superfamilies (Le *et al.*, 2000). The putative transposase of the former family shows sequence homology with that of another well-characterized Class II TE family, TC1/mariner, suggesting a phylogenetic relationship

between the two (Le *et al.*, 2000). Although MITEs are widely distributed in eukaryotes, few were found to have a coding capacity or mobile. Recently, a larger *Tourist*-like element, PIFa, was identified and found to be active in maize (Zhang *et al.*, 2001).

2.2 TE activity regulation in eukaryotes

According to Kidwell (Kidwell and Lisch, 2001), the life cycle of eukaryote TEs can be divided into invasion, maturity and senescence. At the invasion stage, TEs start to emerge from a population, rapidly amplify themselves and finally set the foundation for further evolution within a host genome. At this stage, the host may largely be able to tolerate TE activity. During the maturity stage, TE amplification and loss are relatively balanced. Regulation of TE activity should prevail and most TEs become silent. However, dormant TEs can be awakened and reassume their mobility under various 'genomic shock' (McClintock, 1984). At the senescence stage, TEs have completely lost their mobility. They may also have diversified and become unrecognizable, have evolved to gain a new host function or have simply disappeared from the population. The silencing of TE activity can be achieved by host-mediated mechanisms, TE autoregulation or both (Labrador and Corces, 1997). Based on the scope and specificity, the host-mediated mechanisms can be further categorized as either genome-wide or TE-system-specific regulation.

2.2.1 Genome-wide regulation

Despite the huge diversity of eukaryotic TEs, eukaryotes appear to have evolved several 'universal' means to repress TE activity systematically (Labrador and Corces, 1997). The most-discussed mechanisms include DNA methylation (Bird, 1997),

heterochromatin repression (Pimpinelli *et al.,* 1995), and RNA silencing (Metzke *et al.*, 2001).

DNA methylation-mediated silencing In eukaryotes, DNA methylation occurs in cytosine, mostly at CpG and CpNpG sites (Finnegan, 1998). It appears that TEs are mostly methylated in eukaryotes; however, whenever mobility is reassumed, TEs are found to be hypomethylated (Federoff, 1996; Zhou *et al.,* 2001). This correlation between DNA methylation and TE activity led to the proposal that DNA methylation can suppress TE activity. In principle, DNA methylation-mediated silencing takes effect by restraining the transcription of transposase genes. Methylated DNA can directly prohibit the binding of the basal transcriptional machinery and/or specific transcription factor(s) to the elements. It can also alter chromatin structures, thereby indirectly inhibiting transcriptions (Kass *et al.*, 1997; Costello and Plass, 2001). In addition, cytosine methylation within transposase binding sites can also block the interaction between transposases and the corresponding TEs. The silencing of TE activity by DNA methylation involves TGS and, in many cases, is homology-dependent (Cogoni, 2001).

DNA methylation is the most studied and widely characterized host-silencing system on TE activity. It exists in a wide range of organisms and can repress both Class I and II TEs, such as, *Ac/Ds*, *En/Spm*, *Mu*/MULEs, and *Athila etc.* (Gierls, 1996; Hirochika *et al.*, 2000; Kunze *et al.*, 1997; Martienssen *et al.*, 1994). In *N. crassa*, the transposition of LINE-like elements, *Tad*, occurred only in cultures treated with 5-azacytidine (a cytosine analog that can block cytosine methylation), but was not observed in 5-azacytidine-free tissues (Zhou *et al.*, 2001).

Heterochromatin Euchromatin and heterochromatin are the two important chromatin states of eukaryotic chromosomes. Euchromatin carries the majority of 'native' host genes and shows typical recombination rates; however, heterochromatin contains few such genes and has a repressed recombination rate (Weiler and Wakimato, 1995). Heterochromatin represents a condensed chromatin structure and the genes within it are typically confined and inaccessible to the transcription machinery (Copenhaver *et al.,* 1999; Fransz *et al.,* 2000). These features, together with the evidence that most of eukaryotic TEs accumulate within heterochromatic regions of a host genome, led to the proposal that eukaryotic heterochromatin evolved for the repression of TE activity (Weiler and Wakimato, 1995).

To demonstrate this hypothesis, it is important to rule out the possibility that the heterochromatic regions are merely the graveyard of accumulation of dead TEs. In other words, we should be able to identify TE activity in heterochromatin-deficient mutants. Unfortunately, as molecular characterization of heterochromatin is still in its infancy and just few mutants were generated in the past, only several TEs were tested. Nevertheless, from the limited studies, it was observed that TEs within heterochromatin were functionally competent and could be reactivated in the mutants carrying genes deficient for chromatin remodelling (Singer *et al.*, 2000; also see Chapter 4 and 5).

dsRNA According to the current two-step model (Tijsterman, *et al.*, 2002), dsRNA-mediated silencing occurs first by dicing long dsRNA molecules into 21-23nt siRNAs (catalyzed by a Dicor or dsRNA-specific nuclease) and second by siRNA's guiding the formation of a nuclease-containing protein complex designated as RISC (RNAi-Induced Silencing Complex). The subsequent RNA-RNA pairing between the

antisense strand of siRNA and the target mRNA allows the complex to access to the substrate and trigger mRNA degradation. The dsRNA-mediated silencing of gene activity shows PTGS, TGS and systematic propagation effects.

RNA silencing of TE activity was first discovered in *Caenorhabditis elegans* (Fire *et al.*, 1998), subsequently in several other species (Ketting *et al*, 1999). In *Chlamydomonas reinhardtii*, TOC1 retrotransposons and Gulliver DNA transposons were found to be activated in a RNA silencing-deficient mutant, *Mut*6, background (Wu-Scharf, *et al.*, 2000). In *Trypanosoma brucei*, where RNA silencing system exists, retrotransposon-derived siRNAs were identified (Djikeng *et al.*, 2001).

2.2.2 Developmental and TE system-specific regulation

Of several TE systems, transposon activity was found to be restricted to a narrow window of host development. For example, the P-elements in Drosophila are normally repressed in somatic tissues and reactivated only in germ line (Engel, 1996). This type of TE activity regulation is shown to be determined largely by the participation of host-specific factors. In the case of P-elements, the repression in somatic cells occurs as the result of insufficient fully-spliced transcripts from P-transposase gene, a consequence caused by host-encoded protein on the repression of the transposase gene activity (Tseng *et al.*, 1990).

2.2.3 Autoregulation

In addition to host-mediated regulatory mechanisms, TEs have developed strategies to self-regulate their own copy numbers. For example, after transformation of maize *Ac/Ds* system into tobacco, the TE activity was controlled by a reverse dosage effect: the repression of *Ds* mobility occurred only when the level of the expressed *Ac*

transposase reached above the threshold (Kunze, 1996). In other examples, TE activity can also be down regulated by self-encoded repressor proteins, or homology-dependent gene silencing (Gierl, 1996; Gensen *et al.*, 1999).

2.2.4 *Interaction between host-mediated mechanisms*

Certainly, there is ample evidence to implicate the presence of a host surveillance system of mobile DNA. However, one fundamental question left unanswered is how a host defense system recognizes TEs. Confined by the limited scope and degree of studies on this issue, a widely applicable and concrete statement seems impossible. However, the data from several studies did provide some insight on possible TE signals that can be recognized. They are (1) the repetitiveness of TEs in the genome, (2) the specific TE structures, (3) the TE-generated dsRNAs, and (4) the difference in GC-content between host genome and TEs (Hsieh and Fire, 2000; Tijsterman *et al.*, 2002;)

Can different mechanisms interact with each other and work as a whole on the silencing of TE activity? Unfortunately, so far, few reports have addressed this subject (see Chapter 4). Nevertheless, studies on molecular and genetic characterization of individual systems and their effects on transgenes and/or 'native' genes have confirmed that (1) heterochromatic regions are rich with methylated DNA (Razin and Cedar, 1997), (2) heterochromatin can exert and stabilize the silencing effect induced by CpG methylation *in vitro* (Kass, 1997), (3) some chromatin proteins contain the binding motif exclusive for methylated DNA while some methyltransferases carry a chromatin domain (Fuks *et al.*, 2000; Razin, 1998), (4) dsRNA can trigger heritable DNA methylation (Jones *et al.*, 2001). In fact, several models on the regulation of other repetitive DNA through repression were also recently proposed (Wolffe *et al.*, 1999; Matzke *et al.*, 2001),

which may be applicable to TEs as well. The main steps of these proposed mechanisms involve (1) methylated DNA induces the recruitment of multiple proteins, including histone deacetylases, at methylation sites to form a repression multiprotin complex, (2) the deacetylation of lysine residues on histone H3 and H4 restricts nucleosome mobility, resulting in the formation of a compact heterochromatin structure that can subsequently stabilize and exert methylation-mediated silencing, and (3) siRNA can travel from the cytoplasm to the nucleus where RNA-DNA pairing can trigger cytosine methylation.

2.3 Functional roles of TE activity on host development and evolution

2.3.1 Gene evolution

Increasing evidence suggests that TEs are an important player in the origin and evolution of eukaryotic genes/mutigene families, and the regulation of gene expressions. There are three major pathways by which TEs can participate in eukaryotic gene evolution. By insertion, TEs can provide new regulatory motifs or exon/intron sequences (review sees Kinwell and Lisch, 2001). By abnormal transposition/retrotransposition, TEs can participate in the formation of new mosaic genes and gene duplicates (see Chapter 5). By domain sharing, TEs can directly take on a cellular function (Eickbush, 1997).

Regulatory motifs Several lines of evidence demonstrated that TE insertion into or near a host gene can alter its expression, which may offer certain benefits (Kidwell and Lisch, 2001; Bennetzen, 1999). TEs can provide regulatory motifs, such as promoters, enhances/repressors, or polyadenylation signal to nearby host genes. LTR-retrotransposons and autonomous DNA transposons carry promoters that can be used as an alternative *cis*-element for the transcription of downstream genes. The human IgE

receptor gene is a cell-type-specific gene whose specificity is controlled by the two *Alu* elements inserting upstream of the gene. The element at the further position serves as an enhancer in both basophilic and T-cells, whereas the one at the nearer position functions as an enhancer in only T-cells (Makalowski and Labuda, 1995).

Coding and intron sequences Survey of the human genome reveals the existence of numerous TEs within translated protein sequences (Li *et al.*, 2001). The same case was also observed in other vertebrates (http:/www.ncbi.nlm.nih.gov/Makalowski/Scrap). It was estimated that nearly 10% of human DAF mRNAs contain an *Alu* sequence. The proteins encoded by the *Alu*-free genes are membrane-bound; whereas the proteins encoded by *Alu*-containing mRNAs are soluble, suggesting a functional role of *Alu* elements in DAF protein diversity and evolution (Caras, *et al.*, 1987).

Can TEs function as an intron of inserted genes? Data collected from several lines suggest a possibility (Benntzen, 1999; Wessler, 1987). For example, a *dSpm* insertion into the second exon of maize bronze gene (*Bz*) created a *bz* allele (*bz-m13*); however, it was spliced out when *bz* transcribed (Wessler, 1987). Further comparison with the original *BZ* sequence revealed that the inserted element actually functioned as a fraction of the new intron by providing a new 3'-splicing site (Puruganan and Wessler, 1992; Wessler, 1987). A similar example was also observed in *Drosophila* where a P-element inserted into a *yellow* gene created a new site for alternative splicing (Geyer *et al.*, 1991).

Eukaryotic genes/multigene families Reverse transcription of Class I TEs is the first step for their mobilization. Although they normally use their own polyadenylation signals, on many occasions, Class I TEs also use alternative stronger promoters residing down-stream of the elements, resulting in the transduction of nearby host gene segments

(Goodier *et al.*, 2000; Pickeral, *et al.*, 2000). In the human genome, 15–23% of L1 insertions contain a 3'-transduced host sequence (Pickeral, *et al.*, 2000). The possibility of creation of a novel mosaic gene *via* L1 transduction was recently confirmed under an experimental condition (Moran *et al.*, 1999).

DNA transposons create new genes by abnormal DNA repair after element excisions (see Chapter 5). Generally, Class II TEs mobilize *via* a cut-and–paste mechanism (Plasterk, 1995). Such a process leaves a double-strand DNA break (DSB) at donor sites after TE excisions. It is known that these DSBs are normally repaired by a group of host-encoded repairing enzymes, usually characterized as an error-prone process including the capture of non-homologous DNA segments from ecotypic sites (see Chapter 5).

TEs' beneficial functions Some eukaryote genes may have evolved directly from a TE-specific gene. This was exemplified in RAG gene evolution (Melek, *et al.*, 1998). In vertebrates, generation of immunoglobin diversity is essential for immune response. It is achieved by VDJ recombination, a process of chromosomal breaking-rejoining catalyzed by the RAG genes. The VDJ recombination is similar to a typical transposition process. Furthermore, RAGs were found to maintain certain level of transposase activity *in vitro*. Thus, it is possible that the critical portion of the immune system in vertebrates was evolved from an ancient TE system carrying a RAG gene homologue.

2.3.2 Chromosomal evolution and the dynamics of eukaryote genome

TE activity appears to have contributed to the evolution of eukaryotic sex chromosomes, chromosomal structures (such as, telomeres and centromeres) and new linage relationships (review see Lonnig and Saedler, 2002). Recently, several studies also

provide the evidence suggesting TE contribution to the evolution of heterochromatin (Dimitri and Junakovic, 1999). As such, TEs may play a functional role in chromosomal dynamics and evolution.

Telomeres Telomeres are the structure at chromosomal ends important for chromosomal stabilization. Generally, a telomere is synthesized by host-encoded telomerase-mediated reverse transcription of short RNA molecules. However, in some species, such as *D. melanogaster*, telomeres were created by retrotransposons. Drosphila does not have a typical eukaryotic telomerase gene. Instead, it uses retrotransposon-encoded reverse transcriptase to perform telomerase's function: moving the elements from other regions of the genome to the ends of each replicated chromosome (Eickbush, 1997).

Centromeres The centromeres of eukaryotic chromosomes are essential for the pairing, segregation and inheritance of genetic information. It appears that TEs, especially retrotransposons, are a major component of eukaryotic centromeres and some of them actually contain centromeric-specific domains, suggesting a functional connection between eukaryotic centromeres and TEs (see Chapter 4). Recently, different TEs were also found to be targeted by several centromere-specific proteins (van Steensel *et al.*, 2001). However, the actual TE function in eukaryotic centromeres remains unknown.

Lineage diversities Chromosomal variations play an important role for developing fertility barriers between species. TEs are known to cause all types of chromosomal rearrangements: duplication, translocation, inversion and deletion (Lonnig and Saedler, 2002). For example, in *Drosophila*, several DNA transposon families were identified

at/near the break-points of hybrid dysgenesis-induced chromosomal rearrangements (Lim, 1988). A genome-wide study of *Ty* retrotransposons in the yeast genome revealed that they were often located near recombination hot spots (Kim, 1998). Two possible mechanisms are mainly responsible: ecotypic recombination between elements and transposition-induced rearrangements.

Sex chromosomes Sex chromosomes play an important role in eukaryotic sex determination and fertility. Studies from a number of eukaryotic species revealed that TEs, especially Class I elements, are rich within these chromosomes, suggesting a possible TE role in the origin of the chromosomes (Hackstein and Hochstenbach, 1995; Steinemann and Steinemann, 2001). In *D. miranda*, the massive insertions of different families of class I TEs were correlated with the formation of neo-Y-chromosome (Steinemann and Steinemann, 1997). In addition, TEs can also contribute to Y-chromosome degradation by successive silencing of the chromosomal genes (Steinemann *et al.*, 1993).

Heterochromatin Heterochromatin is an important chromatin structure that may be involved in a number of cellular functions, such as the regulation of recombination rate, gene activity, and cell division. TEs are known to be abundant within heterochromatic regions (International Human Genome Sequencing Consortium, 2001; The Arabidopsis Genome Initiative, 2000). TE accumulation within heterochromatin could be resulted from host selection against their amplification within euchromatin. Alternatively, preferential TE insertions into heterochromatin could also contribute to this unbalanced TE distribution in eukaryotes. Can this distribution pattern be essential for a heterochromatin function? There is some indication suggesting that TEs could play an

important role in the origin and evolution of heterochromatin. In a study by Dorer and Henikoff (1994), a tandem array of modified Drosophila P-elements could induce a *de novo* heterochromatic state. Another example showed that massive insertions of retrotransposons could also create the heterochromatic neo-Y-chromosome (Steinemann and Steinemann, 1997).

Amplification of eukaryotic genomes It is well known that TEs can expand eukaryotic genomes. For example, the maize genome was doubled in size mainly by the amplification of class I TEs (Bennetzen, 2000). The enlargement of genome size doesn't always accompany the increase in the genome complexity (Kidwell, 2002). As such, it was thought that the variation of TE abundance in eukaryotes might merely reflect the fact that they are selfish DNA. However, this point of view was challenged recently by Schulman and his colleagues (Kalendar *et al.*, 2000). They examined the abundance of BARE-1 retrotransposons from the populations distributed in different habitats and observed that the plants growing in higher and dryer areas had nearly 3 times more TEs than those distributed at valley habitats. Their finding suggests that TE abundance in a eukaryotic genome may influence its ability to cope with stressed environments.

2.3.3 TEs in speciation and beyond

Syvanen (1984) ever stated, "I believe that transposons have the potential to induce highly complex changes in a single event". Can TE-mediated changes at the gene and chromosomal levels lead to the speciation and origin of higher systematic categories? Unfortunately, studies on TEs' contribution to the origin of species and beyond are still at an early stage, and as of yet, no evidence so far available to directly demonstrate TE's role(s) in speciation. However, TEs have shown to induce karotype changes, alter both

the expressions and functions of many eukaryotic genes, and create certain level of phenotypic modifications. These changes may provide fundamental genetic flexibility necessary for adaptive evolution and are therefore in accordance with the proposal that TE activity could be one driving force in the origin of certain level of biodiversity (Lonnig and Saedler, 1997). In fact, TE-induced gene inactivation has been related to the origin of cultivated plants and domestic animals (Lonnig and Saedle, 1997; Ingham *et al.*, 1993). TE activity also can create cross-fertilization barriers between different lines of *Pisum sativum* in a relatively short time period.

2.4 *Mutator*/MULEs in higher plants

Mu/MULEs are by far the most diverse and most active DNA transposons in higher plants (Bennetzen, 1996). The first *Mu* element was identified in a maize *Mutator* line 25 years ago (Robertson, 1978). Since then, a number of the elements have been identified in both maize and other higher plant species (Turcotte, *et al.*, 2001; see Chapter 3 and 5). So far, the best studied *Mu*/MULE system is from maize, where the autonomous *Mu* element, or MuDR, has been further characterized (recent review see Walbot and Rudenko, 2001).

2.4.1 Mu/MULE terminal structures

The long TIR (>100 bp) is the prominent feature of the *Mu*/MULEs family. It contains the *Mu* transposase (MURA) binding site (Benito and Walbot, 1997). Within an active MUDR, the TIRs also carry the promoters of *mudrA* and *B* and other *cis*-regulatory elements (Hershberger *et al.*, 1995; Raizada *et al.*, 2000). The long TIR structure was originally thought to be imperative for *Mu* activity. However, recent identification of non-TIR-MULEs in *Arabidopsis* suggests otherwise (see Chapter 3). In *Arabidopsis*,

many MULEs don't form a typical *Mu*-TIR, but still carry a *mudrA*-like gene. Unlike other DNA transposon families that usually share a conserved TIR sequence within intra-family members, *Mu*/MULE family have diversified TIR sequences (see Chapter 3).

2.4.2 MuDR/MuDR-*like elements*

MuDR/MuDR-like elements are autonomous *Mu*/MULEs that contain a *mudrA/mudrA*-like gene (Bennetzen, 1996, see Chapter 3). In addition, MuDRs in maize also contain a *mudrB* and some MuDR-like elements in *Arabidopsis* contain a *ULP* gene or other ORFs (see Chapter 5). There are multiple copies of MuDR/MuDR-like elements in a genome, but only few can catalyze the transposition process (hMuDR, Rudenko and Walbot, 2001).

In maize, *mudrA* and *B* genes are transcribed in active *Mutator* lines and at least *mudrA* transcription is essential for *Mu* activity (Bennetzen, 1996). Some *mudrA* transcripts start at +169 of a MuDR and the transcripts overlap partially with the left TIR sequence; whereas the others start at +252 (outside of the TIR). *mudrB* transcripts start at +163 and overlap with the right TIR (Hershberger *et al.*, 1995). The two genes are transcribed as a convergent MuDR transcript terminating 200-bp away from each other. *mudrA* typical show a 3-introns and 4-exons structure (Hershberger *et al.*, 1995). The fully-spliced transcript is 2.8 kb. *mudrB* also has a 3-intron structural organization and produces 1 kb fully-spliced transcripts. In addition, MuDRs also produce aberrant transcripts by differentially splicing as well as the production of antisense transcripts produced by the failure of proper termination of the *A* or *B* genes (Hershberger *et al.*, 1995). Besides maize, the transcription of *mudrA*-like genes was also observed in several

other higher plant species; however, no further characterization of the transcripts was conducted (Lisch *et al.*, 2001, Singer *et al.*, 2000; also see Chapter 3).

The *mudrA* transcripts from differentially-splicing process presumably encode a group of MURAs (Walbot and Rudenko, 2001). However, only MURA-823 (120 KD) was further characterized to be a nuclear and *Mu*-binding protein (Benito and Walbot, 1997, Walbot and Rudenko, 2001). Like *mudrA*, *mudrB* also produces a pool of MURBs (Lisch and Freeling, 1995; Lisch *et al*, 1999; Walbot and Rudenko, 2001). The majority of the MURB proteins are nuclear, but some may also exist in cytoplasm (Walbot and Rudenko, 2001). Studies on MURA-like proteins in other plant species are limited: following the identification of a shared conserved *Mutator* domain with MURA, none of the MURA homologues was further characterized (Lisch *et al.*, 2001; also see Chapter 4).

2.4.3 Transposition and regulation

The production of high forward mutations *per* generation is a remarkable feature of the *Mu* system. (Bennetzen, 1996; Walbot, 1992). Such a high rate of mutagenesis is caused by (1) multiple copies of MuDRs in an active *Mutator* line, (2) imprecise *Mu* excisions, (3) preferential insertions near/into host genes and (4) broad chromosomal rearrangements induced by *Mu* transposition (Chandler and Hardeman, 1993; Bennetzen, 1996; Walbot and Rudenko, 2001). Insertion of a *Mu* element into a maize gene can disrupt its function and produce phenotypic changes. Mutant phenotypes, however, can be reverted after precise excision of the element. *Mu*-excision-mediated reversions occur within a relatively narrow window during maize development: mainly in somatic but rarely in germinal cells (Chandler and Hardeman, 1993; Bennetzen, 1996; Walbot and Rudenko, 2001). From an outcross between an active *Mu* and a standard maize line, the

progenies often contain more *Mu* copies than their parents and show unique mutant phenotypes (Bennetzen, 1996).

The features of *Mu* transposition in somatic and germinal development suggest the existence of two transposition pathways: cut-and-paste and replicative transposition (Walbot and Rudenko, 2001). But how the switch is controled? One possibility is that the *Mu* system uses different forms of MURA and/or MURB to control different transposition pathways during maize development (Walbot and Rudenko, 2001). So far, three different forms of MURAs have been identified and confirmed to be able to conduct maize *Mu* activity. (Walbot, personal communication). Similarly, different forms of MURBs were also identified (Walbot, personal communication). However, it is currently not clear (1) which MURA is responsible for the *Mu* activity in germinal cells and (2) what are MURBs' roles in determining the switch. Apart from the proposed model, it has been confirmed that *Mu*/MULE activity was also regulated by host factors, such as DNA methylation and chromosomal position effect (Walbot and Stapleton, 1998; Lisch and Freeling 1994; Singer *et al.*, 2001; see Chapter 4). Finally, host proteins were also found to be able to interact with *Mu1* TIRs, suggesting the possible involvement of specific host protein(s) in the regulation of *Mu* activity (Zhao and Sundaresan, 1991).

2.4.4 Mu/MULE *activity and plant development and evolution*

In active *Mutator* lines, *Mu* activity has been associated with a number of maize gene mutations and chromosomal rearrangement (see review Chardler and Hardeman 1994; Robertson *et al.*, 1994). For example, the insertion of *Mu1* at the transcription start site of maize *Shrunken 1* (*SH1*) gene created a *sh1* mutant line (*sh9026*) where the normal *SH1* transcripts were reduced and truncated *sh1* transcripts were increased (Strommer and

Ortiz, 1989). In another example, *Mu* elements inserted into the *YG2* (*Yellow and Green 2*) gene locus near the end of the short arm of maize chromosome 9 (Roberston *et al*, 1994). From 113 putative *Mu*-induced events, 11 were found to produce albino seedlings or the *white deficiency* (*wd*) phenotype. Further cytological analysis revealed that these mutations were created by *Mu*-induced chromosomal deletions including *wd* locus.

With the exception of the studies described in the following Chapters, the long-term impact of *Mu*/MULE activity on host gene and genome evolution has not been broadly examined. In *Arabidopsis*, MURA-like proteins show sequence similarity with functional FAR1, suggesting an evolutionary relationship between these two (Lisch *et al.*, 2001; see Chapter 4). In addition, several Arabidopsis MULEs also contain an Arabidopsis centromere-specific sequence (see Chapter 4), indicating MULE participation in Arabidopsis centremere formation and function. Finally, MULEs may also play an important role in the generation of eukaryotic genes and multigene families (see Chapter 5).

2.4.5 Applications of Mu/MULE systems

A number of eukaryotic TE systems have been developed as efficient transposon-tagging tools for targeted mutagenesis. In this regard, the *Mu*/MULE system should be extremely beneficial, as they create a high rate of mutations and insert frequently into/near a gene. A number of maize genes were characterized through *Mu*-tagging (Walbot and Rudenko, 2001). Under a standard genetic approach, maize plants were crossed with an active *Mutator* line. Subsequently, *Mu*-insert ional mutations were identified based on visualized mutant phenotypes or a PCR-based molecular approach (Walbot *et al.*, 1986, Walbot, 1992). Recently, a genome-wide *RescueMu* tagging system

in mazie was develped in Dr. Walbot's Lab. With this system, *Mu* insertions can be mapped across the maize genome. The sequences flanking individual insertions can be obtained directly through genomic sequencing (review see Walbot and Rudenko, 2001).

2.5 Conclusions

During the past 50 years, not only have numerous eukaryotic TEs been found and characterized, but also have interesting relationships between mobile DNA and the evolution of eukaryotic gene/genome complexity been revealed. In a eukaryote, TE activity (1) depends not only on TE-specific factors (that is the availability of corresponding transposases) but also on their status within the complex architecture of host genome and (2) can indeed regulate eukaryotic gene functions and contribute to gene/genome evolution. It seems also likely that (1) local condensation of TEs (especially retrotransposons) may have assisted the formation of heterochromatin, (2) TE activity may have facilitated the origin and development of host defense systems and (3) TEs may have played an important role in the construction of fertilization barriers, all of which as a whole could have led to the origin of new species. In conclusion, from an evolutionary standpoint, TEs and eukaryotes can co-evolve.

Hypothesis I

As of 1997, eight *Mu* elements had been identified in maize. The most remarkable characteristics shared by these elements are the *Mu*-TIR (Bennetzen, 1996). Despite the fact of finding *Mu*-transposase (MURA) homologues in rice and their sharing homology with Bacteria IS elements (Eisen *et al.*, 1994), the existence of an intact *Mu*-like element in other higher plant organisms than maize remains unclear.

In early 1998, we initiated a pilot study, aiming to identify the MULEs from four sequenced *Arabidopsis* BAC clones released from the project of sequencing Arabidopsis genome at that period of time. We found a total of four MULE-related sequences: one putative *mudrA*-like gene without a TIR, one MuDR-like element with perfect 9-bp TSDs, and two putative non-autonomous MULEs, also with perfect 9-bp TSDs respectively. Based on this primary analysis, we hypothesized that **there may exist a MULE family in the *Arabidopsis* genome, featuring great diversities in terms of element termini and the structure and sequence constitution of the *mudrA*-like genes** (proposed in my first committee meeting in 1998).

CHAPTER 3

MUTATOR-LIKE ELEMENTS (MULES) IN *ARABIDOPSIS THALIANA*: STRUCTURE, DIVERSITY AND EVOLUTION

Zhihui Yu, Steven I. Wright and Thomas E. Bureau

3.1 Abstract

While genome-wide surveys of the abundance and diversity of mobile elements have been conducted for some class I transposable element families, little is known about the nature of class II transposable elements on this scale. In this report, we present the results from analysis of sequence and structural diversity of *MUtator*-Like Elements (MULEs) in the genome of *Arabidopsis thaliana* (Columbia). Sequence similarity searches and subsequent characterization suggest that MULEs exhibit extreme structure, sequence and size heterogeneity. Multiple alignments at the nucleotide and amino acid levels reveal conserved, potentially transposition-related sequence motifs. While many MULEs share common structural features to *Mutator* elements in maize, some groups lack characteristic long terminal inverted-repeats. High sequence similarity and phylogenetic analysis based on nucleotide sequence alignments indicate that many of these elements with diverse structural features may remain transpositionally competent, and that multiple MULE lineages may have been evolving independently over long time scales. Finally, there is evidence that MULEs are capable of capturing host DNA segments, which may have implications for adaptive evolution, both at the element and host levels.

3.2 Introduction

Mutator (*Mu*) is a diverse family of class II transposable elements found in maize. Robertson first identified *Mu* through a heritable high forward mutation rate exhibited by lines derived from a single maize stock (Robertson, 1978). To date, at least six different groups have been identified in maize *Mu* lines (Bennetzen, 1996). *Mu* elements have long (≈200-bp) and highly conserved terminal inverted-repeats (TIRs). However, the internal sequences are often heterogeneous, with little to no sequence similarity with other elements (Chandler and Hardeman, 1992). Upon insertion, *Mu* elements typically generate a 9-bp target site duplication (TSD) of flanking DNA (Bennetzen, 1984; Walbot, 1991). *Mu* Transposition is primarily regulated by *Mu* members designated as MuDR, which contain both *mudrA* and *mudrB* genes (Lisch *et al.,* 1995). *mudrA* has been suggested to encode the *Mu* transposase (MURA; Henikoff *et al.*, 1995; Lisch *et al.*, 1999) and to be related to the transposases of some insertion sequences (*IS*) in bacteria (Eisen *et al.*, 1994), whereas *mudrB* has no known function and shows no significant amino acid sequence similarity with any reported protein (Lisch *et al.*, 1999). As with other mobile elements, some *Mu* elements lacking a functional transposase are capable of transposition, if MURA is supplied *in trans* (Hershberger *et al.,* 1991). *Mu* elements in maize have been demonstrated to be an extremely active agent in creating mutations and have been developed as a highly efficient transposon-tagging tool for maize gene isolation (Walbot, 1992). Despite identification and primary characterization of *Mu* elements in maize, relatively little is known about their distribution, diversity and evolution in other higher plant species.

Arabidopsis thaliana has become a model organism for genetic analysis of many aspects of plant biology and is the first plant species to be targeted for complete genome sequencing (Meinke *et al.,* 1998). This sequence information provides an exceptional opportunity to identify mobile elements and characterize their patterns of diversity at the whole-genome level. The *Arabidopsis* genome has been recently shown to harbor a great number of transposable elements, including various repetitive sequences with amino acid and structural similarity to *Mu* (Le *et al.*, 2000; Lin *et al.*, 1999; Mayer *et al.*, 1999). In this report, we analyze the structural diversity and phylogenetic relationships of the *Mutator*-like element (MULE) groups containing member(s) encoding a putative MURA-like transposase in *A. thaliana.*

3.3 Materials and methods

Data mining Sequences surveyed in this study correspond to selected large-insert DNA clones from the *Arabidopsis* genome project, as described by Le *et al* (2000). Specifically, sequenced clones released before December of 1998 were chosen for systematic screening and classifying MULEs. Additional elements were then periodically mined up to Dec. of 1999. Two computer-based approaches were employed to identify MULEs. The first method involved using *Arabidopsis* genomic sequences as queries in Advanced BLAST searches, as described by Le *et al* (2000). In addition, each DNA segment (typically the sequence from one large-insert clone) was compared against its reverse complement using the program BLAST 2 Sequences (Tatusova and Madden, 1999) to identify long TIR structures. Elements were classified into groups based on shared sequence similarity (BLAST score > 80). Long TIRs were defined as terminal-most regions sharing > 80% sequence identity over ≥ 100 contiguous base pairs. A

detailed description of the mined MULEs presented in this report can be accessed on the Worldwide Web at http://soave.biol.mcgill.ca/clonebase/.

Sequence analysis and molecular cloning Both PCR- and computer-based approaches were employed to document past transposition events and to confirm the position of termini for some elements by identifying RESites, that is sequences which are related to empty sites (Le *et al.*, 2000). In the PCR-based protocol, genomic DNA was isolated from ten ecotypes of *A. thaliana*: No-0, Sn-1, Ws, Nd-1, Tsu-1, Rld-1, Di-G, Tol-0, S96, and Be-0 (Arabidopsis Biological Resource Center; http://aims.cps.msu.edu/aims). PCR primers were designed corresponding to regions flanking putative MULEs. A primer name was composed of three parts, namely i) ATC (*Arabidopsis thaliana* clone), ii) the GI number of the clone harboring the MULE, and iii) the corresponding position in the clone where the primer sequence was derived. The primer pair used to amplify RESites for MULE-1:GI2182289 was ATCGI2182289-38427 (5'-GTGAGGCAACACGTCATCATCTC-3') and ATCGI2182289-40214 (5'-CTGGTCTTGAACCTCGTTCATCC-3'); for MULE-23:GI3063438 was ATCGI3063438-86192 (5'-CCACCTTTAATCCGGGAGAATTC-3') and ATCGI3063438-99055 (5'-CACGATGGAACTCCAGTCAG-3'); and for MULE-24:GI2760316 was ATCGI2760316-88054 (5'-CATGTAACCCTTCATGGGTGG-3') and ATCGI2760316-93177 (5'-TGGGATTCCAATTTGTCAGCCTG-3'). PCR amplifications were carried out using annealing temperatures ranging from 50-65° as previously described (Bureau and Wessler,1994). Amplified fragments were cloned into a pCR2.1 vector (Invitrogen, Carlsbad, CA) and subsequently sequenced using a SequiTherm EXCEL™ II kit (Epicentre, Madison, WI). The resulting DNA sequences

were compared with the corresponding sequences at element insertion sites to confirm the position of element termini and TSDs. Alternatively, the regions flanking putative MULEs were used as BLAST queries to identify related sequences which lacked the element (Le *et al.*, 2000).

Information concerning the position and identity of putative ORFs within mined MULEs was inferred from the annotation of surveyed clones. Multiple sequence alignments of mined MULEs were performed using DIALIGN 2.1 (http://bibiserv.techfak.uni-bielefeld.de/dialign; Morgenstern, 1999), as it has less limitation than other computer programs (e.g., MULTALIN) for aligning very large sequences, such as the elements over 10-kb. Within-group sequence similarity was displayed using PlotSimilarity, implemented with GCG (Wisconsin Package Version 10.0, Genetic Computer Group (GCG), Madison, Wisc). MULE termini were analyzed using MULTALIN (http://www.toulouse.inra.fr/multalin.html, Corpet 1988) and their consensus sequences were derived from these analyses. ProfileScan (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html; Gribskov *et al.*, 1987) and Pfam HMM Search (http://pfam.wustl.edu/hmmsearch.shtml; Bateman *et al.*, 2000) were used to determine the location of conserved domains and motifs. Transposable elements within MULEs were mined using the methodology previously described (Le *et al.*, 2000). Analysis of substitution patterns, and determination of significant deviation from neutral expectations (i.e. Ka/Ks=1) were generated using the program K-Estimator (version 5.3; Comeron, 1995; 1999). Sliding window analysis of sequence diversity (calculated as □, the average pairwise difference) across aligned sequences was conducted using the program DnaSP (version 3.14; Rozas and Rozas, Phylogenetic Analysis Maize *mudrA*

and *Arabidopsis mudrA*-like ORFs were compared by pairwise (BLASTX) and multiple alignment (MULTALIN) to identify the most conserved region to be used in phylogenetic analysis.Using the maize *mudrA* as an outgroup, phylogenetic trees were derived from both distance-based (neighbor-joining) and character-based (parsimony) approaches using programs in the PHYLIP package (version 3.75c; Felsenstein, 1993). Nucleotide distances were computed using the Kimura option of DNADIST. SEQBOOT was used to generate 100 bootstrap replicates, each of which was then analyzed by NEIGHBOR and DNAPARS. The final majority-rule consensus trees were derived using CONSENSE.

3.4 Results

Among 28 mined MULE groups (Le *et al.*, 2000), nine were revealed to contain element(s) encoding a putative protein sharing approximately 25% similarity to MURA in maize. None of the elements were found to harbor a *mudrB*-like ORF. Table 3.1 summarizes the primary features and diversity of these groups. By analyzing flanking DNA sequences between an insertion and its corresponding RESite, both the location of MULE termini and TSDs were confirmed for representative members of eight of the nine MULE groups (Figure 3.1). Moreover, this analysis provides convincing evidence that the mined MULEs are indeed transposable elements (Lee *et al.*, 1999).

Diversity of MULEs Among the nine MULE groups, six contain elements with TIRs (TIR-MULEs, Table 3.1). In general, the TIR-MULEs are structurally similar to *Mu* elements in maize (Bennetzen, 1996), with long TIRs (100 to 408-bp) and typically 9-bp TSDs (among the surveyed elements 49% have 9-bp TSDs, 39% have 10-bp TSDs, 5% have TSDs larger than 10-bp, and 7% have TSDs shorter than 9-bp). As with *Mu*

elements in maize, members within most TIR-MULE groups share the highest sequence similarity only at the TIRs (Figure 3.2), and most lack coding capacity. Significant variation in element abundance is observed among MULE groups. For example, only one member was identified for the MULE-16 group in our survey, compared to 20 members in the MULE-1 group. Within the latter group, 12 members share >90% sequence identity across their entire sequence. They share similarity only with the TIR sequences of the other eight members in the same group.

Although the three other MULE groups also contain elements encoding MURA-like proteins, and 77% of the members within these groups have a 9-bp TSD (Table 3.1, Figure 3.1), MULEs in these groups display characteristics that have not been observed. Namely, (i) the 5' terminus and inverse-complement of the 3' terminus of these individual elements exhibit much lower (<60%) sequence similarity than both the TIR-MULE groups and the *Mu* elements in maize, which typically display >80% sequence similarity between a given element's TIRs (Figure 3.3; Walbot, 1991; Chandler and Hardeman, 1992; Bennetzen, 1996), (ii) the majority of the members are very large in size, ranging from about 7.1-kb to 19.4-kb, (iii) members within a group share relatively high sequence similarity across their entire length (up to 95%) (Figure 3.2), and (iv) approximately two-thirds of these elements contain two or three ORFs, one of which encodes a MURA-like protein, but the others lack high sequence similarity with any characterized ORF in the current database (Table 3.1). Given their consistently low sequence similarity at their termini, we refer to these elements as non-TIR-MULEs.

In addition to structural, size, and element-abundance variation, we also found evidence indicating the apparent acquisition of host DNA segments into the internal

regions of some TIR-MULEs (Figure 3.4, .5, .6, .7, and .8). The size of the acquired DNA fragments range from 97-bp to 492-bp and make up the major portion of the internal sequences of the corresponding MULEs. The acquired DNA sequences are 85-88% identical to the original host DNA segments (Figure 3.4B, .5B, .6B, and .8B; Le *et al.*, 2000). They include 5' flanking sequences of the genes (Figure 3.4A, .5A, and .8A), 5'UTR (Figure 3.4A), exons (Figure 3.4A, .5A, .6A, and .7A), and introns (Figure 3.6A and .7A). It is evident that all acquired gene segments identified are related to the 5'-regions of host transcription factors or developmentally regulated genes.

With one exception, MULE-1:GI2182289 (chromosome 1), the acquired gene sequences do not form ORFs (Figure 3.4, 3.5, 3.6, 3.7, and 3.8). The structure of this element has been previously reported (Le *et al.*, 2000) to share significant sequence similarity with a region spanning the first two exons and first intron of the *Arabidopsis* homeobox-leucine zipper gene *Athb-1* (Ruberti *et al.*, 1991; also referred to as *HAT5* [Schena and Davis, 1994]; Figure 3.7A). The acquisition of an *Athb-1* gene segment results in the formation of a novel putative ORF (Figure 3.7B) encoding a 71 aa polypeptide. This putative protein shares 88% amino acid sequence similarity (Figure 3.7C) with the N-terminal sequence of the Athb-1 that includes an acidic domain (Figure 3.7B). Analysis of sequence diversity across the region of similarity between the putative gene from MULE-1:GI2182289 and the *Athb-1* gene indicates that noncoding regions have diverged more extensively than exons (Figure 3.7D). Calculation of substitution patterns between these two ORFs using the method of Comeron (1995) provides an estimated ratio of nonsynonymous to synonymous substitutions (Ka/Ks) of 0.6733, which is not significantly different from 1 ($p>0.05$). Subsequent analysis has also revealed a

second MULE-1 (GI613649; chromosome 4) with high nucleotide similarity to the same region of *Athb-1* (Figure 3.7A). The *Athb-1*-like region of MULE-1:GI613649 has numerous frameshifts and stop codons relative to *Athb-1* (Figure 3.7C), but the reconstructed amino acid sequence shares 80% similarity to the same region of Athb-1. As with the initially identified segment, a region corresponding to the location of the first intron of *Athb-1* is also present. No expression information of the putative gene in MULE-1:GI2182289 was identified through a survey of the *Arabidopsis* EST database.

Finally, numerous nested transposon insertions also contribute to the diversity of MULE elements. As described in Table 3.2, both class I and II transposable elements have been identified within MULEs. These insertions have variable sizes (ranging from about 0.73-kb to 6.67-kb), display either TIR or LTR structures, and two contain putative transposition-related ORFs. In addition, a novel putative TE insertion was also identified in MULE-23:GI6007863. This sequence has 325-bp long TIR structure and is flanked by a 5-bp direct repeat (table 3.2). Its internal sequence has coding capacity for a *Ty3/gyspy*-like retrotransposon-related protein that is 75% identical to a putative retroelement integrase in *A. thaliana* (Lin *et al.*, 1999), and 42% identical to a characterized probable polyprotein in *A. comosus* (Thomson *et al.*, 1998) in BLASTX surveys. This putative insertion element may reflect a novel class II element that has sustained the insertion of a truncated *Ty3/gyspy*-like retrotransposon. Alternatively, this sequence may represent a novel type of terminal inverted-repeat containing retrotransposon (Zuker *et al.*, 1984; Garrett *et al.*, 1989)

Conserved sequence motifs Figure 3.9 shows the terminal consensus sequences for each of the nine MULE groups. Overall the terminal sequences share no significant

sequence similarity among groups. However, many of the terminal-most sequences tend to fit the general motif 5'-RRR-3' (R=G or A) followed by a short AT-rich region (Figure 3.9). No sequence significantly similar to the maize MURA binding site (5'-GAGGGAAGGGGATTCGACGAAATGGAGGCGTT-3'; Benito and Walbot, 1997) was identified within any of the consensus sequences.

The MURA-like proteins encoded by the mined MULEs were also analyzed for DNA-binding motif(s). Using Profilescan and Pfam HMM, we identified a motif, **CX2CX4HX4C** (X represents any amino acid), at the C-terminal region of sixteen *Arabidopsis* MURA-like proteins (Figure 3.10). This motif also exists in a rice MURA-like protein, a number of known nuclear binding proteins (NBP), and other transposases (Figure 3.10). The C-terminal region of maize MURA has a similar motif, **CX2CX4HX6C**. Analyses of the N-terminal regions of the putative MURA-like proteins did not show significant sequence similarity to any known protein.

Phylogeny of TIR and non-TIR-MULEs A conserved region (~270 nucleotides) was identified within the maize *mudrA* and the *Arabidopsis mudrA*-like genes (Figure 3.11) and used for phylogenetic analysis of the mined MULE groups. We utilized two methods, neighbor-joining and parsimony, to establish evolutionary relationships. Using maize *mudrA* as an outgroup sequence, both methods generated majority-rule trees with similar topologies. The consensus tree derived by the neighbor-joining method is shown in Figure 3.12. These phylogenetic relationships are consistent with our classification of MULE groups based on blast search results, since elements from one group are monophyletic, with high bootstrap support (>93%), and are separated by much shorter branch lengths than found between groups. The phylogeny also indicates that the mined

non-TIR-MULE groups are more closely related to each other than they are to any of the TIR-MULE groups, and that non-TIR MULE elements which encode MURA-like proteins have undergone recent amplification.

3.5 Discussion

Genome sequencing projects allow for detailed analysis of the patterns and extent of transposon diversity in the genomes of model organisms. Our data suggest that the MULE groups in *A. thaliana* exhibit both extreme structural and sequence heterogeneity. In fact, the observed variation indicates that the MULE superfamily may be among the most diverse mobile element families in eukaryotes. The presence of element insertions of varying ages may partly account for MULE diversity. The existence of numerous truncated MULEs (Le *et al.*, 2000; Lin *et al.*, 1999; Mayer *et al.*, 1999), and the high level of divergence between MULE groups indicates that these elements might be an ancient mobile element system in the *Arabidopsis* genome, and many elements may be no longer be transpositionally active. However, the existence of MULEs with significant sequence identity (>90%), and the identification of RESites from the closely related ecotype No-O suggests that many MULEs have in fact been recently mobile. The high level of diversity may also reflect the potential ability of MULEs to remain transpositionally competent with the presence of only a few conserved sequence motifs, if a transposase is supplied *in trans*.

Non-TIR-MULEs are a novel type of plant class II transposable element. In contrast to TIR-MULE groups, as well as *Mu* elements in maize, these elements are characterized by low sequence similarity between termini of individual elements. One might expect that the absence of high sequence similarity at the termini of individual

non-TIR-MULEs suggests that they represent truncated, and presently inactive, elements. However, non-TIR-MULEs are also characterized by their abundance in the genome, high level of homogeneity (up to 99.5%) between members of individual groups, and a relatively high frequency of elements encoding putative MURA-like transposases. These features, combined with phylogenetic analysis indicate that these elements are able to transpose in the absence of long TIRs, and that they may be evolving as an independent lineage. Similar patterns of structural diversity have been observed in a family of unusual *IS* elements (such as *IS*901, *IS*116, and *IS*902; Ohtsubo and Sekine, 1996). These elements have been subgrouped on the basis of presence or absence of TIRs, and have been hypothesized to be regulated by a shared group of transposases. The origin of non-TIR-MULEs, the forces maintaining structural diversity among MULEs, and the functional implications of this diversity for the MULE system in the *Arabidopsis* genome remain unknown.

It seems that acquisition of host DNA sequences to assemble new elements is a frequent event for TIR-MULEs. In addition to our documentation of five acquisition events in *Arabidopsis*, the maize *Mu*2 has also been reported to have acquired a host MRS-A DNA segment (Talbert and Chandler, 1988; Talbert *et al.*, 1989). While the acquisition events by *Arabidopsis* TIR-MULEs involved the 5' ends of cellular genes, the significance of this bias is currently unknown. Acquisition of cellular genes does not appear to necessarily prevent transposition since two MULE-1 elements harboring *Athb-1*-like ORFs on different chromosomes have been identified. Class I elements have also been documented to acquire or transduce cellular genes (Bureau *et al.*, 1994; Boeke and Stoye, 1997). These genes can be expressed by means of a LTR-promoter and in many

cases lead to disease phenotypes (Vogt, 1997). Likewise, acquired and modified host DNA within MULEs could be expressed from either a TIR-promoter, an acquired promoter, or a promoter in the flanking region. However, there is currently no evidence that the putative ORFs are actually expressed *in vivo* or whether these polypeptides have any function. While there is evidence for a lower level of divergence between the MULE-1 *Athb-1*-like gene and *Athb-1* in coding regions, it is unclear whether this pattern reflects selective constraint only on *Athb-1*, or whether there are in fact functional constraints on the coding region of the MULE-1 gene. Furthermore, the Ka/Ks ratio does not provide a strong indication of departure from neutral patterns, suggesting that the acquired exons may be nonfunctional. The ability to capture sequences from their host may not only a mechanism for generating element diversity, but might also be important in generating adaptive changes for MULE groups. On the other hand, considering that genomic DNA segments captured by *Mu* elements and MULEs can transpose, likely be duplicated by means of replicative transposition, and recombine with sequences encoding functional domains, these elements might also play important roles in gene organization and evolution (Henikoff *et al.*, 1997).

We have identified a motif, namely 5'-RRR-3' followed by a short AT-rich region, at the terminal-most ends of the mined MULEs. This motif is reminiscent of a motif (5'-GDTAAA-3'; D=G, T, or A) found in the subterminal regions of the maize *Ac* element which was demonstrated to be the recognition sites for the binding of nuclear proteins in maize (Becker and Kunze, 1996) and tobacco (Levy *et al.*, 1996). The MULE terminal motif is also similar to part of a region (5'-CGGGAACGGTAAA-3') located in the maize *Mu1* TIR that is also recognized by host factors (Zhao and Sundaresan, 1991).

In fact similar motifs have been recognized in a variety of class II plant transposable elements (Levy *et al.*, 1996). It is tempting to speculate that the motif identified in our study may function as a *cis*-acting sequence in regulation of MULE activity.

We have also identified a **CX2CX4HX4C** motif at the C-terminal region of some MURA-like transposases. This motif is characteristic of the zinc finger domain (**CX2-5CX4-12C/HX2-4C/H**) found in RNA binding proteins (Hanano *et al.*, 1996; Rajavashisth *et al.*, 1989), eukaryotic transcription factors (Sherrie and Wolffe, 1993), and transcription splicing factors (Heinrichs and Baker, 1995; Lopato *et al.*, 1999). In addition, this motif has also been found within the *gag*-encoded genes of retroviruses (Berg 1986; Covey, 1986) and retrotransposons, such as *copia*-like retrotransposons from tobacco (Grandbastien *et al.*, 1989), and *Ty* elements in yeast (Jordan and McDonald, 1999). The **CX2-5CX4-12C/HX2-4C/H** motif can be present in a protein sequence from one copy to as many as nine copies (Sherrie and Wolffe, 1993). It has been demonstrated that this motif interacts directly with viral RNA (Covey 1986; Darlix *et al.*, 1995), eukaryotic pre-mRNAs (Fu, 1993; Heirichs and Baker, 1995; Lopato *et al.*, 1999) and single-stranded DNA (Rajavashisth *et al.*,1989; Remacle *et al.*, 1999). Given its RNA- and DNA-binding characteristics, the **CX2CX4HX4C** motif at the C-terminal region of the putative MURA-like transposases might interact with the MULE DNA and/or RNA, possibly playing a role in MULE transposition and/or regulation of MULE mobility in *A. thaliana.*

The initial discovery of *Mu* involved the isolation and characterization of various elements in maize. In this study, we have characterized the sequence and structural diversity of MULEs in *A. thaliana*, thereby extending the range of members of the

MULE superfamily. The apparent success of MULEs in the *Arabidopsis* genome provides an excellent opportunity for learning about the mechanisms driving the diversity and evolution of a class II transposable element system in eukaryotic genomes. The *Mutator* system in maize is a highly effective agent for the creation of *de novo* mutations. In fact, *Mu*-tagging approaches have been extremely effective in the isolation and functional analysis of numerous maize genes (Maes *et al.*, 1999; Walbot, 1992). Introduction of active *Mu* elements into heterologous plant species, however, has not been successful (Walbot, 1992). Identification and characterization of MULEs in species other than maize may thus also facilitate the development of an endogenous element tagging approach.

Table 3.1 Summary of mined MULE groups in *A. thaliana*

group	no. of elements	size range (bp)	TIR size range (bp)	no. of elements with *mudA*-like gene[a]	TSD size (bp)
MULE-1	20	492-3952	103-408	1	9-12
MULE-2	9	444-4809	101-222	1	7-11
MULE-3	2	1213-3771	107-158	1	10
MULE-16	1	3646	292	1	6-7
MULE-24	7	1075-4445	100-319	2	9-10
MULE-27	7	552-4703	141-307	1	9-11
MULE-9	16	2338-17078	na[b]	7	9
MULE-19	4	7119-8188	na	4	8-9
MULE-23	6	12267-19397	na	6	9-8

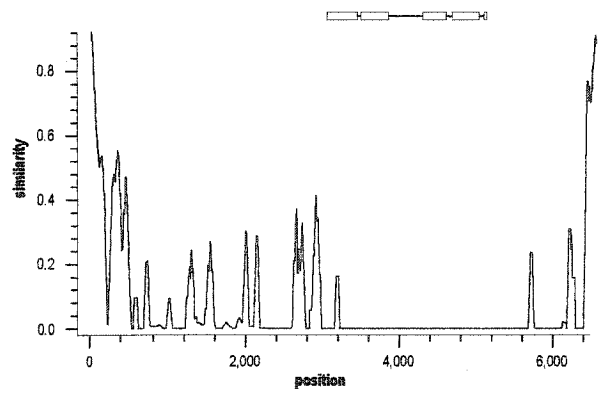
[a]Only one putative *mudrA*-like gene was identified per element.
[b]not applicable.

Table 3.2 Other TE insertions into the MULEs

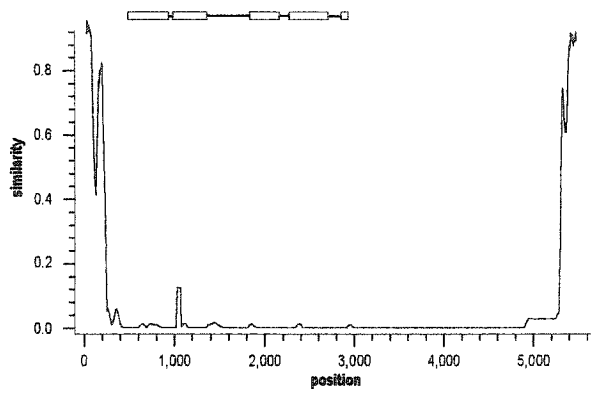
Inserted MULE	TE type	position	size (bp)	coding capacity	TIRsize (bp)	TSD
MULE-9: GI3299824	MULE-3	86859-87925	1066	none	158	gtatgtacct
MULE-9: GI3299824	*En/Spm*	91459-95242	3783	*En/Spm*-like transposase	13	ggt
MULE-9: GI6136349	solo-LTR (*Athila*)	12128-14273	2146	none	5	ccatt
MULE-9: GI3128140	*Tag-1*	50787-51517	731	none	21	cttatgag
MULE-23: GI6007863	unknown	119225-125890	6666	gag-pol polyprotein	325	atttg
MULE-23: GI6007863	solo-LTR (*Tat1*)	117197-118083	983	none	5	ataag

Figure 3.1 RESites of some mined MULE group members. The MULE-associated TSDs are underlined. GI (geninfo) numbers and nucleotide positions in corresponding clones or sequenced DNA segment from *A. thaliana* ecotype No-0 are indicated. Due to the degenerate TSDs of MULE-16, the RESite corresponding to this element does not resolve the precise termini or TSD.

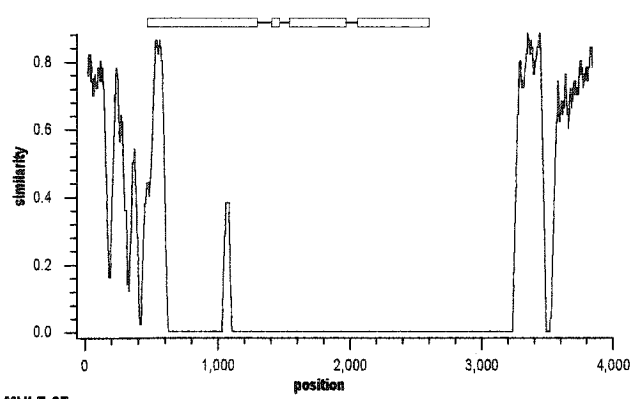
MULE-1



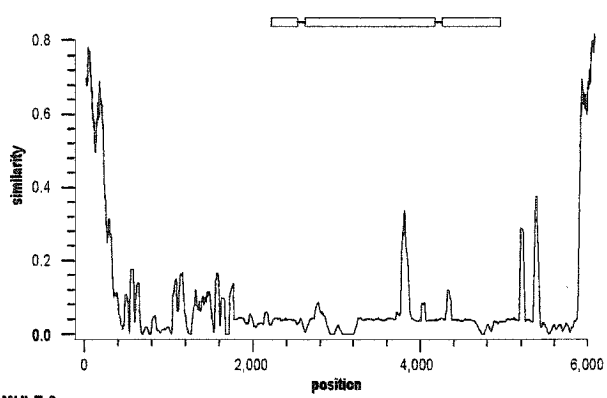
MULE-2



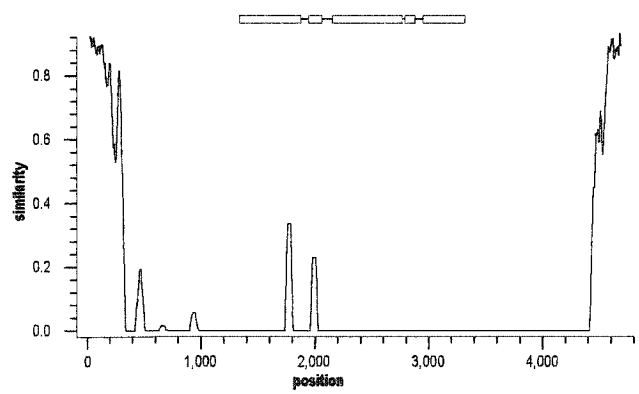
MULE-3



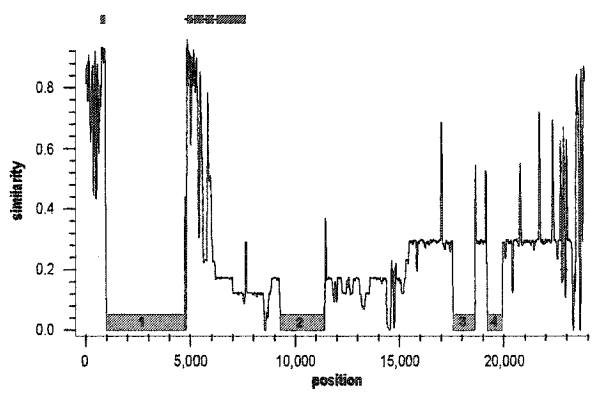
MULE-24



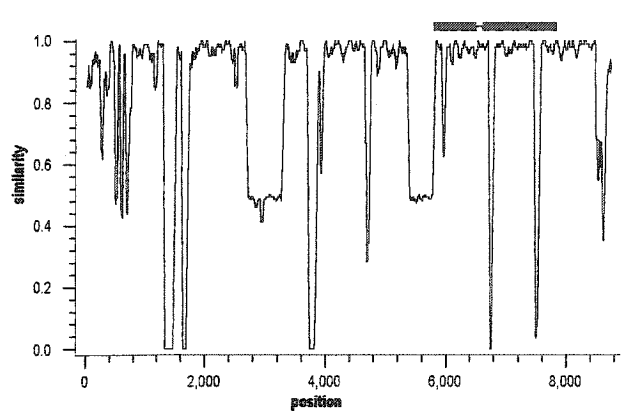
MULE-27



MULE-9



MULE-19



MULE-23

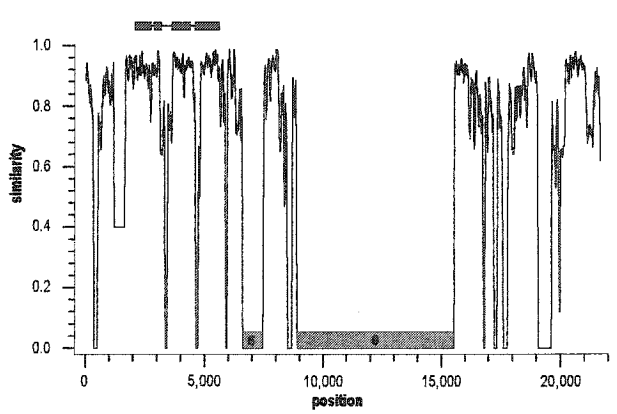


Figure 3.2 Similarity plot of multiple sequence alignments of individual MULE groups (sliding window size: 50-bp). Both nucleotide and indel variation led to a reduction in similarity estimates. The approximate positions of *mudrA*-like genes in individual multiple alignments are indicated with the black bars over the corresponding positions. The shaded regions in MULE-9 and -23 represent the sites where other TE insertions (see Table 3.2) were identified (1: insertion of an *En/Spm*-like element, 2: insertion of an *Athila*-like solo-LTR element, 3: Insertion of a MULE-3 element, 4: insertion of a *Tag-1*-like element, 5: insertion of a *Tat1*-like solo-LTR, and 6: insertion of an unclassifiable element which contains a truncated *Ty3/gypsy*-like integrase domain.). As only one member was identified for MULE-16, no alignment result was available.

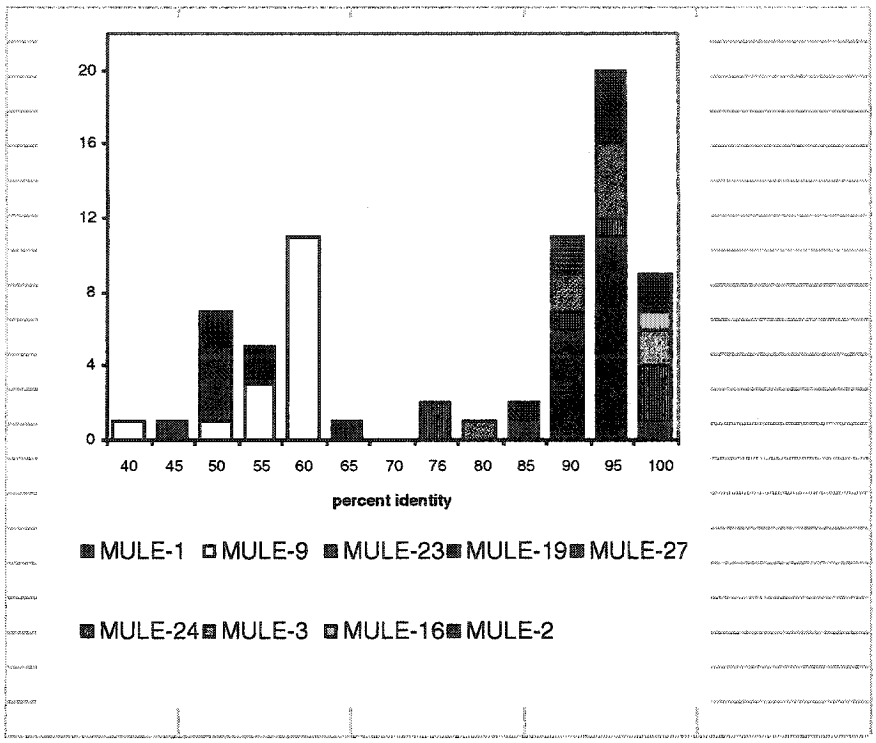


Figure 3.3 Frequency distribution of sequence similarity at the termini of each individual MULE element. The first 100-bp of each element were aligned to the reverse-complement of the last 100-bp, and the percent similarity calculated. MULE-9, -19, and -23 are non-TIR MULEs, while MULE-1, -2, -3, -16, -24 and -27 are TIR-MULEs.

Figure 3.4 Acquisition of *AtHsp101* gene segment by MULE-24:GI3193305. **a**, illustration of *AtHsp101* gene (GI6715467, Hong and Vierling, 2000) and the element structure. The black boxes represent the exons, the white boxes represent the introns, the putative 5'UTR region is labelled, and the dark grey box represents the 5' flanking region, the light grey boxes with an arrow represent the TIRs, and the dash-lined boxes represent the remaining internal sequence of the element. **b**, nucleotide sequence similarity between the acquired DNA segment and corresponding host DNA sequence, with the identical nucleotides being shaded in grey. No complete ORF was identified within the aligned region. **c**, amino acid sequence similarity between AtHsp101 protein segment, encoded by *AtHsp101*, and M-AtHsp101, encoded by the element (derived using BLASTX).

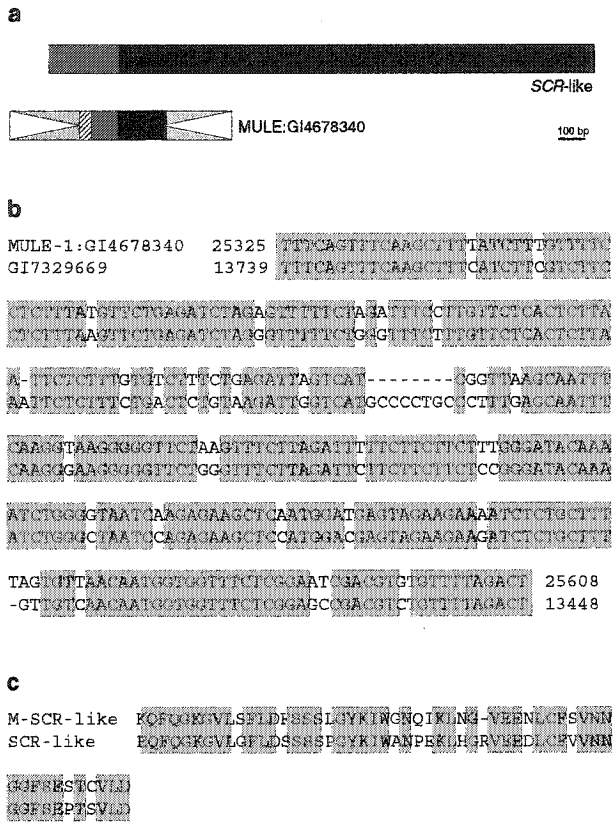


Figure 3.5 Acquisition of *SCR-like* gene segment by MULE-1:GI4678340. **a**, illustration of *SCR-like* gene (annotated as T4C21-40 in the clone, GI7329669) and the element structure. The black boxes represent the exon, the dark gray box represents the 5' flanking region, the light grey boxes with an arrow represent the TIRs and the dash-lined box represents the remaining internal sequence of the element. **b**, nucleotide sequence similarity between the acquired DNA segment and the corresponding host DNA sequence, with identical nucleotides being shaded in grey. No complete ORF was identified within the aligned region. **c**, amino acid sequence similarity between the SCR-like protein segment, encoded by SCR-like gene, and M-SCR-like protein, encoded by the element (derived using BLASTX).

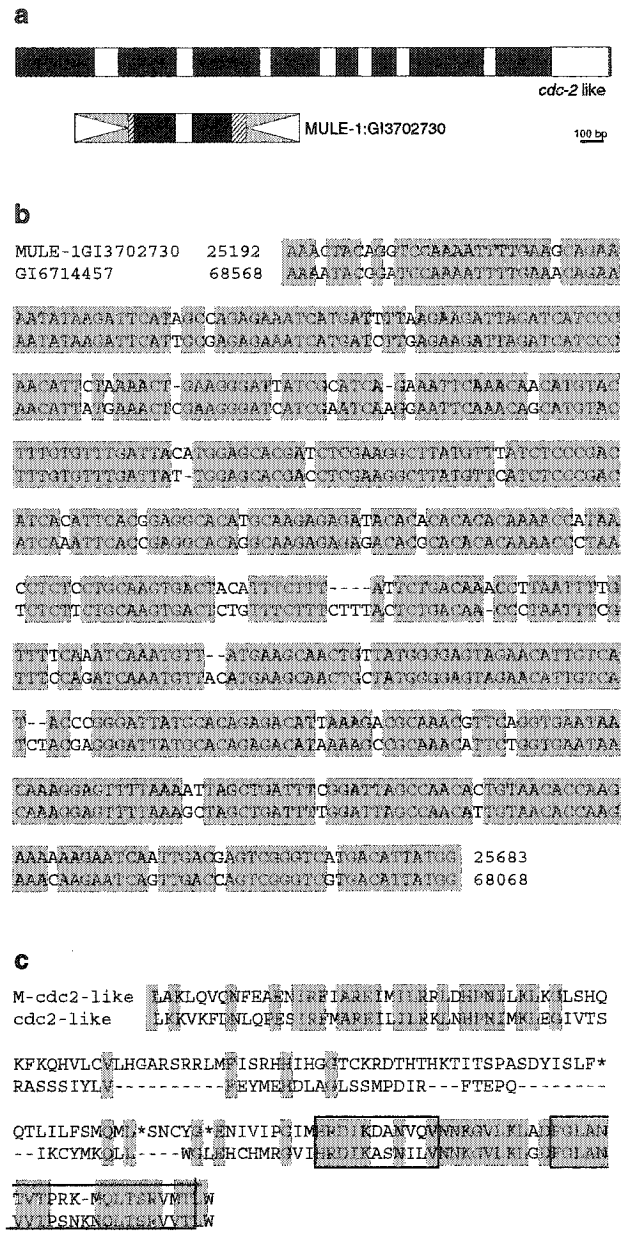


Figure 3.6 Acquisition of *CDC*-like gene segment by MULE-1:GI3702730. **a**, illustration of *CDC*-like gene (annotated as T4P13.23 in the clone, GI6714457) and the element structure. The black boxes represent the exons, the white boxes represent the introns, the light grey boxes with an arrow represent the TIRs, and the dash-lined box represents the remaining internal sequence of the element. **b**, nucleotide sequence similarity between the acquired DNA segment and the corresponding gene sequence, with the identical nucleotides being shaded in grey. No complete ORF was identified within the aligned region. **c**, amino acid sequence similarity between the cdc-like protein segment, encoded by the *CDC*-like gene, and the M-cdc-like protein, encoded by the element (derived using BLASTX). The conserved motifs of the cdc-like protein (Magyar *et al.*, 1997) are boxed. Asterisks represent positions where a frame shift was introduced to achieve an optimal alignment.

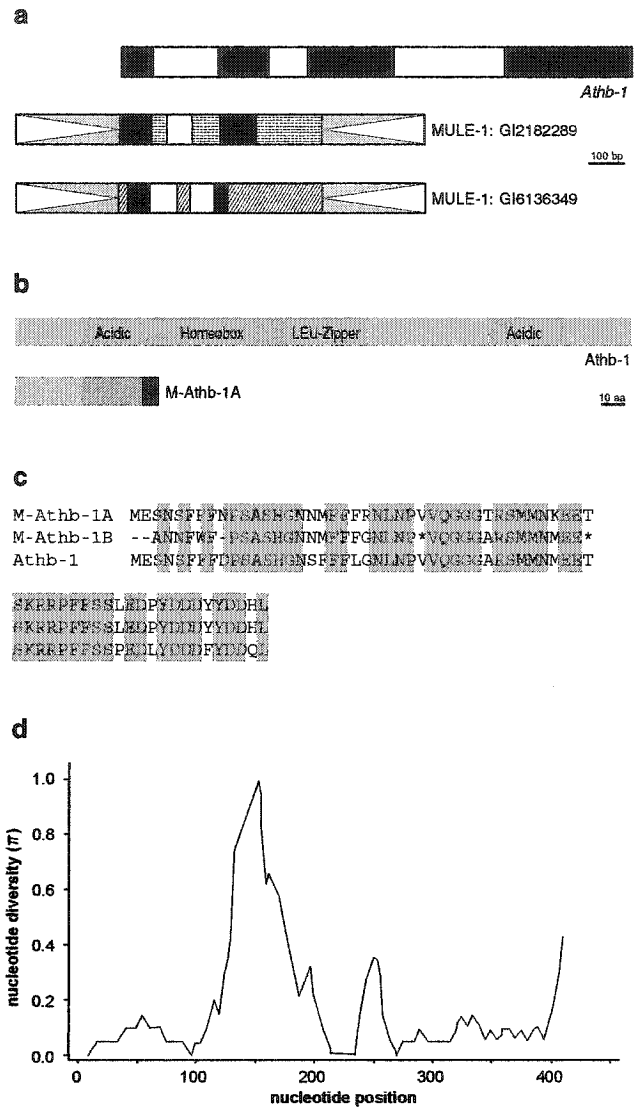


Figure 3.7 Acquisition of *ATHB-1* gene by MULE-1:GI2182289 and MULE-1:GI6136349. **a**, illustration of *ATHB-1* gene and the element structures. Black boxes represent exons; white boxes represent introns; grey boxes with arrows represent TIRS; slash-lined boxes represent the internal region of MULE-1:GI6136349; and dash-lined boxes represent the internal region of MULE-1:GI2182289. The corresponding DNA sequences present in both dashed and slashed boxes have sequence similarity <50%; the corresponding sequences present in grey boxes have sequence similarity >80%; and the DNA sequences present in both black and white boxes of the elements have >86% sequence similarity with the corresponding DNA sequence in the *ATHB-1 gene*. **b**, structural relationship between the Athb-1 and the putative protein, M-Athb1A. **c**, multiple alignment of the amino acid sequence shared between the putative protein encoded by MULE-1:GI2182289 (M-Athb-1A), the derived polypeptide from MULE-1:GI6136439 (M-Athb-1B) and the N-terminal region of the Athb-1. Identical amino acids are shaded in grey. Asterisks represent positions where a frame-shift was introduced to achieve an optimal alignment. **d**, sliding window of nucleotide sequence diversity (þ) across the region of similarity between MULE-1:GI2182289 and *ATHB-1*. Sequences corresponding to an intron are located between positions 88 and 267 while the remaining regions correspond to exons.

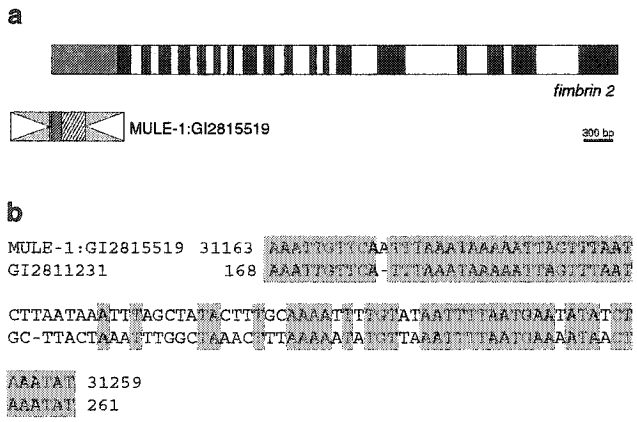


Figure 3.8 Acquisition of 5'-flanking DNA sequence of *FIMBRIN 2* gene by MULE-1:GI2815519. **a**, illustration of *FIMBRIN 2* gene (GI2811231) and the element structure. The black boxes represent the exons, the white boxes represent the introns, the light grey boxes with an arrow represent the TIRs, and the dash-line box represents the remaining internal sequence of the element. **b**, nucleotide sequence similarity between the acquired DNA segment and the corresponding host DNA sequence, with the identical nucleotides being shaded in grey. No complete ORF was identified within the aligned region.

MULE-1
GGGAAAAAAAACGAGAAAATGTCATTTAATCCCCCAACTTTCAAAAAATAGGTCATTTTATACATCAACTTCGTATGTGGCCGTTTAAA
ACATGAACTAAACGTTGACTAATTATTTAAAACATGATTATTCGTTGACCAGGCCAAAATAGAATGCCGTTATCAGTCATTAACGGAAC
TTNTAACTCCGTTANCTAACGTCCGTTATCATCCGTTAACGCGTCGTTTNACTATAATCGAAACANGNAAANTGTCACTTTATTAGTCA
AGTTTCAAATTGTGGCCATTTTAAACATCAACTTCGTTGACCAAGCCAAAATAGACATGTAAATCCCTAAATTCTAAAATTAAAATCTT
GGT

MULE-2
GGGGAAAAATGTTATTTAATACCTGAACTTTCAAAAAATGGCCAAATTAACCGTGAACTCTTGAAATGGCCGTTTTATACCTCAACAAA
AAGTTGACTTCTAATTTTAACCTATAAGTTATCGTTGACTCGGCCTAATNAACCACCGTTAAAAATCCTTCTAACAGCGTAATTNACAG
CCGTTANAGAACATCCGTTAAGGTGATACGATGACGTTTTGTGCTA

MULE-3
GGAAAAATTTTGAATCGGTCTATTTCCCCANGAGAACTATNACATCATGCCAGATANAGCCCGATCTTTGACCTCTNTCTATNTACCCA
CGAAANGAAAGTTAAANGTCTANTTCCCCAGAAATNAAAGTTCNAGAGCTAAATGTCCCCGNTTNNNNATTTCGTC

MULE-16
GGGCTNCCTATNGGAGNCCCCCTCAANTATATTGTTTCAACTGGGTTCCACCCAAACTCTTAGCCTCTTTCTATTTCCCCCTAATCAAC
TTAAATTCTTCTCAATTCCACCTAATATATATAATCCCACGTAATTAACCAACATTAATCAAGCTGCTAAAATNCGATTTAAAATCATA
AAATCACGTTTTAAAATTAATTAGCCTCGATAAACCGATAATCCTACCATAATCTATTAACCCAAATTGAAGCCACTGTATGAGAAAGA
GGCTTCAAATTTTCTTCTTAAAACCCAAAAC

MULE-24
GGGAAAAAATACAACCCGAAAAAACCTCCATTTATTTTTAATTTGGCCGTTTAATACCTGTATTATTAAATTTGGAAGAAAAATACCTA
AGTTAATTTTTATTNCCTTTAAAAACCTTCATTTTTTATTTAATTTTGGTAAAATCAGACTNNCGAGTTAACGACTGTTAACTTTGCTA
ACTTTTTTACACTTTTGCNATTTTTTATTNTNCCNAGTGGTTNTGNGNGNGNTNATCATCNTCATCNNTTCNANATCGATTATNATCTC
CAA

MULE-27
GGGAAAAATGGTCAAAAAAATCACGAACTTTCAAATTTGGGACGATTTAATCTTGAACTTCACGGAAGACAAATAAATCGTAAAGTTTT
TGTTGACTTTCGAAAAAATGACAAAGTTTTTGTTGACTTGCCATTTGAGTCATGTCGTTAAATAGGTTAACAAAATTATTTTACGGCGT
TAATGTTCTCGTTTATTTGCTCTGTTAGAACAAAACNACGTCGTTTATTGCAGAGACAAAGCGAGATAAAACAACGTCGTTTTGTCTGT
TAAAACAAATTTAAACCTTAAATCCCCAAATCGATTTNATTATCT

MULE-9R
GGGTAATTTGCAGGCCACACCCGTTGACCATGTTTTATTTTCAGGGATTGGCAAAGTCAAATGTGGTTTTCACGTTTTGTCTTTTCACC
TGAAATTTGACCATTTTGCCCTTGCTGCTCCAGCTGAGAAACTGTTGAAATGTACGGTTTAGTTAGTCAGGTTTAGTTAGGTACACTGT
TCCGGTTTAGTTGAAGATATTTTTGTTTCACTTGACGCGCTTCGCTGTACAGTTACAAGGG

MULE-9L
AAGAAATTTTCACAGAAGTGGATGAAAAAAGAATTGCTCTCATGGTTTGTCAAAGTGAGATCTGGTTTTCACCGTTTGTCTTTTCATTC
GTTTTCAGACCAAAATACCCTGAGAAAACTTAATAATGGCACAAACCAAACCGTGCAGCAGTTCGTATCGGTTAAGTTAATGGGATCAA
CCAAACCCCACAGCAGTTGGTTTCGGTTTAGTTATGTGGTTCAGTTTTGAAAAAAAAAAAAAAAAAAAAAAAAAATG

MULE-19R
GAGAAAAATCGTCTGGCCAGTCCCTTTTTTGCAAAGAGGAAGCCAAAAAATCCTTGTAAAAAACAGAAATAATTTATTTATTTTTTCTT
AATTCGAAATATATGGATATTACTAGTATATCCCCATCCGAACCTATTAATTACCAGAAAAATCATGGTTCGTTTCCGAACCAAACCAC
CCGAACCACTTTTACTCTTTATTTGAAAATCAAATAAAAAAAACCCTAATATTTCCAGGGATTTCCCAATTCGAATT

MULE-19L
GAATGATCATCTCTTGTGTCCCTTTTGTGGNACAAATAATCTGCAAAACCACTTTAGAGTCTTCATATCCCCTTTGAGTTAATGAATTG
TGTTAAGACGAATTTACCCCTATATTTTATTTTGAATGAATAATTAATATTNTTNNTAGAAAACATTTTGAAANNAAAAAAAANAAAAA
CTGAAAANAAAAAAAAAAAAAAAGCTCGAGCNCGAGCGAGCTCGNNTTCTTCTTCGGGCTTCTTCCTCCGGCGAGC

MULE-23R
GGGTATATGCACTAGATGACCATGAAATCAATAATAATATAGAAACTAACTATGAGTCTATAGTGTCTGTCGGAGACGTATNTGAAATT
ACCCGCAACCCCCTATAACTGTGTTCTTGGTTAGGGGACATATAGGTGTATAGGCGTATGATTGCAGAAATTTGAATTTTGAAAACCTT
GGTAAGAGCAATGCTGTATGTGGACGCAAACTATAGATTTTAATTTCGAGAAATCGATTTATTATCGAATATTTAGNANTANTAGNTNC
TGCAAAAAAAAAACTAGCGCCG

MULE-23L
GGGTTAATTTACTGAATGACCAAATTTGACATCCTTATTAAGAGAATGACCTTGGNCTCNCAGAAAAGTTGGGTATATAACGTTTTTTT
ACCATGCGAAATATCCGATATAGCCTCTACGGATGGGCGGCAAAACCATATTCCCGACAGNNTCAGACTTGTAGATGTAGGGGAGCTAG
AGATTGTAGGGCAATGATTACAACGTATCAAAAAAAAAAATGCATGCATTACTTATGCAGTTACTACTAAATAAAGCATA

Figure 3.9 Primary consensus sequences of the TIRs in individual TIR-MULE groups and the first 250 nucleotides of the non-TIR MULE termini (5'-end, L; reverse complement of 3'-end, R) in individual non-TIR-MULE groups. Only one member of MULE-16 has been identified in our survey.

Tan1 transposases	535	RCSNCFNIGHRRTQ--CS	551
retrotransposon RT1 protein*b*	203	RCYRCLEHGHNARD--CR	219
gag-pol fusion polyprotein*c*	392	KCFNCGKEGHIARN--CR	408
germline RNA helicase-4*d*	639	PCRNCGQEGHFAKD--CQ	655
DEAD box helicasese	655	PCRNCGQEGHFAKD--CQ	671
gag-env fusion protein*f*	470	PCFKCGQLGHIRAQ--CR	480
zinc finger protein 9*g*	156	NCYRCGESGHLARE--CT	172
zinc finger protein*h*	188	TCHYCGELGHKANS--CK	204
splicing factor*i*	90	KCYECGETGHFARE--CR	106
SLU7 splicing factor*j*	20	FCRNCGEAGHKEKD--CM	36
MULE:GI5441872*k*	18805	RCSRCKGYGHNKAT--CK	18852
MULE-24:GI3319339	96586	TCSNCKQIGHNKGS--CK	96633
MULE-24:GI2760316	90038	TCSNCKEIGHNKGT--CK	89991
MULE-16:GI2443899	22930	HCKSCGEAGHNALR--CK	22977
MULE-9:GI3252804	42930	QCSRCRQAGHNKKT--CK	42883
MULE-9:GI4589411	35847	QCSRCRQAGHNKKT--CK	35894
MULE-9:GI3128140	59456	QCSRCRQAGHNKKT--CK	59409
MULE-9:GI6136349	10518	QCSRCRQAGHNKKM--CK	10565
MULE-9:GI4325365	47660	QCSRCRQAGHNKKT--CK	47613
MULE-23:GI3063438	91617	RCSRCTGAGHNRAT--CK	91664
MULE-23:GI3980374	43428	RCSRCTGAGHNRAT--CK	43475
MULE-23:GI2828187	21240	RCSRCTGAGHNRAT--CK	21287
MULE-23:GI5041964	21745	RCSRCTGAGHNRAT--CK	21792
MULE-23:GI6007863	116329	RCSRCTGSDHNRAT--CK	116376
MULE-23:GI4519197	68216	RCSRCTGA*HNRAT--CK	68169
MULE-2:GI5103850	8614	TCSNCLQEGHNKKS--CK	8567
MULE-1:GI3510344	40354	HCGVCGAADHNSRH--HK	40307
MULE-3:GI2832639	33301	HCGVCGAADHNSRH--HK	33348
MULE-27:GI4388816	30968	TCLNC*GEGHNKAG--CK	31015
MURA:GI21301411	696	TCPNCGELGHRQSSYKCP	712

Figure 3.10 Multiple alignment of CX2CX4HX4C motif of MURA/ MURA-like transposases (derived using BLASTX) and representatives of known proteins. Theamino

acid sequences corresponding to MURA-like transposases were derived from a virtual translation of MULE sequences (position indicated). For the remaining proteins, amino acid positions are given. Asterisks represent positions where a frame shift was introduced to achieve optimum alignment. GI numbers corresponding to the sequences that are not related with a Arabidopsis MULE: a. *Aspergillus niger var.awamori* (GI1805251); b. *African malaria mosquito* (GI477117); c. Human immunodeficiency virus (GI4107489); d-e. *Caenorhabditis elegans* (GI3386540); GI2773235 (direct submission to GenBank); f. Avian endogenous retrovirus (GI6048192); g. *Homo sapiens* (GI105602); h. *Drosophila melanogaster* (GI847869); i. *Arabidopsis thaliana* (GI2582645); j. *Saccharomyces cerevisiae* (GI6320293); k. *Oryza sativa* (GI5441872); l. *Zea mays* (GI2130141).

