MUSTANG Is a Novel Family of Domesticated Transposase Genes Found in Diverse Angiosperms

Rebecca K. Cowan, Douglas R. Hoen, Daniel J. Schoen, and Thomas E. Bureau

McGill University, Biology Department, Montreal, Quebec H3A 1B1, Canada

While transposons have traditionally been viewed as genomic parasites or "junk DNA," the discovery of transposon-derived host genes has fueled an ongoing debate over the evolutionary role of transposons. In particular, while mobility-related open reading frames have been known to acquire host functions, the contribution of these types of events to the evolution of genes is not well understood. Here we report that genome-wide searches for Mutator transposase–derived host genes in Arabidopsis thaliana (Columbia-0) and Oryza sativa ssp. japonica (cv. Nipponbare) (domesticated rice) identified 121 sequences, including the taxonomically conserved MUSTANG1. Syntenic MUSTANG1 orthologs in such varied plant species as rice, poplar, Arabidopsis, and Medicago truncatula appear to be under purifying selection. However, despite the evidence of this pathway of gene evolution, MUSTANG1 belongs to only one of only two Mutator-like gene families with members in both monocotyledonous and dicotyledonous plants, suggesting that Mutator-like elements seldom evolve into taxonomically widespread host genes.

Introduction

Mobile elements have traditionally been viewed as "junk DNA" or as genomic parasites (Doolittle and Sapienza 1980; Orgel and Crick 1980). However, reports of transposon-derived genes contributing to cellular processes (Agrawal, Eastman, and Schatz 1998; Mi et al. 2000; Hudson, Lisch, and Quail 2003) have indicated that such genetic elements may influence the evolution of their host by providing a source of novel coding capacity. Other lines of evidence such as sequence analysis and conservation studies also support this view (International Human Genome Sequencing Consortium 2001; Hammer, Strehl, and Hagemann 2005; Zdobnov et al. 2005). This, in turn, has spurred controversy concerning the extent to which transposable elements contribute to host evolution (Hurst and Werren 2001).

In plants, the only transposon-like genes with known host functions, far-red impaired response protein 1 (FAR1), far-red elongated hypocotyl 3 (FHY3), and the FAR1-related sequences (FRS), are related to the family of DNA transposons called Mutator-like elements (MULEs) (Hudson, Lisch, and Quail 2003; Lin and Wang 2004; Yu, Wright, and Bureau 2000). MULEs are DNA transposons, which move via a "cut-and-paste" mechanism (Lisch 2002). The mudrA gene of autonomous elements encodes the transposase MURA that is required for MULE mobility (Lisch et al. 1999). Though most MULEs are bound by terminal inverted repeats (TIRs) of ~200 bp in length, some MULEs lack terminal symmetry and are termed non-TIR MULEs (Yu, Wright, and Bureau 2000). In maize, MURA binds specifically to the terminal sequences of MULEs (Benito and Walbot 1997). This association between the transposase and the terminal sequences is necessary for excision of the element to occur, as it is in other DNA transposons (Steiniger-White, Rayment, and Reznikoff 2004). Upon insertion of the MULE into the genome, 9- to 11-bp target site duplications (TSDs) are created (Bennetzen 1996).

Materials and Methods

Identification of Transposon-Associated and Transposon-Dissociated mudrA Genes

In order to determine the abundance of domesticated mudrA genes in plants, we searched the O. sativa ssp. japonica (cv. Nipponbare) (domesticated rice) and A. thaliana (Columbia-0) genomes for sequences with similarity to mudrA that were not associated with transposon structures such as terminal sequences and TSDs. mudrA-like sequences were identified by using Zea mays mudrA (GI:540581) as a query for a BlastP (Altschul et al. 1997) search (cutoff = E < 1 x 10^-10) of predicted genes from the Arabidopsis TIGR 5.0 release (http://www.tigr.org) and from rice build 2 pseudomolecules (DNA Data Bank of Japan accession numbers AP006867 through AP006877 and National Center for Biotechnology Information [NCBI] accession number AE016959), predicted using FGENESH (http://www.softberry.com/berry.phtml?topic=fgensh&group=help&subgroup=gfind) in

Key words: genome, evolution, transposons, plants.

