Theor Appl Genet (1995) 90:771–775

M. Nordborg · V. Walbot

Estimating allelic diversity generated by excision of different transposon types

Received: 13 January 1994 / Accepted: 18 October 1994

Abstract Methods are presented for calculating the number and type of different DNA sequences generated by base excision and insertion events at a given site in a known DNA sequence. We calculate, for example, that excision of the *Mu1* transposon from the *bz1::Mu1* allele of maize should generate more than 500,000 unique alleles given the extent of base deletion (up to 34 bases removed) and base insertion (0–5 bases) observed thus far in sequenced excision alleles. Analysis of this universe of potential alleles can, for example, be used to predict the frequency of creation of stop codons or repair-generated duplications. In general, knowledge of the distribution of alleles can be used to evaluate models of both excision and repair by determining whether particular events occur more frequently than expected. Such quantitative analysis complements the qualitative description provided by the DNA sequence of individual events. Similar methods can be used to evaluate the outcome of other cases of DNA breakage and repair such as programmed *V(D)J* recombination in immunoglobin genes.

Key words Transposable element · *Mutator* Recombination · Repair

Introduction

Transposable elements (*TE*) generate many novel patterns of gene expression when these mobile elements insert in and excise from a host gene (McClintock 1984). In general, *TE* insertion in the coding region lowers or eliminates gene function. Compensatory mechanisms such as the splicing of most of the *TE* insert from pre-mRNA can restore partial function without *TE* excision from the DNA template (Weil and Wessler 1990). *TE* insertions in gene regulatory regions can alter the timing, pattern and/or quantity of expression. Once inserted, *TEs* also become hot spots for secondary mutations. The abortive transposition of Class II (DNA) elements is common and can result in deletions of gene sequences at the *TE*:host sequence junction (Levy and Walbot 1991). Deletions within a *TE* can, for example, allow a transposon to function both as an intron for which splicing can restore gene expression and as a transposon for which excision will produce new alleles (Menssen et al. 1990; Raboy et al. 1989).

TE insertions also generally alter the host allele by creating host sequence duplications during insertion. When DNA-based elements excise from the host sequence, a precise excision – with regard to the *TE* – and precise repair of the double-strand gap will not restore the original gene sequence, because the host sequence duplication created on insertion will remain (Finnegan 1992). The lengths of these short duplications are characteristic of individual *TE* families; some values are shown in Table 1. Precise excision from exon insertion sites will result in amino acid addition if the duplication is a multiple of three but a frame-shift mutation for other duplications.

Table 1 Length of host sequence duplications created by insertion of Class II (DNA-based) transposable elements. The length of target site duplication is characteristic of *TE* families (reviewed in Blackman and Gelbart 1989; Moerman and Waterston 1989; Walbot 1992). The length of duplication is the same for both the non-autonomous and transposase-encoding element of each family

Element	Organism	Duplication (bp)
Ac	*Zea mays*	8
Em/Spm	*Zea mays*	3
hobo	*Drosophila melanogaster*	8
Mu9	*Zea mays*	9
P	*Drosophila melanogaster*	8
Tam1	*Antirrhinum majus*	3
Tam3	*Antirrhinum majus*	8
Tc1	*Caenorhabditis elegans*	2

Communicated by A. L. Kahler

M. Nordborg · V. Walbot (✉)
Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA

If precise excision and repair were the rule, each insertion event would generate only a single excision allele; this is not the case if excision is imprecise. In several studies of *P* elements in *Drosophila melanogaster* and *Tc1* in *Caenorhabditis elegans* it seems that excision of the *TE* is usually precise; however, the precision depends on the presence of a template for repair that also contains the *TE* (Engels et al. 1990; Plasterk 1991). With regard to the gene, much of the analysis of excision to date has focused on functionally revertant alleles, biasing the sample for particular types of events, i.e. those that restore the reading frame and have minimal impact on gene function (e.g. Engels et al. 1990 for *P* elements). In contrast to functional revertants after *P* element excision, unselected *P* element excision events contain much more extensive insertions and deletions (Takasu-Ishikawa et al. 1992).

In higher plants, the examination of excision events for various elements at numerous insertion sites has shown that imprecise excision events predominate (reviewed in Fedoroff 1989; Walbot 1992). Table 2 is a compilation of excision events from plant genes. Somatic events were recovered from somatically mutable tissue; the sample contained a mixture of alleles and was a less biased sample of allelic types than germinal excision events, most of which were selected for partial function. Although individual somatic events are readily analyzed at the sequence level following polymerase chain reaction (PCR) amplification and cloning, the number and extent of the data sets are limited.