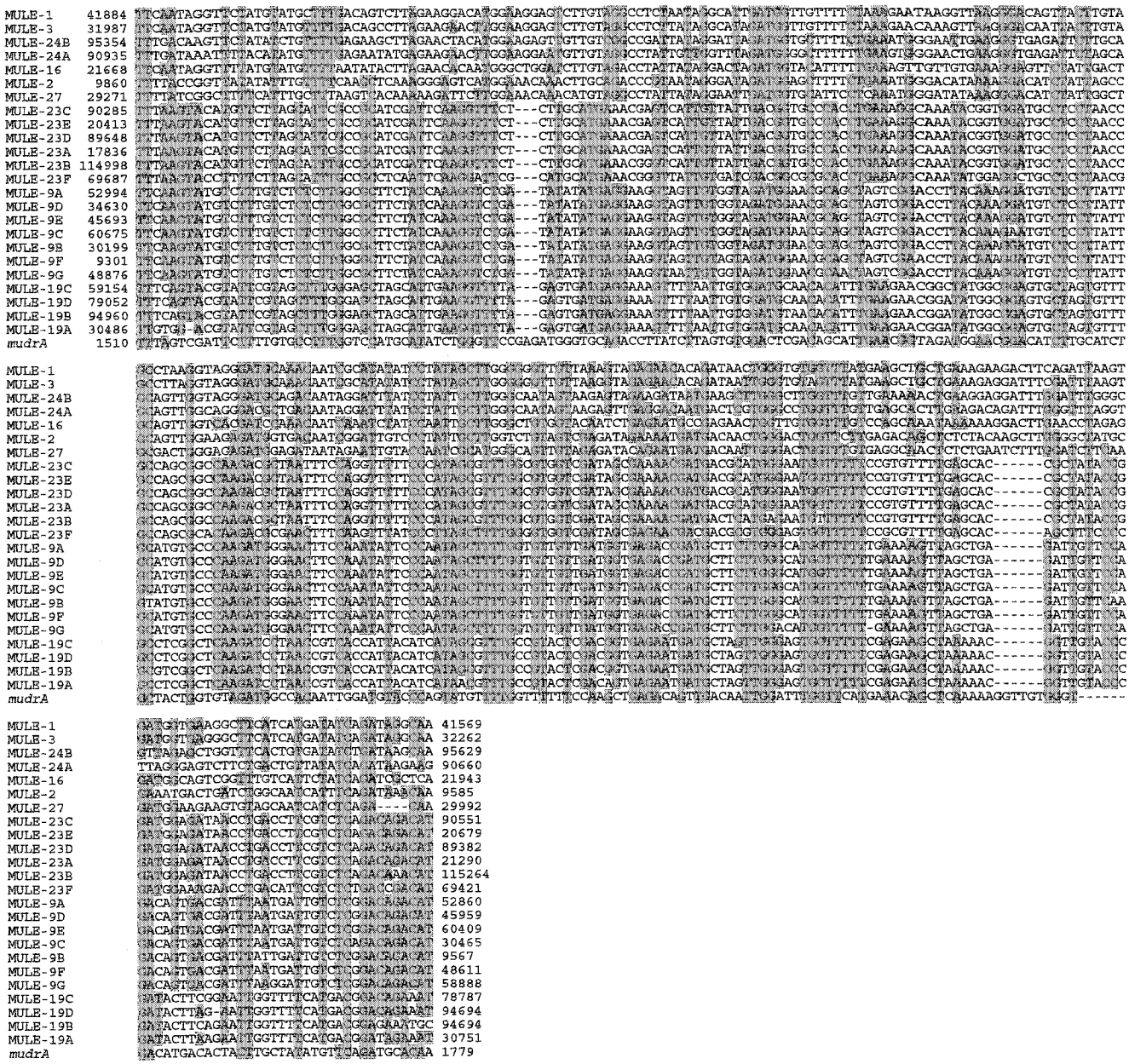


Figure 3.11 Multiple alignment of the most conserved region between the Arabidopsis mudrA-like ORFs and maize mudrA gene. The nucleotides sharing >60% similarity are shaded in gray. The similarity was determined by the conservation mode of the program GeneDoc (Nicholas et al., 1997). The corresponding GI numbers for each MULE are as follows: MULE-1, 3510344; MULE-2, 5103850; MULE-3, 2832639; MULE-16, 2443899; MULE-24A, 2760316; MULE-24B, 3319339; MULE-27, 4388816; MULE-9A, 5672513; MULE-9B, 4185120; MULE-9C, 3128140; MULE-9D, 4589411; MULE-9E, 3252804; MULE-9F, 6136349; MULE-9G, 4325365; MULE-19A, 5041971; MULE-19B, 4585891; MULE-19C, 3242700; MULE-19D, 4914383; MULE-23A, 2828187; MULE-23B, 6007863; MULE-23C, 3063438; MULE-23D, 3980374; MULE-23E, 5041964; MULE-23F, 4519197. The beginning and end nucleotide positions in the corresponding clones are indicated for each sequence used in the alignment.

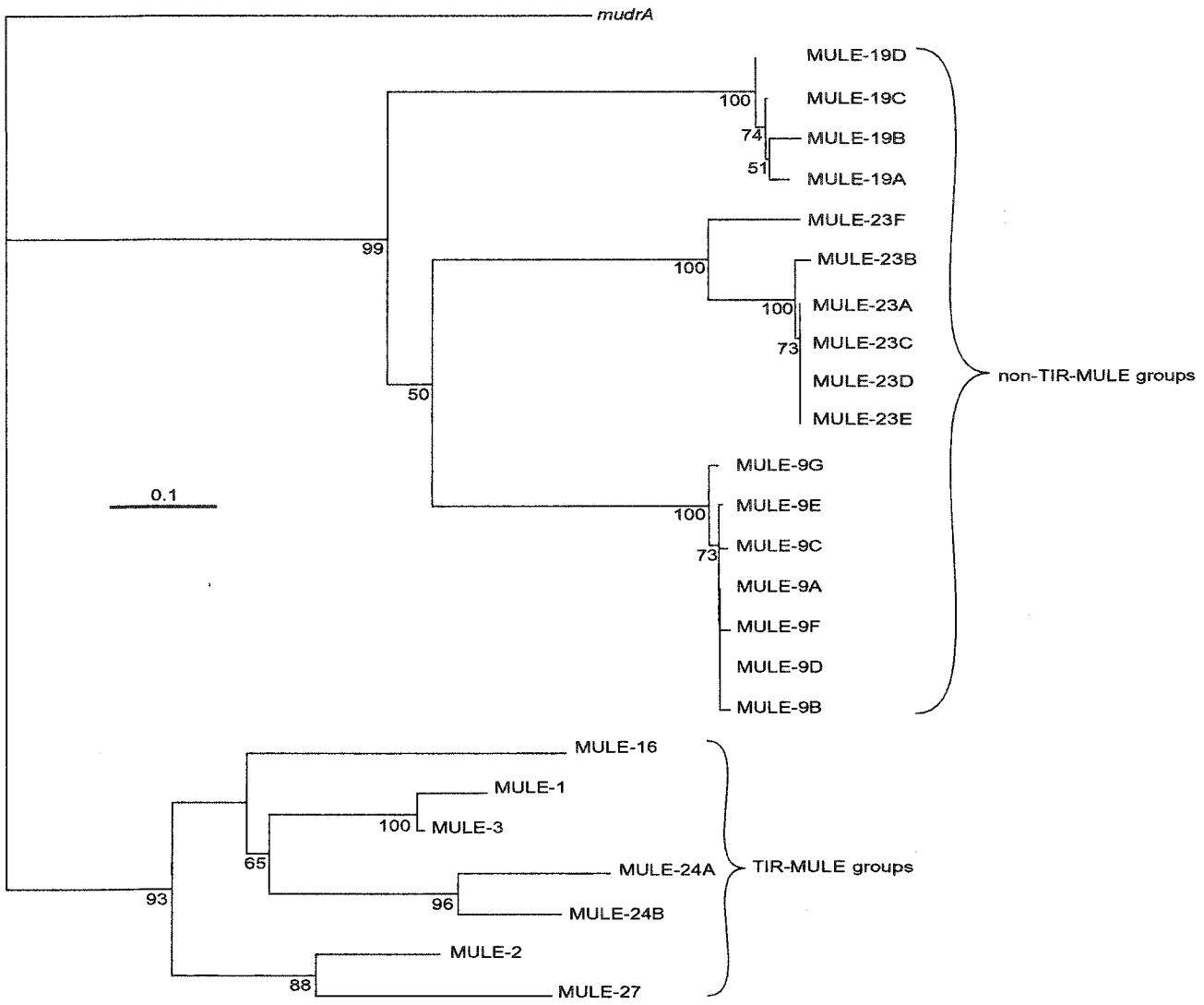


Figure 3.12 A majority-rule and strict consensus tree of *mudrA*-containing MULE elements derived by the neighbor-joining method. The frequencies (>50%) of corresponding branches among 100 derived neighbor-joining trees are indicated. The corresponding GI numbers for each MULE are as indicated in the Figure 11 legend.

Hypothesis II

Our previous data suggested the existence of an Arabidopsis MULE family featuring element abundance in the genome, high sequence and terminal structural diversities between inter-group elements, and potentially mobile competency. Such characteristics raise a fundamental question regarding host-mediated regulation of the MULE evolution. As transposase genes are known playing a central role in the determination of TE mobility and evolution, we thereby approached this question by systematically examining the expression profiles of the *mudrA*-like genes under the regulation of Arabidopsis MET1.

As of 1999, several reports had pointed to a correlation between DNA methylation and the repression of TE activity (Martienssen and Baron, 1994; Chandler and Walbot, 1986). In addition, there existed other means on the silencing of eukaryotic genes, which may well be linked on TE silencing also. For example, the genes in the Drosophila genome are mostly free of DNA methylation and can be repressed by the formation of heterochromatin, as seen from Position-Effect Variegation (PEV) (Weiler and Wakimato, 1995). Several studies on eukaryotic chromatin proteins published that time also suggested a possible link between DNA methylation and the formation of heterochromatin in the regulation of eukaryotic gene activity (Fuks *et al.,* 2000). Taken together, we hypothesized that **the Arabidopsis MULE-transposase genes (or *mudrA*-like genes) can be co-regulated simultaneously by Arabidopsis MET1-mediated genome-wide CpG methylation and the formation of Arabidopsis heterochromatin.**

CHAPTER 4

REGULATION OF *mudrA*-LIKE GENES BY ARABIDOPSIS METHYLTRANSFERASE1 (MET1)

Zhihui Yu and Thomas E. Bureau

4.1 Abstract

Mutator (Mu) and *Mutator*-like elements (MULEs) are DNA transposable elements found in a number of higher plant species including *Arabidopsis thaliana* and *Oryza sativa* (domesticated rice). Like other DNA transposon families, The *Mu*/MULE family is composed of both autonomous and non-autonomous members, with the autonomous ones harboring a transposase gene (*mudrA*) and controlling the mobility of the entire family. By a survey of the Arabidopsis MULEs, we identified a total of 235 *mudrA* homologues from the sequenced genome. They are associated with both Arabidopsis MULE termini, with the majority being within the non-TIR MULEs. We further examined the expression pattern of the gene family utilizing Arabidopsis METHYLTRANSFERASE1 (MET1) mutant (*met1*) plants and found that the Arabidopsis MET1 can differentially regulate individual gene members. Neither the MULE-TIR structure nor the repetitiveness of the elements is found to be necessarily correlated with the MET1-mediated silencing. The efficiency, however, is found to be largely depended on the chromosomal locations of the genes.

4.2 Introduction

Many eukaryotes are capable of modifying their genomes by covalent addition of a methyl group to the 5' position of a cytosine (Gruenbaum *et al.,* 1981). In mammalian cells, this cytosine methylation of DNA occurs within the context of CpG dinucleotides; while in plants, it occurs at both CpG and CpNpG sites (Gruenbaum *et al.,* 1981; Finnegan *et al.,* 1998). DNA methylation is shown to be essential for normal development as reduced levels of methylation and the accumulation of developmental defects have been associated with various biological events, including genomic imprinting and abnormal gene expression (Surani, 1998; Robertson *et al.,* 2000). In principle, DNA methylation can interfere directly with transcription by inhibiting the binding of the basal transcriptional machinery and/or specific transcription factor(s) that require(s) direct contact with unmethylated cytosine (Kass *et al.,* 1997). Alternatively, the presence of methylated DNA may influence chromatin remodeling, indirectly inhibiting gene activity (Costello and Plass, 2001).

DNA methylation has been proposed as one host-mediated mechanism for the repression of transposable elements (TEs), mobile DNA found in almost all eukaryotic organisms examined so far (Bird, 1997, 2002; Fagard and Vaucheret, 200; Martienssen *et. al.,* 2001). Cytosine methylations of a TE promoter can lead to the silencing of the corresponding transposase gene, which, in turn, block the mobility of the entire TE system. Such regulation is associated with the silencing of *Mu* (Martienssen and Baron, 1994; Chandler and Walbot, 1986), *Ac/Ds* (Kunze *et al.,* 1997), and *En/Spm* (Gierl, 1996) elements in maize, and one MULE (Singer *et al.,* 2000) and CACTA element (Miura *et al.,* 2001), and a class of *copia*-like retrotransposons (Hirochika, *et al.,* 2000) in

Arabidopsis. However, exceptions to this pathway also exist. This is most apparent in the genomes that are essentially free of DNA methylation. For example, the nematode *Caenorhabditis elegans* contains approximately the same number of TEs as *Arabidopsis thaliana*, even though the former lacks 5-methylcytosine. TEs in the genome of the invertebrate chordate *Ciona intestinalis* are free of DNA methylation; whereas the host genes in the same genome are predominantly methylated (Simmen *et al.*, 1999). It is thus clear that in addition to DNA methylation, there exist other mechanisms that eukaryotes can utilize to regulate TE activity to ensure the genome stability. However, it is not clear that within a single genome (1) whether or not different mechanisms function coordinately and (2) how individual transposase genes from the same family are open simultaneously to the different pathways.

In this study, we examined the expression profiles of a family of transposase genes, *mudr*A-like genes identified from the sequenced *Arabidopsis* genome. The *Mu* elements were first discovered in maize (Robertson, 1978). Recently, we reported the existence of the *Arabidopsis* MULE family featuring high sequence and structural diversity, yet potentially functional capability (Le *et al.*, 2000; Yu *et al.*, 2000). Subsequently, the *Mu*/MULE homologues were also identified in several other higher plants (Lisch *et al.*, 2001; Turcotte *et al.*, 2001). Like a typical DNA transposon system, the *Arabidopsis* MULEs occur as either autonomous (MuDR-like) harboring a potentially functional transposase (MURA-like) gene (*mudrA*-like) responsible for MULE mobility, or non-autonomous that don't carry a *mudrA*-like gene and whose mobility is conducted *in trans* by the transposases encoded by a related MuDR-like element. In maize, all the identified elements maintain a long TIR structure; and the MURAs were shown to be

able to bind to the TIRs of a MuDR *in cis* or non-autonomous *Mu* elements *in trans*, and are sufficient for *Mu* excision (Benito and Walbot, 1997; Lisch *et al.*, 1999; Raizada and Walbot, 2000). In *Arabidopsis*, in addition to TIR-MULEs, we also identified a number of non-TIR-MULEs, defined as those that don't have *Mu*-specific long-TIRs but carry a *mudrA*-like gene (Yu *et al.*, 2000).

Previously, the expression of seven *Arabidopsis* TIR-MULE-containing *mudrA*-like genes was examined in an *Arabidopsis* hypomethylation mutant line, *ddm1* (*decrease in DNA methylation1*) and the corresponding wild-type strain (Columbia ecotype), and one gene (T3F12.12) was confirmed active in the *ddm1* background but silenced by the gene at DDM1 locus (Singer *et al.*, 2001). As (1) the corresponding gene encodes the proteins actually involving in Arabidopsis chromatin remodeling and (2) the *mudrA* genes examined in their study represent, at most, only 3% of the total *Arabidopsis* homologues, it remains unclear (1) how a genome-wide DNA methylation pattern endorsed by methyltransferases, the enzymes that play the most important role in overall 5-C methylation of eukaryotes, affects the expression of the entire gene family simultaneously and (2) how the genes distributed at different contexts of the Arabidopsis chromatin remodeling respond to the activity of a methyltransferase gene.

4.3 Materials and methods

Plant materials and growth conditions The seeds corresponding to a mutant line of the *met1* and the corresponding wild-type (C24 ecotype) control line were provided by Dr. Finnegan (Commonwealth Scientific and Industrial Research Organization, Australia; Finnegan *et al.*, 1996). The *met1* plants were originally created by transforming wild-type (C24 ecotype) Arabidopsis plants with an antisense cDNA construct of the MET1 gene

(Finnegan *et al.,* 1996). We chose to use the line, *T310.5*, in our study, as it was previously demonstrated to have the lowest level of stable CpG methylation (nearly 15% of the methylation level, compared to the wild-type control plants) for up to three generations (Finnegan *et al.,* 1996). The *T310.5* mutant seeds were vernalized at 4°C for 7 days and then were transferred to a growth chamber set for 22°C and 70% humidity under 16 hours of fluorescent lighting.

Identification of *mudrA*-like genes The *mudrA*-like genes were identified using maize MURA as a query in a TBLASTN search of the sequenced *Arabidopsis* genome (nearly 97% of the estimated 130 Mb genome completed as of May, 2001). All the sequences showing a significant match ($E<10^{-4}$; Li *et al.,* 2001) were selected for further analyses. Multiple alignments of the *mudrA*-like sequences were performed using CLUSTALW (Thompson *et al.,* 1994; http://ca.expasy.org/tools/#align). A BLASTN-based survey was employed to identify *mudrA*-containing MULEs; each of the queries used in this survey was composed of a *mudrA*-like sequence, together with 10-kb DNA flanking segments from each side. The TIRs were identified as the terminal-most regions sharing >80% of nucleotide sequence identity over 100 continuous base pairs (Chapter 3). The identified elements were grouped based on their nucleotide sequence similarity, as described previously (Le *et al.,* 2000).

Analyses of distribution patterns The sequences of Arabidopsis pseudomolecules (contiguous assemblies for each linkage group) and positions of the genes and non-redundant Expressed Sequence Tags (ESTs) on each of the pseudomolecules were obtained from The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/; January 15, 2001 release). The frequency of the expressed

genes was determined using a 1-Mb sliding window. The Arabidopsis TEs were identified previously and can be accessed at www.tebureau.mcgill.ca. The Arabidopsis CENtromeres (CENs) and knobs on the pseudomolecules 4 and 5 were mapped respectively by positioning the clones defining the borders of the corresponding structures. They are F28L22 and T4I21 for CEN-1, T15D9 and T25N22 for CEN-2, T8N9 and T14K23 for CEN-3, T19B17 and F28D6 for CEN-4, and F3F24 and T2L5 for CEN-5 (The Arabidopsis Genome Initiative, 2000, Consortium, C. W. P. A. S., 2000), and T5L23 and T27D20 for the knob-4 (Consortium, C.W.P.A.S., 2000) and F24C7 and F5H8 for the knob-5 (Kazusa DNA Research Institute *et al.*, 2000). A computer program (available upon request) was written to facilitate the analyses of the TE distribution pattern.

Expression analysis Expression profiles of the *mudrA*-like genes were studied by a BLASTN survey of the expressed Arabidopsis sequences (http://www.ncbi.nlm.nih.gov/BLAST/, released before August 2002) as well as with a RT-PCR-based approach. Total RNA was extracted from pooled flowers and siliques of *met1* and the wild-type control plants, using an RNeasy™ Plant Mini Kit (Qiagen, Mississauga, ON) and subsequently treated with DNase (Qiagen, Mississauga, ON) prior to reverse transcription. An Omniscript™ Reverse Transcriptase Kit (Qiagen, Mississauga, ON) was used for the synthesis of the first-strand cDNA from 2 μg of total RNA, as instructed by the manual. cDNA concentration was standardized by quantification of RT-PCR (20 PCR cycles) products of the control gene, the eukaryotic translation initiation factor 4A-1 (Metz *et al.*, 1992). PCR amplifications were performed using a HotStartTaq™ PCR Kit (Qiagen, Mississauga, ON). Of each of the primer pair

used for PCR amplifications of the surveyed *mudrA*-like genes, one was gene specific and the other was a universal primer anchored at the 5'-end of oligo $dT_{(18)}$. The annealing temperatures were ranged from 55°C-65°C and the cycle numbers were set as 32. The amplified fragments were cloned into a pCR2.1™ vector (Invitrogen, Carlsbad, CA) and subsequently sequenced using a SequiTherm™ EXCEL II kit (Epicentre, Madison, WI).

DNA methylation analysis Genomic DNA was extracted from wild-type *Arabidopsis* mature plants using a DNeasy™ Plant Mini kit (Qiagen, Mississauga, ON). A total of 2 μg of genomic DNA was digested either with *Dra*I or *Hph*I, or double digested with *Ssp*I and *Bsty*I prior to a bisulphite treatment. A bisulphite reaction was carried out based on the protocol described by Clark and Frommer (1997). Briefly, the digested DNA fragments were denatured with freshly prepared NaOH (with a final concentration of 0.3 M) at 39°C for 30 minutes. Then, fresh prepared sodium bisulphite and hydroquinone were added to a final concentration of 3.1 and 0.5 mM respectively. All the reactions were carried out in a thermal cycler for 5 cycles of 55°C for 3 hours and 95°C for 5 minutes. Free bisulphate left after the reaction was then removed from the samples with a nucleotide removal kit (Qiagen, Mississauga, ON). The final step of the bisulphite reaction includes alkali (NaOH, with a final concentration of 0.2 M) removal of the bisulphite adduct.

Two genes (T3F12.12 and F9D12.2) were examined in subsequent bisulphite genomic sequencing analyses using a nested PCR approach. Amplifications were conducted with a HotStartTaq™ PCR Kit (Qiagen, Mississauga, ON). First-round PCR was performed in a 25 μl volume reaction, with 5 μl of the bisulphite-treated genomic DNA (the DNA sample digested with *Dra*I was used for the amplification of the control

sequence, a section of the Arabidopsis 180-bp centromeric repeat corresponding to positions 77974 to 78118 in clone GI:18148660; the DNA with *Hph*I digestion was for F9D12.2; and *Ssp*I and *Bsty*I double-digested DNA was for F9D12.2), 200 mM dNTPs, 1 μM primers, 3 mM $MgCl_2$, and 1 unit Taq DNA polymerase (Qiagen, Mississauga, ON). The nested PCR was conducted under the similar conditions except that 5 μl of the diluted PCR products (1:200 dilutions) from the first-round PCR was used as DNA templates. The amplified fragments were sequenced with the same method described for our RNA analysis.

4.4 Results

Using the maize MURA as a query in a TBLASTN search against the sequenced *Arabidopsis* genome, we identified a total of 235 putative *mudrA*-like gene homologues (score of $E<10^{-4}$; Supplementary Table 4.1). Eighty-four percent of these sequences were annotated as a *mudrA*-like, MuDR, MuDR-like, or *Mutator*-like gene (data not shown). The remainder were either not annotated, or annotated as a hypothetical protein. In general, the aligned sequences span only the internal region of the maize MURA (from position 116 to 713), including the putative MURA domain (Eisen *et al.*, 1994).

Subsequent examination of the sequences flanking each of the identified *mudrA*-like gene homologues revealed that 40% (93 out of 235) maintain a known *Arabidopsis* MULE terminus and are associated with a Target Site Duplication (TSD). No obvious TE-associated terminal futures (i.e. TIRs and TSDs) were observed for the remainder, nor are they found to be repetitive at the nucleotide level. It is possible that these genes may be associated with a unique (i.e. non-repetitive) non-TIR-MULE (Yu *et al.*, 2000), a degenerate MULE, or a host gene sharing sequence similarity with the *mudrA*. Based on

the nucleotide sequence diversity, we classified the *mudrA*-contained MULEs into 30 groups, with the majority (82%) being within a non-TIR-MULE. Multiple alignments of the homologues within individual MULE groups revealed that point mutations, deletions, and insertions (including nested TE insertions) were mainly responsible for the sequence diversity of the genes (data not shown). No *mudrB* homologues were identified within any of the *mudrA*-containing MULEs, despite the fact that the non-TIR-MULEs usually maintain one or two other ORFs, as previously described (Yu *et al.,* 2000). A further BLASTN survey of the genome revealed that all the *mudrA*-containing MULEs were associated with non-autonomous members.

Like other eukaryotic genomes, the Arabidopsis centromeric regions represent cytologically deeply staining chromosomal areas that can be differentiated from the surrounding lightly staining regions (Weiler and Wakimoto, 1995; The *Arabidopsis* Initiative, 2000). At the molecular level, these deeply staining areas exhibit low gene density, reduced transcriptional activity, and substantially suppressed recombination rates (Copenhaver *et al.,* 1999). The sequenced BAC clones corresponding to the Arabidopsis CENs and the knobs in chromosomes 4 and 5 was recently defined (The Arabidopsis Genome Initiative, 2000; Consortium, C. W. P. A. S., 2000). In order to compare the *mudrA* distribution pattern with features of the Arabidopsis genome, we mapped the CENs and knob-4 and -5 onto a version of the pseudomolecules corresponding to each linkage group (Table 4.1, Figure 4.1). The Arabidopsis CENs make up ≈7 Mb or 6.0% of the genome, and the two knobs ≈0.6 Mb or 0.5% of the genome. A comparison of the distribution patterns of TEs and non-redundant genes/ESTs reveals a negative correlation (Table 4.1, Figure 4.1). Within the chromosomal regions over nearly 2 Mb from a CEN,

the TE density is less than 30/Mb, whereas the gene density is over 200/Mb (on average, more than 50% of the genes are expressed) (Figure 4.1). The regions proximal to CENs (1-2 Mb) are extremely TE-rich (>100/Mb); however, the gene density is, on average, less than 150/Mb (only 30% of the genes are associated with ESTs; Figure 4.1). Curiously, the number of TEs is generally less within a CEN than within the immediate flanking regions (Figure 4.1; The Arabidopsis Genome Initiative, 2000). We defined the regions containing a CEN and 1-2 Mb of the flanking areas as the *Arabidopsis* centric heterochromatic regions (Table 4.1). This designation is most likely conservative and is consistent with previous reports (Fransz *et al.*, 2000; Kumekawa *et al.*, 2000; Haupt *et al.*, 2001).

Mapping the mined *mudrA*-like genes revealed a distribution pattern similar to that observed for other Arabidopsis TEs (Figure 4.1; Table 4.1). The density of *mudrA*-like genes in centric heterochromatic areas is ≈6/Mb, in contrast to that of ≈0.6/Mb in euchromatic regions. Five of the *mudrA*-like genes distributed within the CENs are associated with intact non-TIR-MULEs flanked by 9-10 bp TSDs (Supplementary Table 4.2). Pair-wise sequence comparisons between these elements and other members of the same MULE groups showed that these CEN-associated MULEs maintain high sequence similarity (up to >90% nucleotide identity) with the members distributed within the euchromatic regions (data not shown). In addition, high sequence similarity (>90% nucleotide identity) was also observed between intra-group MULEs within centric heterochromatic regions (data not shown). MULE-42:GI6598782 (Supplementary Table 4.1) within CEN-2 shares an identical nucleotide sequence (4691-bp) with the

Arabidopsis CEN-specific sequence, Atcss-6, which contains a putative gene potentially encoding a 511 amino acid polypeptide (Copenhaver *et al.*, 1999).

A survey of the expressed Arabidopsis sequences within dbEST databases (http://www.ncbi.nlm.nih.gov/blast/) revealed the expression of a total of 7 *mudrA*-like genes (Table4.2, Supplementary Table 4.2). The origin of the expressed sequences suggests that the corresponding genes were transcribed in roots, aboveground organs, and developing seeds. Of the 7 transcribed genes, only one (F9B22.8; also reported by Singer *et al.*, 2000) was within an element. Further examination of the genome revealed that this element is a member of one TIR-MULE group (MULE-group 27; Yu *et al.*, 2000). No any sort of TE structures was identified within 10-kb flanking sequences of the remaining expressed genes; nor were they found to be repetitive in the genome.

In order to determine expression profiles of the genes under the regulation of *Arabidopsis* MET1, we preformed an RT-PCR assay using the tissues from *met1* and the corresponding MET1 plants (ecotype C24, Finnegan *et al.*, 1996). We chose to study only those *mudrA*-like genes that are within a group containing more than five members. In total, 92 genes were examined and 28 (≈30%) were found to be transcribed (Supplementary Table 4.2). Categorized into 19 MULE groups, five of the expressed genes are within a TIR-MULE, 18 are linked to a non-TIR element and the remaining five are not associated with either of the *Arabidopsis* MULE termini. The sizes of the amplified RT-PCR products ranged from 135 to 560 bp (Figure 4.2; Supplementary Table 4.2). Fifty two percent of the transcribed genes were expressed in both backgrounds; the others were found to be transcribed only in *met1* (Figure 4.2, Supplementary Table 4.2).

Of the seven examined TIR-MULE groups, each contains only one *mudrA*-like gene, which is typical for the Arabidopsis TIR-MULE groups. The genes from five of the groups were transcribed: the one in MULE-16 group was silenced by MET1 (it was also dawn-regulated by DDM1 (Single *et al.*, 2000)); but the others were active in both backgrounds. Of the transcribed genes from 14 non-TIR MULE groups (more than one genes exist and were examined in each group), the level of nucleotide diversity between the intra-group genes allowed us to determine their identities for the most of them except for the groups 19 and 23 where the examined genes within the amplified regions are identical and their expression profiles were not further analyzed. In summary, we observed the following patterns for the remaining 12 non-TIR MULE groups: (1) the genes within the groups 19, 40, 45, and 46 were expressed only in *met1* plants; (2) within the group 42 where six genes were examined, only two genes (F12K21.10 and F3L24.3) were found to be transcribed and in both backgrounds, however, the transcripts from the former gene was dominate (80% *verse* 20% from the later); (3) within the group 41, a total of two genes (T23E23.9 and MFC16.6) were found to be expressed in MET1 plants, but four in *met1* background (the additional two genes were T15G18.X and F27L24.10) (Supplementary Table 4.2) despite the fact that they represented only 20% of the total transcripts in this background; (4) only one gene (T12C22.11) from the group 25 was transcribed in MET1 plants, but in the *met1* background, it accounts for only 13% of the entire transcripts (the remainder were from two additional intra-group members, F3F24.X and F12P23.9, which represent respectively 47% and 40% of the total transcripts).

As summarized in Table 4.2, 87% of the genes expressed in MET1 plants were relatively distant from the Arabidopsis CENs. This distributional feature is in contrast to

that of the genes expressed only in *met1* background, in which 77% were located within the centric heterochromatin. Four of five genes within a TIR-MULE that were transcribed in both MET1 and *met1* plants were all distant from the putative centric heterochromatin (Table 4.2); the solo TIR-MULE-contained gene that was silenced by MET1 was mapped within the putative centric heterochromatin 4. Nine of 12 (75%) genes within a non-TIR-MULE that were transcribed only in *met1* background were also mapped within the putative centric heterochromatic regions, among which three genes, T13E11.12, F3F24.X, and T15E15.X, were mapped within CENs 2 and 5 respectively (Supplementary Table 4.2; Table 4.2).

In order to determine whether this correlation of MET1-mediated regulation of *mudrA*-like gene expression and the distribution pattern of the expressed genes within the genome reflects the feature that the genes within the centric heterochromatic regions are more methylated than those distributed within the euchromatin, we investigated the methylation status of the two TIR-MULEs containing the gene F9D12.2 (expressed in both MET1 and *met1* backgrounds) and T3F12.12 (expressed only in *met1* plants). A region spanning the entire left TIR (upstream of the putative start codon of the genes), the putative 5' untranslated region and a portion of the first exon of the two genes (a total of 794 bp for the former and 578 bp for the later one) were examined. The gene F9D12.2 has 21 CpG sites within the surveyed region, of which 14% were found to be bisulphite sensitive; the gene T3F12.12 has 29 sites correspondingly, of which only 4% was found to be sensitive to the treatment. These CpG sites are evenly distributed within the studied regions and no obvious CpG islands were identified. To monitor the efficiency of the bisulphite treatment, we also examined an Arabidopsis 180-bp centromeric repeat

(corresponding to positions 77974 to 78118 in clone GI: 18148660). Among the 21 methylation sites analyzed, 86% were found to be methylation sensitive.

4.5 Discussion

Our study reveals the existence of more than two hundred *mudrA* homologues in the sequenced Arabidopsis genome. This abundance indicates that the corresponding MULEs replicated successfully in the Arabidopsis genome. One interesting feature of this gene family in *Arabidopsis* is that they are mainly (>80%) carried by non-TIR MULEs. This dominance may reflect the fact that the elements are of recent origin. It may also be possible that distinct features of being a non-TIR-MULE (Yu *et al.,* 2000) assisted the elements spreading in the genome.

The Arabidopsis *mudrA*-like genes are very abundant within the Arabidopsis centric heterochromatic regions. Such a distributional feature was also observed for the other TEs in *Arabidopsis*. Host-mediated selection could result in such a distributional pattern (Kidwell and Lisch, 2001). Alternatively, it could also be the result of preferential TE integrations. We favor the former point of view, as intra-group members were also identified within euchromatic regions. The heterogeneity of the MULEs within individual *Arabidopsis* CENs supports the idea that in a eukaryotic genome, CENs are functionally conserved but divergent in sequence composition (Weiler and Wakimato, 1995; The Arabidopsis Genome Initiative, 2000). The discovery of Arabidopsis CEN-specific sequence, Atcss-6 (Copenhaver *et al.,* 1999), within a non-TIR MULE would suggest a unique correlation between its function, if any, and the element; however, as other members of the same MULE group exist also in other regions of the genome, the sequence itself and its potential function(s) would not thus be CEN-specific. The high

sequence similarity of the intra-group MULEs distributed within the two chromatic regions and their associations with individual TSDs demonstrate that MULE transpositions between the two have occurred relatively recently. Considering the fact that the MULEs can capture host DNA sequences during transposition (Yu et al., 2000; also see chapter 5), such a process may facilitate the redistribution of chromatin-specific information and subsequently induce dynamic changes between euchromatin and heterochromatin. Duplications of euchromatic sequences into the centric heterochromatin were observed recently in the human genome (Guy *et al.*, 2000).

In *Arabidopsis*, at least three proteins, DDM1 (Jeddeloh *et al.*, 1999), MET1 (Finnegan *et al.*, 1996, 1998), and CMT3 (CHROMOMETHYLASE3; Lindroth *et al.*, 2001), have a direct impact on the genome-wide 5-C DNA methylation. DDM1 is a member of a SWI2/SNF2-like ATPase/helicase family, catalyzing ATP-dependent histone-DNA interactions and involving in global DNA methylation in *Arabidopsis* (Jeddeloh *et al.*, 1999). MET1 and CMT3 are two classes of DNA methyltransferases, the former works mainly at CpG sites and the latter at CpNpG sites of the genome (Finnegan *et al.*, 1996; Lindroth *et al.*, 2001; Bartee *et al.*, 2001). MET1 and its homologues in other eukaryotes are also regarded as a class of methylation maintenance proteins. Our discovery that nearly 52% of the examined Arabidopsis *mudrA*-like genes were insensitive to MET1 activity suggests its limitation on the silencing of a multi-transposase-gene family. Despite the fact that a DNA-DNA pairing involving an inverted-repeat can facilitate methylation-mediated gene silencing (Muskens *et al.*, 2000), the MULE-TIR structure is dispensable in MET1-mediated repression. Also, we observed that the silencing effect is homology-independent. Silencing of multiple copies

of duplicated genes simultaneously by DNA methylation is the primary feature of several phenomena, including the inactivation of tandem repeats, *trans*-inactivation of allelic and ecotypic repetitive sequences, and the silencing of transgenes and their corresponding homologous host genes (Muskens *et al.,* 2000; Cogoni, 2001; Meyer and Saedler, 1996). Our finding of differential expression of intra-group *mudrA*-like genes suggests that MET1 may not be the major player triggering homology-dependent gene silencing.

The data gathered from our bisulphate experiment indicate that the genes within a heterochromatic region tend to be capable of maintaining relatively higher level of 5-C methylation at CpG sites than the ones within a euchromatic region. This differential maintenance shown from individual *mudrA*-like genes is consistent to their differential response to the MET1-mediated regulation. As such, the formation of heterochromatin may actually assist MET1-mediated silencing effect. Recent studies on MET1 homologues in mammals (DNMT1) confirmed that its N-terminal non-catalytic domain can bind directly to histone deacetylase complexes (HDAC, the important components in shaping a heterochromatic state) (Fuks *et al.,* 2000), indicating a possibility of the involvement of chromatin remodeling in MET1 activity.

Like other functional transposase genes, the expression of a *mudrA*-like gene can lead to the mobilization of the corresponding element. Such a correlation was observed for the gene, T13F12.12 and the MULE carrying it in a *ddm1* background (Singer *et al.,* 2001). It is also likely that the transcribed genes may be involved in RNA-mediated gene silencing (Matzke *et al.,* 2001). We observed that the Arabidopsis intra-group *mudrA* homologues can be transcribed from both directions (data not shown), which could potentially form double-stranded RNA, subsequently triggering RNA-mediated

repression pathways. The *mudrA*-like genes that were obviously not associated with any type of known MULE structures might be regarded as defective elements. However, the identification of their transcripts may suggest a functional potential, either in the regulation of MULE mobility *in trans* or in a cellular activity unrelated to transposition. Consistently, several annotated *mudrA*-like genes that lack a known MULE terminal structure share high sequence similarity (including the putative functional domain) with an *Arabidopsis* cellular gene that encodes a protein involving in far-red light perception (Yu and Bureau, unpublished data; Lisch *et al.,* 2001; Hudson *et al.,* 1999).

Table 4.1 Distribution of *mudrA*-like genes in the genome of *Arabidopsis thaliana* (Columbia)

chr.	centric heterochromatin				euchromatin		
	position in Mb (CEN position)	size (Mb)	TE no. (no./Mb)	*mudrA*no. (no./Mb)	size (Mb)	TE no. (no./Mb)	*mudrA*no. (no./Mb)
I	13-15 (14.1-15.1)	3	242 (81/Mb)	14 (5/Mb)	27	748 (27/Mb)	39 (1/Mb)
II	1-6 (2.9-4.1)	6	792 (132/Mb)	42 (7/Mb)	14	315 (23/Mb)	9 (0.6/Mb)
III	11-16 (13.2-14.8)	6	743 (123/Mb)	38 (6/Mb)	18	416 (15/Mb)	13 (0.7/Mb)
IV	1-4 (2.1-3.2)	4	523 (131/Mb)	24 (6/Mb)	14	375 (27/Mb)	4 (0.3/Mb)
V	10-14 (10.9-13.2)	5	560 (112/Mb)	30 (6/Mb)	22	600 (27/Mb)	15 (1/Mb)

Table 4.2. Chromosomal positions of expressed *mudrA*-like genes

mudrA-containing MULE designation[a,b] (group no.)	*mudrA* –like gene designation[c]	chr.	expression profile[d]		distance to CEN[e,f] (Mb)
			WT	*met1*	
Centric Heterochromatin					
Non-TIR (46)	T13E11.2	2	-	+	CEN
Non-TIR (25)	F12P23.9	2	-	+	0.6R
Non-TIR (19)	T1O3.28	2	-	+	1.4L
Non-TIR (19)	F1M23.6	3	-	+	0.1L
Non-TIR (40)	MQP15.10	3	-	+	1.1L
Non-TIR (46)	MSJ3.7	3	-	+	1.1L
TIR (16)	T3F12.12[g]	4	-	+	1.2R
Non-TIR (41)	T15G18.X	4	-	+	1.7R
Non-TIR (25)	F3F24.X	5	-	+	CEN
Non-TIR (45)	T8M17.X	5	-	+	0.7L
TIR (1)	MJG14.16	5	+	+	1.4R
? (93)	T15E15.X	5	+	N.T.	CEN
Euchromatin					
Non-TIR (25)	T12C22.11	1	+	+	1.4R
Non-TIR (42)	F12K21.4	1	+	+	1.4L
Non-TIR (41)	T23E23.9	1	+	+	8.5L

TIR (51)	T12C24.24	1	+	+	9.8L
? (92)	F4H5.17	1	+	N.T.	12.0L
? (94)	T12C22.21	1	+	N.T.	1.5R
Non-TIR (46)	F15A23.5	2	-	+	2.6R
Non-TIR (45)	F7H1.17	2	-	+	3.0R
TIR (27)	F9B22.8[h]	2	+	+	3.5R
Non-TIR (41)	F27L4.10	2	-	+	6.2R
? (91)	F17A9.9	3	+	+	11.1L
Non-TIR (42)	F3L24.3	3	+	+	10.5L
? (95)	F7O18.8	3	+	N.T.	12.0L
Non-TIR (41)	MFC16.5	5	+	+	13.1R
TIR (24)	F9D12.2	5	+	+	1.6L
? (96)	MQK4.25	5	+	N.T.	4.4L

[a]TIR: terminal inverted-repeat
[b]?: no known TE terminal structures were identified within the flanking regions; neither were they repetitive in the genome.
[c]X: the corresponding clones were not annotated.
[d]N.T: not tested.
[e]R: right arm of chromosome; L: left arm of chromosome.
[f]CEN: located within centromere.
[g]also transcribed in a *ddm1* line (Singer *et al.*, 2001)
[h]also associated with ESTs (see text).

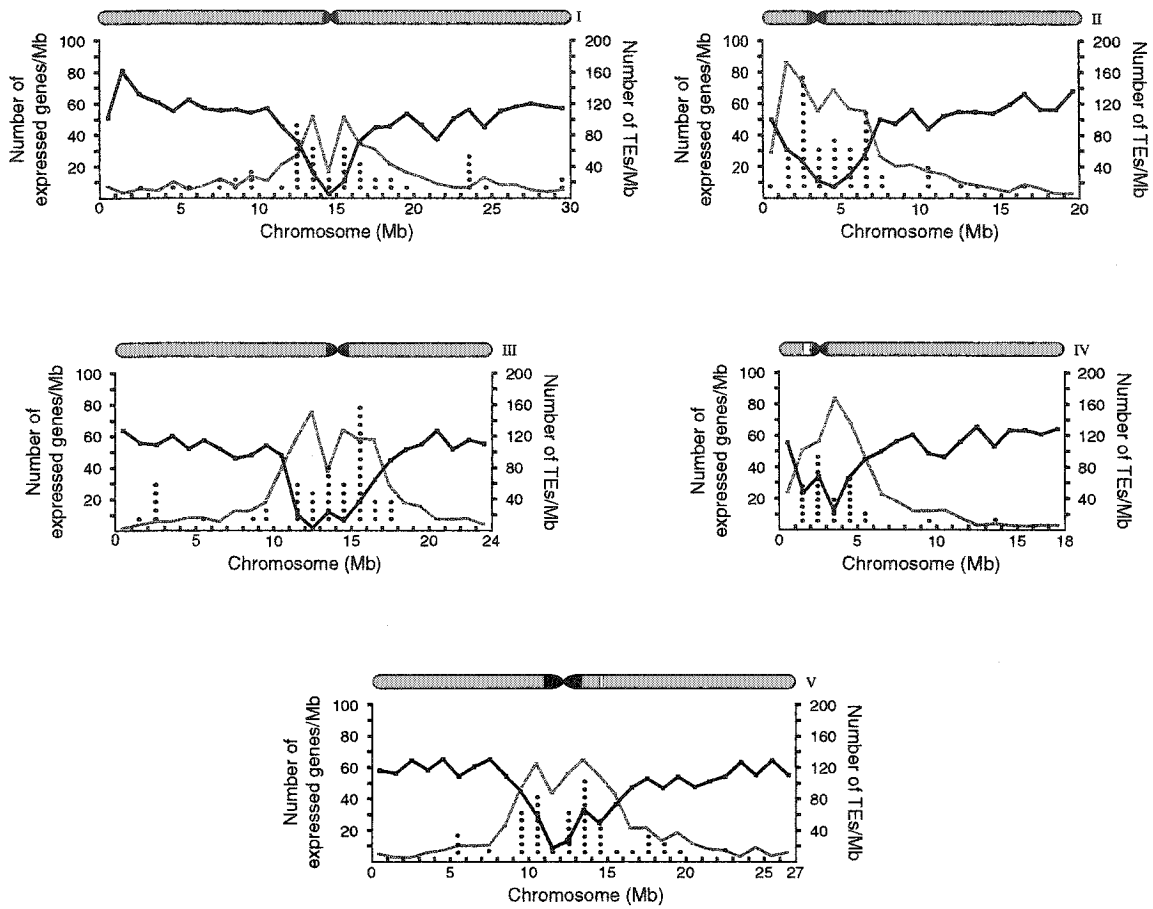


Figure 4.1 *mudrA-like* gene distribution within the Arabidopsis genome. Arabidopsis chromosomes are illustrated above each graph (the black regions corresponding to CENs and white regions for the knobs within chromosome 4 and 5, respectively). Solid black lines indicate the ratio of the expressed genes to the total cellular genes per Mb. TE distributions on individual chromosomes were determined using a 1 Mb sliding window, and are indicated as solid gray lines. Solid black dots indicate the approximate locations of individual *mudrA-like* genes.

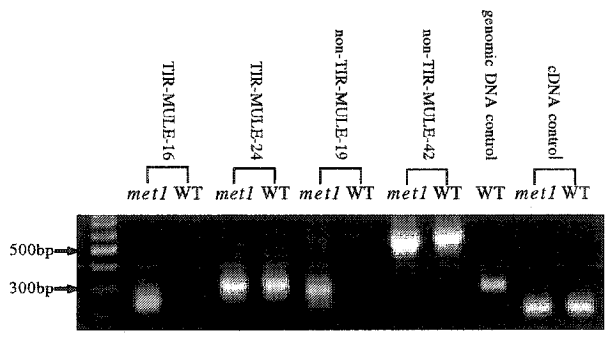


Figure 4.2 RT-PCR results of the *mudrA-like* genes within four MULE groups (the other MULE groups were summarized in Supplementary Figure 4.1) The Arabidopsis eukaryotic translation initiation factor 4A-1 (ETIF4A-1; Metz *et al.*, 1992) was used as an internal control. With genomic DNA as a template,the amplified ETIF4A-1 spans the second intron (82 bp).

Supplementary Table 4.1 *mudrA*-like genes and corresponding MULEs in *A. thaliana*

GI no.	chr.	*MudrA* position[b] start	end	MULE position start	end	target site duplication (TSDs)
6921155	1	48092	51520	46948	51920	AATTAAGTA/AATTAAA
12039051	1	46887	50443	43488	51157	AGGTAATTT/AGGTAATTT
5103850	1	7273	10148	6902	11712	ATAATTTAGT/ATAATTTAAT
12320950	1	88537	91290	71742	93051	ATTAAAATA/ATTAAAATA
12039051	1	59785	61736	59273	64509	ATTTTAAA/ATTTTAAA
2760316	1	89546	92131	88381	92825	GATTCTAAA/GATTCTAAA
9369387	1	28679	32853	27963	36275	GGTTCCAGG/GGTTCCAGG
9438236	1	80642	83582	80238	84733	TTCATTTA/TTCATTTA
9454484	1	88215	91678	86352	98619	TAAAAAAAT/TAAAAAAAT
7159339	1	9844	12812	8981	19452	TAAAAAAAT/TAAAAAAAT
5706738	1	75618	78197	56848	78862	TAGAAAAAAA/TAGAAAAAAA
6957696	1	44612	48698	44376	56770	TATAAGTTAA/TATAAATTAA
6579251	1	53324	56675	51461	63728	TATATTGAA/TATATTGAA
12324242	1	12516	14947	11840	15489	TATTAAAAA/TATTAAAAA
12320878	1	16654	19591	5532	20132	TATTTTCTA/TATTTTCTA
6449476	1	98051	102519	88460	103240	TATTTTCTA/TATTTTCTA
12322371	1	63770	66518	62620	66966	TCTACTAA/TCATACTAA
8778954	1	60376	60864	58778	61749	TCTTTAAAAA/GCTTAAAAA
6715707	1	7922	9910	6890	21239	N.A.
7243815	1	61831	66140	61025	75978	TTGTTTATA/TTGTTTATA
6728952	1	52609	58965	55537	74897	TTTTTTGAA/TTTTTTGAA
8778333	1	41709	44932	N.A.	N.A.	N.A.
6850334	1	7374	10302	N.A.	N.A.	N.A.
5881519	1	45838	50217	N.A.	N.A.	N.A.
6017088	1	4214	5742	N.A.	N.A.	N.A.
12320740	1	14918	19388	N.A.	N.A.	N.A.
12322475	1	70577	71867	N.A.	N.A.	N.A.
12323462	1	93759	95918	N.A.	N.A.	N.A.
8515999	1	36798	39758	N.A.	N.A.	N.A.
5508844	1	1006	2586	N.A.	N.A.	N.A.
5508844	1	11087	12334	N.A.	N.A.	N.A.
5508844	1	7713	9296	N.A.	N.A.	N.A.
5508844	1	3799	5442	N.A.	N.A.	N.A.
6017087	1	26880	28128	N.A.	N.A.	N.A.
9929288	1	33401	37522	N.A.	N.A.	N.A.

5706738	1	23497	27047	N.A.	N.A.	N.A.
12322414	1	22943	25463	N.A.	N.A.	N.A.
12322414	1	28192	309669	N.A.	N.A.	N.A.
6728952	1	95929	97108	N.A.	N.A.	N.A.
11072070	1	4699	7134	N.A.	N.A.	N.A.
3617740	1	60062	61301	N.A.	N.A.	N.A.
3617740	1	72143	73242	N.A.	N.A.	N.A.
3617740	1	103181	103741	N.A.	N.A.	N.A.
12324708	1	53847	56139	N.A.	N.A.	N.A.
9743359	1	31972	34659	N.A.	N.A.	N.A.
7523676	1	8319	9667	N.A.	N.A.	N.A.
6957851	1	46824	49163	N.A.	N.A.	N.A.
2324496	1	58318	59560	N.A.	N.A.	N.A.
12322627	1	37684	39990	N.A.	N.A.	N.A.
5454222	1	39513	43979	N.A.	N.A.	N.A.
5454222	1	74000	78213	N.A.	N.A.	N.A.
6056181	1	61722	63902	N.A.	N.A.	N.A.
6598610	2	37154	39889	36168	53103	AAATTAAAT/AAATTAAAT
6598420	2	41967	47229	38702	46787	AAGTGAAGT/AAGTGAAGT
6598467	2	3406	5543	1682	6206	ATAATTTAAT/ATAATTTAGT
6598597	2	32219	34323	30888	35590	ATATTAATAT/ATATTAATAT
6598480	2	20556	23513	19540	24231	CATTAAAAAACA/CTTTAAAAAA
6598686	2	78119	80177	73741	80859	GATTATTTG/GATTATAAG
6598643	2	3682	6072	2948	11113	GATTTTGGA/GATTTTGGA
6598387	2	57925	60279	52842	61021	GTTCTGT/GTTCTGT
6598529	2	56574	59066	55044	68536	N.A.
6598553	2	41234	43043	3377	45029	N.A.
6598717	2	46337	48532	23208	43452	TAAAAT/TAAAAT
6598473	2	51937	55421	42112	55658	TAAATTATT/TAAATTATT
6598440	2	42850	45808	31230	46540	TAGTATTAA/TAGTATTAA
6598529	2	10055	12559	8782	17240	TATTATTAT/TATTATTAT
6598710	2	48307	50202	36262	51818	TATTTTTTTA/TATTTTTTT
6598747	2	39615	42414	20274	43403	TATTTTTTTT/TTTTTTTTTT
6598674	2	56288	58996	55490	66955	N.A.
6598499	2	65434	67259	64318	80132	TTAAGACAA/TTAAGACAA
6598717	2	25878	27842	24130	39775	TTAATTTTT/TTAATTTTT
6598736	2	29785	32594	14586	33213	TTATATTTT/TTATATTTT
6598426	2	40024	43487	38161	50428	TTCTTTTAA/TTCTTTTAA
6598495	2	35901	38729	24625	39535	TTCTTTTTT/TTCTTGTAATT
6598564	2	10644	12561	9099	13270	TTTTCAAAAAV/TTTTAAAAC
6598567	2	17506	21460	15949	22558	TTTTTA/TTTTTG

6598495	2	17875	20158	N.A.	N.A.	N.A.
6598562	2	43321	46065	N.A.	N.A.	N.A.
6598630	2	103355	106309	N.A.	N.A.	N.A.
6598782	2	13109	13936	N.A.	N.A.	N.A.
6598518	2	27563	28565	N.A.	N.A.	N.A.
6598532	2	21077	23905	N.A.	N.A.	N.A.
6598729	2	18602	20570	N.A.	N.A.	N.A.
6598480	2	79329	80118	N.A.	N.A.	N.A.
6598682	2	2887	4891	N.A.	N.A.	N.A.
6598683	2	19217	21217	N.A.	N.A.	N.A.
6598630	2	16324	19514	N.A.	N.A.	N.A.
6598630	2	40728	42813	N.A.	N.A.	N.A.
6598562	2	25583	28628	N.A.	N.A.	N.A.
6598560	2	55378	57503	N.A.	N.A.	N.A.
7920720	2	21137	24141	N.A.	N.A.	N.A.
6598495	2	53656	56719	N.A.	N.A.	N.A.
6598352	2	6227	8491	N.A.	N.A.	N.A.
6598563	2	81489	84440	N.A.	N.A.	N.A.
6598422	2	68246	70659	N.A.	N.A.	N.A.
6598529	2	36385	38012	N.A.	N.A.	N.A.
6598517	2	50117	51970	N.A.	N.A.	N.A.
6598467	2	12454	13686	N.A.	N.A.	N.A.
6598553	2	11442	13435	N.A.	N.A.	N.A.
6598603	2	987	4080	N.A.	N.A.	N.A.
6598556	2	44112	46542	N.A.	N.A.	N.A.
6598480	2	68831	70535	N.A.	N.A.	N.A.
5541692	3	13828	16477	1990	17385	AAAAAACAA/AAAAAACAA
6899877	3	31013	33060	19383	44033	AAAATTTTA/AAATATTTTA
6899877	3	40839	43564	19383	44033	AAAATTTTA/AAATATTTTA
6782246	3	7273	10148	6902	135392	AAAAAACAA/AAAAAACAA
12408733	3	5748	9002	5005	15475	ATTTTTTTA/ATTTTTTTA
5541692	3	29573	31780	29044	36714	CGGAGAAGA/CGGAGAAGA
5041971	3	29217	31674	28475	36664	GTATGTGAC/GTACGTGAC
5672589	3	23766	26453	12959	27550	N.A.
8347600	3	51538	53485	49819	70996	N.A.
6899954	3	80860	83657	80403	98525	TTATATTAT/TTATATTAT
4185120	3	28536	30873	27794	37106	TTGAAAAAA/TTGAAAAAA
8347620	3	58782	62482	63335	50606	TTTATACTA/TTTATACTA
5672513	3	50935	55351	50211	65469	TTTGTTTAA/TTTGTTTAA
5041964	3	18341	21804	16478	28745	TTTTTTTAA/TTTTTTTAA
8051641	3	25569	27276	N.A.	N.A.	N.A.

5041967	3	46915	49333	N.A.	N.A.	N.A.
3449319	3	55246	56648	N.A.	N.A.	N.A.
5672589	3	53961	55311	N.A.	N.A.	N.A.
6899879	3	70318	73939	N.A.	N.A.	N.A.
12408743	3	28781	30610	N.A.	N.A.	N.A.
12408743	3	1468	1839	N.A.	N.A.	N.A.
12408743	3	25087	26887	N.A.	N.A.	N.A.
6967090	3	67120	69483	N.A.	N.A.	N.A.
6899879	3	31275	32621	N.A.	N.A.	N.A.
8347620	3	87685	89715	N.A.	N.A.	N.A.
8347620	3	99348	100563	N.A.	N.A.	N.A.
8051641	3	88704	88997	N.A.	N.A.	N.A.
12408724	3	26061	28777	N.A.	N.A.	N.A.
6899876	3	7195	19358	N.A.	N.A.	N.A.
5041972	3	14997	17432	N.A.	N.A.	N.A.
4519197	3	68219	71691	N.A.	N.A.	N.A.
6045161	3	49318	50980	N.A.	N.A.	N.A.
7209746	3	3525	5095	N.A.	N.A.	N.A.
6899910	3	47952	49505	N.A.	N.A.	N.A.
6899910	3	105992	106780	N.A.	N.A.	N.A.
8051660	3	32314	33432	N.A.	N.A.	N.A.
8051662	3	13851	15011	N.A.	N.A.	N.A.
6899913	3	63699	69519	N.A.	N.A.	N.A.
6899913	3	13240	14677	N.A.	N.A.	N.A.
6899913	3	22206	23601	N.A.	N.A.	N.A.
6899913	3	51538	53485	N.A.	N.A.	N.A.
6899913	3	63849	64985	N.A.	N.A.	N.A.
1240875	3	48427	50121	N.A.	N.A.	N.A.
8051668	3	29967	32270	N.A.	N.A.	N.A.
7594547	3	76901	78413	N.A.	N.A.	N.A.
7635450	3	94130	96622	N.A.	N.A.	N.A.
6899954	3	99067	100869	N.A.	N.A.	N.A.
6899954	3	30800	33200	N.A.	N.A.	N.A.
12408718	3	29982	31200	N.A.	N.A.	N.A.
12408718	3	73187	73741	N.A.	N.A.	N.A.
6899956	3	55287	57314	N.A.	N.A.	N.A.
9967492	3	15834	16328	N.A.	N.A.	N.A.
4732164	4	62566	63315	51696	65128	AATAAAAAT/AATAAAAAT
2832639	4	31243	33492	30842	34611	ACAATTAATC/ACAATTAATT
4325365	4	46070	50426	37334	51279	ATATGAATAA/ATATGAATAA
4850281	4	99528	101266	99243	101373	ATCGTCAA/ATCGTCAA

4263373	4	101782	103137	100000	108448	CAGACATTT/CAGACATTT
5748495	4	38380	39735	33294	41417	
2443899	4	20819	23250	20148	23792	TAAAA/TAAATAAA
3319365	4	467	2435	4774	23265	TAATTTTAA/TAATTTTAA
2443899	4	30294	33101	29317	45473	TCAAATAAA/TAAAATAAA
3293581	4	13795	18201	12992	25711	TTAATTAAG/TTAATTAAG
3695386	4	24937	26880	23139	37862	TTCTTATAT/TTCTTATAT
6136349	4	8728	10284	6938	24019	TTTATATAA/TTTATATAA
3309276	4	97328	99066	N.A.	N.A.	N.A.
7267276	4	64476	66032	N.A.	N.A.	N.A.
3319339	4	93735	97099	N.A.	N.A.	N.A.
4732164	4	83181	84374	N.A.	N.A.	N.A.
3243214	4	33453	34505	N.A.	N.A.	N.A.
3243214	4	37409	39296	N.A.	N.A.	N.A.
4732168	4	32056	35870	N.A.	N.A.	N.A.
4732168	4	73083	73805	N.A.	N.A.	N.A.
4732168	4	112134	113982	N.A.	N.A.	N.A.
3319365	4	6398	8016	N.A.	N.A.	N.A.
7321075	4	50454	51694	N.A.	N.A.	N.A.
3309276	4	68966	70662	N.A.	N.A.	N.A.
5731752	4	6273	7171	N.A.	N.A.	N.A.
3695386	4	1432	2407	N.A.	N.A.	N.A.
4732169	4	25888	26475	N.A.	N.A.	N.A.
4732169	4	74914	76133	N.A.	N.A.	N.A.
2443899	4	71627	76280	N.A.	N.A.	N.A.
4325365	4	106593	108792	N.A.	N.A.	N.A.
3309259	4	36111	38329	N.A.	N.A.	N.A.
3309259	4	7887	9579	N.A.	N.A.	N.A.
3319359	4	66446	73816	N.A.	N.A.	N.A.
8777493	5	7481	11369	6768	14712	AAATTCTT/AATTCTT
9502397	5	81936	83174	80260	96117	ACTTTTCAC/ACTTTTCAC
2828187	5	17836	21299	15973	28240	ATAAATAAA/ATAAATAAA
4454587	5	43354	45310	42594	46220	ATAATATAA/ATAATATAA
7658323	5	40491	41970	38408	53401	ATAGAAATA/ATAGAAATA
3319339	5	93735	97099	93314	98243	ATATAAAAT/ATATAAAAT
3510341	5	10395	14266	6993	14980	ATCTTGACT/ATCTTGACT
9755607	5	27422	29058	26966	30371	ATTTTCTTT/ATTTTCTTT
3128140	5	57879	62228	47220	63069	CTTTTATAA/CTTTTATCAA
9502158	5	25813	27265	23994	28426	GATTTAGATT/GATTTAGATT
3510344	5	40188	42587	39176	43127	GTTTTTTTTC/GTTTTTTTC
9885848	5	51952	52759	37207	57545	N.A.

4454587	5	113707	112769	109397	114412	N.A.
2656025	5	43880	46363	43085	46864	TAAAAAATA/TAAAAAATA
6579250	5	63232	64833	60623	78537	TAATATTA/TAATATTA
5822965	5	83735	82443	79428	87303	TACATTTAA/TACATTTAA
4220630	5	7482	10352	6738	10748	TAGCATAAT/TAGCATAAT
4680765	5	50695	52841	49883	53854	TAGTATCAAC/TAGTATTAAC
3047060	5	90185	93017	89216	103777	TATATAATA/TATATAATA
4519196	5	27691	30523	16931	31492	TATTATATA/TATTATATA
4589411	5	32968	35940	32249	47458	TTAAGTATA/TTAAGTATA
2264308	5	9021	11438	8313	16268	TTATTTA/ATATTTA
9502158	5	54943	55746	47078	58302	TTATTTTTTT/TTGTTTTTTT
5732428	5	5983	8707	N.A.	N.A.	N.A.
3702730	5	2581	5334	N.A.	N.A.	N.A.
3047060	5	105283	108084	N.A.	N.A.	N.A.
6587796	5	75866	76361	N.A.	N.A.	N.A.
6587796	5	212	840	N.A.	N.A.	N.A.
4753195	5	20416	21358	N.A.	N.A.	N.A.
9755632	5	37354	39641	N.A.	N.A.	N.A.
8051637	5	64340	67167	N.A.	N.A.	N.A.
5732428	5	49155	51238	N.A.	N.A.	N.A.
5732428	5	5983	8707	N.A.	N.A.	N.A.
3046849	5	18742	209724	N.A.	N.A.	N.A.
3510337	5	30274	32633	N.A.	N.A.	N.A.
4159702	5	11382	112640	N.A.	N.A.	N.A.
2264305	5	72784	74771	N.A.	N.A.	N.A.
2264314	5	57169	58962	N.A.	N.A.	N.A.
3985953	5	827	1858	N.A.	N.A.	N.A.
4519196	5	12788	14089	N.A.	N.A.	N.A.
2660661	5	1	3013	N.A.	N.A.	N.A.
9885845	5	69520	70599	N.A.	N.A.	N.A.
9502397	5	46439	48187	N.A.	N.A.	N.A.
7682776	5	29437	30084	N.A.	N.A.	N.A.
9971623	5	9237	10475	N.A.	N.A.	N.A.
9971623	5	44223	45971	N.A.	N.A.	N.A.
6715701	5	35878	37548	N.A.	N.A.	N.A.
9885848	5	25214	26035	N.A.	N.A.	N.A.
2191181	5	52010	55512	N.A.	N.A.	N.A.

[a]X represents the homologues that are not annotated; N. A.: not available.
[b]corresponding to the positions revealed from TBLASTN survey.

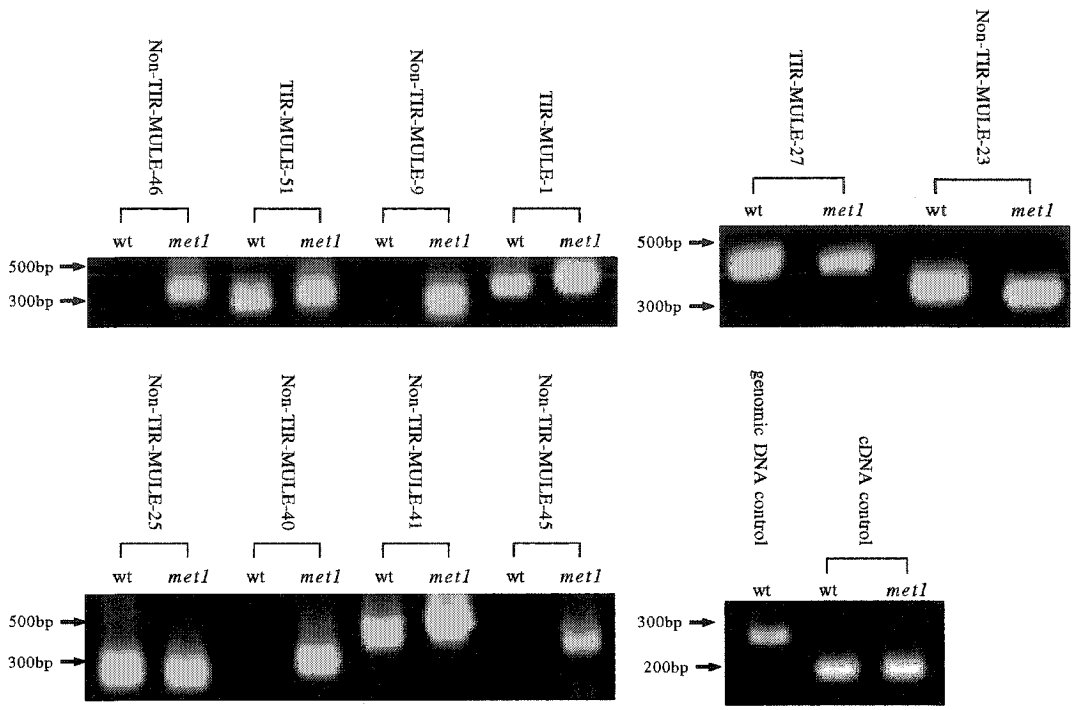
Supplementary Table 4.2 Expression profiles of the *mudrA*-like genes in *A. thaliana*

group no.	termini[a]	no. of MULE	surveyed *mudrA*[b]	expression profile[c]				
				expressed line		no.of clones analyzed		EST[d]
				WT	*met1*	WT	*met1*	
1	TIR	28	MJG14.16	+(2)	+(2)	2	2	-
2	TIR	19	M20D23.2	-	-	-	-	-
3	TIR	6	F28J12.7	-	-	-	-	-
16	TIR	4	T11I11.3 T3F12.12 F21A20.X	- - -	- + (4) -	0	4	-
24	TIR	17	F9D12.2 F1N21.6	+ (4) -	+ (4) -	4	4	-
27	TIR	13	F9B22.8	+ (2)	+ (2)	2	2	AV526976
51	TIR	7	T12C24.24	+ (5)	+ (5)	5	5	-
9	Non-TIR	23	F28L5.16 T22C5.27 F26C24.9 C17L7.X T3H13. 11 F5H8.11 MJE4.X K2M2.17 F5F19.7	-	+	-	7[e]	-
19	Non-TIR	9	T103.28 F1M23.6 MOD1.6 T10J7.7 F26B6.15 MIK22. X	- - - - - -	+ (8) + (4) - - - -	0	12	-
23	Non-TIR	15	F22O13.X K21C13.5 MFE16.4 T22I11.14 F16P2.39 F15O4.15	+	+	3[e]	3[e]	

25	Non-TIR	21	F26C17.4	-	-	10	15	-
			T24G3.X	-	-			
			MAB16.1	-	-			
			F11O6.3	-	-			
			F7M19.90	-	-			
			F3F24.X	-	+ (7)			
			T24M8.2	-	-			
			F12P23.9	-	+ (6)			
			T12C22.11	+ (10)	+ (2)			
			T32A11.X1	-	-			
			F7F22.11	-	-			
			F26C17.4	-	-			
			T32G9.38	-	-			
				-				
40	Non-TIR	16	F7N22.10	-	-	0	13	-
			F10C8.4	-	-			
			F27C12.19	-	-			
			T3F12.3	-	-			
			MQP15.10	-	+ (13)			
			MIF6.10	-	-			
			F3K12.13	-	-			
			T10A2.X	-	-			
			MSK10.13	-	-			
			MSJ3. X	-	-			
41	Non-TIR	14	F1O13.14	-	-	11	10	-
			T29A4.X1	-	-			
			T15G18.X	-	+ (1)			
			T23E23.9	+ (3)	+ (3)			
			F25O24.X	-	-			
			F27L4.10	-	+ (1)			
			T25O11.3	-	-			
			T21L8.60	-	-			
			MFC16.5	+ (8)	+ (5)			
42	Non-TIR	13(7)	F12K21.10	+ (8)	+ (10)	10	12	-
			F3L24.3	+ (2)	+ (2)			
			F18P9.X1	-	-			
			F18P14.20	-	-			
			T26N6.X1	-	-			
			F9A16.13	-	-			
45	Non-TIR	17	F27F5.15	-	-	0	12	-
			T10J7.13	-	-			
			T20G20.16	-	-			
			F7H1.17	-	+ (7)			
			T14A4.16	-	-			
			T8M17.X	-	+ (9)			

46	Non-TIR	28	T21L8.30	-	-	0	22	-
			T22C12.8	-	-			
			F28N16.3	-	-			
			F13A23.5	-	+(8)			
			T2L5.X1	-	-			
			MSJ3.7	-	+(9)			
			F9L11.2	-	-			
			T13E11.2	-	+(5)			
			F5O4.4	-	-			
			T7B9.15	-	-			
			F13M2.X	-	-			
91	?	N.A.	F17A9.9	+	N.T.	N.A.	N.A.	AI998697
92	?.	N.A.	F4H5.17	+	N.T.	N.A.	N.A.	AV544301
93	?	N.A.	T15E15.X	+	N.T.	N.A.	N.A.	AV557094
94	?	N.A.	T12C22.21	+	N.T.	N.A.	N.A.	AV552211
95	?	N.A.	F7O18.8	+	N.T.	N.A.	N.A.	AY059842
96	?	N.A.	MQK4.25	+	N.T.	N.A.	N.A.	AY096512

[a] ?: no known TE terminal structures were identified within the flanking regions; neither [b]
[b]X: the corresponding clones were not annotated.
[c] +: expressed. -: not expressed. The numbers represent transcript abundance within the sequenced samples.
[d] in the case of matching more than one ESTs, only one is listed.
[f] the *mudrA*-like gene sequences were identical within the surveyed regions; therefore, only one representative is listed for each group.
N.A. not available.
N.T. not tested.



Supplementary Figure 4.1 RT-PCR assay of the *mudrA* genes from ten MULE groups (the four other groups were presented in Figure 4.2) The Arabidopsis eukaryotic translation initiation factor 4A-1 (ETIF4A-1; Metz *et al.*, 1992) was used as an internal control. The amplified ETIF4A-1 segment includes the second intron (82 bp) when using a genomic DNA template.

Hypothesis III

Our genome-wide identification of the Arabidopsis MULEs allows us to further study possible TE-host relationships in the course of eukaryotic gene/genome evolution. As of 2000, studies on this subject were largely linked to TE-mediated regulations of gene expression and genome instability (Bennetzen, 1999; Kdiwell and Lisch, 1997). In 1999, Moran and his colleagues (Moran *et al.*, 1999) demonstrated that human L1-mediated transduction could shuffle down-stream host gene segments and subsequently create a new mosaic gene. Consistent with their study on retrotransposons, our previous survey on the MULE diversity within the sequenced Arabidopsis genome (Chapter 3) also suggested a link between MULE mobility and the creation of new eukaryotic genes. Taken together, we hypothesized that **from an evolutionary standpoint, MULE transposition can be viewed as one possible pathway for the creation of novel eukaryotic genes featuring mosaic organization and repetition within the genome. This mobility-mediated mechanism may facilitate the evolution of eukaryotic multigene families.**

CHAPTER 5

ACQUISITION, DUPLICATION, AND DIVERSIFICATION OF EUKARYOTIC GENES BY DNA TRANSPOSONS

Zhihui Yu, Michael Huynh, Nadia Mohabir and Thomas E. Bureau

5.1 Abstract

Mobility or transposition is the most prominent feature of transposable elements (TEs). There is ample evidence directly linking such a feature to the creation of eukaryotic gene mutation and genome instability. As such, TEs were often viewed as selfish DNA. However, recent demonstration of retrotransposition-associated transduction of flanking host gene segments in humans suggests a functional role of TE mobility in gene evolution. Subsequently, it was hypothesized that eukaryotes could utilize the mobility as a vehicle to generate new genes and multigene families. Here we show that by transposition, *Mutator*-like elements (or MULEs, a family of DNA transposons) in the model plants *Arabidopsis thaliana* and *Oryza sativa* (domesticated rice) can capture various host DNA segments, create new genes and expression profiles, and form multigene families. In particular, the MULE mobility in *Arabidopsis* was indispensable in the evolution of two multigene families respectively encoding putative Ubiquitin-like (Ubl)-specific cysteine proteases (AtMULE-Ulps) and related serine proteases. As *ULP*s are known important for specific SUMO-targeting, and the simultaneous reactivation of individual SUMOlation pathways is a fundamental step involved in plant response to environmental stress, we proposed that the diversified *AtMULE-ULPs* may play an important role in the reaction of host defense responses.

5.2 Introduction

Gene duplication and exon shuffling are fundamentally important for the development of eukaryote complexity and adaptive evolution. Gene duplication can reduce phenotypic effects of null alleles and developmental accidents, accomplish a functional compensation of allele genes and establish the groundwork for the evolution of gene complexity (Ohno, 1970; Lynch and Conery, 2001). Remarkably, eukaryotes maintain a large number of duplicated genes (Long, 2001). Studies from five sequenced eukaryotic model genomes revealed that nearly 30 to 60% of the genes belong to identifiable families of duplicates (Long, 2001; Kidwell, 2002). There is evidence suggesting that genome-wide polyploidization and chromosomal duplications may have played a major role in eukaryotic gene duplication (Ohno, 1970; Long, 2001). Exon shuffling is the process of combining exons from different genes (Kolmn and Stememr, 2001). It is an important molecular pathway to create novel genes and gene networks (Eickbush, 1999). Although the majority of the shuffling events were identified in invertebrates, such a process actually has occurred widely in eukaryotes (Long, 2001). In a number of cases, exon shuffling was achieved through host-mediated non-homologous recombination. Recently Moran *et al.* (1999) demonstrated that a human L1 retrotransposon can transduce down-stream host gene segments and create novel mosaic genes under experimental conditions. Their study indicated a mobility-mediated strategy to create new genes.

Potentially, TE mobility facilitates not just the formation of new genes, but also the creation of dispersed multigene families in eukaryotes. However, a genome-wide survey of L1-mediated transductions in humans failed to provide such evidence (Pickeral

et al., 2000; Goodier *et al.*, 2000). Furthermore, it is not clear whether this mobility-based strategy also applies to Class II TEs or DNA transposons. With the vast sequence information provided from the sequencing of the Arabidopsis and rice genomes, we examined the MULE diversity and their potential creation of new genes and gene duplicates in the two genomes. We show that the capture of host DNA segments occurs frequently in the course of MULE evolution. We identified a number of MULE-contained genes that are not associated with TE mobility. Particularly, we provide evidence showing that MULE transposition have contributed to the diversification of Arabidopsis *ULPs*.

5.3 Materials and methods

Identification of the MULEs Two different methodologies were employed for the identification of the MULEs in *A. thaliana* and *O. sativa*. Previously, systematic identification of the Arabidopsis MULEs from 17.5 Mb of the sequenced Arabidopsis genome was conducted and 209 MULEs were identified (Le *et al.*, 2000). Subsequently, a *mudrA*-based strategy (chapter 4) was employed for the identification of the elements carry a *mudrA*-like gene from nearly 130 Mb of the sequenced Arabidopsis genome. In this study, we further mined the new MULEs sharing terminal sequence similarity with the identified elements with a computer script. The rice MULEs were identified from 39 Mb of the sequenced rice genome using the *mudrA*-based strategy developed and employed for the Arabidopsis MULEs. Supplementary Figure 5.1 illustrates the major steps involved in identification of the MULEs from these two genomes.

Identification of MULE acquisition events Both BLASTN and BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/) analyses were employed for the identification of

MULE acquisitions (Supplementary Figure. 5.2). Structural organization of the acquired segments was inferred mainly from the corresponding annotations and confirmed by the expression data. In addition, the ORFs from the rice MULEs were identified using the computer programs for the Rice Genome Automated Annotation System (RGAAS) (http://ricegaas.dna.affrc.go.jp; the corresponding results can be accessed at http://www.tebureau.mcgill.ca). Both PSI- and RPS-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) analyses were conducted for the examination of sequence relationship between the ORFs within the elements and the corresponding host genes.

Sequence analysis at a break point Analyses of junction borders were conducted for those elements whose acquired sequences share over 80% nucleotide sequence identity with their genomic counterparts. Junctions were delineated by aligning the element, an intra-group element nearly identical to the former but lacking the acquired segment, and the host DNA segment (Supplementary Figure. 5.3). The DiAlign program (http://www.genomatix.de/cgi-bin/dialign/dialign.pl) was used for the multiple sequence alignment analyses.

Expression assays The expressions of the ORFs within a MULE was examined through a survey of the expressed Arabidopsis and rice sequences (http://www.ncbi.nlm.nih.gov/dbEST/index.html; as of October, 2002) and by RT-PCR approach, as described previously (Chapter 4). For RT-PCR, total RNAs were extracted from both wild-type (C24 ecotype) and METHYLTRANSFERASE (MET1) mutant (*met1*) *Arabidopsis* plants (T10.5, C24 ecotype; Finnegan *et al.*, 1996) using an RNeasy™ Plant Mini kit (Qiagen, Mississauga, ON). The first-strand cDNA was

synthesized using an Omniscript™ Reverse Transcriptase Kit (Qiagen, Mississauga, ON). Total RNA concentration was standardized by the quantification of RT-PCR products (20 PCR cycles) of the control gene, eukaryotic translation initiation factor 4A-1(Metz *et al.*, 1992).

Mobility assay A display technique (Wright *et al.*, 2001) developed previously was employed for MULE mobility assay. Briefly, genomic DNA from both wild-type (ecotype *Ler*) and CHROMOMETHYLASE3 (CMT3) mutant (*cmt*3) Arabidopsis plants (Lindroth *et al.*, 2001) were extracted using an DNeasy™ Plant Mini kit (Qiagen, Mississauga, ON), subsequently digested with methylation-insensitive restriction enzyme, *Bfa*I and ligated with a universal adapter. A nested-PCR approach was then employed for specific amplification of MULE-mobility associated polymorphisms. The amplification products were cloned into a pCR 2.1™ vector (Invitrogen, Carlsbad, CA) and sequenced using a SequiTherm™ EXCEL II kit (Epicentre, Madison, WI).

5.4 Results

A total of 1392 MULEs, ranging from 0.2 to 20.8 kb, were identified and examined (Table 5.1; Supplementary Tables 5.1 A and B). Fifty one percent of the elements contain at least one host DNA segment with an average size of 1.38 kb (Table 5.1; Supplementary Tables 5.2-5.4). All the acquired sequences were identified to be of nuclear origin (Supplementary Tables 5.2-.4). It is evident that the MULEs can acquire virtually any type of genomic DNA, including both single and repetitive sequences, and as part of coding and non-coding regions (Table 5.1, Supplementary Tables 5.2-.4). Specifically, 73% of the acquired segments share sequence similarity with a host gene of known or predicted function (Table 5.1; Supplementary Tables 5.2-.4). A total of 389

putative ORFs (Table 5.1; Supplementary Tables 5.4A and B) were identified within the MULEs, of which 31% share significant sequence similarity (E-value <10-4, Li *et al.*, 2001) with, for example, a structural RNA gene or a previously functionally-defined gene (Table 5.1; Supplementary Tables 5.4A and B). The remaining ORFs encode proteins for an unknown function. Twenty four of the MULE-contained ORFs match an expressed sequence (Supplementary Tables 5.4A and B).

To elucidate the origin of the MULE-contained genes, we analyzed the sequence and structural organizations of the ORFs that exhibit high sequence similarity (>80%) with a known gene. One such ORF within AtMULE382 encodes an Arabidopsis pectin methylesterase (PME; At4g03930) (*AtMULE-PME*; Figure 5.1A). We determined the *AtMULE-PME* structure, which is identical with that of other higher plant *PMEs*, by comparing its mRNA with the corresponding genomic gene sequence (Figure 5.1B). The conceptually translated AtMULE-PME also maintains the conserved PME domain (Figure 5.1C; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

Twenty-four MULE-contained genes show a mosaic sequence organization, consisting of various DNA segments including different genes (coding and non-coding regions) and intergenic sequences (Table 5.1; Supplementary Tables 5.2 and .3). Figure 2 illustrates one of the two (At3g06940 and At3g05850) expressed mosaic MURAs. Like MURA homologues expressed in higher plants, these two also maintain the *Mutator* domain; but their N-terminal regions contain an additional DNA segment encoding a new peptide sharing sequence homology with PB1 (Phenol and Bem1) domain (Figure 5.2), a recently identified protein scaffold involved in numerous multiprotein-complex formation (Eickbush, 1999).

Redundancy appears to be a common feature of the MULE-contained genes (Supplementary Tables 5.4A and B). The largest gene family we identified encodes Arabidopsis Ubiquitin-like (Ubl)-specific proteases (or Ulps). In animals, *ULP*s are present in only 1 or 2 copies per genome (Melchior, 2000; Suzuki *et al.*, 2000). The Arabidopsis genome, however, contains 24 MULE-contained *ULPs* (*AtMULE-ULPs*) as well as several host counterparts (*At-ULPs*) that are not associated with a MULE (Supplementary Table 5.4A; data not shown). The *AtMULE-ULPs* were associated with three individual MULE groups (Supplementary Tables 5.1 and .4); most (>90%) of the members maintain intact MULE termini and TSDs (Supplementary Table 5.1A; Bennetzen, 1996; chapter 3). Multiple sequence alignments between the *AtMULE-ULPs* revealed over 90% nucleotide sequence similarity between intra-group and lower than 50% similarity between inter-group members (data not shown). Exploring recent observations of the activation of Arabidopsis transposons under various stress conditions, we observed the *de novo* transposition of the *ULP*-containing MULEs in a *cm3* mutant (deficient in CpNpG methylation, Lindroth, *et al.*, 2001) background (Figure 5.3). This is also the first indication of the mobility of a non-TIR MULE.

We also examined the expression profiles of the Arabidopsis *ULPs* by surveying expressed sequence databases at NCBI and by conducting RT-PCR of wild-type (MET1) and *met1* mutant (deficient in CpG methylation, Finnegan *et al.*, 1996) Arabidopsis plants. Our database survey revealed the expression of one *At-ULP* (At4g15880). We further examined the expression of three other *At-ULPs* and all the *AtMULE-ULP*s. Our previous RT-PCR analysis has indicated that Arabidopsis *mudrA*-like genes can be regulated by MET1 activity (Chapter 4). We observed the similar expression pattern of

three *AtMULE-ULPs* from two MULE groups, that is, they were silent in the wild-type but active in *met1* plants (Figure 5.4A). Of the three examined *At-ULPs*, only one (At5g60190) was found to be transcriptionally active. Like the At4g15880 gene, it was transcribed in both MET1 and *met1* backgrounds (Figure 5.4a). The transcripts from the remaining *At-ULPs* and *AtMULE-ULPs* were not observed. This suggests that they may be pseudogenes, be transcribe below detectable threshold of our experiment, or be silenced by mechanism(s) other than CpG methylation. All the expressed ULPs share high amino-acid sequence within the Ulp core domain and maintain the catalytic triad (His-Xn-Asp-Xn-Cys) (Figure 5.4c).

In addition, we identified six MULE-contained ORFs sharing high sequence similarity with the *AtMULE-ULPs* within the encoded putative Ulp core domain except that the Cys residue within the triad was replaced by serine (Ser) (AtMULE-dUlps; Figure 5.4c). All the six ORFs were within the same MULE group (group-23, Supplementary Tables 5.1A), and they share nearly identical nucleotide sequences (99.9% similarity over >10 kb; chapter 3). Similar to the expressed *AtMULE-ULPs*, one of the *AtMULE-dULPs* was transcribed and was also only in *met1* background (Figure 4a). The transcripts from the remaining genes were not observed.

Previous sequence analyses of several *de novo* internal deletions of maize *Mutator* indicated the involvement of an interrupted gap-repair pathway in the origin of *Mutator* diversity (Hasia *et al.*, 1996). This error-prone repair process could facilitate MULE acquisitions by creating single-strand DNA that can be further paired with non-homologous sequences through sequence homology at junctions. To test this possibility, we aligned the sequence at the junctions among the MULEs and the corresponding host

DNA segments (see Methods and Supplementary Figure 5.3). We found that 63% of the aligned junctions maintain sequence homology over as long as 69-bp (Supplementary Table 5.5), which is consistent with the level of the homology generally required for synthesis-dependent strand annealing (SDSA) during DSB repair (Figure 5.5). The introduction of gaps in the multiple alignments reflects (1) incomplete MULE sequence information (Supplementary Figure 5.3), (2) outcomes of exonuclease activity during repair, (3) the accumulation of point mutations over evolutionary time, or (4) the existence of other acquisition pathways.

5.5 Discussion

It is evident that the MULEs in the rice and Arabidopsis genomes can capture host DNA segments (also see Yu *et al.*, 2000; Turcotte *et al.*, 2001). In addition, several other DNA transposons (including one *Mu2* and *dSpm*, and several *Ds*) were also observed to be able to incorporate host DNA sequences (Talbert and Chandler, 1988; Rubin and Levy, 1997; Takahashio *et al.*, 1999). As such, it is plausible to conclude that the amplification of host DNA segments by transposition of Class II TEs is a common phenomenon. This is consistent with the proposed mechanism (Figure 5.5). In eukaryotes, DSBs can occur spontaneously, be induced by DNA-damaging agents, or be created by transposition (Pâques and Haber, 1999; Hasia and Schnable, 1996). The damaging DNA is subsequently restored by a group of host-encoded DNA-repair enzymes (Pâques and Haber, 1999; Haber, 2000). There is plenty of evidence showing that repairing of such a break created by the means unrelated to transposition allows for the 'filling' of non-homologous nucleotide sequences, though most of which are very short (Gorbunova and Levy, 1997). It is well known that the majority of Class II TEs

transpose through a cut-and-paste mechanism, which produces DSBs at donor sites after element excisions (review see Plasterk, 1996). Accordingly, the mobility of Class II TEs can potentially promote the amplification of host DNA segments.

What are the benefit(s) to eukaryotes from such TE mobility-mediated amplifications? It is obvious that new genes can be created directly (Table 5.1; Supplementary Table 5.4). Class I TEs may participate in the creation of intron-free eukaryotic genes *via* retrotransposition-mediated transduction; comparatively, Class II TEs can form a new gene with or without an intron. Furthermore, new genes can also originate following subsequent MULE diversity, as seen from the origin of the *AtMULE-dULPs*. The resultant new triad actually appears to be identical with the functional motif of trypsin-like family of serine proteases (Allaire *et al.*, 1994; Bazan and Fletterick, 1988). Similarly, viral 2a and 3c subclasses of Cys proteases also resemble trypsin-like Ser proteases in bacteria and humans, prompting a potentially evolutionary relationship between these two types of proteases (Bazan and Fletterick, 1988). It is tempting to hypothesize that these expressed novel *AtMULE-dULPs* might encode a group of ULP-derived Ser proteases in *Arabidopsis*.

One important feature of the genes within some of the elements is that they show susceptibility to epigenetic regulation, as observed from the expression of *AtMULE-ULPs* and *AtMULE-dULPs* (Figure 5.4). This may own to the vulnerability of corresponding elements to the host-mediated silencing mechanisms, such as, DNA methylation and heterchromatin-mediated repression. Another important feature is their mosaic sequence and structural constitutions, as exemplified by the expressed *PB1-mudrA*-like genes in *Arabidopsis* (Figure 5.2). The PB1 domain has been found within various proteins in both

plants and animals recently, all of which were functionally unrelated to TE mobility (Ito *et al.*, 2001; Terasawa *et al.*, 2001). In yeast, mutations within the PB1 domain of the Bem1 gene can cause bilateral mating defect and cell temperature-sensitive growth (Terasawa *et al.*, 2001). As all expressed putative *Mu*/MULE transposases (MURAs) in *Arabidopsis* as well as in other higher plants contain the *Mutator* domain exclusively, it is evident that the novel mosaic *PB1-mudrA* genes were created by MUE-shuffling of host DNA segments encoding a PB1 domain precursor. The new PB1-*Mutator* domain architecture within the expressed PB1-MURA proteins may facilitate the assembly of novel multiprotein complexes that consist of both transposase and host proteins. Such protein-protein interactions could be an indication of mutual regulations and functional interactions between mobile DNA and host genes. Built on this point of view, the transposition of DNA transposons not only can creates new genes, but also may facilitate the development of eukaryotic gene complexity in terms of its sequences, structures, expressions, functions and regulations.

