Transposable elements, gene creation and genome rearrangement in flowering plants
Jeffrey L Bennetzen

Plant genome structure is largely derived from the differing specificities, abundances and activities of transposable elements. Recent studies indicate that both the amplification and the removal of transposons are rapid processes in plants, accounting for the general lack of intergenic homology between species that last shared a common ancestor more than 10 million years ago. Two newly discovered transposon varieties, Helitrons and Pack-MULEs, acquire and fuse fragments of plant genes, creating the raw material for the evolution of new genes and new genetic functions. Many of these recently assembled, chimeric gene-candidates are expressed, suggesting that some might escape epigenetic silencing and mutational decay, but a proven case of gene creation by any transposable element activity in plants remains to be demonstrated.

Introduction
The use of DNA markers to generate comparable genetic maps [1–3] led to the suggestion that the nuclear genomes of flowering plants, the angiosperms, were highly similar in the number of genes they contained and in their colinear order on the chromosomes [4]. This apparent genomic colinearity was observed despite more than 800-fold variation in genome size across the angiosperms [5]. Subsequent investigations of local genome structure in plants, primarily by the sequencing of nuclear DNA inserts within bacterial artificial chromosome (BAC) vectors, indicated that most variation in the size of nuclear genomes was caused by the differential amplification and/or retention of retrotransposons that contain long terminal repeats (LTRs) [6,7]. In large genome species such as maize, barley or wheat, genes were commonly found in small islands surrounded by seas of LTR-retrotransposons [6,8,9]. At this scale of comparison, orthologous genes often exhibited extensive colinearity, otherwise known as microcolinearity. However, the transposable elements and other sequences within the intergenic regions were usually completely different, even in comparisons between species such as sorghum and maize or wheat and barley that had last shared common ancestors less than 15 million years ago (mya) [10,11]. In addition, a few non-colinear genes were found in most orthologous genome segments from different plant species, including in those instances in which very closely related species such as maize and sorghum or Arabidopsis thaliana and Capsella rubella were compared [10,12]. Most of these comparative analyses were conducted in the cereals, especially maize, sorghum and rice. Even with conservative criteria for gene identification, approximately 35% of genes appeared to have moved to new locations in the ~12 million years since maize and sorghum diverged from a common ancestor [13,14]. This instability contrasts dramatically with the higher degree of conservation in gene order and content seen in mammals, for instance between mouse and human lineages over ~80 million years of independent evolutionary descent [15].

Most recently, exceptional frequencies of intraspecific gene rearrangement were described within rice and maize [16–19,20]. The conclusions that many rice genes [16,17] and >30% of maize genes [20] were non-colinear even within the species was, seemingly, largely incompatible with the observation that any colinearity could be conserved in more distant comparisons. However, further analyses indicated that most or all of the non-colinear rice genes were transposable elements that had been misannotated as genes [17,21,22], and that most of the non-colinear sequences that had been annotated as genes in maize were actually gene fragments [20]. In this review, I describe recent discoveries that explain the origin of the non-colinearities observed in the genomes of flowering plants. These observations suggest aggressive processes for gene creation in the angiosperms, which compete with persistent mechanisms for gene removal and genome shrinkage.

Adding and removing DNA sequences
As with all other eukaryotes, plants have increased the DNA content of their nuclear genomes by polyploidy, segmental duplication and transposon amplification. For reasons unknown, the rates of at least some of these processes are exceptionally high in plants. Essentially 100% of flowering plants are current polyploids or can
be traced to one or more ‘paleopolyploid’ (see Glossary) events within the past 200 million years [23]. Comparative genetic maps indicate segmental duplications, some comprising whole chromosomes or chromosome arms [24], and BAC sequence analyses indicate a wealth of tandem gene families. In all angiosperms with haploid genomes larger than 2000 Mb that have been investigated to date, more than 50% of the nuclear DNA has been found to consist of LTR-retrotransposons and other repeats [25]. Given that the mean 1C angiosperm genome size is about 5600 Mb (I J Leitch, unpublished), it can be concluded that the majority of plant DNA on the planet is composed of transposable elements. These great abundances and recent genome size contributions of transposable elements are also indicated by the fact that most are recent insertions. For instance, more than 80% of the intact LTR-retrotransposons in all analyzed angiosperms can be dated as insertions that occurred within the past five million years [25–28].

This exceptional rate of growth in the DNA content of plant genomes is in competition with a very high rate of sequence removal. Unequal homologous recombination, commonly by intra-chromatid events, can convert an intact LTR-retrotransposon into a solo-LTR (see Glossary) [6,8,21,27,28] or can remove the chromosomal sequences between two LTR-retrotransposons of the same family [27,28]. Most DNA sequence removal in *Arabidopsis*, rice and wheat, the only species investigated at this level, appears to be associated with tiny deletions that can be ascribed to illegitimate recombination [21*,27–29]. The precise mechanism(s) of illegitimate recombination are not known in these cases, but deletion associated with the repair of double-strand breaks [30] is a likely candidate.