E-mail: thomas.bureau@mcgill.ca.

doi:10.1093/molbev/msi202
Advance Access publication June 29, 2005

© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oupjournals.org
the monocot trained settings. Similarly, Z. mays mudrA (GI:540581) was used as a query for a Psi-Blast (Altschul et al. 1997) search against the rice predicted gene set and The Arabidopsis Information Resource (TAIR) UNIPROT data set (http://www.arabidopsis.org) (cutoff = E < 10\(^{-12}\)) second Psi-Blast iteration) in order to identify more divergent mudrA-like sequences. Mined mudrA genes are listed in supplementary table 1 (Supplementary Material online).

Whether the mined mudrA genes were associated with a transposon structure was determined by (a) identifying correctly oriented MULE termini separated by less than 30 kbp via homology searches with models and queries based on the TIR and non-TIR termini of previously identified families using HMMER (Eddy 1998) and BlastN (Altschul et al. 1997); (b) using BL2SEQ (Tatusova and Madden 1999) with an expect value of 100 and a mismatch penalty of −1 to identify TIRs in the 10-kbp flanking regions; (c) using the genes and flanking sequences as BlastN (Altschul et al. 1997) queries to identify repetitiveness extending beyond the putative-coding region; and (d) scanning extreme terminal sequences of TIRs and repetitive regions for TSDs. Genes with TSDs were considered transposon-associated mudrA genes (TAMs), those lacking TSDs but with other transposon features were considered potentially transposon-associated mudrA genes (pTAMs), while those lacking all transposon features were considered transposon-dissociated mudrA genes (TDMs).

Arabidopsis genes were considered to have open reading frames (ORFs) if they were so annotated (http://www.arabidopsis.org) or if they corresponded to a gene in the TAIR UNIPROT data set. Rice genes were considered to have ORFs if they corresponded to a predicted gene from the rice build 2 pseudomolecules. Sequences for predicted rice mudrA genes are listed in supplementary table 2 (Supplementary Material online). The TAIR sequence viewer tool was used to determine if Arabidopsis genes were expressed. To determine which rice genes were expressed, rice cDNAs were mapped to the build 2 rice pseudomolecules using BLAT (Kent 2002) with default parameters and the “fine” option. Similarly, rice expressed sequence tags (ESTs) were mapped using BLAT (Kent 2002) with trimmed trailing poly-A and leading poly-T tails, 3-kbp maximum intron size, and 90% overall identity (matches/lengths), rejecting ESTs with alignments to more than one location with less than 3% identity difference. cDNAs and ESTs that overlapped predicted mudrA genes by ≥ 100 bp were considered to be associated with them.

Clustering and Phylogenetic Analyses

Interrelationships among TDMs and TAMs were examined through clustering and phylogenetic analyses based on the conserved MULE-specific domain, MUDR (Hershberger et al. 1995; Marchler-Bauer et al. 2003). To identify MUDR domains in the mined mudrA genes, the mudrA genes were used as queries for TBlastN (Altschul et al. 1997) and BlastP (Altschul et al. 1997) searches against the 21 diverse MUDR domains listed in the NCBI Entrez (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed) entry for pfam03108. Where both protein and nucleotide sequences were available, the BlastP-identified MUDR domain was retained, though it was manually edited if a predicted intron eliminated a large portion of the domain. More divergent mined mudrA genes were used as Psi-Blast (Altschul et al. 1997) queries in order to identify regions within them that aligned with the MUDR domain of conserved mudrA genes. See supplementary table 1 (Supplementary Material online) for a list of detected MUDR domain regions.

For the clustering analysis, MUDR regions were iteratively clustered at every 5% change in identity using BlastCLUST (Altschul et al. 1997) with default parameters (fig. 1 and supplementary table 3, Supplementary Material online).

For the phylogenetic analysis, ClustalW (Higgins, Thompson, and Gibson 1994) was used to generate a multiple alignment of the MUDR domain regions, omitting truncated sequences less than 100 aa residues in length. In addition, the sequences were edited so that gaps in the alignment caused by less than 1% of the sequences were eliminated. A neighbor-joining tree with 100 bootstrap repetitions was generated from this alignment using PHYLIP (Retief 2000; Felsenstein 2004) with default parameters (supplementary fig. 1 [Supplementary Material online] and fig. 2).