Ulps belong to the C48 family of cysteine proteases (Li and Hochstrasser, 1999 and 2000). They are essential for SUMOlation (Small Ubiquitin MOdifier) activity (Melchior, 2000; Kim *et al.*, 2000; Suzuki *et al.*, 2000). The core domain surrounding the catalytic triad (His-Xn-Asp-Xn-Cys) represents the hallmark of Ulp family (Li and Hochstrasser, 1999, 2000; Kim *et al.*, 2000; Suzuki, *et al.*, 2000). We have shown that in Arabidopsis (1) the diversity of the *AtMULE-ULPs* mirrors that of the MULEs carrying them, (2) the *ULP*-containing MULEs maintain the intact MULE terminal structure and perfect TSDs and (3) some of these elements were mobile under a hypomethylation stress (Figure 5.3). It is thus evident that the amplification and diversification of the

Arabidopsis ULP family is, in large part, the result of MULE activity over evolutionary time.

Recent studies from both *Arabidopsis* and tomato have demonstrated that reactivation of different SUMOlation pathways is a fundamental step involved in plant response to environmental changes, such as temperature variation, and exposing to different DNA damage reagents, toxic chemicals or virus infections (Hanania *et al.*, 1999; Kurepa *et al.*, 2003). Different SUMO pathways require the involvement of individual Ulps whose specificity is largely determined by their N-terminal sequences (Cabrita *et al.*, 1997). The diversified *AtMULE-ULPs* could play a role in such responses given that they are active only under stressed conditions and the encoded inter-group proteins show diversified N-termini (data not shown). It is also not impossible that they participate in the regulation of MULE activity in *Arabidopsis*.

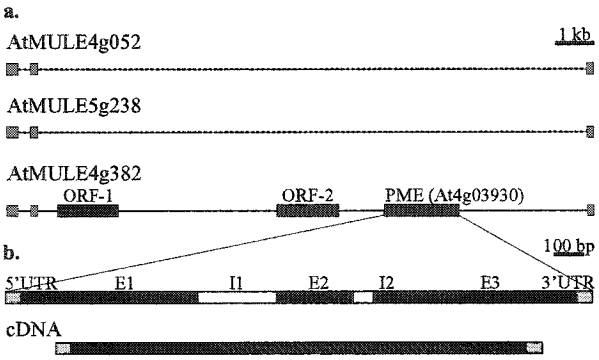
In conclusion, this study demonstrates a new functional role of DNA transposons in the evolution of eukaryotic genes and multigene families. This TE-mediated mechanism does not merely generate gene redundancy; it may facilitate the development of eukaryote gene complexity as well. As demonstrated in plants (Kunze *et al.*, 1997), gene redundancy in combination with differential gene inactivation and release can be implicated in adaptive evolution, specifically in the origin of cultivated plants and domesticated animals, and likely in explosive species radiations as well.

Table 5.1 Summary of MULE acquisitions in *A.thaliana* and *O.sativa*

Category	*A. thaliana*	*O. sativa*
MULEs (MULE groups)	557 (44)	835 (49)
Acquisitions (BLASTN/BLASTX)	306 (219/87)	409 (158/251)
MULE-contained ORFs (known gene-related)	123 (61)	266 (61)
classification of acquired segments (BLASTN/BLASTX):		
I. intergenic regions	141/na	50/na
II. genes or gene segments	165/59	359/61
1. RNA genes	1/0	14/0
2. transcription & regulation	9/1	11/7
3. signal transduction	13/6	15/10
4. cell cycle regulation	2/0	1/0
5. cell structure	1/2	2/4
6. chromatin modulation	0/1	2/2
7. transport	3/1	14/8
8.protein synthesis, modification & degradation	2/32	7/5
9. metabolism	3/1	113/18
10. photosynthesis	4/0	3/1
12. DNA replication & modification	21/11	0/0
13. pathogen & environmental response	1/0	2/2

14. other	0/3	4/4
15. unclassified (hypothetical & unknown)	106/63	271/205

na: not applicable.



c.

```
Le-PME     DGSTTFRSATLAAVGQGFILQDICIQNTAGPAKDQAVALRV
St-PME     DGTTTFRSATLAAVGQGFILQDICIQNTAGPEKHQAVALRV
Ps-PME     DGTTTFNSATVAAVGDGFIAQDIGFQNTAGPEKHQAVALRV
Vv-PME     DGWTTYASATFAVKGKGFIARDMTFENTAGPEKHQAVALRV
At-PME     DGSTTFHSATVAAVGERFLARDITFQNTAGPSKHQAVALRV
MULE-PME NVRRTFYTATFASNGKGFIGVDMCFRNTAGPSKHQAVALRV
```

Figure 5.1 Acquisition of a *PME* gene by AtMULE-382. **a**, illustration of structural and sequence relationships between the *PME*-containing MULE and two other elements of the same group that lack the gene. The grey-shaded regions represent the sequences exhibiting over 80%nucleotide sequence similarity. Dashed lines represent the gaps created from the original alignments (generated by DIALIGN 2; http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/dialign). **b**, structural organization of the expressed *AtMULE-PME*. **c**, Partial alignment of the conserved PME domain.

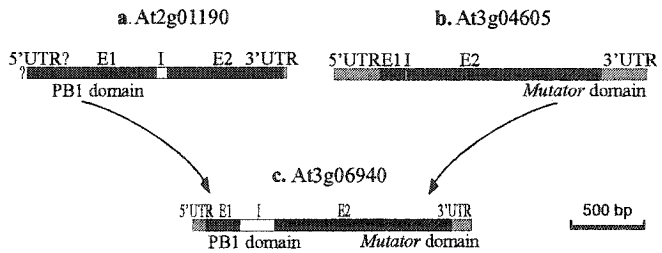


Figure 5.2 Formation of a *PB1-mudrA* mosaic gene in Arabidopsis. Gene structure was determined comparing genomic sequence with the corresponding cDNA. The *Mutator* and PB1 domains within the conceptual translated protein were identified from survey of Conserved Domain Database at NCBI.

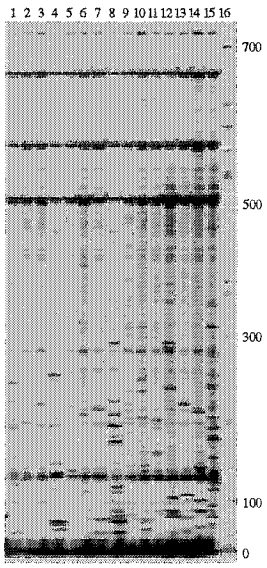


Figure 5.3 Autograph of a Display gel, conducted with a transposon display protocol, of a group of *ULP*-containing MULEs. Genomic DNA used in this assay was extracted from Arabidopsis *cmt3* hypomethylation mutant plants (lane 1-14) as well as the wild-type control plants (lane 15). The *de novo* MULE mobility was displayed as the novel bands shown in the *cmt3* background. Lane 16 represents a 1-kb DNA ladder.

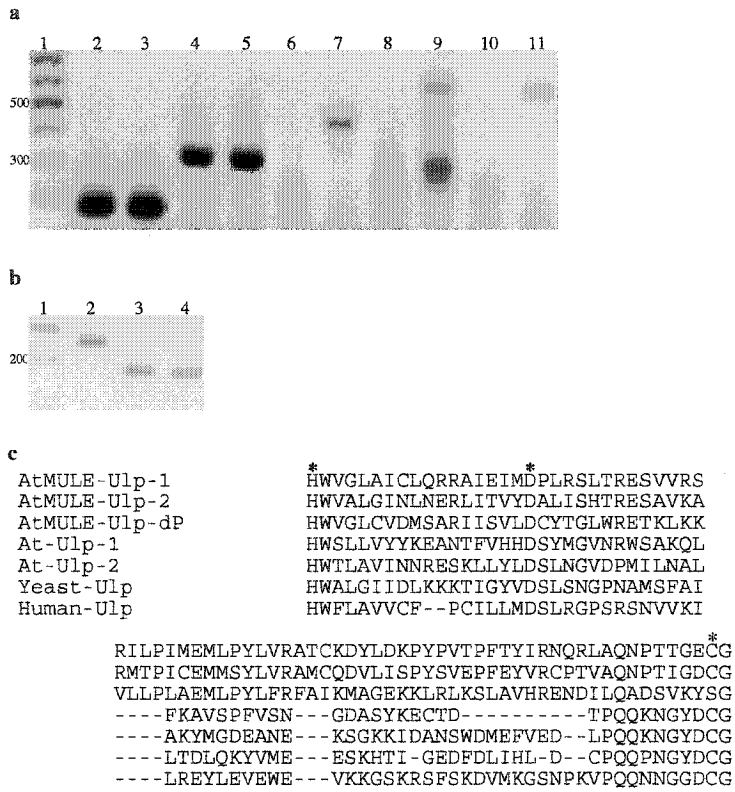


Figure 5.4 Expression of Arabidopsis *ULPs*. **a**, RT-PCR (32 cycles) assay results. Lane 1: 1-kb DNA ladder; lane 2-3: positive control using Arabidopsis eukaryotic translation initiation factor 4A-1 (*ETIF4A*) gene; lane 4-5: *At-ULPs;* lane 6-7: *AtMULE-ULPs* within group-9; lane 8-9: *AtMULE-ULP*-derived ORFs (MULE-23 group); lane 10-11: *AtMULE-ULPs* within group-45. **b**, standardization of RNA concentrations by PCR of the *ETIF4A* gene (19 cycles). Lane 1: a 1-kb DNA marker; lane 2: genomic DNA from the wild–type plants as the template; lane 3-4: cDNAs derived from the wild-type (lane 3) as well as the *met1* mutant (lane 4) plants as the template. **c**, multi-alignment of the Ulp core domain encoded by the expressed *ULPs* in Arabidopsis, yeast and humans respectively. AtMULE-Ulp-dP: the protein encoded by *ULP*-derived ORF. The catalytic triad is labeled with *. GI numbers of the aligned Ulps: 18396116 (*AtMULE-ULP*s within group-9), 18398039 (*AtMULE-ULPs* within group-45), 18424328 (*AtULP-1*), 18414542 (*At-ULP-2*), 18404980 (*AtMULE-ULP*-derived ORFs), 17380332 (Yeast *ULP1*) and 17380330 (Human *ULP*).

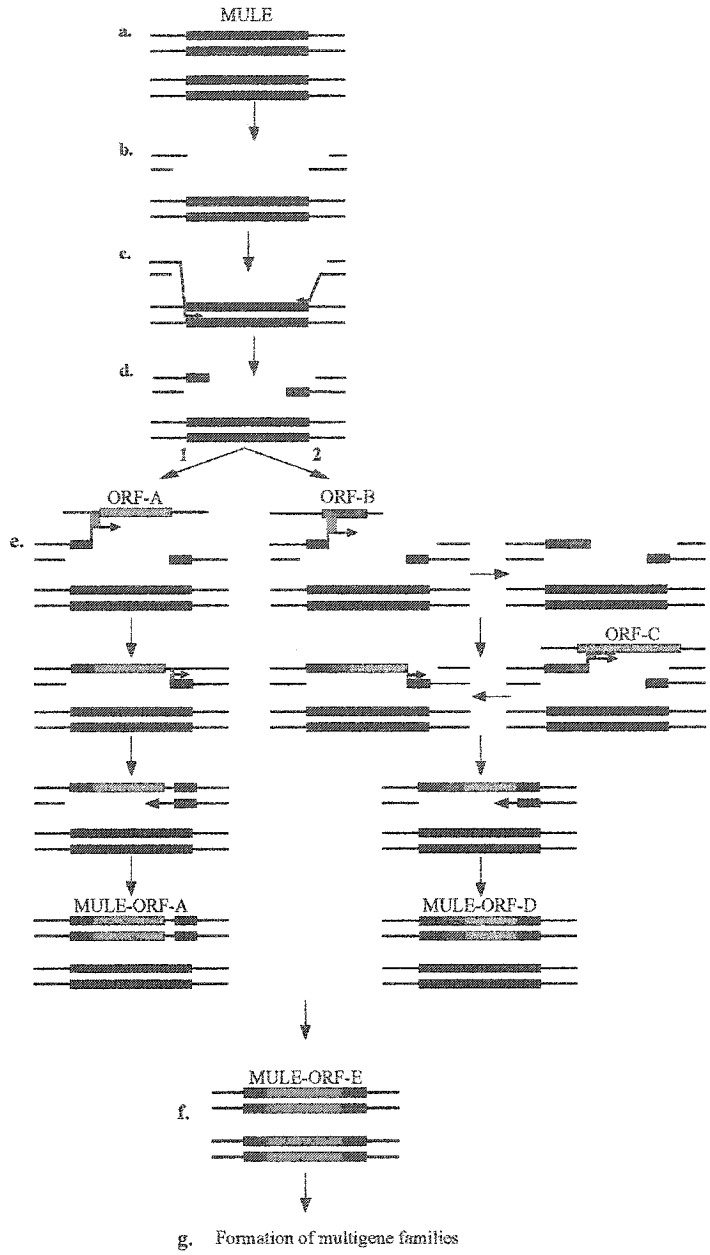


Figure 5.5 A proposed model of generation of eukaryotic genes and multigene families *via* MULE transposition activity. **a**, genomic DNA with a MULE insertion. **b**, excision of the element results in a DSB with 3'-protruding end at the donor site. **c**, the gap is repaired by copying homologous element. **d**, the synthesis process is prematurely terminated. **e**, newly elongated single-strand DNA invades a host DNA segment located at ecotypic sites by pairing homologous sequences to restore the DNA synthesis. **e-1**, single gene is captured. **e-2**, more than one ecotypic sites are visited. **f**, formation of a new gene upon further sequence diversification within the element. **g**, creation of multiple gene families *via* MULE transposition.

Supplementary Table 5.1A MULEs in the sequenced *Arabidopsis* genome

name	group	chr.	position*		TSD			terminal structure
			start	end	size	left	right	
AtMULE001	XXVIII	1	257710	258629	920	TTTACAAAT	TTTACAAAT	TIR
AtMULE002	II	1	512692	513137	446	AAAAAAATA	AAAAAAATA	TIR
AtMULE003	VIII	1	1613289	1614679	1391	TTTATTTAA	TTTATTTAA	non-TIR
AtMULE004	XXIII	1	2788449	2800716	12268	TATTCTTTC	TATTCTTTC	non-TIR
AtMULE005	XXVIII	1	4110126	4111043	918	TTGTTTATA	TTGTTTATA	TIR
AtMULE006	II	1	4209186	4209627	442	TAGATGTTC	TCGATGTTC	TIR
AtMULE007	I	1	4329918	4334413	4496	TTCATATTTA	TTCATATTTA	TIR
AtMULE009	II	1	5911742	5916554	4813	TAAATTAT	TAAATTAT	TIR
AtMULE010	IX	1	6156420	6158881	2462	TATTTTTAAA	TATTTTAAA	non-TIR
AtMULE011	I	1	6210907	6212004	1098	GAACATTAC	GAACATTAC	TIR
AtMULE012	XXII	1	6906162	6906482	321	TTATAATAA	TTATAATAA	TIR
AtMULE013	I	1	7023560	7024543	984	TTTCCAAATTGGTCC	TTTCCCAAATTGGTCC	TIR
AtMULE014	XXXXV	1	7254973	7257241	2269	TAAAACATAA	TAAAACATAA	non-TIR
AtMULE015	XXIII	1	7352944	7365211	12268	TTCAATATA	TTCAATATA	non-TIR
AtMULE016	XXIII	1	7392288	7393409	1122	ATTAATATT	ATTAATATT	non-TIR
AtMULE017	XXVIII	1	7761221	7762125	905	TACTTTTAAT	TACTTTTAAT	TIR
AtMULE018	VI	1	7960260	7961288	1029	TTTTTCATT	TATTTCCTT	non-TIR
AtMULE019	XXXXI	1	8453499	8461810	8312	GGTTCCAGG	GGTTCCAGG	non-TIR
AtMULE020	IX	1	9690787	9705567	14781	TATTTTCTA	TATTTTCTA	non-TIR
AtMULE021	X	1	9929684	9930790	1107	CAAAAAAAA	CAAAAAAAA	TIR
AtMULE022	XXXXV	1	10167080	10171095	4016	TTTAAAAAAAA	TTTAAAAAAAA	non-TIR
AtMULE023	XXXIV	1	10361418	10363660	2243	GCAAATAAA	GCAAATAAA	non-TIR
AtMULE024	VIII	1	10511865	10512156	292	TTTTTTATCA	TTTTTTATGA	non-TIR
AtMULE025	XII	1	10541979	10543089	1111	TTGTTTTTT	TTGTTTTTT	non-TIR
AtMULE026	XXVIII	1	10816732	10817647	916	TATAAATTA	TATAAATTA	TIR
AtMULE027	IX	1	11063005	11065769	2765	TAATTTTTAA	TAATTTTTAA	non-TIR
AtMULE028	XXXIV	1	11848955	11851197	2243	AATATTTTA	AATATTTTA	non-TIR
AtMULE029	XXXXXXXXXII	1	12153169	12157515	4347	TCTACTAA	TCATACTAA	TIR
AtMULE030	XXXXXXXXXIII	1	12300881	12301905	1025	AAAAGAGATAA	AAAAGATAA	non-TIR
AtMULE031	XII	1	12351036	12352151	1116	TATATTTTT	TCTATTTTT	non-TIR
AtMULE032	XXXXII	1	12688767	12699237	10471	ATTTTTTTA	ATTTTTTTA	non-TIR
AtMULE033	XXIII	1	12705687	12706809	1123	TATAATTAA	TATAATTAA	non-TIR
AtMULE034	XXXIII	1	12832579	12853888	21310	TATTTTAAT	TATTTTAAT	non-TIR
AtMULE035	X	1	12908994	12909498	505	AAAAAACTAAA	AAAAAATAAA	TIR
AtMULE036	X	1	12909636	12910800	1165	ATTTTATATT	ATTTTATATT	TIR

AtMULE037	XXV	1	12920484	12922162	1679	TTCATTAAA	TTCATTAAA	non-TIR
AtMULE038	XII	1	13071488	13072454	967	TATATTTTA	TATATTTTA	non-TIR
AtMULE039	I	1	13128534	13129140	607	GAAGAGAAATG	GAAGAGAATG	TIR
AtMULE040	XXIII	1	13187229	13206665	19437	GTATTATTA	GTATTATTA	non-TIR
AtMULE041	XXXXII	1	13558756	13560047	1292	AAAATTCAT	AAAATTCAT	non-TIR
AtMULE042	XII	1	13581225	13584454	3230	TTAATTTTT	TTAATTTTT	non-TIR
AtMULE043	XII	1	13586590	13587477	888	TTCATTAAA	TTCATTAAA	non-TIR
AtMULE044	XXXIV	1	13865048	13866661	1614	TATATACTT	TATATACTT	non-TIR
AtMULE045	XII	1	13919383	13921168	1786	TAAACTATT	TAAACTATT	non-TIR
AtMULE047	XII	1	15212440	15213563	1124	CTTGATAAGA	CTTGATAAGA	non-TIR
AtMULE048	XXI	1	15829175	15830209	1035	TACCAAAACC	TACTAAAACC	non-TIR
AtMULE049	XV	1	15837907	15838676	770	ATTATAAAAAGG	ATTATAAAAAGG	TIR
AtMULE050	XXIV	1	15919205	15920254	1050	ATAACATATA	ATACCATATA	TIR
AtMULE051	XII	1	16173755	16175772	2018	TTAATAAAA	TTAATAAAA	non-TIR
AtMULE052	XXIX	1	16202118	16203395	1278	AAACTTAAA	AAACTTAAA	non-TIR
AtMULE053	XII	1	16299939	16301423	1485	AAGTATATA	AAGTATATA	non-TIR
AtMULE054	XXXIII	1	16540156	16559515	19360	TTTTTTGAA	TTTTTTGAA	non-TIR
AtMULE055	IX	1	16632384	16647338	14955	TTGTTTATA	TTGTTTATA	non-TIR
AtMULE056	XII	1	17169748	17171063	1316	AATATTTTA	AATATTGTA	non-TIR
AtMULE057	XXXXXXXIX	1	17329198	17333091	3894	TAAATATTTG	TAAATATTTG	non-TIR
AtMULE058	XXIX	1	17386096	17387318	1223	TATTAGTTA	TATTAGTTA	non-TIR
AtMULE059	XXXXXXXIX	1	17410692	17411725	1034	TTTCTTTAA	TTTCTTTAA	non-TIR
AtMULE060	II	1	17533331	17534947	1617	TATAATTTTT	TATAATTTTT	TIR
AtMULE061	I	1	17657989	17659056	1068	GTAAAACAA	GTAAAACAA	TIR
AtMULE062	XXII	1	17809699	17810001	303	TAATATAAA	TAATATAAA	TIR
AtMULE063	XXXXXXXXXII	1	18107475	18108551	1077	AAAATCTAA	AAAATCTAA	TIR
AtMULE064	XXIV	1	18116643	18119611	2969	ATTAGAAAAC	ATTAGAAAAC	TIR
AtMULE065	X	1	18158270	18159201	932	AATTTTTCTT	AATTTTTTTCT	TIR
AtMULE066	I	1	18177844	18178416	573	GAAAAAAA	GAAAAAAAA	TIR
AtMULE067	XXII	1	18269151	18269453	303	TTAAAAAAT	TTAAAAAAT	TIR
AtMULE068	II	1	18320041	18321634	1594	GTTATTAAAA	GTTATTAAAA	TIR
AtMULE069	XXVII	1	18989396	18989934	539	GATACTAGATG	GATACTAGATG	TIR
AtMULE070	IX	1	19057373	19066685	9313	TTGAAAAAA	TTGAAAAAA	non-TIR
AtMULE071	VII	1	19073870	19074548	679	TATTTATTA	TATTTATTA	non-TIR
AtMULE072	XXXIV	1	19095701	19101814	6114	ATAATTTAT	ATAATTTAT	non-TIR
AtMULE073	VI	1	19449586	19451058	1473	AAAAAATT	AAAAAATT	non-TIR
AtMULE074	VIII	1	20050434	20051731	1298	CTTTTATTA	CTTTTATTA	non-TIR
AtMULE075	XII	1	20211349	20212435	1087	TATATTTAAA	TATATTCAAA	non-TIR
AtMULE076	II	1	20344796	20345240	445	CTATCTAAAT	CTATCTAAAT	TIR

AtMULE077	IX	1	20405636	20408069	2434	TTATATAAA	TTATATAAA	non-TIR
AtMULE078	XXV	1	20701856	20703245	1390	ATTCTTTAA	ATTCTTTAA	non-TIR
AtMULE079	XXXXII	1	21113976	21115408	1433	TTTTTTCAA	TTTTTTCAA	non-TIR
AtMULE080	VII	1	21820764	21821440	677	TAAAATAAA	TAAAATAAA	non-TIR
AtMULE081	I	1	22271371	22272389	1019	TCTTCTTAAA	TCTTCTTAAA	TIR
AtMULE082	VI	1	22580899	22582473	1575	ATATATAAAAG	AAATATAAAAG	non-TIR
AtMULE083	XXV	1	23004346	23005944	1599	ATAAAATTT	ATAAAATTT	non-TIR
AtMULE084	XXIV	1	23132724	23133708	985	TAAATAGATG	TAAATAGATG	TIR
AtMULE085	XII	1	24449814	24450890	1077	AAAAATAAAGA	AAAAATAAAGA	non-TIR
AtMULE086	XXXXVI	1	24888831	24889885	1055	TAACATTTA	TAACATTTA	non-TIR
AtMULE087	VI	1	26083053	26084642	1590	AAAATTTA	AAAATTTA	non-TIR
AtMULE088	I	1	28235538	28236349	812	AATTATATTT	AATTATATTT	TIR
AtMULE089	XVI	1	29085648	29089297	3650	TATTAAAAA	TATTAAAAA	TIR
AtMULE090	XXII	1	3933195	3933505	311	TATAAACCTC	TTTAAATCTC	TIR
AtMULE091	XII	1	4111127	4113084	1958	TAGAATA	TAGAAAATA	non-TIR
AtMULE092	XXXXXXXVIII	1	4682379	4686167	3789	AACTAAAA	AACTAAAA	TIR
AtMULE093	XII	1	6236563	6237792	1230	TTTACATTA	TTTATTTTTA	non-TIR
AtMULE094	XII	1	8238648	8239815	1168	GTTGTTAAA	GTTTTAAA	non-TIR
AtMULE095	XXXXXXIV	1	8475190	8476367	1178	TACATACTT	TATATACTT	TIR
AtMULE096	XXIV	1	9201739	9202757	1019	TTTACTTTCA	TATAGTTTCA	TIR
AtMULE097	XII	1	9250061	9251243	1183	TATAATG	TATAAG	non-TIR
AtMULE098	I	1	10521056	10521424	369	GTTTCGA	GTTTTTGCGA	TIR
AtMULE099	VIII	1	10638291	10639192	902	TTTTATATTTA	TTTTAATAATTA	non-TIR
AtMULE100	XXXX	1	11718147	11719318	1172	CAAAAA	CTAAAAA	non-TIR
AtMULE101	IX	1	11766589	11769385	2797	TTTAGTTAA	TTTTTTTAA	non-TIR
AtMULE102	IX	1	12727794	12747642	19849	ATAATCAA	ATAAATAAA	non-TIR
AtMULE103	XXXXXXXXXIV	1	13277965	13290359	12395	TATAAGTTAA	TATAAATTAA	non-TIR
AtMULE104	XXXXXXXXXV	1	13620841	13622131	1291	GAGCCCATTCA	GAACCCTTCA	non-TIR
AtMULE105	XIX	1	13729205	13730199	995	TGCCCA	TGCCAA	non-TIR
AtMULE106	XXVII	1	13762026	13762295	270	AGAAAAT	AAAAAT	TIR
AtMULE107	X	1	13941186	13942227	1042	CCATTATTAAA	CCATTAAAAA	TIR
AtMULE108	XII	1	14019065	14019421	357	ATATGG	ATTTGG	non-TIR
AtMULE110	X	1	15468280	15468490	211	AAATATTTTT	AACATATTTTT	TIR
AtMULE112	VI	1	15581880	15582553	674	ATTCATT	ATTGATT	non-TIR
AtMULE113	XXVII	1	15694056	15694284	229	GTTTTCAT	GTTTTCT	TIR
AtMULE114	XXIV	1	15917484	15918555	1072	ATAACATATA	ATAGCATATA	TIR
AtMULE115	XII	1	15938783	15938996	214	AAGATTTT	AAAATTTT	non-TIR
AtMULE116	XII	1	15940103	15940322	220	AAATTTTTTG	AAACTTTTTG	non-TIR
AtMULE117	XXV	1	16035131	16035400	270	CCAATTTTCT	CCCTTTTCT	non-TIR

AtMULE118	XXXX	1	16052201	16054544	2344	ACTTTAATTTTT	ACCTAATTTGT	non-TIR
AtMULE120	I	1	16674283	16674624	342	TTTCTAG	TTTTTAG	TIR
AtMULE121	XII	1	18022378	18023482	1105	CAAAATAATA	CAAAAAAACA	non-TIR
AtMULE122	XII	1	19664692	19665568	877	TTATTATTT	TTGTTATTT	non-TIR
AtMULE123	XIII	1	20902712	20903407	696	TTATTAATT	TTATAAATT	non-TIR
AtMULE124	XXIX	1	21769189	21769418	230	AATTAAAA	AATCAAAA	non-TIR
AtMULE125	XII	1	21873731	21874597	867	CTATTTAGTTT	CTTTTAGTTT	non-TIR
AtMULE126	XIII	1	22596981	22597804	824	TTTTTTTTA	TTTCTTTTA	non-TIR
AtMULE127	XXXXV	1	24059426	24061672	2247	TTGTTTTTTT	TTGTTTCTTCT	non-TIR
AtMULE128	XXII	1	24314030	24314328	299	ATTTGAAAA	ATTCGAAAA	TIR
AtMULE129	II	1	25321996	25322397	402	CATTGATTAAC	CATTCATTAAC	TIR
AtMULE130	XXV	1	25350047	25350536	490	CAAATAGAAA	CAAATAGCAA	non-TIR
AtMULE131	VI	1	15265291	15266851	1561	TATTA	TTGTATTA	non-TIR
AtMULE132	XXIV	1	18111932	18112217	286	TATTTTT	TATTTTT	TIR
AtMULE134	XI	1	12332176	12332631	456	na	na	TIR
AtMULE135	X	1	12752599	12754311	1713	CAAATTAAT	CTAATTAAT	TIR
AtMULE136	IX	2	6278324	6293630	15307	TAGTATTAAAA	TAGTATTAAA	non-TIR
AtMULE137	XII	2	6609910	6611760	1851	ATAAACAAA	ATAAACAAA	non-TIR
AtMULE138	XXIV	2	151015	152197	1183	GTTGGATATA	GTTGGATATA	TIR
AtMULE139	XXXXVI	2	565757	581571	15815	TTAAGACAA	TTAAGACAA	non-TIR
AtMULE140	II	2	712345	712790	446	AAAAGAAATA	AAAAGAAATA	TIR
AtMULE141	XII	2	870447	871486	1040	AATATAAAA	AAGATAAAA	non-TIR
AtMULE142	XXXX	2	1009607	1011992	2386	TACATATTT	TATATATTT	non-TIR
AtMULE143	XXV	2	1084232	1085873	1642	TAATTATAA	TAATTATAA	non-TIR
AtMULE144	VII	2	1379616	1380293	678	ATAATCTA	ATAATCTA	non-TIR
AtMULE145	II	2	1651282	1651725	444	GTTTAGATTT	GTTTAGATTT	TIR
AtMULE146	XXXXI	2	1739510	1765561	26052	TAAAATGT	TAAAATGT	non-TIR
AtMULE147	VII	2	2084825	2085504	680	TTGTTTTAT	TTGTTTTAT	non-TIR
AtMULE148	VII	2	2139437	2140033	597	TTTAATTTTA	TTTAATTTTA	non-TIR
AtMULE149	XXVIII	2	2148944	2149862	919	TAGTTTAAA	TAGTTTAAA	TIR
AtMULE150	I	2	2208831	2209863	1033	GGGAGAAAT	GGGAGAAAT	TIR
AtMULE151	VI	2	2310064	2311454	1391	TTTTTTTA	TTTTTTTA	non-TIR
AtMULE152	XXIV	2	2589775	2590780	1006	GAACATAA	GAACATAA	TIR
AtMULE153	XXXXXXXXXVI	2	2764877	2778423	13547	TAAATTATT	TAAATTATT	non-TIR
AtMULE154	XII	2	3012496	3013408	913	ATATTAAAA	ATATTAAAA	non-TIR
AtMULE155	XXXXVI	2	3013552	3022010	8459	TATTATTAT	TATTATTAT	non-TIR
AtMULE156	XXVI	2	3056684	3057270	587	TAAAGGAAT	TAAAGGAAT	TIR
AtMULE157	XXVI	2	3056685	3057271	587	TAAAGGAATG	TAAAGGAATG	TIR
AtMULE158	XXXXII	2	3120463	3130232	9770	TTTTCTTTT	TTTTCTTTT	non-TIR

AtMULE159	XXXIII	2	3935166	3953793	18628	TTATATTTT	TTATATTTT	non-TIR
AtMULE160	XII	2	4046575	4047617	1043	AATTTTGTT	AATTTTTTT	non-TIR
AtMULE161	X	2	4052633	4053728	1096	CCTGGTAATT	CCTAGTAATT	TIR
AtMULE162	XXXX	2	4161370	4163740	2371	ATAAAATAA	ATAAAATAA	non-TIR
AtMULE163	VIII	2	4196362	4196640	279	ATTATAGAT	ATTATAGAT	non-TIR
AtMULE164	XII	2	4302280	4303180	901	AAAAAAGG	AAAAAAGG	non-TIR
AtMULE165	XXV	2	4598698	4599235	538	TTATTATAAA	TTATAAA	non-TIR
AtMULE166	XXVII	2	5005127	5005644	518	CATATTTAAATG	CATATTTAAATG	TIR
AtMULE167	XXXX	2	5139431	5162560	23130	TATTTTTTT	TTTTTTTTT	non-TIR
AtMULE168	XXVIII	2	5506042	5506889	848	AACATCATT	AACATCATT	TIR
AtMULE169	XII	2	5559633	5560713	1081	TTTAATAAA	TTTAATAAA	non-TIR
AtMULE170	XXXIV	2	5605561	5608753	3193	TATTTATCC	TATTTATCC	non-TIR
AtMULE171	XXIX	2	5681646	5682696	1051	TTATTTAA	TTATTTAA	non-TIR
AtMULE172	XII	2	5688188	5690083	1896	ATATTTTTG	ATATTTTTG	non-TIR
AtMULE173	I	2	5700317	5701035	719	ATATAAAAAC	ATTTAAAAAC	TIR
AtMULE174	XII	2	5709786	5711435	1650	AAAAATAAAA	AAAATAAAA	non-TIR
AtMULE175	XXVII	2	5809886	5814588	4703	ATATTAATAT	ATATTAATAT	TIR
AtMULE176	XXIII	2	5835003	5848806	13804	AATAAATTA	AATAAATTA	non-TIR
AtMULE177	XXXXVI	2	5908947	5924502	15556	TATTTTTTT	TATTTTTTT	non-TIR
AtMULE178	XXV	2	5992663	5996128	3466	TATTTTCAA	TATTTTCAA	non-TIR
AtMULE179	XXVIII	2	6077491	6090444	12954	CTATTTTA	CTATTTTTA	TIR
AtMULE180	XXXXVI	2	6541051	6556541	15491	TAATTCTTT	TAATTCTTT	non-TIR
AtMULE181	XII	2	6607613	6609648	2036	TATATTATT	TATATTATT	non-TIR
AtMULE182	XXIV	2	6829942	6836553	6612	CTTTTTTTTTA	CTTTTTTGTA	TIR
AtMULE183	XXII	2	6840896	6841864	969	ATTTTTAAA	ATTTTTAAA	non-TIR
AtMULE184	VII	2	6871130	6871804	675	AAAAATATA	AAAAATATA	non-TIR
AtMULE185	X	2	6943315	6944446	1132	ATAAAAAATT	ATAAAAAATT	TIR
AtMULE186	XXXXV	2	7015409	7019632	4224	TTGATAAAT	TTGATAAAT	non-TIR
AtMULE187	I	2	7104192	7105289	1098	ATTAGAATTT	ATTAGAATTT	TIR
AtMULE188	IX	2	7530124	7532829	2706	TAATAATAA	TAATAATAA	non-TIR
AtMULE189	XXXXII	2	7854498	7856189	1692	TTGAGATAA	TTGAAATAA	non-TIR
AtMULE190	XXIV	2	8150651	8151830	1180	TTATGCAAAA	TTATGCAAAA	TIR
AtMULE191	XII	2	8832770	8833891	1122	CTTTTATTT	CTTTTATTT	non-TIR
AtMULE192	I	2	9603858	9604944	1087	CTATTTTTTAC	CTATTTTTTAC	TIR
AtMULE193	XVI	2	14759491	14760598	1108	CCAATTTG	CCAAATTTG	TIR
AtMULE194	XXXXVI	2	1740436	1756081	15646	TTAATTTTT	TTAATTTTT	non-TIR
AtMULE195	XXXXVI	2	1206580	1207616	1037	TTTTTATAA	TTTTTATAA	non-TIR
AtMULE196	XII	2	1240374	1241520	1147	ATAAA	ATAAA	non-TIR
AtMULE197	VII	2	1443089	1443679	591	TAATTATAA	TAATATTA	non-

								TIR
AtMULE198	XXVII	2	1460260	1460490	231	ATACTA	ATACTA	TIR
AtMULE199	XXXXXXVIII	2	1508392	1508898	507	na	na	TIR
AtMULE200	XII	2	1730638	1730897	260	CTAAAAA	CTAAAAA	non-TIR
AtMULE202	XXV	2	1888983	1889245	263	ATTTTGT	ATTTTAGT	non-TIR
AtMULE203	XXIV	2	1898187	1898460	274	TTACTAA	TTCACTAA	TIR
AtMULE204	XXXXV	2	2003020	2009491	6472	na	na	non-TIR
AtMULE205	IX	2	2040960	2053220	12261	TTTTTGT	TGTTTTGAT	non-TIR
AtMULE206	XII	2	2064087	2065100	1014	AAAAATTAA	AAAAATTAA	non-TIR
AtMULE213	XII	2	2963945	2965161	1217	CATACTT	CTACTT	non-TIR
AtMULE215	XXXXII	2	3059811	3073310	13500	TTTAATAAA	TTTAACA	non-TIR
AtMULE216	XII	2	3189100	3189561	462	AAAAATCT	AAAAAGTTT	non-TIR
AtMULE217	XII	2	3192984	3193240	257	AATTTTTTTT	AATTTTTGTT	non-TIR
AtMULE218	XXXIV	2	4566540	4566799	260	ATTCCTT	ATTCCCT	non-TIR
AtMULE219	XXXXXXXXXVII	2	4592510	4594311	1802	ACTCATCA	ACTCATCCA	TIR
AtMULE220	X	2	4896327	4896959	633	AATTTATAAT	AATTTATAAT	TIR
AtMULE221	XII	2	5068331	5069632	1302	AATATTTTA	AATTATTT	non-TIR
AtMULE223	X	2	5697649	5698157	509	ATATATGAAT	ATATATTAAT	TIR
AtMULE224	XXIV	2	5714359	5714604	246	TTTTATAA	TTTTATAA	TIR
AtMULE225	X	2	5723222	5723430	209	CAAAA	CAAAA	TIR
AtMULE226	XXV	2	5748784	5751305	2522	TTGAAAAA	TTGAAAAA	non-TIR
AtMULE227	I	2	5771107	5772193	1087	ATGGAATATA	ATGGAATATA	TIR
AtMULE228	XXXXXXXXXVIII	2	6009003	6013920	4918	ATACAACTATC	ATACCACTATC	non-TIR
AtMULE230	XII	2	6527464	6533320	5857	AATTTT	AATTTT	non-TIR
AtMULE231	XII	2	6665778	6678755	12978	TTTAAAA	TTTAAAA	non-TIR
AtMULE233	IX	2	7237891	7244272	6382	TTTTATA	TTTTATA	non-TIR
AtMULE234	VII	2	7367428	7368125	698	CATATAAAAA	CATATATATA	non-TIR
AtMULE235	I	2	9298934	9299137	204	GACATATTT	GACATTTTT	TIR
AtMULE236	XXXXII	2	9435364	9435930	567	CCTCCATCCA	CCTCCATCAA	non-TIR
AtMULE238	X	2	10476187	10476395	209	TTTTA	TTTTA	TIR
AtMULE239	I	2	10538395	10539485	1091	TGGATTCAT	TGGATACAT	TIR
AtMULE240	X	2	11214698	11215900	1203	CACAAAT	CCCAAAT	TIR
AtMULE241	VI	2	11931214	11932599	1386	AAAATA	AAAATA	non-TIR
AtMULE242	VI	2	12687390	12688981	1592	ATATATATA	ATATATATA	non-TIR
AtMULE243	XXV	2	14350899	14354584	3686	ATTTATATA	ATTTATGTA	non-TIR
AtMULE244	I	2	14900488	14901600	1113	AAATTTTTGG	AAATTTTTGG	TIR
AtMULE245	XII	2	15972097	15972300	204	TTATTTTTT	TTTATTTTTT	non-TIR
AtMULE246	XXII	2	16544910	16545186	277	ATAAGTGTTA	ATTAAGTTTA	TIR
AtMULE249	XXV	2	1778106	1778401	296	ATTTA	TTTA	non-TIR
AtMULE250	XII	2	2652947	2654143	1197	TAAATAATA	TAAAAATA	non-

								TIR
AtMULE251	XXXXV	2	6957088	6964819	7732	TACATCAT	TACAGTCGT	non-TIR
AtMULE252	XXXXV	2	4821811	4836727	14917	TTCTTTTTTAA	TTCTTGTAA	non-TIR
AtMULE253	XXV	2	9008020	9008309	290	na	na	non-TIR
AtMULE254	V	2	14502396	14504597	2202	CTAGATTAG	CTAGACTAG	non-TIR
AtMULE255	XXVIII	2	16099080	16099999	920	TTTTATTAA	TTTTATTAA	TIR
AtMULE256	XXII	3	1924024	1924328	305	TTCTTTTTT	TTCTTTTTT	TIR
AtMULE257	XXIII	3	2772633	2773764	1132	TTTTAATTAA	TTTTAATTA	non-TIR
AtMULE258	XXXXII	3	2808433	2818903	10471	ATTTTTTTA	ATTTTTTTA	non-TIR
AtMULE259	I	3	3538077	3539124	1048	ATTTTTATTTG	ATTTTTATTTG	TIR
AtMULE260	XV	3	5377918	5378156	239	AAAATTTT	AAGATTTTT	non-TIR
AtMULE261	XXV	3	5929361	5930637	1277	TCTTGTTTTTG	TCTTGTTTTTG	non-TIR
AtMULE262	X	3	7978614	7979841	1228	CGATTTATTTC	CGATTTATTTC	TIR
AtMULE263	XXV	3	8100954	8102825	1872	TGAAAAAAA	TGAAAAAAA	non-TIR
AtMULE264	XXIV	3	8425709	8426240	532	TTTATAAAAA	TTTATAAAAA	TIR
AtMULE265	IX	3	8842885	8858142	15258	TTTTGTTTAA	TTTTGTTTAA	non-TIR
AtMULE266	XI	3	9474500	9474935	436	TATTGTAAAA	TATTGTAAAA	TIR
AtMULE267	XXIII	3	9723654	9735921	12268	TTTTTTTAA	TTTTTTTAA	non-TIR
AtMULE268	I	3	9807892	9808959	1068	ATTAATAAAT	ATTAATAAAT	TIR
AtMULE631	XXXXXXXXXIX	3	2189494	2192626	3133	na	na	non-TIR
AtMULE632	XXXXXXXXIX	3	1743543	1745949	2407	na	na	non-TIR
AtMULE269	VII	3	9828480	9829158	679	ATATTAGTT	ATATTAGTT	non-TIR
AtMULE270	IX	3	11658286	11660996	2711	TAAAAACAA	TAAAAACAA	non-TIR
AtMULE271	XII	3	13129588	13129805	218	AAAAAAAA	AAAAAAAAA	non-TIR
AtMULE272	XII	3	15032489	15038897	6409	CGTATTTT	CGTATTTT	non-TIR
AtMULE273	X	3	18343394	18344011	618	GAATATCATT	GAATATCATT	TIR
AtMULE274	I	3	20529497	20530584	1088	AAACTATAAA	AAACTATAAA	TIR
AtMULE275	XXV	3	2259600	2260970	1371	TATTTAAA	TATCTAAA	non-TIR
AtMULE276	I	3	2970571	2971585	1015	GTAATA	GTAATA	TIR
AtMULE278	XV	3	5377134	5377370	237	ACAAATACGT	ATAAATATGT	TIR
AtMULE279	XXVIII	3	6486262	6487180	919	CAAAGAA	CAAAAGAA	TIR
AtMULE280	XXII	3	8003280	8003580	301	TATTT	TAATTT	TIR
AtMULE281	X	3	8343570	8343800	231	TAAAG	TAAAG	TIR
AtMULE282	XXV	3	10008100	10008300	201	ATGTATTA	ATGTATA	non-TIR
AtMULE283	XXXXII	3	10440304	10441400	1097	TAAAGTATTT	TAAAGTATTT	non-TIR
AtMULE284	XII	3	10764600	10764800	201	TTTACAA	TTTACAA	non-TIR
AtMULE285	XXVIII	3	10863410	10865500	2091	TATAACAAA	TAATACAAA	TIR
AtMULE286	I	3	11021485	11022396	912	TTTTATTTG	TATTTTTTG	TIR
AtMULE287	XXXXVI	3	11176600	11181400	4801	ATTTGC	ATTTAC	non-TIR
AtMULE288	X	3	11184492	11184700	209	CTTTTTT	CTTTTTT	TIR

AtMULE289	XI	3	11263100	11263599	500	GATATATC	GATTTATC	TIR
AtMULE290	XXIV	3	11294500	11294800	301	TACATA	TACTA	TIR
AtMULE291	XXXXII	3	11295000	11296600	1601	GATTA	GATTA	non-TIR
AtMULE292	XII	3	11315900	11317100	1201	CCTTTT	CCTTTT	non-TIR
AtMULE293	XII	3	11320200	11320400	201	AAAAGAA	AAAATAAA	non-TIR
AtMULE294	XII	3	11373600	11375397	1798	GTAAAAT	GTAAAAAAT	non-TIR
AtMULE295	X	3	11675400	11676400	1001	TTTGGGAC	TTTGTGAC	TIR
AtMULE296	XXXXVI	3	11705500	11705800	301	CTATT	CTATT	non-TIR
AtMULE297	XXXXVI	3	11706200	11706500	301	ATATCGA	ATATCTA	non-TIR
AtMULE298	XXXXVI	3	11706900	11707200	301	TCCCT	TCCCT	non-TIR
AtMULE299	XXXXVI	3	11707700	11708000	301	AGAAGA	AGAATA	non-TIR
AtMULE305	XXXIV	3	11716200	11716495	296	AGAATCCAC	ATAATGAC	non-TIR
AtMULE306	XXIV	3	11727103	11727400	298	TTTTGATT	TTTTTTT	TIR
AtMULE307	XII	3	11742500	11743600	1101	TTGTTTT	TTTTTTT	non-TIR
AtMULE308	XXXXVI	3	11765100	11766895	1796	GAACCACAAAAA	GACAACAAATA	non-TIR
AtMULE309	I	3	11863500	11864000	501	ATTTTTTTCTC	ATTTCTC	TIR
AtMULE310	XII	3	12017100	12017800	701	CTTATGAAA	CTTTTGTAAA	non-TIR
AtMULE311	XII	3	12078595	12094300	15706	AAGAA	AAGAA	non-TIR
AtMULE312	XXXX	3	12098600	12098800	201	TAATAATAA	TATTAATAA	non-TIR
AtMULE314	XXXXXXXXV	3	12466901	12467300	400	AAATTACCTTT	AAATACTTT	non-TIR
AtMULE315	XXXXXXXXVI	3	12468392	12468694	303	AAAAAG	AAAAAG	non-TIR
AtMULE316	XXIV	3	12573898	12574795	898	CTTTTT	CTTTTT	TIR
AtMULE317	XXV	3	12670103	12671785	1683	TTTAA	TTTAA	non-TIR
AtMULE319	XXIV	3	12821805	12824691	2887	TTTATTTTT	TTTTTT	TIR
AtMULE323	XII	3	13262898	13264596	1699	TAAATTGAT	TAAATTGT	non-TIR
AtMULE324	XXXXVI	3	13335707	13346000	10294	TCTTTATTTATG	TCTTATTTATG	non-TIR
AtMULE325	IX	3	13346204	13358900	12697	CGTTTGTCA	CTTTTGTCTA	non-TIR
AtMULE326	IX	3	13546500	13574300	27801	CACTC	CATCTC	non-TIR
AtMULE331	XIX	3	14316403	14318190	1788	AAACATA	AAACAATA	non-TIR
AtMULE332	XXXIV	3	14600994	14602300	1307	TCAATAG	TCTAAGAG	non-TIR
AtMULE333	XII	3	14692396	14692600	205	TAAAAG	TAAAAAAAG	non-TIR
AtMULE336	IX	3	15152300	15154900	2601	TTCACCT	TTCTCCT	non-TIR
AtMULE337	XXXX	3	15176900	15177800	901	ATATTTTTC	AAATTTTTC	non-TIR
AtMULE338	XXXXII	3	15305600	15310900	5301	AGTTTT	AGTTTT	non-TIR
AtMULE339	IX	3	15616000	15627400	11401	GTAAAATTGC	GAAAATTTC	non-TIR
AtMULE341	XXXX	3	15730401	15730707	307	AAATTTCATTA	AAATCATCA	non-TIR
AtMULE342	X	3	15748800	15749595	796	CAAATAAA	CAATTAAA	TIR

AtMULE343	X	3	15955100	15963500	8401	CAATAATTTTA	CAATAAATATTA	TIR
AtMULE344	XII	3	16160300	16160500	201	GAATC	GAATC	non-TIR
AtMULE345	XXXXVI	3	16396805	16402090	5286	TTTGTTAAAATG	TTTTTTAATG	non-TIR
AtMULE346	XII	3	16437596	16437900	305	GTTTTAA	GTTTTAA	non-TIR
AtMULE347	I	3	16439500	16440200	701	TCTTTA	TTTTTA	TIR
AtMULE348	XXXXXXXIX	3	16536800	16537804	1005	TATAATTT	TATAATTT	non-TIR
AtMULE349	XII	3	16565000	16565900	901	CTTTTTTTTAC	CTTTTTTTTTAC	non-TIR
AtMULE350	X	3	16627809	16628998	1190	ATTTTTTTAT	ATTTTTTTAT	TIR
AtMULE351	XII	3	16712900	16714300	1401	TTTTCA	TTTTACA	non-TIR
AtMULE352	IX	3	16806400	16808900	2501	TTTTA	TTTTA	non-TIR
AtMULE353	XII	3	17022512	17022900	389	AAATCTGT	AAATCTCT	non-TIR
AtMULE355	VII	3	17343909	17344600	692	TAAGATTA	TAAGTTTA	non-TIR
AtMULE356	XXXXVI	3	17710400	17713000	2601	CAAGTT	CAAATT	non-TIR
AtMULE357	XXVI	3	17714707	17715308	602	ATTTAATTTT	ATTTTTATTT	TIR
AtMULE359	XXXX	3	18406200	18408600	2401	GAAATTT	GAAAGTT	non-TIR
AtMULE360	VII	3	18754197	18754891	695	TGGATATG	TGGATATG	non-TIR
AtMULE361	VI	3	18925700	18927981	2282	CTTTT	CTTTT	non-TIR
AtMULE362	XXIII	3	19155992	19157116	1125	TAAAAAA	TAAAAAA	non-TIR
AtMULE363	XXVIII	3	19474400	19474894	495	AAAAAC	AAAAAC	TIR
AtMULE364	XI	3	20320100	20320498	399	GAAATGTT	GAAAATGATT	TIR
AtMULE365	XII	3	20950700	20952000	1301	AAATATAA	AAAAAAA	non-TIR
AtMULE366	XII	3	20952993	20954000	1008	TTTTTTTT	TATTTGT	non-TIR
AtMULE367	XI	3	20981600	20982794	1195	TCATTTATT	TCATTATT	non-TIR
AtMULE368	X	3	21426500	21426700	201	TAAGTTTA	TATGTTTA	TIR
AtMULE370	X	3	3837773	3838982	1210	CAAAA	CAAAA	TIR
AtMULE371	XII	3	11891489	11891709	221	AAAATGTTT	AAAAGTTT	non-TIR
AtMULE372	XXXX	3	12106699	12122791	16093	TTATAAAA	TTAAAAAA	non-TIR
AtMULE373	XII	3	12137006	12138185	1180	CTTTT	CTTTT	non-TIR
AtMULE374	X	3	13393582	13394184	603	AAATTAT	AAATTT	TIR
AtMULE375	XXXIV	3	15947000	15949295	2296	AATTCTTGTA	ATTTATTTTA	non-TIR
AtMULE376	XXV	3	18071592	18071893	302	na	na	non-TIR
AtMULE377	XXIX	4	696411	697629	1219	TTAATTTT	TTAATTTT	non-TIR
AtMULE378	XXV	4	1089531	1090749	1219	ATATAAAAAT	ATATAAAAT	non-TIR
AtMULE379	I	4	1232446	1233635	1190	CAAAAAAAAAAC	CAAAAAAAAAAAC	TIR
AtMULE380	III	4	1452890	1453956	1067	GTATGTACCT	GTATGTACCT	TIR
AtMULE381	VIII	4	1486168	1487410	1243	TTAAGCAAA	TTAAGCAAA	non-TIR
AtMULE382	XII	4	1865637	1881269	15633	TAATTAAAA	TAATTAAAA	non-TIR
AtMULE383	XXXIII	4	1889531	1908022	18492	TTAAAATTA	TTAAAATTA	non-TIR

AtMULE384	XII	4	2070974	2071998	1025	TAAAAGGTAT	TAAAAGTAT	non-TIR
AtMULE385	VII	4	2167866	2168540	675	TTAAATATA	TTAAATATA	non-TIR
AtMULE386	XXVIII	4	2495414	2495655	242	TTTATATTA	TTTATATTA	TIR
AtMULE387	I	4	2603433	2604531	1099	ATTTTACTAA	ATTTTACTAA	TIR
AtMULE388	VI	4	2638391	2639959	1569	TACTATA	TACTATA	non-TIR
AtMULE389	I	4	2688302	2689427	1126	ATTTTATCTT	ATTTTATCTT	TIR
AtMULE390	IX	4	2695882	2712963	17082	TTATATAAA	TTATATAAA	non-TIR
AtMULE391	XII	4	3059896	3061914	2019	AACATATTA	AACATATTA	non-TIR
AtMULE392	XII	4	4141327	4143327	2001	TATTCGAAA	TATTCGAAA	non-TIR
AtMULE393	XXV	4	4165359	4168894	3536	TATTTTTAA	TATTTTTAA	non-TIR
AtMULE394	X	4	4286080	4286631	552	TTTAATTTTTT	TTTAATTTTT	TIR
AtMULE395	XXIV	4	4323065	4324174	1110	GTGAAATAAT	GTGAAATAAT	TIR
AtMULE396	XXII	4	4421250	4421553	304	TTATATATA	TTATATATA	TIR
AtMULE397	VI	4	4435811	4437283	1473	TATAATTTA	TATAATTTA	non-TIR
AtMULE398	XXXX	4	4485093	4501249	16157	TTTATTTTA	TTTATTTGA	non-TIR
AtMULE399	XI	4	4521819	4522230	412	TATTTAATTT	TATATAATTT	TIR
AtMULE400	XXI	4	4522366	4524553	2188	TCACCATATG	TCACCATATG	non-TIR
AtMULE401	XXVIII	4	4646660	4647577	918	ATATATAAA	ATATATAAA	TIR
AtMULE402	IX	4	4648787	4662732	13946	ATATGAATAA	ATATGAATAA	non-TIR
AtMULE403	XI	4	4692705	4693136	432	AATCTCTATC	AATCTATATC	TIR
AtMULE404	I	4	4755688	4756767	1080	CATTTCTTGTC	CATTTCTTGGC	TIR
AtMULE405	XII	4	4809898	4811071	1174	AAAATTCT	AAAATTCT	non-TIR
AtMULE406	XII	4	4830906	4832038	1133	ATATAATCG	ATATAATCG	non-TIR
AtMULE407	XII	4	4874497	4875568	1072	ATTTAATCG	ATATAATCG	non-TIR
AtMULE408	XXXXI	4	4908130	4916475	8346	AAATGTCTGGAG	AAATGTCTGGAG	non-TIR
AtMULE409	XXVIII	4	5075196	5076043	848	AAAAAGATA	AAAAAGATA	TIR
AtMULE410	V	4	5078594	5081369	2776	CAAAATGAAG	CAAAATGAAG	non-TIR
AtMULE411	XXV	4	5993433	5997447	4015	AAATAAATA	AAATAAATA	non-TIR
AtMULE412	XXIII	4	6048920	6049595	676	TAGTTTAAG	TAGTTTAAG	non-TIR
AtMULE413	XII	4	6526132	6530187	4056	CAAGAACAA	CAAGAACAA	non-TIR
AtMULE414	VII	4	7095491	7096168	678	TTGAATATA	TTGAATATA	non-TIR
AtMULE415	XXII	4	7683144	7683448	305	TAAAAAAAA	TAAAAAAAA	TIR
AtMULE416	III	4	9136437	9140207	3771	ACAATTAAT	ACAATTAAT	TIR
AtMULE417	XII	4	9235416	9236528	1113	TATAAATAA	TATAAATAA	non-TIR
AtMULE418	XXIII	4	9365541	9366661	1121	ATAAAAAAATGG	ATAAAAAAATGG	non-TIR
AtMULE419	XIII	4	9502333	9503164	832	ATTATAAAA	ATTATAAAA	non-TIR
AtMULE420	V	4	9924300	9926462	2163	TGGTTATGT	TGGTTATGT	non-TIR
AtMULE421	I	4	296631	297716	1086	TTGTCTTATGGG	TTCGCCTTATGGG	TIR
AtMULE422	XXIV	4	590477	591793	1317	ATATAATAA	ATAATATAA	TIR

AtMULE423	XXXXXXXXXIX	4	857434	861278	3845	CATCTTTTT	CATATTGTTT	non-TIR
AtMULE425	II	4	1022630	1023074	445	GATAAAGAATA	GATAAGCATA	TIR
AtMULE426	XXII	4	1115380	1115686	307	TTAAA	TTAAA	TIR
AtMULE429	IX	4	1448782	1462177	13396	TTGTGTAAAA	TTGCGTAAAA	non-TIR
AtMULE430	XXXXVI	4	1908629	1913169	4541	TATATA	TATATA	non-TIR
AtMULE431	XXXXII	4	1910360	1912327	1968	AGATAGAT	AGATATAT	non-TIR
AtMULE432	XXIV	4	2077701	2077929	229	TTTTTTCT	TATTTTCT	TIR
AtMULE433	IX	4	2233870	2236576	2707	CTTATTAAA	CTTACAAA	non-TIR
AtMULE435	XXXXXXI	4	2274570	2276790	2221	TACTCT	TACTAT	non-TIR
AtMULE436	XXXXXXXXXII	4	2429558	2430888	1331	ATTCTACTT	ATTCTACTT	TIR
AtMULE437	XXXX	4	2442210	2444360	2151	AATTTTT	AATTTTT	non-TIR
AtMULE440	XII	4	3074797	3075016	220	AAATTTTTTTT	AAATATTTTT	non-TIR
AtMULE441	XXXX	4	3256521	3257460	940	GTTAGAGAT	GTTAGTGAT	non-TIR
AtMULE442	XXXX	4	3432131	3445733	13603	AATAAAAAT	AATAAAAAT	non-TIR
AtMULE443	XXXXVI	4	3774666	3787203	12538	CAAAAAAAA	CAAAAACAAA	non-TIR
AtMULE444	XII	4	3921458	3922160	703	TGTTTAAA	TGTTTATA	non-TIR
AtMULE445	XI	4	4155705	4156141	437	AATTATAATTG	AATTGTAATTG	TIR
AtMULE446	XII	4	4346983	4348250	1268	AAAAGAT	AAAAGAT	non-TIR
AtMULE447	XII	4	4358416	4358632	217	AAAAATTG	AAAAATTC	non-TIR
AtMULE448	XXV	4	4387227	4388568	1342	TTTCATTG	TTCCATTG	non-TIR
AtMULE449	XXVIII	4	4482710	4483615	906	ATTAAGAGTGGGA	ATTAAGAGTGGGA	TIR
AtMULE450	XVI	4	4506770	4510419	3650	TGTTAT	TGATTAT	TIR
AtMULE451	XXVI	4	4518174	4518522	349	CAATTTTTAGTGGTAA	CAATTTTAGTGGTAA	TIR
AtMULE452	XII	4	4576825	4578480	1656	AAAGTTT	AAAGTTT	non-TIR
AtMULE454	XXIV	4	4744434	4745060	627	GATTTTTTTT	GATTTTTATT	TIR
AtMULE455	XXXVI	4	4899780	4901230	1451	AAAGATT	AAAGTTT	non-TIR
AtMULE456	XXXXXXXIX	4	5080230	5080790	561	TTTTAAATC	TTTTAAAC	non-TIR
AtMULE457	XII	4	5132689	5132930	242	CTTTTAAA	CTTTTAAA	non-TIR
AtMULE458	XXXXV	4	5162580	5164820	2241	TATGTTA	TATGTTA	non-TIR
AtMULE459	XXVII	4	5217324	5217955	632	CCTAAACAAAT	CCTTAAAAAAT	TIR
AtMULE460	XXXXXX	4	5995205	5996950	1746	TCCAA	TCCTAA	non-TIR
AtMULE461	X	4	7598170	7599360	1191	AAAGTTA	AAAGTTA	TIR
AtMULE462	XII	4	9596147	9596430	284	TAATATGA	TAATAGA	non-TIR
AtMULE463	X	4	9596862	9597070	209	TCATAAAA	TCTAAAA	TIR
AtMULE464	XXIX	4	9757554	9758840	1287	ATACATCTA	ATACATCTA	non-TIR
AtMULE465	II	4	12281511	12281988	478	TTTCTCAAAG	TTTGTCAAAG	TIR
AtMULE467	XXXX	4	13977400	13979800	2401	TCTTTCT	TCTTTTT	non-TIR
AtMULE468	X	4	14362692	14363600	909	TCAAAT	TCAAAT	TIR
AtMULE469	XXIX	4	14887709	14888915	1207	AACTAGA	AACTAGTA	non-TIR

AtMULE470	II	4	15795712	15796588	877	na	na	TIR
AtMULE471	I	4	1232678	1233410	733	GAAACAAG	GATTATAG	TIR
AtMULE472	XII	4	2797639	2798033	395	AAAAAAAA	ATAAAAA	non-TIR
AtMULE473	VII	4	7407757	7408434	678	AAAAAAAAA	AAAAAAATA	non-TIR
AtMULE474	XXXXXXXV	4	1983240	1987650	4411	na	na	non-TIR
AtMULE475	I	4	7823600	7823960	361	na	na	TIR
AtMULE476	XXII	5	112815	113991	1177	TATTATTAT	TATTATTAT	TIR
AtMULE477	IX	5	1065397	1067896	2500	TTTATCTAA	TTTATCTAA	non-TIR
AtMULE478	XXIV	5	2347738	2348655	918	ATAAAATTA	ATAAAATTA	TIR
AtMULE479	I	5	3005881	3007474	1594	TTTCTTTTT	TTTCTTTTT	TIR
AtMULE480	V	5	3944950	3947154	2205	TAAGGTAAGG	TAAGCTAAGG	non-TIR
AtMULE481	XXII	5	4790701	4791003	303	ATTAATAAT	ATTAATAAT	TIR
AtMULE482	XXXXXXXXXI	5	5126038	5129443	3406	ATTTTCTTT	ATTTTCTTT	TIR
AtMULE483	XXXXXXXXX	5	5232021	5235932	3912	TAGTTATAA	TAGTTATAA	TIR
AtMULE484	XXIX	5	5797851	5799067	1217	ACGAGACTA	ATGAGACTA	non-TIR
AtMULE485	XXXXXXXXV	5	5981757	5983347	1591	TTAATTTA	TTAATTTA	non-TIR
AtMULE486	II	5	6473543	6473957	415	CAATTTAAATA	CAATTTAAATA	TIR
AtMULE487	II	5	6938692	6939137	446	AATTAACACC	AATTAACATCC	TIR
AtMULE488	XXV	5	7483241	7486510	3270	TTCTTTTAA	TTCTTTTAA	non-TIR
AtMULE489	XXXXXXXXIX	5	8193533	8193980	448	ATTTAACAAA	ATTTAACAAA	TIR
AtMULE490	XXVII	5	8218031	8218662	632	TTTTTTAT	TTTTTTAT	TIR
AtMULE491	XII	5	8745730	8746917	1188	TTGATTATT	TTGATTATT	non-TIR
AtMULE492	XXXXXXXXVIII	5	9076587	9079791	3205	TTTTTTTAA	TTTTTATAA	non-TIR
AtMULE493	XXV	5	9125549	9126905	1357	AAAAAAATA	AAAAAATTA	non-TIR
AtMULE494	XXIV	5	9305265	9310193	4929	ATATAAAATG	ATATAAAATG	TIR
AtMULE495	XXIV	5	9315405	9316595	1191	TGTTAAGAG	TGTTAATAG	TIR
AtMULE496	XII	5	9362829	9363915	1087	TATTAACAA	TATTAACAA	non-TIR
AtMULE497	XXXXXXXXV	5	9437101	9437401	301	TTTTTAA	TTTTTAA	non-TIR
AtMULE498	XXXXXXXXVI	5	9961751	9965722	3972	TAGTATCAAC	TAGTATCAAC	TIR
AtMULE499	XXXXXXXXVII	5	10194430	10202305	7876	TACATTTAA	TACATTTAA	non-TIR
AtMULE500	XXXXII	5	10281969	10302320	20352	TATATATATAT	TATATATAT	non-TIR
AtMULE503	XXXXII	5	10474279	10488200	13922	TTAAATCAAA	TTAAATCAAA	non-TIR
AtMULE504	XII	5	10572573	10574743	2171	TATATTTTA	TATATTTA	non-TIR
AtMULE505	XXXXXXXXII	5	10725283	10725552	270	CTCAATAT	CCCAATAT	TIR
AtMULE506	XXXXII	5	10885817	10887111	1295	ATGATTATT	ATGATTATT	non-TIR
AtMULE507	XII	5	10909182	10910362	1181	TTTAAACTA	TTTAAACTA	non-TIR
AtMULE508	XII	5	10957394	10958532	1139	AAAAAATAA	AAAAAATAA	non-TIR
AtMULE509	X	5	10959305	10959777	473	CTGTTAAAATAT	CTGTTAAAATAT	TIR
AtMULE510	XXXXXXXXV	5	10960474	10960836	363	TTATTTTTTA	TTATTTTTA	non-TIR
AtMULE511	XXXXII	5	11042425	11060339	17915	ATAATATTA	ATAATATTA	non-TIR

AtMULE512	XXXXII	5	12518896	12534752	15857	GTGAAAAGTG	GTGAAAAGTG	non-TIR
AtMULE513	XXXXII	5	13209005	13224749	15745	TTTCTTATAT	TTTCTATAT	non-TIR
AtMULE514	IX	5	13262664	13264874	2211	GATAAAAA	GATAATAA	non-TIR
AtMULE515	XXXXXXXXVI	5	13759234	13763244	4011	TAGCATAATT	TAGCATAATT	TIR
AtMULE516	IX	5	13816735	13818203	1469	AATTGTATT	AATTGTATT	non-TIR
AtMULE517	XV	5	14016130	14017313	1184	CTTTCTAATC	CTTTCTAATC	TIR
AtMULE518	XXXXII	5	14162166	14185836	23671	TATCTATTA	TATCTATTA	non-TIR
AtMULE519	I	5	14203627	14204695	1069	TTTTTGTAA	TTTTTGTAA	TIR
AtMULE520	XXIX	5	14436962	14438156	1195	TACATATA	TACATATA	non-TIR
AtMULE521	IX	5	14506286	14521495	15210	TTAAGTATA	TTAAGTATA	non-TIR
AtMULE522	XXXXXXXXV	5	14527572	14529490	1919	TAAAATTAA	TAAAATTAA	non-TIR
AtMULE523	XII	5	14578299	14585145	6847	TCCCAATTATAAA	TCCCAATTATAAA	non-TIR
AtMULE524	XXV	5	14589648	14593888	4241	AGAAAAC	AGAAAAC	non-TIR
AtMULE525	XII	5	14629466	14630987	1522	TAGATTTAA	TAGATTTAA	non-TIR
AtMULE526	I	5	14681179	14685130	3952	GTTTTTTTC	GTTTTTTTTC	TIR
AtMULE527	XXXXII	5	14830170	14838113	7944	CAAATTCTT	CAAATTCTT	non-TIR
AtMULE528	XXXXXXXXV	5	14875719	14875928	210	AAAAAAATT	AACAAAATT	non-TIR
AtMULE529	XXVII	5	14950464	14951083	620	ATTTAATTTT	ATTTAATTTT	TIR
AtMULE530	I	5	15018094	15019181	1088	GTTTCTTGTT	GTTTCTTGTT	TIR
AtMULE531	XII	5	15109677	15110883	1207	AAAGTTTAAA	AAATTCAAA	non-TIR
AtMULE532	XXXXXXXXIV	5	15209416	15211433	2018	CTAATCAAAA	CTAATCAAAA	non-TIR
AtMULE533	VI	5	15304930	15306401	1472	TAATTCTAA	TAATTCTAA	non-TIR
AtMULE534	XXXXXXXXII	5	15457871	15459559	1689	TTATAATTGGG	TTATAATTGGG	TIR
AtMULE535	XII	5	16564271	16565346	1076	AAACAAAAA	AAACAAAAA	non-TIR
AtMULE536	XXII	5	16645923	16646226	304	TTTATTTTA	TTTATTTTA	TIR
AtMULE537	II	5	16974482	16974927	446	AAATGAGAAG	AAATGAGAAG	TIR
AtMULE538	XX	5	17055149	17055476	328	CATCCTTAAC	CATCCTTAAC	TIR
AtMULE539	XXV	5	17089406	17090762	1357	ATATAGAAT	ATATAGAAT	non-TIR
AtMULE540	XXXXXXXVIII	5	17255111	17258890	3780	ATTTTTTTA	ATTTTTTTA	TIR
AtMULE541	X	5	17278637	17280791	2155	AATTGTAAAT	AATTGTAAAT	TIR
AtMULE542	XXII	5	17350874	17351178	305	TTTTTTTTA	TTTTTTTTA	TIR
AtMULE543	XXXXXXXX	5	17409342	17411094	1753	TTGAGAATT	TCGAGAATT	non-TIR
AtMULE544	IX	5	17854170	17856508	2339	TTTCCATAA	TTTCCATAA	non-TIR
AtMULE545	XXIII	5	18123318	18135585	12268	ATAAATAAA	ATAAATAAA	non-TIR
AtMULE546	XXXXXXXIX	5	18175140	18176160	1021	CTAAATAATA	CTAAATAATA	non-TIR
AtMULE547	XXV	5	18671226	18673821	2596	AATAAACAA	AATAAACAA	non-TIR
AtMULE548	XXII	5	18812788	18813091	304	TAAAAAAAA	TACAAAAAA	TIR
AtMULE549	XXVI	5	21425968	21426246	279	AATATCGAT	AATATTGAT	TIR
AtMULE550	XXXXXXXVIII	5	22833815	22837594	3780	TAAAAAATA	TAAAAAATA	TIR
AtMULE551	XI	5	23326747	23327181	435	GTTAAAAT	GTTAAAAT	TIR

AtMULE552	XXV	5	23778172	23779549	1378	CAAAAAAAAAAA	CAAAAAAAAAA	non-TIR
AtMULE555	XXXXXXXVII	5	6643393	6648012	4620	CCATATTGACTAA	CCATATTACTAA	TIR
AtMULE556	XXV	5	6926466	6928341	1876	TTAAAAAA	TTTAAAAAA	non-TIR
AtMULE557	XXXXXXXVI	5	7048279	7051143	2865	TCCCATC	TCCCATTC	non-TIR
AtMULE558	I	5	8383522	8384590	1069	ATTTATTGA	ATTAATTGA	TIR
AtMULE559	VII	5	8474866	8475525	660	TTTATTTT	TTTATTAT	non-TIR
AtMULE560	XXXXXXXV	5	9437635	9438521	887	AAATATTTT	AAATTCTTTT	non-TIR
AtMULE562	IX	5	9780135	9783300	3166	TTAATTAAAA	TAAAATTAAAA	non-TIR
AtMULE563	I	5	10043510	10043725	216	ATTATAA	ATTACAA	TIR
AtMULE564	XXXXXXXI	5	10238691	10247445	8755	ATTTAG	ATTTAG	non-TIR
AtMULE565	VI	5	10239419	10240432	1014	TTAAAA	TGAAAA	non-TIR
AtMULE568	XXXXXXXIV	5	10295406	10296732	1327	CCTCCT	CCTCCT	non-TIR
AtMULE570	XXXXXXXIII	5	10617396	10618914	1519	TTTTCCTT	TTTTTCTT	non-TIR
AtMULE571	XXXXXXXII	5	10638727	10649951	11225	TTTGTTTTGTTTTTT	TTATTTTTTTTGT	non-TIR
AtMULE572	IX	5	10685539	10688225	2687	TTTATAATA	TTTATAATA	non-TIR
AtMULE573	XXXXXXXII	5	11023220	11025543	2324	CATCACTTT	CATTCATTTT	non-TIR
AtMULE574	XXXXXXVII	5	11540353	11540618	266	AGTATTATAG	AGTTTATG	non-TIR
AtMULE575	XXXXXXVI	5	12565111	12568430	3320	GTATAAAT	GTATACAT	non-TIR
AtMULE576	XIV	5	12590352	12591080	729	TAAAGAGTTTTTA	TAAATTGTTTTA	non-TIR
AtMULE577	XXXXXXXV	5	12982066	12985330	3265	TTTAAA	TTATAAA	non-TIR
AtMULE579	IX	5	13282342	13287913	5572	ACAACGG	ACAACGG	non-TIR
AtMULE581	XII	5	13816735	13818203	1469	AATTGTATT	AATTGTATT	non-TIR
AtMULE582	XIX	5	13964499	13972451	7953	CATTATTTA	CATATTTA	non-TIR
AtMULE583	IX	5	14446986	14453398	6413	ATTAAGTATA	ATTAGTAA	non-TIR
AtMULE585	I	5	14682160	14684590	2431	AAGCT	AAGCT	TIR
AtMULE586	XXVII	5	14944758	14944979	222	TTTAATG	TTTATTG	TIR
AtMULE587	X	5	15052874	15053591	718	AATTAGTATA	AATTAATATA	TIR
AtMULE588	XXXXXXIV	5	15085227	15085452	226	AAACGA	AAACCA	TIR
AtMULE589	XXVII	5	15123228	15123432	205	AACATAAA	AACAAAA	TIR
AtMULE591	XXXXXXIII	5	15720904	15721551	648	TAAAACTCT	TAAAAAACTTT	non-TIR
AtMULE592	XII	5	15721707	15722067	361	TTTTATTGAA	TTTTTTTGAA	non-TIR
AtMULE593	XII	5	15815885	15816507	623	TAAATATAT	TAAAACATAT	non-TIR
AtMULE594	XVIII	5	15879728	15880405	678	TTGTTTCCA	TTGTTCATA	non-TIR
AtMULE595	XXXXXXII	5	16018787	16019034	248	CCAAAAAAAAAT	CCAAAAAAAAT	non-TIR
AtMULE596	VIII	5	16052367	16052677	311	CAAATTCTCTT	CAATTCTCTT	TIR
AtMULE597	IX	5	16053003	16055488	2486	TTTTATTAAA	TTTTATTAAA	non-TIR
AtMULE598	X	5	16112679	16113857	1179	ATATTTTTC	ACATTTTTC	TIR
AtMULE599	XXXXXXXXXII	5	16911949	16914162	2214	AATATGT	AATATCT	non-TIR

AtMULE600	XII	5	17033522	17034488	967	TAAATCAC	TAAAATAC	non-TIR
AtMULE601	X	5	17278996	17279837	842	TATATAACATAC	TATATAATAAGATAC	TIR
AtMULE602	XXXXXXXXI	5	17870903	17873240	2338	AGAATTA	AGGAATTA	non-TIR
AtMULE603	XXXXXXXI	5	18480846	18483414	2569	AAAGACA	AAAGAGA	non-TIR
AtMULE604	VI	5	18483569	18484582	1014	TTTTTTTCA	TTATTTTTA	non-TIR
AtMULE605	XXII	5	18645834	18646121	288	TTTTTATCTATTATC	TTTTTTCATTTATC	non-TIR
AtMULE607	VII	5	20643076	20644232	1157	AAAAAAT	AAAAAAT	non-TIR
AtMULE610	VI	5	10867994	10868979	986	AATATTTAAT	AATATTTAAT	non-TIR
AtMULE611	X	5	10881334	10881899	566	TTTTTCTTT	TTTTTACTAT	TIR
AtMULE613	XXXXII	5	13575761	13576662	902	na	na	non-TIR
AtMULE615	II	5	8966561	8966812	252	na	na	TIR
AtMULE616	XXVII	5	13386664	13386977	314	na	na	TIR
AtMULE617	I	5	18984248	18984511	264	na	na	TIR
AtMULE618	XII	4	16414620	16415801	1182	AAATTTAA	AAATTTAA	non-TIR
AtMULE619	XXIV	3	1067	12320042	1067	TTAATTTTT	TTAATTTTG	TIR
AtMULE620	XXIII	2	5926314	5937029	10716	TTTTTGAA	ATTTTTGAA	non-TIR

*Positions on the Arabidopsis pseudomolecules (Jan. 16th, 2000 version).