Although the most common size of deletions in flowering plants — at least, as measured in the rice genome — is 1–2 bp [21*], these and larger deletions can lead to a very effective removal of all classes of DNA that are not retained by natural selection. For instance, the half-life of LTR-retrotransposon sequences in rice is less than three million years and can be associated with removal of >194 Mb of LTR-retrotransposon DNA in the past eight million years [21*,28]. Despite this, the genomes of the two rice subspecies *japonica* and *indica* have grown >2% over the past few hundred thousand years, primarily by a frequency of LTR-retrotransposon amplification that has outstripped the progressive removal of DNA by small deletions [21*].

A simple explanation for the great variation in angiosperm genome size is either that lineages will differ in the frequency of genome growth, for instance, as a result of rare polyploidy and/or episodic transposon bursts, or that they might differ in the qualitative and/or quantitative properties of DNA sequence removal (Figure 1). For instance, the repair of double-strand breaks in the small genome of *A. thaliana* is accompanied by fewer insertions and a larger average size of deletions than in the larger genome of *Nicotiana tabacum* [30]. It is also possible that the activities of these competing genome-growth and -shrinkage mechanisms could be significantly influenced by dramatic changes in the internal or external environment, as indicated by the activation of transposable elements by genome stress (e.g. chromosome breakage) or the induction of sequence removal processes by the establishment of a de novo polyploid state [31–33].

**Gene creation**

Most closely related organisms share the same types of genes, although gene copy numbers and regulation can vary over short times of evolutionary divergence. Even apparently ‘novel’ or ‘orphan’ genes can often be traced to extensive primary sequence divergence from a clear ancestral gene [34,35]. Hence, most truly novel genes were probably created, perhaps from raw genomic sequence, hundreds of millions of years ago. The more recent creation of chimeric genes with novel genetic functions has been proposed by the process of ‘exon shuffling’, wherein fragments of genes are fused together, partly relying on the fact that chimeric introns would often be processed to yield intact exons in a final mRNA product [36].

Recent discoveries in flowering plants suggest a very high rate of gene creation by transposon capture and exon shuffling [37**,**38**,**39**,**40**]. Early studies in maize had shown that transposons could acquire specific gene sequences and amplify them across the genome [41–43]. In the case of the *Bs1* LTR-retrotransposon of maize, a portion of a plasma membrane proton ATPase gene had been acquired — presumably at the RNA level, because all introns were missing — whereas subsequent divergence had selectively retained the transmembrane domains [41]. Although maize *Bs1* was the first LTR-retrotransposon observed, in any organism, to acquire a
gene fragment, numerous instances of the acquisition of genes by a closely related mobile DNA, the retroviruses of animals, have been described [44]. Mutator-like DNA elements (MULEs) had also been seen to acquire genic sequences in both maize and Arabidopsis [43,45]. However, Jiang and et al. [37**] found that MULE capture of gene fragments was not an interesting oddity but, rather, a major feature of rice genome structure. More than 3000 MULEs containing fragments of genes were found in rice genome sequence and were named ‘Pack-MULEs’. The small genomic fragments (averaging 325 bp, with a range of 47–986 bp) had been acquired at the DNA level, with both exons and introns present, but various rearrangements were sometimes identified within the acquired sequence. These rearrangements could have occurred either during or after the acquisition process. Gene fragments from different rice genes were found together in ~23% of Pack-MULEs, and at least 5% of Pack-MULEs were found to be expressed, as evidenced by full-length cDNAs with an identical DNA sequence match. More than 90% of these expressed Pack-MULEs appear to have been transcriptionally initiated within the element itself [37**]. Hence, by the criterion of expression at the RNA level, many of these Pack-MULEs are already new genes.

A mechanism for gene fragment acquisition by Pack-MULEs has not been proven, but ectopic gene-conversion across a nicked cruciform structure [46] has been proposed. Although MULEs are abundant components of most or all angiosperm genomes, the frequency of Pack-MULEs has not been calculated in any species other than rice [37**]. However, preliminary analysis of genomic sequence data from Lotus japonicum, a distant angiosperm relative of rice, indicates a similar abundance of Pack-MULEs (J Jiang and SR Wessler, unpublished), thereby suggesting that they will be major components of most or all flowering plant genomes.