Phylogenetic and Syntenic Relationships Within the MUSTANG (MUG) Gene Family

MUG1 genes in other plant species were identified by using rice and Arabidopsis MUG1 as queries for TBlastN (Altschul et al. 1997) searches against the non-redundant NCBI (http://www.ncbi.nlm.nih.gov/BLAST), JGI (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html), and Plant Genome Database (http://www.plantgdb.org) databases. A multiple alignment of the amino acid sequences of fungal MULE hop78 (Chalvet et al. 2003), TAMs At2g07100 and At1g12720, and the MUG family proteins was generated using ClustalW (Higgins, Thompson, and Gibson 1994). The alignment was edited so that the ends corresponded to those of MUG1. A maximum parsimony tree with 100 bootstrap replications was constructed from this alignment using PHYLIP (Retief 2000; Felsenstein 2004) with default parameters (fig. 3).

Predicted proteins encoded by the 50-kbp regions flanking the MUG1 genes of Arabidopsis, rice, and poplar were identified from the TAIR (http://www.arabidopsis.org), TIGR (http://www.tigr.org), and JGI (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html) data sets, respectively. Predicted proteins encoded by the Medicago contig (GI:50284582) were determined using the GENSCAN web server with Arabidopsis trained settings (Burge and Karlin 1997). These predicted genes were used for an unfiltered all-against-all BlastP (Altschul et al. 1997) search with an effective database length of 740.307.180 in order to uncover microsynteny (fig. 4, supplementary fig. 2, and supplementary table 4, Supplementary Material online).

Nonsynonymous to Synonymous Substitution Rate Ratio Estimates

A multiple alignment of MUG1 orthologs was generated using ClustalW (Higgins, Thompson, and Gibson
1994). Corresponding nucleotide sequences were manually aligned using the amino acid alignment as a guide. After eliminating ambiguity sites, the CODEML program in the PAML package (Yang 1997) was used to estimate the nonsynonymous to synonymous substitution rate ratio ($d_{N}/d_{S}$) for the provided nucleotide sequence alignment and maximum parsimony tree (constructed using the PHYLIP [Retief 2000; Felsenstein 2004] package with default parameters). $d_{N}/d_{S}$ values for the N-terminal, MUDR domain, central, zinc finger domain, and C-terminal subsections of MUG1 were similarly estimated.

### Results and Discussion

Transposon-Dissociated mudrA-like Genes in Rice and Arabidopsis

Our search for sequences with similarity to mudrA that were not associated with transposon structures revealed 45 such TDMs in Arabidopsis and 76 in rice (table 1). Approximately nine-tenths of these have predicted ORFs (table 1). Strikingly, 40% (18/45) and 37% (28/76) of TDMs in Arabidopsis and rice, respectively, match ESTs and/or cDNAs as compared to only 1% (2/143) and 6% (32/498) of Arabidopsis and rice TAMs (table 1), hinting that the expression profile of TDMs is more similar to that of host genes than to that of mobile elements.

#### Taxonomic Distribution of TDMs and TAMs

Clustering analysis, based on the MUDR domains of the mined TAMs and TDMs, found that most of these sequences clustered exclusively with other MUDR domains from the same species until the identity threshold was $<40\%$ (fig. 1). A neighbor-joining tree also indicated that the MUDR domains form mainly species-specific clades (fig. 2). This suggests that many TDMs have evolved from MULEs since the monocot-dicot divergence 140–150 MYA (Chaw et al. 2004). Nonetheless, two exceptional groups with both monocot and dicot members were identified, and both of these involve TDMs (figs. 1 and 2 and supplementary fig. 1, Supplementary Material online).