Supplementary Table 5.1B MULEs in the sequenced rice genome

name	group	chr	GI	position	size (bp)	TSD	
						L	R
OsMULE0101	I	6	5734616	151920-153340	1421	gttgggagag	gttgggagag
OsMULE0102	I	1	11244751	151118-152469	1352	gaaacagaca	gaaacagaca
OsMULE0103	I	4	10241657	64558-65958	1401	agatccgcac	agaatccccac
OsMULE0104	I	4	12140341	3277-4624	1348	taatgaagga	taatgaagga
OsMULE0105	I	1	12060505	134122-135515	1394	cgtcaaaaag	cgtcaaaaag
OsMULE0106	I	1	10697236	74892-76285	1394	ctttttgacg	ctttttgacg
OsMULE0107	I	1	6630680	20841-22210	1370	cgacaagtga	cgacaagtga
OsMULE0108	I	1	6815051	149466-150835	1370	cgacaagtga	cgacaagtga
OsMULE0109	I	6	7363267	117988-119378	1391	na	na
OsMULE0110	I	10	13384337	66123-67163	1041	aaggtgagtc	aatgtgagtc
OsMULE0111	I	1	8570076	79320-80598	1279	aaatctgct	aaatctgct
OsMULE0112	I	1	7340902	99993-101479	1487	gcacttaac	gcacttaac
OsMULE0113	I	1	9795252	27880-29246	1367	cctaaccgcc	cctaaccgcc
OsMULE0114	I	10	13184992	74795-75864	1070	na	na
OsMULE0115	I	1	8467930	23683-30464	6782	gcgagggtag	gcgagggtag
OsMULE0116	I	1	8096366	85443-92224	6782	gcgagggtag	gcgagggtag
OsMULE0117	I	1	13872907	17791-21899	4109	ccacacgag	ccacacgag
OsMULE0118	I	3	12039270	25203-32573	7371	ttgtagcagg	ttgtagcagg
OsMULE0119	I	6	11602826	20145-27495	7351	cgcgctacgc	cgcgctacgt
OsMULE0120	I	1	13366120	30443-39655	9213	agcaaaatca	agcaaaatca
OsMULE0121	I	10	7363410	96540-98954	2415	tgcctcgt	tgcctcgt
OsMULE0122	I	10	10122030	41962-45910	3949	gtgcgcctcc	gtgcgcctcc
OsMULE0123	I	10	12044845	84923-87387	2465	gatgctgga	gatgttgga
OsMULE0124	I	10	13185022	84923-87387	2465	gatgctgga	gatgttgga
OsMULE0125	I	4	5679837	16937-17337	401	ttcaccaga	ttcaccaga
OsMULE0126	I	10	12039314	47917-49297	1381	tctctttt	tctctttt
OsMULE0127	I	3	13957655	72236-73597	1362	tcactaac	tcactaac
OsMULE0201	II	6	8698574	103735-104695	961	tggatgag	tggatgag
OsMULE0301	III	1	10179050	117186-118501	1316	aatcaagat	aatcaagat
OsMULE0302	III	10	13184872	86717-88209	1493	tagtaagta	tagtaagta
OsMULE0303	III	1	13161438	117055-118190	1136	na	na
OsMULE0304	III	1	13122417	58858-59993	1136	na	na
OsMULE0305	III	1	8096366	141096-142562	1467	actaaacta	actaaacta
OsMULE0306	III	1	6630680	58323-59998	1676	na	na
OsMULE0307	III	3	12039270	93235-96949	3715	na	na
OsMULE0308	III	1	11191986	46089-48653	2565	ttttgataa	ttttgataa
OsMULE0309	III	1	13936418	11456-12828	1373	attctgatt	attctgatt
OsMULE0401	IV	1	13486822	52552-54235	1684	aaccgacaa	aaccgacaa
OsMULE0402	IV	?	5042437	20014-21697	1684	aaccgacaa	aaccgacaa
OsMULE0403	IV	?	5042437	163589-164904	1316	aatcaagat	aatcaagat
OsMULE0404	IV	6	8096397	104634-106183	1550	agaaggaat	agaaggaat
OsMULE0405	IV	1	12641875	108824-110411	1588	tattgccta	tattgccta
OsMULE0406	IV	1	13366120	41323-41910	588	atcataaaa	atcataaaa
OsMULE0407	IV	10	13677079	132291-133687	1397	gggactggc	gggactggc
OsMULE0408	IV	10	10122030	101279-102675	1397	gccagtccc	gccagtccc
OsMULE0409	IV	1	13620983	132105-133407	1303	tctctttt	tctctttt

OsMULE0410	IV	1	13620983	127535-129247	1713	acagtcc	tcagtcc
OsMULE0411	IV	10	10140721	8410-10241	1832	acggccccgt	acggccccgt
OsMULE0412	IV	1	6539576	77902-78779	878	tgccctgt	tgccctgt
OsMULE0413	IV	1	8570080	104267-104713	447	gctttttaa	gctttttaa
OsMULE0414	IV	3	12039270	94818-96173	1356	atcaagaaa	atcaagaaa
OsMULE0415	IV	10	10140721	8410-10241	1832	acggccccgt	acggccccgt
OsMULE0416	IV	10	13184072	35520-36946	1427	na	na
OsMULE0501	V	1	5042437	101744-102502	759	tcgccctg	tcgccctg
OsMULE0502	V	?	5042437	170433-171006	574	taaaaaatt	taaaaaatt
OsMULE0503	V	1	13486822	134274-135032	759	tcgccctg	tcgccctg
OsMULE0504	V	1	10179050	55341-56099	759	tcgccctg	tcgccctg
OsMULE0505	V	1	7242897	135903-142198	6296	tattgggt	tattgggt
OsMULE0506	V	1	6815051	57155-63450	6296	tattgggt	tattgggt
OsMULE0507	V	3	13957655	96987-99380	2394	ccagccaga	ccggccgga
OsMULE0508	V	10	12039362	65579-73299	7721	cagaaacag	cagaaacag
OsMULE0509	V	1	13442956	73489-76028	2540	accaatt	acctatt
OsMULE0510	V	1	11079211	19622-20720	1099	na	na
OsMULE0511	V	4	12140340	74102-77478	3377	ggaactaata	ggagctaaca
OsMULE0601	VI	?	5042437	170433-171006	574	taaaaaatt	taaaaaatt
OsMULE0602	VI	1	10179050	124030-124603	574	taaaaaatt	taaaaaatt
OsMULE0603	VI	1	7106505	243-816	574	taaaaaatt	taaaaaatt
OsMULE0604	VI	1	10934069	83141-83671	531	ttttttt	ttttttt
OsMULE0605	VI	1	9558536	17628-18221	594	taataaaaa	taataaaaa
OsMULE0701	VII	6	8698574	85970-87390	1421	cttagaaaa	cttagaaaa
OsMULE0702	VII	6	8698574	102615-104978	2364	na	na
OsMULE0703	VII	6	6069643	57823-59799	1977	tagaataga	tagaataga
OsMULE0704	VII	1	8096629	124595-126014	1420	taagcaagt	taagcaagt
OsMULE0705	VII	1	8096629	95575-97545	1971	ttttcataa	tttacaaaa
OsMULE0706	VII	1	8570080	83383-84612	1230	atgtatcaa	atgtatcaa
OsMULE0707	VII	6	11875148	87002-88401	1400	tttttatat	tttttatat
OsMULE0708	VII	1	11034662	136539-137827	1289	gcagaaaaa	gcagaaaaa
OsMULE0709	VII	3	12039270	50376-53235	2860	na	na
OsMULE0710	VII	3	6063530	103152-105271	2120	atgaaaaaa	atgaaaaaa
OsMULE0711	VII	3	12039270	112006-113409	1404	tctagaaag	tccagaaag
OsMULE0712	VII	6	8096305	87435-88796	1362	ttctcttgg	ttctcttgg
OsMULE0713	VII	6	8096305	125491-126852	1362	ttctcttgg	ttctcttgg
OsMULE0714	VII	1	11967924	107869-109549	1681	cttatcatt	cttatcatt
OsMULE0715	VII	1	11761068	46876-48400	1525	cgtaacactt	cgtaacgctt
OsMULE0716	VII	10	10140627	111151-112997	1847	ttttccagt	ttttccagt
OsMULE0717	VII	5	9755370	80807-82813	2007	atgtggg	atatcgg
OsMULE0718	VII	1	12382000	12935-14590	1656	tcaaccgga	taaatcgga
OsMULE0719	VII	10	13992707	137317-139675	2359	taagcacgc	taagcacgc
OsMULE0720	VII	1	7630233	108387-109951	1565	taatgcctt	taatgcctt
OsMULE0721	VII	1	7340902	15169-16733	1565	taatgcctt	taatgcctt
OsMULE0722	VII	1	10179051	102735-104483	1749	aactcccca	aactcccca
OsMULE0723	VII	1	7637397	40949-42224	1276	ttttctctt	ttttctctt
OsMULE0724	VII	1	10945235	9693-11067	1375	na	na
OsMULE0725	VII	1	10697187	99516-100890	1375	na	na
OsMULE0726	VII	1	13486738	152042-153966	1925	aaggattaa	aaggattaa

OsMULE0727	VII	1	13486690	119099-120371	1273	accacagaa	accacagaa
OsMULE0728	VII	1	13486738	126030-127778	1749	cagccaaa	cagccaaa
OsMULE0729	VII	10	13184072	29574-30964	1391	cccaaatac	cccaaatat
OsMULE0730	VII	1	12249128	119252-120591	1340	agggtcatt	agggtcatt
OsMULE0731	VII	1	13486858	11913-13768	1856	aactccgac	aactccgac
OsMULE0732	VII	1	11034690	119908-121763	1856	aactccgac	aactccgac
OsMULE0733	VII	1	12313694	48275-49959	1685	tttcttgag	cttcctgag
OsMULE0734	VII	10	8439785	56477-58024	1548	tcttatcta	tcttatcca
OsMULE0735	VII	3	13592189	61880-63763	1884	na	na
OsMULE0736	VII	6	7363267	95390-97432	2043	aagtcggac	atgtcggac
OsMULE0737	VII	1	10945235	43420-45230	1811	ttttgtttg	ttttgtttg
OsMULE0738	VII	1	10697187	133243-135053	1811	ttttgtttg	ttttgtttg
OsMULE0739	VII	1	11862978	12677-15035	2359	ctttggaa	ctttggaa
OsMULE0740	VII	1	9711848	54740-56022	1283	na	na
OsMULE0741	VII	1	11526599	109283-111350	2068	ttcggccta	ttcggccta
OsMULE0742	VII	10	13677105	119609-120873	1265	atcatgctcc	atcgtgctcc
OsMULE0743	VII	1	7340902	15169-16733	1565	taatgcctt	taatgcctt
OsMULE0744	VII	1	7630233	108387-109951	1565	taatgcctt	taatgcctt
OsMULE0746	VII	4	10241613	33304-34944	1641	aattctgacgg	tattctgacgg
OsMULE0801	VIII	6	8698574	96840-98855	2016	ttaatcgca	ttaatcgca
OsMULE0802	VIII	6	5091496	1031-3046	2016	ttaatcgca	ttaatcgca
OsMULE0803	VIII	1	13442957	135505-136953	1449	tattccttg	tattccttg
OsMULE0804	VIII	1	13442957	92943-94654	1712	tttaaatat	tttaaatat
OsMULE0805	VIII	10	10140660	81283-82519	1237	aaaaacta	aaaaacaa
OsMULE0806	VIII	10	10140660	41982-43807	1826	tttaaacaa	tttaaacaa
OsMULE0807	VIII	1	13365563	62937-72557	9621	ttgttttct	ttgttttcc
OsMULE0808	VIII	1	9795252	39100-40849	1750	ttttgcaaa	ttttgcaaa
OsMULE0809	VIII	1	13486711	133889-135578	1690	taatttttat	taatttttat
OsMULE0810	VIII	1	13486738	74604-76293	1690	taatttttat	taatttttat
OsMULE0811	VIII	1	13486711	1161-2852	1692	aaacttag	aaatag
OsMULE0812	VIII	10	10140627	123757-125057	1301	taacttggg	taagttggg
OsMULE0813	VIII	10	10140627	101511-103270	1760	tccacaaaa	tccacaaaa
OsMULE0814	VIII	10	10140627	122088-123693	1606	atcagaattg	atcagaattg
OsMULE0815	VIII	10	13357269	86022-87810	1789	na	na
OsMULE0816	VIII	1	7637397	138964-140764	1801	tatgtgtta	tatgtgtta
OsMULE0817	VIII	1	7637397	137200-138920	1721	ttgttcaat	ttttcaaat
OsMULE0818	VIII	1	7637397	164118-165723	1606	tgggtgaaa	tgggtgaaa
OsMULE0819	VIII	1	13359042	112058-113876	1819	cagaaaaaaaa	cagaaaaaaaa
OsMULE0820	VIII	10	12331454	87547-89080	1534	agggcggaa	aaggcggaa
OsMULE0821	VIII	1	10697187	136198-136414	217	cacaaatattt	cacaaattt
OsMULE0822	VIII	1	10945235	46375-46591	217	cacaaatattt	cacaaattt
OsMULE0823	VIII	10	13346556	68113-69760	1648	tttatctgg	tttatctgg
OsMULE0824	VIII	3	12583790	88697-101848	13152	aattagaaa	aattagaaa
OsMULE0825	VIII	6	11875148	103071-103417	347	ttcttggaa	ttcttggaa
OsMULE0826	VIII	6	11875148	85293-86930	1638	tggtacat	tggtacat
OsMULE0827	VIII	10	13185078	170301-171781	1481	tttaatact	tttaatccc
OsMULE0828	VIII	1	10800055	15447-17052	1606	tgggtgaaa	tgggtgaaa
OsMULE0829	VIII	1	9049451	51378-53211	1834	tttttttt	tttttttt
OsMULE0830	VIII	1	13366212	54100-55896	1797	na	na

OsMULE0831	VIII	1	10179052	18080-19876	1797	na	na
OsMULE0832	VIII	1	10179052	126388-133332	6945	gcaaatataa	gcaaatataa
OsMULE0833	VIII	1	11071975	39159-46103	6945	gcaaatataa	gcaaatataa
OsMULE0834	VIII	4	10241423	4639-6532	1894	tttttggaa	tttttggaa
OsMULE0835	VIII	10	13677079	130973-132274	1302	acccggaggt	acccagaggg
OsMULE0836	VIII	10	13677079	74482-79809	5328	tagttatag	tagttctag
OsMULE0837	VIII	10	10122030	102692-103993	1302	ccctctgggt	acctccgggt
OsMULE0838	VIII	1	13620983	135783-136151	369	na	na
OsMULE0839	VIII	1	13620983	126021-127441	1421	ccttgcgtt	ccttgcgtt
OsMULE0840	VIII	3	13384340	3893-6768	2876	tcaatctggt	tcaacctggt
OsMULE0841	VIII	1	11602829	83156-84783	1628	ttccgggaga	ttccgggaga
OsMULE0842	VIII	1	12641878	131273-132710	1438	tataagactt	tataagacat
OsMULE0843	VIII	1	11967924	57689-58990	1302	aactttcta	aactttcta
OsMULE0844	VIII	1	11967924	74010-75829	1820	aatagtaat	aatagtaat
OsMULE0845	VIII	1	10179051	143162-144200	1039	ttttttcc	ttttttcc
OsMULE0846	VIII	X	12061428	7056-9350	2295	aacccaata	aacccaata
OsMULE0847	VIII	1	9558510	30968-33877	2910	cacctacgt	cacctacgt
OsMULE0848	VIII	3	12656799	42434-44293	1860	acccactgc	acccactgc
OsMULE0849	VIII	1	11526599	64131-65665	1535	cctgggcta	cctgggcta
OsMULE0850	VIII	1	13442956	44923-46000	1078	caagcagaac	caaccagaac
OsMULE0851	VIII	1	7340852	109690-109986	297	aacaggagc	aacaggagc
OsMULE0852	VIII	1	11034690	66528-68731	2204	tttttttt	tttttttt
OsMULE0853	VIII	1	11034690	133621-135082	1462	gccgaaaa	ggccgaata
OsMULE0854	VIII	1	13486858	25626-27087	1462	agccgaaaa	ggccgaata
OsMULE0855	VIII	1	12328452	37111-38456	1346	tcgg	tcgg
OsMULE0856	VIII	1	9711848	22089-23740	1652	taaatgtgg	taaatgtgg
OsMULE0857	VIII	1	11761068	95004-96655	1652	taaatgtgg	taaatgtgg
OsMULE0858	VIII	1	13486765	101336-102957	1622	gtgagcacc	gtgagcacc
OsMULE0859	VIII	1	12382000	72439-74003	1565	aggaa	aggaa
OsMULE0860	VIII	10	10140779	14068-15657	1590	ttctgtttt	ttctgtttt
OsMULE0861	VIII	10	7363412	10301-11933	1633	tcccactaaa	tcccactaaa
OsMULE0862	VIII	6	9845048	21770-23316	1547	na	na
OsMULE0863	VIII	1	11136555	22932-24899	1968	actagaccct	actagaccct
OsMULE0864	VIII	6	6069643	33468-35187	1720	atcccattc	atcccattc
OsMULE0865	VIII	10	12280907	121913-123153	1241	tttcattt	tttcattt
OsMULE0866	VIII	1	7340902	63022-64819	1798	gctgcaagc	gctgcaagc
OsMULE0867	VIII	8	5257255	80366-82034	1669	gaaagaatt	gaaggaatt
OsMULE0868	VIII	1	6815051	1805-2719	915	na	na
OsMULE0869	VIII	1	8096366	63565-64479	915	na	na
OsMULE0870	VIII	10	13185093	107915-109672	1758	aaatatggg	aaatatggg
OsMULE0871	VIII	1	6721534	120178-121877	1700	aaaataaaa	aaaataaaa
OsMULE0872	VIII	1	6815077	13911-15610	1700	aaaataaaa	aaaataaaa
OsMULE0873	VIII	1	8468009	90027-91213	1187	na	na
OsMULE0874	VIII	4	10241657	34087-35194	1108	cttactaat	cttactaat
OsMULE0875	VIII	3	13249436	43889-45732	1844	aaaaggatt	aaaaggatt
OsMULE0876	VIII	10	11545729	130080-131387	1308	aaatataaat	aaatataaat
OsMULE0877	VIII	1	14021067	66279-68113	1835	gctgctattt	gctgctattt
OsMULE0878	VIII	4	5777612	30114-30815	702	cacttttta	cagttttta
OsMULE0879	VIII	1	14021066	20365-21968	1604	na	na

OsMULE0880	VIII	1	13936418	14329-16390	2062	na	na
OsMULE0881	VIII	3	13699786	25130-25965	836	na	na
OsMULE0882	VIII	1	7106505	36084-37422	1339	ttcggaggac	ttcggagtct
OsMULE0883	VIII	1	13699092	77046-78880	1835	tataagcaat	tataggcaat
OsMULE0901	IX	10	11527448	89887-90451	565	na	na
OsMULE0902	IX	10	13184944	73523-74081	559	tttaactaa	tttaactaa
OsMULE0903	IX	5	12025551	10439-10965	527	tttatttt	tttatttt
OsMULE0904	IX	5	12025551	29874-30106	233	atatggtag	atatggtag
OsMULE0905	IX	1	13620983	48868-49396	529	ttttttctc	ttttttctc
OsMULE0906	IX	1	11136555	76794-77315	522	aaaaat	aaaaat
OsMULE0907	IX	?	10140618	144137-144701	565	agattta	agatatttt
OsMULE0908	IX	10	10140753	59729-60127	399	taatttcaa	taatttcaa
OsMULE0909	IX	10	13491223	124433-124860	428	atataattt	atataattt
OsMULE0910	IX	10	12597872	7212-7639	428	atataattt	atataattt
OsMULE0911	IX	6	5803242	33118-33572	455	tttttttta	tttttttta
OsMULE0912	IX	1	10336600	32394-32821	428	aaaaa	aaaaa
OsMULE0913	IX	1	9558536	73657-74056	400	aataaataa	aataaataa
OsMULE0914	IX	1	15212126	141379-141807	429	aatatcaaa	aatatcaaa
OsMULE0915	IX	1	14021066	69389-69812	424	atttttata	attttttt
OsMULE0916	IX	1	6630680	104193-104529	337	tttttttct	tttttttagt
OsMULE0917	IX	10	13184927	59759-60163	405	aaaaattt	aaaaattt
OsMULE0918	IX	1	10179053	23797-24196	400	taaattata	taaattata
OsMULE0919	IX	3	12656799	119678-120056	379	taaatcaaa	taaaccaaa
OsMULE0920	IX	3	13346555	43305-43683	379	taaatcaaa	taaaccaaa
OsMULE0921	IX	1	13486660	82414-82803	390	tttttta	tttttc
OsMULE0922	IX	1	8570080	137103-137504	402	gatcaaata	gatcaaata
OsMULE0923	IX	1	8096629	61727-62064	338	tttatttt	tttatttt
OsMULE0924	IX	1	8096563	166479-166869	391	acttttaa	acttttaa
OsMULE0925	IX	6	5734616	31305-31705	401	ttatttt	ttatttt
OsMULE0926	IX	1	12641876	27568-28002	435	ttctttata	ttctttata
OsMULE0927	IX	6	11875148	59800-60361	562	aaaacatta	aaaacatta
OsMULE0928	IX	6	6006355	137822-138383	562	aaaacatta	aaaacatta
OsMULE0929	IX	1	12328485	13912-14315	404	aatatatta	aatatatta
OsMULE0930	IX	1	7339690	51883-52285	403	taagatttt	taagatttt
OsMULE0931	IX	4	10241614	53850-54275	426	aaaaaagta	agaaaaata
OsMULE0932	IX	5	12056976	36851-37125	275	ctaaat	ctaaat
OsMULE0933	IX	6	11602826	143387-143940	554	aaaaaata	aaaaaata
OsMULE0934	IX	10	13185022	144849-145278	430	tttaaggat	tttaatgat
OsMULE0935	IX	10	12044845	144849-145278	430	tttaaggat	tttaatgat
OsMULE0936	IX	1	9711819	10047-10369	323	ttttaattt	ttttaatta
OsMULE0937	IX	1	9558485	57018-57340	323	ttttaattt	ttttaatta
OsMULE0938	IX	3	13677104	21358-21922	565	agattta	agatattt
OsMULE0939	IX	?	10140618	144137-144701	565	agattta	agatattt
OsMULE0940	IX	1	8467930	7029-7410	382	atcatatga	atcttataa
OsMULE0941	IX	1	8096366	68789-69170	382	atcatatga	atcttataa
OsMULE0942	IX	3	13112227	2250-2616	367	aaaattta	aaagtttt
OsMULE0943	IX	1	6539576	146856-147339	484	aaaaaaata	aaaagaata
OsMULE0944	IX	1	6539551	51430-51913	484	aaaaaaata	aaaagaata
OsMULE0945	IX	10	12039362	35391-35673	283	aaatttata	aaatttata

OsMULE0946	IX	1	14020922	127353-127745	393	ttttattta	ttttattta
OsMULE0947	IX	1	6815051	77061-77454	394	ttaatttta	ttaatttta
OsMULE0948	IX	1	10945235	85046-85443	398	aaaagttaa	aacagttaa
OsMULE0949	IX	1	12328452	106452-106831	380	ttttttttaa	ttttttttaa
OsMULE0950	IX	1	13872872	1806-2185	380	ttttttttaa	ttttttttaa
OsMULE0951	IX	1	9049451	91714-92105	392	taaatgtta	taaacgtta
OsMULE0952	IX	5	6863078	164068-164459	392	aaaaaacaa	aaaaaacaa
OsMULE0953	IX	1	13486711	48007-48400	394	aaaaaaaaa	aaaaaaaaa
OsMULE0954	IX	1	9967269	86465-86823	359	tcttttt	tttttt
OsMULE1001	X	3	13384340	133474-134806	1333	tttatataa	tttagataa
OsMULE1002	X	10	13185093	118424-119635	1212	ttttatatt	ttttatttt
OsMULE1003	X	10	7363409	64276-65799	1524	tatttaaaaa	tatttaaaat
OsMULE1004	X	10	10140779	95918-97642	1725	ttaaacta	ttaaacta
OsMULE1005	X	1	13548708	76951-77752	802	tatatttaaa	tatatttaaa
OsMULE1006	X	1	13548708	106457-107534	1078	acctagtta	acctagtta
OsMULE1007	X	1	8096366	8078-9573	1496	ttatttta	ttatttta
OsMULE1008	X	1	12082352	18442-22567	4126	tattcaata	taatcaata
OsMULE1009	X	1	13365563	95868-97101	1234	na	na
OsMULE1010	X	1	6815077	3578-5038	1461	aaaaaaaaa	aaaaaaaaa
OsMULE1011	X	1	6721534	109845-111305	1461	aaaaaaaaa	aaaaaaaaa
OsMULE1012	X	2	8468047	126080-132330	6251	tacataatatttc	taataatatttc
OsMULE1013	X	1	11967925	52367-53613	1247	na	na
OsMULE1014	X	1	13430000	116716-117761	1046	na	na
OsMULE1015	X	1	13620985	121109-122450	1342	aaataaaaa	aaataaaaa
OsMULE1016	X	10	10140660	69608-70899	1292	ttttcttaa	ttttcttaa
OsMULE1017	X	1	13486738	123103-124392	1290	attctgtta	attctgtta
OsMULE1018	X	1	8468009	130463-131903	1441	ttttaattt	ttttaattt
OsMULE1019	X	10	13184902	94849-96109	1261	aatttaaa	aatttaaa
OsMULE1020	X	10	13184872	39935-41195	1261	aatttaaa	aatttaaa
OsMULE1021	X	10	10440613	44856-46142	1287	atataaaa	atataaaa
OsMULE1022	X	1	7630233	138442-139865	1424	aaaaaaaaaaa	aaaaaaaaaaa
OsMULE1023	X	1	7340902	45224-46647	1424	aaaaaaaaaaa	aaaaaaaaaaa
OsMULE1024	X	1	12082350	71627-73077	1451	tttataaaa	tttataaaa
OsMULE1025	X	6	10178254	2926-3953	1028	ttaaataaa	ttaaataaa
OsMULE1026	X	6	5091496	148956-149983	1028	ttaaataaa	ttaaataaa
OsMULE1027	X	1	8570079	100511-101263	753	tttattt	tttttt
OsMULE1028	X	10	12597733	41964-43138	1175	ttatgtaaa	ttatgtaaa
OsMULE1029	X	10	11545729	25845-26982	1138	tttgaaaa	tttgaaaa
OsMULE1030	X	10	13185022	89111-90364	1254	tacacaa	tacacaa
OsMULE1031	X	10	12044845	89111-90364	1254	tacacaa	tacacaa
OsMULE1032	X	1	7340902	45224-46647	1424	aaaaaaaaa	aaaaaaaaa
OsMULE1033	X	1	13620986	139221-139695	475	ataatgatt	ataatgatt
OsMULE1034	X	1	10257386	110897-112032	1136	ttttatttt	ttttatttt
OsMULE1035	X	1	7637397	85734-86704	971	tattttaa	tattttaa
OsMULE1036	X	3	13957655	94357-95773	1417	aaaactatt	aaaactatt
OsMULE1037	X	10	13185078	143003-146700	3698	attctaata	attctaata
OsMULE1101	XI	1	13810565	11293-11720	428	ttatttttta	tatttttta
OsMULE1102	XI	1	13442958	96769-97196	428	ttatttttta	tatttttta
OsMULE1103	XI	10	13185093	143490-143894	405	atttttt	tttttta

OsMULE1104	XI	X	10140611	44371-44830	460	aaaaaaata	taaaaaata
OsMULE11b01	XIB	3	12583790	40646-41339	694	ataaatt	ataaatt
OsMULE11b02	XIB	1	9988419	56929-57779	851	atccaca	atccaca
OsMULE11b03	XIB	10	13184902	41512-42565	1054	gttcgggag	gtccgggag
OsMULE11b04	XIB	5	9755370	1068-1860	793	gggaaag	tggaagt
OsMULE11b05	XIB	6	10178254	28423-34032	5610	gttaccacg	gttaccacg
OsMULE11b06	XIB	10	10716598	57466-58404	939	gtaaatatc	gtaaatatc
OsMULE11b07	XIB	1	6815077	122203-122918	716	gtgtgtgaa	gtgtgtgaa
OsMULE1201	XII	1	13161334	110579-111576	998	ataaggata	ataaggata
OsMULE1202	XII	1	9558510	31639-32731	1093	ttctttctt	ttctttctt
OsMULE1203	XII	10	10140660	85487-86633	1147	na	na
OsMULE1204	XII	1	14020922	90512-91698	1187	aagaaataa	aggaaataa
OsMULE1301	XIII	1	12641878	25869-30708	4840	tatatttta	tatatttta
OsMULE1302	XIII	1	13810564	94604-95504	901	ttctttttt	ttctttttt
OsMULE1303	XIII	1	11967924	86434-87416	983	ttcttcttt	tcttctttc
OsMULE1304	XIII	1	10697187	25172-26740	1569	ctattctat	ctattctat
OsMULE1305	XIII	10	13184927	53123-54516	1394	taaaaaata	taaaaaata
OsMULE1306	XIII	6	6907081	126385-127950	1566	tattt	tattt
OsMULE1307	XIII	10	13184992	64289-65664	1376	ttaaagt	ttaaagt
OsMULE13b01	XIIIB	6	7363267	93259-94013	755	tttctattg	tttctgctg
OsMULE13b02	XIIIB	1	7630233	40171-40913	743	ttttttctt	ttttttctt
OsMULE13b03	XIIIB	10	13185078	141465-142223	759	tatgcgatt	tatgcgatt
OsMULE13b04	XIIIB	10	13184902	104847-105446	600	na	na
OsMULE13b05	XIIIB	10	13184872	49864-67155	17292	taaataaaa	taaataaaa
OsMULE13b06	XIIIB	10	13491223	106808-107518	711	catatttaa	catatttaa
OsMULE13b07	XIIIB	1	11136555	91925-92514	590	na	na
OsMULE13b08	XIIIB	10	7363410	70957-71194	238	atttaa	attttt
OsMULE13b09	XIIIB	1	10179051	66044-66949	906	tcatttaa	tcatttaa
OsMULE13b10	XIIIB	10	13185038	2953-3991	1039	na	na
OsMULE13b11	XIIIB	10	7363409	127751-128789	1039	na	na
OsMULE13b12	XIIIB	1	10257386	28709-29485	777	atttaagat	atttaagat
OsMULE13b13	XIIIB	10	13184072	10750-11521	772	ttgtactgg	ttgtactgg
OsMULE13b14	XIIIB	1	11034662	91558-92174	617	na	na
OsMULE13b15	XIIIB	1	8468046	117238-117973	736	ttttaaa	ttttcag
OsMULE13b16	XIIIB	3	13249436	26186-26961	776	aaagaaata	aaagaaata
OsMULE13b17	XIIIB	1	8096563	156539-157299	761	tagttggaa	tacttggaa
OsMULE13b18	XIIIB	6	9711842	3010-3776	767	tataaacta	tataaacta
OsMULE13b19	XIIIB	1	12328562	20273-21022	750	tagtctt	atgtctt
OsMULE13b20	XIIIB	3	12039331	70438-71177	740	taaacaata	taaacaata
OsMULE13b21	XIIIB	1	10934069	99626-100394	769	na	na
OsMULE13b22	XIIIB	1	6016845	131523-132805	1283	aagaataag	aagaataag
OsMULE13d01	XIIID	10	8439785	71189-72051	863	aagaacata	aagaacata
OsMULE13d02	XIIID	1	12328562	100930-101792	863	tttttttta	tttttttta
OsMULE13d03	XIIID	1	9909165	131682-132545	864	aaagatcaa	aaagatcaa
OsMULE13d04	XIIID	1	12060505	73368-74224	857	aaaaatata	aaaaatata
OsMULE1401	XIV	1	12641875	84320-86218	1899	ctaataaaaa	ctaaaagaaa
OsMULE1402	XIV	10	10140660	118934-120190	1257	actcgaata	actcgaata
OsMULE1403	XIV	X	10140611	66373-68952	2580	tttttcaaa	tttttccaa
OsMULE1404	XIV	1	9558485	65861-67188	1328	tattatata	tattatata

OsMULE1405	XIV	1	9711819	18890-20217	1328	tattatata	tattatata
OsMULE1406	XIV	1	13359042	127510-128794	1285	acttgat	atcctgat
OsMULE1407	XIV	X	12656792	30426-31746	1321	ttgaataaa	ttgaataaa
OsMULE1408	XIV	10	13384337	44014-45128	1115	na	na
OsMULE1409	XIV	1	13365488	8724-9619	896	taaagttta	taaagttta
OsMULE1501	XV	1	6721501	86139-89558	3420	cctcggtct	cctcggtct
OsMULE1701	XVII	1	6721501	30557-31109	553	attctagaa	attctagaa
OsMULE1702	XVII	1	11244751	159043-159568	526	ttttgttaa	ttttgttaa
OsMULE1703	XVII	10	13357243	152115-152652	538	ttgag	ttgaag
OsMULE1704	XVII	10	13249425	135730-136264	535	aattataaa	aattataaa
OsMULE1705	XVII	X	9087173	31408-31945	538	taacaata	aaaaaata
OsMULE1706	XVII	2	5441876	26554-27091	538	ttaatataa	ttaatataa
OsMULE1707	XVII	1	13810564	41056-41578	523	acaaaattt	aaaaaattt
OsMULE1708	XVII	1	8570076	89887-90414	528	ttttttta	ttttttta
OsMULE1709	XVII	1	11526599	90233-90752	520	aaataatcg	aaataattg
OsMULE1710	XVII	1	12328452	12271-12784	514	ttttttt	ttttttt
OsMULE1711	XVII	1	13872872	107231-107666	436	na	na
OsMULE1712	XVII	10	10140660	14472-14967	496	tagataaaa	tatataaaa
OsMULE1713	XVII	1	13603463	5061-5592	532	ttaacataa	ttatcataa
OsMULE1714	XVII	10	13184902	12382-12965	584	taaaaataa	taaaaataa
OsMULE1715	XVII	?	10140618	149560-150067	508	tttagaaaa	tttagaaaa
OsMULE1716	XVII	3	13677104	26781-27288	508	tttagaaaa	tttagaaaa
OsMULE1717	XVII	1	14021066	67949-68454	506	ttatataaa	ttatataaa
OsMULE1718	XVII	10	13184072	38969-39468	500	taaaaataa	taaaaataa
OsMULE1719	XVII	4	10241423	59077-59595	519	aaaattcaaaa	aaaattcaata
OsMULE1720	XVII	10	7636307	70106-71693	1588	tttaatttt	tttaatttt
OsMULE1721	XVII	10	9186821	25649-27226	1578	ctatactta	ctatactta
OsMULE1722	XVII	3	12583790	45727-47309	1583	tatatttaa	tatatttaa
OsMULE1723	XVII	10	7243640	102140-103717	1578	ctatactta	ctatactta
OsMULE1724	XVII	10	13384337	41388-41865	478	atgataaaa	atgataaaa
OsMULE1725	XVII	1	7228436	112834-113369	536	aattaaata	aattaaata
OsMULE1726	XVII	1	7228436	85616-86088	473	tactttaaa	tactttaaa
OsMULE1727	XVII	3	6063530	116239-116745	507	tttcataag	tttcataaa
OsMULE1728	XVII	1	13620985	120290-120799	510	na	na
OsMULE1729	XVII	3	13592189	26479-26980	502	tataagtt	tcgtaagtt
OsMULE1730	XVII	10	12597733	56194-56711	518	tttcattgta	tttcattgta
OsMULE1731	XVII	10	14165313	140873-141407	535	tttataatt	tttataatt
OsMULE1801	XVIII	3	13677104	148741-153066	4326	na	na
OsMULE1901	XIX	3	13384340	148764-149803	1040	na	na
OsMULE1902	XIX	10	13605984	100085-101249	1165	ttgttttta	ttgttttta
OsMULE1904	XIX	10	12331454	45062-48469	3408	ttaaaataa	ttaaaataa
OsMULE1905	XIX	1	11034662	11497-12818	1322	tataatta	gataatta
OsMULE1906	XIX	1	10179051	67069-68146	1078	tatatttta	tatatttta
OsMULE1907	XIX	1	11136555	14828-15884	1057	aatttttct	ttttttct
OsMULE1908	XIX	1	9049451	26960-31989	5030	aaaattgc	aaatttgc
OsMULE1909	XIX	?	10140618	74115-75130	1016	ctgtccaattt	ctccaattt
OsMULE1910	XIX	X	10140611	66561-68056	1496	taaataaaa	aaaataaaa
OsMULE1911	XIX	X	10140611	65439-65667	229	na	na
OsMULE1912	XIX	1	13359042	102071-103327	1257	aatttttaa	aaatttgaa

OsMULE1913	XIX	1	13359042	7298-8772	1475	taaatatat	taaatatat
OsMULE1914	XIX	1	12382000	29841-30710	870	tttataata	tttataata
OsMULE1915	XIX	10	13384339	19543-20220	678	ttttacgaa	ttttacaa
OsMULE1916	XIX	10	13249425	116273-117244	972	aatttaata	aatttaata
OsMULE1917	XIX	1	13430000	39511-46368	6858	tttatatta	tttatatta
OsMULE1918	XIX	1	11244751	133848-143562	9715	atgtataa	agtgtataa
OsMULE1919	XIX	1	13486690	111007-112279	1273	ttttctaaa	ttttctaaa
OsMULE1920	XIX	1	13365995	7113-8382	1270	ttaaaaata	ttaaaaata
OsMULE1921	XIX	5	12056976	94946-95796	851	ttttttaa	tttaatattt
OsMULE1922	XIX	5	12056976	106071-108081	2011	ttaaaaaaaaaa	ttaaaaaaaaaa
OsMULE1923	XIX	10	12280907	20815-21698	884	ctacta	ctacta
OsMULE1924	XIX	1	12328562	73234-74352	1119	aa	aa
OsMULE1925	XIX	1	12082350	97964-99022	1059	tttttctt	tttcttctt
OsMULE1926	XIX	1	11034690	86446-86558	113	aaacaatgaa	aaaaaggaa
OsMULE1927	XIX	1	13366177	146082-147374	1293	aaaaagtta	aaaaagtta
OsMULE1928	XIX	1	13486797	44228-45503	1276	aaaaaaaaa	aaaaaataa
OsMULE1929	XIX	10	13899389	89715-91032	1318	ttatataaaa	ttatataaaa
OsMULE1930	XIX	1	8468009	41324-42424	1101	atactacta	atactacta
OsMULE19b01	XIXB	1	14149144	62136-63701	1566	tttagtata	tttagtata
OsMULE19b02	XIXB	10	12044845	47977-49229	1253	atatttata	atatttttta
OsMULE19b03	XIXB	10	13185022	47977-49229	1253	atatttata	atatttttta
OsMULE19b04	XIXB	1	14164523	121904-123184	1281	tttttaaa	tttttaaa
OsMULE19c01	XIXC	10	13185078	5002-6284	1283	na	na
OsMULE19c02	XIXC	1	13936418	97636-99017	1382	aatctataa	aacctataa
OsMULE19c03	XIXC	6	5295936	1528-3126	1599	taatgtgaa	taatgtgaa
OsMULE19c04	XIXC	1	7106505	133106-134341	1236	atgttctaa	atgttctat
OsMULE19c05	XIXC	1	6498456	3801-5036	1236	atgttctaa	atgttctat
OsMULE19c06	XIXC	1	13620987	135185-136318	1134	atttttatt	atttttgtt
OsMULE19c07	XIXC	1	8096629	17972-18954	983	aataaaaatt	aataaaaatt
OsMULE19c08	XIXC	1	9711791	3029-4453	1425	ttccaaaaa	ttgcaagaa
OsMULE19c09	XIXC	1	8096463	73120-74544	1425	ttccaaaaa	ttgcaagaa
OsMULE19c10	XIXC	10	13185038	5656-6373	718	na	na
OsMULE19c11	XIXC	10	7363409	130454-131307	854	attattatt	attattata
OsMULE19c12	XIXC	1	11761068	52506-53417	912	tttttttt	ttttttat
OsMULE19d01	XIXD	?	12039331	118093-119296	1204	ctacta	ctacta
OsMULE19e01	XIXE	4	12140340	13980-15211	1232	tttggaata	tttggaata
OsMULE19f01	XIXF	X	12061428	121002-122197	1196	tttttcatt	tttttcata
OsMULE19f02	XIXF	10	13185050	121311-122082	772	aagtaaatt	aagttaatt
OsMULE19f03	XIXF	1	13161334	29390-30499	1110	tacattaaa	tacattaaa
OsMULE19f04	XIXF	1	8570079	104848-105998	1151	tagcccaaa	tagcccaaa
OsMULE19f05	XIXF	1	13442958	100059-101211	1153	taggaatta	taggaatta
OsMULE19f06	XIXF	6	6907081	60212-61260	1049	na	na
OsMULE19f07	XIXF	1	12082350	103933-105117	1185	attaaccaa	attaaggaa
OsMULE19f08	XIXF	3	13699786	22948-24313	1366	tatg	tatg
OsMULE19f09	XIXF	6	8096305	50717-51935	1219	taaataaa	taaataaa
OsMULE19f10	XIXF	10	13184886	66301-67403	1103	na	na
OsMULE19f11	XIXF	10	7363409	49152-50121	970	ttataaaaa	ttataaaaa
OsMULE19f12	XIXF	6	9845048	29262-30778	1517	tcttgcaagac	tcttgcaaagac
OsMULE19f13	XIXF	10	7636307	76218-77270	1053	tttataata	tttataata

OsMULE19f14	XIXF	10	13184944	33068-37695	4628	gtctaaaaa	atctaaaaa
OsMULE19f15	XIXF	6	11862946	150191-151247	1057	tttaaatagt	tttaaatact
OsMULE2001	XX	1	9558455	29168-33640	4473	na	na
OsMULE2101	XXI	1	13486765	4405-5326	922	gtcgagggag	gtcgagggaag
OsMULE2102	XXI	4	12140295	13603-14797	1195	ctctcaa	ctctcaa
OsMULE2103	XXI	1	11034662	30842-31901	1060	ctctggaact	ccttgtaaca
OsMULE2201	XXII	X	10140611	49254-54403	5150	tttattt	tttttt
OsMULE2301	XXIII	X	10140611	27582-35564	7983	ggaaaagctt	ggaaaagctt
OsMULE2302	XXIII	1	13094247	109084-118236	9153	aaaaaatctg	aaaaaatctg
OsMULE2303	XXIII	1	8096629	25115-34266	9152	ttactacgc	ttactacgc
OsMULE2304	XXIII	3	14141756	61137-70263	9127	cggagagcg	cggagagcg
OsMULE2305	XXIII	10	10716598	91546-100581	9036	gagttccta	gagttccta
OsMULE2306	XXIII	1	13486690	31987-41106	9120	gagtttacag	gagtttacag
OsMULE2307	XXIII	10	10140660	73365-81240	7876	aaggaggag	aaggaggag
OsMULE2308	XXIII	1	13548708	139351-147296	7946	ttctctcctc	ttctcccctc
OsMULE2309	XXIII	1	12328562	57348-65403	8056	aaataaatc	aaataaatc
OsMULE2310	XXIII	1	9049451	83148-89868	6721	gtcctatatg	gtcctatatg
OsMULE2311	XXIII	10	13185093	30446-38346	7901	gcgatgattc	gcgatgattc
OsMULE2312	XXIII	1	8467930	31460-39498	8039	aattttgtga	aattttatga
OsMULE2313	XXIII	1	8096366	93220-101258	8039	aattttgtga	aattttatga
OsMULE2314	XXIII	1	8096366	40883-41545	663	cttctctag	cttctccag
OsMULE2315	XXIII	2	5441876	17584-25619	8036	gatggagtt	gatggagtt
OsMULE2316	XXIII	1	9558510	91099-97903	6805	ctattgctc	ctattgctc
OsMULE2317	XXIII	1	9558510	120744-125814	5071	gccatctata	gccatctata
OsMULE2318	XXIII	1	7228436	9670-16474	6805	ctattgctc	ctattgctc
OsMULE2319	XXIII	1	7228436	39315-44385	5071	gccatctata	gccatctata
OsMULE2320	XXIII	1	13486880	26735-34423	7689	gtcagtgctc	gtcagtgctc
OsMULE2321	XXIII	1	8468045	81269-88957	7689	gtcagtgctc	gtcagtgctc
OsMULE2322	XXIII	1	13486822	1755-5819	4065	caagagaag	caagagaag
OsMULE2323	XXIII	1	13365489	50251-53915	3665	gacgacaatg	gacgacaatg
OsMULE2324	XXIII	1	12313696	47419-51001	3583	gaatttgaag	gaatttgaag
OsMULE2325	XXIII	1	14090356	76948-95277	18330	aatatt	aatatt
OsMULE2326	XXIII	1	12328514	28518-46847	18330	catatgctcat	catatcctcct
OsMULE2327	XXIII	5	12025551	33323-38293	4971	gtttcaag	gtttcaag
OsMULE2328	XXIII	1	13486627	20661-22660	2000	taataggag	taataggag
OsMULE2329	XXIII	1	12641878	66367-69296	2930	aaaattagt	aaaattagt
OsMULE2330	XXIII	1	6721534	54764-58215	3452	gcatcttgga	gcatcttgga
OsMULE2331	XXIII	1	13359042	117957-118632	676	cgtgaataga	cgtgaataga
OsMULE2332	XXIII	1	14149144	73178-73861	684	gagctgtcaa	gagctgtcaa
OsMULE2333	XXIII	1	13620984	16209-16893	685	cgtttgcac	cgtttgcac
OsMULE2334	XXIII	1	12249128	147411-148075	665	gatttcaga	gattccaga
OsMULE2335	XXIII	6	5803242	140348-140952	605	ctctctccc	ctctctcac
OsMULE2336	XXIII	6	13548703	144068-144672	605	gtgagagag	gggagagag
OsMULE2337	XXIII	1	13366212	21528-22117	590	cgaaagcatc	tgaaagcatc
OsMULE2338	XXIII	1	11967907	100928-101595	668	cctgccaa	cctgccga
OsMULE2339	XXIII	1	13365973	112036-112660	625	gagaaatcc	gagaaatgc
OsMULE2340	XXIII	10	12597733	83076-83714	639	na	na
OsMULE2341	XXIII	3	12957695	124161-124772	612	tgcatatgct	tgcatatgct
OsMULE2342	XXIII	1	13161359	30044-30505	462	gtttggcg	gtttggtacg

OsMULE2343	XXIII	1	14090356	82085-92448	10364	na	na
OsMULE2344	XXIII	1	12328514	33655-44018	10364	na	na
OsMULE2401	XXIV	1	13603463	86811-87307	497	aaaaaaata	aaaaaaata
OsMULE2402	XXIV	3	13096049	9750-10234	485	attaatttt	attaatttt
OsMULE2403	XXIV	3	13957655	120482-120977	496	atttttt	atttttt
OsMULE2404	XXIV	10	13357243	22208-29820	7613	atattagaa	atattagaa
OsMULE2405	XXIV	1	13486627	4325-4803	479	aaaataaaa	aaaataaaa
OsMULE2406	XXIV	1	9795252	5368-5846	479	ttttatttt	ttttatttt
OsMULE2407	XXIV	1	12641878	109116-110655	1540	tttttatt	tttttatt
OsMULE2408	XXIV	1	11526599	10763-12542	1780	aaagtaaa	aaagtaaa
OsMULE2409	XXIV	10	13899391	139722-140223	502	tatat	taaatatat
OsMULE2410	XXIV	10	13185078	51845-53385	1541	tatatataaaa	tatttataaaa
OsMULE2411	XXIV	10	10716598	32208-32684	477	ttttt	taattt
OsMULE2412	XXIV	1	14020922	99880-101390	1511	ttttaacaa	ttttaacaa
OsMULE2413	XXIV	1	6498418	88669-90519	1851	tagtcaaaa	tagtcaaaa
OsMULE2414	XXIV	1	11138053	47396-49246	1851	tagtcaaaa	tagtcaaaa
OsMULE2415	XXIV	1	13873000	33185-35035	1851	tagtcaaaa	tagtcaaaa
OsMULE2416	XXIV	10	13384337	84500-85431	932	tttctgttt	tttctgttt
OsMULE2417	XXIV	1	12313696	130704-132869	2166	na	na
OsMULE2501	XXV	1	13366211	36148-44672	8525	cgcagtagca	aacagtagca
OsMULE2502	XXV	10	13185050	18617-23394	4778	na	na
OsMULE2503	XXV	1	13810566	15505-23587	8083	gtacttttta	gtacttttt
OsMULE2504	XXV	1	12641878	121619-127250	5632	ggggagtggg	ggggagtggg
OsMULE2505	XXV	1	13365489	69260-80575	11316	catat	gatat
OsMULE2506	XXV	10	13384339	71580-79447	7868	gtcccgtttc	gtcctgtttc
OsMULE2507	XXV	1	13620984	58503-65259	6757	gtgagaagag	gtgagaagag
OsMULE2508	XXV	5	11990635	133397-137594	4198	ctacatcgt	ctacatcgt
OsMULE2509	XXV	4	10241423	104411-107390	2980	cagattcatc	cagattcatc
OsMULE2510	XXV	10	8439785	59972-62028	2057	gaaggcgaac	gaaggcgaac
OsMULE2512	XXV	4	12140342	33330-33813	484	tatcttcgtg	tatcttcgtg
OsMULE2513	XXV	10	13184927	89454-90000	547	gagatttcag	gagatttcag
OsMULE2514	XXV	1	11862978	85561-104071	18511	gtgccatgtc	gtgccatgtc
OsMULE2515	XXV	5	12025551	99287-99614	328	cttcatcac	cttcatcac
OsMULE2516	XXV	1	9757660	33934-34253	320	taacgtctct	taacgtctct
OsMULE2517	XXV	1	8570044	9682-10001	320	taacgtctct	taacgtctct
OsMULE2518	XXV	1	12249128	122106-122583	478	cacatggct	catcatggct
OsMULE2519	XXV	10	10140737	31877-32212	336	gctgttact	gctgattacgt
OsMULE2520	XXV	1	7637397	63880-64208	329	cacattcggc	cacattcggc
OsMULE2521	XXV	X	10241688	16225-16549	325	na	na
OsMULE2901	XXIX	1	5922603	86581-88191	1611	tttaaaaaa	tttaaaaaa
OsMULE2902	XXIX	1	6016845	46779-48389	1611	tttaaaaaa	tttaaaaaa
OsMULE2903	XXIX	1	8468046	62308-62850	543	ttaaaaaaa	ttaaaaaaa
OsMULE2904	XXIX	1	7637397	71975-72512	538	tatttttt	tatttttt
OsMULE2905	XXIX	1	13161334	141618-142167	550	cattctaaga	cattctaaga
OsMULE2906	XXIX	1	10179053	99184-99711	528	tatttttt	tatttttt
OsMULE3101	XXXI	10	8439785	103485-108399	4915	na	na
OsMULE3102	XXXI	1	13603463	104119-109038	4920	na	na
OsMULE3103	XXXI	6	11862946	81827-82374	548	aaaaattag	aaaaattag
OsMULE3104	XXXI	1	9967269	87357-87909	553	tttacttta	tttaccta

OsMULE3105	XXXI	1	9711791	125822-126371	550	atatataaa	acatataaa
OsMULE3106	XXXI	1	13027337	11485-12034	550	atatataaa	acatataaa
OsMULE3107	XXXI	10	13184927	6787-7433	647	gaacgatat	caaggatat
OsMULE3108	XXXI	1	13161359	101794-102318	525	ttcttttaa	ttttttaa
OsMULE3109	XXXI	1	8096629	83188-83698	511	ttatataaaa	ttacataaaa
OsMULE3110	XXXI	10	7363410	136584-136978	395	tttaattaa	tttaattaa
OsMULE3111	XXXI	10	13491223	94781-95287	507	na	na
OsMULE3112	XXXI	3	13592189	96374-96912	539	ttaattatt	ttaattatt
OsMULE3201	XXXII	1	11862978	56381-57947	1567	tttattttc	tttattttc
OsMULE3202	XXXII	4	10241657	63000-64356	1357	tcctctttt	tcctctttt
OsMULE3203	XXXII	1	7339690	72841-74363	1523	acaaattcg	acaaagtcg
OsMULE3204	XXXII	1	7339690	74685-76028	1344	attcccatc	attcccatc
OsMULE3205	XXXII	1	7339690	143578-144797	1220	cctttttt	ttttttt
OsMULE3206	XXXII	10	13357243	117085-118654	1570	tttgcggtt	tttgcggtt
OsMULE3207	XXXII	1	13810566	90123-91574	1452	tttctttgt	tttctttgt
OsMULE3208	XXXII	1	13486627	58069-59290	1222	catatgggc	catatgggc
OsMULE3209	XXXII	1	8468009	52126-54054	1929	ttgttt	ttgtta
OsMULE3210	XXXII	10	13185050	137868-140234	2367	tttaaattt	tttaaattt
OsMULE3211	XXXII	6	14270108	68252-70312	2061	gcgtggaaa	gcgtggaaa
OsMULE3212	XXXII	6	14270108	66858-68170	1313	tttcatcta	tttcatctt
OsMULE3213	XXXII	1	13365488	83502-84792	1291	gaata	gaata
OsMULE3214	XXXII	1	12082351	130352-131799	1448	na	na
OsMULE3215	XXXII	3	14522999	103376-105079	1704	gagcgatcg	gagcgatcg
OsMULE3216	XXXII	6	11875148	110582-110964	383	tttggtttt	tttggtttt
OsMULE3217	XXXII	10	13899391	48002-49571	1570	taattctta	taattctta
OsMULE3218	XXXII	10	13184944	65624-72406	6783	tttcttttt	tttctttct
OsMULE3219	XXXII	1	14522861	100330-102010	1681	aaaagaaaa	aaaagaaaa
OsMULE3220	XXXII	1	11244751	113808-115780	1973	tataaattt	taaaaaatt
OsMULE3221	XXXII	1	9988419	78116-80019	1904	ccgcgcata	ccgcgcata
OsMULE3222	XXXII	X	10241688	45347-46835	1489	gtgcgggtg	gtgcgggtg
OsMULE3223	XXXII	1	12641876	20999-22730	1732	ttgttgacg	ttgttgacg
OsMULE3224	XXXII	1	14164523	28031-29269	1239	na	na
OsMULE3225	XXXII	1	10257386	114661-115899	1239	na	na
OsMULE3226	XXXII	10	10140627	18321-19896	1576	aatatgaac	aatatgaac
OsMULE3227	XXXII	1	8467981	58882-60050	1169	caaccaacat	cgaccaacat
OsMULE3228	XXXII	10	10140753	109191-110737	1547	tacgtagat	tacatagat
OsMULE3229	XXXII	10	14290377	67018-68068	1051	ttatgatga	ttaggagga
OsMULE3230	XXXII	1	9049451	1741-3171	1431	atttcaaga	atttcatga
OsMULE3231	XXXII	1	9711819	117337-118767	1431	atttcaaga	atttcatga
OsMULE3232	XXXII	1	9711819	83081-84894	1814	gccccaaat	gccccaaat
OsMULE3233	XXXII	10	10140737	79679-81444	1766	cacctatcat	cacctatcat
OsMULE3234	XXXII	10	12280907	81809-83552	1744	ttttttc	ttttttt
OsMULE3235	XXXII	10	10716598	49427-50632	1206	gaaatctaa	gaaatctaa
OsMULE3236	XXXII	1	11034690	5362-6581	1220	ttttt	ttttt
OsMULE3237	XXXII	?	6041757	77073-78241	1169	ttttccttt	ttttccttt
OsMULE3238	XXXII	1	14149142	23080-24248	1169	aaaggaaaa	aaaggaaaa
OsMULE3239	XXXII	1	13620983	115752-125707	9956	tttacaacatt	cttacaacatc
OsMULE3240	XXXII	1	13365489	48086-49468	1383	acacaaaaa	acacagaaa
OsMULE3241	XXXII	10	13384337	16161-17476	1316	aaaataggg	aaaataggg

OsMULE3242	XXXII	10	13677079	2386-3756	1371	tttcatttc	tttcatttc
OsMULE3243	XXXII	1	13620986	139753-141444	1692	ttttttatt	ttttttatt
OsMULE3244	XXXII	1	12328452	103353-105035	1683	na	na
OsMULE3245	XXXII	10	10122030	83004-84483	1480	aagcagtaa	aagcagtaa
OsMULE3246	XXXII	10	12039314	6350-7926	1577	tttcgattta	ttttgattta
OsMULE3247	XXXII	3	12039270	40244-43423	3180	ccttagaaa	ccttagaaa
OsMULE3248	XXXII	4	12140339	1271-2807	1537	aaaaagaaa	aaaaagaaa
OsMULE3249	XXXII	1	12641878	119777-121390	1614	ttcaaaata	ttcgaaata
OsMULE3250	XXXII	1	13161438	120696-126031	5336	ttaaccgaa	ttaaccgaa
OsMULE3251	XXXII	1	13122417	62499-67834	5336	ttaaccgaa	ttaaccgaa
OsMULE3252	XXXII	1	9558485	130052-131865	1814	gccccaaat	gccccaaat
OsMULE3253	XXXII	10	14290377	109591-112441	2851	aaaggggaaa	aaaggggaaa
OsMULE3254	XXXII	1	13486660	15589-17256	1668	tccaaatct	tccaaatct
OsMULE3255	XXXII	1	11967924	75924-77765	1842	tattccgtt	tattccgtt
OsMULE3256	XXXII	3	13096049	41102-45156	4055	gcgccgcta	atgccgcta
OsMULE3257	XXXII	1	13872907	68987-70455	1469	attaaaatc	tttaaaatg
OsMULE3258	XXXII	6	9845048	23319-25282	1964	tttttcctt	tttttcctt
OsMULE3259	XXXII	2	8468047	77682-79143	1462	tttcggtgg	tttcggtgg
OsMULE3261	XXXII	?	5042437	118411-119270	860	taaccaatt	taaataatt
OsMULE3262	XXXII	?	5042437	148748-149161	414	tattctctagg	tattttctaagg
OsMULE3263	XXXII	1	10179050	72008-72867	860	taaccaatt	taaataatt
OsMULE3264	XXXII	1	10179050	102344-102758	415	tattctctagg	tattttctaagg
OsMULE3265	XXXII	1	12328485	29334-31530	2197	ttcttgtga	ttcttcaga
OsMULE3266	XXXII	1	13620987	76087-78159	2073	tttaggtcg	tttaggttg
OsMULE3267	XXXII	1	7106505	58555-61421	2867	aattatttt	acaaatttt
OsMULE3268	XXXII	3	18640658	41552-42892	1341	tttttttta	tttttttta
OsMULE3301	XXXIII	1	13603463	120507-127599	7093	atactcctа	atactcctg
OsMULE3302	XXXIII	1	12060505	77532-85185	7654	cacatttac	cacatttac
OsMULE3601	XXXVI	1	13699092	15606-18272	2667	ttttcca	ttttcca
OsMULE3701	XXXVII	4	5777612	59940-61961	2022	tttcgtcct	tttcgtcct
OsMULE3702	XXXVII	6	8096397	76057-77803	1747	atttcaaac	atttcagac
OsMULE3703	XXXVII	6	14270108	153661-155621	1961	ttttaacac	ttttaacac
OsMULE3704	XXXVII	1	13603463	9433-11330	1898	gaaccctat	gaaccctat
OsMULE3705	XXXVII	10	14578149	116367-118192	1826	tttgggtta	tttgggtta
OsMULE3706	XXXVII	10	10140779	102757-104256	1500	aaaataaa	aaaataaa
OsMULE3707	XXXVII	4	10241423	78781-80469	1689	ccaatttgta	ccaatttgta
OsMULE3708	XXXVII	1	11071975	4492-5504	1013	aatataatgc	aatataatgc
OsMULE3709	XXXVII	1	10179052	91721-92733	1013	aatataatgc	aatataatgc
OsMULE3710	XXXVII	1	10179052	19943-21042	1100	na	na
OsMULE3711	XXXVII	1	14021067	68142-69524	1383	gccaagaaa	gccaggaaa
OsMULE3712	XXXVII	10	10716598	58673-65833	7161	ggacatata	ggacatata
OsMULE3713	XXXVII	1	13365563	130715-132508	1794	ttatattta	ttatagtta
OsMULE3714	XXXVII	1	13027337	109976-111390	1415	gcagtgaaa	gcagtgaaa
OsMULE3715	XXXVII	1	8468048	55848-57172	1325	cctgctgt	cctgctgt
OsMULE3716	XXXVII	3	13957655	6071-7465	1395	ccgtgctttt	ccgtgatttt
OsMULE3717	XXXVII	3	14141756	75109-76503	1395	ccgtgctttt	ccgtgatttt
OsMULE3718	XXXVII	1	11761068	20013-21217	1205	caatatttt	caatatttt
OsMULE3719	XXXVII	1	11136555	120740-122263	1524	caggaaata	caggaaata
OsMULE3720	XXXVII	1	13872907	122017-123558	1542	tcggcgctaa	tctgcgctaa

OsMULE3721	XXXVII	1	13366212	105823-107618	1796	atctcgtta	atctcgtta
OsMULE3722	XXXVII	1	13366212	55963-57062	1100	na	na
OsMULE3723	XXXVII	1	10179052	69803-71598	1796	atctcgtta	atctcgtta
OsMULE3724	XXXVII	1	13486822	36365-38106	1742	ttgtgaggg	ttatgaggg
OsMULE3725	XXXVII	?	5042437	3827-5568	1742	ttgtgaggg	ttatgaggg
OsMULE3726	XXXVII	10	10140660	87171-88503	1333	na	na
OsMULE3727	XXXVII	1	8468009	133069-134529	1461	tttgcatgc	ttcgcatgc
OsMULE3728	XXXVII	3	13112227	93619-94921	1303	atttcctat	gtttactat
OsMULE3729	XXXVII	?	13940635	50507-52056	1550	catagtgatt	catagtgatt
OsMULE3730	XXXVII	10	10440613	125889-127364	1476	ttccgttgc	ttccgttgc
OsMULE3731	XXXVII	1	11602829	10790-12140	1351	attctgatt	attattatg
OsMULE3732	XXXVII	10	11545729	168510-169614	1105	aaaatt	aaaatt
OsMULE3733	XXXVII	10	8439785	7626-8730	1105	aaaatt	aaaatt
OsMULE3734	XXXVII	10	13702835	17333-18320	988	acaaccaat	acatccaat
OsMULE3735	XXXVII	1	9558536	123118-124472	1355	ttgcgtgcac	tggcgtgcaa
OsMULE3736	XXXVII	10	14670086	99633-100875	1243	ttacgtaga	ttacgtgga
OsMULE3801	XXXVIII	1	5091597	76508-77079	572	ttctcttaaa	ttctcttaat
OsMULE3802	XXXVIII	1	11526599	21300-21853	554	aaaaataaa	aaaaataaa
OsMULE3803	XXXVIII	1	11138053	112138-112630	493	aaaaatcaa	aaaaaaaaa
OsMULE3804	XXXVIII	1	13873000	97927-98419	493	aaaaatcaa	aaaaaaaaa
OsMULE3805	XXXVIII	1	9229986	45202-45694	493	aaaaatcaa	aaaaaaaaa
OsMULE3806	XXXVIII	10	13185078	147969-148515	547	aaagaaaaa	aaagaaaaa
OsMULE3901	XXXIX	1	13486822	103324-104033	710	ttaatttaa	ttaatttaa
OsMULE3902	XXXIX	?	5042437	70786-71495	710	ttaatttaa	ttaatttaa
OsMULE3903	XXXIX	1	10179050	24391-25100	710	ttaatttaa	ttaatttaa
OsMULE3904	XXXIX	1	13366212	33886-36139	2254	aaggaatta	aaggaatta
OsMULE3905	XXXIX	1	13620985	15492-18863	3372	tagttcatt	tagttcata
OsMULE3906	XXXIX	3	12039331	25985-26910	926	ttaaaagaa	ttaaaagaa
OsMULE3907	XXXIX	1	10336600	37397-38546	1150	na	na
OsMULE4001	XL	?	5042437	74732-76102	1371	ttattttaa	ttattttaa
OsMULE4002	XL	1	13486822	107270-108640	1371	ttattttaa	ttattttaa
OsMULE4003	XL	1	10179050	28337-29707	1371	ttattttaa	ttattttaa
OsMULE4004	XL	1	12328562	120826-122199	1374	ttaaaaaac	ttaaaaaaa
OsMULE4005	XL	1	12328562	118174-118954	781	tttttagaa	tttttaagaa
OsMULE4006	XL	6	6907081	158419-159898	1480	gggatcac	gggatcac
OsMULE4007	XL	1	14495189	26946-32450	5505	tttaatata	tttaatata
OsMULE4008	XL	1	12082352	17624-23128	5505	tatattaaa	tatattaaa
OsMULE4009	XL	6	7363267	80136-81432	1297	tttttt	tttttt
OsMULE4010	XL	10	13185022	27630-30422	2793	ttgtttctct	ttctttctct
OsMULE4011	XL	10	13185022	18910-27600	8691	ttttgc	ttttgc
OsMULE4012	XL	10	12044845	27630-30422	2793	ttgtttctct	ttctttctct
OsMULE4013	XL	10	12044845	18910-27600	8691	ttttgc	ttttgc
OsMULE4014	XL	1	9558455	127995-129046	1052	tctatcttta	tctatcttta
OsMULE4015	XL	1	8570080	44810-45662	853	tgaaaaat	tgaaaaat
OsMULE4016	XL	1	8570080	33107-34158	1052	tctatcttta	tctatcttta
OsMULE4017	XL	X	12656792	72800-73719	920	gaaaaagg	gaaaaagg
OsMULE4018	XL	10	10140737	55047-58310	3264	cctcttcaaa	cctcttaaaa
OsMULE4019	XL	X	10140611	42257-43170	914	tataaag	tataag
OsMULE4020	XL	10	13184992	102255-103107	853	na	na

OsMULE4022	XL	4	12140299	32413-33472	1060	ccttggtttgta	cgttgtttgta
OsMULE4023	XL	4	12140340	16545-17862	1318	na	na
OsMULE4024	XL	1	13365491	82904-83835	932	tttgtct	tttctgt
OsMULE4025	XL	1	8096366	61036-61936	901	aattataat	aattataat
OsMULE4026	XL	6	5091496	124666-125641	976	catgataaa	catgataaa
OsMULE4027	XL	4	5852077	122120-123110	991	tttaatata	ttaatata
OsMULE4028	XL	4	5679837	36381-37370	990	tatattaaa	tatattaaa
OsMULE4029	XL	3	6063530	141897-142850	954	tttagcaaa	tttagcaaa
OsMULE4030	XL	6	13548703	14687-15697	1011	ttaattagg	ttaattagg
OsMULE4031	XL	X	9087173	115897-116846	950	tttttctat	tttttcaat
OsMULE4032	XL	1	6630680	123537-124268	732	ttcttataa	ttcttataa
OsMULE4033	XL	1	11244751	11137-11850	714	atttgaaat	atttgaaat
OsMULE4034	XL	1	13365995	135417-136313	897	attaaaaaaa	attaaaaaaa
OsMULE4035	XL	1	12382000	2312-3205	894	attccttgg	attccttgg
OsMULE4036	XL	1	13699092	21151-22067	917	tacttgctg	tatttgctg
OsMULE4037	XL	10	13184872	100267-101124	858	cctttttt	cctttttt
OsMULE4038	XL	10	10440613	55962-56926	965	ataatttttt	atatttttt
OsMULE4039	XL	?	10140618	116153-117040	888	ccctcttttcg	ccctcttttcg
OsMULE4040	XL	10	13185093	146193-147047	855	na	na
OsMULE4042	XL	?	5670155	35348-45741	10394	tttttcttc	tttttctcc
OsMULE4043	XL	4	4680488	27791-38426	10636	tttttcttc	tttttctcc
OsMULE4044	XL	10	4680335	19831-24763	4933	tttttcttc	tttttctcc
OsMULE4045	XL	1	12328485	103103-104599	1497	ctatttgct	ctatttgct
OsMULE4046	XL	1	12082351	83103-83813	711	agcaaatag	agcaattag
OsMULE4047	XL	10	7363409	118463-119266	804	na	na
OsMULE4048	XL	10	11527448	50688-51972	1285	ttaaaaa	ttaaaaa
OsMULE4049	XL	3	13384340	72100-72945	846	atttggttg	atttggttg
OsMULE4050	XL	3	13384340	12776-13786	1011	attttctaa	ttttctaa
OsMULE4051	XL	1	14021067	8135-8947	813	tagttgaaa	tagttgaaa
OsMULE4052	XL	1	13810565	66379-67192	814	aagaaaaag	aagaaaaag
OsMULE4053	XL	1	11071975	9771-13672	3902	aaaaaaaaag	aagaaaaaag
OsMULE4054	XL	1	10179052	97000-100901	3902	aaaaaaaaag	aagaaaaaag
OsMULE4055	XL	1	13486711	119861-120780	920	atgatgaaa	atgaagaaa
OsMULE4056	XL	1	13486738	60576-61495	920	atgatgaaa	atgaagaaa
OsMULE4057	XL	10	13184944	26015-26598	584	aaaatttt	agaatttt
OsMULE4058	XL	10	9087173	115897-116846	950	tttttcaat	tttttcaat
OsMULE4059	XL	10	18056686	104518-105055	538	na	na
OsMULE4101	XLI	1	13365563	64311-68475	4165	tctcctc	tctcctc
OsMULE4301	XLIII	6	9711842	4741-5845	1105	atataataa	atataataa
OsMULE4302	XLIII	3	12039270	51991-52962	972	tttttatta	tttttatta
OsMULE4401	XLIV	6	9711842	36206-36559	354	ggtgcttat	ggtgcttat
OsMULE4402	XLIV	1	8096366	143830-144207	378	atttattag	atttattag
OsMULE4403	XLIV	1	8467930	82070-82447	378	atttattag	atttattag
OsMULE4404	XLIV	6	9229995	14209-14643	435	atatcaat	atatcaat
OsMULE4405	XLIV	6	7363267	4054-4488	435	atatcaat	atatcaat
OsMULE4406	XLIV	1	6815051	101835-102259	425	gtgcttttg	gtgcttttg
OsMULE4407	XLIV	1	6815051	37778-38202	425	ttaaaacat	ttaaaacat
OsMULE4408	XLIV	1	13620983	84992-85407	416	tttataaaa	tttataaaa
OsMULE4409	XLIV	?	13677079	27364-27762	399	tgaaaca	tgaaaca

OsMULE4410	XLIV	?	13677079	86230-86653	424	tgcgatat	tgcgatat
OsMULE4411	XLIV	3	13384340	30626-31004	379	tttaagtgt	tttaagtgt
OsMULE4412	XLIV	1	13365498	55667-56067	401	ctatcttt	ctatcttt
OsMULE4413	XLIV	10	13129494	41909-42305	397	tttatatgt	tttatatgt
OsMULE4414	XLIV	3	13899390	4196-4619	424	tatcacat	tatcacat
OsMULE4415	XLIV	1	8096463	131828-132380	553	taaactaa	taaactaa
OsMULE4416	XLIV	1	9711791	61737-62289	553	taaactaa	taaactaa
OsMULE4417	XLIV	1	11136555	53361-53753	393	gagaaaaac	gagaaaaac
OsMULE4418	XLIV	3	12039270	53553-53946	394	ttcgtatt	tcgagtt
OsMULE4419	XLIV	10	13491223	83442-83865	424	gaatttat	gaatttat
OsMULE4420	XLIV	1	9049451	91456-92242	787	taaccttgt	taaccttgt
OsMULE4421	XLIV	3	13957655	37532-37953	422	ttaatttggc	ttaatttggc
OsMULE4422	XLIV	10	9186821	46409-46797	389	gaaaaaaat	gaaaataat
OsMULE4423	XLIV	10	7243640	122900-123288	389	gaaaaaaat	gaaaataat
OsMULE4424	XLIV	1	12082350	17393-17793	401	aatcgata	aatcgata
OsMULE4425	XLIV	1	13365488	11649-11876	228	acaaacaat	acaaacaat
OsMULE4426	XLIV	1	13620987	75530-75968	439	aagcaattactgca	aagcaattacagca
OsMULE4427	XLIV	1	13429999	43918-44340	423	atataatta	atataatta
OsMULE4428	XLIV	3	13249436	71813-72917	1105	na	na
OsMULE4429	XLIV	1	13442956	65042-65553	512	tttcggatc	tttcggatc
OsMULE4430	XLIV	1	13366177	80709-81131	423	aaacacaaa	aactaacaa
OsMULE4431	XLIV	3	6063530	140821-141230	410	aatgcttt	aatgcttt
OsMULE4432	XLIV	10	10140693	122325-127521	5197	ttctac	ttctac
OsMULE4433	XLIV	10	10140693	94812-95204	393	tatttttgttc	tgttctgttc
OsMULE4434	XLIV	10	13899389	100548-100979	432	na	na
OsMULE4435	XLIV	3	12583790	31204-31632	429	ttctagatc	ttctagatc
OsMULE4436	XLIV	1	13486711	111611-112055	445	na	na
OsMULE4437	XLIV	1	13486711	92220-92641	422	na	na
OsMULE4438	XLIV	1	13486738	52326-52770	445	na	na
OsMULE4439	XLIV	1	13486738	32935-33356	422	na	na
OsMULE4440	XLIV	1	7242897	116526-116950	425	ttaaaacat	ttaaaacat
OsMULE4441	XLIV	1	11034662	119327-119749	423	tattgcgg	tattgcgg
OsMULE4442	XLIV	1	13366120	45317-45739	423	ttattaat	ttattaat
OsMULE4443	XLIV	1	13442958	37754-38153	400	ctcgggttt	ctcgggttt
OsMULE4444	XLIV	1	12328452	21175-21597	423	na	na
OsMULE4445	XLIV	1	11034690	85959-86381	423	tatggacca	tatggacca
OsMULE4446	XLIV	3	12039331	5831-6253	423	gtcgtttat	gtcgtttat
OsMULE4447	XLIV	1	13366211	31363-32832	1470	tttggtagg	atcggtagg
OsMULE4448	XLIV	3	12656799	95271-95626	356	tgattgt	tgattgg
OsMULE4449	XLIV	3	13346555	18898-19253	356	tgattgt	tgattgg
OsMULE4450	XLIV	10	13384337	116418-116799	382	ttattttt	ttattttt
OsMULE4451	XLIV	10	13184886	88525-88864	340	aaataatt	aaacaatt
OsMULE4452	XLIV	10	12039314	87605-88930	1326	na	na
OsMULE4453	XLIV	1	7340902	122503-122847	345	acaacctaa	acaacctaa
OsMULE4501	XLV	1	8467930	118411-119395	985	gggctggag	gggctggag
OsMULE4701	XLVII	1	6721501	25156-25838	683	gacatataatattacta	gacatataatattacta
OsMULE4702	XLVII	3	13957655	52154-52843	690	gttactt	tttaatt
OsMULE4703	XLVII	10	13184992	69985-70641	657	aagaaaatt	aagaaaatt
OsMULE4704	XLVII	4	10241613	2531-3216	686	ttaaatctt	ttaaaactt

OsMULE4705	XLVII	4	5852077	38702-39387	686	aagttttaa	aagatttaa
OsMULE4706	XLVII	10	13184072	119832-120510	679	aagattcaa	aatattcaa
OsMULE4707	XLVII	10	9186821	98017-98691	675	cacaacaat	aaaaataat
OsMULE4708	XLVII	10	7363409	30562-31236	675	cacaacaat	aaaaataat
OsMULE4709	XLVII	4	10241657	140108-140775	668	tataattaa	tataattaa
OsMULE4710	XLVII	1	13027337	54485-55163	679	tttattaaa	tttattaaa
OsMULE4711	XLVII	1	9711848	93878-94559	682	aatataataatt	aatataatttat
OsMULE4712	XLVII	1	9558536	122165-122832	668	taatttatt	taatttatt
OsMULE4713	XLVII	1	11526599	38440-39890	1451	aaggttgaa	aaagttgaa
OsMULE4714	XLVII	1	6721534	99964-100695	732	ttgcaattg	ttgcagttg
OsMULE4715	XLVII	3	13699786	106963-107623	661	na	na
OsMULE4716	XLVII	3	13699786	65476-66085	610	na	na
OsMULE4717	XLVII	1	13486690	109901-110450	550	tttagctat	tttactat
OsMULE4801	XLVIII	1	11526599	38975-39714	740	na	na
OsMULE4802	XLVIII	1	11526599	151954-152696	743	gttttttta	tttttttta
OsMULE4803	XLVIII	1	11526599	90233-90752	520	aaataatcg	aaataattg
OsMULE4804	XLVIII	1	12328452	56802-57543	742	taaccaaaa	taactaaaa
OsMULE4805	XLVIII	1	13365995	128468-129208	741	tataaattt	tataaattt
OsMULE4806	XLVIII	3	12656799	50287-51030	744	ctaaaagaa	ctaaaagaa
OsMULE4807	XLVIII	3	13346555	89448-90188	741	ataataata	ataataata
OsMULE4808	XLVIII	3	13957655	95870-96597	728	na	na
OsMULE4809	XLVIII	3	12039270	42461-43204	744	tttcgtaaa	tttcgtaaa
OsMULE4810	XLVIII	3	13096049	75807-76548	742	aaaaaaaaaa	aaaaaaaaaa
OsMULE4811	XLVIII	1	8570077	32846-33579	734	tttaatttt	tttagtttt
OsMULE4812	XLVIII	10	12039314	20114-20857	744	ttaaaagtt	ttaaaagtt
OsMULE4813	XLVIII	10	13357269	25900-26633	734	atatagaaa	atctagaaa
OsMULE4814	XLVIII	4	10241646	19916-20648	733	tttgaaaaa	tttgaaaaa
OsMULE4815	XLVIII	1	10934069	119783-120482	700	na	na
OsMULE4816	XLVIII	1	9795252	34658-35422	765	caaataaaa	caaataaaa
OsMULE4817	XLVIII	1	6630680	42428-43147	720	tatatattca	tttatattga
OsMULE4818	XLVIII	6	11875148	107036-107754	719	aaaaaaaaaa	aaaaaaagaa
OsMULE4819	XLVIII	3	13699786	103457-104046	590	taaactatt	taaactatt
OsMULE4901	XLIX	3	13957655	99631-99980	350	taattaagc	taattaagc
OsMULE4902	XLIX	1	13486627	70784-71128	345	cgacacaaa	cgacacaaa
OsMULE4903	XLIX	1	12082351	143486-143830	345	aagaaaaca	aagaaaaca
OsMULE4904	XLIX	1	13810566	144218-144567	350	tgtttg	tgtttg
OsMULE4905	XLIX	10	13184992	87676-88027	352	atagaatac	atagaatac
OsMULE4906	XLIX	1	13442958	50411-50755	345	caaaatttt	caaaatttt
OsMULE4907	XLIX	1	8468049	47940-49756	1817	na	na
OsMULE4908	XLIX	1	8467930	83468-83791	324	ctgaaacgg	ctgaaacgg
OsMULE4909	XLIX	1	8096366	145228-145551	324	ctgaaacgg	ctgaaacgg
OsMULE4910	XLIX	10	12039314	127032-127377	346	gagggaggg	gagggaggg
OsMULE4911	XLIX	10	10140753	23194-23539	346	gagggaggg	gagggaggg
OsMULE4912	XLIX	1	11761068	138907-139249	343	ttaa	ttaa
OsMULE4913	XLIX	1	9711848	65992-66334	343	ttaa	ttaa
OsMULE4914	XLIX	4	10241646	57564-57908	345	taatattca	taatattca
OsMULE4915	XLIX	1	13365973	99128-99435	308	cagctagcg	cagctagcg
OsMULE4916	XLIX	1	11071975	131792-132135	344	ccgtactat	ccgtactat
OsMULE4917	XLIX	3	12039270	89819-90160	342	acctatgtg	acatatggg

OsMULE4918	XLIX	1	11967907	84732-86385	1654	aagtttgca	aagtttgca
OsMULE4919	XLIX	1	13486880	149632-150237	606	na	na
OsMULE5101	LI	10	13384339	42203-43116	914	gatgttaca	gatgttaca
OsMULE5102	LI	1	6630680	75391-76288	898	ccgatggaa	ccaatggaa

Supplementary Table 5.2 Representatives of MULE acquisition of host DNA in *Arabidopsis* and rice

MULE (size bp)	Captured DNA			MULE-contained ORF	
	description	organization	size (bp)	ID	expression
AtMULE382 (15633)	Pectin methylesterase gene	entire gene	1965	22328304	+ (RT-PCR)
AtMULE205 (17915)	ubiquitin-like (Ubl) specific protease	entire gene	8175	18396116	+ (RT-PCR)
AtMULE170 (3193)	a. intergenic region b. 2 expressed unknown genes c. 2-oxoglutarate dehydrogenase E2 subunit d. cytochrome c oxidase subunit	naa 5'UTRs & exon exon 5'UTR	933 448 &146 90 96	18397275	na
AtMULE136 (5307)	protein phosphatase 2C	exon &intron	9047	18397624	na
OsMULE0801 (2016)	DRE/CRT binding factor	entire gene	904	12581491	na
OsMULE0827 (1481)	N-type calcium channel alpha-1B/cdB3	entire gene	704	na‡	+(5038870ʃ)
OsMULE1028 (1175)	amino acid transporter	entire gene	650	na‡	+(16579242ʃ)
OsMULE3235 (1206)	mitochondrial alternative oxidase 1c	exon & intron	549	16902305	na
OsMULE0714 (1681)	S-adenosyl-L-methionine synthetase	exon	584	1778820	na

† GI: Gene Identifier.
‡predicted by the Rice Genome Automated Annotation System (http://RiceGAAS.dna.affrc.go.jp/).
ʃEST GI.
na: not available.

Supplementary Table 5.3A Acquistion of host DNA sequences by *Arabidopsis* MULEs

MULE	host gene	blastN/X	clone GI*/ORF ID†	position of acquired segment within the clone‡ start
AtMULE013	homeobox-leucine zipper protein HAT5/Athb-1 mRNA	N	na/16327	328
AtMULE015	hypothetical protein	X	4836903/At2g29240	1
AtMULE018	intergenic region	N	6056182/na	1152
AtMULE019	hypothetical protein	X	9369387/At3g47320	1
AtMULE020	hypothetical protein	X	12320878/T10D10.10	14
AtMULE020	hypothetical protein	X	12320878/F28L5.2	1
AtMULE020	hypothetical protein	X	12320878/F28L5.16	1
AtMULE023	putative protein phosphatase 2C	N	20197227/At2g30020	26298
AtMULE027	intergenic region	N	7267319/na	8528
AtMULE028	unknown protein	X	12322371/At1g33450	1
AtMULE029	hypothetical protein	X	12322381/F10C21.12	1
AtMULE030	intergenic region	N	2853071/na	40603
AtMULE030	intergenic region	N	2853071/na	41403
AtMULE032	t5r gene	N	na/1632775	542
AtMULE034	hypothetical protein	X	12320970/T32G9.37	37
AtMULE034	hypothetical protein	X	12320970/T32G9.33	130
AtMULE035	intergenic region	N	20197580/na	86636
AtMULE037	intergenic region	N	20197898/na	7197
AtMULE040	cytochrome P450-like protein	X	8778333/F18O14.38	1
AtMULE040	hypothetical protein	X	8778333/F18O14.37	1
AtMULE041	intergenic region	N	3402747/na	41388
AtMULE044	intergenic region	N	4757398/na	34651
AtMULE054	hypothetical protein	X	6728952/T12C22.12	124
AtMULE054	hypothetical protein	X	6728952/At1g44880	1
AtMULE055	hypothetical protein	X	7767661/F27F5.16	1
AtMULE055	putative replication protein A1	X	20197475/At2g12140	1
AtMULE059	intergenic region	N	2853071/na	40602
AtMULE059	intergenic region	N	3449318/na	33305
AtMULE060	hypothetical protein	X	8778507/F21D18.3	1

		position of acquired segment within the MULE			feature of the acquired segment
end	size[§]	start	end	size (bp)	
526	199	281	647	367	exon & intron
992	992	614	5049	4436	ORF
1300	149	230	378	149	na
142	142	4193	4697	505	ORF
211	198	8401	8994	594	exon
1198	1198	567	5442	4876	ORF
238	238	9595	10540	946	ORF
26520	223	511	733	223	exon
8710	183	927	1116	190	na
160	160	367	916	550	ORF
160	160	367	916	550	ORF
40854	252	165	415	251	na
41504	102	739	840	102	na
767	226	5529	5751	223	5' UTR
442	406	6685	8500	1816	exon & intron
1311	1182	15129	20625	5497	exon & intron
86755	120	170	291	122	na
7678	482	446	929	484	na
1887	1887	8559	15167	6609	ORF
724	724	15745	19249	3505	ORF
41961	574	193	770	578	na
35476	826	255	1082	828	na
337	214	4407	5048	642	exon & intron
1075	1075	11553	17260	5708	ORF
1745	1745	5514	17498	11985	ORF
228	228	4172	5114	943	ORF
40853	252	164	414	251	na
33511	207	435	646	212	na
338	338	245	1258	1014	exon
99	99	458	754	297	ORF
1131	1073	682	1761	1080	na
16169	536	658	1194	537	na

AtMULE063	hypothetical protein	X	6693374/At1g49680	1
AtMULE064	intergenic region	N	12323691/na	59
AtMULE068	intergenic region	N	20198237/na	15634
AtMULE070	hypothetical protein	X	4185120/At5g35000	1
AtMULE070	hypothetical protein	X	4185120/At5g36860	1
AtMULE072	putative replication protein A1	N	20197574/At2g11170	32213
AtMULE072	putative replication protein A1	N	20197574/At2g11170	31743
AtMULE072	putative replication protein A1	N	20197574/At2g11170	32321
AtMULE077	intergenic region	N	7635467/na	87630
AtMULE078	similar to Homo sapiens TAK1 binding protein	N	5002514/AT4g11860	50011
AtMULE079	intergenic region	N	3402747/na	41347
AtMULE081	homeobox gene Athb-1 mRNA	N	na/16327	328
AtMULE083	intergenic region	N	4757398/na	34651
AtMULE095	intergenic region	N	5041974/na	11317
AtMULE101	hypothetical protein	X	7549621/At2g16820	1
AtMULE127	intergenic region	N	8347605/na	1284
AtMULE127	intergenic region	N	8347605/na	1864
AtMULE127	intergenic region	N	8347605/na	2313
AtMULE136	hypothetical protein	X	20197260/At2g14770	1
AtMULE136	putative replication protein A1	X	20197260/At2g14780	1
AtMULE138	putative cyclin mRNA	N	12325394/At1g47210	29089
AtMULE139	hypothetical protein	X	20198197/At2g04960	17
AtMULE142	hypothetical protein	X	6451837/F21A17.12	6
AtMULE143	intergenic region	N	14475930/na	80072
AtMULE143	unknown protein	X	20197898/At2g03570	1
AtMULE153	hypothetical protein	X	20197405/At2g06860	1
AtMULE155	adenylate kinase -like protein	X	na/At5g34895	418
AtMULE158	hypothetical protein	X	20198265/At2g07505	1
AtMULE158	putative lumen protein retaining receptor	X	15218700/At1g34610	1
AtMULE159	hypothetical protein	X	15219471/At1g44880	1
AtMULE159	hypothetical protein	X	12320950/At1g35100	1
AtMULE159	putative replication protein A1	X	20198226/At2g10370	1
AtMULE167	hypothetical protein	X	20198242/At2g12700	1
AtMULE170	intergenic region	N	20197600/na	2727
AtMULE170	intergenic region	N	12324708/na	41379

119	119	3168	3524	357	ORF
1204	1204	3893	8804	4912	ORF
34482	2270	726	3077	2352	exon & intron & intergenic region
32014	272	3077	3347	271	exon & intron
32480	160	3360	3515	156	exon & intron
87772	143	503	646	144	na
50144	134	930	1063	134	exon
41994	648	528	1159	632	na
526	199	370	740	371	5' UTR & exon
35257	607	768	1377	610	na
11749	433	337	759	423	na
245	245	972	1842	871	ORF
1470	187	1290	1474	185	na
2219	356	1285	1642	358	na
2812	500	619	1150	532	na
1756	1756	501	9545	9045	ORF
235	235	10111	11056	946	ORF
29281	193	429	621	193	exon
291	275	3685	4998	1314	exon & intron
149	144	1015	1562	548	exon & intron
80365	294	1104	1397	294	na
59	59	420	755	336	exon
938	938	479	6565	6087	ORF
490	73	5002	5319	318	exon
143	143	4208	4704	497	ORF
995	995	5065	8803	3739	ORF
1038	1038	2199	7419	5221	exon & intron
220	220	13459	14210	752	ORF
267	267	10150	11161	1012	ORF
151	151	21368	21917	550	ORF
3003	277	2580	2935	356	na
41570	192	820	1018	199	na
6629	456	1806	2261	456	exon & intron
161	161	60684	61169	486	ORF
833	833	55218	59923	4706	ORF

AtMULE170	putative protein	N	5262205/F21C20.40	6174
AtMULE176	hypothetical protein	X	20197835/At2g14020	1
AtMULE176	hypothetical protein	X	20197835/At2g14010	1
AtMULE176	intergenic region	N	5823567/na	57253
AtMULE177	hypothetical protein	X	20198187/At2g14130	1
AtMULE178	putative protein phosphatase 2C	N	6671816/6714273	45298
AtMULE180	unknown protein	X	5672589/At2g14140	1
AtMULE182	hypothetical protein	X	22325701/At2g15815	1
AtMULE184	intergenic region	N	3293583/na	11016
AtMULE186	intergenic region	N	13677103/na	43787
AtMULE186	intergenic region	N	13677103/na	46415
AtMULE186	putative replication protein A1	X	22326553/At2g16330	1
AtMULE188	intergenic region	N	7635467/na	87630
AtMULE189	intergenic region	N	3402747/na	41343
AtMULE215	chloroplast RNA-binding protein cp31	N	na/681905	138
AtMULE215	DNA methyltransferase pseudogene	N	na/6017900	831
AtMULE224	intergenic region	N	3449329/na	75533
AtMULE226	ATP-dependent RNA helicase	N	8778333/F15O4.40	129058
AtMULE226	intergenic region	N	20197792/na	72668
AtMULE226	intergenic region	N	12323851/na	34830
AtMULE226	unknown protein	X	22331929/At3g15030	217
AtMULE228	hypothetical proteins	N	12322814/At2g14240 & At2g14230	85377
AtMULE228	intergenic region	N	12322814/na	82171
AtMULE228	intergenic region	N	20197425/na	59341
AtMULE228	intergenic region	N	20197425/na	59605
AtMULE228	intergenic region	N	20197425/na	6109
AtMULE230	putative protein	N	7649372/18407749	49671
AtMULE230	putative RNA-binding protein	N	20197957/At2g42890	31096
AtMULE233	intergenic region	N	7635467/na	87630
AtMULE233	intergenic region	N	7635467/na	87630
AtMULE233	plasma membrane intrinsic protein	N	20197457/At2g16850	106666
AtMULE234	putative protein	N	7594547/T15B3_180	102243
AtMULE239	intergenic region	N	3510344/na	38611
AtMULE239	intergenic region	N	4519187/na	14431
AtMULE243	hypothetical protein	N	12323851/F7P12.4	72496
AtMULE243	intergenic region	N	3449311/na	43060

58653	1401	312	1668	1357	na
733	733	6913	11219	4307	ORF
46864	1567	1801	3352	1552	ORF
733	733	6941	10870	3930	ORF
131	131	5818	6282	465	exon
11279	264	351	615	265	na
44652	866	2745	3618	874	na
46531	117	3634	3749	116	na
233	233	1957	2883	927	ORF
87772	143	2055	2198	144	na
41994	652	275	905	631	na
500	363	32	398	367	5' upstream region
1660	830	417	1246	830	5' upstream region
75647	115	37	151	115	na
129489	432	1187	1569	383	exon & intron
72790	123	896	1019	124	na
34944	115	2161	2275	115	na
364	148	214	546	333	exon
86522	1146	6887	8034	1148	ORF & exon & intergenic region
83461	1291	8951	10189	1239	na
59489	149	1058	1207	150	na
59863	259	1308	1567	260	na
6265	157	1837	1992	156	na
50045	375	3224	3592	369	exon & intron
31669	574	3607	4183	577	exon & intron
87772	143	5730	5873	144	na
87772	143	507	650	144	na
106849	184	3300	3481	182	exon & intron
102488	246	371	620	250	intron
38935	325	343	666	324	na
14765	335	795	1129	335	na
72738	243	217	457	241	exon & intron
44117	1058	1351	2397	1047	na
44414	177	2680	2854	175	na
95283	217	881	1100	220	exon & intron

AtMULE243	intergenic region	N	3449311/na	44238
AtMULE243	similar to Homo sapiens Werner WRN	N	5748493/F18A5.260	95067
AtMULE243	similar to membrane-associated salt-inducible protein isolog	N	7549541/F1L3.33	118792
AtMULE251	intergenic	N	6899910/na	47917
AtMULE251	putative replication protein A1	N	6899910/T12K4_70	30479
AtMULE251	putative replication protein A1	N	6899910/T12K4_70	30145
AtMULE251	hypothetical protein	N	6899910/T12K4_40	26080
AtMULE251	hypothetical protein	N	6899910/T12k4_30	25235
AtMULE252	hypothetical protein	X	20198021/At2g12100	1
AtMULE252	hypothetical protein	X	20198021/At2g12130	1
AtMULE258	putative lumen protein retaining receptor	N	22326553/7330791	4369
AtMULE258	putative lumen protein retaining receptor	N	22326553/7330791	3810
AtMULE258	putative lumen protein retaining receptor	N	22326553/7330791	2260
AtMULE258	putative lumen protein retaining receptor	N	22326553/7330791	1820
AtMULE258	intergenic region	N	7330791/na	750
AtMULE258	t5r gene	N	na/1632775	542
AtMULE259	unknown protein & ABC transporter-like protein	N	3869067/MCK7.13&MCK7.14	34908
AtMULE261	intergenic region	N	20197606/na	47299
AtMULE261	similar to Homo sapiens TAK1 binding protein	N	5002514/AT4g11860	50011
AtMULE265	hypothetical protein	X	22330780/At1g27800	1
AtMULE265	hypothetical protein	X	22331929/At3g24380	1
AtMULE265	hypothetical protein	X	22331929/At3g24390	1
AtMULE265	hypothetical protein	X	22331929/At3g24390	1
AtMULE265	putative replication protein A1	X	20197260/At2g14780	1
AtMULE267	putative MADS-box protein fragment	N	20198021/At2g12000	54603
AtMULE267	putative MADS-box protein fragment	N	20198021/At2g11990	52073
AtMULE267	intergenic region	N	20198021/na	55529
AtMULE270	intergenic region	N	7635467/na	87630
AtMULE270	intergenic region	N	20198197/na	55878
AtMULE272	14-3-3-like protein GF14 phi (GRF4) gene	N	na/2232145	184
AtMULE272	intergenic region	N	3243214/na	66807
AtMULE275	putative protein	N	5002514/At4g11860	50011

119000	209	2883	3091	209	exon
50414	2498	963	3477	2515	na
31686	1208	3627	4777	1151	exon & intron & intergenic region
30419	275	4827	5145	319	exon & intron & intergenic region
26628	549	5277	5829	553	exon & intergenic region
25873	639	6074	6712	639	exon & intron
1248	1248	93432	99476	6045	ORF
209	209	101725	102354	630	ORF
4512	144	5637	5780	144	exon & intron
4227	418	5924	6342	419	exon & intron
2403	144	7763	7901	139	exon & intron
2117	298	8076	8374	299	exon & intron
1629	880	8491	9371	881	na
767	226	4721	4943	223	promoter region
35339	432	316	750	435	exon & intergenic region
47486	188	1042	1232	191	na
50144	134	818	951	134	exon
211	211	5742	6374	633	ORF
230	230	7171	7860	690	ORF
65	65	9273	9467	195	ORF
1115	1115	9836	14661	4826	ORF
235	235	4196	5141	946	ORF
54977	375	6199	6696	498	gene segment
52263	191	5395	5584	190	gene segment
56022	494	7113	7832	720	na
87772	143	512	655	144	na
56077	200	109	307	199	na
321	138	1980	2116	137	5' untranslated region & exon
66943	137	1284	1420	137	na
50144	134	911	1044	134	exon
526	199	366	732	367	5' UTR & exon & intron
83096	111	26	134	109	intron
69399	364	131	494	364	na

AtMULE276	homeobox-leucine zipper	N	na/16327	328
AtMULE281	similar to wax synthase	N	7159339/F12K21.19	82986
AtMULE283	intergenic region	N	20197571/na	69036
AtMULE284	intergenic region	N	5822965/na	49007
AtMULE285	intergenic region	N	20197457/na	51991
AtMULE291	intergenic region	N	9885848/na	64017
AtMULE294	intergenic region	N	20197457/na	51949
AtMULE308	intergenic region	N	20197457/na	52043
AtMULE309	intergenic region	N	4589432/na	42939
AtMULE311	intergenic region	N	4538895/na	66144
AtMULE311	intergenic region	N	4732170/na	48791
AtMULE317	unknown protein	N	20197898/At2g03570	7016
AtMULE319	intergenic region	N	8051676/na	49154
AtMULE323	intergenic region	N	20197457/na	51958
AtMULE324	intergenic region	N	4732170/na	37201
AtMULE325	hypothetical protein	X	22330780/At1g27800	1
AtMULE325	putative protein	X	6899956/T5C2_90	1
AtMULE325	putative replication protein A1	X	20197260/At2g14780	1
AtMULE339	putative protein	X	20197260/At2g14780	1
AtMULE339	putative protein	X	22331929/At3g43380	1
AtMULE339	putative protein	X	18407555/At3g43390	1
AtMULE345	integenic region	N	12321079/na	9939
AtMULE345	intergenic region	N	12321079/na	11824
AtMULE345	hypothetical protein	N	12321079/F19C17.27	22335
AtMULE345	hypothetical protein	N	12321079/F19C17.27	22901
AtMULE345	hypothetical protein	N	12321079/F19C17.27	23205
AtMULE345	hypothetical protein	N	12321079/F19C17.27	23755
AtMULE352	intergenic region	N	20198197/na	55952
AtMULE352	intergenic region	N	7635467/na	87630
AtMULE355	intergenic region	N	3293583/na	11018
AtMULE359	intergenic region	N	14475930/na	80066
AtMULE359	intergenic region	N	14475930/na	80070
AtMULE360	intergenic region	N	3047074/na	18137
AtMULE361	intergenic region	N	9954738/na	136357
AtMULE372	hypothetical protein	X	22331929/At3g30480	1
AtMULE372	putative replication protein A1	X	22331929/At3g30500	1
AtMULE375	cytochrome P450	X	22330780/At1g47630	280