Research in maize has shown the presence of numerous clusters of gene fragments, with all fragments in the same predicted transcriptional and translational orientation [47]. Lal et al. [48] found that one such cluster was, in fact, a Helitron that had recently inserted into the maize Sh2 gene to cause an inactivational mutation. Helitrons are a new class of eukaryotic transposable element, initially discovered through database analyses in A. thaliana, rice and Caenorhabditis elegans [49], and now found in insects, vertebrates and fungi as well [50,51]. Helitrons, unlike other transposons in eukaryotes, do not have terminal repeats and do not cause duplications of target site DNA, so their detection can be quite challenging. Helitrons often contain open reading frames (ORFs) that are predicted to encode a protein with replication initiator and helicase activities associated with rolling-circle replication of bacterial transposons [52], plus an RPA (replication protein A)-like protein that could also provide a singled-stranded DNA binding activity needed for DNA replication. Hence, rolling-circle replication seems a likely model for Helitron transposition [49]. The 3' ends of Helitrons contain a region that could form a 12–18 bp hairpin (with a 2–4 bp unpaired loop) that is followed 5–8 bp downstream by the consensus sequence CTRR at the end of the element, whereas the 5' end contains the terminal sequence TC. These conserved components might be the only cis sequences required for Helitron transposition. Helitron insertions are inserted within the target sequence AT, so that the 5' end is always A/TC and the 3' end is always CTRR/T. The mechanism of gene sequence acquisition by Helitrons is not known, although it appears that gene fragments are incorporated progressively at several possible locations, including the 5' end or near the 3' end.
Both Pack-MULEs and Helitrons appear to be abundant in many plant genomes, but the copy numbers of individual elements are quite low within any single genome. For instance, although there were >3000 Pack-MULEs found in rice [37**, the most individual elements had copy numbers of less than five. This is in stark contrast to the more abundant plant LTR-retrotransposons that can have copy numbers exceeding 10,000 per genome [7, 25]. This observation suggests that individual Pack-MULEs and Helitrons are quickly silenced [54] before they can amplify to high abundances within the nucleus. Given their structure and frequent expression, many of these elements could produce chimeric RNAs and fused peptides that would interact with all of those genes and gene products from which they have borrowed. Hence, it is likely that many chimeric Helitron or Pack-MULE RNAs will not only induce their own epigenetic regulation (e.g., silencing) but also contribute to the epigenetic regulation of the intact genes that have donated gene fragments to the element [40**]. If translated, chimeric Pack-MULE or Helitron peptides might also alter cellular enzymology and/or regulation, for instance by serving as dominant-negative inhibitors through the poisoning of multiprotein complexes. In cases in which these chimeric RNAs or proteins affect phenotype, selection for or against their expression will become a significant factor in their retention or removal, respectively.

If the recently assembled chimeric sequences inside Helitrons or Pack-MULEs do not provide a trait beneficial to the host plant, point mutations and indels (see Glossary) will eventually occur within cis-essential components of the elements, thereby making permanent an inactivation that might have been initially epigenetic. As with all other sequences in plant genomes, Pack-MULEs and Helitrons will be exposed to the persistent processes of sequence removal, which are primarily associated with the small deletions created by illegitimate recombination [21**, 27–29]. Hence, most potential chimeric genes created by Pack-MULEs or Helitrons will be lost within a few million years.

Those rare exceptions in which a chimeric gene survives over a long time period, and is thus shared in a conserved state by descendant genomes, will provide compelling evidence for a significant contribution of these transposons to gene creation and resultant biological diversification. The gene fragment inserted into \textit{Bs1}, and some (>10%) of the predicted chimeric genes within Pack-MULEs exhibit the DNA sequence characteristics associated with selection for a conserved protein function [37**, 41]; however, this selection could be for element function (e.g., more effective transposition) and not for any host biological process. Hence, convincing evidence for retained function at the sequence level in any chimeric gene would need to be manifested in Pack-MULE- or Helitron-derived genes that had lost their mobility (e.g., by terminal deletions). Equally convincing proof for the creation of a new gene by Pack-MULE- or Helitron-mediated exon-shuffling would be the identification of a mutant phenotype in the plant by a mutation (e.g., inactivation) in any element-derived chimeric gene. Neither of these forms of evidence has yet been described for any predicted new gene created by a Helitron or Pack-MULE.
MULE, at least partly because such statistical and experimental analyses have not been pursued to any comprehensive degree in any plant species.

Gene and genome rearrangement

The hyper-variability of the non-genic sequences that make up the majority of angiosperm genomic DNA is primarily because most of these are mobile sequences, and because plants very rapidly remove nuclear DNA that is not retained by natural selection (Figure 1). Some transposable elements also stimulate other types of genome rearrangement, including inversion, duplication or deletion of adjacent DNA, by chromosome-breaking, by aborted transposition, or by ectopic recombination between homologous transposable elements at different chromosomal locations. Hence, genome structure in any organism is, largely, the outcome of transposable element action and of the cellular processes that act on transposons.