The first of these exceptional groups includes members of the FAR1 gene family, including two TDMs, FAR1 and PHY3, that have proven host functions in far-red light response (Hudson, Lisch, and Quail 2003). The second group, a previously unreported family henceforth referred to as the MUSTANG (MUG) gene family, is composed, with one exception, entirely of TDMs (supplementary fig. 1 and supplementary table 3, Supplementary Material online). However, while the neighbor-joining tree indicates that a TAM gene, At2g07100, is a part of the MUG gene family, the clustering analysis places it with other Arabidopsis TAMs. To resolve this discrepancy as well as to gain greater resolution of the internal branches of the MUG gene family, a maximum parsimony tree was constructed using the entire protein sequences (fig. 3).

The maximum parsimony tree places At2g07100 with other TAMs, rather than inside the MUG gene family, with high bootstrap support of 97% (fig. 3). The shorter sequences used to construct the original neighbor-joining tree, as well as the fact that the predicted ORF of At2g07100 has accumulated several stop codons, may account for the differing placement of At2g07100 and lower associated bootstrap values in this tree (supplementary fig. 1, Supplementary Material online). The maximum parsimony tree divides the remaining TDMs into two primary clades, both of which include monocot and dicot members. This indicates that there were several members of the MUG gene family before the divergence of the monocot and dicot lineages.

**MUSTANG1** is a Novel, Taxonomically Widespread Gene

Interestingly, a pair of rice/Arabidopsis MUG genes are not only highly similar to each other at the amino acid level...
The maximum parsimony tree supports the hypothesis that these genes are orthologs, which we have designated MUG1 (fig. 3). The tree also indicates that the two poplar MUG1 genes probably arose through a duplication event sometime after the Medicago and poplar lineages diverged. The duplication is unlikely to be the result of transposition as the ~20-kbp regions flanking the poplar genes share significant nucleotide-level similarity (BL2SEQ E < 10^{-120}, identity >78%). The predicted genes in the 50-kbp flanking regions of the rice and Arabidopsis MUG1 genes are similar to the predicted genes in the poplar flanking regions (fig. 4 and supplementary fig. 2, Supplementary Material online). Even the flanking regions of the Medicago MUG1 gene, which is located on an unordered contig, display microsynteny with the flanking regions of the other plant species (fig. 4 and supplementary fig. 2, Supplementary Material online). It is impossible to determine whether the maize MUG1-like gene is located in a syntenic region as the available sequence is only 3,547 bp in length. However, the synteny and phylogenetic relationships between the other MUG1 genes support their orthology and eliminate the possibility that the distribution of MUG1 genes arose through horizontal transfer.

MUG1 Functionality

The rice, Arabidopsis, maize, and one of the poplar MUG1 orthologs are associated with cDNAs, ESTs, or both

![Fig. 2.—Relationships between monocot and dicot mudrA genes as determined by phylogenetic analysis, depicted as a majority rule neighbor-joining tree rooted at MULE hop78 (Chalvet et al. 2003). Monocot leaves are labeled in gray and dicot in black.](https://academic.oup.com/mbe/article-abstract/22/10/2084/1138507)

![Fig. 3.—Maximum parsimony tree of the MUG gene family. This majority rule consensus tree is rooted with fungal MULE hop78 (Chalvet et al. 2003). Bootstrap values are shown at nodes. All genes shown are TDMs and belong to the MUG gene family except for hop78 and At2g07100 and At1g12720 which are TAMs. Genes indicated by one or two asterisks are associated with ESTs or cDNAs, respectively. Oryza sativa ssp. japonica gene names are shown beginning with “R.” O. sativa ssp. indica with “Ri,” Arabidopsis thaliana with “A,” Zea mays with “Z,” Medicago truncatula with “M,” and Populus trichocarpa with “P.”](https://academic.oup.com/mbe/article-abstract/22/10/2084/1138507)