49143	137	38	173	136	na
52176	186	1571	1756	186	na
64587	571	643	1213	571	na
52176	228	1287	1508	222	na
52176	134	310	443	134	na
43147	209	207	415	209	na
66354	211	14596	14803	208	na
49095	305	7225	7529	305	na
7851	836	289	1140	852	exon & intron & intergenic region
50511	1358	1256	2597	1342	na
52176	219	298	516	219	na
37597	397	7817	8212	396	na
211	211	6950	7582	633	ORF
1113	1113	506	6143	5638	ORF
235	235	8177	8912	736	ORF
235	235	3319	4264	946	ORF
215	215	4797	5441	645	ORF
1113	1113	6248	10927	4680	ORF
10341	403	2253	2655	403	na
12212	389	2927	3313	387	na
22589	255	3308	3557	250	exon & intergenic region
23023	123	3765	3884	120	exon
23691	487	4465	4950	486	exon
23920	166	5020	5185	166	exon
56077	126	185	304	120	na
87772	143	488	630	143	na
11315	298	316	613	298	na
80179	114	954	1067	114	na
80365	296	1064	1359	296	na
18295	159	323	477	155	na
137041	685	254	966	713	na
156	156	8947	9483	537	ORF
257	257	11287	12294	1008	ORF
397	118	1698	2030	333	exon
41616	238	747	993	247	na
41570	189	1009	1206	198	na
41571	190	1224	1422	199	na

AtMULE375	intergenic region	N	12324708/na	41379
AtMULE375	intergenic region	N	12324708/na	41382
AtMULE375	intergenic region	N	12324708/na	41382
AtMULE375	putative protein phosphatase 2C	N	20197227/18397479	26298
AtMULE380	putative protein	N	6899953/T21C14_30	51200
AtMULE382	putative pectinesterase	N	7243815/F27F5.7	30232
AtMULE382	intergenic region	N	4760411/na	64455
AtMULE382	intergenic region	N	4760411/na	91184
AtMULE382	glucose-6-phosphate/ phosphate-translocator	N	na/7229674	1762
AtMULE382	glucose-6-phosphate/	N	na/7229674	2666
AtMULE382	hypothetical protein	X	22329272/At4g03940	1
AtMULE383	unknown protein	N	12322979/F13K9.7	31273
AtMULE383	hypothetical protein	N	4263038/18399481	52051
AtMULE383	hypothetical protein	X	22330780/At1g44880	1
AtMULE383	putative replication protein A1	X	20198226/At2g10370	1
AtMULE388	intergenic region	N	2264320/na	1152
AtMULE389	homeobox-leucine zipper protein HAT5 mRNA	N	12408717/At3g01470	5724
AtMULE390	putative disease resistance protein	X	22331929/At3g30816	22
AtMULE390	putative protein	X	22331929/At3g43390	1
AtMULE390	putative protein	X	22331929/At3g43370	1
AtMULE391	intergenic region	N	20197457/na	51958
AtMULE392	intergenic region	N	20197457/na	51949
AtMULE392	intergenic region	N	20197457/na	51972
AtMULE393	protein phosphatase 2C	X	22326553/At2g40180	707
AtMULE393	putative protein	N	7362771/T28A8_70	38188
AtMULE398	replication protein A1-like	X	5672589/MSJ3.14	1
AtMULE402	putative protein	X	22331929/At3g43390	1
AtMULE411	intergenic region	N	3449311/na	43171
AtMULE411	putative protein phosphatase 2C	N	20197227/At2g30020	26298
AtMULE411	similar to membrane-associated salt-inducible protein isolog	N	7549541/F1L3.33	118325
AtMULE411	intergenic region	N	12324708/na	41379
AtMULE411	intergenic region	N	12324708/na	41382

26520	223	511	732	222	exon
51314	115	611	725	115	exon & intron
33362	3131	3062	5119	2058	ORF
65582	1128	10735	12385	1651	n/a
91469	286	14451	14740	290	n/a
1959	198	12995	13192	198	exon
2889	224	13659	13881	223	exon & intron
310	310	7883	8398	516	ORF
31806	534	674	1864	1191	exon & intron
54363	2313	7802	13691	5890	part of gene T5L23.13, intergenic region, and part of gene T5L23.14
998	998	2164	7241	5078	exon
266	266	10007	11015	1009	ORF
1330	179	636	813	178	na
5999	276	350	648	299	exon & intron
342	321	10093	11656	1564	exon
1113	1113	507	6461	5955	ORF
235	235	11915	12860	946	ORF
52176	219	334	553	220	na
52187	239	275	512	238	na
52187	216	670	884	215	na
1389	683	160	389	230	exon
38440	253	1888	2223	336	exon & intron
248	248	10780	11792	1013	ORF
1087	1087	510	6344	5835	exon & intron
44119	949	722	1657	936	na
26520	223	3283	3505	223	exon
118740	416	219	634	416	exon
41570	192	2592	2790	199	na
41570	189	2809	3006	198	na
41616	238	3022	3268	247	na
31994	145	1773	1917	145	exon

AtMULE411	intergenic region	N	12324708/na	41379
AtMULE411	hypothetical protein	N	3080352/At4g19320	31850
AtMULE413	hypothetical protein	N	5733889/F18B13.31	113352
AtMULE413	intergenic region	N	6899954/na	57616
AtMULE413	phosphoenolpyruvate translocator PPT1	N	na/7546828	73329
AtMULE416	hypothetical protein	N	6899953/T21C14_30	51133
AtMULE418	intergenic region	N	20197369/na	49705
AtMULE419	hypothetical protein	X	3080352/T5K18.50	28
AtMULE422	heat shock protein 101 mRNA	N	12324896/At1g74310	12711
AtMULE423	intergenic region	N	6899953/na	27879
AtMULE423	intergenic region	N	3319339/na	11040
AtMULE423	unknown protein	N	12322414/F8D11.12	66268
AtMULE425	intergenic region	N	20197969/na	56226
AtMULE429	hypothetical protein	N	7268604/AT4g18420	141990
AtMULE429	intergenic region	N	7268604/na	142446
AtMULE433	intergenic region	N	7635467/na	87630
AtMULE436	hypothetical protein	X	5706727/T4B21.8	1
AtMULE437	intergenic region	N	5881769/na	40136
AtMULE442	hypothetical protein	X	22331929/At3g30480	18
AtMULE442	hypothetical protein	X	22331929/At3g30500	46
AtMULE442	unknown protein	X	4519196/MSK10.9	1
AtMULE452	intergenic region	N	20197457/na	51951
AtMULE454	hypothetical protein	N	4455262/F17L22.200	77953
AtMULE455	intergenic region	N	4538990/na	38568
AtMULE458	intergenic region	N	13677103/na	43625
AtMULE458	intergenic region	N	13677103/na	46456
AtMULE458	similar to phosphate transporter proteins	X	4757678/F9H16.16	302
AtMULE458	unknown protein	X	3985955/MTH16.13	1
AtMULE460	cytochrome P450	X	22330780/At1g47630	280
AtMULE460	intergenic region	N	2564051/na	52545
AtMULE460	intergenic region	N	2564051/na	52545
AtMULE460	intergenic region	N	2564051/na	52558
AtMULE460	putative protein phosphatase 2C	N	20197227/At2g30020	26298
AtMULE465	intergenic region	N	20197969/na	56226
AtMULE467	intergenic region	N	14475930/na	80066

113758	407	2065	2472	408	exon & intron
58860	1245	802	2037	1236	na
73688	360	388	742	355	5' UTR
51695	563	2884	3437	554	ORF
49935	231	436	667	232	na
179	152	136	615	480	exon
13274	564	386	939	554	5' UTR
27984	106	1647	1752	106	na
11408	369	2347	2720	374	na
66827	560	2741	3300	560	exon & intron
56415	190	63	252	190	na
142192	203	4648	4851	204	exon & intron
142573	128	5045	5172	128	na
87772	143	528	671	144	na
160	160	368	916	549	exon
40481	346	1335	1680	346	na
155	138	7114	7596	483	exon
257	212	9497	10284	788	exon
142	142	4694	5281	588	exon
52156	206	253	457	205	na
78086	134	305	438	134	exon
38881	314	809	1108	300	na
44694	1070	602	1669	1068	na
46531	76	473	548	76	na
366	65	1976	2170	195	exon
43	43	97	225	129	exon
397	118	215	547	333	exon
52757	213	1248	1458	211	na
52747	203	818	1015	198	na
52747	190	1045	1231	187	na
26520	223	1511	1733	223	exon
56415	190	195	384	190	na
80179	114	986	1099	114	na
80365	296	1096	1391	296	na
76	76	349	576	228	ORF

AtMULE467	intergenic region	N	14475930/na	80070
AtMULE470	hypothetical protein	X	22329272/At4g35400	1
AtMULE476	intergenic region	N	20198240/na	21208
AtMULE476	intergenic region	N	20198240/na	21220
AtMULE476	intergenic region	N	20198240/na	21627
AtMULE477	intergenic region	N	7635467/na	87630
AtMULE477	intergenic region	N	20198197/na	55879
AtMULE479	intergenic region	N	9971625/na	87908
AtMULE479	intergenic region	N	2656030/na	64329
AtMULE479	intergenic region	N	5041971/na	43706
AtMULE479	unknown protein	X	3985955/MTH16.13	1
AtMULE480	intergenic region	N	7267134/na	86695
AtMULE482	putative protein	X	9755607/F14F8_70	1
AtMULE483	putative protein	X	9755632/F1N13_130	1
AtMULE485	intergenic region	N	5391457/na	76771
AtMULE485	intergenic region	N	6899941/na	52982
AtMULE486	intergenic region	N	20197969/na	56225
AtMULE487	intergenic region	N	20197969/na	56229
AtMULE488	intergenic region	N	5430744/na	50987
AtMULE488	intergenic region	N	6850877/na	49024
AtMULE493	similar to Homo sapiens TAK1 binding protein	N	5002514/AT4g11860	50011
AtMULE494	hypothetical protein	X	9828633/F1N21.7	146
AtMULE495	subunit 6b of cytochrome c oxidase	N	6056182/F12K8.20	82017
AtMULE499	hypothetical protein	N	8576188/F6I1.2	41967
AtMULE499	intergenic region	N	5430744/na	50767
AtMULE499	putative protein phosphatase 2C	N	20197227/At2g30020	26298
AtMULE499	putative replication protein A1	N	20197574/At2g11170	34254
AtMULE499	putative replication protein A1	N	20197574/At2g11170	31743
AtMULE499	putative replication protein A1	N	20197574/At2g11170	32263
AtMULE499	putative replication protein A1	N	20197574/At2g11170	31790
AtMULE500	intergenic region	N	20197710/na	16729
AtMULE500	intergenic region	N	20197710/na	16729
AtMULE500	intergenic region	N	20197710/na	17238
AtMULE500	intergenic region	N	20197710/na	17238
AtMULE500	intergenic region	N	20197710/na	17238
AtMULE500	intergenic region	N	20197710/na	16729

21392	185	208	393	186	na
21351	132	771	902	132	na
21729	103	536	639	104	na
87772	143	1848	1991	144	na
56077	199	2197	2398	202	na
88103	196	584	794	211	na
64462	134	412	545	134	na
43943	238	1139	1375	237	na
65	65	1307	1501	195	ORF
87307	613	243	858	616	na
169	169	2463	3039	577	ORF
115	115	429	918	490	ORF
77349	579	844	1410	567	na
53107	126	677	801	125	na
56410	186	69	254	186	na
56409	181	70	250	181	na
51197	211	799	1011	213	na
49498	475	1776	2261	486	na
50144	134	897	1030	134	exon
303	158	4076	4549	474	exon
82440	424	382	818	437	intron
42144	178	6227	6412	186	exon & intron
51228	462	6773	7234	462	na
26640	343	7256	7590	335	exon
34482	229	726	953	228	exon & intron
31883	141	966	1105	140	exon & intron
32480	218	1118	1331	214	exon & intron
32211	422	1365	1772	408	exon & intron
18542	1814	2520	4331	1812	na
18361	1633	13091	14717	1627	na
18542	1305	11400	12710	1311	na
18542	1305	8607	9917	1311	na
18542	1305	5814	7124	1311	na
17242	514	10298	10802	505	na
17242	514	7505	8009	505	na
17242	514	4712	5216	505	na

AtMULE500	intergenic region	N	20197710/na	16729
AtMULE500	intergenic region	N	20197710/na	16729
AtMULE500	intergenic region	N	20197710/na	91575
AtMULE500	putative protein	X	22328163/At5g28270	1
AtMULE500	putative replication protein A1	X	15227179/At2g16330	1
AtMULE501	intergenic region	N	20197835/na	48732
AtMULE502	intergenic region	N	20197835/na	48732
AtMULE503	hypothetical protein	X	20198021/At2g12130	1
AtMULE504	intergenic region	N	18149191/na	56074
AtMULE504	intergenic region	N	18149191/na	56074
AtMULE506	putative protein	X	7635467/F14L2_50	1
AtMULE507	hypothetical protein	X	5430745/F13F21.12	3
AtMULE511	intergenic region	N	7270670/na	73746
AtMULE511	intergenic region	N	7270670/na	96972
AtMULE512	unknown protein	N	8051665/T19N8.8	30079
AtMULE513	intergenic region	N	20197574/na	20497
AtMULE513	intergenic region	N	20197574/na	20784
AtMULE513	hypothetical protein	N	20197574/At2g11160	21026
AtMULE513	hypothetical protein	N	20197574/At2g11160	22172
AtMULE513	hypothetical protein	N	20197574/At2g11160	22378
AtMULE513	hypothetical protein	N	20197574/At2g11160	23046
AtMULE513	intergenic region	N	20197574/na	23646
AtMULE513	intergenic region	N	20197574/na	24360
AtMULE516	intergenic region	N	5391457/na	76823
AtMULE518	intergenic region	N	8347620/na	95758
AtMULE518	putative replication protein A1	N	20197574/At2g11170	31140
AtMULE518	putative replication protein A1	N	20197574/At2g11170	29842
AtMULE518	putative replication protein A1	N	20197574/At2g11170	27384
AtMULE518	putative replication protein A1	N	20197574/At2g11170	29092
AtMULE518	putative replication protein A1	N	20197574/At2g11170	26334
AtMULE518	unknown protein	N	8347620/F1D9.22	96361
AtMULE518	intergenic region	N	8347620/na	95758
AtMULE518	unknown protein	N	8347620/F1D9.21	95122
AtMULE518	unknown protein	N	8347620/F1D9.21	95173
AtMULE519	similar to cyclin-dependent protein kinase cdc2MsC	N	12408709/T4P13.23	68086
AtMULE520	hypothetical protein	N	5091531/T19E23.17	82108

92071	497	2917	3417	501	na
526	526	17523	19901	2379	ORF
165	165	16366	17019	654	exon
48861	130	190	320	131	na
48861	130	190	320	131	na
209	209	7760	8386	627	ORF
56393	320	621	941	321	na
56436	363	1272	1621	350	na
114	114	430	771	342	exon
63	61	107	331	225	exon
74197	452	11221	11667	447	na
97504	533	7703	8249	547	na
31938	1860	10950	12804	1855	exon & intron
20681	185	9379	9564	186	na
20949	166	9562	9727	166	na
21862	837	9807	11036	1230	exon & intergenic region
22319	148	11239	11384	146	exon & intron
22795	418	11476	11893	418	intron
23448	403	12092	12495	404	exon & intergenic region
24260	615	12539	13153	615	na
24568	209	13808	14022	215	na
77290	468	266	732	467	na
95873	116	9610	9725	116	na
34482	3343	726	4431	3706	exon & intron & intergenic region
31076	1235	4437	11123	6687	exon & intron
27598	215	10732	10949	218	exon
29336	245	11317	12020	704	exon & intron
27844	1511	13293	14798	1506	exon & intron & intergenic region
97662	1302	5063	6362	1300	exon & intron & intergenic region
95987	230	6482	6721	240	na
95690	569	10465	11030	566	ORF & intergenic region
95421	249	13539	13786	248	exon & intron & intergenic region
68558	473	297	778	482	exon & intron
82251	144	964	1106	143	exon
82251	144	90	233	144	exon
63	61	107	307	201	exon

AtMULE520	hypothetical protein	N	5091531/T19E23.17	82108
AtMULE522	hypothetical protein	X	5430745/F13F21.12	3
AtMULE523	intergenic region	N	4538895/na	66130
AtMULE524	intergenic region	N	3449311/na	42085
AtMULE526	putative protein	N	6899953/T21C14_30	51400
AtMULE527	unknown protein	N	8051665/T19N8.8	30115
AtMULE527	unknown protein	N	8051665/T19N8.8	31138
AtMULE532	intergenic region	N	20197457/na	51958
AtMULE533	intergenic region	N	2264320/na	1149
AtMULE534	hypothetical protein	X	9438236/T12C24.25	27
AtMULE537	intergenic region	N	20197969/na	56226
AtMULE539	similar to Homo sapiens TAK1 binding protein	N	5002514/At4g11860	50011
AtMULE541	intergenic region	N	7649372/na	56885
AtMULE543	intergenic region	N	20197457/na	51991
AtMULE544	intergenic region	N	7635467/na	87630
AtMULE544	intergenic region	N	20198197/na	55914
AtMULE545	intergenic region	N	20198021/na	52073
AtMULE545	putative MADS-box protein fragment	N	20198021/At2g12000	54603
AtMULE545	intergenic region	N	20198021/na	55789
AtMULE545	unknown protein	X	2828187/K21C13.7	1
AtMULE547	intergenic region	N	7671394/na	31786
AtMULE547	similar to Homo sapiens TAK1 binding protein	N	7267843/At4g11860	187311
AtMULE552	similar to Homo sapiens TAK1 binding protein	N	5002514/At4g11860	50011
AtMULE555	intergenic region	N	9954737/na	5188
AtMULE555	intergenic region	N	9954737/na	5188
AtMULE555	transparent testa glabra 1 Ttg1 protein	N	2760164/K18P6.4	6889
AtMULE555	transparent testa glabra 1 Ttg1 protein	N	2760164/K18P6.4	6889
AtMULE562	intergenic region	N	18265367/na	87630
AtMULE562	intergenic region	N	20197478/na	32558
AtMULE564	intergenic region	N	6056182/na	46777
AtMULE568	hypothetical protein	X	6957851/T32E20.4	1
AtMULE570	hypothetical protein	X	18399881/At1g35770	1
AtMULE571	hypothetical protein	N	5732428/F26C17.7 & F22O13.22	31177

66293	164	5983	6141	159	na
45025	2941	1099	4044	2946	na
51662	263	345	607	263	exon
30900	786	4109	4904	796	exon & intron
31938	801	3051	3837	787	exon & intron
52176	219	1469	1688	220	na
1330	182	661	841	181	na
156	130	438	888	451	exon
56415	190	65	254	190	na
50144	134	328	461	134	exon
57733	849	390	1215	826	na
52176	186	1321	1505	185	na
87772	143	468	611	144	na
56077	164	2039	2195	157	na
52263	191	5395	5584	190	na
54786	184	6199	6382	184	gene segment
56022	234	7599	7837	239	na
1015	1015	7220	11658	4439	ORF
31976	191	1397	1586	190	na
187444	134	2136	2269	134	exon
50144	134	328	461	134	exon
5392	205	1118	1320	203	na
5392	205	3303	3505	203	na
6994	106	2639	2744	106	exon
6994	106	1879	1984	106	exon
87772	143	507	650	144	na
33063	506	2428	2929	502	na
48108	1332	730	2063	1334	na
94	94	620	967	348	exon
967	967	481	6119	5639	exon & intron
31881	705	1564	2267	704	exons & intergenic region
32911	914	446	1357	912	exon & intergenic region

	& hypothetical protein pseudogene			
AtMULE571	hypothetical protein pseudogene	N	5732428/F22O13.22	31998
AtMULE572	intergenic region	N	7635467/na	87630
AtMULE572	intergenic region	N	20198197/na	55879
AtMULE573	putative replication protein	X	2232816/At5g28940	7
AtMULE581	intergenic region	N	5391457/na	76823
AtMULE591	intergenic region	N	20197494/na	14132
AtMULE592	intergenic region	N	20197494/na	14129
AtMULE593	intergenic region	N	20197916/na	98728
AtMULE597	intergenic region	N	20198197/na	55937
AtMULE599	intergenic region	N	8347605/na	1324
AtMULE599	intergenic region	N	8347605/na	1864
AtMULE599	intergenic region	N	8347605/na	2439
AtMULE599	intergenic region	N	5041971/na	43706
AtMULE601	intergenic region	N	8778333/na	63192
AtMULE602	intergenic region	N	5041967/na	46220

*: Clone GI for unknown genes or, where unavailable, corresponding to chromosome GI.
†: Gene GI, MIPS code or locus as provided by NCBI (http://www.ncbi.nlm.nig.gov/BLAST/).
‡: Nucleotide positions within the corresponding clone /chromosome for BLASTN or amino acid positions within the ORF for BLASTX.
§: Size in bp for BALSTN results, or in aa for BALSTX results.
na: not applicable.

87772	143	468	611	144	na
56077	199	99	281	183	na
466	460	10	2112	2103	exon
77376	554	196	732	537	na
14242	111	456	566	111	na
14274	146	133	278	146	na
98892	165	187	354	168	na
56077	141	2129	2262	134	na
1470	147	872	1017	146	na
2227	364	665	1030	366	na
2812	374	1220	1593	374	na
43944	239	223	455	233	na
63507	316	58	375	318	na
47063	844	278	1119	842	na

Supplementary Table 5.3B Acquistion of host DNA sequences by rice MULEs

MULE	rice gene	blastN/X	clone GI*/ORF ID†
OsMULE0101	5S ribosomal RNA	N	na/1813883
OsMULE0102	5S ribosomal RNA	N	na/1813883
OsMULE0103	5S ribosomal RNA	N	na/1813883
OsMULE0104	5S ribosomal RNA	N	na/1813883
OsMULE0105	5S ribosomal RNA	N	na/1813883
OsMULE0106	5S ribosomal RNA	N	na/1813883
OsMULE0107	5S ribosomal RNA	N	na/1813883
OsMULE0108	5S ribosomal RNA	N	na/1813883
OsMULE0109	5S ribosomal RNA	N	na/1813883
OsMULE0110	5S ribosomal RNA	N	na/1813883
OsMULE0111	5S ribosomal RNA	N	na/1813883
OsMULE0114	5S ribosomal RNA	N	na/1813883
OsMULE0120	Aegilops crassa large subunit of ribulose 1,5 bisphosphate carboxylase/oxygenase & apoprotein I	N	na/11308
OsMULE0120	intergenic region	N	15623922/na
OsMULE0125	hypothetical protein	X	na/20805022
OsMULE0126	5S ribosomal RNA	N	na/1813883
OsMULE0127	5S ribosomal RNA	N	na/1813884
OsMULE0201	hypothetical protein	X	na/5091500
OsMULE0301	hypothetical protein	X	na/11320854
OsMULE0302	hypothetical protein	N	15004914 / P0583G08.7
OsMULE0302	hypothetical protein	N	15004914 / P0583G08.7
OsMULE0303	intergenic region	N	5042437
OsMULE0304	intergenic region	N	5042437
OsMULE0305	hypothetical protein	X	na/8467942
OsMULE0306	intergenic region	N	10179050/na

position of acquired segment within the clone			position of acquired segment within the MULE			feature of the acquired segment
start	end	size	start	end	size (bp)	
160	268	109	556	664	109	nontranscribed spacer
169	271	103	577	679	103	nontranscribed spacer
165	271	107	583	689	107	nontranscribed spacer
160	271	112	585	696	112	nontranscribed spacer
171	271	101	596	696	101	nontranscribed spacer
171	271	101	699	799	101	nontranscribed spacer
160	271	112	596	707	112	nontranscribed spacer
160	271	112	596	707	112	nontranscribed spacer
170	271	102	692	793	102	nontranscribed spacer
165	284	120	548	667	120	nontranscribed spacer
170	271	102	690	791	102	nontranscribed spacer
165	280	116	464	579	116	nontranscribed spacer
3444	3686	243	917	1159	243	exon
5820	6017	198	416	609	194	na
1	60	60	75	254	180	exon
170	271	102	685	786	102	nontranscribed spacer
174	271	98	599	696	98	nontranscribed spacer
66	201	136	471	878	408	exon
1	143	143	538	1051	514	ORF
45781	45921	141	1082	1222	141	exon
46109	46185	77	858	934	77	exon & intron
107754	107857	104	647	754	108	na
107754	107857	104	647	754	108	na
1	187	187	365	1080	716	ORF
61354	61454	101	711	815	105	na
19	173	155	862	1278	417	exon
18062	18220	159	976	1133	158	na

OsMULE0306	hypothetical protein	X	na/8096389
OsMULE0307	intergenic region	N	16904683/na
OsMULE0307	intergenic region	N	15290074/na
OsMULE0307	hypothetical protein	X	na/12039282
OsMULE0308	intergenic region	N	20143588/na
OsMULE0308	hypothetical protein	X	na/12328548
OsMULE0309	hypothetical protein	N	5803242 / 5803251
OsMULE0401	PCF1 and PCF2 mRNA	N	na/2580439
OsMULE0401	intergenic region	N	20161904/na
OsMULE0401	intergenic region	N	20161904/na
OsMULE0401	intergenic region	N	20161904/na
OsMULE0402	PCF1 and PCF2 mRNA	N	na/2580439
OsMULE0402	intergenic region	N	20161904/na
OsMULE0402	intergenic region	N	20161904/na
OsMULE0402	intergenic region	N	20161904/na
OsMULE0403	hypothetical protein	X	na/11320854
OsMULE0404	hypothetical protein	X	na/8096419
OsMULE0405	intergenic region	N	18542908/na
OsMULE0405	intergenic region	N	18542908/na
OsMULE0405	intergenic region	N	18542908/na
OsMULE0407	hypothetical protein	X	na/19571149
OsMULE0408	hypothetical protein	X	na/19571149
OsMULE0410	hypothetical protein	X	na/15528609
OsMULE0411	hypothetical protein	X	na/15144317
OsMULE0414	hypothetical protein	X	na/12039282
OsMULE0415	hypothetical protein	X	na/15144317
OsMULE0416	intergenic region	N	16904677/na
OsMULE0416	intergenic region	N	16904677/na
OsMULE0501	hypothetical protein	X	na/13486853

110196	110306	111	1581	1691	111	na
1	222	222	246	2495	2250	ORF
2599	2712	114	238	347	110	na
25	231	207	3	2330	2328	exon
44070	44187	118	188	310	123	intron
334	492	159	1081	1239	159	exon
122084	122401	318	459	767	309	na
121257	121534	278	1051	1326	276	na
121825	122036	212	820	1037	218	na
334	492	159	1081	1239	159	exon
122084	122401	318	459	767	309	na
121257	121534	278	1051	1326	276	na
121825	122036	212	820	1037	218	na
1	126	126	538	906	369	ORF
1	355	355	243	1355	1113	ORF
73293	73825	533	921	1509	589	na
73969	74106	138	680	827	148	na
74290	74409	120	476	594	119	na
8	313	306	212	1218	1007	exon
8	313	306	180	1186	1007	exon
214	504	291	351	1479	1129	exon
1	223	223	848	1516	669	ORF
1	82	82	558	912	355	exon
1	223	223	849	1517	669	ORF
96834	96963	130	842	968	127	na
97173	97382	210	493	713	221	na
6	98	93	79	435	357	exon
6	98	93	79	435	357	exon
6	98	93	79	435	357	exon

OsMULE0503	hypothetical protein	X	na/13486853
OsMULE0504	hypothetical protein	X	na/13486853
OsMULE0702	hypothetical protein	N	10241657 / H0711G06.18
OsMULE0702	hypothetical protein	N	10241657 / H0711G06.18
OsMULE0702	intergenic region	N	14572678/na
OsMULE0703	hypothetical protein	X	na/6069657
OsMULE0704	Zea mays tonoplast membrane integral protein	N	na/13447826
OsMULE0704	putative beta-tonoplast intrinsic protein	N	15341601 / OSJNBa0051D19.19
OsMULE0705	heat shock protein 82	N	na/20255
OsMULE0705	Avena sativa permatin precursor mRNA	N	na/1373391
OsMULE0705	Avena sativa permatin precursor mRNA	N	na/1373391
OsMULE0707	Zea mays tonoplast membrane integral protein	N	na/13447826
OsMULE0707	putative beta-tonoplast intrinsic protein	N	15341601 / OSJNBa0051D19.19
OsMULE0709	hypothetical protein	X	na/12039286
OsMULE0711	putative mitochondrial inner membrane protein	N	10122030 / OSJNBb0018B10.5
OsMULE0714	S-adenosyl-L-methionine synthetase	N	na/1778820
OsMULE0715	hypothetical protein	X	na/11761078
OsMULE0716	putative sucrose-phosphate synthase	X	na/10140657
OsMULE0717	hypothetical protein	X	na/9945059
OsMULE0718	hypothetical protein	X	na/12382002
OsMULE0719	hypothetical protein	X	na/14165329
OsMULE0722	hypothetical protein	X	na/11034577
OsMULE0728	hypothetical protein	X	na/13486758
OsMULE0729	hypothetical protein	X	na/13194234
OsMULE0730	hypothetical protein	X	na/13872991
OsMULE0731	hypothetical protein	X	na/11034577
OsMULE0731	putative cinnamoyl-CoA reductase	X	na/10140638
OsMULE0732	hypothetical protein	X	na/11034577
OsMULE0732	putative cinnamoyl-CoA reductase	X	na/10140638
OsMULE0733	unannotated	N	na/21623777

84446	84652	207	1126	1332	207	exon & intron
84828	84976	149	1432	1585	154	exon & intron
24178	24440	263	1819	2080	262	na
1	304	304	372	1693	1322	ORF
630	794	165	459	623	165	intron
26014	26243	230	405	634	230	exon & intron
3075	3221	147	885	1030	146	intron
587	682	96	644	739	96	exon
515	577	63	1201	1265	65	exon
630	794	165	765	929	165	intron
26009	26243	235	754	988	235	exon & intron
1	184	184	302	1027	726	ORF
33893	34333	441	475	903	429	exon
115	698	584	732	1315	584	exon
49	132	84	785	1036	252	exon
7	100	94	355	989	635	exon
1	193	193	563	1252	690	exon
695	940	246	301	1321	1021	exon
1	215	215	416	1563	1148	ORF
39	219	181	922	1464	543	exon
19	313	295	361	1419	1059	exon
20	241	222	355	1020	666	exon
43	184	142	176	809	634	exon
59	219	161	924	1400	477	exon
1	193	193	340	1166	827	exon
59	219	161	924	1400	477	exon
1	193	193	340	1166	827	exon
122335	122580	246	1142	1340	199	na
122923	123024	102	991	1092	102	na
176	240	65	513	707	195	exon

OsMULE0733	unannotated	N	na/21623777
OsMULE0733	hypothetical protein	X	na/14164464
OsMULE0735	intergenic region	N	19386744/na
OsMULE0736	3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase	N	na/5059024
OsMULE0736	3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase	N	na/5059024
OsMULE0736	S-receptor kinase	X	na/10178100
OsMULE0737	putative auxin transport protein-like	N	16904684 / P0678F11.18
OsMULE0737	putative auxin transport protein-like	N	16904684 / P0678F11.18
OsMULE0738	putative auxin transport protein-like	N	16904684 / P0678F11.18
OsMULE0738	putative auxin transport protein-like	N	16904684 / P0678F11.18
OsMULE0739	intergenic region	N	na/19571079
OsMULE0739	Zea mays unknown protein	N	na/21207554
OsMULE0739	Zea mays unknown protein	N	na/21207554
OsMULE0741	unknown protein	X	na/21592457
OsMULE0741	putative mitochondrial processing peptidase α subuunit	X	na/14090341
OsMULE0746	hypothetical protein	X	na/5852080
OsMULE0801	C-repeat/dehydration-responsive element binding factor	X	na/12581492
OsMULE0802	C-repeat/dehydration-responsive element binding factor	X	na/12581492
OsMULE0804	hypothetical protein	X	na/15128225
OsMULE0806	putative peptide transport protein	N	13486880 / P0024G09.4
OsMULE0806	putative peptide transport protein	N	13486880 / P0024G09.4
OsMULE0806	Hordeum vulgare peptide transporter (ptr1) mRNA	N	na/2655097
OsMULE0806	Hordeum vulgare peptide transporter (ptr1) mRNA	N	na/2655097
OsMULE0807	intergenic region	N	10140779/na
OsMULE0807	intergenic region	N	10140779/na
OsMULE0807	intergenic region	N	10140779/na
OsMULE0807	hypothetical protein	N	20330521 / P0460E08.14
OsMULE0808	group 3 late embryogenesis abundant type I protein	N	na/1235566
OsMULE0808	group 3 late embryogenesis abundant type I protein	N	na/1235566
OsMULE0808	group 3 late embryogenesis abundant type I protein	N	na/1235566

38148	38808	661	659	1331	673	na
756	897	142	1124	1265	142	exon
577	638	62	1472	1540	69	exon
708	783	76	742	969	228	exon
78442	78603	162	1046	1199	154	exon
78221	78443	223	516	748	233	exon
78442	78603	162	1046	1199	154	exon
78221	78443	223	516	748	233	exon
68891	69385	495	434	910	477	na
97	165	69	366	434	69	exon
158	661	504	1012	1616	605	exon
142	281	140	462	812	351	exon
1	65	65	916	1110	195	exon
1	67	67	578	778	201	exon
1	253	253	459	1217	759	ORF
1	253	253	459	1217	759	ORF
64	249	186	237	847	611	exon
24848	25076	229	408	633	226	exon
24689	24840	152	941	1092	152	exon
849	1042	194	443	633	191	exon
714	841	128	941	1068	128	exon
69541	69740	200	4580	4779	200	na
69844	70466	623	3855	4476	622	na
70565	70833	269	3528	3796	269	na
59739	60029	291	763	1049	287	exon & intron
298	434	137	437	573	137	exon
432	528	97	993	1089	97	exon
527	611	85	1173	1256	84	exon
149091	149380	290	579	866	288	exon
746	812	67	253	453	201	exon

OsMULE0809	hypothetical protein	N	20161295 / B1131G08.24
OsMULE0809	Arabidopsis thaliana hypothetical protein	X	na/13365573
OsMULE0810	hypothetical protein	N	20161295 / B1131G08.24
OsMULE0810	Arabidopsis thaliana hypothetical protein	X	na/13365573
OsMULE0811	hypothetical protein	X	na/15408816
OsMULE0812	hypothetical protein	X	na/10140628
OsMULE0813	unknown protein	N	17155067 / OSJNBb0060I05.1
OsMULE0813	unknown protein	N	17155067 / OSJNBb0060I05.1
OsMULE0813	putative zinc finger protein	X	na/3236255
OsMULE0814	hypothetical protein	X	na/10140640
OsMULE0816	putative phytochrome-associated protein	N	9049451 / P0710E05.9
OsMULE0817	hypothetical protein	N	15146360 / P0046E05.27
OsMULE0818	hypothetical protein	N	18056689 / OSJNBb0042K08.7
OsMULE0819	unannotated	N	16756248 / OSJNBa0057L21.8
OsMULE0823	pyruvate decarboxylase (pdc3)	N	na/476283
OsMULE0825	hypothetical protein	X	na/11875169
OsMULE0828	hypothetical protein	N	18056689 / OSJNBb0042K08.7
OsMULE0829	hypothetical protein	X	na/9049457
OsMULE0830	hypothetical protein	X	na/13366223
OsMULE0831	hypothetical protein	X	na/13366223
OsMULE0834	putative wall-associated kinase 2	X	na/10241425
OsMULE0835	hypothetical protein	X	na/10122031
OsMULE0836	putative receptor kinase	N	5441876 / 5441881
OsMULE0837	hypothetical protein	X	na/10122031
OsMULE0840	hypothetical protein	X	na/14029012
OsMULE0841	hypothetical protein	N	13486627 / P0439B06.17
OsMULE0842	hypothetical protein	X	na/13873040
OsMULE0843	hypothetical protein	X	na/20804473
OsMULE0844	hypothetical protein	X	na/20804469
OsMULE0844	hypothetical protein	X	na/20804468

149091	149380	290	579	866	288	exon
746	812	67	253	453	201	exon
1	145	145	853	1439	587	exon
1	214	214	233	874	642	ORF
34885	34977	93	471	571	101	exon
34707	34755	49	587	635	49	exon
115	249	135	676	1285	610	exon
1	104	104	889	1200	312	ORF
60953	61444	492	761	1255	495	exon
138385	138497	113	1142	1254	113	exon
62190	62337	148	435	568	134	exon
26	248	223	695	908	214	exon
1114	1342	229	657	885	229	exon
46	103	58	172	345	174	exon
62190	62337	148	435	568	134	exon
1	260	260	257	1040	784	ORF
87	395	309	104	1552	1449	exon
87	395	309	104	1552	1449	exon
347	490	144	878	1556	679	exon
21	271	251	272	1024	753	exon
29274	29417	144	349	492	144	exon
21	271	251	279	1031	753	exon
1	185	185	1677	2475	799	exon
69699	69892	194	904	1089	186	exon & intron
8	197	190	348	1137	790	exon
378	604	227	385	1065	681	exon
101	187	87	475	735	261	exon
11	199	189	1025	1591	567	exon
1	164	164	259	894	636	ORF
1	374	374	608	2113	1506	ORF

OsMULE0845	hypothetical protein	X	na/11034588
OsMULE0846	hypothetical protein	X	na/20279444
OsMULE0847	hypothetical protein	X	na/9558516
OsMULE0848	Bacillus subtilis unknown protein	N	6069643 / na
OsMULE0848	Bacillus subtilis unknown protein	N	6069643 / na
OsMULE0849	auxin transport protein REH1 mRNA	N	na/3377508
OsMULE0849	unknown protein	X	na/14090345
OsMULE0850	hypothetical protein	N	12039314 / OSJNBb0064P21.6
OsMULE0852	hypothetical protein	X	na/11034700
OsMULE0853	hypothetical protein	X	na/10122031
OsMULE0854	hypothetical protein	X	na/10122031
OsMULE0856	hypothetical protein	N	13872907 / P0684B02.25
OsMULE0856	putative esterase	X	na/21537287
OsMULE0857	hypothetical protein	N	13872907 / P0684B02.25
OsMULE0857	putative esterase	X	na/21537287
OsMULE0858	hypothetical protein	X	na/20161664
OsMULE0859	hypothetical protein	X	na/12382009
OsMULE0860	hypothetical protein	X	na/10140796
OsMULE0860	hypothetical protein	X	na/20160652
OsMULE0861	hypothetical protein	N	10140737 / OSJNBa0055P24.15
OsMULE0862	Arabidopsis thaliana hypothetical protein	N	15004914 / P0583G08.7
OsMULE0862	Arabidopsis thaliana hypothetical protein	N	15004914 / P0583G08.7
OsMULE0862	Arabidopsis thaliana hypothetical protein	N	15004914 / P0583G08.7
OsMULE0863	GT-binding nuclear protein (GT-2)	N	na/20248
OsMULE0863	GT-binding nuclear protein (GT-2)	N	na/20248
OsMULE0863	hypothetical protein	X	na/13603424
OsMULE0864	hypothetical protein	N	20503059 / OSJNBb0011A08.21
OsMULE0864	hypothetical protein	X	na/6069651
OsMULE0866	unknown protein	X	na/20521254
OsMULE0867	intergenic region	N	15528744/na

79	590	512	957	2734	1778	exon
1015	1521	507	1001	1496	496	exon
1719	1858	140	835	971	137	exon
1640	1788	149	913	1061	149	exon
607	674	68	516	719	204	exon
24162	24481	320	496	803	308	exon
59	145	87	438	698	261	exon
34	271	238	275	982	708	exon
26	271	246	289	982	694	exon
125186	125395	210	405	614	210	exon
87	327	241	589	1272	684	exon
125186	125395	210	405	614	210	exon
87	327	241	589	1272	684	exon
122	368	247	310	1272	963	exon
700	1001	302	145	1295	1151	exon
17	218	202	300	905	606	exon
187	273	87	1067	1339	273	exon
80381	80553	173	873	1047	175	exon
45628	45802	175	371	537	167	exon & intron
45073	45175	103	544	645	102	exon & intron
46130	46176	47	323	369	47	exon
343	456	114	1188	1301	114	exon
565	602	38	876	913	38	exon
5	157	153	1206	1710	505	exon
56905	57010	106	955	1060	106	exon
29	116	88	1009	1398	390	exon
35	201	167	533	1225	693	exon
50435	50543	109	1274	1384	111	na
24027	24220	194	307	490	184	exon
24027	24220	194	307	490	184	exon

OsMULE0868	hypothetical protein	N	6815051 / 6815056
OsMULE0869	hypothetical protein	N	6815051 / 6815056
OsMULE0871	hypothetical protein	N	20503059 / OSJNBb0011A08.6
OsMULE0871	hypothetical protein	N	20503059 / OSJNBb0011A08.6
OsMULE0871	intergenic region	N	19919056/na
OsMULE0872	hypothetical protein	N	20503059 / OSJNBb0011A08.6
OsMULE0872	hypothetical protein	N	20503059 / OSJNBb0011A08.6
OsMULE0872	intergenic region	N	19919056/na
OsMULE0873	hypothetical protein	N	15408621 / B1039D07.23
OsMULE0873	hypothetical protein	N	na/15408621
OsMULE0874	hypothetical protein	X	na/10241663
OsMULE0875	unknown protein	X	na/13324791
OsMULE0879	intergenic region	N	16756247/na
OsMULE0880	unknown protein	N	17155067 / OSJNBb0060I05.1
OsMULE0880	unknown protein	N	17155067 / OSJNBb0060I05.1
OsMULE0882	hypothetical protein	X	na/7106514
OsMULE0883	intergenic region	N	11602826/na
OsMULE0883	hypothetical protein	X	na/19571110
OsMULE1001	unannotated	N	13184944/na
OsMULE1001	unannotated	N	13184944/na
OsMULE1004	hypothetical protein	X	na/10140787
OsMULE1008	hypothetical protein	X	na/20804502
OsMULE1010	hypothetical protein	X	na/6815078
OsMULE1011	hypothetical protein	X	na/6815078
OsMULE1012	putative GDP-L-fucose synthetase	X	na/20465283
OsMULE1013	unannotated	N	19110518/na
OsMULE1014	hypothetical protein	N	17298574 / OSJNBa0095C06.3
OsMULE1014	hypothetical protein	N	17298574 / OSJNBa0095C06.3
OsMULE1014	hypothetical protein	N	17298574 / OSJNBa0095C06.3
OsMULE1015	putative dihydroorotase	N	20975281 / P0481E12.10

19508	19823	316	1059	1380	322	exon
18996	19167	172	476	647	172	exon
54647	54785	139	704	842	139	na
19508	19823	316	1059	1380	322	exon
18996	19167	172	476	647	172	exon
54647	54785	139	704	842	139	na
116884	117358	475	513	969	457	exon
116885	117140	256	244	499	256	exon
33	141	109	327	863	537	exon
1	286	286	338	1493	1156	ORF
147013	147346	334	646	985	340	na
34840	34977	138	1291	1427	137	exon & intron
35121	35248	128	1445	1566	122	exon*
133	273	141	559	981	423	exon
133037	133181	145	1143	1287	145	na
279	559	281	173	1232	1060	exon
63207	63478	272	354	625	272	na
63539	63721	183	628	809	182	na
84	170	87	524	878	355	exon
4	234	231	3158	3850	693	exon
79	311	233	478	1190	713	exon
79	311	233	478	1190	713	exon
5	144	140	5476	5881	406	exon
14991	15220	230	226	575	350	na
12387	12625	239	204	427	224	exon
13859	13992	134	489	631	143	intron
12680	12768	89	643	731	89	exon
31600	31927	328	573	885	313	exon & intron
121	173	53	701	859	159	exon
36	152	117	834	1184	351	exon

OsMULE1016	hypothetical protein	X	na/10140676
OsMULE1018	hypothetical protein	X	na/8468033
OsMULE1021	hypothetical protein	X	na/10440627
OsMULE1022	hypothetical protein	X	na/7630256
OsMULE1023	hypothetical protein	X	na/7630256
OsMULE1024	hypothetical protein	X	na/7630256
OsMULE1025	hypothetical protein	X	na/5091528
OsMULE1026	hypothetical protein	X	na/5091528
OsMULE1028	putative amino acid transporter	X	na/14029035
OsMULE1029	intergenic region	N	16904698/na
OsMULE1029	intergenic region	N	16904698/na
OsMULE1029	hypothetical protein	X	na/18542926
OsMULE1032	hypothetical protein	X	na/7340909
OsMULE1035	hypothetical protein	X	na/7523501
OsMULE1036	hypothetical protein	X	na/14018070
OsMULE11b01	hypothetical protein	X	na/12583791
OsMULE11b02	hypothetical protein	X	na/9988430
OsMULE11b03	hypothetical protein	N	6069643 / 6069651
OsMULE11b04	hypothetical protein	X	na/9945048
OsMULE11b06	intergenic region	N	18461245/na
OsMULE11b06	hypothetical protein	X	na/10716612
OsMULE1201	unknown protein	N	15290074 / P0004A09.25
OsMULE1202	hypothetical protein	X	na/9558516
OsMULE1203	putative chalcone synthase	X	na/13310890
OsMULE1302	hypothetical protein	X	na/15528559
OsMULE1306	hypothetical protein	X	na/6907104
OsMULE13b22	hypothetical protein	X	na/10140663
OsMULE1401	hypothetical protein	N	13899391 / OSJNBa0045C13.17
OsMULE1401	hypothetical protein	X	na/15289793
OsMULE1402	hypothetical protein	X	na/10140683

4	123	120	671	1030	360	exon
99	256	158	893	1366	474	exon
99	256	158	893	1366	474	exon
99	150	52	61	213	153	exon
2	151	150	295	744	450	ORF
2	151	150	295	744	450	ORF
110	214	105	272	586	315	exon
70947	71092	146	176	331	156	na
70947	71081	135	873	988	116	na
1	137	137	376	786	411	ORF
99	256	158	893	1366	474	exon
1	126	126	384	761	378	ORF
1	266	266	284	1081	798	ORF
1	91	91	268	635	368	ORF
97	281	185	177	731	555	exon
34444	34545	102	602	702	101	exon
1	138	138	58	471	414	ORF
141069	141271	203	326	528	203	na
1	90	90	186	727	542	exon
122459	122769	311	265	559	295	exon
414	590	177	286	816	531	exon
18	193	176	193	928	736	exon
6	150	145	176	646	471	exon
1	64	64	1054	1245	192	exon
508	620	113	237	574	338	exon
137529	137673	145	1007	1150	144	exon & intron
164	372	209	414	1567	1154	exon
5	164	160	318	1019	702	exon
170	451	282	475	1379	905	exon
1	121	121	313	982	670	ORF

OsMULE1403	putative receptor-like protein kinase	X	na/13129438
OsMULE1404	hypothetical protein	X	na/9711823
OsMULE1405	hypothetical protein	X	na/9711823
OsMULE1407	hypothetical protein	X	na/14018048
OsMULE1715	unannotated segment	N	9795249/na
OsMULE1716	unannotated segment	N	9795249/na
OsMULE1720	hypothetical protein	X	na/12583816
OsMULE1721	hypothetical protein	X	na/20042967
OsMULE1722	hypothetical protein	X	na/12583816
OsMULE1723	hypothetical protein	X	na/20042967
OsMULE1901	unknown protein	X	na/14029022
OsMULE1902	hypothetical protein	N	10140737 / OSJNBa0055P24.1
OsMULE1904	putative cytokinin-regulated kinase 1	X	na/13940607
OsMULE1905	hypothetical protein	X	na/11034664
OsMULE1906	ammonium transporter	X	na/7140936
OsMULE1907	hypothetical protein	X	na/21740560
OsMULE1908	intergenic region	N	21686922/na
OsMULE1910	putative receptor-like protein kinase	X	na/13129438
OsMULE1913	putative wall-associated kinase	N	15408719 / P0443D08.12
OsMULE1914	hypothetical protein	X	na/12382005
OsMULE1915	unannotated	N	21741122/na
OsMULE1916	hypothetical protein	X	na/13992693
OsMULE1917	putative receptor kinase	N	13249436 / OSJNBb0033N16.4
OsMULE1919	hypothetical protein	X	na/13486701
OsMULE1920	serine/threonine protein kinase	X	na/14009296
OsMULE1922	intergenic region	N	12331454/na
OsMULE1922	intergenic region	N	12331454/na
OsMULE1923	hypothetical protein	X	na/12328572
OsMULE1924	hypothetical protein	X	na/12328572
OsMULE1925	basic zipper bZIP mRNA	N	na/13124870

1	121	121	313	982	670	ORF
1	295	295	189	1128	940	ORF
36514	36656	143	138	286	149	na
36514	36656	143	138	286	149	na
1	185	185	408	1210	803	exon
111	338	228	111	1187	1077	exon
1	190	190	391	1211	821	ORF
111	338	228	111	1187	1077	exon
1	216	216	155	802	648	ORF
2685	2844	160	347	505	159	exon
1	148	148	2763	3206	444	ORF
109	238	130	587	1073	487	exon
340	473	134	287	702	416	ORF
256	340	85	319	558	240	exon
61254	61359	106	431	545	115	na
170	451	282	287	1191	905	exon
41740	41905	166	783	946	164	intron
1	116	116	190	537	348	exon
3840	3978	139	453	594	142	na
1	116	116	289	725	437	exon
97850	97975	126	6336	6461	126	exon
4	137	134	633	1034	402	exon
334	435	102	567	890	324	exon
49322	49513	192	1179	1375	197	na
50045	50117	73	1848	1920	73	na
91	156	66	613	810	198	exon
91	156	66	613	810	198	exon
278	524	247	603	835	233	exon
30264	30371	108	607	714	108	exon
1	149	149	506	1045	540	ORF

OsMULE1927	hypothetical protein	N	20804776 / P0704D04.8
OsMULE1928	hypothetical protein	X	na/13486805
OsMULE1929	putatitve ethylene-responsive protein	N	18056688 / OSJNBa0046L02.5
OsMULE1929	putatitve ethylene-responsive protein	N	18056688 / OSJNBa0046L02.5
OsMULE1930	hypothetical protein	X	na/8468017
OsMULE19b01	putative senescence-associated protein 15	N	15281152 / OSJNBa0036D19.6
OsMULE19b01	hypothetical protein	X	na/21104861
OsMULE19b02	unknown protein	N	17026065 / P0406G08.30
OsMULE19b02	unknown protein	N	17026065 / P0406G08.30
OsMULE19b03	unknown protein	N	17026065 / P0406G08.30
OsMULE19b03	unknown protein	N	17026065 / P0406G08.30
OsMULE19b05	putative apyrase	X	na/16905183
OsMULE19c01	intergenic region	N	21693905/na
OsMULE19c01	intergenic region	N	21693905/na
OsMULE19c01	hypothetical protein	X	na/9386778
OsMULE19c02	hypothetical protein	X	na/4495227
OsMULE19c03	hypothetical protein	N	20160679 / P0468B07.8
OsMULE19c03	hypothetical protein	N	10241657 / H0711G06.2
OsMULE19c03	hypothetical protein	N	10241657 / H0711G06.2
OsMULE19c04	hypothetical protein	X	na/6498458
OsMULE19c05	hypothetical protein	X	na/6498458
OsMULE19c06	hypothetical protein	N	20270142 / OSJNBa0026L12.30
OsMULE19c06	hypothetical protein	N	20270142 / OSJNBa0026L12.30
OsMULE19c07	unknown protein	X	na/21327956
OsMULE19c08	putative histone deacetylase	N	14588673 / P0712E02.21
OsMULE19c08	putative histone deacetylase	N	14588673 / P0712E02.21
OsMULE19c09	putative histone deacetylase	N	14588673 / P0712E02.21
OsMULE19c09	putative histone deacetylase	N	14588673 / P0712E02.21
OsMULE19c12	hypothetical protein	X	na/11761080
OsMULE19d01	hypothetical protein	X	na/13310888

46754	46970	217	739	959	221	exon
46591	46690	100	964	1063	100	exon
161	236	76	631	858	228	exon
71819	71970	152	878	1029	152	exon
1	247	247	277	1525	1249	ORF
139625	139847	223	678	885	208	exon
139257	139304	48	929	976	48	exon
139625	139847	223	678	885	208	exon
139257	139304	48	929	976	48	exon
1	189	189	524	1247	724	ORF
75251	75329	79	446	524	79	na
75053	75160	108	615	721	107	na
115	225	111	834	1283	450	exon
1	117	117	255	1203	949	exon
44749	44912	164	332	491	160	exon & intron
9345	9618	274	515	791	277	exon
8901	9049	149	1180	1328	149	intron
160	352	193	203	1027	825	exon
160	352	193	203	1027	825	exon
27793	27898	106	363	468	106	exon
27737	27932	196	710	893	184	exon
137	288	152	238	803	566	exon
123965	124069	105	235	339	105	exon
123742	123797	56	484	550	67	exon
123965	124069	105	235	339	105	exon
123742	123797	56	484	550	67	exon
1	105	105	225	666	442	ORF
1	132	132	237	631	395	exon
1	228	228	195	878	684	exon
1	96	96	277	902	626	exon

OsMULE19f01	hypothetical protein	X	na/13129485
OsMULE19f03	hypothetical protein	X	na/20521293
OsMULE19f04	hypothetical protein	X	na/9711907
OsMULE19f05	hypothetical protein	N	13486822 / OSJNBa0004G10.32
OsMULE19f06	hypothetical protein	X	na/6907092
OsMULE19f08	hypothetical protein	N	14209589 / P0419B01.9
OsMULE19f08	hypothetical protein	N	14209589 / P0419B01.9
OsMULE19f12	putative receptor kinase	X	na/13324792
OsMULE19f13	unknown protein	X	na/20042914
OsMULE19f14	hypothetical protein	N	7630233 / 7630248
OsMULE19f14	hypothetical protein	N	7630233 / 7630248
OsMULE19f14	unknown protein	X	na/20161681
OsMULE19f14	hypothetical protein	X	na/21104534
OsMULE19f15	hypothetical protein	X	na/11862974
OsMULE2001	hypothetical protein	X	na/9558461
OsMULE2101	hypothetical protein	X	na/14587342
OsMULE2102	putative non-phototropic hypocotyl	X	na/12322729
OsMULE2103	hypothetical protein	X	na/15623914
OsMULE2302	intergenic region	N	21624134/na
OsMULE2302	intergenic region	N	16197550/na
OsMULE2303	intergenic region	N	21624134/na
OsMULE2304	intergenic region	N	21624134/na
OsMULE2305	intergenic region	N	21624134/na
OsMULE2306	intergenic region	N	21624134/na
OsMULE2307	intergenic region	N	21624134/na
OsMULE2310	intergenic region	N	21624134/na
OsMULE2311	intergenic region	N	21624134/na
OsMULE2316	intergenic region	N	21624134/na
OsMULE2318	intergenic region	N	21624134/na
OsMULE2320	intergenic region	N	21624134/na

1	133	133	247	644	398	exon
130659	130898	240	354	595	242	exon
1	152	152	194	718	525	ORF
56366	56768	403	271	673	403	exon
57160	57272	113	674	783	110	exon
71	231	161	415	906	492	exon
1	114	114	185	526	342	exon
88758	89189	432	4096	4522	427	exon & intron
88520	88707	188	242	439	198	exon
1	54	54	3197	3358	162	exon
102	540	439	911	2314	1404	exon
11	144	134	317	762	446	exon
57	136	80	3839	4280	442	exon
3	149	147	297	759	463	exon
293	496	204	292	900	609	exon
104	266	163	302	790	489	exon
116020	116311	292	428	708	281	na
83894	84338	445	729	1169	441	na
116020	116311	292	8444	8724	281	na
116020	116311	292	8418	8698	281	na
116020	116311	292	8326	8606	281	na
116020	116311	292	431	711	281	na
116020	116228	209	7156	7353	198	na
116020	116228	209	6006	6203	198	na
116020	116228	209	503	700	198	na
116020	116311	292	6094	6374	281	na
116020	116311	292	6094	6374	281	na
116020	116228	209	6872	7069	198	na
116020	116228	209	6872	7069	198	na
116020	116311	292	3355	3635	281	na

OsMULE2321	intergenic region	N	21624134/na
OsMULE2322	intergenic region	N	21624134/na
OsMULE2323	intergenic region	N	16197550/na
OsMULE2323	intergenic region	N	21624134/na
OsMULE2324	intergenic region	N	21624134/na
OsMULE2325	hypothetical protein	X	na/15289876
OsMULE2326	hypothetical protein	X	na/15289876
OsMULE2327	hypothetical protein	X	na/21742776
OsMULE2328	hypothetical protein	X	na/18844797
OsMULE2343	intergenic region	N	13603463/na
OsMULE2407	intergenic region	N	21624301/na
OsMULE2407	hypothetical protein	X	na/13873035
OsMULE2408	phosphate transporter mRNA	N	na/15983298
OsMULE2412	putative protein kinase	X	na/15128440
OsMULE2413	hypothetical protein	N	9049451 / P0710E05.11
OsMULE2413	hypothetical protein	N	9049451 / P0710E05.11
OsMULE2414	hypothetical protein	N	9049451 / P0710E05.11
OsMULE2414	hypothetical protein	N	9049451 / P0710E05.11
OsMULE2415	hypothetical protein	N	9049451 / P0710E05.11
OsMULE2415	hypothetical protein	N	9049451 / P0710E05.11
OsMULE3201	putative transmembrane protein (MtN3)	N	15408754 / P0454A11.16
OsMULE3202	unannotated	N	21742177/na
OsMULE3203	hypothetical protein	X	na/20279462
OsMULE3204	putative anthocyanin 5-aromatic acyltransferase	X	na/21671911
OsMULE3205	hypothetical protein	X	na/7339718
OsMULE3206	hypothetical protein	X	na/20160945
OsMULE3207	hypothetical protein	X	na/15289987
OsMULE3208	hypothetical protein	X	na/13486640
OsMULE3209	hypothetical protein	X	na/8468018
OsMULE3211	hypothetical protein	X	na/20161435

83894	84115	222	2712	2933	222	na
116020	116311	292	2955	3235	281	na
116020	116228	209	528	725	198	na
2	82	81	7118	7426	309	exon
2	82	81	7118	7426	309	exon
255	404	150	1359	1884	526	exon
6	75	70	1584	1800	217	exon
87409	87451	43	1132	1174	43	na
102887	103478	592	293	900	608	na
3	110	108	1007	1330	324	exon
1262	1549	288	975	1265	291	exon
1	368	368	189	1349	1161	ORF
71347	71687	341	774	1132	359	exon
71017	71130	114	469	581	113	exon
71347	71687	341	774	1132	359	exon
71017	71130	114	469	581	113	exon
71347	71687	341	774	1132	359	exon
71017	71130	114	469	581	113	exon
81705	82147	443	562	1005	444	exon & intron
41154	41407	254	535	795	261	na
2	129	128	426	887	462	exon
3	100	98	519	833	315	exon
1	119	119	594	950	357	exon
6	141	136	432	836	405	exon
22	149	128	418	801	384	exon
1	257	257	33	931	899	exon
313	651	339	327	1848	1522	exon
193	273	81	804	1043	240	exon
127731	127830	100	828	927	100	na
1	101	101	744	1079	336	exon

OsMULE3212	intergenic region	N	6539551/na
OsMULE3214	hypothetical protein	X	na/14209528
OsMULE3214	hypothetical protein	X	na/15451590
OsMULE3215	putative asparaginyl-tRNA synthetase	N	14587202 / B1045D11.4
OsMULE3215	hypothetical protein	X	na/14626293
OsMULE3217	putative phosphate transporters & hypothetical protein	X	na/15217301
OsMULE3218	putative SCARECROW1 protein	N	6983854 / 1497986
OsMULE3218	putative SCARECROW1 protein	N	6983854 / 1497986
OsMULE3218	hypothetical protein	N	13702835 / OSJNBb0011A08.3
OsMULE3219	intergenic region	N	15144390/na
OsMULE3219	hypothetical protein	N	20279375 / OJ1175C11.13
OsMULE3220	intergenic region	N	14141756/na
OsMULE3220	intergenic region	N	14141756/na
OsMULE3220	hypothetical protein	X	na/20161247
OsMULE3220	hypothetical protein	X	na/20161248
OsMULE3221	hypothetical protein	X	na/21741788
OsMULE3224	hypothetical protein	X	na/8467991
OsMULE3225	hypothetical protein	X	na/8467991
OsMULE3226	hypothetical protein	X	na/10140645
OsMULE3227	hypothetical protein	X	na/8467991
OsMULE3228	homeodomain-leucine zipper transcription factor	N	na/6635776
OsMULE3228	homeodomain-leucine zipper transcription factor	N	na/6635776
OsMULE3229	putative receptor protein kinase	X	na/18642676
OsMULE3230	hypothetical protein	X	na/10934030
OsMULE3231	hypothetical protein	X	na/10934030
OsMULE3232	putative iron-phytosiderophore transporter	X	na/15624064
OsMULE3233	hypothetical protein	N	13872872 / P0434B04.23
OsMULE3233	unannotated segment	N	7363412/na
OsMULE3234	hypothetical protein	X	na/12643043
OsMULE3236	hypothetical protein	X	na/7339718

153	282	130	267	545	279	exon
19959	20077	119	886	1004	119	exon*
80	139	60	1292	1471	180	exon
1	169	169	437	1031	595	exon
173885	174034	150	454	601	148	exon
174425	174499	75	619	690	72	exon
9491	9888	398	6108	6513	406	exon
26904	27087	184	1002	1189	188	na
66829	67009	181	443	622	180	exon
54594	54826	233	860	1093	234	na
54435	54542	108	745	852	108	na
2	159	158	245	718	474	exon
1	133	133	1323	1773	451	exon
26	240	215	403	1044	642	exon
110	275	166	402	872	471	exon
110	275	166	402	872	471	exon
1	102	102	823	1167	345	exon
122	291	170	346	855	510	exon
2029	2261	233	865	1093	229	exon & intron
2267	2469	203	450	621	172	exon & intron
1	65	65	494	688	195	exon
8	169	162	349	1040	692	exon
8	169	162	349	1040	692	exon
357	516	160	481	960	480	exon
112684	112850	167	635	801	167	exon & intron
11173	11347	175	703	875	173	na
1	217	217	238	1336	1099	exon
1	193	193	256	950	695	na
104	248	145	372	806	435	na
104	248	145	364	798	435	na

OsMULE3237	hypothetical protein	X	na/14091837
OsMULE3238	hypothetical protein	X	na/14091837
OsMULE3239	hypothetical protein	X	na/15528605
OsMULE3240	hypothetical protein	X	na/15289877
OsMULE3242	putative kinase	X	na/20303574
OsMULE3243	putative l-asparaginase	N	10241423 / H0212B02.5
OsMULE3243	putative l-asparaginase	N	10241423 / H0212B02.5
OsMULE3244	putative UDP-glucuronyltransferase-like protein	N	10241423 / H0212B02.8
OsMULE3244	putative UDP-glucuronyltransferase-like protein	N	10241423 / H0212B02.8
OsMULE3245	hypothetical protein	X	na/10122050
OsMULE3246	Arabidopsis thaliana putative B' regulatory subunit of PP2A	X	na/21280992
OsMULE3248	hypothetical protein	X	na/21740795
OsMULE3249	hypothetical protein	X	na/13873038
OsMULE3250	hypothetical protein	X	na/13161458
OsMULE3250	hypothetical protein	X	na/13161460
OsMULE3251	hypothetical protein	X	na/13161458
OsMULE3251	hypothetical protein	X	na/13161460
OsMULE3252	putative iron-phytosiderophore transporter (yellow stripe 1 mutant)	X	na/20160640
OsMULE3253	mitochondrial alternative oxidase 1c	N	na/16902305
OsMULE3253	mitochondrial alternative oxidase 1c	N	na/16902305
OsMULE3253	putative cytochrome P450	N	20270142 / OSJNBa0026L12.14
OsMULE3254	hypothetical protein	X	na/15528765
OsMULE3255	ethylene-insensitive-3-like protein mRNA	N	na/17221606
OsMULE3255	hypothetical protein	X	na/20804467
OsMULE3257	hypothetical protein	X	na/19571080
OsMULE3258	putative cytochrome P450	N	20270142 / OSJNBa0026L12.14
OsMULE3258	mitochondrial alternative oxidase 1c	N	na/16902305
OsMULE3259	hypothetical protein	N	10716598 / OSJNBa0079L16.22
OsMULE3259	intergenic region	N	10716598 / na
OsMULE3261	intergenic region	N	na/13384337

2	160	159	284	798	515	na
253	418	166	32	1371	1340	na
1	234	234	333	1034	702	ORF
30184	30548	365	341	666	326	exon
30735	30875	141	766	905	140	exon & intron
54927	55074	148	921	1065	145	exon
55428	55606	179	1064	1242	179	exon & intron
1	99	99	370	666	297	exon
176	388	213	346	949	604	exon
1	197	197	333	989	657	ORF
75	328	254	382	1296	915	exon
1	209	209	280	1689	1410	exon
55	118	64	4788	4979	192	exon
1	209	209	280	1689	1410	exon
55	118	64	4788	4979	192	exon
357	516	160	481	960	480	exon
1580	1977	398	624	1057	434	exon & intron
1358	1583	226	1957	2182	226	exon
84831	85019	189	411	599	189	exon
1	124	124	721	1092	372	exon
1361	1470	110	1409	1522	114	exon
1	274	274	388	1404	1017	exon
1	173	173	419	1247	829	ORF
84831	85026	196	403	598	196	exon
1358	1902	545	729	1271	543	exon & intron
109990	110195	206	620	820	201	exon & intron*
110294	110443	150	835	984	150	na
49701	50108	408	235	640	406	na
49701	50108	408	235	640	406	na
7563	8404	842	509	1335	827	na

OsMULE3263	intergenic region	N	na/13384337
OsMULE3265	intergenic region	N	na/13346556
OsMULE3267	hypothetical protein	X	na/7106516
OsMULE3267	hypothetical protein	X	na/7106516
OsMULE3268	hypothetical protein	X	na/18921320
OsMULE3601	putative far-red impaired response protein	X	na/20804794
OsMULE3701	hypothetical protein	X	na/5777623
OsMULE3702	hypothetical protein	X	na/8096412
OsMULE3703	hypothetical protein	X	na/8096412
OsMULE3704	hypothetical protein	X	na/8096412
OsMULE3705	hypothetical protein	X	na/15451636
OsMULE3706	hypothetical protein	X	na/15451636
OsMULE3707	hypothetical protein	X	na/15451636
OsMULE3708	hypothetical protein	X	na/10241435
OsMULE3709	hypothetical protein	X	na/10241435
OsMULE3710	hypothetical protein	X	na/13366223
OsMULE3712	hypothetical protein	X	na/10716612
OsMULE3713	putative amino acid transport protein	X	na/20161442
OsMULE3714	putative protoporphyrinogen oxidase	X	na/10440619
OsMULE3715	protoporphyrinogen oxidase I mRNA	N	na/13429991
OsMULE3716	hypothetical protein	X	na/14018072
OsMULE3717	hypothetical protein	X	na/14018072
OsMULE3718	hypothetical protein	X	na/11761074
OsMULE3719	hypothetical protein	X	na/21901977
OsMULE3719	hypothetical protein	X	na/21901978
OsMULE3720	hypothetical protein	X	na/20804786
OsMULE3722	hypothetical protein	X	na/13366223
OsMULE3724	homeodomain-leucine zipper transcription factor	N	na/6635776
OsMULE3724	unannotated	N	17226685/na
OsMULE3725	homeodomain-leucine zipper transcription factor	N	na/6635776

203	344	142	315	771	457	exon
364	507	144	2090	2518	429	exon
1	129	129	674	1060	387	exon
73	484	412	597	1811	1215	exon
1	390	390	457	1846	1390	exon
111	372	262	263	1351	1089	exon
110	372	263	388	1654	1267	exon
87	372	286	300	1501	1202	exon
1	395	395	205	1528	1324	ORF
1	395	395	208	1200	993	ORF
7	279	273	275	1359	1085	exon
175	275	101	281	580	300	exon
175	275	101	281	580	300	exon
1	113	113	384	689	306	exon
101	150	50	260	409	150	exon
164	400	237	636	1364	729	exon
98	267	170	488	1087	600	exon
490	618	129	700	828	129	exon
128	420	293	238	1116	879	exon
128	420	293	238	1116	879	exon
3	218	216	211	977	767	exon
24	121	98	267	560	294	exon
1	153	153	669	1176	508	ORF
18	201	184	173	1285	1113	exon
1	113	113	384	689	306	exon
1751	2051	301	1059	1372	314	exon & intron
61192	61305	114	776	881	106	na
1751	2051	301	1059	1372	314	exon & intron
61192	61305	114	776	881	106	na
60337	60385	49	1644	1692	49	na

OsMULE3725	unannotated	N	17226685/na
OsMULE3725	unannotated	N	17226685/na
OsMULE3726	putative chalcone synthase	X	na/13310890
OsMULE3727	hypothetical protein	X	na/8468033
OsMULE3727	hypothetical protein	X	na/20804930
OsMULE3728	unknown protein	X	na/13236652
OsMULE3729	hypothetical protein	X	na/19571645
OsMULE3730	protoporphyrinogen oxidase I mRNA	N	na/13429991
OsMULE3731	hypothetical protein	X	na/17385712
OsMULE3732	unknown protein	X	na/12643036
OsMULE3733	unknown protein	X	na/12643036
OsMULE3734	hypothetical protein	X	na/14140282
OsMULE3735	hypothetical protein	X	na/12313688
OsMULE3736	hypothetical protein	X	na/10140724
OsMULE3737	putative protoporphyrinogen oxidase	X	na/10440619
OsMULE3901	hypothetical protein	X	na/11320835
OsMULE3902	hypothetical protein	X	na/11320835
OsMULE3903	hypothetical protein	X	na/11320835
OsMULE3904	hypothetical protein	X	na/21741706
OsMULE3905	hypothetical protein	X	na/20160460
OsMULE3905	vegetative cell wall protein gp1	X	na/12018147
OsMULE4001	intergenic region	N	13384337/na
OsMULE4002	intergenic region	N	13384337/na
OsMULE4003	intergenic region	N	13384337/na
OsMULE4007	hypothetical protein	X	na/20087083
OsMULE4008	hypothetical protein	X	na/20087083
OsMULE4039	Anopheles gambiae hypothetical protein	X	na/21289881
OsMULE4042	hypothetical protein	X	na/4680493
OsMULE4042	hypothetical protein	X	na/4680503
OsMULE4042	hypothetical protein	X	na/4680494

209	466	258	307	1086	780	exon
171	237	67	261	461	201	exon
19	85	67	495	680	186	exon
139	279	141	290	712	423	exon
182	503	322	329	1315	987	exon
437	610	174	882	1055	174	exon
1	53	53	721	879	159	exon
2	134	133	432	830	399	ORF
2	134	133	432	830	399	ORF
1	192	192	221	796	576	ORF
1	224	224	276	1082	807	ORF
1	82	82	642	986	345	exon
87	370	284	277	1176	900	exon
1	96	96	247	534	288	exon
1	96	96	247	534	288	exon
1	96	96	247	534	288	exon
74	175	102	1367	1677	311	exon
1	75	75	143	425	283	exon
70	193	124	2751	3143	393	exon
115797	115936	140	441	580	140	na
115797	115936	140	441	580	140	na
115797	115936	140	441	580	140	na
81	162	82	3773	4024	252	exon
81	162	82	1482	1733	252	exon
3	58	56	265	432	168	exon
1	252	252	1285	2042	758	ORF
1	154	154	4029	5331	1303	ORF
10	146	137	7038	7515	478	exon
201	252	52	2591	2746	156	exon
1	252	252	1245	2000	756	ORF

OsMULE4042	hypothetical protein	X	na/4680344
OsMULE4043	hypothetical protein	X	na/4680493
OsMULE4043	hypothetical protein	X	na/4680503
OsMULE4043	Drosophila melanogaster hypothetical protein	X	na/7293084
OsMULE4043	hypothetical protein	X	na/4680344
OsMULE4043	hypothetical protein	X	na/4680494
OsMULE4044	hypothetical protein	X	na/4680494
OsMULE4045	hypothetical protein	X	na/20161302
OsMULE4301	hypothetical protein	X	na/10140700
OsMULE4452	hypothetical protein	N	13486660 / P0028E10.6
OsMULE4452	hypothetical protein	N	13486660 / P0028E10.6
OsMULE4902	hypothetical protein	X	na/21743362
OsMULE4910	hypothetical protein	X	na/21743362
OsMULE4911	hypothetical protein	X	na/21743362
OsMULE5102	hypothetical protein	X	na/5091602

[+] certain genes denoted as "clone GI# / locus"

1	154	154	4029	5331	1303	ORF
6	121	116	5343	5690	348	exon
201	252	52	2550	2705	156	exon
10	146	137	7281	7758	478	exon
27	131	105	1694	2022	329	exon
1	166	166	321	900	580	exon
219	356	138	451	921	471	exon
25500	26186	687	379	1075	697	exon & intron
25313	25431	119	1095	1213	119	exon & intron
80	121	42	94	219	126	exon
80	121	42	127	252	126	exon
80	121	42	127	252	126	exon
1	83	83	62	310	249	exon

Supplementary Table 5.4A MULE-contained *Arabidopsis* ORFs

MULE	gene	strand	gene position within MULE		size (bp)	CD/PSI BLAST search*	expresssion
			start	end			
AtMULE004	At1g08740	-	11655	7217	4439	dUlp/hypothetical protein	na
AtMULE010	At1g17900	-	1729	967	763	na/hypothetical protein	na
AtMULE015	At1g21020	+	611	3671	3061	na/hypothetical protein	na
AtMULE015	At1g21030	-	4248	5052	805	dUlp/hypothetical protein	na
AtMULE019	At1g23930	-	7977	5599	2379	domain of unknown function/hypothetical protein	na
AtMULE020	At1g27780	+	597	6506	5910	peptidase-C48/hypothetical protein	na
AtMULE020	At1g27790	+	7205	7567	363	na/hypothetical protein	na
AtMULE020	At1g27800	+	8362	8997	636	na/hypothetical protein	na
AtMULE020	At1g27810	-	10540	9592	949	na/hypothetical protein	na
AtMULE023	At1g29620	+	253	1907	1655	na/hypothetical protein	na
AtMULE028	At1g32720	+	253	1982	1730	na/putative protein phosphatase 2C	na
AtMULE029	At1g33450	+	367	919	553	na/MtN20	na
AtMULE032	At1g34610	+	1125	4844	3720	peptidase-C48/hypothetical protein	na
AtMULE102	At1g34710	+	4243	5191	949	domain of unknown function/hypothetical protein	na
AtMULE102	At1g34720	-	5340	5870	763	na/hypothetical protein	na
AtMULE102	At1g34730	+	8737	14702	5966	na/hypothetical protein	na
AtMULE102	At1g34740	-	8736	4843	3894	peptidase-C48/hypothetical protein	na
AtMULE034	At1g35070	-	6680	4843	1838	na/hypothetical protein	na
AtMULE034	At1g35080	+	7853	9930	2078	na/hypothetical protein	na
AtMULE034	At1g35090	-	11609	10601	1009	domain of unknown function/hypothetical protein	na
AtMULE034	At1g35100	+	12280	13067	788	na/hypothetical protein	na
AtMULE040	At1g35650	-	18827	14363	4465	dUlp/hypothetical protein	na
AtMULE103	At1g35770	-	11916	5804	6113	peptidase-C48/hypothetical protein	na
AtMULE054	At1g44850	+	4404	6581	2178	na/replication protein - like	na

AtMULE054	At1g44860	+	7468	8482	1015	domain of unknown function/hypothetical protein	na
AtMULE054	At1g44870	-	9601	9116	486	na/replication protein - like	na
AtMULE054	At1g44880	-	17206	11550	5657	peptidase-C48/hypothetical protein	na
AtMULE055	At1g45080	+	4172	5117	946	na/lumen protein retaining receptor	na
AtMULE055	At1g45090	-	14498	9478	5021	peptidase-C48/hypothetical protein	na
AtMULE060	At1g48250	+	194	1250	1057	na/hypothetical protein	na
AtMULE063	At1g49680	+	458	757	300	na/hypothetical protein	na
AtMULE070	At1g52010	+	743	3080	2338	na/hypothetical protein	na
AtMULE070	At1g52020	+	3036	8807	5772	peptidase-C48/hypothetical protein	na
AtMULE072	At1g52090	-	3854	3510	345	na/hypothetical protein	na
AtMULE077	At1g55400	-	1735	1003	733	na/hypothetical protein	na
AtMULE081	At1g61200	-	737	350	388	na/Athb-1 protein	na
AtMULE126	At1g61920	+	164	475	312	na/hypothetical protein	na
AtMULE138	At2g01310	-	627	466	162	na/na	na
AtMULE143	At2g03570	-	755	327	429	na/hypothetical protein	na
AtMULE204	At2g05480	-	2321	1327	995	domain of unknown function/hypothetical protein	na
AtMULE205	At2g05560	+	508	8682	8175	peptidase-C48/hypothetical protein	yes
AtMULE153	At2g06860	+	479	6568	6090	peptidase-C48/hypothetical protein	na
AtMULE155	At2g07260	-	7146	5104	2043	na/adenylate kinase -like protein	na
AtMULE158	At2g07510	-	8685	6071	2615	na/ putative lumen protein retaining receptor	na
AtMULE159	At2g10350	+	2436	7422	4987	peptidase-C48/hypothetical protein	na
AtMULE159	At2g10360	+	9025	9516	492	na/hypothetical protein	na
AtMULE159	At2g10370	+	10150	11164	1015	na/replication protein	na
AtMULE159	At2g10380	+	12053	13002	950	na/hypothetical protein	na
AtMULE159	At2g10390	+	13459	14135	677	na/replication protein	na
AtMULE252	At2g12100	+	439	8059	7621	peptidase-C48/hypothetical protein	na
AtMULE252	At2g12110	+	6482	4843	1640	peptidase-C48/hypothetical protein	na
AtMULE252	At2g12120	+	7364	8059	696	na/hypothetical protein	na
AtMULE252	At2g12130	+	8732	9361	630	na/hypothetical protein	na
AtMULE252	At2g12140	-	10724	9780	945	na/replication protein - like	na

AtMULE167	At2g12690	-	7099	6562	538	na/hypothetical protein	na
AtMULE170	At2g13580	+	1577	2885	1309	2-oxoacid-dh/2-oxoglutarate dehydrogenase E2 subunit	na
AtMULE176	At2g14010	+	2000	6705	4706	dUlp/hypothetical protein	na
AtMULE176	At2g14020	+	7466	7951	486	na/hypothetical protein	na
AtMULE177	At2g14130	+	535	5180	4646	peptidase-C48/hypothetical protein	na
AtMULE177	At2g14140	+	6913	11222	4310	na/hypothetical protein	na
AtMULE178	At2g14270	+	1598	2064	467	na/hypothetical protein	na
AtMULE136	At2g14770	+	501	9548	9048	na/protein phosphatase 2C	na
AtMULE136	At2g14780	-	5189	539	4651	domain of unknown function/hypothetical protein	na
AtMULE180	At2g15190	+	539	5189	4651	na/replication protein - like	na
AtMULE180	At2g15200	+	6941	10873	3933	na/hypothetical protein	na
AtMULE231	At2g15420	-	11086	6853	4234	na/SMC2-like condensin	na
AtMULE182	At2g15800	+	556	1353	798	na/hypothetical protein	na
AtMULE251	At2g16160	+	4164	5064	901	domain of unknown function/hypothetical protein	na
AtMULE251	At2g16180	-	6061	11034	4974	peptidase-C48/hypothetical protein	yes
AtMULE186	At2g16330	+	1957	2886	930	na/replication protein - like	na
AtMULE233	At2g16830	+	3302	3497	196	na/plasma membrane intrinsic protein	na
AtMULE632	AT3g05850	-	1	2406	2406	PB1-Mutator/hypothetical protein	yes
AtMULE631	AT3g06940	+	1	3117	3117	PB1-Mutator/hypothetical protein	yes
AtMULE265	AT3g24370	-	6374	5811	564	na/hypothetical protein	na
AtMULE265	AT3g24380	-	7860	7168	693	na/hypothetical protein	na
AtMULE265	AT3g24390	-	14661	9833	4829	peptidase-C48/hypothetical protein	na
AtMULE267	AT3g26530	-	11658	7217	4442	dUlp/hypothetical protein	yes
AtMULE287	AT3g29210	+	3142	4843	1702	na/hypothetical protein	na
AtMULE372	AT3g30470	-	7184	6595	590	na/MtN20	na
AtMULE372	AT3g30480	+	8947	9486	540	na/hypothetical protein	na
AtMULE324	AT3g32389	+	5040	4843	198	peptidase-C48/hypothetical protein	na
AtMULE326	AT3g33010	+	5040	4843	198	na/hypothetical protein	na
AtMULE339	AT3g43390	+	6232	4843	1390	na/putative lumen protein retaining receptor	na
AtMULE375	AT3g43780	+	253	1983	1731	na / putative protein phosphatase 2C	na

AtMULE345	AT3g44500	+	538	5202	4665	peptidase-C48/hypothetical protein	na
AtMULE352	AT3g45120	-	1568	1248	321	na/hypothetical protein	na
AtMULE361	AT3g50250	-	733	287	447	na/hypothetical protein	na
AtMULE425	AT4g02320	+	2294	4843	2550	pectinesterase/pectin methylesterase	na
AtMULE382	AT4g03930	+	3590	5555	1966	pectin esterase/pectin esterase	yes
AtMULE382	AT4g03950	+	12686	14279	1594	na / probable glucose-6-phosphate/phosphate-translocator	na
AtMULE390	AT4g05280	+	507	6431	5925	peptidase-C48/hypothetical protein	na
AtMULE390	AT4g05290	+	7151	7792	642	na/hypothetical protein	na
AtMULE390	AT4g05300	-	8596	7702	895	peptidase-S35/hypothetical protein	na
AtMULE442	AT4g07680	+	735	2055	1321	peptidase-C48/hypothetical protein	na
AtMULE442	AT4g07690	+	9270	10283	1014	domain of unknown function/hypothetical protein	na
AtMULE443	AT4g07970	+	7633	4843	2791	na/adenylate kinase -like protein	na
AtMULE398	AT4g08650	+	8567	9621	1055	peptidase-C48/hypothetical protein	na
AtMULE402	AT4g08880	+	1111	7423	6313	peptidase-C48/hypothetical protein	na
AtMULE419	AT4g19270	-	599	133	467	na/hypothetical protein	na
AtMULE470	AT4g35400	+	349	579	231	na/hypothetical protein	na
AtMULE482	AT5g15690	-	3039	1560	1480	na/hypothetical protein	na
AtMULE483	AT5g15990	+	429	921	493	na/hypothetical protein	na
AtMULE556	AT5g20460	+	512	1064	553	na/hypothetical protein	yes
AtMULE494	AT5g26350	-	4602	4073	530	na/hypothetical protein	na
AtMULE500	AT5g28270	-	19901	17520	2382	na/hypothetical protein	na
AtMULE503	AT5g28480	+	458	5413	4956	na/hypothetical protein	na
AtMULE571	AT5g28600	+	480	6122	5643	na/hypothetical protein	na
AtMULE511	AT5g28970	+	1005	8272	7268	peptidase-C48/hypothetical protein	na
AtMULE511	AT5g28980	+	8904	11167	2264	domain of unknown function/hypothetical protein	na
AtMULE511	AT5g28990	+	11842	12483	642	na/hypothetical protein	na
AtMULE512	AT5g33230	+	10193	10685	493	na/hypothetical protein	na
AtMULE518	AT5g36020	+	1009	4136	3128	na/replication protein - like	na
AtMULE518	AT5g36030	+	4732	6211	1480	peptidase-C48/hypothetical protein	na
AtMULE518	AT5g36040	+	10317	11050	734	na/hypothetical protein	na

AtMULE518	AT5g36050	+	11666	12673	1008	na/replication protein - like	na
AtMULE518	AT5g36060	+	13403	13960	558	na/hypothetical protein	na
AtMULE518	AT5g36070	+	14541	15423	883	na/replication protein - like	na
AtMULE521	AT5g36830	+	4241	5082	842	na/replication protein - like	na
AtMULE521	AT5g36840	-	7771	5655	2117	na/hypothetical protein	na
AtMULE521	AT5g36850	-	9433	8738	696	na/hypothetical protein	na
AtMULE521	AT5g36860	-	14677	9799	4879	peptidase-C48/hypothetical protein	na
AtMULE545	AT5g44880	+	6201	6686	486	na/hypothetical protein	na
AtMULE545	AT5g44890	-	11658	7213	4446	dUlp/hypothetical protein	na

na: not applicable.
*dUlp: the putative proteins encoded by *ULP*-derived ORFs.