Two conclusions can be drawn from these data. First, the magnitude of deletion and addition is greater for *TE* excision from insertion sites in regulatory regions than for sites in transcription units. This probably reflects the bias to select and recover partially functional germinal revertants; large deletions, for example, might be tolerated in promoters but destroy gene function if they occurred in an exon. If we restrict analysis to events within the coding region (Table 2A), a second trend emerges, namely that the magnitude of deletion and addition events appears to vary with element family. Sequences of multiple excision alleles exist for only a few insertion locations of *Mu1*, *Ds1*, *Tam21* and *Spm-18* as reported in Table 2, but individual revertant events have been sequenced from a larger number of *TE* insertion sites for these elements (summarized in Fedoroff 1989 and Walbot 1992). Based on all available data, it appears that host sequence changes are minimal for *Ds*, intermediate for *Tam2* and *En/Spm* and large for *Mu1* excisions. In fact, the range of *Ds* excision alleles from the promoter of *Bz-wm* is also very small, although only three events have been examined.

The present paper has two aims. First, we develop an exact method for calculating the number of excision/repair events as well as the resulting number of unique alleles that can be formed. We show that the number of alleles that can be formed is usually very large, and subsequently discuss the practical implications of this.

Second, we present some results from a computer program that enumerates all resulting alleles. Our main goal here was to illustrate the possible use of programs such as ours for the statistical testing of various models of how host sequence changes are generated by transposition. We wish to emphasize that the method is readily modified to account for similar processes. Excision of *TEs* can be analyzed as well as any process that involves the breakage and imperfect repair of sequences.

Methods and results

Allele number

Given the range in the size of deletion and addition events, we can use the available data to predict the allelic diversity to be expected from transposon excision events involving individual *TE* families. For the

Table 2 Deletion and excision events associated with imprecise *TE* excision events

A: Insertions in the coding region

Allele (Number sequenced)	*TE* inserted	Number of bases[a]		Reference
		Deleted	Added	
bz1::Mu1 (18)	*Mu1*	4–34	0–5	Britt and Walbot (1991)
bz-Mum9 (11)	*Mu1*	6–44	0–2	Doseff et al. (1991)
adh1-Fm335 (4)	*Ds1*	0–2	0–3	Fedoroff (1989)
niv-44 (2)	*Tam21*	0–35	0–1	Coen et al. (1989)
wx-m8 (9)	*Spm-18*	0–9	0–2	Schwarz-Sommer et al. (1985)

B: Insertions in the upstream regulatory region

Allele (Number sequenced)	*TE* inserted	Number of bases[a]		Reference
		Deleted	Added	
pal^{rec}-2 (7)	*Tam3*	4–100	7	Coen et al. (1989)
niv^{rec}-98 (3)	*Tam3*	0–3	0–199	Coen et al. (1989)
Bz-wm (3)	*Ds1*	0–2	0–2	Sullivan et al. (1989)

[a] The number of bases deleted is listed including both the host sequence duplication and flanking gene DNA, i.e. if insertion created a 9-bp host sequence duplication but a revertant allele retains only 5 of these, the deletion is 4 bp. Bases added refers only to non-templated bases found at the excision site. Some of these are inverted duplications that are proposed to be created as a consequence of normal repair processes, however, some base additions remain unexplained

purposes of this analysis, the insertion allele (that is the original allele plus the *TE* and the associated host sequence duplications) will be the initial allele. Deletions and additions to the host sequence following the *TE* excision will be expressed relative to this starting point (Fig. 1).

When calculating the total number of possible deletion events, we will distinguish between two scenarios. If m bases can be excised independently from *both* sides, for a maximum deletion length of $2m$ bases, there are $(m+1)^2$ possible deletion events. If, on the other hand, m is the maximum *total* number of bases that can be excised (and the deleted bases have to be apportioned between the two sides), there are

$$\sum_{i=0}^{m}(i+1)=\frac{(1+m)(2+m)}{2} \tag{1}$$

possible excision events. The number of insertion sequences of length 0 to n is,

$$\sum_{i=0}^{n}4^i=\frac{4^{n+1}-1}{3}, \tag{2}$$

which follows from the possibility of one of four bases at each position.