![Fig. 4.—Syntenic 100-kbp region surrounding MUG1 genes. The MUG1 genes are depicted in red. Genes shown in the same color share protein-level similarity according to an all-against-all BlastP search. See supplementary fig. 2 and table 4 (Supplementary Material online) for a color version of this figure and E values, respectively. Vertical lines indicate gaps in the sequence. The Medicago sequence represents an unordered contig. Poplar I and II refer, respectively, to the molecules scaffold_201 containing the MUG1 gene eugene3.02010019 and LG_XIII containing the MUG1 gene grail3.0016035401. Note that the diagram is not to scale.](https://academic.oup.com/mbe/article-abstract/22/10/2084/1138507)
Like FAR1 and FHY3, MUG1 does have a highly conserved MUG1 also functions in far-red light response. However, or has until recently been, maintained by purifying selection d expressed gene is performing a host function, we calculated a coding functionality (Hurst 2002). The dN/dS of 0.0683 Spearman stem, and seedling leaves. As an indication of whether this for the six identified orthologs. A S of significantly less than 1 that a coding sequence is being, or has until recently been, maintained by purifying selection and is thus a good indication of whether a gene has protein coding functionality (Hurst 2002). The dN/dS of 0.0683 for the MUG1 genes is significantly less than 1 (P < 0.001; log likelihood ratio test).

Like the MUG1 genes, the transcription factors FAR1 and FHY3, which function as regulators of far-red light response (Hudson, Lisch, and Quail 2003), have also evolved from a MULE transposase to perform a host function. Given how dissimilar MUG1 is to these proteins (BL2SEQ E > 0.17, identity <24%), it seems unlikely that MUG1 also functions in far-red light response. However, like FAR1 and FHY3, MUG1 does have a highly conserved zinc finger domain. In fact, with a dN/dS of 0.0072, the zinc finger domain appears to be the most highly conserved region of the MUG1 gene. Therefore, as with FAR1, MUG1 may bind DNA in a manner that evolved directly from the way in which MURA transposase binds to TIRs (Hudson, Lisch, and Quail 2003). This would be consistent with other transposon-derived host genes whose mode of action is reminiscent of that of the transposon gene they evolved from. For instance, the vertebrate immune system proteins RAG1 and RAG2 assist V(D)J recombination of antibody genes by binding to and nicking DNA in a manner highly similar to DNA transposases, from which they are likely derived (Agrawal, Eastman, and Schatz 1998).

In addition, the co-opted human endogenous retroviral envelope (env) protein syncytin mediates trophoblast fusion during placental development (Chang et al. 2004), while env proteins in general are thought to promote membrane fusion for viral entry into cells (Colman and Lawrence 2003).

Conclusions

The rarity of taxonomically widespread MULE-dissociated mdrA genes is such that they appear to represent an unusual detour rather than a common mechanism for generating novel proteins. These results, however, do not exclude the possibility that more recently domesticated mdrA genes may be common among individual monocot and dicot lineages not investigated here. The six rice and four Arabidopsis expressed TDMs that do not belong to either the FAR1 or MUG gene families may be representative of such cases. Moreover, TDMs may frequently be “born” but “die out” soon after. One reason for this could be that many domesticated genes may benefit their host under specific environmental conditions. Another possible explanation is that newly domesticated genes, which have not yet lost their mobility-enabling structures, may transpose, thus losing the cis-regulatory factors required for their host function and thereby reverting to selfish DNA. In fact, in evolutionary terms, it is likely that such sequences have been more successful as transposons than as host genes.

Supplementary Material

Supplementary figure 1 and tables 1–4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Supplementary Fig. 1—Neighbor-joining tree of the mdrA superfamily. The tree depicted is a majority rule consensus tree constructed from an alignment of the MUDR domain regions of mdrA-like genes. The tree is rooted with fungal MULE hop78 (Chalvet et al. 2003). Bootstrap values are shown at nodes. Oryza sativa ssp. japonica gene names are shown beginning with “R,” O. sativa ssp. indica with “Ri,” Arabidopsis thaliana with “A,” Zea mays with “Z,” Medicago truncatula with “M,” and Populus trichocarpa with “P.”

Supplementary Fig. 2—Color version of fig. 1. Genes shown in the same color share protein-level similarity according to an all-against-all BlastP search. See supplementary table 4 for E values.

Acknowledgments

We would like to thank Nikoleta Juretic and Ken Hastings for their helpful comments on the manuscript. This study was supported by Discovery grants from the National Science and Engineering Research Council (NSERC) of Canada to T.E.B. and D.J.S. and a grant from the McGill University William Dawson Scholar Chair to T.E.B.

Literature Cited


Dan Graur, Associate Editor

Accepted June 21, 2005