Supplementary Table 5.4B MULE-contained rice ORFs

MULE	gene	gene position within MULE	size (bp)	strand	CD/PSI Blast	expression
OsMULE0105	MORF0105	115-1272	1158	-	na / hypothetical protein	na
OsMULE0106	MORF0106	123-1280	1158	+	na / hypothetical protein	na
OsMULE0110	MORF0110	464-967	504	-	na / na	na
OsMULE0112	7340920	63-1167	1105	-	na / hypothetical protein	na
OsMULE0120	MORF0120a	120-2395	2276	+	na / hypothetical protein	na
OsMULE0127	MORF0127	76-667	592	+	na / hypothetical protein	na
OsMULE0201	MORF0201a	166-399	234	+	na / na	na
OsMULE0201	MORF0201b	468-905	438	-	na / hypothetical protein	na
OsMULE0301	11320854	537-1053	517	+	na / hypothetical protein	na
OsMULE0305	MORF0305	362-1060	699	-	na / hypothetical protein	na
OsMULE0307	12039282	242-2494	2253	-	na/translation factor EF-1 alpha-like protein	na
OsMULE0309	MORF0309	265-1000	736	-	na / na	na
OsMULE0403	5042460	538-1051	514	+	na / hypothetical protein	na
OsMULE0404	MORF0404	240-1355	1116	-	na / hypothetical protein	na
OsMULE0409	MORF0409	191-1220	1030	+	na / na	na
OsMULE0411	MORF0411	270-1516	1247	-	na / hypothetical protein	na
OsMULE0411	10140730*	830-1429	600	-	na / hypothetical protein	na
OsMULE0414	MORF0414	243-1116	874	-	na / hypothetical protein	na
OsMULE0415	MORF0415	271-1517	1247	-	na / hypothetical protein	na
OsMULE0701	MORF0701	154-1017	864	-	na / na	na
OsMULE0702	MORF0702	1588-2025	438	-	na / hypothetical protein	na
OsMULE0703	MORF0703	336-1693	1358	-	na / hypothetical protein	na
OsMULE0703	6069657+	336-1693	1358	-	na / hypothetical protein	na
OsMULE0709	MORF0709a	1981-2368	388	+	na / na	na
OsMULE0709	12039286	298-728	431	+	na / hypothetical protein	na
OsMULE0709	MORF0709b	1200-1828	629	-	na / hypothetical protein	na

OsMULE0709	12039288*	1200-2313	1114	-	na / hypothetical protein	na
OsMULE0710	MORF0710	294-638	345	+	serine/threonine protein kinases, catalytic domain	na
OsMULE0711	12039278	444-1118	675	-	mitochondrial import inner membrane translocase subunit Tim17 / translocase of inner mitochondrial membrane TIM23 (At)	na
OsMULE0712	MORF0712	92-419	328	-	na / na	na
OsMULE0713	MORF0713	92-419	328	-	na / na	na
OsMULE0714	MORF0714	744-1238	495	+	S-adenosylmethionine synthetase 2, N-terminal and central domain	na
OsMULE0715	MORF0715	831-1208	378	+	na / na	na
OsMULE0716	10140629	1223-1543	321	+	na / hypothetical protein	na
OsMULE0719	14165329	416-1596	1181	+	na / hypothetical protein	na
OsMULE0720	MORF0720	354-1257	904	-	na / na	na
OsMULE0721	MORF0721	354-1257	904	-	na / na	na
OsMULE0722	MORF0722	449-709	261	+	na / hypothetical protein	na
OsMULE0722	11034576+	449-709	261	+	na / hypothetical protein	na
OsMULE0724	MORF0724	383-1064	682	+	na/putative lipase	na
OsMULE0725	MORF0725	383-1064	682	+	na/putative lipase	na
OsMULE0726	MORF0726	32-993	962	-	na/putative protein	na
OsMULE0731	MORF0731	903-1781	879	-	na/cinnamoyl-CoA reductase (Pinus taeda)	na
OsMULE0732	MORF0732	903-1781	879	-	na/cinnamoyl-CoA reductase (Pinus taeda)	na
OsMULE0734	MORF0734	201-1171	971	-	na / na	na
OsMULE0735	MORF0735	189-974	786	-	na / hypothetical protein	na
OsMULE0735	13702831*	741-974	234	-	na / na	na
OsMULE0739	MORF0739	989-1886	898	+	na / hypothetical protein	na
OsMULE0740	MORF0740	91-370	280	-	na / hypothetical protein	na
OsMULE0740	11761118+	91-370	280	-	na / hypothetical protein	na
OsMULE0743	MORF0743	354-1257	904	-	na / na	na
OsMULE0744	MORF0744	354-1257	904	-	na / na	na
OsMULE0801	MORF0801	185-1021	837	+	DNA-binding domain in plant proteins such as APETALA2 and EREBPs / CRT/DRE binding factor	na
OsMULE0802	MORF0802	185-1021	837	+	DNA-binding domain in plant proteins such as APETALA2	na

					and EREBPs / CRT/DRE binding factor	
OsMULE0804	MORF0804	360-850	491	+	na / hypothetical protein	na
OsMULE0806	MORF0806	940-1508	569	-	PTR2, proton-dependent oligopeptide transport family barley	na
OsMULE0806	10140678*	207-1627	1421	-	PTR2, proton-dependent oligopeptide transport family barley	na
OsMULE0807	MORF0807	2904-4476	1573	-	FAR1 & murA / far-red impaired response protein; A mutator-like transposase-like protein; phytochrome signaling protein-like	na
OsMULE0807	13365573*	31-4796	4766	-	FAR1 & murA / far-red impaired response protein; mutator-like transposase-like protein; phytochrome A signaling protein-like	na
OsMULE0808	MORF0808a	955-1401	447	-	na / na	na
OsMULE0808	MORF0808b	401-631	231	-	na / na	yes
OsMULE0809	13486736	177-1339	1163	-	FAD binding domain / berberine bridge enzyme-like protein	na
OsMULE0810	13486752	177-1339	1163	-	FAD binding domain / berberine bridge enzyme-like protein	na
OsMULE0811	MORF0811	431-1505	1075	+	na / hypothetical protein	na
OsMULE0812	10140628	229-873	645	-	na / hypothetical protein	na
OsMULE0814	MORF0814	343-1390	1048	+	na / na	na
OsMULE0814	10140634*	270-602	333	-	na / na	na
OsMULE0814	10140640*	886-1200	315	-	na / hypothetical protein	na
OsMULE0816	MORF0816	522-1668	1147	+	na / hypothetical protein	na
OsMULE0816	7523489+	522-1668	1147	+	na / hypothetical protein	na
OsMULE0817	MORF0817	247-1345	1099	-	na / hypothetical protein	na
OsMULE0817	7523490*	107-1549	1443	-	na / hypothetical protein	na
OsMULE0818	MORF0818	438-1366	929	+	na/purine permease-like protein	na
OsMULE0819	MORF0819	352-795	444	+	na / na	na
OsMULE0823	MORF0823a	1126-1564	439	+	na / na	na
OsMULE0823	MORF0823b	419-940	522	-	thiamine pyrophosphate enzyme, N-terminal TPP binding domain / pyruvate decarboxylase isozyme 3 (PDC)	na
OsMULE0827	MORF0827	405-1108	704	+	na / N-type calcium channel alpha-1B cdB3	yes

					variant (Gallus gallus)	
OsMULE0828	MORF0828	438-1366	929	+	na/purine permease-like protein	na
OsMULE0829	MORF0829	229-615	387	-	na / na	na
OsMULE0829	9049457*	257-1043	787	+	na / hypothetical protein	na
OsMULE0830	MORF0830	437-1597	1161	+	na / putative plant disease resistance polyprotein	na
OsMULE0831	MORF0831	437-1597	1161	+	na / putative plant disease resistance polyprotein	na
OsMULE0832	MORF0832	6267-6839	573	-	na / similar to Oryza sativa serine carboxypeptidase-like protein	na
OsMULE0833	MORF0833	6267-6839	573	-	na / similar to Oryza sativa serine carboxypeptidase-like protein	na
OsMULE0836	MORF0836a	275-1028	754	-	na / similar to putative receptor kinase	na
OsMULE0836	MORF0836b	2832-3861	1030	+	na / na	na
OsMULE0840	MORF0840	797-2687	1891	+	na / hypothetical protein	yes
OsMULE0842	MORF0842	555-1200	646	+	na / na	na
OsMULE0844	MORF0844	17-786	770	+	na / hypothetical protein	na
OsMULE0845	MORF0845	202-894	693	-	na / hypothetical protein	na
OsMULE0845	11034588+	202-894	693	-	na / hypothetical protein	na
OsMULE0846	MORF0846	202-894	693	-	na / hypothetical protein	na
OsMULE0846	20279444	608-2113	1506	-	na/protein kinase Xa21	na
OsMULE0848	MORF0848	359-946	588	+	na / hypothetical protein	na
OsMULE0848	12656809*	359-1451	1093	+	na / hypothetical protein	na
OsMULE0852	MORF0852	767-1111	345	-	na / hypothetical protein	na
OsMULE0852	11034700*	173-701	529	+	na / hypothetical protein	na
OsMULE0861	MORF0861	910-1452	543	+	na / hypothetical protein	na
OsMULE0862	MORF0862a	310-585	276	+	na / hypothetical protein similar to Arabidopsis thaliana	na
OsMULE0862	MORF0862b	612-902	291	-	na / hypothetical protein	na
OsMULE0863	MORF0863	704-1045	342	+	na / na	na
OsMULE0866	MORF0866	431-1643	1213	+	na / vacuolar targeting receptor bp-80 (Triticum aestivum)	na
OsMULE0866	7340913+	431-1643	1213	+	na / vacuolar targeting receptor bp-80 (Triticum aestivum)	na
OsMULE0870	MORF0870a	1246-1506	261	+	na / hypothetical protein	yes
OsMULE0870	MORF0870b	253-648	396	-	na / na	na
OsMULE0871	MORF0871	274-630	357	-	na / hypothetical protein	na

OsMULE0872	MORF0872	274-630	357	-	na / hypothetical protein	na
OsMULE0875	MORF0875	237-1590	1354	-	na / hypothetical protein	yes
OsMULE0875	MORF0875	237-1590	1354	-	na / hypothetical protein	yes
OsMULE0876	MORF0876	275-1033	759	+	na / na	na
OsMULE0877	MORF0877	531-1562	1032	-	na / na	na
OsMULE0880	MORF0880	1152-1535	384	-	na/unknown protein	na
OsMULE0881	MORF0881	98-422	325	-	na / hypothetical protein	na
OsMULE0883	MORF0883	170-1515	1346	-	na / hypothetical protein	na
OsMULE1003	MORF1003	328-1282	955	+	na / na	na
OsMULE1006	MORF1006	448-831	384	+	na / na	na
OsMULE1007	MORF1007	529-930	402	+	na / hypothetical protein	na
OsMULE1010	MORF1010	239-1193	955	+	na / hypothetical protein	na
OsMULE1011	MORF1011	239-1193	955	+	na / hypothetical protein	na
OsMULE1012	MORF1012a	3887-4184	298	+	na / hypothetical protein	na
OsMULE1012	MORF1012b	4395-5096	702	+	na / na	na
OsMULE1012	MORF1012c	5455-5817	363	+	na / GDP-4-keto-6-deoxy-D-mannose-3, 5-epimerase-4-reductase (GER1)	na
OsMULE1013	MORF1013	248-1011	764	+	na / hypothetical protein	na
OsMULE1016	MORF1016	341-1063	723	+	na/reverse transcriptase	na
OsMULE1017	MORF1017	477-740	264	+	na / na	na
OsMULE1022	MORF1022a	207-785	579	-	na / na	yes
OsMULE1022	MORF1022b	898-1344	447	-	na / na	na
OsMULE1023	MORF1023a	207-785	579	-	na / na	na
OsMULE1023	MORF1023b	898-1344	447	-	na / na	na
OsMULE1025	MORF1025	292-663	372	-	na / na	na
OsMULE1026	MORF1026	292-663	372	-	na / na	na
OsMULE1026	5091528*	292-836	545	-	na / hypothetical protein	na
OsMULE1028	MORF1028	206-855	650	-	na / putative amino acid transporter	yes
OsMULE1028	MORF1028	206-855	650	-	na / putative amino acid transporter	yes
OsMULE1029	MORF1029	373-822	450	-	na / hypothetical protein	na
OsMULE1029	12643021+	373-822	450	-	na / hypothetical protein	na

OsMULE1032	MORF1032a	207-785	579	-	na / na	yes
OsMULE1032	MORF1032b	898-1344	447	-	na / na	na
OsMULE1036	MORF1036	284-1084	801	+	na/acyltransferase	na
OsMULE11b01	12583791	267-637	371	+	na / hypothetical protein	na
OsMULE11b02	MORF11b02	149-668	520	-	na / hypothetical protein	na
OsMULE11b03	MORF11b03	152-810	659	-	na / na	na
OsMULE11b04	MORF11b04	55-471	417	-	na / hypothetical protein	na
OsMULE11b04	9945048[+]	55-471	417	-	na / hypothetical protein	na
OsMULE11b05	MORF11b05	1350-5288	3939	-	na/putative methyl-CpG binding protein	na
OsMULE11b07	MORF11b07	254-546	293	-	na / na	na
OsMULE1201	MORF1201	192-838	647	+	na / hypothetical protein	na
OsMULE1202	MORF1202a	274-792	519	-	na / hypothetical protein	na
OsMULE1302	MORF1302b	274-792	519	-	na / hypothetical protein	na
OsMULE1305	MORF1305	375-645	271	+	na / na	na
OsMULE13b05	MORF13b05a	1931-2320	390	+	na / hypothetical protein	na
OsMULE13b05	MORF13b05c	13732-16773	3042	+	na / hypothetical protein	na
OsMULE1402	10140683	160-1021	862	+	na / hypothetical protein	na
OsMULE1403	MORF1403	963-2314	1352	-	na / hypothetical protein	na
OsMULE1404	MORF1404	313-985	673	+	glycosyl hydrolases family 17 / similar to glucan endo-1,3-beta-D-glucosidase precursor (Nicotiana tabacum)	na
OsMULE1404	9558496[+]	313-985	673	+	glycosyl hydrolases family 17 / similar to glucan endo-1,3-beta-D-glucosidase precursor (Nicotiana tabacum)	na
OsMULE1405	MORF1405	313-985	673	+	glycosyl hydrolases family 17 / similar to glucan endo-1,3-beta-D-glucosidase precursor (Nicotiana tabacum)	na
OsMULE1405	9711823[+]	313-985	673	+	glycosyl hydrolases family 17 / similar to glucan endo-1,3-beta-D-glucosidase precursor (Nicotiana tabacum)	na
OsMULE1406	MORF1406	408-1091	684	+	na / na	na
OsMULE1407	MORF1407	186-805	620	-	na / hypothetical protein	na
OsMULE1407	14018048[*]	186-1128	943	-	na / hypothetical protein	na
OsMULE1408	MORF1408	167-639	473	-	na / na	na
OsMULE1501	MORF1501a	277-943	667	+	na / na	na
OsMULE1720	MORF1720	164-1210	1047	-	na / hypothetical protein	na
OsMULE1720	8099230[*]	390-1210	821	-	na / hypothetical protein	na
OsMULE1721	MORF1721	921-1190	270	+	na/putative reverse transcriptase	na
OsMULE1722	12583816	387-1210	824	-	na / hypothetical protein	na

OsMULE1901	MORF1901	211-744	534	+	na / putative AMP-binding protein	na
OsMULE1901	14029022*	155-805	651	+	na / hypothetical protein	na
OsMULE1902	13786460	269-716	448	+	na / hypothetical protein	na
OsMULE1904	13940607	2763-3206	444	+	cytokinin-regulated kinase 1 (Nicotiana tabacum)	na
OsMULE1906	MORF1906	284-844	561	-	ammonium transporter family / ammonium transporter 2 (AtAMT2)	yes
OsMULE1906	11034567+	284-844	561	-	ammonium transporter family / ammonium transporter 2 (AtAMT2)	yes
OsMULE1907	MORF1907	328-819	492	-	na / hypothetical protein	
OsMULE1913	MORF1913	513-1020	508	-	na / na	na
OsMULE1914	MORF1914	202-570	369	-	na / na	na
OsMULE1917	MORF1917b	6229-6624	396	+	na/ proline-rich APG - like protein	na
OsMULE1919	MORF1919	247-459	213	+	na / hypothetical protein	na
OsMULE1922	MORF1922	445-1002	558	+	na / na	na
OsMULE1925	MORF1925	213-731	519	-	na / na	na
OsMULE1927	MORF1927	195-1052	858	-	na / hypothetical protein	na
OsMULE1927	13366207+	195-1052	858	-	na / hypothetical protein	na
OsMULE1928	MORF1928	503-1045	543	-	na / hypothetical protein	na
OsMULE1928	13486805+	503-1045	543	-	na / hypothetical protein	na
OsMULE1929	MORF1929	181-1042	862	-	na / na	yes
OsMULE1929	MORF1929	181-1042	862	-	na / na	yes
OsMULE19b01	MORF19b01	277-1528	1252	+	na/beta-ketoacyl-CoA synthase (Simmondsia chinensis)	na
OsMULE19b04	MORF19b04	362-1080	719	+	na / na	na
OsMULE19b05	MORF19b05	224-1002	779	-	na/ Cell Wall-bound Apyrase 2 (Pisum sativum)	na
OsMULE19c02	MORF19c02	255-1151	897	+	na / hypothetical protein	na
OsMULE19c03	MORF19c03	204-821	618	-	na / hypothetical protein	na
OsMULE19c03	5295937*	204-1421	1218	-	na / hypothetical protein	na
OsMULE19c04	MORF19c04	200-954	755	-	na / hypothetical protein	na
OsMULE19c06	MORF19c06	435-1007	573	+	na / na	na
OsMULE19c08	9711792	237-1024	788	+	histone deacetylase family / histone deacetylase HDA101 (ZM)	na
OsMULE19c09	7573618	237-1024	788	+	histone deacetylase family /	na

					histone deacetylase HDA101 (ZM)	
OsMULE19c12	MORF19c12	222-666	445	-	na / hypothetical protein	na
OsMULE19c12	11761080[+]	222-666	445	-	na / hypothetical protein	na
OsMULE19e01	MORF19e01	471-1048	578	-	na / na	na
OsMULE19f01	MORF19f01	185-768	584	+	na / na	na
OsMULE19f04	MORF19f04	243-644	402	-	pyridine nucleotide-disulphide oxidoreductase / lycopene beta cyclase	na
OsMULE19f05	MORF19f05	338-799	462	+	na / hypothetical protein	na
OsMULE19f06	MORF19f06	356-718	363	-	na / hypothetical protein	na
OsMULE19f06	6907092[*]	191-718	528	-	na / hypothetical protein	na
OsMULE19f07	MORF19f07	352-759	408	+	na / na	na
OsMULE19f13	MORF19f13	185-538	354	+	na/putative presenilin	na
OsMULE19f13	8099235[*]	185-700	516	+	na / hypothetical protein	na
OsMULE19f14	MORF19f14a	138-485	348	+	na/putative stachyose synthase	na
OsMULE19f14	MORF19f14b	865-4313	3449	-	na / hypothetical protein	na
OsMULE19f15	MORF19f15	215-792	578	-	na / hypothetical protein	na
OsMULE19f15	11862974[+]	215-792	578	-	na / hypothetical protein	na
OsMULE2201	MORF2201a	175-1247	1073	+	na / na	na
OsMULE2302	MORF2302	176-1925	1750	+	na / mudra	na
OsMULE2303	MORF2303b	7227-8976	1750	+	na / mudra	na
OsMULE2311	MORF2311a	161-344	184	-	na / mudra	na
OsMULE2322	MORF2322a	772-1911	1140	+	na / mudra	na
OsMULE2322	13486823[*]	2138-3235	1098	+	na / mutator-like	na
OsMULE2322	MORF2322b	2138-3958	1821	+	na/Similar to maize transposon muDR mudrA	na
OsMULE2323	MORF2323a	725-1509	785	+	na/mutator-like transposase	na
OsMULE2323	MORF2323b	1736-3487	1752	+	na/mudrA	na
OsMULE2325	MORF2325a	1190-2247	1058	+	Sina, Seven in absentia protein family/ SIAH1 protein (Brassica napus var. napus)	na
OsMULE2325	MORF2325d	14056-14298	243	-	na/mutator-like transposase	na
OsMULE2326	MORF2326a	1190-2247	1058	+	Sina, Seven in absentia protein family/ SIAH1 protein (Brassica napus var. napus)	na
OsMULE2326	MORF2326d	14056-14298	243	-	na/mutator-like transposase	na

OsMULE2327	MORF2327b	2992-3741	750	-	na/mutator-like transposase&Putative Ulp1 protease	na
OsMULE2328	MORF2328	256-561	306	-	na/mudrA	na
OsMULE2328	13486633	921-1346	426	+	na / hypothetical protein	na
OsMULE2329	MORF2329	313-1819	1507	-	na/mudrA	na
OsMULE2330	MORF2330	985-3220	2236	-	na/mudrA&hypothetical protein	na
OsMULE2330	6721543*	2236-3220	985	-	na / mutator-like	na
OsMULE2343	MORF2343b	8919-9443	525	-	na/mudrA	na
OsMULE2344	MORF2344b	8919-9443	525	-	na/mudrA	na
OsMULE2404	MORF2404	844-2434	1591	+	na / na	na
OsMULE2410	MORF2410	283-1301	1019	-	na / na	na
OsMULE2412	MORF2412	186-1349	1164	-	protein kinase domain/receptor kinase	na
OsMULE2413	MORF2413	183-1166	984	+	na/putative calmodulin-binding protein & hypothetical protein	na
OsMULE2413	6498435*	444-1166	723	+	na / hypothetical protein	na
OsMULE2414	MORF2414	183-1166	984	+	na/putative calmodulin-binding protein & hypothetical protein	na
OsMULE2414	11138063*	444-1166	723	+	na / hypothetical protein	na
OsMULE2415	MORF2415	183-1166	984	+	na/putative calmodulin-binding protein & hypothetical protein	na
OsMULE2415	13873006*	444-1166	723	+	na / hypothetical protein	na
OsMULE2417	MORF2417	515-1993	1479	-	na / na	na
OsMULE2505	MORF2505b	6301-10506	4206	+	na/mudrA&polyprotein	na
OsMULE2505	MORF2505d	11042-11226	185	-	na / na	na
OsMULE2506	MORF2506b	5214-5441	228	-	na / hypothetical protein	na
OsMULE2507	MORF2507b	74-262	189	+	na / hypothetical protein	na
OsMULE2508	MORF2508b	74-272	199	-	na / na	na
OsMULE2510	MORF2510	84-1122	1039	-	na / hypothetical protein	na
OsMULE2514	MORF2514	12436-12588	153	+	na / na	na
OsMULE2901	MORF2901	394-1550	1157	+	na / hypothetical protein	na
OsMULE2902	MORF2902	394-1550	1157	+	na / hypothetical protein & hypothetical protein	na
OsMULE3102	MORF3102	315-1662	1348	-	na / na	na
OsMULE3102	MORF3102	2353-3264	912	-	na / na	na

OsMULE3204	MORF3204	288-836	549	-	na / putative anthocyanin 5-aromatic acyltransferase	na
OsMULE3206	MORF3206	339-958	620	-	na / hypothetical protein	na
OsMULE3208	MORF3208	547-920	374	-	na / na	na
OsMULE3209	MORF3209	282-1646	1365	+	na / hypothetical protein	na
OsMULE3211	MORF3211	346-982	637	-	na / na	na
OsMULE3214	MORF3214	460-677	218	+	na / na	na
OsMULE3219	MORF3219	306-845	540	+	na/putative receptor serine/threonine kinase	na
OsMULE3223	MORF3223	502-1355	854	-	ubiquitin / polyubiquitin 4 - Arabidopsis thaliana	na
OsMULE3226	MORF3226	727-1167	441	-	na / hypothetical protein	na
OsMULE3226	10140645*	359-1167	809	-	na / hypothetical protein	na
OsMULE3229	MORF3229	151-752	602	+	na / na	na
OsMULE3232	MORF3232	1261-1401	141	-	na / na	na
OsMULE3233	10140749	237-1504	1268	-	Exo70 exocyst complex subunit / putative leucine zipper protein	na
OsMULE3239	MORF3239a	508-2000	1493	+	na / hypothetical protein & hypothetical protein	na
OsMULE3239	MORF3239b	2682-3421	740	+	na / similar to profilaggrin (Rattus norvegicus)	na
OsMULE3242	20303574	333-1034	702	-	serine/threonine protein kinases / putative protein kinase	na
OsMULE3246	MORF3246	224-1286	1063	+	protein phosphatase 2A regulatory B subunit (B56 family) /B' regulatory subunit of PP2A (A. thaliana)	na
OsMULE3248	MORF3248	330-982	653	-	na/GRS protein (A. thaliana)	na
OsMULE3252	MORF3252	1261-1401	141	-	na / na	na
OsMULE3253	MORF3253	822-1131	310	-	alternative oxidase/alternative oxidase 1c	na
OsMULE3254	MORF3254	721-1375	655	+	na / hypothetical protein	na
OsMULE3256	MORF3256	237-3690	3454	-	na / hypothetical protein	na
OsMULE3257	13872921	418-1271	854	+	na / hypothetical protein	na
OsMULE3258	MORF3258	311-718	408	+	na / cytochrome p450	na
OsMULE3266	MORF3266	938-1865	928	+	na / na	na
OsMULE3267	MORF3267	355-780	426	+	na / hypothetical protein	na
OsMULE3301	MORF3301	615-4472	3858	-	na/mudrA	na
OsMULE3302	MORF3302	615-3177	2563	-	na / mudrA & putative serine/threonine phosphatase PP7	na
OsMULE3601	MORF3601	609-2265	1657	+	FAR1 & mutator family / far-red impaired response protein overlaps	na

					w/ mutator-like transposase	
OsMULE3705	MORF3705	202-828	627	-	na / hypothetical protein	na
OsMULE3707	MORF3707	960-1491	532	+	na / hypothetical protein	na
OsMULE3708	11071977	198-940	743	-	na / hypothetical protein	na
OsMULE3709	11034606	198-940	743	-	na / hypothetical protein	na
OsMULE3711	MORF3711	390-1163	774	+	na / na	na
OsMULE3712	MORF3712a	714-1447	734	-	na / hypothetical protein	na
OsMULE3712	MORF3712b	4158-6854	2697	-	na / hypothetical protein	na
OsMULE3712	10716614*	4158-6367	2210	-	na / hypothetical protein	na
OsMULE3713	MORF3713	1164-1543	380	-	na / hypothetical protein	na
OsMULE3714	MORF3714	54-1342	1289	-	na / hypothetical protein	na
OsMULE3718	MORF3718	235-782	548	-	na / hypothetical protein	na
OsMULE3719	MORF3719	669-1179	511	+	na / hypothetical protein	na
OsMULE3720	MORF3720	917-1274	358	+	na / hypothetical protein	na
OsMULE3720	13872931+	917-1274	358	+	na / hypothetical protein	na
OsMULE3724	MORF3724	280-1545	1266	+	na / putative DNA-binding protein	na
OsMULE3724	13486832*	280-1585	1306	+	na / putative DNA-binding protein	na
OsMULE3725	MORF3725	280-1545	1266	+	na / putative DNA-binding protein	na
OsMULE3725	5042439*	280-1585	1306	+	na / putative DNA-binding protein	na
OsMULE3726	MORF3726	247-1089	843	+	na/putative chalcone synthase	na
OsMULE3728	na	na	na	na	na	yes
OsMULE3730	MORF3730	216-1190	975	-	na / hypothetical protein	na
OsMULE3732	MORF3732	216-1190	975	-	na / hypothetical protein	na
OsMULE3733	MORF3733	216-1190	975	-	na / hypothetical protein	na
OsMULE3734	14140282	217-795	579	-	lectin legB / hypothetical protein	na
OsMULE3735	MORF3735	276-1085	810	+	na / hypothetical protein	na
OsMULE3736	MORF3736	234-829	596	-	na / na	na
OsMULE3904	MORF3904	1301-2029	729	-	na / hypothetical protein	na
OsMULE3906	MORF3906	161-808	648	-	auxin responsive protein / auxin-induced protein X15	na
OsMULE4011	MORF4011	4200-7689	3490	-	na / hypothetical protein	na
OsMULE4042	MORF4042a	2626-3023	398	+	na / na	yes

OsMULE4042	MORF4042b	1282-2368	1087	-	na / hypothetical protein	yes
OsMULE4042	MORF4042c	4026-5331	1306	-	na / hypothetical protein	na
OsMULE4043	MORF4043a	2585-2982	398	+	na / na	na
OsMULE4043	MORF4043b	1242-2000	759	-	na / hypothetical protein	na
OsMULE4043	4680493[+]	1242-2000	759	-	na / hypothetical protein	na
OsMULE4043	MORF4043c	3984-5286	1303	-	na / hypothetical protein	na
OsMULE4043	4680503[+]	3984-5286	1303	-	na / hypothetical protein	na
OsMULE4044	MORF4044	1424-2031	608	+	na / na	na
OsMULE4053	MORF4053	2634-2792	159	-	na / na	na
OsMULE4054	MORF4054	2634-2792	159	-	na / na	na
OsMULE4101	MORF4101	1083-2387	1305	-	na / hypothetical protein	na
OsMULE4302	MORF4302	366-753	388	+	na / na	na
OsMULE4432	MORF4432a	642-1365	724	-	na / na	na
OsMULE4432	10140712[*]	323-1098	776	+	na / hypothetical protein	na
OsMULE4432	MORF4432b	4114-5098	985	-	na / na	na
OsMULE4907	MORF4907	1051-1593	543	+	na / na	na
OsMULE5102	MORF5102	376-799	424	+	na / na	na

[+] annotated ORF is identical to a predicted ORF

[*] annotated ORF overlaps only in portion with a predicted ORF and thus is counted as a separate ORF

Supplementary Table 5.5A Homology analysis at the break points of *Arabidopsis* MULEs

MULE and aligned host DNA		homology length (bp)		sequence	
		L	R	L	R
cMULE:	AtMULE215	gap	17	na	AATATAATACAGGGGTAT
eMULE:	AtMULE442				AATATAATATAGGGGTAT
acquisition:	4732166				A- TGATTTATAGGGGTAT
cMULE:	AtMULE243a	2	32	GT	TTCAAGTATAATCTTATGCATTACCTCTCGTT
eMULE:	AtMULE553			GT	TTGTGATATTCTCCTAAACACACCCTAGCATT
acquisition:	9885845			GT	TTCAGGTATAATCTTATGCACTACCTCTCGTT
cMULE:	AtMULE243b	2	12	CA	GAGAAGCGTGAA
eMULE:	AtMULE178			CA	GAGGACCGTGTA
acquisition:	7549541			CA	GAGGAAGCGTGAA
cMULE:	AtMULE259	2	7	GC	CATAGCT
eMULE:	AtMULE492			GC	CATAGCT
acquisition:	76279			GC	CATAGCT
cMULE:	AtMULE276	gap	7	na	TTTTGGA
eMULE:	AtMULE519				TTTTGAA
acquisition:	12408717				TTTTTGA
cMULE:	AtMULE291	gap	7	na	AAAGATT
eMULE:	AtMULE283				AAAGATT
acquisition:	20198147				AAAGAGT
cMULE:	AtMULE361	8	5	AACAGAAT	TATAAA

eMULE:	AtMULE242			AACAGAAT	TATAAA
acquisition:	9954738			AATACA -T	TATAAA
cMULE:	AtMULE392	37	38	CCAAAAAAAATGCTTTTAAAAATCCTTGTAAATTTTT	TTTTACAATAGTTTTACAAGATTTACAATCGTTTTTAA
eMULE:	AtMULE593			CCAAAAAAATTAATTTTTAAAAT CCTTGTAAATATTT	TTTTAAAAGGGTTTTAAAAGATTTACAAGAGATTTTAA
acquisition:	20197457			CCAAAAAAA-TGCTTTTAAAAATCCTTGTAAATTTTT	TTTTACAATAGGTTTACAAGATTTACAAGAGTTTTTAA
cMULE:	AtMULE059	16	9	ACTTTTGTTTTAGTCT	AGAGAATTT
eMULE:	AtMULE c-sequence-1			AATTTTGTTTTAGTCT	AGAGAATTT
acquisition:	6449044			ACTTTTGTTTTAGTCT	AGAGAATTT
cMULE:	AtMULE064	7	gap	TAAAAAA	na
eMULE:	AtMULE c-sequence-2			TAAAAAA	
acquisition:	12324196			TAAAAAT	
cMULE:	AtMULE081	gap	5	na	ATTTT
eMULE:	AtMULE519				ATTTT
acquisition:	12408717				ATTTT
cMULE:	AtMULE095	5	gap	AAAAA	na
eMULE:	AtMULE c-sequence-3			AAAAA	
acquisition:	5041974			AAAAA	
cMULE:	AtMULE127	1	1	A	C
eMULE:	AtMULE055			A	C
acquisition:	13677103			A	C
cMULE:	AtMULE143	gap	3	na	TTT
eMULE:	AtMULE552				TTT
acquisition:	5041974				TTT

cMULE:	AtMULE170a	0	8	0	TGTTTATT
eMULE:	AtMULE511				TGTTTATT
acquisition:	20197755				CGTTTATT
cMULE:	AtMULE170b	gap	4	na	TTTT
eMULE:	AtMULE511				TTTT
acquisition:	5262205				TTTT
cMULE:	AtMULE170c	5	1	TCGAC	A
eMULE:	AtMULE511			TCGAC	A
acquisition:	20197730			TCGAC	A
cMULE:	AtMULE413	6	na	TTTATT	na
eMULE:	AtMULE618			TTTATT	
acquisition:	6899954			TTTATT	
cMULE:	AtMULE 422	4	na	GATT	na
eMULE:	AtMULE619			GATT	
acquisition:	12324896			GATT	
cMULE:	AtMULE458	1	8	G	TACAAATA
eMULE:	AtMULE251			G	TACAAATA
acquisition:	13677103			G	TTCAACTA
cMULE:	AtMULE477	24	na	TCACAAACAAAAATGTACGTATAT	na
eMULE:	AtMULE402			TTAtAACAAGGAATGTACGTATAT	
acquisition:	7635467			TCACCAACAAAGATGTACGTACAT	
cMULE:	AtMULE477	5	1	TCGAC	na

eMULE:	AtMULE402			TCGAC	
acquisition:	7635467			TCGAC	
cMULE:	AtMULE493	5	na	TGCCT	na
eMULE:	5822965			TGCCT	
acquisition:	5002514			TGTCT	
cMULE:	AtMULE499	6	na	TGGATA	na
eMULE:	AtMULE072			TGGATA	
acquisition:	4185120			TGGATA	
cMULE:	AtMULE516	24	6	TTTTAA-AATCCTTGTAAATTTTT	TAAAAT
eMULE:	AtMULE593			TTTTAAAAATCCTTGTAAATTTTT	TAGAAT
acquisition:	5391457			TTTTAAAAATCCTTGTAAATATTT	TAAAAT
cMULE:	AtMULE519	gap	7	na	ATTATGG
eMULE:	AtMULE389				ATTAGGG
acquisition:	12408709				ATTATGG
cMULE:	AtMULE524	gap	3	na	TAA
eMULE:	AtMULE190				TAA
acquisition:	3449311				TAA
cMULE:	AtMULE541	gap	2	na	TT
eMULE:	AtMULE135				TT
acquisition:	7649372				TT
cMULE:	AtMULE571	46	gap	ATAACCTCTCCTTCAGATCTGGGTTTCTTCATT CG-------TTTT--GGGTAAT	na
eMULE:	AtMULE620			ATAACCTCTCTTTCAGATTTGGGTTTTGTTCATT CGACTTTTTGGGTAAT	

acquisition:	5732428			ATAACCTCTCTCTTAGATCTGGGTTTTGCTCCTA	
				AGACTTTTTAGGTAAT	
cMULE:	AtMULE599	2	8	AG	TACAAATA
eMULE:	AtMULE252			AG	TACAACTA
acquisition:	8347605			AG	TTCAACTA
cMULE:	AtMULE562	gap	2	na	TT
eMULE:	AtMULE233				TT
acquisition:	8347605				TT

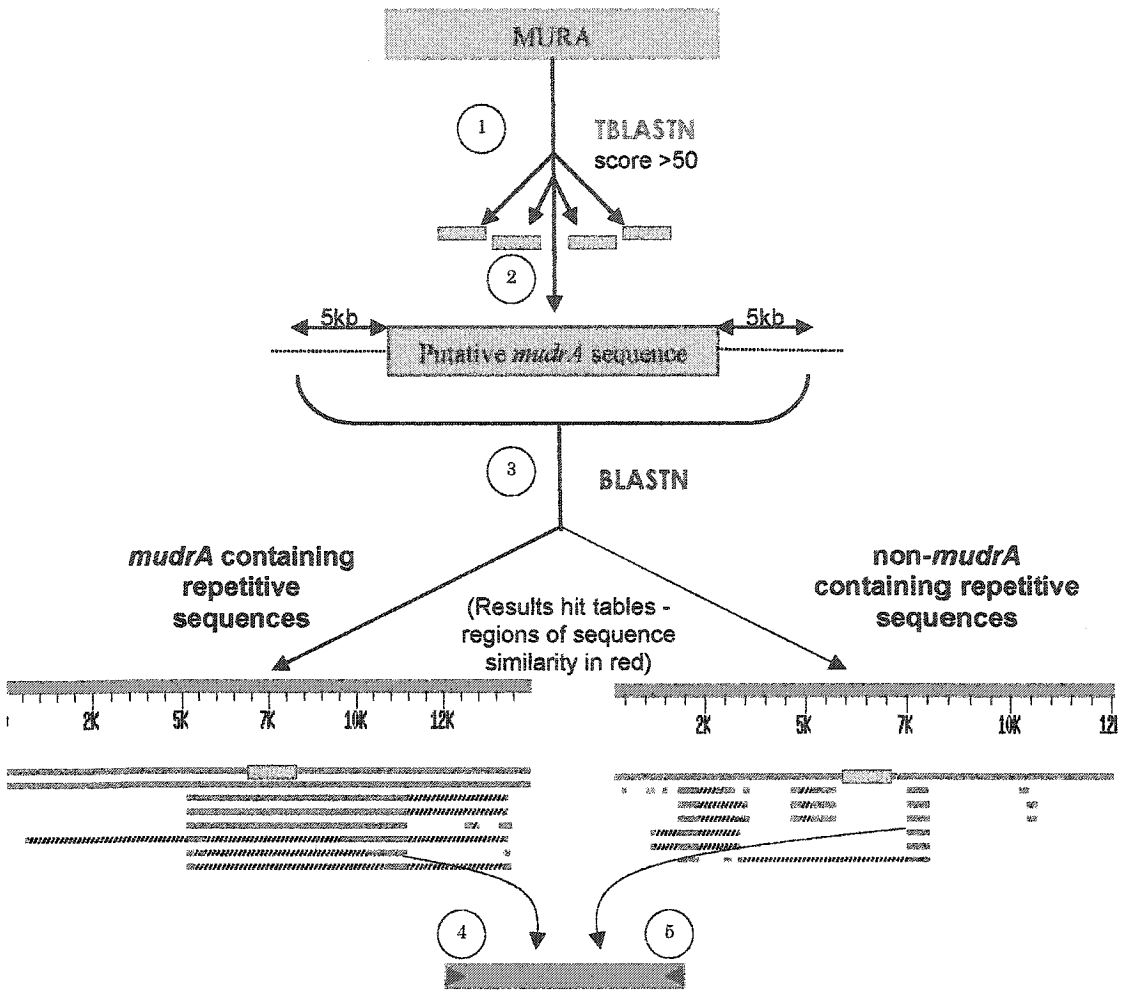
L: left side of the elements.
R: right side of the elements.
na: not applicable.
cMULE: composite MULE or the MULE containing at least one acquired host DNA segment.
eMULE: empty MULE or the element that does not contain host DNA but shares
a high sequence identity with the cMULE.
AtMULE c-sequence: consensus sequence of the corresponding group.

Supplementary Table 5.5B Homology analysis at the break points of rice MULEs

MULE and aligned host DNA		homology length (bp)		sequence	
		L	R	L	R
cMULE:	OsMULE0714	na	18	gap	TTGTCCCCGTCCGCGTCC
eMULE:	OsMULE0715	na	18	gap	TGGCCGCAGTCAGTGTCC
acquisition:	1778820	na	18	gap	TGGTCCCCGTCCGCGTCC
cMULE:	OsMULE0873	9	na	CTCTCCCTC	gap
eMULE:	OsMULE0866	9	na	CTCTCTCTC	gap
acquisition:	15408621	9	na	CTCGCCCTC	gap
cMULE:	OsMULE1001	10	na	CGAGCGGAGG	gap
eMULE:	OsMULE1038	10	na	CGAGCGGAGG	gap
acquisition:	13184944	11	na	CCAGCTGGAGG	gap
cMULE:	OsMULE1013	11	na	CTCTCTCTCTC	gap
eMULE:	OsMULE1039	11	na	CTTTCTCTCTC	gap
acquisition:	19110518	11	na	CTCGCTCTCTC	gap
cMULE:	OsMULE11b06	2	na	GG	gap
eMULE:	OsMULE11b04	2	na	GG	gap
acquisition:	18461245	2	na	GG	gap
cMULE:	OsMULE1201	1	na	T	gap
eMULE:	OsMULE1205	1	na	T	gap
acquisition:	15290074	1	na	T	gap
cMULE:	OsMULE1715	69	na	GTTCTAACCCACTCGATGCGCTGACGTGGCAATGCC	gap

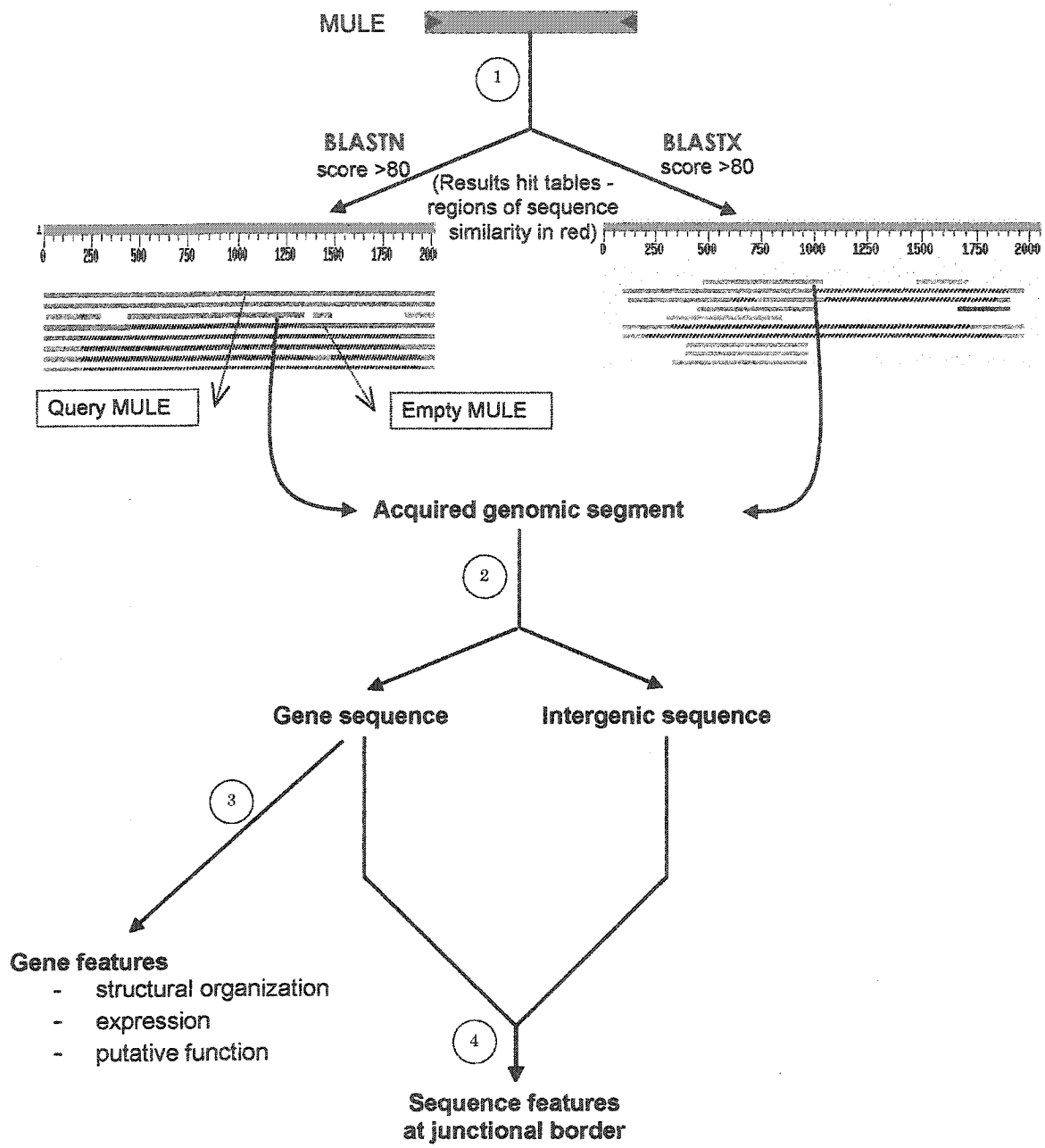
				ACGGTGACACCTCCGTGTTAGTGGAATCCACAT	
eMULE:	OsMULE1729	69	na	GCTCCAACCCGCTCCATGAGACAGCGTGGC-ATGTC	gap
				ATGGTGACGCTTAAGTGTCAGCGGAATCCACAT	
acquisition:	9795249	69	na	GCTCCGTCACGCTCCATATGCTGACGTGGCATTGCC	gap
				ATGATGACATCTACATGTCAGTGAAATCCATAT	
cMULE:	OsMULE1716	69	na	GTTCTAACCCACTCGATGCGCTGACGTGGCAATGC	gap
				CACGGTGACACCTCCGTGTTAGTGGAATCCACAT	
eMULE:	OsMULE1729	69	na	GCTCCAACCCGCTCCATGAGACAGCGTGGC-ATGTC	gap
				ATGGTGACGCTTAAGTGTCAGCGGAATCCACAT	
acquisition:	9795249	69	na	GCTCCGTCACGCTCCATATGCTGACGTGGCATTGCC	gap
				ATGATGACATCTACATGTCAGTGAAATCCATAT	
cMULE:	OsMULE1902	1	na	C	gap
eMULE:	OsMULE1901	1	na	C	gap
acquisition:	10140737	1	na	C	gap
cMULE:	OsMULE1925	na	3	gap	GAA
eMULE:	OsMULE1924	na	3	gap	GAA
acquisition:	13124870	na	3	gap	GAA
cMULE:	OsMULE19c03	8	13	CGCACCGC	TCGAGGAGGCCGT
eMULE:	OsMULE19c13	8	13	CGCACCGC	TCGAGGAGGCCGT
acquisition:	10241657	8	13	CGCACCGC	TCGTGCAGTCCGT
cMULE:	OsMULE2407	3	na	ACC	gap
eMULE:	OsMULE2402	3	na	ACC	gap
acquisition:	21624301	3	na	ACC	gap
cMULE:	OsMULE3212	na	1	gap	A

eMULE:	OsMULE3268	na	1	gap	A
acquisition:	6539551	na	1	gap	A
cMULE:	OsMULE3219	na	1	gap	C
eMULE:	OsMULE3270	na	1	gap	C
acquisition:	15144390	na	1	gap	C
cMULE:	OsMULE3730	na	20	gap	GGCATTCGACCCCACCCCC
eMULE:	OsMULE3737	na	20	gap	GGCGTTCGACTGCCACCCCC
acquisition:	13429991	na	20	gap	GGCGTTCGAGCGCCACCTCC



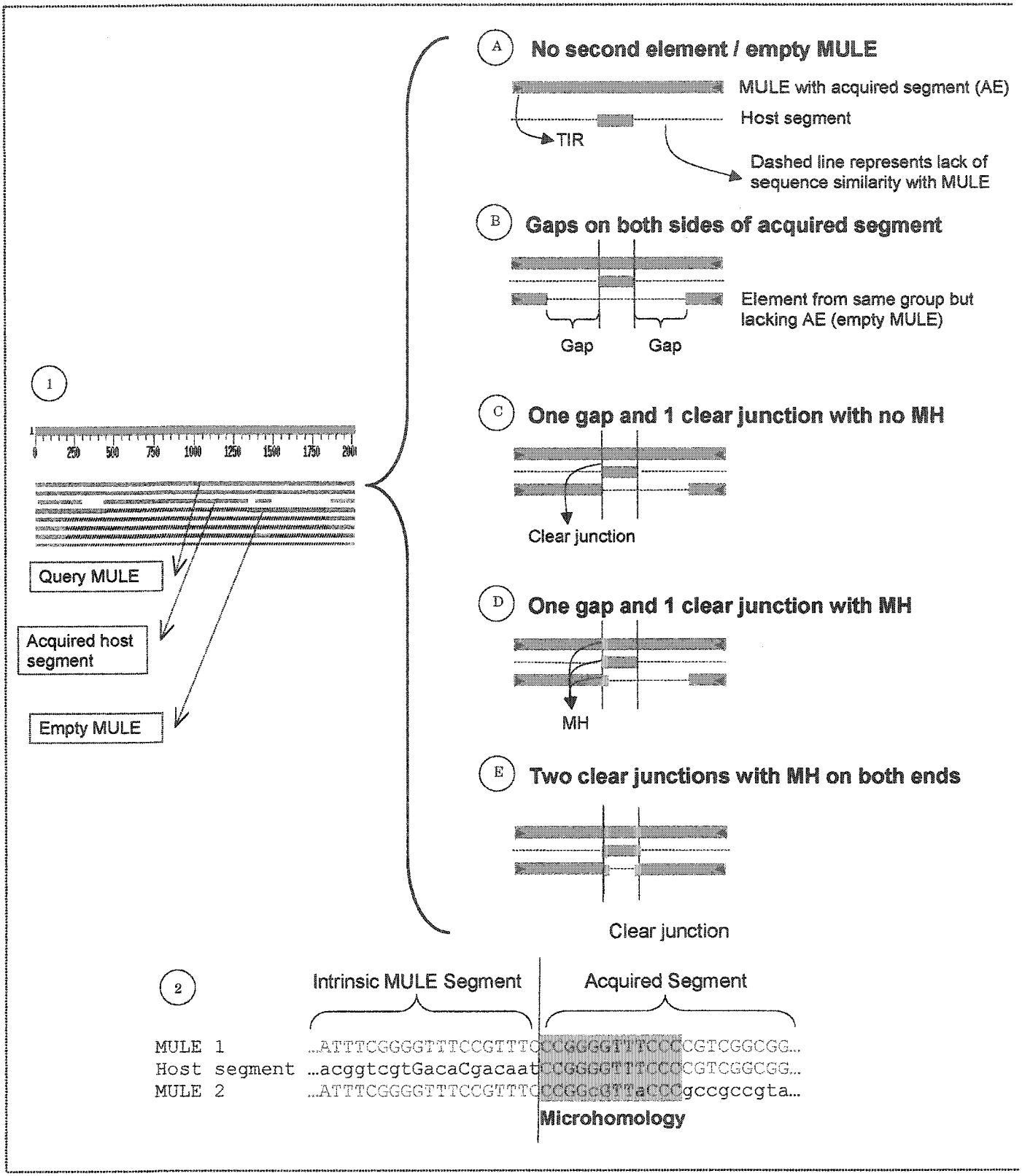
Bonafide MULE with TSDs

Supplementary Figure 5.1 A MURA-based strategy for the identification of MULEs in rice and *Arabidopsis*. Maize MURA and annotated rice MURA-like sequences were used as queries in a translated BLAST search (TBLASTN) to generate a list of putative *mudrA* nucleotide sequences within our database (step 1). Only putative *mudrA*-like sequences above a score cut-off of 50 were kept. These putative *mudrA*-like sequences plus 5kb flanking regions (step 2) were then used as queries in nucleotide BLAST (BLASTN) searches (step 3). BLASTN results were inspected for such distinguishing features such as TIRs defined ends, and TSDs (step 4). Having the terminal structure but missing the sequence similarity with a *mudrA*-like query within the internal regions defines the corresponding elements as non-autonomous MULEs. Repetitive structures that did not extend over but correspond to only one side of the putative *mudrA* sequence (within the flanking 5kb) were also analyzed for TIR structures utilizing BLAST2 (step 5). Elements that shared a BLASTN score of 80 or greater were grouped together.



Supplementary Figure 5.2 Identification and analysis of MULE acquisitions in rice and Arabidopsis. Each verified MULE was checked for the presence of acquisition events using BLASTN and translated BLAST, BLASTX (step 1). Acquired segments at the nucleotide level (BLASTN) represent relatively recent events. To be included in our analysis, segments must meet the following criteria: not be transposase or transposon-related, exist in the genome as a copy that is unassociated with any MULE, be at least 100-bp in continuous length, and have a BLASTN score of greater than 80. Often a MULE will acquire several segments from the same region of a clone. If these segments exist on the clone no further than 50-bp apart, then they are

joined together and taken as a single acquisition event. Acquired segments at the amino acid level (BLASTX) represent relatively ancient events whereby the origin of the segment can not be determined. Acquisition events were categorized as either gene-related or intergenic (step 2). BLASTX results were only recorded if they had a score greater than 80. Hits from the same protein and in the same frame (+/-) were joined together and considered as a single acquisition event. For gene-related acquisitions, annotation and ORF prediction were used to determine the gene structure of the acquired segment. Possible expression of these gene segments was revealed by a BLASTN search against NCBI's EST database. Only hits above 95% identity along the entire EST were noted. The putative function was deduced through conserved-domain (RPS-BLAST) searches and BLASTP analysis (step 3). Further characterization of the acquired segments' junction borders revealed information regarding potential acquisition mechanisms (step 4).


A
No second element / empty MULE
MULE with acquired segment (AE)
Host segment
TIR
Dashed line represents lack of sequence similarity with MULE
B
Gaps on both sides of acquired segment
Element from same group but lacking AE (empty MULE)
Gap
Gap
1
0
250
500
750
1000
1250
1500
1750
2000
Query MULE
Acquired host segment
Empty MULE
C
One gap and 1 clear junction with no MH
Clear junction
D
One gap and 1 clear junction with MH
MH
E
Two clear junctions with MH on both ends
Clear junction
2
Intrinsic MULE Segment
Acquired Segment
MULE 1 ...ATTTCGGGGTTTCCGTTTCCCGGGGTTTCCCCGTCGGCGG...
Host segment ...acggtcgtGacaCgacaatCCGGGGTTTCCCCGTCGGCGG...
MULE 2 ...ATTTCGGGGTTTCCGTTTCCCGGcGTTaCCCgccgccgta...
Microhomology

Supplementary Figure 5.3 Analysis of junctional borders.Clear junctions were defined as the point where similarity with another element lacking the AE (empty MULE) ends and the similarity with the acquired segment begins. The sequence similarity between the MULE with the AE, the empty MULE, and the acquired host segment are revealed in the graphical and text outputs. (1). Certain elements did not have a corresponding empty MULE to align with: (A) the vast majority of elements don't share sufficient internal sequence similarity with other members of the group, resulting in gaps (gaps could be present on both sides as described in (B) or one side (C and D)). Once clear breakpoints are identified, they were further analyzed homologies defined as short sequence (1-46-bp) similarities (>50% identity) between all three segments: the MULE with the acquired segment, the host copy of the acquired segment with 50-bp of the flanking sequence, and an empty MULE. Homologies shown in green could be found on one end of the acquired segment (D) or on both sides (E). A typical example of the homology at a junction (5') is depicted in (2). Sequence similarity is shown in red. Sequences in black or dashed lines represent lack of sequence similarity.

CHAPTER 6

GENERAL CONCLUSIONS

Exploring the vast information provided from the sequencing of the Arabidopsis and rice genomes, we performed a genome-wide study of the MULE family in these two species. As the first study on a DNA transposon family in higher plants and one of the few in eukaryotes at such a scale, we revealed multi-directional TE-host relationships that would not be otherwise uncovered by a classic way of TE studies.

The *Mu*/MULE system exists in not only maize but other higher plants as well. A TIR structure has long been regarded as one hallmark in the classification of Class II TEs and imperative for their activity (Plasterk, 1995). However, our discovery of the large number of non-TIR MULEs and their *de novo* activity (both transcriptional and transpositional) challenges this point of view. The co-existence of active TIR- and non-TIR MULEs in the Arabidopsis genome suggests that the TIR structure may just be the one *cis*-factor in the regulation of Class II TE mobility. Families within this Class should not be classified based only on TIRs.

The MULEs in the Arabidopsis genome represent one of the most abundant, diversified, yet still functional TE families in eukaryotes. Such a presentation was achieved through TE-host co-regulations over evolutionary time. Certainly, MET1-mediated CpG methylation can trigger *mudrA*-like gene silencing; however, its effect mostly requires the formation of the Arabidopsis heterochromatin. The co-repression of *mudrA*-like genes by methylation and heterochromatin formation enhances the quelling

of the MULE activity in wild-type *Arabidopsis*. However, the silent state is reversible, as seen from the *de novo* MULE mobility in methylation- and chromatin-remodeling-deficient Arabidopsis plants (also see Single *et al.*, 2001). These phenomena suggest that heterochromatic regions are not merely a graveyard for the accumulating of dead TEs; instead, heterochromatin formation may have played a crucial role in eukaryotic TE evolution. The dynamics of a eukaryotic genome can cause quelling-reactivation cycles of TE mobility, which consequently regulates TE amplifications within the genome.

Study of the MULE diversity revealed several lines of new evidence indicating a beneficial function of the elements in *Arabidopsis* and rice. The shared sequence homology between a functional *FAR1* and the expressed *mudrA*-like genes in both Arabidopsis and rice suggests that transposase genes could have evolved to take on a host function directly. The large accumulation of the elements within heterochromatic regions and the existence of MULE-contained centromeric-specific sequences suggest a link between the MULE activity and the evolution of *Arabidopsis* chromosomal remodelling. Finally, the MULE creation of numerous host genes with futures of redundancy, mosaic organization, and differential expression profiles demonstrates that MULE activity may have played a significant role in the evolution of eukaryotic genes and multiple gene families.

In conclusion, as the first eukaryotic TE family examined at a genome-wide scale for the study of TE-host relationships, the data presented in this thesis demonstrate that from an evolutionary perspective, eukaryotes and mobile DNA can be co-regulated. It is this co-regulation that drives TE-host co-evolution.

REFERENCES

Alarie, M., M. M. Chernaia, B. A. Malcolm, and M. N. James (1994). Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinases. *Nature* **369:** 72-76.

Bartee, L., F. Malagnac, and J. Bender (2001). Arabidopsis *cmt3* chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. *Genes Dev.* **15**: 1753-1758.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer (2000). The Pfam Protein Families Database. *Nucleic Acids Res.* **28**: 263-266.

Bazan, J. F. and R. J. Fletterick (1988). Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications. *Proc. Natl. Acad. Sci. USA* **85**: 7872-7876.

Becker, H. and R. Kunze (1996). Binding site for maize nuclear proteins in the subterminal regions of the transposable element *Activator*. *Mol. Gen. Genet.* **251**: 428-435.

Benito, M. -I., and V. Walbot (1997). Characterization of the maize *Mutator* transposable element MURA transposase as a DNA-binding protein. *Mol. Cell. Biol.* **17**: 5161-5175.

Bennetzen, J. L. (1984) Transposable elements *Mu*1 is found in multiple copies only in Robertson's *Mutator* maize lines. *J. Mol. Appl. Genet.* **2**: 519-524.

Bennetzen, J. L. (1996). The *Mutator* transposable element system of maize. In *Transposable elements*, edited by H. Saedler and A.Gierl. Springer-Verlag, Berlin.

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

Berg, J. M. (1986). Potential metal-binding domains in nucleic acid binding proteins. *Science* **232**: 485-487.

Besansky, N. J., S. M. Paskewitz, D. M. Hamm, and F. H. Collins (1992). Distinct families of site-specific retrotransposons occupy identical positions in the rRNA genes of *Anopheles gambiae*. *Mol. Cell. Biol.* **12**: 5102-5110.

Bird, A. (1997). Does DNA methylation control transposition of selfish elements in the gremlins? *Trends Genet.* **13:** 469-470.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**: 21.

Bird, A. and A. P. Wolffe (1999). Methylation-induced repression: belts, braces and chromatin. *Cell* **99**: 451-454.

Britten, J. R. (1996). DNA sequence insertion and evolutionary variation in gene regulation *Proc. Natl. Acad. Sci. USA.* **93**: 9374-9377.

Boeke, J. D., and J. P. Stoye (1997). Retrotransposons, Endogenous retroviruses, and the evolution of the retroelements. In *Retroviruses*, edited by J. M. Coffin *et al.*Cold Spring Harbor Laboratory Press, U.S.A.

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**: 1823-1294.

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

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. USA* **91**: 1411-1413.

Bureau, T. E., S. E. White, and S. R. Wessler (1994). Transduction of a cellular gene by a plant retroelement. *Cell* **77**: 479-480.

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.

Caras, I. W., M. A. Davitz, L. Rhee, G. Weddell, D. W. Martin Jr. *et al.* (1987). Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins. *Nature*, **325**: 545-549.

Cabrita, G. C., M. Iqbal, H. Reddy and G. Kemp (1997). Activation of the Adenovirus Protease Requires Sequence Elements from Both Ends of the Activating Peptide. *J. Biol. Chem.* **272**: 5635-5639.

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

Chandler, V. L. and V. Walbot (1986). DNA modification of a maize transposable element correlates with loss of activity. *Proc. Natl. Acad. Sic. USA.* **83**: 1767-1771.

Clark, S. J. and M. Frommer (1997). Bisulphite genomic sequencing of methylated cytosines. In *Laboratory Methods for the Detection of Mutations and Polymorphisms in DNA*, edited by G. R. Taylor. Boca Raton: CRC Press.

Cogoni, C. (2001). Homology-dependent gene silencing mechanisms in fungi. *Annu. Rev. Microbiol.* **55**: 381-406.

Comeron, J. (1995). A method for estimating the numbers of synonymous and nonsynonmous substitution per site. *J. Mol. Evol.* **41**: 1152-1159.

Comeron, J. M., M. Kreitman, and M. Aguade (1999). Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. *Genetics* **151**: 239-249.

Consortium and C. W. P. A. S. (2000). The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100**: 377-386.

Copenhaver, G. P., K. Nickel, T. Kuromori, M. -I. Benito, S. Kaul, *et al.* (1999). Genetic definition and sequence analysis of Arabidopsis centromeres. *Science* **286**: 2468-2474.

Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res. Nucleic Acids Res.* **16**: 10881-10890.

Costello, J. F. and C. Plass (2001). Methylation matters. *J. Med. Genet.* **38:** 285-303.

Covey, S. N. (1986). Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. *Nucleic Acids Res.* **14**: 623-633.

Djikeng A, H. Shi, C. Tschudi and E. Ullu (2001). RNA interference in Trypanosoma brucei: cloning of small interfering RNAs provides evidence for retrotransposon-derived 24-26-nucleotide RNAs. *RNA* **7**: 1522–1530.

Dimitri, P., B. Arcà, L. Berghella and E. Mei (1997). High genetic instability of heterochromatin after transposition of the LINE-like I factor in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*. **94**: 8052-8057.

Dimitra, P. and N. Junakvic (1999). Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Tren. In Genet.* **15**: 123-124.

Darlix, J. L., M. Lapadat, H. Tapolsky, H. De Rocquigny, and B. P. Roques (1995). First glimpses at structure-function relationships of nucleocapsid protein of retroviruses. *J. Mol. Biol.* **254**: 523-537.

Dorer, D. and S. Henikoff 1994 Transgene repeat arrays interact with distant heterochromatin and cause silencing in *cis* and *trans*. *Cell* **77**: 993-1002.

Eickbush. T. (1997). Molecular biology: telomerase and retrotransposons: which came first. *Science* **277**: 911-912.

Eickbush, T. (1999). Transcription: exon shuffling in retrospect. *Science* **283**: 11465-1467.

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. *Nucleic Acids Res.* **22**: 2634-2636.

Engels, W. R. (1996). P-elements in *Drosophila*. In *Transposable Elements*, edited by H. Saedler and A. Gierl, A. Springer, Germani.

Eline. T., L. Prak and H. H. Kazazian Jr. (2000). Mobile elements and the human genome. *Nature Rev. Genet.* **1**: 134-143.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Fagard, M. and H. Vaucheret (2000). (*trans*)Gene silencing in plants: how many mechanisms? *Annu. Rev. Plant Physiol. Plant Mol. Biol. Annu. Rev. Plant Physiol. Plant Mol. Biol.* **51**: 167-194.

Federoff, N. (1996). DNA methylation and activation of the maize *Spm* transposable element. *Curr. Top. Microbiol. Immunol.* **197**: 143-164.

Feng, Q., J. V. Moran, H. H. Kazazian, and J. D. Boeke (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905-916.

Finnegan, E. J., W. J. Peacock and E. S. Dennis (1996). Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development. *Proc. Natl. Acad. Sci. USA.* **93**: 8449-8454.

Finegan, E. J., R. K. Genger, W. J. Peacock and E. S. Dennis (1998). DNA methylation in plants. *Rev. Plant Physiol. Plant Mol. Biol.* **49**: 223-247.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, and S. E. Driver (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**: 806-811.

Fransz, P. F., S. Armstrong, J. H. de Jong, L. D. Parnell, C. van Drunen (2000). Genetic definition and sequence analysis of Arabidopsis centromeres. *Cell* **100**: 367-376.

Feschotte, C., N. Jing and S. R. Wessler (2002). Plant transposable: where genetics meet genomics. *Nature Rev.Genet.* **3**: 329-341.

Fu, X. -D. (1993). Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* **365**: 82-85.

Fuks, F., W. A. Burgers, A. Brehm, L. Hughes-Davies and T. Kouzarides (2000). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat Genet.* **24**: 88-91.

Gao, F. (1998). A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol.* **72**: 5680-5698.

Garrett, J. E., D. S. Knutzon, and D. Carroll (1989). Composite transposable elements in the *Xenopus laevis* genome. *Mol. Cell. Biol.* **9**: 3018-3027.

Geyer, P., A. Chien, V. Corces, and M. Green (1991). Mutations in the su(s) gene affect RNA processing in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA.* **88**: 7116-7120.

Gierl, A. (1996). The *En/Spm* transposable elements of maize. In *Transposable Elements*, edited by H. Saedler and A. Gierl, A. Berlin: Springer.

Goodier, J. L., E. M. Ostertag, and H. H. Kazazian (2000).Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**: 653–657.

Gorbunova, V. and A. V. Levy (1997). Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. *Nucleic Acids Res.* **25**: 4650-4657.

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.

Green, E. D. and A. Chakravarti (2001). The human genome sequence expedition: views from the 'base camp.' *Genome Res.* **11**: 645–651.

Gribskow, M., A. D. Mclachlan, and D. Eisenberg (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad.Sci. USA.* **84**: 4355-4358.

Gruenbaum, Y., R. Stein, H. Cedar and A. Razin (1981). A methylation of CpG sequences in eukaryotic DNA. *FEBA Lett.* **124**: 67-71.

Guy, J., C. Spalluto, A. McMurray, T. Hearn, M. Crosier, *et al.* (2000). Genomic sequence and transcriptional profile of the boundary pericentromeric satellites and genes on human chromosome arm 10q. *Human Mol. Genet.* **9**: 2029-2042.

Haber, J. E. (2000). Partners and pathways repairing a double-strand break. *Trends in Genet.* **16**: 259-264.

Hackstein, J. H., R. Hochstenbach and P. L. Pearson (2000). Towards an understanding of the genetics of human male infertility: lessons from flies. *Trends Genet.* **16**: 562-572.

Hanano, S., M. Sugita, and M. Sugiura (1996). Isolation of a novel RNA-binding protein and its association with a large ribonucleic protein particle present in the nucleoplasm of tobacco cells. *Plant Mol. Biol.* **31**: 57-68.

Hanania, U., N. Furman-Matarasso, M. Ron and A. Avni (1999). Isolation of novel SUMO protein from tomato that suppresses EIX-induced cell death. *The Plant J.* **19**: 533-451.

Haupt, W., T. C. Fishcher, S. Winderl, P. Fransz and R. A. Torres-Ruiz (2001). The CENTROMERE1 (CEN1) region of *Arabidopsis thaliana*: architecture and functional impact of chromatin. *The Plant J.* **27**: 285-296.

Hsia, A-P. and P. S. Schnable (1996). DNA sequence analyses support the role of interrupted gap repair in the origin of internal deletions of the maize transposon, MuDR. *Genetics* **142**: 603-618.

Heirichs, V., and B. S. Baker (1995). The *Drosophila* SR protein RBP1 contributes to the regulation of *Doublesex* alternative splicing by recognizing RBP1 RNA target sequences. *EMBO J.* **14**: 3987-4000.

Henikoff, S., E. A. Greene, S. Pietrokovski, P. Bork, T. K. Attwood, and L. Hood (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**: 609-614.

Henning, W. (1999). Heterochromatin. *Chromosoma* **108**: 1-9.

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.

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

Hong, S. -W, and E. Vierling (2000). Mutants of *Arabidopsis thaliana* defective in the acquisition of tolerance to high temperature stress. *Proc. Natl. Acad. Sci. USA.* **97**: 4392-4397.

Hsia, A. -P. and P. S. Schnable (1996). DNA sequence analyses support the role of interrupted gap-repair in the origin of internal deletions of the maize transposon, MuDR. *Genetics* **142**: 603-618.

Hsieh, J. and A. Fire (2000). Recognization and silencing of repeated DNA. *Annu. Rev. Genet.* **34**: 187-204.

Hudson, M., C. Ringli, M. T. Boylan and P. H. Quail (1999). The FAR1 locus encodes a novel nuclear protein specific to phytochromeA signaling. *Genes Dev.* **13**: 2017-2027.

Ingham, L. D, W. W. Hanna, J. W. Baier and L. C. Hannah (1993). Origin of the main class of repetitive DNA within selected *Pennistum* species. *Mol. Gen. Genet.* **238**: 350–356.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Ito, T., Y. Matsui, T. Ago, K. Ota and H. Sumimoto (2001). Novel modular domain PB1 recognizes PC motif to mediate functional protein–protein interactions. *EMBO J.* **20**: 3938-3946.

Jacq, J. Alt-Morbe, B. Andre, W. Arnold, A. Bahr, *et al.* (1997). The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. *Nature* **387** (6632 Suppl): 75-78.

Jeddeloh, J. A., J. Bender and E. J. Richards (1999). The DNA methylation locus *DDM1* is required for maintenance of gene silencing in *Arabidopsis*. *Genes Dev.* 12: 1714-1725.

Jensen, S., M-P Gassama, and T. Heidmann (1999). Taming of transposable elements by homology-dependent gene silencing. *Nature Gent.* **21:** 209–212.

Jin, Y. -K. and J. L. Bennetzen (1994). Integration and nonrandom mutation of plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize. *Plant Cell* **6**: 1177-1186.

Jones, L., F. Patcliff and D. C. Baulcombe (2001). RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires *Met1* for maintenance. *Curr. Boil.* **11**: 747-757.

Jordan, I. K., and J. F. Macdonald (1999). Tempo and mode of *Ty* element evolution in *Saccharomyces cerevisiae. Genetics* **151**: 1341-1351.

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. USA.* **97**: 6603-6607.

Kass, S. U., D. Pruss and A. P. Wolffe (1997). How does DNA methylation repress transcription? *Trends Genet.* **13**: 444-449.

Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University Sequencing Consortium, The European Union Arabidopsis Genome Sequencing Consortium and Institute of Plant Genetics and Crop Plant Research (IPK) (2000). Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana. Nature* **408**: 823-826.

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.

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

Kidwell, M. G. and D. R. Lisch (2001). Transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**: 1-24.

Kidwell, M. G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49-63.

Kim, K., S. H. Baek, Y. J. Jeon, S. Nishimori, T. Suzuki *et al.* (2000). A new SUMO-1-specific protease, SUSP1, that is highly expressed in reproductive organs. *J. Biol. Chem.* **275**: 14102-14106.

Kim J. M., S. Vanguri, J. D. Boeke, A. Gabriel, and D. F. Voytas (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8**: 464-474.

Kolkmn, J. A. W. P. C. Stememr (2001). Directed evolution of proteins by exon shuffling. *Nature biotechnology* **19**: 423-428.

Kumekawa, N., T. Hosouchi, H. Tsuruoka and H. Kotani (2000) The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5. *DNA Res.* **7**: 315-321.

Kumar, A. and J. L. Bennetzen (1999). Plant retrotransposons. *Annu Rev Genet.* **33**: 479-532.

Kunze, R, H. Saedler and W. E. Lönnig (1997). Plant transposable elements. *Adv. Bot. Res.* **27**: 331–470.

Kunze, R. H. (1996). The maize transposable elements *Activator* (*Ac*). *Curr. Top. Microbiol. Immunol.* **204**: 161-194.

Kurepa, J., J. M. Walker, J. Smalle, M. M. Gosink, S. J. Davis *et al.* (2003). The small Ubiquitin-like Modifier (SUMO) protein modification system in *Arabidopsis*. J. *Biol.Chem.* **278:** 6862 – 6872.

Labrador, M. and V. G. Corces (1997). Transposable elements-host interactions: regulation of insertion and excision. *Annu. Rev. Genet.* **31:** 3811-404.

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

Levy, A. A., M. Fridlender, U. H. E. Rubin, and Y. Sitrit (1996). Binding of *Nicotiana* nuclear proteins to the subterminal regions of the *Ac* transposable element. **251**: 436-441.

Li, S. -J. and M. Hochstrasser (1999). A new protease required for cell-cycle progression in yeast. *Nature* **138**:246-251.

Li, S. -J. and M. Hochstrasser (2000). The yeast ULP2 (SMT4) gene encodes a novel protease specific for the ubiquitin-like Smt3 protein. *Mol. Cell. Biol.* **20**: 2367-2377.

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko (2001). Evolutionary analyses of the human genome. *Nature* **409**: 847-849.

Lim, J. K. (1988). Intrachromosomal rearrangements mediated by *Hobo* transposons in *Drosphila melanogaster*. *Proc. Natl. Acad. Sci. USA.* **85***:* 9153-9157.

Lin, X., S. Kaul, S. Roundsley, T. P. Shea, M. -I. Benito *et al.* (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**: 761-768.

Lindroth, A. M., X. Cao, J. P. Jackson, D. Zilberman, C. M. McCallum (2001). Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**: 2077-2080.

Lisch, D., P. Chomet and M. Freeling (1995). Genetic characterization of the *Mutator* system in maize: behavior and regulation of *Mu* transposons in a minimal line. *Genetics* **139**: 1777-1796

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.

Lisch, D. R., M. Freeling, R. J. Langham and M. Y. Choy (2001). *Mutator* transposase is widespread in the grasses. *Plant Physiol.* **125**: 1293-1303.

Long, M. (2001). Evolution of novel genes. *Curr. Opin. Genet. Dev.* **11**: 673-680.

Lonnig W-E. and H. Saedler (1997). Plant transposons: contributors to evolution? *Gene* **205**: 245–253.

Lonnig, W-E and H. Saedler (2002). Chromosomal rearrangements and transposonable elements. *Annu. Rev. Genet.* **36**: 389–410.

Lopato, S., R. Gattoni, G. Fabini, J. Stevenin, and A. Barta (1999). A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. *Plant Mol. Biol.* **39**: 761-773.

Lynch, M. and J. Conery (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* **226**: 792–801.

McClintock, B. (1948). Mutable loci in maize. *Carnegie Institution of Washington Year Book* **45**: 176-186.

MacDonald, J. (1993). Evolution and consequence of transposable elements. *Curr. Opin. Genet.* Dev. **3**: 855-864.

Makalowski, W., G. A., and D. Labuda (1994). *Alu* sequences in the coding regions of mRNA, s source of protein variability. *Trends Genet.* **10**: 188-193.

Maes, T., P. De Keukeleire, and T. Gerats (1999). Plant tagnology. *Trends in Plant Science* **4**: 90-96.

Magyar, Z, T., Meszaros, P. Miskolczi, M. Deak, A. Feher, S. Brown, E. Kondorosi, *et al.* (1997). Cell cycle phase specificity of putative cyclin-depedent kinase variants in synchronized alfalfa cells. *Plant Cell* **9**: 223-235.

Martienssen, R. A. and A. Baron (1994). Coordinate suppression of mutations caused by Robertson's *Mutator* transposons in maize. *Genetics* **136**: 1157-1170.

Martienssen R. A. and V. Colo (2001). DNA methylation and epigenetic inheritance in plant and filamentous fungi. *Science* **293**: 1070-1074.

Matzke, M., A. J. M. Matzke and J. M. Kooter (2001). RNA: guiding gene silencing. *Science* **293**: 1080-1083.

Mayer, K., C. Schuller, R. Wambutt, G. Murphy, G. Volckaert *et al.* (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402**: 769-777.

McDonald, J. F. (1993). Evolution and consequences of transposable elements. *Curr. Opin. Genet. Dev.* **3**: 855-864.

Meinke, D. W., J. M. Cherry, C. Dean, S. D. Rounsley, and M. Koornneef (1998). *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282**: 662-682.

Melchior, F. (2000). SUMO-nonclassical ubiquitin. *Annu. Rev. Cell Dev. Biol.* **16**: 591-626.

Melek, M., M. Gellert, M. and D. C. van Gent (1998). Rejoining of DNA by the RAG1 and RAG2 proteins. *Science* **280**: 301–303.

Metz, A. M., R. T. Timmer and K. S. Browning (1992). Sequences for two cDNAs encoding *Arabidopsis thaliana* eukaryotic protein synthesis initiation factor 4A. *Gene* **120:** 313-314.

Meyer, P. and H. Saedler (1996). Homology-dependent gene silencing in plants. *Plant Mol. Biol.* **47**: 23-48.

Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada *et al.* (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* **411**: 212-214.

Morgenstern, B. (1999). DIALIGN 2, improvement of the segment-tosegment approach to multiple sequence alignment. *Bioinformatics* **15**: 211-218.

Moran, J. V., R. J. DeBerardinis, and H. H. Kazazian (1999). Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.

Muskens, M. W., A. P. Vissers and J. M. Kooter (2000). Role of inverted DNA repeats in transcriptional and post-transcriptional gene silencing. *Plant Mol. Biol.* **43**: 243-260.

Nakayashil, I., H., K. Ikeda, Y. Hashimoto, Y. Tosa and S. Mayama (2001). Methylation is not the main force repressing the retrotransposon MAGGY in *Magnaporthe grisea*. *Nucleic Acids Res.* **29**: 1278-1284.

Nari, J., G. Noat, and J. Ricard (1991). Pectin methylesterase, metal ions and plant cell-wall extension. Hydrolysis of pectin by plant cell-wall pectin methylesterase. *Biochem. J.* **279:** 343-350.

Nicholas, K. B., H. B. Nicholas Jr., and D. W. Deerfield (1997). GeneDOC: Analysis and visualization of genetic variation. *EMB. News* **4**: 14.

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.

Ohno, S. (1970) Evolution by gene duplication. Springer-Verlag, New York.

Ohtsubo, E., and Y. Sekine (1996). Bacterial insertion sequences. In *Transposable Elements*, edited by H. Saedler and A. Gierl. Springer-Verlag, Berlin.

Palmgren, M. G. (1994). Capture of host DNA by a plant retroelement:*BS1* encodes plasma membraneH^+-ATPase domains. *Plant Mol. Biol.* **25**: 137-140.

Pâques, F. and J. E. Haber (1999). Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **63**: 349-404.

Pickeral, O. K., W., Makalowski, M S. Boguski, and J. D. Boeke (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genet. Res.* **10**: 411–415.

Pimpinelli S, M. Berloco, L. Fanti, P. Dimitri P, and S. Bonaccorsi (1995). Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. *Proc Natl Acad Sci USA*. **92**: 3804-3808.

Plasterk, R. A. (1994). Capturing of host DNA by a plant retroelement: *Bs1* encodes plasma membrane H+-ATPase domains. In *mobile genetic elements*, edited by D. J. Sherratt. IRS press, Oxford.

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

Rajavashisth, T. B., A. K. Taylor, A. Analibi, K. L. Svenson, and L. J. Lusis (1989). Identification of a Zinc Finger protein that binds to sterol regulatory elements. *Science* **245**: 640-643.

Raizada, M. N. and V. Walbot (2000). The late developmental pattern of *Mu* transposon excision Is conferred by a cauliflower mosaic virus 35S–driven MURA cDNA in transgenic Maize. *Plant Cell* **12**: 5-22.

Raizada, M. N., M. –I. Benito and V. Walbot (2000). The late developmental pattern of *Mu* transposon excision is conferred by a cauliflower mosaic virus 35S-driven MURA cDNA in transgenic maize. Plant Cell **12:** 5–21.

Raizin, A., and H. Cedar (1977). Distribution of 5-methylcytosine in chromatin. *Proc. Natl. Acad. Sci. USA*. **74**: 2725-2728.

Rajavashisth, T. B., A. K. Taylor, A. Analibi, K. L. Svenson, and L. J. Lusis (1989). Identification of a Zinc Finger protein that binds to the sterol regulatory elements. *Science* **245**: 640-643.

Razin, A. (1998). CpG methylation, chromatin structure and gene silencing. *EMBO J.* **17**: 4905-4908.

Remacle, J. E., H. Kraft, W. Lerchner, G. Wuytens, C. Collart *et al.* (1999). New mode of DNA binding of multi-zinc finger transcription factors: EF1 family members bind with two hands to two target sites. *EMBO J.* **18**: 5073-5084.

Richard, L., L-X. Qin, P. Gadal, and R. Goldberg (1994). Molecular cloning and characterization of a putative pectin methylesterase cDNA in *Arabidopsis thaliana* (L.). *FEBS Lett.* **355**: 135-139.

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

Roberston, D. S. and P. S. Stinard and M. P. Maguire (1994). Genetic evidence of *Mutator*-induced deletions in the short arm of chromosome 9 of maize. II. *wd* deletions. *Genetics* **136**: 1143-1149.

Robertson, K. D., S. Ait-Si-Ali, T. Yokochi, P. A. Wade, P. L. Jones and A. P. Wolffe (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. *Nat. Genet.* **25**: 338-342.

Robertson, K. D. (2001). DNA methylation, methyltransferases, and cancer. *Oncogene* **20**:3139-3155.

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

Ruberti, I, G. Sessa, S. Lucchetti, and G. Morelli (1991). A novel class of plant proteins containing a homeodomain with a closely linked leucine zipper motif. *EMBO J.* **10**: 1787-1791.

Rubin, E. and A. A. Levy (1997). Abortive gap repair: underlying mechanism for *Ds* element formation. *Mol Cell Biol.* **17**: 6294-6302.

Rudenko, G. N. and V. Walbot (2001). Expression and post-transcriptional regulation of maize transposable element MuDR and its derivatives. *Plant Cell* **13**: 553-570.

Sacco, M. A., D. M. Flannery, K. Howes and K. Venugopal (2000). Avian endogenous retrovirus EAV-HP shares regions of identity with avian leukosis virus subgroup J and the avian retrotransposon ART-CH. *J. Virol.***74**:1296-1306.

Schena, M. and R. W. Davis (1994). Structure of homeobox-leucine zipper genes suggests a model for the evolution of gene families. *Proc. Natl. Acad. Sci. USA.* **91**: 8393-8397.

Shapiro, J. A. (1995). The discovery and significance of mobile genetic elements. In *mobile genetic elements*, edited by D. J. Sherratt. IRS press, Oxford.

Sherrie, R. T. and P. Wolffe (1993). Dual roles for transcription and translation factors in the RNA storage particles of Xenopus oocytes. *Trends in Cell Bio.* **3**: 94-98.

Singer, T., C. Yordan and R. A. Martienssen (2001). Robertson's *Mutator* transposons in *A. thaliana* are regulated by the chromatin-remodeling gene decrease in DNA methylation (DDM1). *Genes Dev.* **15**:591-602.

Steinemann, S. and M. Steinemann (2001). Biased distribution of repetitive elements: a landmark for neo-Y chromosome evolution in *Drosophila Miranda*. *Cytogenet. Cell Genet.* **93**: 228-233.

Steinemann, M, S. Steinemann and F. Lottspeich (1993). How Y chromosomes become genetically inert. *Proc. Natl. Acad.USA.* **15**: 5737-541.

Steinemann, M. and S. Steinmann (1997). The enigma of Y chromosome dgeneration: TRAM, a novel retrotransposon is preferentially located on the neo-Y chromosome of *Drosophila miranda*. *Genetics* **145**: 261-266.

Strommer, J. and D. Ortiz (1989). *Mu1*-mediated mutant alleles of maize exhibit background-dependent changes in expression and processing. *Dev. Genet.* **10**: 452-459.

Surani, M. A. (1998). Imprinting and the initiation of gene silencing in the germ line. *Cell* **93**:309-312.

Suzuki, T., A. Ichiyama, H. Saitoh, T. Kawakami, and M. Omata, *et al.* (2000). A new 30-kDa ubiquitin-related SUMO-1 hydrolase from bovine brain. *Biol. Chem.* **274**: 3131-3134.

Syvanen, M. (1984). The evolutionary implications of mobile genetic elements. *Annu. Rev. Genet.* **18**: 271–293.

Takahashi, S, Y. Inagaki, H. Satoh, A. Hoshino and S. Iida (1999). Capture of a genomic HMG domain sequence by the *En/Spm*-related transposable element Tpn1 in the Japanese Morning Glory. *Mol. Gen. Gent.* **261**: 447-451.

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

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

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.

Terasawa, I., Y. Noda, T. Ito, H. Hatanaka, and S. Ichikawa *et al.* (2001). Structure and ligand recognition of the PB1 domain: a novel protein module binding to the PC motif. *EMBO J.* **20**: 3947-3956.

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

Thompson, J. D., D. G. Higgins and T. J. Ginson (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.

Thomson, K. G., J. E. Thomas, and R. G. Dietzgen (1998). Retrotransposon-like sequences integrated into the genome of pineapple, *Ananas comosus*. *Plant Mol. Biol.* **38**: 461-465.

Tijsterman, M., R. F. Ketting, and R. H. A. Plasterk (2002). The genetics of gene silencing. *Annu. Rev. Genet.* **36**: 489–519.

Tseng, J. C., S. Zollman, S., A. C. Chain, and F. A. Laski (1991). Splicing of the Drosophila P element ORF2-ORF3 intron is inhibited in a human cell extract. *Mech. Dev.* **35**: 65-72.

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

van Steensel, B. and S. Henikoff (2001). Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27**: 304-308.

Vogt, V. M. (1997). Retroviral virions and genomes. In *Retroviruses*, edited by J. M. Coffin *et al.*, Cold Spring Harbor Laboratory Press, USA.

Vongs, A., T. Kakutani, R. A. Martienssen and E.J. Richards (1993). *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**: 1926-1928.

Walbot, V. (1986). Inheritance of *Mutator* activity in *Zea mays* as assayed by somatic instability of the bz2-mu1 allele. *Genetics* **114**: 1293-1312.

Walbot, V. (1991). The *Mutator* transposable element family of maize. *Curr. Top. Genet. Eng. Curr. Top. Genet. Eng.* **13**: 1-37.

Walbot, V. (1992). Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43**: 49-82.

Walbot, V. and A. Stapleton (1998). Reaction of potential of epigenetically inactive *Mu* transposonable elements of *Zea mays* L. *Maydica* **43**: 183-193.

Walbot, V. (2000). Saturation mutagenesis using maize transposons. *Curr. Opp. in Plant Biol.* **3**: 103-107.

Walbot, V. and G. N. Rudenk (2002). MuDR/*Mu* transposable elements of maize. In: *Mobile DNA II*, eds. N. L. Craig, R, *etc.* Amer. Soc. Microbiology,Washington, D. C.

Weiler, K. S. and B. T. Wakimato (1995). Heterochromatin and gene expression in Drosophila. *Annu. Rev. Genet.* **29**: 577-605.

Wright, S. I., Q. -H. Le, Q.-H., D. J. Schoen and T. E. Bureau (2001). Population dynamics of an Ac-like transposable element in self- and cross-pollinating *Arabidopsis*. *Genetics* **158**: 1279-1288.

Woffle, A. P. and M. A. Matzke (1999). Epigenetics: regulation through repression. *Sciecne* **286**: 481-486.

Wu-Scharf, D., B. Jeong, C. Zhang, C. and H. Cerutti (2000). Transgene and transposon silencing in *Chlamydomonas reinhardtii* by a DEAH-Box RNA helicase. *Science* **290**: 1159–1162.