Now if all sequences thus formed were unique, we could obtain the number of alleles by multiplying Eqs. 2 and 1 [or $(m+1)^2$, if appropriate]. Because there are many ways of obtaining a particular sequence (deleting and adding an **A** is equivalent to doing nothing at all), this will only give us an upper bound to the number of possible alleles.

We would like to develop an exact expression for the number of alleles formed that takes all overlap into account. It is clear that the exact number will depend on the starting sequence in addition to m and n, and it would seem that calculating all the possible alleles would be difficult because of this. This is not the case. We will show that the dependence on the starting sequence is extremely weak and that it is possible to give an exact expression.

Consider the case of all possible events resulting from up to two total deletions (in the second sense mentioned above – two bases can be deleted from one side and zero from the other, or one from both sides, etc.) and up to two insertions. We let the initial sequence by AC-[transposon]-GT (Fig. 1). The first thing to realize is that alleles of lengths from $+2$ to -2 will be generated and that, of course, there can only be overlap (i.e. alleles can only be the same) within each length class. We can therefore count the unique alleles in each such class and add the results.

The class $+2$ consists of all alleles formed by no deletions and by insertions of two bases. There are $4^2=16$ alleles in this class, all of them unique. Next consider the class $+1$. It consists of alleles formed by the insertion of two and deletion of one, and the insertion of one and deletion of zero. All alleles formed by only one insertion will also be formed by the deletion of one and insertion of two, so we will only need to count the latter. In general, it is true that we will only need to concern ourselves with the alleles formed by the maximum number of insertions in each class.

There is still redundancy among these alleles, and the number of unique alleles can be calculated as follows. There are two types of these alleles, ANNGT and ACNNT, depending on whether the single deleted base was in the 5′ or 3′ flanking sequence. Because there are $4^2=16$ of each, we would have a total of 32 alleles, *if* there were no overlaps between the two groups. How many such overlaps are there? For alleles to overlap between these two groups, they must look like ACNGT, and the number of such alleles is the same as the number of alleles formed by no deletions and one insertion (and is equal to four). We can repeat exactly the same procedure for each of the classes $+1$ to -1.

The last class (-2) is different in that it consists of alleles formed by *no* insertions (and two deletions). It contains the alleles AC, CG, and AT, a total of three. Now consider what the effects of an extremely redundant initial sequence, AA-[transposon]-AA, would be. The three alleles in this last class would all be identical, for a total of one allele. *This last class, of maximal deletion, is the only one affected by the starting sequence.*

Define $f(i,j)=4^i(j+1)$ as the number of combinations resulting from i insertions and j deletions. We define $f(i,j)=0$ if either i or j is negative. Then, using exactly the same reasoning as above we obtain for the general case of a maximum deletion length of m, a maximum insertion length of n, and initial sequence $\mathscr{S}$ the total number of alleles $\mathscr{N}$ as

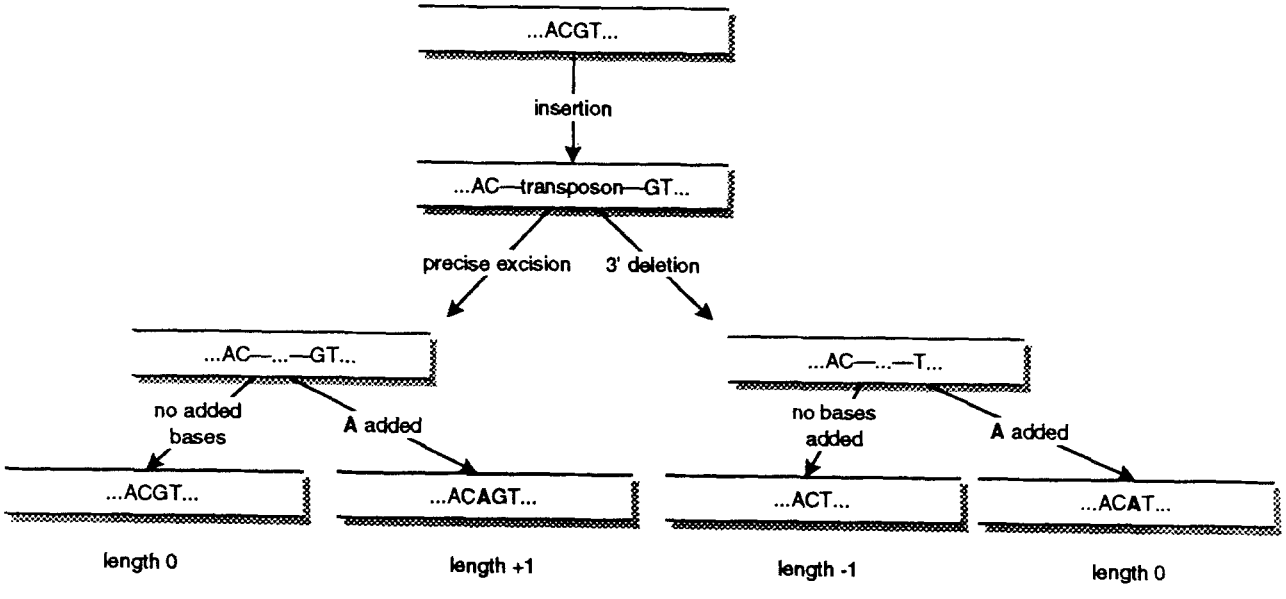
$$\mathscr{N}(m,n,\mathscr{S})=\sum_{k=-m}^{n}g(k,m,n,\mathscr{S}) \tag{3}$$