In plants, most genes appear to retain similar or identical function despite their very different chromosomal environments in different plant species. For instance, the adh1 gene and one adjacent gene of unknown function were moved to a new chromosome in a common ancestor of maize and sorghum [13], and thus adh1 is found in a non-syntenic location when compared with its location in more distantly related grasses like rice or wheat. Despite this movement, the subsequent deletion of the adjacent mobilized gene, and the accumulation of surrounding seas of LTR-retrotransposons, the tissue specificity, induction profile and developmental timing of adh1 gene expression appear to be unchanged. This routine observation suggests that plant genes are very well insulated from their surrounding chromosomal environments. Hence, movement of an intact gene to a new chromosomal location will often yield a functional locus, thereby permitting its retention under natural selection.

How often do transposable elements mediate these gene movements? Although gross chromosomal rearrangements are commonly traced to the action of transposable elements in maize, for instance, the more common, single-gene rearrangements observed in comparisons of different plant species [13,14] are not yet associated with any proven molecular mechanism. Reciprocal tandem duplications and/or deletions, and small inversions, are likely to be caused by unequal recombination, but it is not clear whether gene movement from chromosome to chromosome often occurs by this same process in plants. Given the random loss of unselected DNA in plants, the sequence hallmarks for a transposable element vector in gene movement would be rapidly lost, and thus only observed in very recent gene-rearrangements [11].

Conclusions

The frequent acquisition of gene fragments by Pack-MULEs and Helitrons suggests a possible role for these elements in the redistribution of genes across the genome. However, in both element types, the fragments acquired are significantly smaller than most intact genes. It is possible that rare fragment-acquisitions will include a complete gene, but no such case has yet been found. Also significant, at an analytical level, is the fact that predicted chimeric gene fusion products within Pack-MULEs or Helitrons will be mis-annotated as genes in assessments of microcolinearity [14,16–19], thereby predicting less genetic colinearity than truly exists. Hence, reassessments of microcolinearity in plants are warranted now that the presence and properties of Pack-MULEs and Helitrons have been demonstrated.

Future studies will investigate the frequency with which plant mobile DNAs, especially Pack-MULEs and Helitrons, have contributed to the genetic repertoire and variable arrangement of angiosperm genomes. It is astounding that so much interesting structural novelty has been discovered in plants genomes, given that many fewer plant genomes have been subjected to comprehensive sequence analysis than have genomes from the prokaryotic, fungal and animal kingdoms. As more plant genomes are sequenced, more raw material for genome analysis will be generated, and a wealth of unexpected outcomes can be predicted.

Acknowledgements

The author thanks Ning Jiang and Susan Wessler for providing unpublished information, and Katrien Devos for assistance with the production of the manuscript. The preparation of this article was supported by National Science Foundation grant DBI031724.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

● of special interest

● of outstanding interest


21. The authors compared BAC sequences from orthologous regions in two maize inbreds, Mo17 and B73, covering more than 2.8 Mb of DNA on four chromosomes. The data indicated that more than one-third of the sequences annotated as genes were absent in one of the two inbreds at the locations sequenced. The authors found that these `missing genes’ were usually gene fragments found in clusters that were oriented in the same direction.


39. An insertion mutation in the maize barren stalk1 gene was found to be caused by the insertion of a Helitron, and the authors also found that Helitrons containing different gene fragments are common in the maize genome.


41. The authors find that the apparent genetic non-colinearity that they previously reported in the bz1 region is caused by gene fragments acquired by Helitrons. Shaded gene fragment that is in the opposite orientation to most gene segments acquired by these cells.


43. Comparison of over 20 000 gene models (see Glossary) across B73 and Mo17 orthologous regions predicted that more than 20% are not shared.
Inspection of nine such non-colinear annotated genes found that eight were actually gene fragments within Helitrons. This result suggests that the maize genome will contain a minimum of several thousand Helitrons carrying fragments of cellular genes. Some, but not all, Helitrons were found to produce transcripts that fused fragments from different genes. Insertions of other transposable elements were found in some Helitrons, including one LTR-retrotransposon insertion dated to 2 mya, suggesting that these elements have been active over a very long time period.


The authors describe a family of four closely related Helitrons. One element is found to be transcribed and produces a chimeric transcript that joins fragments from several predicted genes. This transcript was correctly spliced at most introns but employed alternative splicing at one normal intron and one chimeric intron. The structure of the elements in this family suggest how these non-autonomous Helitrons can accumulate both terminal and internal additional gene fragments by a serial acquisition process.