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

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

Zhao, Z-Y., and V. Sundaresan (1991). Binding sites for maize nuclear proteins in the terminal inverted repeats of the *Mu*1 transposable element. *Mol. Gen. Genet.* **229**: 17-26.

Zhou, A., E. B. Cambareri, and J. A. Kinesy (2001). DNA methylation inhibits expression and transposition of *Neurospora Tad* retrotransposition. *Mol. Genet. Genomics* **265**: 748-754.

Zuker, C., J. Cappello, H. F. Lodish, P. George, and S. Chung (1984). Dictyostelium transposable element *DIRS-1* has 350-base-pair inverted terminal repeats that contain a heat shock promoter. *Proc.Natl. Acad. Sci. USA.* **81**: 2660-2664.

Transposon diversity in *Arabidopsis thaliana*

Quang Hien Le*, Stephen Wright*, Zhihui Yu, and Thomas Bureau†

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

Communicated by Margaret G. Kidwell, University of Arizona, Tucson, AZ, April 7, 2000 (received for review January 7, 2000)

Recent availability of extensive genome sequence information offers new opportunities to analyze genome organization, including transposon diversity and accumulation, at a level of resolution that was previously unattainable. In this report, we used sequence similarity search and analysis protocols to perform a fine-scale analysis of a large sample (≈17.2 Mb) of the *Arabidopsis thaliana* (Columbia) genome for transposons. Consistent with previous studies, we report that the *A. thaliana* genome harbors diverse representatives of most known superfamilies of transposons. However, our survey reveals a higher density of transposons of which over one-fourth could be classified into a single novel transposon family designated as *Basho*, which appears unrelated to any previously known superfamily. We have also identified putative transposase-coding ORFs for miniature inverted-repeat transposable elements (MITEs), providing clues into the mechanism of mobility and origins of the most abundant transposons associated with plant genes. In addition, we provide evidence that most mined transposons have a clear distribution preference for A + T-rich sequences and show that structural variation for many mined transposons is partly due to interelement recombination. Taken together, these findings further underscore the complexity of transposons within the compact genome of *A. thaliana*.

Transposons are fundamental components of most eukaryotic genomes, with important contributions to their size, structure, and variation. Based on mode of mobility, transposons are divided into two classes. Class I transposons move through an RNA intermediate and are reverse transcribed before their integration at another location in the genome. Retroelements can be further subdivided into retrotransposons (e.g., *copia*-like and *gypsy*-like) which are flanked by long terminal repeats (LTRs) and non-LTR retroelements (e.g., long and short interspersed nuclear elements). Class II elements are characterized by terminal inverted repeats (TIRs) and move directly through a DNA form by a "cut and paste" mechanism (1). Structural features, shared sequence similarity, and the size and sequence of the target site duplication (TSD) generated upon insertion, serve to further distinguish transposon superfamilies. As mutational agents, much of transposon impact may be deleterious to their hosts, although some insertions appear to play a significant role in adaptive evolution (2–4).

Historically, transposon discovery and analysis have been primarily conducted through the molecular genetic characterization of transposon-induced morphological mutations (1). Whereas these studies have allowed for the characterization of many mobile element groups, they do not allow for fine-scale investigation into the extent of transposon diversity and into the forces driving this variability. As genome sequencing projects increase in scale and number, detailed analysis of the patterns of transposon diversity and abundance, and their contribution to genome organization, becomes possible (5). The evidence thus far indicates that these patterns may be extremely variable among eukaryotic genomes (6, 7), suggesting the importance of such studies across diverse organisms.

The genome of the model plant *Arabidopsis thaliana* is small, with a correspondingly low repetitive DNA content relative to other higher plants (8), and is currently targeted for complete sequencing. Previous studies using computer-based sequence similarity searches have revealed numerous repetitive elements within the *Arabidopsis* genome (9, 10). Similarly, the complete sequences of *Arabidopsis* chromosomes 2 and 4 have also allowed for the identification of transposon-related sequences (11, 12). In this report, we performed a systematic and fine-scale search of *Arabidopsis* genome sequences to identify and characterize transposons. Our study differs from previous mining attempts in that we not only have compiled repetitive sequences but also provide evidence that many are in fact transposons by structural analysis and demonstration of past mobility. In this way, we provide insight into *Arabidopsis* transposon structure, mobility, distribution, and diversity.

Materials and Methods

Transposon Mining. *Arabidopsis* genomic sequences from 243 clones (approximately 17.2 Mb) with representation from all five linkage groups, were accessed from GenBank (National Center for Biotechnology Information, NCBI; http://www.ncbi.nlm.nih.gov/), between June and December 1998. Intergenic and intron sequences longer than 500 bp in length were used as BLAST search queries (version 2.0, http://www.ncbi.nlm.nih.gov/blast/) (13). Repetitive sequences with similarity to the query (score >80) were compiled into groups and used as queries in additional database searches. A total of 770 repetitive sequences belonging to 197 groups were identified. Of these, 444 mined sequences belonging to 135 groups were determined to be members of previously described transposon superfamilies by virtue of shared sequence composition and/or structural features such as TIRs, LTRs, a terminal poly(A)-rich sequence, coding capacity for mobility-related proteins, and flanking direct repeats (i.e., TSDs). Another 7 repetitive sequence groups representing 179 elements were determined to be mobile elements by documentation of a mobile history (see below). In many cases, some members of each mobile element group lacked one or both terminal sequences and were defined as truncated. Lastly, 147 repetitive sequences belonging to 55 groups could not be classified as transposons based on structural and sequence analysis. A detailed description of the mined mobile elements in our survey can be accessed from our relational database (http://soave.biol.mcgill.ca/clonebase/). A subset of the mined transposons was previously identified and/or annotated to be repetitive DNA, putative transposons, or transposon-related sequences by other researchers (6, 9–12, 14–22).

Data Analysis. When necessary, further sequence analysis and alignments were performed by using CLUSTAL W (23), DIVERGE, TRANSLATE, PILEUP, BESTFIT, and GAP from the University of Wisconsin Genetics Computing Group suite of programs (version 10.0) or additional blast search tools provided at NCBI (13,

Abbreviations: gi, geninfo identifier; IS, insertion sequence; MITE, miniature inverted-repeat transposable element; MLE, *Mariner*-like element; MULE, *Mutator*-like element; RESite, related to empty site; TIR, terminal inverted repeat; TSD, target site duplication; LTR long terminal repeat.

*Q.H.L. and S.W. contributed equally to this work.

†To whom reprint requests should be addressed. E-mail: thomas_bureau@maclan.mcgill.ca.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

24). Visualization of amino acid alignments was done by using GENEDOC (25). A PERL program was written to compile sequences immediately flanking transposon insertion and to calculate the average G + C content using a 20-bp sliding window (written by C. Olive, McGill University). Only flanking sequences of intact, and not truncated, elements were included in the calculations. As a control, 50 positions within intergenic regions were randomly selected and submitted to the program. A copy of this program is available upon request.

Documentation of Mobile History. Sequences immediately flanking the element were used as queries in database searches. Sequences sharing similarity to the queries typically represented either orthologous or, more frequently, paralogous regions (e.g., multigene families, transposons, duplicated genomic regions, or other repetitive sequences). In many cases, a pairwise comparison could be used to identify a gap corresponding to the absence of the insertion. This mined sequence with high nucleotide sequence similarity to the original query but lacking the insertion is referred to as a *r*elated to *e*mpty *s*ite (RESite). Examination of RESites also served to delimit element termini and to identify corresponding TSDs when this information could not be obtained from sequence and structural analysis. Alternatively, when a computer-assisted approach failed, RESites were amplified from other ecotypes of *Arabidopsis thaliana*, *Arabidopsis lyrata*, or *Brassica* spp. by using a previously described PCR approach (ref. 26; data not shown).

Results and Discussion

To obtain a complete transposon profile of the *Arabidopsis* genome, we systematically surveyed a large sample (≈17.2 Mb) of genomic sequences for transposon insertions by using sequence similarity search algorithms (13, 24). Demonstration of past mobility of the mined elements was provided through the identification of RESites, which are sequences similar to the empty site of an insertion (Fig. 1). This approach allows a rapid and efficient means to delimit element termini and highlight putative target site duplication events. We suggest that RESites provide convincing evidence that many mined interspersed repetitive sequences in *Arabidopsis* are in fact transposons. Together, mined repetitive sequences and their corresponding RESites strongly support a transposition-based insertion mechanism.

Based on shared sequence and structural similarities and the analysis of RESites, the majority of repetitive sequences mined in our survey (82%) could be compiled and categorized into 142 groups of putative transposons. The majority of the mined transposons corresponded to known superfamilies according to shared structural and sequence similarities (Table 1, Fig. 1*A*). Among the groups of mined transposons, 28 have members that are structurally reminiscent to the maize *Mutator* element and/or harbored ORFs with significant similarity to the maize *Mutator* transposase (i.e., MURA). *Mutator-*like *e*lements (MULEs) were in part defined as transposons with long TIRs (TIR-MULEs) as in the case of the maize *Mu* family of elements (30). However, some MULEs lack long TIRs (non-TIR-MULES) but have other features characterizing them as MULEs such as a 8- to 10-bp TSD and, for 5 elements, an ORF encoding a MURA-like transposase. In fact, sequence comparison and analysis of *Arabidopsis* TIR- and non-TIR-MULEs indicate that they share a common evolutionary history (Z.Y., S.W., and T.B., unpublished data).

Interestingly, over one-fourth of the elements identified from 7 distinct groups did not belong to any of the previously described superfamilies but appeared related based on common structural features. The identification of 9 RESites indicate that these elements, which we have named *Basho* (after the nomadic Japanese haiku poet), have defined termini (5′-CHH . . . CTAG-3′, where H = A, T, or C) and a target site preference for the mononucleotide "T." *Basho*-like elements have been previously described as repetitive sequences and putative insertion sequences (9, 16), but no evidence (e.g., RESites, defined element termini, and target site sequence) was presented to indicate whether they are in fact mobile elements. Curiously, *Basho* elements appear to insert in a preferred orientation relative to the sequence context of the target site, namely 5′-AT-3′ (Fig. 1*B*). In addition, we have identified a *Basho*-like group in maize (data not shown), suggesting that *Basho* defines a new superfamily of transposons. Although high sequence similarity and RESites attest to past mobility, the mechanism by which *Basho* elements have transposed is presently unknown.

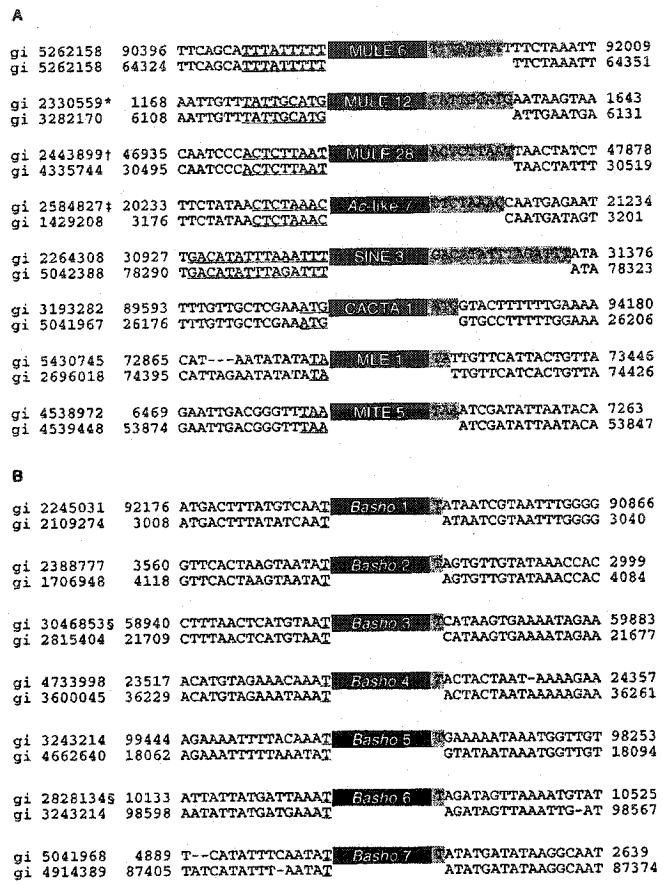


Fig. 1. RESites corresponding to mined *Arabidopsis* elements. (*A*) Examples of RESites for different groups of mined elements. All of the RESites illustrated were identified by computer-assisted database searches. In total, 34 RESites were found by computer-assisted database searches, and 13 were identified by cross-ecotype PCR analysis. The target sequences are underlined, and the TSDs are shaded. GenBank geninfo identifier (gi) numbers and nucleotide position on clones are indicated. *, Inserted into a *Basho* III element; †, inserted into a *Basho* III element; ‡, inserted into a MITE IX element. (*B*) RESites found for *Basho* insertions confirm mononucleotide TSD (shaded). §, Inserted into a *Basho* V element.

Our survey permitted an examination of the genomic patterns of transposons within *Arabidopsis.* For many eukaryotes, there appears to be a relationship between the number of transposons, primarily class I elements (i.e., LTR and non-LTR retrotransposons), and genome size (31). In very large genomes, for instance, the transposon content can account for the majority of the genome (32). In agreement with the small size of the *Arabidopsis* genome, approximately 5% of the genomic sequences surveyed were composed of transposons of which only 2% were class I elements. Consistent with previous estimates of

Table 1. Transposons in 17.2 Mb of the *Arabidopsis thaliana* (Columbia) genome

Type	Superfamily	Number of groups	Number of transposons
Class I	SINEs*	3	16
	LINEs†	28	31
	copia-like Retrotransposons	27	40
	gypsy-like Retrotransposons	23	45
	Undetermined‡	2	2
Class II	*Ac*-like	7	38
	CACTA-like	1	3
	MULEs	28	108
	MITEs	15	105
	MLEs	1	56
Class?	*Basho*	7	179
Total		142	623

*Short interspersed nuclear elements (SINEs) are defined as elements that lack coding capacity or have no similarity to coding regions and have either a putative pol III promoter, a long TSD, and/or a poly A+T-rich tail at one terminus (27, 28).

†Long interspersed nuclear elements (LINEs) are defined as having members with sequence similarity to the coding domains of previously reported LINEs.

‡These elements structurally resemble LTR-retrotransposons (i.e., an element with LTRs and 4-bp TSDs but no coding capacity and a putative solo-LTR) but lack signature sequences typical of *copia*-like or *gypsy*-like retrotransposons (29).

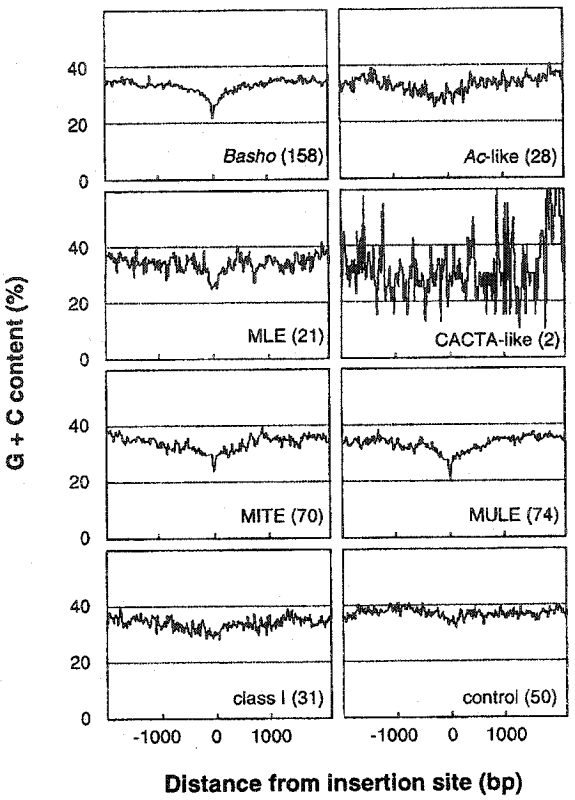


Fig. 2. A majority of mined transposons show an insertion preference for A + T-rich regions. The number of transposons used in the calculations is indicated in parentheses. The average G + C content was determined by using a 20-bp sliding window over 2000 bp of sequences flanking the transposon insertion sites. Only two CACTA-like elements were used, resulting in a low signal-to-noise ratio. Individual groups (see Table 1) of class I elements showed a similar profile (data not shown) as indicated for the entire class I average.

the retrotransposon content in *Arabidopsis* (10, 11, 18), we found significantly fewer class I elements than reported in the genomes of other higher plants (29). This contrast in abundance of transposon type between various genomes may suggest differential success depending on genomic environment.

The mined transposons were predominantly found in intergenic regions, with 5% located within introns of predicted genes, and approximately 8% were nested insertions. The prevalence of mined transposons in noncoding regions is similar to the patterns which have been observed in other organisms; for example, heterochromatic regions in *Drosophila* and intergenic regions in maize typically contain numerous nested transposons (32, 33). A previous report also suggests that transposons are highly enriched within centromeric heterochromatin in *Arabidopsis* (34). In this report, only four of the clones surveyed in our study correspond to centromeric heterochromatin. Therefore, a meaningful comparison between these regions and euchromatic regions could not be achieved. Whereas many insertions appeared to have a random target site sequence context, preferences for specific A + T-rich target sites were observed for three of the most abundant types of elements: *m*iniature *i*nverted-repeat *t*ransposable *e*lements (MITEs) (TA, TAA, ATA), *Basho* (T), and *Mariner*-like elements (MLE; TA) (Fig. 1). In addition to sequence-specific target sites, these elements appear to be distributed preferentially in A + T-rich regions. Up to 300 bp of sequences immediately flanking the site of insertion show a G + C content of ≈25% (Fig. 2), which is lower than previous estimations for the *Arabidopsis* genome [36.7% (35) and 35.8% (11)] and lower than observed in our control (Fig. 2). *Arabidopsis* MULEs, which do not appear to have target sequence preference are also preferentially distributed in A + T-rich regions.

For a subset of mined transposons, the high level of shared nucleotide sequence similarity observed (>90%) suggests that elements have been recently active. Shared sequence identity between flanking regions of elements and matching RESites also attest to recent mobility (Fig. 1). Although the majority of mined elements lacked any coding capacity, some members of mined groups were found harboring corresponding ORFs, such as reverse transcriptase, *Ac*-like transposase, CACTA-like transposase, or maize *Mutator* transposase (29, 30, 36). In addition, one group of mined elements (MLE I) not only shares structural features with the *Tc1/Mariner* transposon superfamily (Fig. 3*A*), but also has at least one member located on chromosome 2 that harbors an ORF with up to 46% amino acid sequence similarity with the transposase of *Tc1/Mariner*-like elements, *Pogo*R11, and *Tigger*1 (Fig. 3*B*). Furthermore, MLE I elements have the conserved terminal bases (5′-CAGT-3′) necessary for the efficient transposition of other typical *Tc1/Mariner*-like elements (37). Some members of the MLE I have previously been reported to belong to a novel family of MITEs, referred to as *Emigrant* (14), based on their small size and target site preference for the dinucleotide TA. However, the MLE I elements clearly have more in common with transposons of the *Tc1/Mariner* superfamily (Fig. 3) than to elements belonging to the MITE superfamily (38). The mined MLE I transposase shares no significant sequence similarity with two *Tc1/Mariner*-like transposases reported by Lin *et al.* (11) also on chromosome 2.

Whereas the mechanism of MITE mobility was previously unclear, we found evidence that some groups of MITEs have unusually large members that potentially encode for a transposase (Fig. 4*A*). Specifically, putative ORFs found in members of two distinct MITE groups (X and XI) share 51% amino acid sequence similarity (Fig. 4*B*). In addition, the ratio of synonymous to nonsynonymous nucleotide substitutions of 2.82 suggests the possibility of functional constraint on this coding sequence. These putative ORFs also share sequence similarity with the transposases of cyanobacterial and other prokaryotic *i*nsertion *s*equences (*IS*) (Fig. 4*C*). Because MITE X and XI

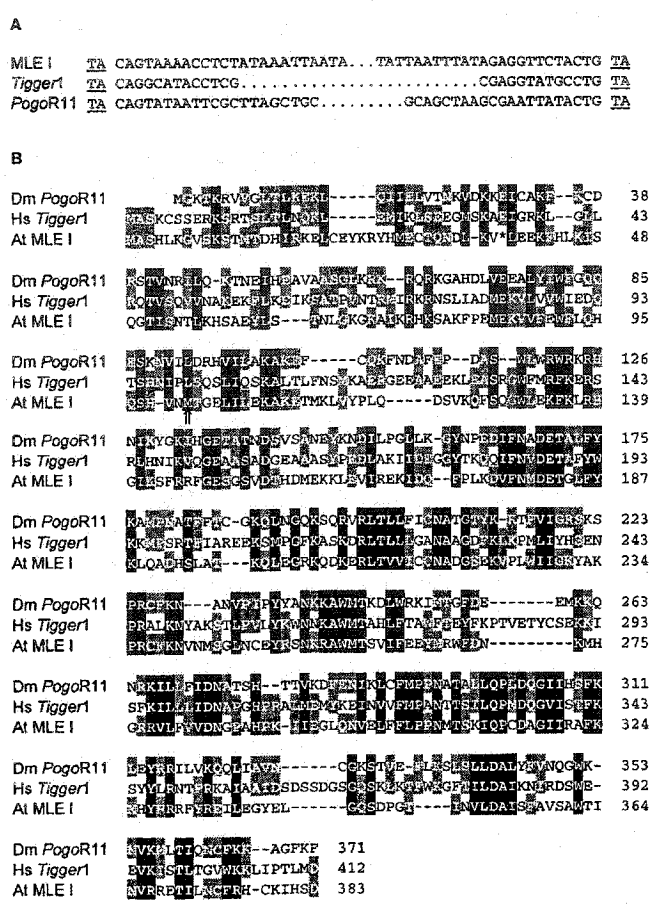


Fig. 3. Structure of an *Arabidopsis* Tc1/*Mariner*-like transposon. (*A*) Similarities between TIRs and TSDs (underlined) of an *Arabidopsis* MLE I member and Tc1/*Mariner*-like elements *Pogo* (Drosophila, gi 8354) and *Tigger* (human, gi 2226003). (*B*) Putative transposase for the *Arabidopsis* MLE I (gi 4262216) aligned with transposases from *Drosophila melanogaster Pogo*R11 (gi 2133672) and from human *Tigger*1 (gi 2226004). Amino acid residues are shaded based on the level of structural and functional similarities. Residues conserved between three or two sequences are shaded black and gray, respectively. The arrow (⇑) indicates the predicted start of the *Arabidopsis* MLE I ORF as annotated in GenBank. The first methionine of the *Arabidopsis* MLE I transposase was inferred from the reading frame and sequence similarity with the human *Tigger*1 element. The stop (*) was introduced by a single nucleotide substitution (at position 85709 in gi 4262209) from GAG (glutamine) to TAG (stop).

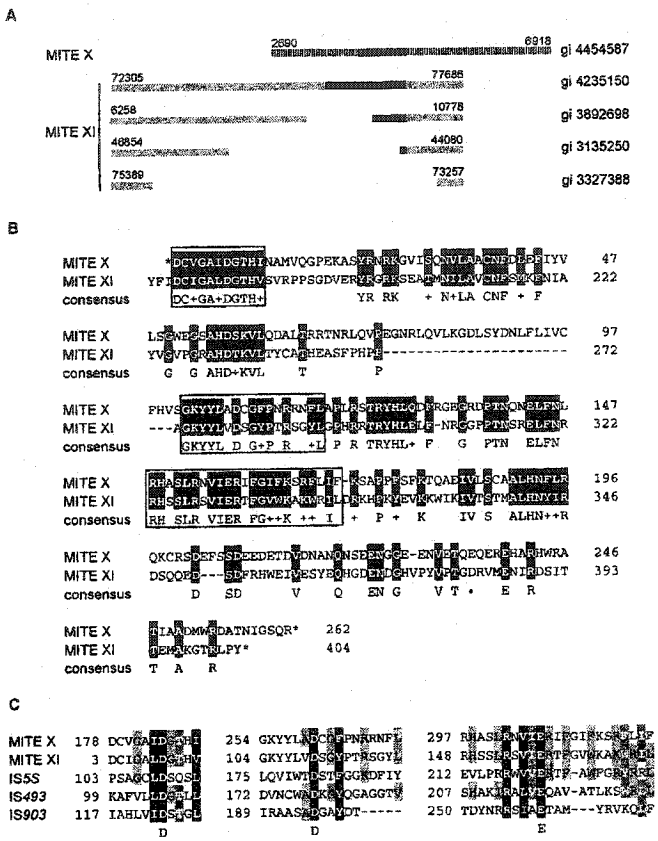


Fig. 4. MITE transposases (*A*). Diagram depicting structural similarities between members of MITE XI elements (gray boxes) and a member of MITE X (striped box). Black boxes represent ORFs corresponding to a putative transposase: MITE XI ORFs share >98% amino acid sequence similarity. MITE X and XI are distinct groups because no significant internal nucleotide sequence similarity is observed between MITE X and XI nor for the consensus sequence of their TIRs (5′-GG(G/T)GGTGTTATTGGTT-3′ for MITE X and 5′-GGCCCTGTTTTGTTTG-3′ for MITE XI). However, putative transposase ORFs, length of TIRs (16 bp), and TSD (5′-TTA-3′) of MITE X and XI are similar, indicating a related mechanism of transposition and possibly a common origin for both groups. GenBank gi numbers of the clones from which elements were mined are indicated to the right. Nucleotide positions of transposon termini are indicated above each element. (*B*) Amino acid sequence similarity between the conceptual translation for MITE X (gi 4454587, nucleotide position 4650–5865) and the corresponding region of MITE XI ORF (gi 4585884). Identical amino acids and functionally or structurally related residues are shaded in black. *, Translational stops. Boxed sequences indicate the region containing the DDE motif. (*C*) Alignment of conserved regions corresponding to the functionally important DDE motif found in transposases and integrases of many transposable elements (39–42). Transposases are from MITE X (gi 4454587, conceptual translation of nucleotides 4650–5865), MITE XI (gi 4585884), IS*5*5 (gi 1256580) from *Synechocystis* sp., IS*493* (gi 1196467), and IS*903* (gi 136129) from *Escherichia coli*. Amino acid residues are shaded based on the level of structural and functional similarities. Residues conserved between all, four, or three sequences are shaded black, dark gray, and light gray, respectively.

members are structurally similar to other MITE groups in *Arabidopsis*, as well as the *Tourist*-like elements in grasses (43), we suggest that the MITE X and XI ORFs represent the transposases characteristic of the *Tourist*-like family of elements.

We observe extensive diversity among mined elements, both within and among groups. The large number of mined transposon groups (Table 1) illustrates a high level of evolutionary divergence and is indicative of long-term persistence or the result of horizontal transfer (44). Differences in size between related elements suggests that insertions and deletions may account for a significant proportion of variation within groups and is reflected by the high occurrence of apparently truncated elements (38%). In addition, interelement recombination appears to generate sequence diversity among some of the mined elements. For example, the mosaic structure of members of some MULE groups is due to the exchange of terminal sequences (Fig. 5). Such mosaic elements are apparently capable of transposition, as suggested by both their copy number, sequence similarity, and presence of perfect TSDs. Another possible factor driving sequence diversity in MULEs is the acquisition of truncated cellular genes. For example, a member of MULE I harbors sequences that share ≈85% nucleotide sequence similarity to a region spanning exon 1, intron 1, and exon 2 of the *Arabidopsis* homeobox gene *Athb*-1 (Fig. 6; ref. 45; Z.Y., S. W., and T.B., unpublished data). Despite the sequence diversity observed, some conserved motifs among elements exist, which suggests a common interaction with host factors (e.g., general transcription factors and gene regulatory proteins). We have identified a subterminal sequence motif (5′-TTTTCCCGCCAAAA-3′) shared between MITE XI and *Ac*-like VII. This sequence may be

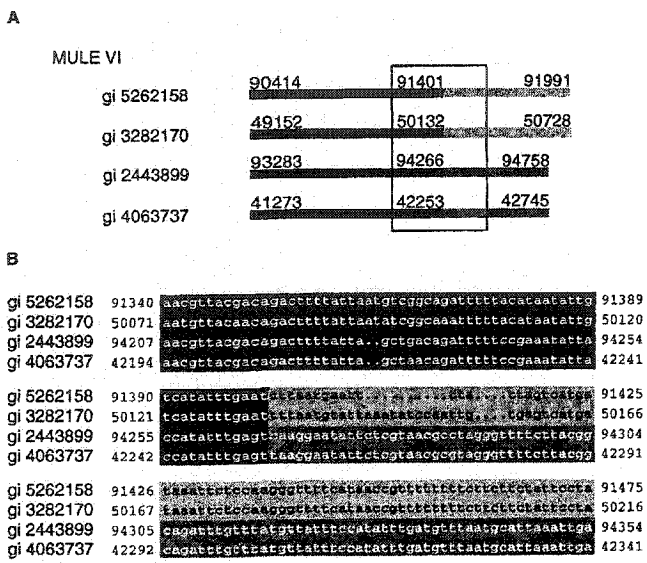


Fig. 5. Interelement recombination in non-TIR-MULEs generates mosaic transposons. (*A*) Diagrammatic alignment of four members of MULE VI. Black and gray boxes delimit regions of nucleotide sequence similarity (>90%) between members of MULE groups VI. GenBank gi numbers are marked to the right. Nucleotide positions on clones are noted above each element. (*B*) Expanded view (as boxed in *A*) indicating the sequence of the putative recombination region.

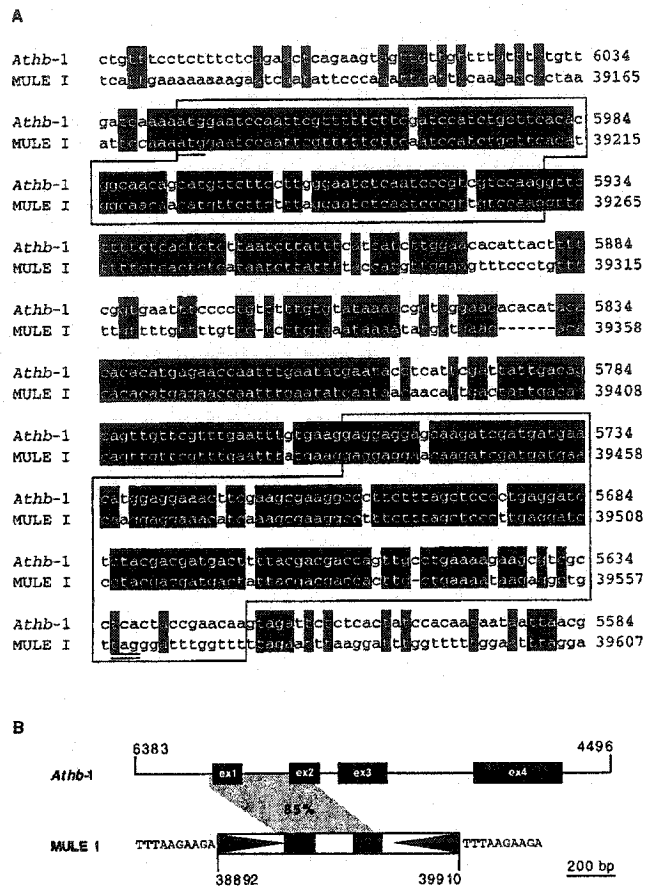


Fig. 6. Acquisition of a truncated cellular gene by a member of MULE-I. (*A*) MULE-I (gi 2182289) shares 85% nucleotide similarity with the first two exons and first intron of the homeobox gene *Athb*-1 (gi 6016704). Conserved nucleotides are shaded in black, and positions on clones are indicated on the right of the alignment. Sequences for the first two exons are boxed. The first ATG is underlined, and an in-frame stop codon of the MULE-I ORF is double underlined. (*B*) Diagram illustrating the regions of nucleotide similarity (85%) between the MULE-I and the genomic sequence of *Athb*-1 (shaded in gray). Positions for the *Athb*-1 gene and MULE-I element on their respective clones are indicated. MULE-I TIRs are represented by black triangles, and the TSD sequences are indicated flanking both termini.

analogous to the motifs shared between the maize *Ac* and *Mutator*, which are thought to be important in the regulation of mobility (46). Furthermore, a motif determined to be a matrix attachment region (5′-CAAATTTATTTTTATA-3′) (47) within a member of MULE XXIX is conserved among all group members.

Our report documents the features and characteristics of the surprisingly diverse forms of transposons residing in, for the most part, noncoding regions of the relatively small and compact genome of *Arabidopsis*. Lack of insertions into coding regions and prevalence in A + T-rich sequences/regions may reflect purifying selection against deleterious mutations and/or insertion preference of many mined elements. In either case, richness of elements in nongenic regions may represent merely genomic "junk," or, in some cases, elements may have a host function, as has been observed in other organisms (2). Clearly, the actual impact of transposons in genome evolution will require further molecular dissection of the factors important in gene expression, as well as analysis of population dynamics of element insertions. Our survey also provides additional clues as to the origin and evolution of transposons. Significant similarity between transposases of bacterial insertion sequences and the mobility-related proteins found in this (MITE X and XI) and other studies (*Mutator* or MULEs, *Tc1*/*Mariner*, and retroelements) (39–41, 48) suggests that the majority of element superfamilies are ancient components of genomes. In addition, recombination appears to be important in giving rise to novel forms of MULEs and possibly other types of elements in *Arabidopsis*.

By providing a fine-scale approach in the mining of transposons from sequence databases, we have demonstrated that the abundance and variation of transposons in *Arabidopsis* is higher than previously reported. This observation obviously highlights the need for continued analysis of genomic sequence information to provide a high-resolution image of the *Arabidopsis* genome (49). Finally, information on transposon structures, mobile histories, and genomic coordinates presented in our report will expedite the development of novel biotechnologies for mapping purposes, systematic gene disruption, and functional analysis.

We thank Dr. Daniel J. Schoen for comments on the manuscript. We are grateful to Chris Olive, Chengbin Feng, Hala Razi Khan, and Sylva Petrova for setting up and maintaining our complete transposon profile database, which can be accessed at http://soave.biol.mcgill.ca/clonebase/. This work was funded by a National Science and Engineering Research Council of Canada grant to T.E.B.

1. Berg D. E. & Howe, M. M., eds. (1989) *Mobile DNA* (American Society for Microbiology, Washington, DC).
2. Britten, R. J. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 9374–9377.
3. Kidwell, M. G. & Lisch, D. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 7704–7711.
4. McDonald, J. F. (1995) *Trends Ecol. Evol.* **10**, 123–126.
5. Jordan, I. K. & McDonald, J. F. (1999) *Genetics* **151**, 1341–1351.
6. Jurka, J. (1998) *Curr. Opin. Struct. Biol.* **8**, 333–337.
7. Smit, A. F. A. (1996) *Curr. Opin. Genet. Dev.* **6**, 743–748.
8. Leutwiler, L. S., Hough-Evans, B. R. & Meyerowitz, E. M. (1984) *Mol. Gen. Genet.* **194**, 15–23.

9. Surzycki, S. A. & Belknap, W. R. (1999) *J. Mol. Evol.* **48,** 684–691.
10. Wright, D. A. & Voytas, D. F. (1998) *Genetics* **149,** 703–715.
11. Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M.-I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., *et al.* (1999) *Nature (London)* **402,** 761–768.
12. Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., Terryn, N., *et al.* (1999) *Nature (London)* **402,** 769–777.
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
14. Casacuberta, E., Casacuberta, J. M., Puigdomènech, P. & Monfort, A. (1998) *Plant J.* **16,** 79–85.
15. Bevan, M., Bancroft, I. Bent, E., Love, K., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema W., *et al.* (1998) *Nature (London)* **391,** 485–488.
16. Doutriaux, M. P., Couteau, F., Bergounioux, C. & White, C. (1998) *Mol. Gen. Genet.* **257,** 283–291.
17. Chye, M.-L., Cheung, K.-Y. & Xu, J. (1997) *Plant Mol. Biol.* **35,** 893–904.
18. Konieczny, A., Voytas, D. F., Cummings, M. P. & Ausubel, F. M. (1991) *Genetics* **127,** 801–809.
19. Pélissier, T., Tutois, S., Deragon, J. M., Tourmente, S., Genestier, S. & Picard G. (1995) *Plant Mol. Biol.* **29,** 441–452.
20. Tsay, Y.-F., Frank, M. J., Page, T., Dean, C. & Crawford, N. M. (1993) *Science* **260,** 342–344.
21. Klimyuk, V. I. & Jones, J. D. (1997) *Plant J.* **11,** 1–14.
22. Wright, D. A., Ke, N., Smalle, J., Hauge, B. M., Goodman, H. M. & Voytas, D. F. (1996) *Genetics* **142,** 569–578.
23. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
24. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
25. Nicholas, K. B., Nicholas, H. B., Jr. & Deerfield, D. W. II (1997) *EMBNEW. NEWS* **4,** 14.
26. Bureau, T. E. & Wessler, S. R. (1992) *Plant Cell* **4,** 1283–1294.
27. Gilbert, N. & Labuda, D. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2869–2874.
28. Yoshioka, Y., Matsumoto, S., Kojima, S., Oshima, K., Okada, N. & Machida, Y. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6562–6566.
29. Xiong, Y. & Eickbush, T. H. (1990) *EMBO J.* **9,** 3353–3362.
30. Walbot, V. (1991) in *Genetic Engineering*, Setlow, J. K., ed. (Plenum, New York), pp. 1–37.
31. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284,** 604–607.
32. SanMiguel, P., Tikhonov, A., Y.-K. Jin, Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., *et al.* (1996) *Science* **274,** 765–767.
33. Vaury, C., Bucheton, A. & Pelisson, A. (1989) *Chromosoma* **98,** 215–224.
34. Copenhaver, G. P., Nickel, K., Kuromori, T., Benito, M.-I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L. D., *et al.* (1999) *Science* **286,** 2468–2474.
35. Barakat, A., Matassi, G. & Bernardi, G. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 10044–10049.
36. Fedoroff, N. V. (1989) *Cell* **56,** 181–191.
37. Fischer, S. E. J., van Luenen, H. G. A. M. & Plasterk, R. H. A. (1999) *Mol. Gen. Genet.* **262,** 268–274.
38. Wessler, S. R., Bureau, T. E. & White, S. E. (1995) *Curr. Opin. Genet. Dev.* **5,** 814–821.
39. Grindley, N. D. F. & Leschziner, A. E. (1995) *Cell* **83,** 1063–1066.
40. Lohe, A. R., De Aguiar, D. & Hartl, D. L. (1997) *Proc. Natl. Acad. Sci. USA.* **94,** 1293–1297.
41. Capy, P., Vitalis, R., Langin, T., Higuet, D. & Bazin, C. (1996) *J. Mol. Evol.* **42,** 359–368.
42. Tavakoli, N. P., DeVost, J. & Derbyshire, K. M. (1997) *J. Mol. Biol.* **274,** 491–504.
43. Bureau, T. E. & Wessler, S. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1411–1415.
44. Kidwell, M. (1992) *Curr. Opin. Genet. Dev.* **2,** 868–873.
45. Ruberti, I., Sessa, G., Lucchetti, S. & Morelli, G. (1991) *EMBO J.* **10,** 1787–1791.
46. Becker, H.-A. & R. Kunze, R. (1996) *Mol. Gen. Genet.* **251,** 428–435.
47. van Drunen, C. M., Oosterling, R. W., Keultjes, G. M., Weisbeek, P. J., van Driel, R. & Smeekens, S. C. M. (1997) *Nucleic Acids Res.* **25,** 3904–3911.
48. Eisen, J. A., Benito, M.-I. & Walbot, V. (1994) *Nucleic Acids Res.* **22,** 2634–2636.
49. Karp, P. D. (1998) *Bioinformatics* **14,** 753–754.

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*

The Arabidopsis Genome Initiative*

Authorship of this paper should be cited as 'The Arabidopsis Genome Iniative'. A full list of contributors appears at the end of this paper

The flowering plant *Arabidopsis thaliana* is an important model system for identifying genes and determining their functions. Here we report the analysis of the genomic sequence of *Arabidopsis*. The sequenced regions cover 115.4 megabases of the 125-megabase genome and extend into centromeric regions. The evolution of *Arabidopsis* involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications, giving rise to a dynamic genome enriched by lateral gene transfer from a cyanobacterial-like ancestor of the plastid. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of *Drosophila* and *Caenorhabditis elegans*— the other sequenced multicellular eukaryotes. *Arabidopsis* has many families of new proteins but also lacks several common protein families, indicating that the sets of common proteins have undergone differential expansion and contraction in the three multicellular eukaryotes. This is the first complete genome sequence of a plant and provides the foundations for more comprehensive comparison of conserved processes in all eukaryotes, identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement.

The plant and animal kingdoms evolved independently from unicellular eukaryotes and represent highly contrasting life forms. The genome sequences of *C. elegans*[1] and *Drosophila*[2] reveal that metazoans share a great deal of genetic information required for developmental and physiological processes, but these genome sequences represent a limited survey of multicellular organisms. Flowering plants have unique organizational and physiological properties in addition to ancestral features conserved between plants and animals. The genome sequence of a plant provides a means for understanding the genetic basis of differences between plants and other eukaryotes, and provides the foundation for detailed functional characterization of plant genes.

Arabidopsis thaliana has many advantages for genome analysis, including a short generation time, small size, large number of offspring, and a relatively small nuclear genome. These advantages promoted the growth of a scientific community that has investigated the biological processes of *Arabidopsis* and has characterized many genes[3]. To support these activities, an international collaboration (the Arabidopsis Genome Initiative, AGI) began sequencing the genome in 1996. The sequences of chromosomes 2 and 4 have been reported[4,5], and the accompanying Letters describe the sequences of chromosomes 1 (ref. 6), 3 (ref. 7) and 5 (ref. 8).

Here we report analysis of the completed *Arabidopsis* genome sequence, including annotation of predicted genes and assignment of functional categories. We also describe chromosome dynamics and architecture, the distribution of transposable elements and other repeats, the extent of lateral gene transfer from organelles, and the comparison of the genome sequence and structure to that of other *Arabidopsis* accessions (distinctive lines maintained by single-seed descent) and plant species. This report is the summation of work by experts interested in many biological processes selected to illuminate plant-specific functions including defence, photomorphogenesis, gene regulation, development, metabolism, transport and DNA repair.

The identification of many new members of receptor families, cellular components for plant-specific functions, genes of bacterial origin whose functions are now integrated with typical eukaryotic components, independent evolution of several families of transcription factors, and suggestions of as yet uncharacterized metabolic pathways are a few more highlights of this work. The implications of these discoveries are not only relevant for plant biologists, but will also affect agricultural science, evolutionary biology, bioinformatics, combinatorial chemistry, functional and comparative genomics, and molecular medicine.

Overview of sequencing strategy

We used large-insert bacterial artificial chromosome (BAC), phage (P1) and transformation-competent artificial chromosome (TAC) libraries[9–12] as the primary substrates for sequencing. Early stages of genome sequencing used 79 cosmid clones. Physical maps of the genome of accession Columbia were assembled by restriction fragment 'fingerprint' analysis of BAC clones[13], by hybridization[14] or polymerase chain reaction (PCR)[15] of sequence-tagged sites and by hybridization and Southern blotting[16]. The resulting maps were integrated (http://nucleus/cshl.org/arabmaps/) with the genetic map and provided a foundation for assembling sets of contigs into sequence-ready tiling paths. End sequence (http://www.tigr.org/tdb/at/abe/bac_end_search.html) of 47,788 BAC clones was used to extend contigs from BACS anchored by marker content and to integrate contigs.

Ten contigs representing the chromosome arms and centromeric heterochromatin were assembled from 1,569 BAC, TAC, cosmid and P1 clones (average insert size 100 kilobases (kb)). Twenty-two PCR products were amplified directly from genomic DNA and sequenced to link regions not covered by cloned DNA or to optimize the minimal tiling path. Telomere sequence was obtained from specific yeast artificial chromosome (YAC) and phage clones, and from inverse polymerase chain reaction (IPCR) products derived from genomic DNA. Clone fingerprints, together with BAC end sequences, were generally adequate for selection of clones for sequencing over most of the genome. In the centromeric regions, these physical mapping methods were supplemented with genetic mapping to identify contig positions and orientation[17].

Selected clones were sequenced on both strands and assembled using standard techniques. Comparison of independently derived sequence of overlapping regions and independent reassembly sequenced clones revealed accuracy rates between 99.99 and 99.999%. Over half of the sequence differences were between genomic and BAC clone sequence. All available sequenced genetic markers were integrated into sequence assemblies to verify sequence contigs[4–8]. The total length of sequenced regions, which extend from either the telomeres or ribosomal DNA repeats to the 180-base-pair

(bp) centromeric repeats, is 115,409,949 bp (Table 1). Estimates of the unsequenced centromeric and rDNA repeat regions measure roughly 10 megabases (Mb), yielding a genome size of about 125 Mb, in the range of the 50–150 Mb haploid content estimated by different methods[18]. In general, features such as gene density, expression levels and repeat distribution are very consistent across the five chromosomes (Fig. 1), and these are described in detail in reports on individual chromosomes[4–8] and in the analysis of centromere, telomere and rDNA sequences.

We used tRNAscan-SE 1.21 (ref. 19) and manual inspection to identify 589 cytoplasmic transfer RNAs, 27 organelle-derived tRNAs and 13 pseudogenes—more than in any other genome sequenced to date. All 46 tRNA families needed to decode all possible 61 codons were found, defining the completeness of the functional set. Several highly amplified families of tRNAs were found on the same strand[6]; excluding these, each amino acid is decoded by 10–41 tRNAs.

The spliceosomal RNAs (U1, U2, U4, U5, U6) have all been experimentally identified in *Arabidopsis*. The previously identified sequences for all RNAs were found in the genome, except for U5 where the most similar counterpart was 92% identical. Between 10 and 16 copies of each small nuclear RNA (snRNA) were found across all chromosomes, dispersed as singletons or in small groups.

The small nucleolar RNAs (snoRNAs) consist of two subfamilies, the C/D box snoRNAs, which includes 36 *Arabidopsis* genes, and the H/ACA box snoRNAs, for which no members have been identified in *Arabidopsis*. U3 is the most numerous of the C/D box snoRNAs, with eight copies found in the genome. We identified forty-five additional C/D box snoRNAs using software (www.rna.wustl.edu/snoRNAdb/) that detects snoRNAs that guide ribose methylation of ribosomal RNA.

A combination of algorithms, all optimized with parameters based on known *Arabidopsis* gene structures, was used to define gene structure. We used similarities to known protein and expressed sequence tag (EST) sequence to refine gene models. Eighty per cent of the gene structures predicted by the three centres involved were completely consistent, 93% of ESTs matched gene models, and less than 1% of ESTs matched predicted non-coding regions, indicating

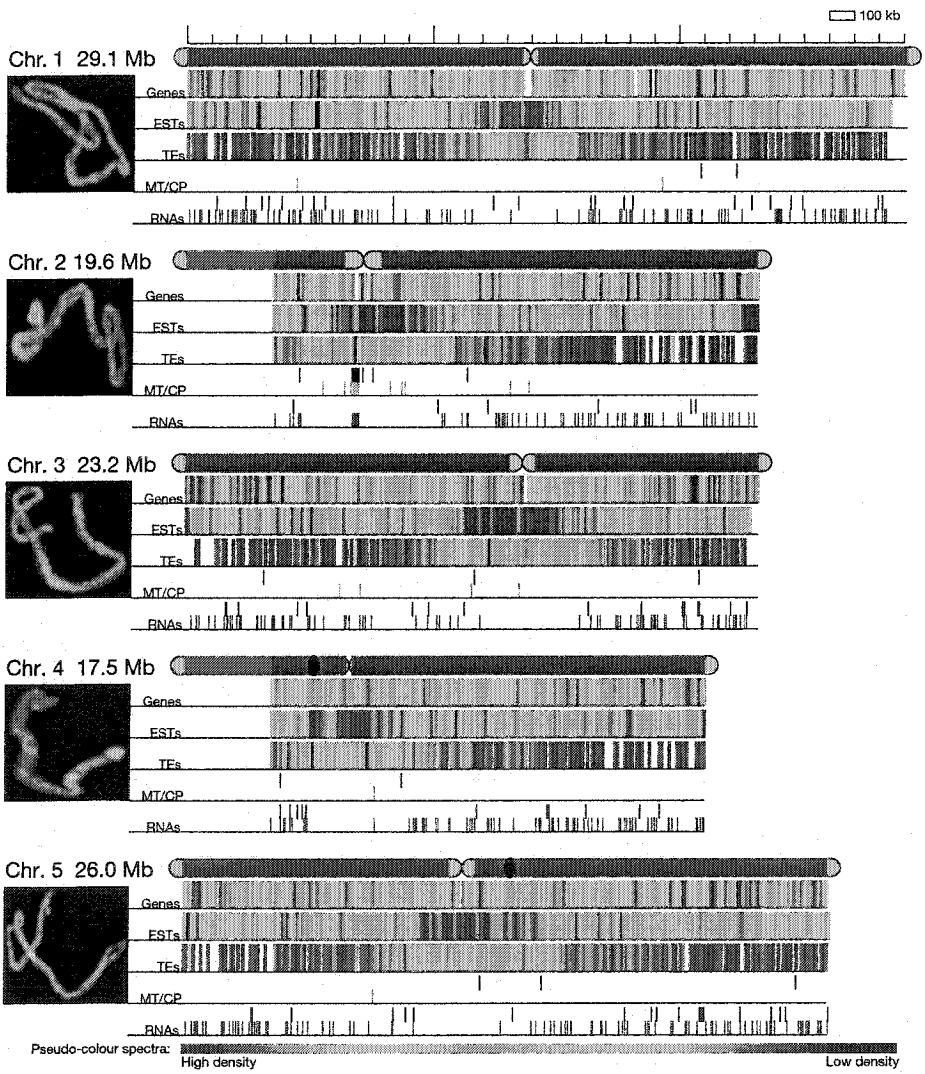


Figure 1 Representation of the *Arabidopsis* chromosomes. Each chromosome is represented as a coloured bar. Sequenced portions are red, telomeric and centromeric regions are light blue, heterochromatic knobs are shown black and the rDNA repeat regions are magenta. The unsequenced telomeres 2N and 4N are depicted with dashed lines. Telomeres are not drawn to scale. Images of DAPI-stained chromosomes were kindly supplied by P. Fransz. The frequency of features was given pseudo-colour assignments, from red (high density) to deep blue (low density). Gene density ('Genes') ranged from 38 per 100 kb to 1 gene per 100 kb; expressed sequence tag matches ('ESTs') ranged from more than 200 per 100 kb to 1 per 100 kb. Transposable element densities ('TEs') ranged from 33 per 100 kb to 1 per 100 kb. Mitochondrial and chloroplast insertions ('MT/CP') were assigned black and green tick marks, respectively. Transfer RNAs and small nucleolar RNAs ('RNAs') were assigned black and red ticks marks, respectively.

that most potential genes were identified. The sensitivity and selectivity of the gene prediction software used in this report has been comprehensively and independently assessed[20].

The 25,498 genes predicted (Table 1) is the largest gene set published to date: *C. elegans*[1] has 19,099 genes and *Drosophila*[2] 3,601 genes. *Arabidopsis* and *C. elegans* have similar gene density, whereas *Drosophila* has a lower gene density; *Arabidopsis* also has a significantly greater extent of tandem gene duplications and segmental duplications, which may account for its larger gene set.

The rDNA repeat regions on chromosomes 2 and 4 were not sequenced because of their known repetitive structure and content. The centromeric regions are not completely sequenced owing to large blocks of monotonic repeats such as 5S rDNA and 180-bp repeats. The sequence continues to be extended further into centromeric and other regions of complex sequence.

Characterization of the coding regions

To assess the similarities and differences of the *Arabidopsis* gene complement compared with other sequenced eukaryotic genomes, we assigned functional categories to the complete set of *Arabidopsis* genes. For chromosome 4 genes and the yeast genome, predicted functions were previously manually assigned[5,21]. All other predicted proteins were automatically assigned to these functional categories[22], assuming that conserved sequences reflect common functional relationships.

The functions of 69% of the genes were classified according to sequence similarity to proteins of known function in all organisms; only 9% of the genes have been characterized experimentally (Fig. 2a). Generally similar proportions of gene products were predicted to be targeted to the secretory pathway and mitochondria in *Arabidopsis* and yeast, and up to 14% of the gene products are

Table 1 Summary statistics of the *Arabidopsis* genome

Feature	Value					
a) The DNA molecules						
	Chr. 1	Chr. 2	Chr. 3	Chr. 4	Chr. 5	Σ
Length (bp)	29,105,111	19,646,945	23,172,617	17,549,867	25,953,409	115,409,949
Top arm (bp)	14,449,213	3,607,091	13,590,268	3,052,108	11,132,192	
Bottom arm (bp)	14,655,898	16,039,854	9,582,349	14,497,759	14,803,217	
Base composition (%GC)						
Overall	33.4	35.5	35.4	35.5	34.5	
Coding	44.0	44.0	44.3	44.1	44.1	
Non-coding	32.4	32.9	33.0	32.8	32.5	
Number of genes	6,543	4,036	5,220	3,825	5,874	25,498
Gene density (kb per gene)	4.0	4.9	4.5	4.6	4.4	
Average gene length (bp)	2,078	1,949	1,925	2,138	1,974	
Average peptide length (bp)	446	421	424	448	429	
Exons						
Number	35,482	19,631	26,570	20,073	31,226	13,2982
Total length (bp)	8,772,559	5,100,288	6,654,507	5,150,883	7,571,013	33,249,250
Average per gene	5.4	4.9	5.1	5.2	5.3	
Average size (bp)	247	259	250	256	242	
Introns						
Number	28,939	15,595	21,350	16,248	25,352	107,484
Total length (bp)	4,828,766	2,768,430	3,397,531	3,030,649	4,030,045	18,055,421
Average size (bp)	168	177	159	186	159	
Number of genes with ESTs (%)	60.8	56.9	59.8	61.4	61.4	
Number of ESTs	30,522	14,989	20,732	16,605	22,885	105,733
b) The proteome						
Classification/function						
Total proteins	6,543	4,036	5,220	3,825	5,874	25,498
With INTERPRO	4,194	1,205	2,989	1,545	3,136	13,069
domains	64.1%	29.9%	57.8%	40.4%	53.4%	51.3%
Genes containing at	2,334	1,322	1,615	1,402	1,940	8,613
least one TM domain	35.7%	32.8%	30.9%	36.7%	33.0%	33.8%
Genes containing at	2,513	1,424	1,664	1,304	2,121	9,026
least one SCOP domain	38.4%	35.3%	31.9%	34.1%	36.1%	35.4%
With putative signal peptides						
Secretory pathway	1,242 19.0%	675 16.7%	877 17.0%	659 17.2%	1,014 17.3%	4,467 17.6%
>0.95 specificity	1,146 17.5%	632 15.7%	813 15.7%	632 16.5%	964 16.4%	4,167 16.4%
Chloroplast	866 13.2%	535 13.2%	754 14.6%	532 13.9%	887 15.1%	3,574 14.0%
>0.95 specificity	602 9.2%	290 7.2%	420 8.1%	298 7.8%	475 8.1%	2,085 8.2%
mitochondria	901 13.8%	425 10.5%	554 10.7%	390 10.2%	627 10.7%	2,897 11.4%
>0.95 specificity	113 1.7%	49 1.2%	63 1.2%	59 1.5%	65 1.1%	349 1.4%
Functional classification						
Cellular metabolism	1,188 22.7%	620 23.3%	745 22.8%	588 22.9%	868 21.1%	4,009 22.5%
Transcription	880 16.8%	474 17.8%	566 17.3%	335 13.1%	763 18.6%	3,018 16.9%
Plant defence	640 12.2%	276 10.4%	354 10.8%	295 11.5%	490 11.9%	2,055 11.5%
Signalling	573 11.0%	296 11.1%	356 10.9%	210 8.2%	420 10.2%	1,855 10.4%
Growth	542 10.4%	263 9.9%	357 10.9%	448 17.5%	469 11.4%	2,079 11.7%
Protein fate	520 9.9%	273 10.2%	314 9.6%	264 10.3%	395 9.6%	1,766 9.9%
Intracellular transport	435 8.3%	214 8.9%	269 8.2%	220 8.6%	334 8.1%	1,472 8.3%
Transport	236 4.5%	139 5.2%	155 4.7%	113 4.4%	206 5.0%	849 4.8%
Protein synthesis	216 4.1%	111 4.2%	148 4.5%	90 3.5%	165 4.0%	730 4.1%
Total	5,230	2,666	3,264	2,563	4,110	17,833

The features of *Arabidopsis* chromosomes 1–5 and the complete nuclear genome are listed. Specialized searches used the following programs and databases: INTERPRO[23]; transmembrane (TM) domains by ALOM2 (unpublished); SCOP domain database[121]; functional classification by the PEDANT analysis system[22]. Signal peptide prediction (secretory pathway, targeted to chloroplast or mitochondria) was performed using TargetP[122] and http://www.cbs.dtu.dk/services/TargetP/.
* Default value.

likely to be targeted to the chloroplast (Table 1). The significant proportion of genes with predicted functions involved in metabolism, gene regulation and defence is consistent with previous analyses[5]. Roughly 30% of the 25,498 predicted gene products, (Fig. 2a), comprising both plant-specific proteins and proteins with similarity to genes of unknown function from other organisms, could not be assigned to functional categories.

To compare the functional catagories in more detail, we compared data from the complete genomes of *Escherichia coli*[23], *Synechocystis* sp.[24], *Saccharomyces cerevisiae*[21], *C. elegans*[1] and *Drosophila*[2], and a non-redundant protein set of *Homo sapiens*, with the *Arabidopsis* genome data (Fig. 2b), using a stringent BLASTP threshold value of $E < 10^{-30}$. The proportion of *Arabidopsis* proteins having related counterparts in eukaryotic genomes varies by a factor of 2 to 3 depending on the functional category. Only 8–23% of *Arabidopsis* proteins involved in transcription have related genes in other eukaryotic genomes, reflecting the independent evolution of many plant transcription factors. In contrast, 48–60% of genes involved in protein synthesis have counterparts in the other eukaryotic genomes, reflecting highly conserved gene functions. The relatively high proportion of matches between *Arabidopsis* and bacterial proteins in the categories 'metabolism' and 'energy' reflects both the acquisition of bacterial genes from the ancestor of the plastid and high conservation of sequences across all species. Finally, a comparison between unicellular and multicellular eukaryotes indicates that *Arabidopsis* genes involved in cellular communication and signal transduction have more counterparts in multicellular eukaryotes than in yeast, reflecting the need for sets of genes for communication in multicellular organisms.

Pronounced redundancy in the *Arabidopsis* genome is evident in segmental duplications and tandem arrays, and many other genes with high levels of sequence conservation are also scattered over the genome. Sequence similarity exceeding a BLASTP value $E < 10^{-20}$ and extending over at least 80% of the protein length were used as parameters to identify protein families (Table 2). A total of 11,601 protein types were identified. Thirty-five per cent of the predicted proteins are unique in the genome, and the proportion of proteins belonging to families of more than five members is substantially higher in *Arabidopsis* (37.4%) than in *Drosophila* (12.1%) or

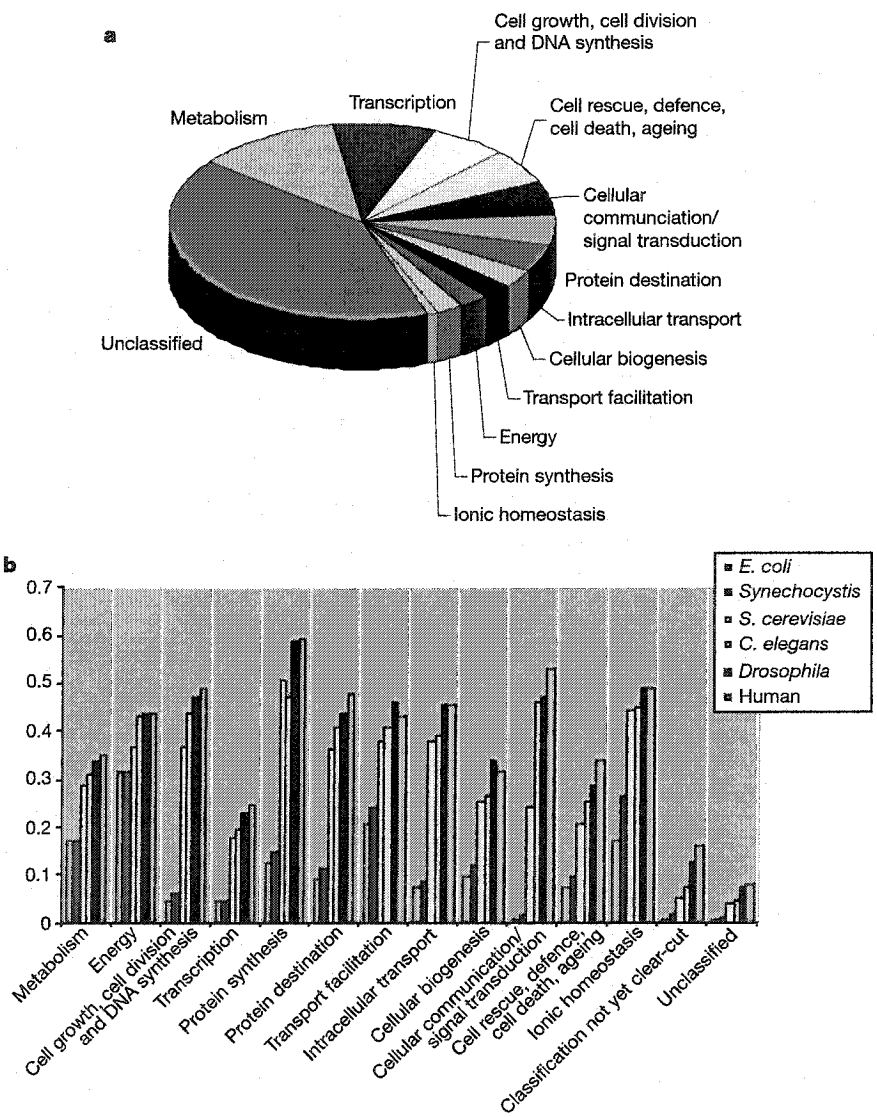


Figure 2 Functional analysis of *Arabidopsis* genes. **a**, Proportion of predicted *Arabidopsis* genes in different functional categories. **b**, Comparison of functional categories between organisms. Subsets of the *Arabidopsis* proteome containing all proteins that fall into a common functional class were assembled. Each subset was searched against the complete set of translations from *Escherichia coli*, *Synechocystis* sp. PCC6803, *Saccharomyces cerevisae*, *Drosophila*, *C. elegans* and a *Homo sapiens* non-redundant protein database. The percentage of *Arabidopsis* proteins in a particular subset that had a BLASTP match with $E \leq 10^{-30}$ to the respective reference genome is shown. This reflects the measure of sequence conservation of proteins within this particular functional category between *Arabidopsis* and the respective reference genome. *y* axis, 0.1 = 10%.

Table 2 Proportion of genes in different organisms present as either singletons or in paralogous families

	No of singletons and distinct gene families	Unique	Gene families containing				
			2 members	3 members	4 members	5 members	>5 members
H. influenzae	1,587	88.8%	6.8%	2.3%	0.7%	0.0%	1.4%
S. cerevisiae	5,105	71.4%	13.8%	3.5%	2.2%	0.7%	8.4%
D. melanogaster	10,736	72.5%	8.5%	3.4%	1.9%	1.6%	12.1%
C. elegans	14,177	55.2%	12.0%	4.5%	2.7%	1.6%	24.0%
Arabidopsis	11,601	35.0%	12.5%	7.0%	4.4%	3.6%	37.4%

The number of genes in the genomes of *Haemophilus influenzae*, *S. cerevisiae*, *Drosophila*, *C. elegans* and *Arabidopsis* that are present either as singletons or in gene families with two or more members are listed. To be grouped in a gene family, two genes had to show similarity exceeding a BLASTP value $E < 10^{-20}$ and a FASTA alignment over at least 80% of the protein length. In column 1, the number of genes that are unique plus the number of gene families are listed. Columns 2 to 6 give the percentage of genes present as singletons or in gene families of *n* members.

C. elegans (24.0%). The absolute number of *Arabidopsis* gene families and singletons (types) is in the same range as the other multicellular eukaryotes, indicating that a proteome of 11,000–15,000 types is sufficient for a wide diversity of multicellular life. The proportion of gene families with more than two members is considerably more pronounced in *Arabidopsis* than in other eukaryotes (Fig. 3). As segmental duplication is responsible for 6,303 gene duplications (see below), the extent of tandem gene duplications accounts for a significant proportion of the increased family size. These features of the *Arabidopsis*, and presumably other plant genomes, may indicate more relaxed constraints on genome size in plants, or a more prominent role of unequal crossing over to generate new gene copies.

Conserved protein domains revealed more informative differences through INTERPRO[25] analysis of the predicted gene products from *Arabidopsis*, *S. cerevisiae*, *C. elegans* and *Drosophila*. Statistically over-represented domains, and those that are absent from the *Arabidopsis* genome, indicate domains that may have been gained or lost during the evolution of plants (Supplementary Information Table 1). Proteins containing the Pro-Pro-Arg repeat, which is involved in RNA stabilization and RNA processing, are over-represented as compared to yeast, fly and worm; 400 proteins containing this signature were detected in *Arabidopsis* compared with only 10 in total in yeast, *Drosophila* and *C. elegans*. Protein kinases and associated domains, 169 proteins containing a disease resistance protein signature, and the Toll/IL-1R (TIR) domain, a component of pathogen recognition molecules[26], are also relatively abundant. This suggests that pathways transducing signals in response to pathogens and diverse environmental cues are more abundant in plants than in other organisms.

The RING zinc finger domain is relatively over-represented in *Arabidopsis* compared with yeast, *Drosophila* and *C. elegans*, whereas the F-box domain is over-represented as compared with yeast and *Drosophila* only. These domains are involved in targeting proteins to the proteasome[27] and ubiquitinylation[28] pathways of protein degradation, respectively. In plants many processes such as hormone and defence responses, light signalling, and circadian rhythms and pattern formation use F-box function to direct negative regulators to the ubiquitin degradation pathway. This mode of regulation appears to be more prevalent in plants and may account for a higher representation of the F box than in *Drosophila* and for the over-representation of the ubiquitin domain in the *Arabidopsis* genome. RING finger domain proteins in general have a role in ubiquitin protein ligases, indicating that proteasome-mediated degradation is a more widespread mode of regulation in plants than in other kingdoms.

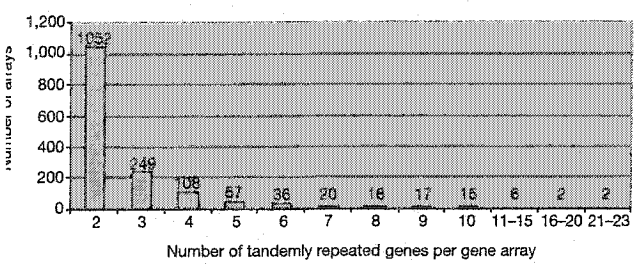


Figure 3 Distribution of tandemly repeated gene arrays in the *Arabidopsis* genome. Tandemly repeated gene arrays were identified using the BLASTP program with a threshold of $E < 10^{-20}$. One unrelated gene among cluster members was tolerated. The histogram gives the number of clusters in the genome containing 2 to *n* similar gene units in tandem.

Most functions identified by protein domains are conserved in similar proportions in the *Arabidopsis*, *S. cerevisiae*, *Drosophila* and *C. elegans* genomes, pointing to many ubiquitous eukaryotic pathways. These are illustrated by comparing the list of human disease genes[29] to the complete *Arabidopsis* gene set using BLASTP. Out of 289 human disease genes, 139 (48%) had hits in *Arabidopsis* using a BLASTP threshold $E < 10^{-10}$. Sixty-nine (24%) exceeded an $E < 10^{-40}$ threshold, and 26 (9.3%) had scores better than $E < 10^{-100}$ (Table 3). There are at least 17 human disease genes more similar to *Arabidopsis* genes than yeast, *Drosophila* or *C. elegans* genes (Table 3).

This analysis shows that, although numerous families of proteins are shared between all eukaryotes, plants contain roughly 150 unique protein families. These include transcription factors, structural proteins, enzymes and proteins of unknown function. Members of the families of genes common to all eukaryotes have undergone substantial increases or decreases in their size in *Arabidopsis*. Finally, the transfer of a relatively small number of cyanobacteria-related genes from a putative endosymbiotic ancestor of the plastid has added to the diversity of protein structures found in plants.

Genome organization and duplication

The *Arabidopsis* genome sequence provides a complete view of chromosomal organization and clues to its evolutionary history. Gene families organized in tandem arrays of two or more units have been described in *C. elegans*[1] and *Drosophila*[2]. Analysis of the *Arabidopsis* genome revealed 1,528 tandem arrays containing 4,140 individual genes, with arrays ranging up to 23 adjacent members (Fig. 3). Thus 17% of all genes of *Arabidopsis* are arranged in tandem arrays.

Large segmental duplications were identified either by directly aligning chromosomal sequences or by aligning proteins and searching for tracts of conserved gene order. All five chromosomes were aligned to each other in both orientations using MUMmer[30], and the results were filtered to identify all segments at least 1,000 bp in length with at least 50% identity (Supplementary Information Fig. 1). These revealed 24 large duplicated segments of 100 kb or larger, comprising 65.6 Mb or 58% of the genome. The only duplicated segment in the centromeric regions was a 375-kb segment on chromosome 4. Many duplications appear to have undergone further shuffling, such as local inversions after the duplication event.

We used TBLASTX[5] to identify collinear clusters of genes residing in large duplicated chromosomal segments. The duplicated regions encompass 67.9 Mb, 60% of the genome, slightly more than was

found in the DNA-based alignment (Fig. 4), and these data extend earlier findings[4,5,31]. The extent of sequence conservation of the duplicated genes varies greatly, with 6,303 (37%) of the 17,193 genes in the segments classified as highly conserved ($E < 10^{-30}$) and a further 1,705 (10%) showing less significant similarity up to $E < 10^{-5}$. The proportion of homologous genes in each duplicated segment also varies widely, between 20% and 47% for the highly conserved class of genes. In many cases, the number of copies of a gene and its counterpart differ (for example, one copy on one chromosome and multiple copies on the other; see Supplementary Information Fig. 2); this could be due to either tandem duplication or gene loss after the segmental duplication.

What does the duplication in the *Arabidopsis* genome tell us about the ancestry of the species? Polyploidy occurs widely in plants and is proposed to be a key factor in plant evolution[32]. As the majority of the *Arabidopsis* genome is represented in duplicated (but not triplicated) segments, it appears most likely that *Arabidopsis*, like maize, had a tetraploid ancestor[33]. A comparative sequence analysis of *Arabidopsis* and tomato estimated that a duplication occurred ~112 Myr ago to form a tetraploid[34]. The degrees of conservation of the duplicated segments might be due to divergence from an ancestral autotetraploid form, or might reflect differences present in an allotetraploid ancestor. It is also possible, however, that several independent segmental duplication events took place instead of tetraploid formation and stabilization.

The diploid genetics of *Arabidopsis* and the extensive divergence of the duplicated segments have masked its evolutionary history. The determination of *Arabidopsis* gene functions must therefore be pursued with the potential for functional redundancy taken into account. The long period of time over which genome stabilization has occurred has, however, provided ample opportunity for the divergence of the functions of genes that arose from duplications.

Comparative analysis of *Arabidopsis* accessions

Comparing the multiple accessions of *Arabidopsis* allows us to identify commonly occurring changes in genome microstructure. It also enables the development of new molecular markers for genetic mapping. High rates of polymorphism between *Arabidopsis* accessions, including both DNA sequence and copy number of tandem arrays, are prevalent at loci involved in disease resistance[35]. This has been observed for other plant species, and such loci are thought to serve as templates for illegitimate recombination to create new pathogen response specificities[36]. We carried out a comparative analysis between 82 Mb of the genome sequence of *Arabidopsis* accession Columbia (Col-0) and 92.1 Mb of non-redundant low-pass (twofold redundant) sequence data of the genomic DNA of accession Landsberg *erecta* (*Ler*). We identified two classes of differences between the sequences: single nucleotide polymorphisms (SNPs), and insertion–deletions (InDels). As we used high stringency criteria, our results represent a minimum estimate of numbers of polymorphisms between the two genomes.

In total, we detected 25,274 SNPs, representing an average density of 1 SNP per 3.3 kb. Transitions (A/T–G/C) represented 52.1% of the SNPs, and transversions accounted for the remainder: 17.3% for A/T–T/A, 22.7% for A/T–C/G and 7.9% for C/G–G/C. In total, we detected 14,570 InDels at an average spacing of 6.1 kb. They ranged from 2 bp to over 38 kilobase-pairs, although 95% were smaller than 50 bp. Only 10% of the InDels were co-located with simple sequence repeats identified with the program Sputnik. An analysis of 416 relative insertions greater than 250 bp in Col-0 showed that 30% matched transposon-related proteins, indicating that a substantial proportion of the large InDels are the result of transposon insertion or excision. Many InDels contained entire active genes not related to transposons. Half of such genes absent from corresponding positions in the Col-0 sequence were found elsewhere on the genome of *Ler*. This indicates that genes have been transferred to new genomic locations.

Gene structures are often affected by small InDels and SNPs. The positions of SNPs and InDels were mapped relative to 87,427 exons and 70,379 introns annotated in the Col-0 sequence. SNPs were found in exons, introns and intergenic regions at frequencies of 1 SNP per 3.1, 2.2 and 3.5 kb, respectively. The frequencies for InDels were 1 per 9.3, 3.1 and 4.3 kb, respectively. Polymorphisms were detected in 7% of exons, and alter the spliced sequences of 25% of the predicted genes. For InDels in exons, insertion lengths divisible by three are prevalent for small insertions (< 50 bp), indicating that many proteins can withstand small insertions or deletions of amino acids without loss of function.

Our analyses show that sequence polymorphisms between accessions of *Arabidopsis* are common, and that they occur in both coding and non-coding regions. We found evidence for the relocation of genes in the genome, and for changes in the complement of transposable elements. The data presented here are available at http://www.arabidopsis.org/cereon/.

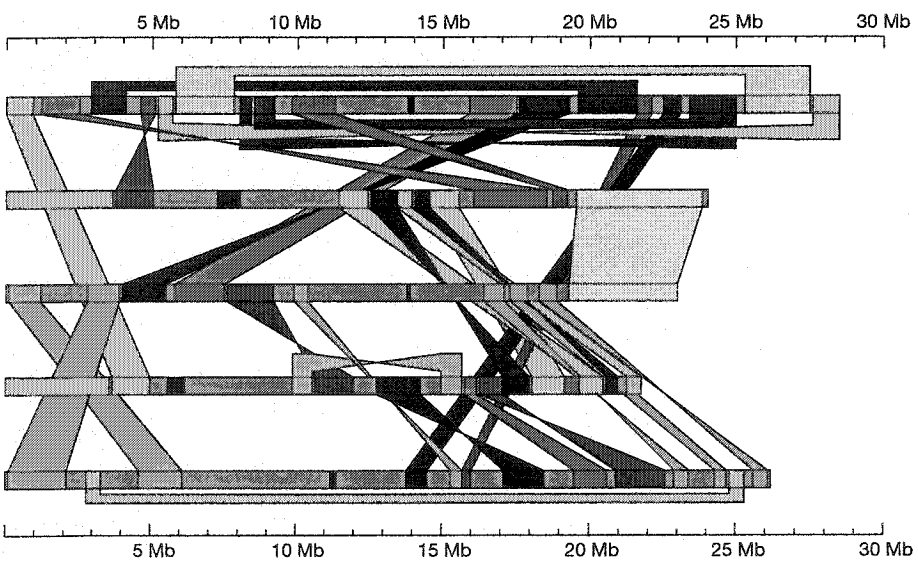


Figure 4 Segmentally duplicated regions in the *Arabidopsis* genome. Individual chromosomes are depicted as horizontal grey bars (with chromosome 1 at the top), centromeres are marked black. Coloured bands connect corresponding duplicated segments. Similarity between the rDNA repeats are excluded. Duplicated segments in reversed orientation are connected with twisted coloured bands. The scale is in megabases.

Comparison of *Arabidopsis* and other plant genera

Comparative genetic mapping can reveal extensive conservation of genome organization between closely related species[37,38]. The comparative analysis of plant genome microstructure reveals much about the evolution of plant genomes and provides unprecedented opportunities for crop improvement by establishing the detailed structures of, and relationships between, the genomes of crops and *Arabidopsis*.

The lineages leading to *Arabidopsis* and *Capsella rubella* (shepherd's purse) diverged between 6.2 and 9.8 Myr ago, and the gene content and genome organization of *C. rubella* is very similar to that of *Arabidopsis*[39], including the large-scale duplications. Alignment of *Arabidopsis* complementary DNA and EST sequences with genomic DNA sequences of *Arabidopsis* and *C. rubella* showed conservation of exon length and intron positions. Coding sequences predicted from these alignments differed from the annotated *Arabidopsis* gene sequences in two out of five cases.

The ancestral lineages of *Arabidopsis* and the *Brassica* (cabbage and mustard) genera diverged 12.2–19.2 Myr ago[40]. *Brassica* genes show a high level of nucleotide conservation with their *Arabidopsis* orthologues, typically more than 85% in coding regions[40]. The structure of *Brassica* genomes resembles that of *Arabidopsis*, but with extensive triplication and rearrangement[41], and extensive divergence of microstructure (Supplementary Information Fig. 3). The divergence between the genomes of *Arabidopsis* and *Brassica oleracea* is in striking contrast to that observed between *Arabidopsis* and *C. rubella*, although the time since divergence is only twofold greater. This accelerated rate of change in triplicated segments of the genome of *B. oleracea* indicates that polyploidy fosters rapid chromosomal evolution.

The *Arabidopsis* and tomato lineages diverged roughly 150 Myr ago, and comparative sequence analysis of segments of their genomes has revealed complex relationships[34]. Four regions of the *Arabidopsis* genome are related to each other and to one region in the tomato genome, suggesting that two rounds of duplication may have occurred in the *Arabidopsis* lineage. The extensive duplication described here supports the proposal that the more recent of these duplications, estimated to have occurred ~112 Myr ago, was the result of a polyploidization event. The lineages of *Arabidopsis* and rice diverged ~200 Myr ago[42]. Three regions of the genome of *Arabidopsis* were related to each other and to one region in the rice genome, providing further evidence for multiple duplication events[43,44].

The frequent occurrence of tandem gene duplications and the apparent deletion of single genes, or small groups of adjacent genes, from duplicated regions suggests that unequal crossing over may be a key mechanism affecting the evolution of plant genome microstructure. However, the segmental inversions and gene translocations in the genomes of both rice and *B. oleracea* that are not found in *Arabidopsis* indicate that additional mechanisms may be involved[40].

Integration of the three genomes in the plant cell

The three genomes in the plant cell—those of the nucleus, the plastids (chloroplasts) and the mitochondria—differ markedly in gene number, organization and stability. Plastid genes are densely packed in an order highly conserved in all plants[45], whereas mitochondrial genes[46] are widely dispersed and subjected to extensive recombination.

Organellar genomes are remnants of independent organisms—plastids are derived from the cyanobacterial lineage and mitochondria from the α-Proteobacteria. The remaining genes in plastids include those that encode subunits of the photosystem and the electron transport chain, whereas the genes in mitochondria encode essential subunits of the respiratory chain. Both organelles contain sets of specific membrane proteins that, together with housekeeping proteins, account for 61% of the genes in the chloroplast and 88 % in the mitochondrion (Table 4). The balances are involved in transcription and translation.

The number of proteins encoded in the nucleus likely to be found

Table 3 *Arabidopsis* genes with similarities to human disease genes

Human disease gene	*E* value	Gene code	*Arabidopsis* hit
Barier–White, SERCA	5.9×10^{-272}	T27I1_16	Putative calcium ATPase
Xeroderma Pigmentosum, D-XPD	7.2×10^{-228}	F15K9_19	Putative DNA repair protein
Xeroderma pigment, B-ERCC3	9.6×10^{-214}	AT5g41360	DNA excision repair cross-complementing protein
Hyperinsulinism, ABCC8	7.1×10^{-188}	F20D22_11	Multidrug resistance protein
Renal tubul. acidosis, ATP6B1	1.0×10^{-182}	AT4g38510	Probable H+-transporting ATPase
HDL deficiency 1, ABCA1	2.4×10^{-181}	At2g41700	Putative ABC transporter
Wilson, ATP7B	7.6×10^{-181}	AT5g44790	ATP-dependent copper transporter
Immunodeficiency, DNA Ligase 1	8.2×10^{-172}	T6D22_10	DNA ligase
Stargardt's, ABCA4	2.8×10^{-166}	At2g41700	Putative ABC transporter
Ataxia telangiectasia, ATM	3.1×10^{-168}	AT3g48190	Ataxia telangiectasia mutated protein AtATM
Niemann–Pick, NPC1	1.2×10^{-166}	F7F22_1	Niemann–Pick C disease protein-like protein
Menkes, ATP7A	1.1×10^{-153}	F2K11_17	ATP-dependent copper transporter, putative
HNPCC*, MLH1	1.5×10^{-150}	AT4g09140	MLH1 protein
Deafness, hereditary, MYO15	2.7×10^{-150}	At2g31900	Putative unconventional myosin
Cam, cardiac myopathy, MYH7	6.5×10^{-147}	T1G11_14	Putative myosin heavy chain
Xeroderma Pigmentosum, F-XPF	1.4×10^{-146}	AT5g41150	Repair endonuclease (gb|AAF01274.1)
G6PD deficiency, G6PD	7.6×10^{-137}	AT5g40760	Glucose-6-phosphate dehydrogenase
Cystic fibrosis, ABCC7	2.3×10^{-135}	AT3g62700	ABC transporter-like protein
Glycerol kinase defic, GK	7.9×10^{-135}	T21F11_21	Putative glycerol kinase
HNPCC, MSH3	6.6×10^{-134}	AT4g25540	Putative DNA mismatch repair protein
HNPCC, PMS2	5.1×10^{-126}	AT4g02460	No title
Zellweger, PEX1	4.1×10^{-125}	AT5g08470	Putative protein
HNPCC, MSH6	9.6×10^{-122}	AT4g02070	G/T DNA mismatch repair enzyme
Bloom, BLM	4.4×10^{-109}	T19D16_15	DNA helicase isolog
Finnish amyloidosis, GSN	2.2×10^{-107}	AT5g57320	Villin
Chediak–Higashi, CHS1	5.8×10^{-99}	F10O3_11	Putative transport protein
Xeroderma Pigmentosum, G-XPG	7.1×10^{-89}	AT3g28030	Hypothetical protein
Tare lymphocyte, ABCB3	1.3×10^{-84}	AT5g39040	ABC transporter-like protein
Citrullinemia, type I, ASS	3.2×10^{-83}	AT4g24830	Argininosuccinate synthase-like protein
Coffin–Lowry, RPS6KA3	5.2×10^{-81}	AT3g08720	Putative ribosomal-protein S6 kinase (ATPK19)
Keratoderma, KRT9	8.5×10^{-81}	AT3g17050	Unknown protein
Myotonic dystrophy, DM1	1.4×10^{-76}	At2g20470	Putative protein kinase
Cartter's, SLC12A1	1.6×10^{-75}	F26G16_9	Cation-chloride co-transporter, putative
Dents, CLCN5	3.3×10^{-74}	AT5g26240	CLC-d chloride channel protein
Diaphanous 1, DAPH1	1.9×10^{-73}	68069_m00158	Hypothetical protein
AKT2	6.9×10^{-72}	AT3g08730	Putative ribosomal-protein S6 kinase (ATPK6)

in organelles was predicted using default settings on TargetP (Table 1). Many nuclear gene products that are targeted to either (or both) organelles were originally encoded in the organelle genomes and were transferred to the nuclear genome during evolutionary history. A large number also appear to be of eukaryotic origin, with functions such as protein import components, which were probably not required by the free-living ancestors of the endosymbionts.

To identify nuclear genes of possible organellar ancestry, we compared all predicted *Arabidopsis* proteins to all proteins from completed genomes including those from plastids and mitochondria (Supplementary Information Table 2). This search identified proteins encoded by the *Arabidopsis* nuclear genome that are most similar to proteins encoded by other species' organelle genomes (14 mitochondrial and 44 plastid). These represent organelle-to-nuclear gene transfers that have occurred sometime after the divergence of the organelle-containing lineages[47]. There is a great excess of nuclear encoded proteins most similar to proteins from the cyanobacteria *Synechocystis* (Supplementary Information Fig. 4; 806 *Arabidopsis* predicted proteins matching 404 different *Synechocystis* proteins, providing further evidence of a genome duplication). These 806 *Arabidopsis* predicted proteins, and many others of greatly diverse function, are possibly of plastid descent. Through searches against proteins from other cyanobacteria (with incompletely sequenced genomes), we identified 69 additional genes of possibly plastid descent. Only 25% of these putatively plastid-derived proteins displayed a target peptide predicted by TargetP, indicating potential cytoplasmic functions for most of these genes.

The difference between predicted plastid-targeted and predicted plastid-derived genes indicates that there is a probable overestimation by *ab initio* targeting prediction methods and a lack of resolution with respect to destination organelles, the possible extensive divergence of some endosymbiont-derived genes in the nuclear genome, the co-opting of nuclear genes for targeting to organelles, and cytoplasmic functions for cyanobacteria-derived proteins. Clearly more refined tools and extensive experimentation is required to catalogue plastid proteins.

The transfer of genes between genomes still continues (Supplementary Information Table 3). Plastid DNA insertions in the nucleus (17 insertions totalling 11 kb) contain full-length genes encoding proteins or tRNAs, fragments of genes and an intron as well as intergenic regions. Subsequent reshuffling in the nucleus is illustrated by the *atpH* gene, which was originally transferred completely, but is now in two pieces separated by 2 kb. The 13 small mitochondrial DNA insertions total 7 kb in addition to the large insertion close to the centromere of chromosome 2 (ref. 3). The high level of recombination in the mitochondrial genome may account for these events.

Transposable elements

Transposons, which were originally identified in maize by Barbara McClintock, have been found in all eukaryotes and prokaryotes. A subset of transposons replicate through an RNA intermediate (class I), whereas others move directly through a DNA form (class II). Transposons are further classified by similarity either between their mobility genes or between their terminal and/or internal motifs, as well as by the size and sequence of their target site. Internally deleted elements can often be mobilized in *trans* by fully functional elements.

Transposons in *Arabidopsis* account for at least 10% of the genome, or about one-fifth of the intergenic DNA. The *Arabidopsis* genome has a wealth of class I (2,109) and II (2,203) elements, including several new groups (1,209 elements; Supplementary Information Table 4). Mobile histories for many elements were obtained by identifying regions of the genome with significant similarity to 'empty' target sites (RESites) thus providing high-resolution information concerning the termini and target site duplications[48,49]. These regions were readily detected because of the propensity of transposons to integrate into repeats and because of duplications in the genome sequence. In several cases, genes appear to have been included as 'passengers' in transposable units[48]. In some cases, shared sequence similarity, coding capacity and RESites attest to recent activity of transposable elements in the *Arabidopsis* genome. Only about 4% of the complete elements identified correspond to an EST, however, suggesting that most are not transcribed.

Transposable elements found in many other plant genomes are well represented in *Arabidopsis*, including *copia*- and *gypsy*-like long terminal repeat (LTR) retrotransposons, long interspersal nuclear elements (LINEs); short interspersed nuclear elements (SINEs), *hobo*/*Activator*/Tam3 (*hAT*)-like elements, CACTA-like elements and miniature inverted-repeat transposable elements (MITES). Although usually small in size, some larger *Tourist*-like MITEs contain open reading frames (ORFs) with similarity to the transposases of bacterial insertion sequences[48]. *Basho* and many *Mutator*-like elements (MULEs), first discovered in the *Arabidopsis* sequence, represent structurally unique transposons[48–50]. *Basho* elements have a target site preference for mononucleotide 'A' and wide distribution among plants[48,51]. MULEs exhibit a high level of sequence diversity and members of most groups lack long terminal inverted repeats (TIRs). Phylogenetic analysis of the *Arabidopsis* MURA-like transposases suggests that TIR-containing MULEs are more closely related to one another than to MULEs lacking TIRs[49,52].

For many plants with large genomes, class I retrotransposons contribute most of the nucleotide content[53]. In the small *Arabidopsis* genome, class I elements are less abundant and primarily occupy the centromere. In contrast, *Basho* elements and class II transposons such as MITEs and MULEs predominate on the periphery of pericentromeric domains (Fig. 5). In class II transposons, MULEs and CACTA elements are clustered near centromeres and heterochromatic knobs, whereas MITEs and *hAT* elements have a less pronounced bias. The distribution pattern of transposable elements observed in *Arabidopsis* may reflect different types of pericentromeric heterochromatin regions and may be similar to those found in animals.

Numerous centromeric satellite repeats are located between each chromosome arm and have not yet been sequenced, but are represented in part by unanchored BAC contigs (R. Martienssen and M. Marra, unpublished data). End sequence suggests that these domains contain many more class I than class II elements, consistent with the distribution reported here (K. Lemcke and R. Martienssen, unpublished data). We do not know the significance of the apparent paucity of elements in telomeric regions and in the region flanking the rDNA repeats on chromosome 4 (but not on chromosome 2).

Overall, transposon-rich regions are relatively gene-poor and have lower rates of recombination and EST matches, indicating a correlation between low gene expression, high transposon density and low recombination[51]. The role of transposons in genome

Table 4 General features of genes encoded by the three genomes in *Arabidopsis*

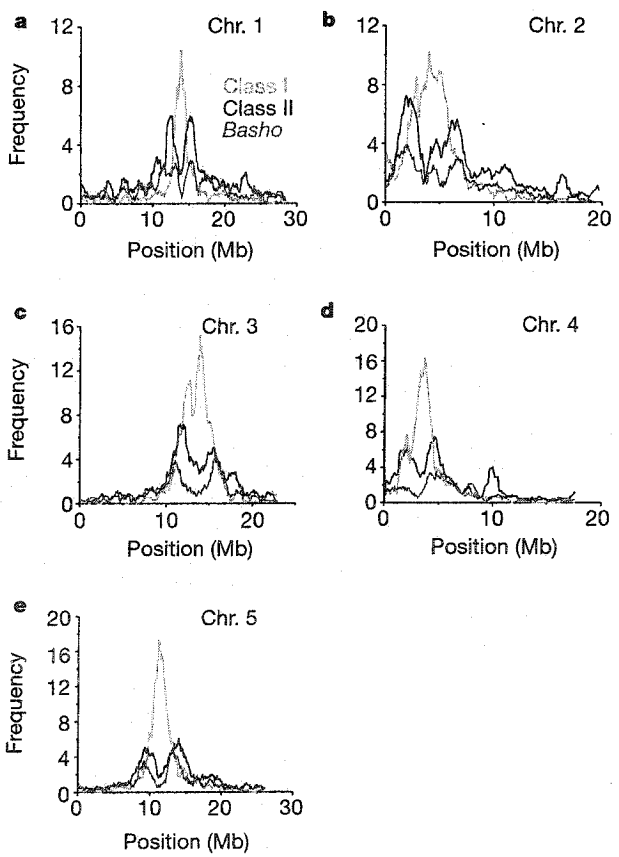
	Nucleus/cytoplasm	Plastid	Mitochondria
Genome size	125 Mb	154 kb	367 kb
Genome equivalent/cell	2	560	26
Duplication	60%	17%	10%
Number of protein genes	25,498	79	58
Gene order	Variable, but syntenic	Conserved	Variable
Density (kb per protein gene)	4.5	1.2	6.25
Average coding length	1,900 nt	900 nt	860 nt
Genes with introns	79%	18.4%	12%
Genes/pseudogenes	1/0.03	1/0	1/0.2–0.5
Transposons (% of total genome size)	14%	0%	4%

organization and chromosome structure can now be addressed in a model organism known to undergo DNA methylation and other forms of chromatin modification thought to regulate transposition[52].

DNA, telomeres and centromeres

Nucleolar organizers (NORs) contain arrays of unit repeats encoding the 18S, 5.8S and 25S ribosomal RNA genes and are transcribed by RNA polymerase I. Together with 5S RNA, which is transcribed by RNA polymerase III, these rRNAs form the structural and catalytic cores of cytoplasmic ribosomes. In *Arabidopsis*, the NORs juxtapose the telomeres of chromosomes 2 and 4, and comprise uninterrupted 18S, 5.8S and 25S units all orientated on the chromosomes in the same direction[54]. In contrast, the 5S rRNA genes are localized to heterogeneous arrays in the centromeric regions of chromosomes 3, 4 and 5 (ref. 55; and Fig. 6). Both NORs are roughly 3.5–4.0 megabase-pairs and comprise ~350–400 highly methylated rRNA gene units, each ~10 kb (ref. 54). The sequence between the euchromatic arms and NORs has been determined. Elsewhere in the genome, only one other 18S, 5.8S, 25S rRNA gene unit was identified in centromere 3. Although minor variations in sequence length and composition occur in the NOR repeats, these variants are highly clustered, supporting a model of sequence maintenance through concerted evolution[55].

Arabidopsis telomeres are composed of CCCTAAA repeats and average ~2–3 kb (ref. 56). For TEL4N (telomere 4 North), consensus repeats are adjacent to the NOR; the remaining telomeres are typically separated from coding sequences by repetitive subtelomeric regions measuring less than 4 kb. Imperfect telomere-like arrays of up to 24 kb are found elsewhere in the genome, particularly near centromeres. These arrays might affect the expression of nearby genes and may have resulted from ancient rearrangements, such as inversions of the chromosome arms.



gure 5 Distribution of class I, II and *Basho* transposons in *Arabidopsis* chromosomes. ..ıe frequency of class I retroelements (green), class II DNA transposons (blue) and *Basho* elements (purple) are shown at 100-kb intervals along the five chromosomes (**a**–**e**) of *Arabidopsis*.

Centromere DNA mediates chromosome attachment to the meiotic and mitotic spindles and often forms dense heterochromatin. Genetic mapping of the regions that confer centromere function provided the markers necessary to precisely place BAC clones at individual centromeres[17]; 69 clones were targeted for sequencing, resulting in over 5 Mb of DNA sequence from the centromeric regions. The unsequenced regions of centromeres are composed primarily of long, homogeneous arrays that were characterized previously with physical[57] and genetic mapping[17] and contain over 3 Mb of repetitive arrays, including the 180-bp repeats and 5S rDNA[51] (Fig. 6).

Arabidopsis centromeres, like those of many higher eukaryotes, contain numerous repetitive elements including retroelements, transposons, microsatellites and middle repetitive DNA[17]. These repeats are rare in the euchromatic arms and often most abundant in pericentromeric DNA. The repeats, affinity for DNA-binding dyes, dense methylation patterns and inhibition of homologous recombination indicate that the centromeric regions are highly heterochromatic, and such regions are generally viewed as very poor environments for gene expression. Unexpectedly, we found at least 47 expressed genes encoded in the genetically defined centromeres of *Arabidopsis* (http://preuss.bsd.uchicago.edu/arabidopsis.genome.html). In several cases, these genes reside on islands of unique sequence flanked by repetitive arrays, such as 180-bp or 5S rDNA repeats. Among the genes encoded in the centromeres are members of 11 of the 16 functional categories that comprise the proteome. The centromeres are not subject to recombination; consequently, genes residing in these regions probably exhibit unique patterns of molecular evolution.

The function of higher eukaryotic centromeres may be specified by proteins that bind to centromere DNA, by epigenetic modifications, or by secondary or higher order structures. A pairwise comparison of the non-repetitive portions of all five centromeres showed they share limited (1–7%) sequence similarity. Forty-one families of small, conserved centromere sequences (AtCCS, see http://preuss.bsd.uchicago.edu/arabidopsis.genome.html) are enriched in the centromeric and pericentromeric regions and differ from sequences found in the centromeres of other eukaryotes. Molecular and genetic assays will be required to determine whether these conserved motifs nucleate *Arabidopsis* centromere activity. Apart from the AtCCS sequences, most centromere DNA is not shared between chromosomes, complicating efforts to derive clear evolutionary relationships. In contrast, genetic and cytological assays indicate that homologous centromeres are highly conserved among *Arabidopsis* accessions, albeit subject to rearrangements such as inversions to form knobs[5,58,59] and insertions[4]. Further investigation of centromere DNA promises to yield information on the evolutionary forces that act in regions of limited recombination, as well as an improved understanding of the role of DNA sequence patterns in chromosome segregation.

Membrane transport

Transporters in the plasma and intracellular membranes of *Arabidopsis* are responsible for the acquisition, redistribution and compartmentalization of organic nutrients and inorganic ions, as well as for the efflux of toxic compounds and metabolic end products, energy and signal transduction, and turgor generation. Previous genomic analyses of membrane transport systems in *S. cerevisiae* and *C. elegans* led to the identification of over 100 distinct families of membrane transporters[60,61]. We compared membrane transport processes between *Arabidopsis*, animals, fungi and prokaryotes, and identified over 600 predicted membrane transport systems in *Arabidopsis* (http://www-biology.ucsd.edu/~ipaulsen/transport/), a similar number to that of *C. elegans*

(~700 transporters) and over twofold greater than either *S. cerevisiae* or *E. coli* (~300 transporters).

We compared the transporter complement of *Arabidopsis*, *C. elegans* and *S. cerevisiae* in terms of energy coupling mechanisms (Fig. 7a). Unlike animals, which use a sodium ion P-type ATPase pump to generate an electrochemical gradient across the plasma membrane, plants and fungi use a proton P-type ATPase pump to form a large membrane potential (−250 mV)[62]. Consequently, plant secondary transporters are typically coupled to protons rather than to sodium[63]. Compared with *C. elegans*, *Arabidopsis* has a surprisingly high percentage of primary ATP-dependent transporters (12% and 21% of transporters, respectively), reflecting increased numbers of P-type ATPases involved in metal ion transport and ABC ATPases proposed to be involved in sequestering unusual metabolites and drugs in the vacuole or in other intracellular compartments. These processes may be necessary for pathogen defence and nutrient storage.

About 15% of the transporters in *Arabidopsis* are channel proteins, five times more than in any single-celled organism but half the number in *C. elegans* (Fig. 7b). Almost half of the *Arabidopsis* channel proteins are aquaporins, and *Arabidopsis* has 10-fold more Mfamily major intrinsic protein (MIP) family water channels than any other sequenced organism. This abundance emphasizes the importance of hydraulics in a wide range of plant processes, including sugar and nutrient transport into and out of the vasculature, opening of stomatal apertures, cell elongation and epinastic movements of leaves and stems. Although *Arabidopsis* has a diverse range of metal cation transporters, *C. elegans* has more, many of which function in cell–cell signalling and nerve signal transduction. *Arabidopsis* also possesses transporters for inorganic anions such as phosphate, sulphate, nitrate and chloride, as well as for metal cation channels that serve in signal transduction or cell homeostasis. Compared with other sequenced organisms, *Arabidopsis* has 10-fold more predicted peptide transporters, primarily of the proton-dependent oligopeptide transport (POT) family, emphasizing the importance of peptide transport or indicating that there is broader substrate specificity than previously realized. There are nearly 1,000 *Arabidopsis* genes encoding Ser/Thr protein kinases, suggesting that peptides may have an important role in plant signalling[64].

Virtually no transporters for carboxylates, such as lactate and pyruvate, were identified in the *Arabidopsis* genome. About 12% of the transporters were predicted to be sugar transporters, mostly consisting of paralogues of the MFS family of hexose transporters. Notably, *S. cerevisiae*, *C. elegans* and most prokaryotes use APC family transporters as their principle means of amino-acid transport, but *Arabidopsis* appears to rely primarily on the AAAP family of amino-acid and auxin transporters. More than 10% of the transporters in *Arabidopsis* are homologous to drug efflux pumps; these probably represent transporters involved in the sequestration into vacuoles of xenobiotics, secondary metabolites, and breakdown products of chlorophyll.

Surprisingly, *Arabidopsis* has close homologues of the human ABC TAP transporters of antigenic peptides for presentation to the major histocompatability complex (MHC). In *Arabidopsis*, these transporters may be involved in peptide efflux, or more speculatively, in some form of cell-recognition response. *Arabidopsis* also has 10-fold more members of the multi-drug and toxin extrusion (MATE) family than any other sequenced organism; in bacteria, these transporters function as drug efflux pumps. Curiously, *Arabidopsis* has several homologues of the *Drosophila* RND transporter family Patched protein, which functions in segment polarity, and more than ten homologues of the *Drosophila* ABC family eye pigment transporters. In plants, these are presumably involved in intracellular sequestration of secondary metabolites.

DNA repair and recombination

DNA repair and recombination pathways have many functions in different species such as maintaining genomic integrity, regulating mutation rates, chromosome segregation and recombination, genetic exchange within and between populations, and immune system development. Comparing the *Arabidopsis* genome with other species[65] indicates that *Arabidopsis* has a similar set of DNA repair and recombination (RAR) genes to most other eukaryotes. The pathways represented include photoreactivation, DNA ligation, non-homologous end joining, base excision repair, mismatch excision repair, nucleotide excision repair and many aspects of DNA recombination (Supplementary Information Table 5). The *Arabidopsis* RAR genes include homologues of many DNA repair genes that are defective in different human diseases (for example, hereditary breast cancer and non-polyposis colon cancer, xeroderma pigmentosum and Cockayne's syndrome).

One feature that sets *Arabidopsis* apart from other eukaryotes is the presence of additional homologues of many RAR genes. This is seen for almost every major class of DNA repair, including recombination (four RecA), DNA ligation (four DNA ligase I), photoreactivation (one class II photolyase and five class I photolyase homologues) and nucleotide excision repair (six RPA1, two RPA2, two Rad25, three TFB1 and four Rad23). This is most striking for genes with probable roles in base excision repair. *Arabidopsis* encodes 16 homologues of DNA base glycosylases (enzymes that

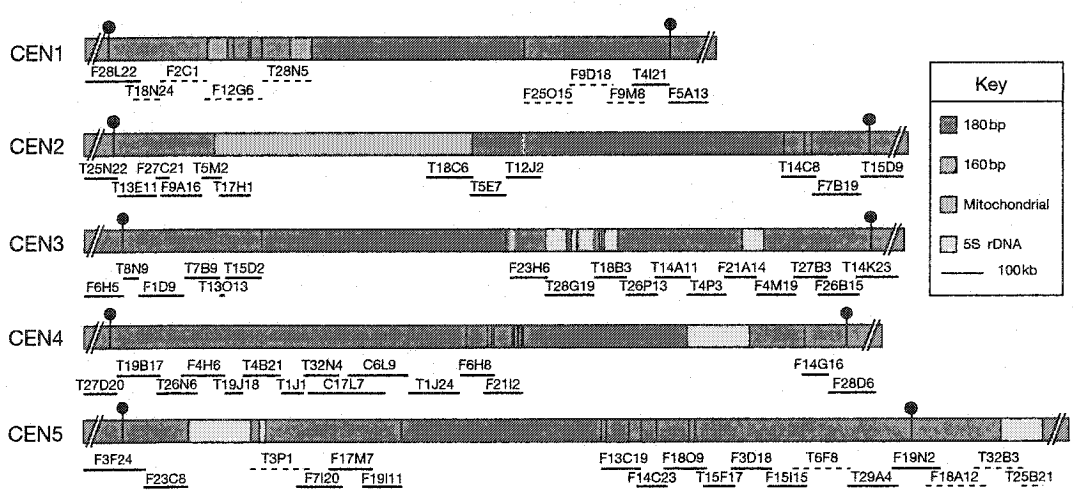


Figure 6 Predicted centromere composition. Genetically defined centromere boundaries are indicated by filled circles; fully and partially assembled BAC sequences are represented by solid and dashed black lines, respectively. Estimates of repeat sizes within the centromeres were derived from consideration of repeat copy number, physical mapping and cytogenetic assays.

recognize abnormal DNA bases and cleave them from the sugar-phosphate backbone)—more than any other species known. This includes several homologues of each of three families of alkylation damage base glycosylases: two of the *S. cerevisiae* MPG; six of the *E. coli* TagI; and two of the *E. coli* AlkA. *Arabidopsis* also encodes three homologues of the apurinic-apyrimidinic (AP) endonuclease Xth. AP endonucleases continue the base excision repair started by glycosylases by cleaving the DNA backbone at abasic sites.

Evolutionary analysis indicates that some of the extra copies of RAR genes in *Arabidopsis* originated through relatively recent gene duplications—because many of the sets of genes are more closely related to each other than to their homologues in any other species. As duplication is frequently accompanied by functional divergence, the duplicate (paralogous) genes may have different repair specificities or may have evolved functions that are outside RAR functions (as is the case for two of the five class I photolyase homologues, which function as blue-light receptors). In most cases, it is not known whether the paralogous gene copies have different functions. The presence of multiple paralogues might also allow functional redundancy or a greater repair or recombination capacity.

The multiplicity of RAR genes in *Arabidopsis* is also partly due to the transfer of genes from the organellar genomes to the nucleus. Repair gene homologues that appear to be of chloroplast origin (Supplementary Information Tables 2 and 5) include the recombination proteins RecA, RecG and SMS, two class I photolyase homologues, Fpg, two MutS2 proteins, and the transcription-repair coupling factor Mfd. Two of these (RecA and Fpg) are involved in RAR functions in the plastid, suggesting that the others may be as well. The finding of an Mfd orthologue of cyanobacterial descent is surprising. In *E. coli*, Mfd couples nucleotide excision repair carried out by UvrABC to transcription, leading to the rapid repair of DNA damage on the transcribed strand of transcribed genes[66] The absence of orthologues of UvrABC in *Arabidopsis* renders the function of Mfd difficult to predict. The presence of Mfd but not UvrABC has been reported for only one other species, a bacterial endosymbiont of the pea aphid.

Other nuclear-encoded *Arabidopsis* DNA repair gene homologues are evolutionarily related to genes from α-Proteobacteria, and thus may be of mitochondrial descent. In particular, the six homologues of the alkyl-base glycosylase TagI appear to be the result of a large expansion in plants after transfer from the mitochondrial genome. Whether any of these TagI homologues function in the repair and maintenance of mitochondrial DNA has not been determined. More detailed phylogenetic analysis may reveal additional *Arabidopsis* RAR genes to be of organellar ancestry.

There are some notable absences of proteins important for RAR in other species, including alkyltransferases, MSH4, RPA3 and many components of TFIIH (TFB2, TFB3, TFB4, CCL1, Kin28). Nevertheless, *Arabidopsis* shows many similarities to the set of DNA repair genes found in other eukaryotes, and therefore offers an experimental system for determining the functions of many of these proteins, in part through characterization of mutants defective in DNA repair[67].

Gene regulation

Eukaryotic gene expression involves many nuclear proteins that modulate chromatin structure, contribute to the basal transcription machinery, or mediate gene regulation in response to developmental, environmental or metabolic cues. As predicted by sequence similarity, more than 3,000 such proteins may be encoded by the *Arabidopsis* genome, suggesting that it has a comparable complexity of gene regulation to other eukaryotes. *Arabidopsis* has an additional level of gene regulation, however, with DNA methylation potentially mediating gene silencing and parental imprinting.

Plants have evolved several variations on chromatin remodelling proteins, such as the family of HD2 histone deacetylases[68]. Although *Arabidopsis* possesses the usual number of SNF2-type chromatin remodelling ATPases, which regulate the expression of nearly all genes, there are significant structural differences between yeast and metazoan SNF2-type genes and their orthologues in *Arabidopsis*. DDM1, a member of the SNF2 superfamily, and MOM1, a gene with similarity to the SNF2 family, are involved in transcriptional gene silencing in *Arabidopsis*. MOM1 has no clear orthologue in fungal or metazoan genomes.

Consistent with its methylated DNA, *Arabidopsis* possesses eight DNA methyltransferases (DMTs). Two of the three types are orthologous to mammalian DMT[69] whereas one, chromomethyltransferase[70], is unique to plants. No DMTs are found in yeast or *C. elegans*, although two DMT-like genes are found in *Drosophila*[71]. *Arabidopsis* also encodes eight proteins with methyl-DNA-binding domains (MBDs). Despite lacking methylated DNA, *Drosophila* encodes four MBD proteins and *C. elegans* has two. These differences in chromatin components are likely to reflect important differences in chromatin-based regulatory control of gene expression in eukaryotes (Supplementary Information Table 6; http://Ag.Arizona.Edu/chromatin/chromatin.html).

The *Arabidopsis* genome encodes transcription machinery for the three nuclear DNA-dependent RNA polymerase systems typical of eukaryotes (Supplementary Information Table 6). Transcription by RNA polymerases II and III appears to involve the same machinery as is used in other eukarotes; however, most transcription factors for RNA polymerase I are not readily identified. Only two polymerase I regulators (other than polymerase subunits and TATA-binding protein) are apparent in *Arabidopsis*, namely homologues of yeast RRN3 and mouse TTF-1. All eukaryotes examined to date have distinct genes for the largest and second largest subunits of polymerase I, II and III. Unexpectedly, *Arabidopsis* has two genes encoding a fourth class of largest subunit and second-largest subunit (Supplementary Information Fig. 5). It will be interesting to determine whether the atypical subunits comprise a polymerase that has a plant-specific function. Four genes encoding single-subunit plastid or mitochondrial RNA polymerases have been identified in *Arabidopsis* (Supplementary Information Table 6). Genes for the bacterial β-, β′- and α-subunits of RNA polymerase are also present, as are homologues of various σ-factors, and these proteins may regulate chloroplast gene expression. Mutations in the *Sde-1* gene, encoding RNA-dependent RNA polymerase (RdRp), lead to defective post-transcriptional gene silencing[72]. We also identified five more closely related RdRp genes.

Our analysis, using both similarity searches and domain matches, has identified 1,709 proteins with significant similarity to known classes of plant transcription factors classified by conserved DNA-binding domains. This analysis used a consistent conservative threshold that probably underestimates the size of families of diverse sequence. This class of protein is the least conserved among all classes of known proteins, showing only 8–23% similarity to transcription factors in other eukaryotes (Fig. 2b). This reduced similarity is due to the absence of certain classes of transcription factors in *Arabidopsis* and large numbers of plant-specific transcription factors. We did not detect any members of several widespread families of transcription factors, such as the REL (Rel-like DNA-binding domain) homology region proteins, nuclear steroid receptors and forkhead-winged helix and POU (Pit-1, Oct- and Unc-8b) domain families of developmental regulators. Conversely, of 29 classes of *Arabidopsis* transcription factors, 16 appear to be unique to plants (Supplementary Information Table 6). Several of these, such as the AP2/EREBP-RAV, NAC and ARF-AUX/IAA families, contain unique DNA-binding domains, whereas others contain plant-specific variants of more widespread domains, such as the DOF and WRKY zinc-finger families and the two-repeat MYB family.

Functional redundancy among members of large families of closely related transcription factors in *Arabidopsis* is a significant potential barrier to their characterization[73]. For example, in the

SHATTERPROOF and SEPALLATA families of MADS box transcription factors, all genes must be defective to produce visible mutant phenotypes[74,75]. These functionally redundant genes are found on the segmental duplications described above. Our analyses, together with the significant sequence similarity found in large families of transcription factors such as the R2R3-repeat MYB and WRKY families, suggest that strategies involving overexpression will be important in determining the functions of members of transcription factor families.

Arabidopsis has two or over three times more transcription factors than identified in *Drosophila*[29] or *C.elegans*[1], respectively. The significantly greater extent of segmental chromosomal and local tandem duplications in the *Arabidopsis* genome generates larger gene families, including transcription factors. The partly overlapping functions defined for a few transcription factors are also likely to be much more widespread, implicating many sequence-related transcription factors in the same cellular processes. Finally, the expanded number of genes involved in metabolism, defence and environmental interaction in *Arabidopsis* (Fig. 2a), which have few counterparts in *Drosophila* and *C. elegans*, all require additional numbers and classes of transcription factors to integrate gene function in response to a vast range of developmental and environmental cues.

Cellular organization

Plant cells differ from animal cells in many features such as plastids, vacuoles, Golgi organization, cytoskeletal arrays, plasmodesmata linking cytoplasms of neighbouring cells, and a rigid polysaccharide-rich extracellular matrix—the cell wall. Because the cell wall maintains the position of a cell relative to its neighbours, both changes in cell shape and organized cell divisions, involving cytoskeleton reorganization and membrane vesicle targeting, have major roles in plant development. Plant cytokinesis is also unique in that the partitioning membrane is formed *de novo* by vesicle fusion. We compared the *Arabidopsis* genome with those of *C. elegans*, *Drosophila* and yeast to glimpse the genetic basis of plant-cell-specific features.

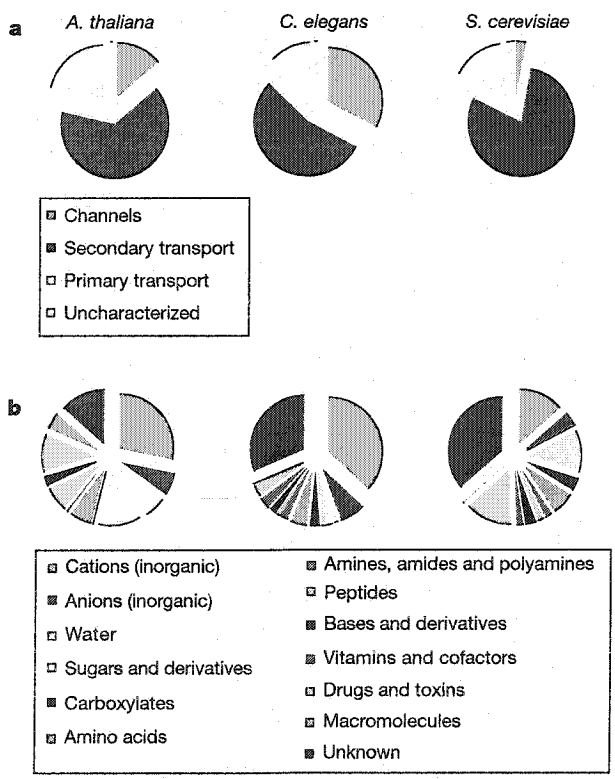


Figure 7 Comparison of the transport capabilities of *Arabidopsis*, *C. elegans* and *S. cerevisiae*. Pie charts show the percentage of transporters in each organism according to bioenergetics (**a**) and substrate specificity (**b**).

The principal components of the plant cytoskeleton are microtubules (MTs) and actin filaments (AFs); intermediate filaments (IFs) have not been described in plants. *Arabidopsis* appears to lack genes for cytokeratin or vimentin, the main components of animal IFs, but has several variants of actin, α- and β-tubulin. The *Arabidopsis* genome also encodes homologues of chaperones that mediate the folding of tubulin and actin polypeptides in yeast and animal cells, such as the prefoldin and cytosolic chaperonin complexes and tubulin-folding cofactors. The dynamic stability of MTs and AFs is influenced by MT-associated proteins and actin-binding proteins, respectively, several of which are encoded by *Arabidopsis* genes. These include the MT-severing ATPase katanin, AF-cross-linking/bundling proteins, such as fimbrins and villins, and AF-disassembling proteins, such as profilin and actin-depolymerizing factor/cofilin. The *Arabidopsis* proteome appears to lack homologues of proteins that, in animal cells, link the actin cytoskeleton across the plasma membrane to the extracellular matrix, such as integrin, talin, spectrin, α-actinin, vitronectin or vinculin. This apparent lack of 'anchorage' proteins is consistent with the different composition of the cell wall and with a prominence of cortical MTs at the expense of cortical AFs in plant cells.

Plant-specific cytoskeletal arrays include interphase cortical MTs mediating cell shape, the preprophase band marking the cortical site of cell division, and the phragmoplast assisting in cytokinesis[76]. Although plant cells lack structural counterparts of the yeast spindle pole body and the animal centrosome, *Arabidopsis* has homologues of core components of the MT-nucleating γ-tubulin ring complex, such as γ-tubulin, Spc97/hGCP2 and Spc98/hGCP3. *Arabidopsis* has numerous motor molecules, both kinesins and dyneins with associated dynactin complex proteins, which are presumably involved in the dynamic organization of MTs and in transporting cargo along MT tracks. There are also myosin motors that may be involved in AF-supported organelle trafficking. Essential features of the eukaryotic cytoskeleton appear to be conserved in *Arabidopsis*.

The *Arabidopsis* genome encodes homologues of proteins involved in vesicle budding, including several ARFs and ARF-related small G-proteins, large but not small ARF GEFs (adenosine ribosylation factor on guanine nucleotide exchange factor), adapter proteins, and coat proteins of the COP and non-COP types. *Arabidopsis* also has homologues of proteins involved in vesicle docking and fusion, including SNAP receptors (SNAREs), *N*-ethylmaleimide-sensitive factor (NSF) and Cdc48-related ATPases, accessory proteins such as Sec1 and soluble NSF attachment protein (SNAP), and Rab-type GTPases. The large number of *Arabidopsis* SNAREs can be grouped by sequence similarity to yeast and animal counterparts involved in specific trafficking pathways, and some have been localized to the trans-Golgi and the pre-vacuolar pathway[77]. *Arabidopsis* also has a receptor for retention of proteins in the endoplasmic reticulum, a cargo receptor for transport to the vacuole and several phragmoplastins related to animal dynamin GTPases. Thus, plant cells appear to use the same basic machinery for vesicle trafficking as yeast and animal cells.

Animal cells possess many functionally diverse small G-proteins of the Ras superfamily involved in signal transduction, AF reorganization, vesicle fusion and other processes. Surprisingly, *Arabidopsis* appears to lack genes for G-proteins of the Ras, Rho, Rac and Cdc42 subfamilies but has many Rab-type G-proteins involved in vesicle fusion and several Rop-type G-proteins, one of which has a role in actin organization of the tip-growing pollen tube[78]. The significance of this divergent amplification of different subfamilies of small G-proteins in plants and animals remains to be determined.

Arabidopsis possesses cyclin-dependent kinases (CDKs), including a plant-specific Cdc2b kinase expressed in a cell-cycle-dependent manner, several cyclin subtypes, including a D-type cyclin that

mediates cytokinin-stimulated cell-cycle progression[79], a retinoblastoma-related protein and components of the ubiquitin-dependent proteolytic pathway of cyclin degradation. In yeast and animal cells, chromosome condensation is mediated by condensins, sister chromatids are held together by cohesins such as Scc1, and metaphase–anaphase transition is triggered by separin/Esp1 endopeptidase proteolysis of Scc1 on APC-mediated degradation of its inhibitor, securin/Psd1. Related proteins are encoded by the *Arabidopsis* genome. Thus, the basic machinery of cell-cycle progression, genome duplication and segregation appears to be conserved in plants. By contrast, entry into M phase, M-phase progression and cytokinesis seem to be modified in plant cells. *Arabidopsis* does not appear to have homologues of Cdc25 phosphatase, which activates Cdc2 kinase at the onset of mitosis, or of polo kinase, which regulates M-phase progression in yeast and animals. Conversely, plant-specific mitogen-actived protein (MAP) kinases appear to be involved in cytokinesis.

Cytokinesis partitions the cytoplasm of the dividing cell. Yeast and animal cells expand the membrane from the surface towards the centre in a cleavage process supported by septins and a contractile ring of actin and type II myosin. By contrast, plant cytokinesis starts in the centre of the division plane and progresses laterally. A transient membrane compartment, the cell plate, is formed *de novo* by fusion of Golgi-derived vesicles trafficking along the phragmoplast MTs[80]. Consistent with the unique mode of plant cytokinesis, *Arabidopsis* appears to lack genes for septins and type II myosin. Conversely, cell-plate formation requires a cytokinesis-specific syntaxin that has no close homologue in yeast and animals. Although syntaxin-mediated membrane fusion occurs in animal cytokinesis and cellularization, the vesicles are delivered to the base of the cleavage furrow. Thus, the plant-specific mechanism of cell division is linked to conserved eukaryotic cell-cycle machinery.

Two main conclusions are suggested by this comparative analysis. First, *Arabidopsis* and eukaryotic cells have common features related to intracellular activities, such as vesicle trafficking, cytoskeleton and cell cycle. Second, evolutionarily divergent features, such as organization of the cytoskeleton and cytokinesis, appear to relate to the plant cell wall.

Development

The regulation of development in *Arabidopsis*, as in animals, involves cell–cell communication, hierarchies of transcription factors, and the regulation of chromatin state; however, there is no reason to suppose that the complex multicellular states of plant and animal development have evolved by elaborating the same general processes during the 1.6 billion years since the last common unicellular ancestor of plants and animals[81,82]. Our genome analyses reflect the long, independent evolution of many processes contributing to development in the two kingdoms.

Plants and animals have converged on similar processes of pattern formation, but have used and expanded different transcription factor families as key causal regulators. For example, segmentation in insects and differentiation along the anterior–posterior and limb axes in mammals both involve the spatially specific activation of a series of homeobox gene family members. The pattern of activation is causal in the later differentiation of body and limb axis regions. In plants the pattern of floral whorls (sepals, petals, stamens, carpels) is also established by the spatially specific activation of members of a family of transcription factors, but in this instance the family is the MADS box family. Plants also have homeobox genes and animals have MADS box genes, implying that each lineage invented separately its mechanism of spatial pattern formation, while converging on actions and interactions of transcription factors as the mechanism. Other examples show even greater divergence of plant and animal developmental control. Examples are the AP2/EREBP and NAC families of transcription factors, which have important roles in flower and meristem development; both families are so far found only in plants (Supplementary Information Table 6).

A similar story can be told for cell–cell communication. Plants do not seem to have receptor tyrosine kinases, but the *Arabidopsis* genome has at least 340 genes for receptor Ser/Thr kinases, belonging to many different families, defined by their putative extracellular domains (Supplementary Information Table 7). Several families have members with known functions in cell–cell communication, such as the CLV1 receptor involved in meristem cell signalling, the S-glycoprotein homologues involved in signalling from pollen to stigma in self-incompatible *Brassica* species, and the BRI1 receptor necessary for brassinosteroid signalling[83]. Animals also have receptor Ser/Thr kinases, such as the transforming growth factor-β (TGF-β) receptors, but these act through SMAD proteins that are absent from *Arabidopsis*. The leucine-rich repeat (LRR) family of *Arabidopsis* receptor kinases shares its extracellular domain with many animal and fungal proteins that do not have associated kinase domains, and there are at least 122 *Arabidopsis* genes that code for LRR proteins without a kinase domain. Other *Arabidopsis* receptor kinase families have extracellular domains that are unfamiliar in animals. Thus, evolution is modular, and the plant and animal lineages have expanded different families of receptor kinases for a similar set of developmental processes.

Several *Arabidopsis* genes of developmental importance appear to be derived from a cyanobacteria-like genome (Supplementary Information Table 2), with no close relationship to any animal or fungal protein. One salient example is the family of ethylene receptors; another gene family of apparent chloroplast origin is the phytochromes—light receptors involved in many developmental decisions (see below). Whereas the land plant phytochromes show clear homology to the cyanobacterial light receptors, which are typical prokaryotic histidine kinases, the plant phytochromes are histidine kinase paralogues with Ser/Thr specificity[84]. Similarly to the ethylene receptors, the proteins that act downstream of plant phytochrome signalling are not found in cyanobacteria, and thus it appears that a bacterial light receptor entered the plant genome through horizontal transfer, altered its enzymatic activity, and became linked to a eukaryotic signal transduction pathway. This infusion of genes from a cyanobacterial endosymbiont shows that plants have a richer heritage of ancestral genes than animals, and unique developmental processes that derive from horizontal gene transfer.

Signal transduction

Being generally sessile organisms, plants have to respond to local environmental conditions by changing their physiology or redirecting their growth. Signals from the environment include light and pathogen attack, temperature, water, nutrients, touch and gravity. In addition to local cellular responses, some stimuli are communicated across the plant body, with plant hormones and peptides acting as secondary messengers. Some hormones, such as auxin, are taken up into the cell, whereas others, such as ethylene and brassinosteroids, and the peptide CLV3, act as ligands for receptor kinases on the plasma membrane. No matter where the signal is perceived by the cell, it is transduced to the nucleus, resulting in altered patterns of gene expression.

Comparative genome analysis between *Arabidopsis*, *C. elegans* and *Drosophila* supports the idea that plants have evolved their own pathways of signal transduction[85]. None of the components of the widely adopted signalling pathways found in vertebrates, flies or worms, such as Wingless/Wnt, Hedgehog, Notch/lin12, JAK/STAT, TGF-β/SMADs, receptor tyrosine kinase/Ras or the nuclear steroid hormone receptors, is found in *Arabidopsis*. By contrast, brassinosteroids are ligands of the BRI1 Ser/Thr kinase, a member of the largest recognizable class of transmembrane sensors encoded by 340 receptor-like kinase (RLK) genes in the *Arabidopsis* genome (Supplementary Information Table 7). With a few notable exceptions, such as CLV1, the types of ligands sensed by RLKs are

completely unknown, providing an enormous future challenge for plant biologists. G-protein-coupled receptors (GPCRs)/ seven-transmembrane proteins are an abundant class of proteins in mammalian genomes, instrumental in signal transduction. INTERPRO detected 27 GPCR-related domains in *Arabidopsis* (Supplementary Information Table 1), although there is no direct experimental evidence for these. *Arabidopsis* contains a family of 18 seven-transmembrane proteins of the mildew resistance (MlO) class, several of which are involved in defence responses. Notably, only single Gα (GPA1) and Gβ (AGB1) subunits are found in *Arabidopsis*, both previously known[86].

Although cyclic GMP has been proposed to be involved in signal transduction in *Arabidopsis*[87], a protein containing a guanylate cyclase domain was not identified in our analyses. Nevertheless, cyclic nucleotide-binding domains were detected in various proteins, indicating that cNMPs may have a role in plant signal transduction. Thus, although cNMP-binding domains appear to have been conserved during evolution, cNMP synthesis in *Arabidopsis* may have evolved independently.

We were unable to identify a protein with significant similarity to known Gγ subunits, but recent biochemical studies suggest that a protein with this functional capacity is likely to be present in plant cells (H. Ma, personal communication). Therefore, there is potential for the formation of only a single heterotrimeric G-protein complex; however, its functional interaction with any of the potential GPCR-related proteins remains to be determined.

Modules of cellular signal pathways from bacteria and animals have been combined and new cascades have been innovated in plants. A pertinent example is the response to the gaseous plant hormone ethylene[88]. Ethylene is perceived and its signal transmitted by a family of receptors related to bacterial-type two-component histidine kinases (HKs). In bacteria, yeast and plants, these proteins sense many extracellular signals and function in a His-to-Asp phosphorelay network[89]. In turn, these proteins physically interact with the genetically downstream protein CTR1, a Raf/MAPKKK-related kinase, revealing the juxtaposition of bacterial-type two-component receptors and animal-type MAP kinase cascades. Unlike animals, however, *Arabidopsis* does not seem to have a Ras protein to activate the MAP kinase cascade. MAP kinases are found in abundance in *Arabidopsis*: we identified ~20, a higher number than in any other eukaryote. As potentially counteracting components, we found ~70 putative PP2C protein phosphatases. Although this group is largely uncharacterized functionally, several members are related to ABI1/ABI2, key negative regulators in the signalling pathway for the plant hormone abscisic acid. Additional components of the His-to-Asp phosphorelay system were also found in *Arabidopsis*, including authentic response regulators (ARRs), pseudoresponse regulators (PRRs) and phosphotransfer intermediate protein (HPt)[90]. We found 11 HKs in the proteome (3 new), 16 RRs (2 new) and 8 PRRs (2 new). The biological roles of most ARRs, PRRs and HPts are largly unknown, but several have been found to have diverse functions in plants, including transcriptional activation in response to the plant hormone cytokinin[91], and as components of the circadian clock[92].

Plants seem to have evolved unique signalling pathways by combining a conserved MAP kinase cascade module with new receptor types. In many cases, however, the ligands are unknown. Conversely, some known signalling molecules, such as auxin, are still in search of a receptor. Auxin signalling may represent yet another plant-specific mode of signalling, with protein degradation through the ubiquitin-proteasome pathway preceding altered gene expression. With many *Arabidopsis* genes encoding components of the ubiquitin-proteasome pathway, elimination of negative regulators may be a more widespread phenomenon in plant signalling.

Recognizing and responding to pathogens

Plants are constantly exposed to pests, parasites and pathogens and have evolved many defences. In mammals, polymorphism for parasite recognition encoded in the MHC genes contributes to resistance. In plants, disease resistance (*R*) genes that confer parasite recognition are also extremely polymorphic. This polymorphism has been proposed to restrict parasites, and its absence may explain the breakdown of resistance in crop monocultures[93]. In contrast to MHC genes, plant resistance genes are found at several loci, and the complete genome sequence enables analysis of their complement and structure. Parasite recognition by resistance genes triggers defence mechanisms through various signalling molecules, such as protein kinases and adapter proteins, ion fluxes, reactive oxygen intermediates and nitric oxide. These halt pathogen colonization through transcriptional activation of defence genes and a form of programmed cell death called the hypersensitive response[94]. The *Arabidopsis* genome contains diverse resistance genes distributed at many loci, along with components of signalling pathways, and many other genes whose role in disease resistance has been inferred from mutant phenotypes.

Most resistance genes encode intracellular proteins with a nucleotide-binding (NB) site typical of small G proteins, and carboxy-terminal LRRs[95]. Their amino termini either carry a TIR domain, or a putative coiled coil (CC). There are 85 TIR–NB–LRR resistance genes at 64 loci, and 36 CC–NB–LRR resistance genes at 30 loci. Some NB–LRR resistance genes express neither obvious TIR nor CC domains at their N termini. This potential class is present seven times, at six loci. There are 15 truncated TIR–NB genes that lack an LRR at 10 loci, often adjacent to full TIR–NB–LRR genes. There are also six CC–NB genes, at five loci. These truncated products may function in resistance. Intriguingly, two TIR–NB–LRR genes carry a WRKY domain, found in transcription factors that are implicated in plant defence, and one of these also encodes a protein kinase domain.

Resistance gene evolution may involve duplication and divergence of linked gene families[36]; however, most (46) resistance genes are singletons; 50 are in pairs, 21 are in 7 clusters of 3 family members, with single clusters of 4, 5, 7, 8 and 9 members, respectively. Of the non-singletons, ~60% of pairs are in direct repeats, and ~40% are in inverted repeats. Resistance genes are unevenly distributed between chromosomes, with 49 on chromosome 1; 2 on chromosome 2; 16 on chromosome 3; 28 on chromosome 4; and 55 on chromosome 5.

In other plant species, resistance genes encode both transmembrane receptors for secreted pathogen products and protein kinases, and some other classes are also found. The *Cf* genes in tomato encode extracellular LRRs with a transmembrane domain and short cytoplasmic domain. Mutation in an *Arabidopsis* homologue, *CLAVATA2*, results in enlarged meristems, but to date no resistance function has been assigned to the 30 *Arabidopsis CLV2* homologues. *CLAVATA1*, a transmembrane LRR kinase, is also required for meristem function. *Xa21*, a rice LRR-kinase, confers *Xanthomonas* resistance, and the *Arabidopsis FLS2* LRR kinase confers recognition of flagellin. It has been proposed that CLV1 and CLV2 function as a heterodimer; perhaps this is also true for Xa21, FLS2 and Cf proteins. There are 174 LRR transmembrane kinases in *Arabidopsis*, with only *FLS2* assigned a role in resistance. A unique resistance gene, beet *Hs1pro-1*, which confers nematode resistance, has two *Arabidopsis* homologues.

The tomato Pto Ser/Thr kinase acts as a resistance protein in conjunction with an NB–LRR protein, so similar kinases might do the same for *Arabidopsis* NB–LRR proteins. There are 860 Ser/Thr kinases in the *Arabidopsis* sequence. Fifteen of these share 50% identity over the Pto-aligned region. The Toll pathway in *Drosophila* and mammals regulates innate immune responses through LRR/TIR domain receptors that recognize bacterial lipopolysaccharides[96]. Pto is highly homologous to *Drosophila* PELLE and mammalian IRAK protein kinases that mediate the TIR pathway.

Additional genes have been defined that are required for resistance by our analysis of the genome sequence. The *ndr1* mutation defines a gene required by the CC–NB–LRR gene *RPS2* and *RPM1*. *NDR1* is 1 of 28 *Arabidopsis* genes that are similar both to each other and to the tobacco *HIN1* gene that is transcriptionally induced early during the hypersensitive response. *EDS1* is a gene required for TIR–NB–LRR function, and like *PAD4*, encodes a protein with a putative lipase motif. *EDS1*, *PAD4* and a third gene comprise the *EDS1/PAD4* family. The *NPR1/NIM1/SAI1* gene is required for systemic acquired resistance, and we found five additional *NPR1* homologues. Recessive mutations at both the barley Mlo and *Arabidopsis LSD1* loci confer broad-spectrum resistance and derepress a cell-death program. There are at least 18 Mlo family members that resemble heterotrimeric GPCRs in *Arabidopsis*, and only two *LSD1* homologues.

One of the earliest responses to pathogen recognition is the production of reactive oxygen intermediates. This involves a specialized respiratory burst oxidase protein that transfers an electron across the plasma membrane to make superoxide. *Arabidopsis* encodes eight apparently functional *gp91* homologues, called *Atrboh* genes. Unlike *gp91*, they all carry an ~300 amino-acid N-terminal extension carrying an EF-hand Ca^{2+}-binding domain. In mammals, activation of the respiratory oxidative burst complex in the neutrophil, which includes gp91, requires the action of Rac proteins. As no Rac or Ras proteins are found in *Arabidopsis*, members of the large rop family of G proteins may carry this out. Similarly, we did not detect any *Arabidopsis* homologues of other mammalian respiratory burst oxidase components (p22, p47, p67, p40).

There are no clear homologues of many mammalian defence and cell-death control genes. Although nitric oxide production is involved in plant defence, there is no obvious homologue of nitric oxide synthase. Also absent are apparent homologues of the REL domain transcription factors involved in innate immunity in both *Drosophila* and mammals. We found no similarity to proteins involved in regulating apoptosis in animal cells, such as classical caspases, bcl2/ced9 and baculovirus p35. There are, however, 36 cysteine proteases. There are also eight homologues of a newly defined metacaspase family[97], two of which, along with LSD1, have a clear GATA-type zinc-finger.

Photomorphogenesis and photosynthesis

Because nearly all plants are sessile and most depend on photosynthesis, they have evolved unique ways of responding to light. Light serves as an energy source, as well as a trigger and modulator of complex developmental pathways, including those regulated by the circadian clock. Light is especially important during seedling emergence, where it stimulates chlorophyll production, leaf development, cotyledon expansion, chloroplast biogenesis and the coordinated induction of many nuclear- and chloroplast-encoded genes, while at the same time inhibiting stem growth. The goal of this process, called photomorphogenesis, is the establishment of a body plan that allows the plant to be an efficient photosynthetic machine under varying light conditions[98]. The signal transduction cascade leading to light-induced responses begins with the activation of photoreceptors. Next, the light signal is transduced via positively and negatively acting nuclear and cytoplasmic proteins, causing activation or derepression of nuclear and chloroplast-encoded photosynthetic genes and enabling the plant to establish optimal photoautotrophic growth. Although genetic and biochemical studies have defined many of the components in this process, the genome sequence provides an opportunity to identify comprehensively *Arabidopsis* genes involved in photomorphogenesis and the establishment of photoautotrophic growth. We identified at least 100 candidate genes involved in light perception and signalling, and 139 nuclear-encoded genes that potentially function in photosynthesis.

The roles have been described of only 35 of the 100 candidate photomorphogenic genes (Supplementary Information Table 8). All of the light photoreceptors had been discovered previously, including five red/far-red absorbing phytochromes (PHYA-E), two blue/ultraviolet-A absorbing cryptochromes (CRY1 and CRY2), one blue-absorbing phototropin (NPH1) and one NPH1-like (or NPL1). In contrast, we uncovered many new proteins similar to the photomorphogenesis regulators COP/DET/FUS, PKS1, PIF3, NDPK2, SPA1, FAR1, GIGANTEA, FIN219, HY5, CCA1, ATHB-2, ZEITLUPE, FKF1, LKP1, NPH3 and RPT2.

Both the phytochromes and NPH1 contain chromophores for light sensing coupled to kinase domains for signal transmission. Phytochromes have an N-terminal chromophore-binding domain, two PAS domains, and a C-terminal Ser/Thr kinase domain[99], whereas NPH1 has two LOV domains (members of the PAS domain superfamily) for flavin mononucleotide binding and a C-terminal Ser/Thr kinase domain[100]. PAS domains potentially sense changes in light, redox potential and oxygen energy levels, as well as mediating protein–protein interactions[99,100]. We searched for uncharacterized proteins with the combination of a kinase domain and either a phytochrome chromophore-binding site or PAS domains. Although we found no new phytochrome-like genes, we did identify four predicted proteins that contain PAS and kinase domains (Supplementary Information Fig. 6). These proteins share 80% amino-acid identity, but, unlike NPH1 and NPL1, have only one PAS domain. The combination of potential signal sensing and transmitting domains makes it tempting to speculate that these proteins may be receptors for light or other signals.

Our screen included searches for components of photosynthetic reaction centres and light-harvesting complexes, enzymes involved in CO_2 fixation and enzymes in pigment biosynthesis. We identified 11 core proteins of photosystem I, including the eukaryotic-specific components PsaG and PsaH[101], and 8 photosystem II proteins, including a single member (psbW) of the photosystem II core. We also found 26 proteins similar to the Chlorophyll-a/b binding proteins (8 Lhca and 18 Lhcb). Of the seven subunits of the cytochrome b_6f complex (PetA–D, PetG, PetL, PetM), only one (*PetC*) was found in the nuclear genome, whereas the remainder are probably encoded in the chloroplast. Similarly, of the nine subunits of the chloroplast ATP synthase complex, three are encoded in the nucleus, including the II- , γ- and δ-subunits; the remaining subunits (I, III, IV, α, β, ϵ) are encoded in the chloroplast[102]. Ten genes were related to the soluble components of the electron transfer chain, including two plastocyanins, five ferredoxins and three ferredoxin/NADP oxidoreductases. Forty genes are predicted to have a role in CO_2 fixation, including all of the enzymes in the Calvin–Benson cycle. For pigment biosynthesis, 16 genes in chlorophyll biosynthesis and 31 genes in carotenoid biosynthesis were found (Supplementary Information Table 8). Our analyses have identified several potential components of the light perception pathway, and have revealed the complex distribution of components of the photosynthetic apparatus between nuclear and plastid genomes.

Metabolism

Arabidopsis is an autotrophic organism that needs only minerals, light, water and air to grow. Consequently, a large proportion of the genome encodes enzymes that support metabolic processes, such as photosynthesis, respiration, intermediary metabolism, mineral acquisition, and the synthesis of lipids, fatty acids, amino acids, nucleotides and cofactors[103]. With respect to these processes, *Arabidopsis* appears to contain a complement of genes similar to those in the photoautotropic cyanobacterium *Synechocystis*[45], but, whereas *Synechocystis* generally has a single gene encoding an enzyme, *Arabidopsis* frequently has many. For example, *Arabidopsis* has at least seven genes for the glycolytic enzyme pyruvate kinase, with an

additional five for pyruvate kinase-like proteins. Whatever the reason for this high level of redundancy, it varies from gene to gene in the same pathway; the 11 enzymes of glycolysis are encoded by up to 51 genes that are present in as few as one or as many as eight copies. Similarly, of the 59 genes encoding proteins involved in glycerolipid metabolism, 39 are represented by more than one gene[104]. Genome duplication and expansion of gene families by tandem duplication have contributed to this diversity.

This high degree of apparent structural redundancy does not necessarily imply functional redundancy. For instance, although there are seven genes for serine hydroxymethyltransferase, a mutation in the gene for the mitochondrial form completely blocks the photorespiratory pathway[105]. Although there are 12 genes for cellulose synthase, mutations in at least 2 of the 12 confer distinct phenotypes because of tissue-specific gene expression[106].

The metabolome of *Arabidopsis* differs from that of cyanobacteria, or of any other organism sequenced to date, by the presence of many genes encoding enzymes for pathways that are unique to vascular plants. In particular, although relatively little is known about the enzymology of cell-wall metabolism, more than 420 genes could be assigned probable roles in pathways responsible for the synthesis and modification of cell-wall polymers. Twelve genes encode cellulose synthase, and 29 other genes encode 6 families of structurally related enzymes thought to synthesize other major polysaccharides[106]. Roughly 52 genes encode polygalacturonases, 20 encode pectate lyases and 79 encode pectin esterases, indicating a massive investment in modifying pectin. Similarly, the presence of 39 β-1,3-glucanases, 20 endoxyloglucan transglycosylases, 50 cellulases and other hydrolases, and 23 expansins reflects the importance of wall remodelling during growth of plant cells. Excluding ascorbate and glutathione peroxidases, there are 69 genes with significant similarity to known peroxidases and 15 laccases (diphenol oxidases). Their presence in such abundance indicates the importance of oxidative processes in the synthesis of lignin, suberin and other cell-wall polymers. The high degree of apparent redundancy in the genes for cell-wall metabolism might reflect differences in substrate specificity by some of the enzymes.

The high degree of apparent redundancy in the genes for cell wall metabolism might reflect differences in substrate specificity by some of the enzymes. It is already known that cell types have different wall compositions, which may require that the relevant enzymes be subject to cell-type-specific transcriptional regulation. Of the 40 or so cell types that plants make, almost all can be identified by unique features of their cell wall[107]. A large number of genes involved in wall metabolism have yet to be defined. Although more than 60 genes for glycosyltransferases can be found in the genome sequence, most of these are probably involved in protein glycosylation or metabolite catabolism and do not seem to be adequate to account for the polysaccharide complexity of the wall. For instance, at least 21 enzymes are required just to produce the linkages of the pectic polysaccharide RGII, and none of these enzymes has been identified at present. Thus, if these and related enzymes involved in the synthesis of other cell-wall polymers are also represented by multiple genes, a substantial number of the genes of currently unknown function may be involved in cell-wall metabolism.

Higher plants collectively synthesize more than 100,000 secondary metabolites. Because flowering plants are thought to have similar numbers of genes, it is apparent that a great deal of enzyme creation took place during the evolution of higher plants. An important factor in the rapid evolution of metabolic complexity is the large family of cytochrome P450s that are evident in *Arabidopsis* (Supplementary Information Table 1). These enzymes represent a superfamily of haem-containing proteins, most of which catalyse NADPH- and O_2-dependent hydroxylation reactions. Plant P450s participate in myriad biochemical pathways including those devoted to the synthesis of plant products, such as phenylpropanoids, alkaloids, terpenoids, lipids, cyanogenic glycosides and glucosinolates, and plant growth regulators, such as gibberellins, jasmonic acid and brassinosteroids. Whereas *Arabidopsis* has ~286 P450 genes, *Drosophila* has 94, *C. elegans* has 73 and yeast has only 3. This low number in yeast indicates that there are few reactions of basic metabolism that are catalysed by P450s. It seems likely that many animal P450s are involved in detoxification of compounds from food plant sources. The role of endogenous enzymes is poorly understood; only a few dozen P450 enzymes from plants have been characterized to any extent. The discrepancy between the number of known P450-catalysed reactions and the number of genes suggests that *Arabidopsis* produces a relatively large number of metabolites that have yet to be identified.

In addition to the large number of cytochrome P450s, *Arabidopsis* has many other genes that suggest the existence of pathways or processes that are not currently known. For instance, the presence of 19 genes with similarity to anthranilate *N*-hydroxycinnamoyl/benzoyl transferase is currently inexplicable. This enzyme is involved in the synthesis of dianthramide phytoalexins in Caryophyllaceae and Gramineae. No phytoalexins of this class have been described in *Arabidopsis* as yet. Similarly, the presence of 12 genes with sequence similarity to the berberine bridge enzyme, ((*S*)-reticuline:oxygen oxidoreductase (methylene-bridge-forming); EC 1.5.3.9), and 13 genes with similarity to tropinone reductase, suggests that *Arabidopsis* may have the ability to produce alkaloids. In other plants, the berberine bridge enzyme transforms reticuline into scoulerine, a biosynthetic precursor to a multitude of species-specific protopine, protoberberine and benzophenanthridine alkaloids. The discovery of these and many other intriguing genes in the *Arabidopsis* genome has created a wealth of new opportunities to understand the metabolic and structural diversity of higher plants.

Concluding remarks

The twentieth century began with the rediscovery of Mendel's rules of inheritance in pea[108], and it ends with the elucidation of the complete genetic complement of a model plant, *Arabidopsis*. The analysis of the completed sequence of a flowering plant reported here provides insights into the genetic basis of the similarities and differences of diverse multicellular organisms. It also creates the potential for direct and efficient access to a much deeper understanding of plant development and environmental responses, and permits the structure and dynamics of plant genomes to be assessed and understood.

Arabidopsis, *C. elegans* and *Drosophila* have a similar range of 11,000–15,000 different types of proteins, suggesting this is the minimal complexity required by extremely diverse multicellular eukaryotes to execute development and respond to their environment. We account for the larger number of gene copies in *Arabidopsis* compared with these other sequenced eukaryotes with two possible explanations. First, independent amplification of individual genes has generated tandem and dispersed gene families to a greater extent in *Arabidopsis*, and unequal crossing over may be the predominant mechanism involved. Second, ancestral duplication of the entire genome and subsequent rearrangements have resulted in segmental duplications. The pattern of these duplications suggests an ancient polyploidy event, and mutant analysis indicates that at least some of the many duplicate genes are functionally redundant. Their occurrence in a functionally diploid genetic model came as a surprise, and is reminiscent of the situation in maize, an ancient segmental allotetraploid. The remarkable degree of genome plasticity revealed in the large-scale duplications may be needed to provide new functions, as alternative promoters and alternative splicing appear to be less widely used in plants than they are in animals. Apart from duplicated segments, the overall chromosome structure of *Arabidopsis* closely resembles that of *Drosophila*; transposons and other repetitive sequences are concentrated in the heterochromatic regions surrounding the centromere,

whereas the euchromatic arms are largely devoid of repetitive equences. Conversely, most protein-coding genes reside in the uchromatin, although a number of expressed genes have been dentified in centromeric regions. Finally, *Arabidopsis* is the first nethylated eukaryotic genome to be sequenced, and will be invalu-ble in the study of epigenetic inheritance and gene regulation.

Unlike most animals, plants generally do not move, they can erpetuate indefinitely, they reproduce through an extended hap-oid phase, and they synthesize all their metabolites. Our compar-son of *Arabidopsis*, bacterial, fungal and animal genomes starts to lefine the genetic basis for these differences between plants and ther life forms. Basic intracellular processes, such as translation or esicle trafficking, appear to be conserved across kingdoms, reflect-ng a common eukaryotic heritage. More elaborate intercellular rocesses, including physiology and development, use different sets f components. For example, membrane channels, transporters and ignalling components are very different in plants and animals, and he large number of transcription factors unique to plants contrasts vith the conservation of many chromatin proteins across the three ukaryotic kingdoms. Unexpected differences between seemingly imilar processes include the absence of intracellular regulators of ell division (Cdc25) and apoptosis (Bcl-2). On the other hand, NA repair appears more highly conserved between plants and nammals than within the animal kingdom, perhaps reflecting ommon factors such as DNA methylation. Our analysis also hows that many genes of the endosymbiotic ancestor of the plastid ave been transferred to the nucleus, and the products of this rich rokaryotic heritage contribute to diverse functions such as photo-utotrophic growth and signalling.

The sequence reported here changes the fundamental nature of lant genetic analysis. Forward genetics is greatly simplified as nutations are more conveniently isolated molecularly, but at the ame time extensive gene duplications mean that functional redun-lancy must be taken into account. At a biochemical level, the pecificity conferred by nucleotide sequence, and the completeness f the survey allow complex mixtures of RNA and protein to be esolved into their individual components using micro-arrays and nass spectrometry. This specificity can also be used in the parallel nalysis of genome-wide polymorphisms and quantitative traits in atural populations[109]. Looking ahead, the challenge of determining he function of the large set of predicted genes, many of which are lant-specific, is now a clear priority, and multinational programs ave been initiated to accomplish this goal using site-selected nutagenesis among the the necessary tools[110]. Finally, productive aths of crop improvement, based on enhanced knowledge of *Arabidopsis* gene function, will help meet the challenge of sustaining ur food supply in the coming years.

Note added in proof: at the time of publication 17 centromeric BACs nd 5 sequence gaps in chromosome arms are being sequenced. □

Methods

he three centres used similar annotation approaches involving in silico gene-finding ethods, comparison to EST and protein databases, and manual reconciliation of that ata. Gene finding involved three steps: (1) analysis of BAC sequences using a computa-onal gene finder; (2) alignment of the sequence to the protein and EST databases; (3) ssignment of functions to each of the genes. Genscan[111], GeneMark.HMM[112], Xgrail[113] enefinder (P. Green, unpublished software) and GlimmerA[114] were used to analyse BAC quences. All of these systems were specially trained for *Arabidopsis* genes. Splice sites ere predicted using NetGene2[115], Splice Predictor[116] and GeneSplicer (M. Pertea and Salzberg, unpublished software). For the second step, BACs were aligned to ESTs and to e *Arabidopsis* gene index[117] using programs such as DDS/GAP2[118] or BLASTN[119]. egmental duplications were analysed and displayed using a modified version of IALIGN2 (ref. 120).

eceived 20 October; accepted 15 November 2000.

The *C. elegans* Sequencing Consortium. Sequence and analysis of the genome of *C. elegans*. *Science* **282,** 2012–2018 (1998).

Adams, M. D. The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).

Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282,** 662–665 (1998).

Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402,** 761–768 (1999).

Mayer, K. *et al.* Sequence and analysis of chromsome 4 of the plant *Arabidopsis thaliana*. *Nature* **402,** 769–777 (1999).

Theologis, A. *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408,** 816–820 (2000).

Salanoubat, M. *et al.* Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408,** 820–822 (2000).

Tabata, S. *et al.* Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* **408,** 820–822 (2000).

Choi, S. D., Creelman, R., Mullet, J. & Wing, R. A. Construction and characterisation of a bacterial artificial chromosome library from *Arabidopsis thaliana*. *Weeds World* **2,** 17–20 (1995).

Mozo, T., Fischer, S., Shizuya, H. & Altmann, T. Construction and characterization of the IGF *Arabidopsis* BAC library. *Mol. Gen. Genet.* **258,** 562–570 (1998).

Lui, Y. -G., Mitsukawa, N., Vazquez-Tello, A. & Whittier, R. F. Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.* **7,** 351–358 (1995).

Lui, Y. -G. *et al.* Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. *Proc. Natl Acad. Sci. USA* **96,** 6535–6540 (1999).

Marra, M. *et al.* A map or sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).

Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 271–275 (1999).

Sato, S. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1. 6 Mb regions covered by twenty physically assigned P1 clones. *DNA Res.* **4,** 215–230 (1997).

Bent, E., Johnson, S. & Bancroft, I. BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana*. *Plant J.* **13,** 849–855 (1998).

Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).

Meyerowitz, E. M. & Somerville, C. R. *Arabidopsis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1994)

Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

Pavy, N. *et al.* Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *BioInformatics* **15,** 887–900 (1999).

Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387** (Suppl.) 7–65 (1997).

Frishman, D. *et al.* Functional and structural genomics using PEDANT. *BioInformatics* (in the press).

Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1462 (1997).

Kotani, H. & Tabata, S. Lessons from the sequencing of the genome of a unicellular cyanobacterium, *Synechocystis* SP. PCC6803. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49,** 151–171 (1998).

Apweiler, R. *et al. INTERPRO* (http://www. ebi. ac. uk/interpro/). Collaborative Computer Project 11 Newsletter no. 10 (Cambridge, 2000).

Bent, A. F. *et al. RPS2* of *Arabidopsis thaliana* a leucine-rich repeat class of plant disease resistance genes. *Science* **265,** 1856–1860 (1994).

Skowyra, D. *et al.* F box proteins are receptors that recruit phsphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* **91,** 209–219 (1997).

Joazeiro, C. A. P. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* **102,** 549–552 (2000).

Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).

Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27,** 2369–2376 (1999).

Blanc, G. *et al.* Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12,** 1093–1102 (2000).

Wendel, J. F. Genome evolution in polyploids. *Plant Mol. Biol.* **42,** 225–249 (2000).

Gaut, B. S. & Doebley, J. F. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl Acad. Sci. USA* **94,** 6809–6814 (1997).

Ku, H. -M., Vision, T., Liu, J. & Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci. USA* **97,** 9121–9126 (2000).

Noel, L. *et al.* Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. *Plant Cell* **11,** 2099–2111 (1999).

Ellis, J., Dodds, P. & Pryor, T. Structure, function, and evolution of plant disease resistance genes. *Trends Plant Sci.* **3,** 278–284 (2000).

Tanksley, S. D. *et al.* High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132,** 1141–1160 (1992).

Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Grasses, line up and form a circle. *Curr. Biol.* **5,** 737–739 (1995).

Acarkan, A., Rossberg, M., Koch, M. & Schmidt, R. Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. *Plant J.* **23,** 55–62 (2000).

Cavell, A., Lydiate, D., Parkin, I., Dean, C. & Trick, M. A 30 centimorgan segment of *Arabidopsis thaliana* chromosome 4 has six collinear homologues within the *Brassica napus* genome. *Genome* **41,** 62–69 (1998).

O'Neill, C. & Bancroft, I. Comparative physical mapping of segments of the genome of *Brassica oleracea* var *alboglabra* that are homologous to sequenced regions of the chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* **23,** 233–243 (2000).

Wolfe, K. H., Gouy, M., Yang, Y. -W., Sharp, P. M. & Li, W. -H. Date of the monocot-dicot divergence estimated from the chloroplast DNA sequence data. *Proc. Natl Acad. Sci. USA* **86,** 6201–6205 (1989).

van Dodeweerd, A. -M. *et al.* Identification and analysis of homologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome* **42,** 887–892 (1999)

Mayer, K. Sequence level analysis of homologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome Res.* (submitted).

Sato, S. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research* **6,** 283–290 (1999).

Unseld, M., Marienfeld, J., Brandt, P. & Brennicke, A. The mitochondrial genome in *Arabidopsis*

thaliana contains 57 genes in 366,924 nucleotides. *Nature Genet.* **15,** 57–61 (1997).
47. Palmer, J. D. *et al.* Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl Acad. Sci. USA* **97,** 6960–6966 (2000).
48. Le, Q. -H. *et al.* Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **97,** 7376–7381 (2000).
49. Yu, Z., Wright, S. & Bureau, T. *Mutator*-like elements (MULEs) in *Arabidopsis thaliana*: Structure, diversity and evolution. *Genetics* (in the press).
50. Feschotte, C. & Mouches, C. 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 (2000).
51. Martienssen, R. Transposons, DNA methylation and gene control. *Trends Genet.* **14,** 263–264 (1998).
52. Singer, T., Yordan, C. & Martienssen, R. Robertson's *Mutator* transposons in *Arabidopsis* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation (DDM1)*. *Genes Dev.* (in the press).
53. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768 (1996).
54. Copenhaver, G. P. & Pikaard, C. S. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9,** 273–282 (1996).
55. Fransz, P. *et al.* Cytogenetics for the model system *Arabidopsis thaliana*. *Plant J.* **13,** 867–876 (1998).
56. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukarotic telomere from *Arabidopsis thaliana*. *Cell* **53,** 127–136 (1988).
57. Round, E. K., Flowers, S. K. & Richards, E. J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7,** 1045–1053 (1997).
58. The CSHL/WUGSC/PEB *Arabidopsis* Sequencing Consortium. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100,** 377–386 (2000).
59. Fransz, P. F. *et al.* Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: Structural organization of heterochromatic knob and centromere region. *Cell* **100,** 367–376 (2000).
60. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301,** 75–101 (2000).
61. Paulsen, I. T., Sliwinski, M. K., Nelissen, B., Goffeau, A. & Saier, M. H. Jr Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett.* **430,** 116–125 (1998).
62. Hirsch, R. E., Lewis, B. D, Spalding, E. P. & Sussman, M. R. A role for the AKT1 potassium channel in plant nutrition. *Science* **280,** 918–921 (1998).
63. Slayman, C. L. & Slayman, C. W. Depolarization of the plasma membrane of Neurospora during active transport of glucose: evidence for a proton-dependent cotransport system. *Proc. Natl Acad. Sci. USA* **71,** 1035–1939 (1974).
64. Ryan, C. A. & Pearce, G. Systemin: a polypeptide signal for plant defensive genes. *Annu. Rev. Cell. Dev. Biol.* **14,** 1–17 (1998).
65. Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435,** 171–213 (1999).
66. Selby, C. P. & Sancar, A. Structure and function of transcription-repair coupling factor. Structural domains and binding properties. *J. Biol. Chem.* **270,** 4882–4889 (1995).
67. Britt, A. B. Molecular genetics of DNA repair in higher plants. *Trends Plant Sci.* **4,** 20–25 (1999).
68. Dangl, M. Response to Aravind, L. & Koonin, E. V. Second Family of Histone Deacetylases. *Science* **280,** 1167 (1998).
69. Cao, X. *et al.* Conserved plant genes with similarity to mammalian de novo DNA methyltransferases. *Proc. Natl Acad. Sci. USA* **97,** 4979–4984 (2000).
70. Henikoff, S. & Comai, L. A DNA methyltransferase homologue with a chromodomain exists in multiple polymorphic forms in *Arabidopsis*. *Genetics* **149,** 307–318 (1998).
71. Hung, M. -S. *et al. Drosophila* proteins related to vertebrate DNA (5-cytosine) methyltransferases. *Proc Natl Acad. Sci. USA* **96,** 11940–11945 (1999).
72. Dalmay, T., Hamilton, A. J., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent-RNA polymerase in *Arabidopsis* is required for post transcriptional gene silencing mediated by a transgene but not by a virus—the truth. *Cell* **101,** 543–553 (2000).
73. Riechmann, J. L. & Ratcliffe, O. J. A genomic perspective on plant transcription factors. *Curr. Opin. Plant Biol.* **3,** 423–434 (2000).
74. Liljegren, S. J. *et al.* SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* **404,** 766–770 (2000).
75. Pelaz, S. *et al.* B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* **405,** 200–203 (2000).
76. Canaday, J., Stoppin-Mellet, V., Mutterer, J., Lambert, A. M. & Schmit, A. C. Higher plant cells: gamma-tubulin and microtubule nucleation in the absence of centrosomes. *Microsc. Res. Technol.* **49,** 487–495 (2000).
77. Bassham, D. C. & Raikhel, N. V. Unique features of the plant vacuolar sorting machinery. *Curr. Opin. Cell Biol.* **12,** 491–495 (2000).
78. Zheng, Z. L. & Yang, Z. The Rrop GTPase switch turns on polar growth in pollen. *Trends Plant Sci.* **5,** 298-303 (2000).
79. den Boer, B. G. & Murray, J. A. Triggering the cell cycle in plants. *Trends Cell Biol.* **10,** 245–250 (2000).
80. Heese, M., Mayer, U. & Jurgens, G. Cytokinesis in flowering plants: cellular process and developmental integration. *Curr. Opin. Plant Biol.* **1,** 486–491 (1998).
81. Meyerowitz, E. M. Plants, animals, and the logic of development. *Trends Genet.* **15,** M65–M68 (1999).
82. Wang, D. Y. C. *et al.* Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Bio.* **266,** 63–171 (1999).
83. Torii, K. Receptor kinase activation and signal transduction in plants: an emerging picture. *Curr. Opin. Plant Biol.* **3,** 362–367 (2000).
84. Yeh, K. C. & Lagarias, J. C. Eukaryotic phytochromes: Light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl Acad. Sci. USA* **95,** 13976–13981 (1998).
85. McCarty, D. R. & Chory, J. Conservation and innovation in plant signaling pathways. *Cell* **103,** 201–211 (2000).
86. Weiss, C. A., Garnaat, C., Mukai, K., Hu, Y. & Ma, H. Molecular cloning of cDNAs from maize and *Arabidopsis* encoding a G protein beta subunit. *Proc. Natl Acad. Sci. USA* **91,** 9554–9558 (1994).
87. Bowler, C. *et al.* Cyclic GMP and calcium mediate phytochrome phototransduction. *Cell* **77,** 73–81 (1994).
88. Stepanova, A. & Ecker, J. R. Ethylene signaling: from mutants to molecules. *Curr. Opin. Plant Biol.* **3,** 353–360 (2000).
89. Urao, T., Yamaguchi-Shinozaki, K. & Shinozaki, K. Two-component systems in plant signal transduction. *Trends Plant Sci.* **5,** 67–74 (2000).
90. Makino, S. *et al.* Genes encoding pseudo-response regulators: Insight into His-to-Asp phosphorelay and circadian rhythm in *Arabidopsis thaliana*. *Plant Cell Physiol.* **41,** 791–803 (2000).
91. D'Agostino, I. B. & Kieber, J. J. Phosphorelay signal transduction: the emerging family of plant response regulators. *Trends Biol. Sci.* **24,** 452–456 (1999).
92. Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene TOC1, an autoregulatory response regulator homologue. *Science* **289,** 768–771 (2000).
93. Stahl, E. A. & Bishop, J. G. Plant-Pathogen arms races at the molecular level. *Curr. Opin. Plant Biol.* **3,** 299–304 (2000).
94. McDowell, J. M. & Dangl, J. L. Signal transduction in the plant innate immune response. *Trends Biochem. Sci.* **25,** 79–82 (2000).
95. Van der Biezen, E. A. & Jones, J. D. Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem Sci.* **23,** 454–456 (1998).
96. Belvin, M. P. & Anderson, K. V. A conserved signaling pathway: the *Drosophila* toll-dorsal pathway. *Annu. Rev. Cell. Dev. Biol.* **12,** 393–416 (1996).
97. Uren, A. G. *et al.* Identification of paracaspases and metacaspases: Two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol. Cell* **6,** 961–967 (2000).
98. Fankhauser, C. & Chory, J. Light control of plant development. *Annu. Rev. Cell. Dev. Biol.* **13,** 203–229 (1997).
99. Briggs, W. R. & Huala, E. Blue-light photoreceptors in higher plants. *Annu. Rev. Cell. Dev. Biol.* **15,** 33–62 (1999).
100. Christie, J. M., Salomon, M., Nozue, K., Wada, M. & Briggs, W. R. LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): binding sites for the chromophore flavin mononucleotide. *Proc. Natl Acad. Sci. USA* **96,** 8779–8783 (1999).
101. Golbeck, J. H. Structure and function of photosystem I. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43,** 293–324 (1992).
102. Maier, R. M., Neckermann, K., Igloi, G. L. & Kossel, H. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol.* **251,** 614–28 (1995).
103. Buchanan, B. B., Gruissem, W. & Jones, R. L. in *Biochemistry and Molecular Biology of Plants* 1367 (Am. Soc. Plant Physiol., Rockville, Maryland, 2000).
104. Mekhedov, S., Martínez de Ilárduya, O. & Ohlrogge, J. Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. *Plant Physiol.* **122,** 389–401 (2000).
105. Somerville, C. R., & Ogren, W. L. Photorespiration deficient mutants of *Arabidopsis thaliana* lacking mitochrondrial serine transhydroxymethylase activity. *Plant Physiol.* **67,** 666–671 (1981).
106. Richmond, T., & Somerville, C. R. The cellulose synthase superfamily. *Plant Physiol* **124,** 495–499 (1999).
107. Carpita, N. Vergara C: A recipe for cellulose. *Science* **279,** 672–673 (1998).
108. De Vries, H. Sur la loi de disjonction des hybrides. *C. R. Acad. Sci. Paris* **130,** 845–847 (1900).
109. Alonso-Blanco, C. & Koornneef, M. Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci.* **5,** 1360–1385 (1999).
110. Chory, J. Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiology* **123,** 423–425 (2000).
111. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
112. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).
113. Uberbacher, E. C. & Mural, R. J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88,** 11261–11265 (1991).
114. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
115. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452 (1996).
116. Brendel, V. & Kleffe, J. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.* **26,** 4748–4757 (1998).
117. Quackenbush, J., Liang, F., Holt, I., Pertea, G. & Upton, J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28,** 141–145 (2000).
118. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45 (1997).
119. Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
120. Morgenstern, B. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *BioInformatics* **15,** 211–218 (1999).
121. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).
122. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

This work was supported by the National Science Foundation (NSF) Cooperative Agreements (funded by the NSF, the US Department of Agriculture (USDA) and the US Department of Energy (DOE)), the Kazusa DNA Research Institute Foundation, and by the European Commission. Additional support from the USDA, Ministère de la Recherche, GSF-Forschungszentrum f. Umwelt u. Gesundheit, BMBF (Bundesministerium f. Bildung, Forschung und Technologie), the BBSRC (Biotechnology and Biological

Sciences Research Council) and the Plant Research International, Wageningen, is also gratefully acknowledged. The authors wish to thank E. Magnien, D. Nasser and J. D. Watson for their continual support and encouragement.

Correspondence and requests for materials should be addressed to The Arabidopsis Genome Initiative (e-mail: genomeanalysis@tigr.org or genomeanalysis@gsf.de).

The Arabidopsis Genome Initiative

Three groups contributed to the work reported here. The Genome Sequencing groups, arranged here in order of sequence contribution, sequenced and annotated assigned chromosomal regions. The Genome Analysis group carried out the analyses described. The Contributing Authors interpreted the genome analyses, incorporating other data and analyses, with respect to selected biological topics.

Genome Sequencing Groups

Samir Kaul, Hean L. Koo, Jennifer Jenkins, Michael Rizzo, Timothy Rooney, Luke J. Tallon, Tamara Feldblyum, William Nierman, Maria-Ines Benito, Xiaoying Lin, Christopher D. Town, J. Craig Venter & Claire M. Fraser

The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Satoshi Tabata, Yasukazu Nakamura, Takakazu Kaneko, Shusei Sato, Erika Asamizu, Tomohiko Kato, Hirokazu Kotani & Shigemi Sasamoto

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan

Joseph R. Ecker[1]*†, Athanasios Theologis[2]*, Nancy A. Federspiel[3]*†, Curtis J. Palm[3], Brian I. Osborne[2], Paul Shinn[1], Aaron B. Conway[3], Valentina S. Vysotskaia[2], Ken Dewar[1], Lane Conn[3], Catherine A. Lenz[2], Christopher J. Kim[1], Nancy F. Hansen[3], Shirley X. Liu[2], Eugen Buehler[1], Hootan Altafi[3], Hitomi Sakano[2], Patrick Dunn[1], Bao Lam[3], Paul K. Pham[2], Qimin Chao[1], Michelle Nguyen[3], Guixia Yu[2], Huaming Chen[1], Audrey Southwick[3], Jeong Mi Lee[2], Molly Miranda[3], Mitsue J. Toriumi[2] & Ronald W. Davis[3]

*1, Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104 USA; 2, Plant Gene Expression Center/USDA-U.C.Berkeley, 800 Buchanan Street, Albany, California 94710, USA; 3, Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA. * These authors contributed equally to this work. † Present addresses: The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA (J.R.E.); Exelixis, Inc., 170 Harborway, P.O. Box 511, South San Francisco, California 94083-0511, USA (N.A.F)*

European Union Chromosome 4 and 5 Sequencing Consortium: R. Wambutt[1], G. Murphy[2], A. Düsterhöft[3], W. Stiekema[4], T. Pohl[5], K.-D. Entian[6], N. Terryn[7] & G. Volckaert[8]

1, AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany; 2, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 3, QIAGEN GmbH, Max-Volmer-Str. 4, D-40724 Hilden, Germany; 4, Greenomics, Plant Research International, Droevendaalsesleeg 1, NL 6700, AA Wageningen, The Netherlands; 5, GATC GmbH, Fritz-Arnold Strasse 23, D-78467 Konstanz, Germany; 6, SRD GmbH, Oberurseler Str. 43, Oberursel 61440, Germany; 7, Department for Plant Genetics, (VIB), University of Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium; 8, Katholieke Universiteit Leuven, Laboratory of Gene Technology, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium

European Union Chromosome 3 Sequencing Consortium: M. Salanoubat[1], N. Choisne[1], M. Rieger[2], W. Ansorge[3], M. Unseld[4], B. Fartmann[5], G. Valle[6], F. Artiguenave[1], J. Weissenbach[1] & F. Quetier[1]

1, Genoscope and CNRS FRE2231, 2 rue G. Crémieux, 91057 Evry Cedex, France; 2, Genotype GmbH Angelhofweg 39, D-69259 Wilhemlsfeld, Germany; 3, European Molecular Biology Laboratory, Biochemical Instrumentation Program, Meyerhoftstr. 1, D-69117 Heidelberg, Germany; 4, LION Bioscience AG, Im Neuenheimer Feld 515-517, 69120 Heidelberg, Germany; 5, MWG-Biotech AG, Anzinger Strasse 7a, 85560 Ebersberg, Germany; 6, CRIBI, Università di Padova, via G. Colombo 3, Padova 35131, Italy

The Cold Spring Harbor and Washington University Genome Sequencing Center Consortium: Richard K. Wilson[1], Melissa de la Bastide[2], M. Sekhon[1], Emily Huang[2], Lori Spiegel[2], Lidia Gnoj[2], K. Pepin[1], J. Murray[1], D. Johnson[1], Kristina Habermann[2], Neilay Dedhia[2], Larry Parnell[2], Raymond Preston[2], L. Hillier[1], Ellson Chen[3], M. Marra[2], Robert Martienssen[4] & W. Richard McCombie[2]

1, Washington University Genome Sequencing Center, Washington University in St Louis School of Medicine, 4444 Forest Park Blvd., St. Louis, Missouri 63108 USA; 2, Lita Annenberg Hazen Genome Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 3, Celera Genomics, 850 Lincoln Center Drive, Foster City, California 94494, USA; 4, Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Genome Analysis Group

Klaus Mayer[1]*, Owen White[2]*, Michael Bevan[3], Kai Lemcke[1], Todd H. Creasy[2], Cord Bielke[2], Brian Haas[1], Dirk Haase[1], Rama Maiti[2], Stephen Rudd[1], Jeremy Peterson[2], Heiko Schoof[1], Dimitrij Frishman[1], Burkhard Morgenstern[1], Paulo Zaccaria[1], Maria Ermolaeva [2], Mihaela Pertea[2], John Quackenbush[2], Natalia Volfovsky[2], Dongying Wu[2], Todd M. Lowe[4], Steven L. Salzberg [2] & Hans-Werner Mewes[1]

*1, GSF-Forschungszentrum f. Umwelt u. Gesundheit, Munich Information Center for Protein Sequences, am Max-Planck-Institut f. Biochemie, Am Klopferspitz 18a, D-82152, Germany; 2, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; 3, Molecular Genetics Deartment, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 4, Dept Genetics, Stanford University Medical School, Stanford, California 94305-5120, USA. * These authors contributed qually to this work*

Contributing Authors

Comparative analysis of the genomes of *A. thaliana* accessions. S. Rounsley, D. Bush, S. Subramaniam, I. Levin & S. Norris

Cereon Genomics LLC, 45 Sidney St, Cambridge, Massachussetts 02139, USA

Comparative analysis of the genomes of *A. thaliana* and other genera. R. Schmidt[1], A. Acarkan[1] & I. Bancroft[2]

1, Max-Delbrück-Laboratorium in der Max-Planck-Gesellschaft, Carl-von-Linné-Weg 10, 50829 Cologne, Germany; 2, Brassicas and Oilseeds Research Department, John Innes Centre, Norwich NR4 7UJ, UK

Integration of the three genomes in the plant cell: the extent of protein and nucleic acid traffic between nucleus, plastids and mitochondria. F. Quetier[1], A. Brennicke[2] & J. A. Eisen[3].
1, Genoscope, Centre Nationale de Sequencage, 2 rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France; 2, Molekulare Botanik, Universität Ulm, 89069 Ulm, Germany; 3, The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Transposable elements. T. Bureau[1], B.-A. Legault[1], Q.-H. Le[1], N. Agrawal[1], Z. Yu[1] & R. Martienssen[2]
1, McGill University, Dept of Biology, 1205 rue Dr Penfield, Montreal, Quebec, H3A 1B1, Canada; 2, Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

rDNA, telomeres and centromeres. G. P. Copenhaver[1], S. Luo[1], C. S. Pikaard[2] & D. Preuss[1]
1, Howard Hughes Medical Institute, The University of Chicago, 1103 East 57th Street, Chicago, Illiois, USA; 2, Biology Department, Washington University in St Louis, St Louis, Missouri 63130, USA

Membrane transport. I. T. Paulsen[1] & M. Sussman[2]
1, The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA; 2, University of Wisconsin Biotechnology Center, 425 Henry Mall, Madison, Wisconsin 53706, USA

DNA repair and recombination. A. B. Britt[1] & J. A. Eisen[2]
1, Section of Plant Biology, University of California, Davis, California 95616, USA; 2, The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Gene regulation. D. A. Selinger[1], R. Pandey[1], D. W. Mount[2], V. L. Chandler[1], R. A. Jorgensen[1] & C. Pikaard[3]
1, Department of Plant Sciences, University of Arizona, 303 Forbes Hall; and 2, Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA; 3, Biology Department, Washington University in St Louis, St Louis, Missouri 63130, USA

Cellular organization. G. Juergens
Entwicklungsgenetik, ZMBP-Centre for Plant Molecular Biology, auf der Morgenstelle 1, Tuebingen D-72076, Germany

Development. E. M. Meyerowitz.
Division of Biology, California Institute of Biology, Pasadena, California 91125, USA

Signal transduction. J. R. Ecker[1] & A. Theologis[2].
1, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA; 2, Plant Gene Expression Center/USDA-UC Berkeley, 800 Buchanan Street, Albany, California 94710, USA

Recognition of and response to pathogens. J. Dangl[1] & J. D. G. Jones[2]
1, Biology Department, Coker Hall, University of North Carolina, Chapel Hill, North Carolina 27599, USA; 2, Sainsbury Laboratory, John Innes Centre, Colney Lane, Norwich NR4 7UJ, UK

Photomorphogenesis and photosynthesis. M. Chen & J. Chory
Howard Hughes Medical Institute and Plant Biology Laboratory, The Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

Metabolism. C. Somerville
Carnegie Institution, 260 Panama Street, Stanford, California 94305, USA