where g is the number of alleles in each class, calculated as

$$g(k,m,n,\mathscr{S})=\begin{cases}f(l,l-k)-f(l-1,l-k-1) & \text{if } k>-m\\ \xi(m,\mathscr{S}) & \text{if } k=-m\end{cases} \tag{4}$$

where l is the largest possible number of insertions given that the resulting length is to be k, and $\xi(m,\mathscr{S})$ is the number of alleles in the class with no insertions (which will equal $m+1$ if there is no re-

Fig. 1 Example of addition and deletion of bases following *TE* deletion



dundancy, and always be less otherwise). The effects of ignoring the initial sequence will always be less than m, which is clearly negligible for reasonably large values of m and n.

The same reasoning works when deletions happen independently on either side of the transposable element; because the class of maximum deletion will always contain exactly one allele, there is never any dependence on the initial sequence. The corresponding formulæ for this case are

$$\mathcal{N}(m,n) = \sum_{k=-2m}^{n} g(k,m,n) \tag{5}$$

where g as before is the number of alleles in each class, here calculated as

$$g(k,m,n) = \begin{cases} f(l,l-k) - f(l-1,l-k-1) & \text{if } k \geq 0 \\ f(l,2m-[l-k]) - f(l-1,2m-[l-k]-1] & \text{if } k < 0 \end{cases} \tag{6}$$

where l is found as in Eq. 4. Note that m here stands for the maximum number of bases that can be deleted from either side of the *TE*, so the maximum total number of bases that can be deleted in $2m$.

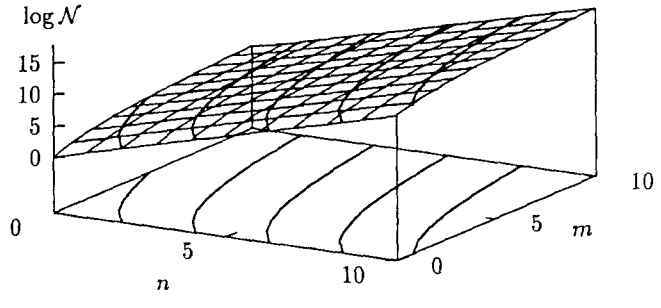
These expressions may look formidable, but they are actually quite simple to compute and are extremely well suited for a computer[1]. We will now discuss a few properties of functions 3 and 5.

First, the range of allelic diversity that can be created by different *TE* families is considerable. Two examples will illustrate this point. First we consider *Ds1*, for which deletions of between zero and two bases and insertions of up to three bases have been observed (Table 2). Given a non-redundant flanking sequence (e.g. AC-[transposon]-GT), this generates 389 unique alleles. The second example is *Mu1* in *bz1::Mu1*. When excised from the *Bz1* allele, with deletions and insertions as large as 34 and 5 bases, respectively, there are 501837 possible alleles, a truly enormous number. Both numbers were calculated using Eq. 3.

Given the universe of potential products from the *Ds1* excision, current surveys of products generated samples of 1–5% of the expected alleles per host gene insertion site. The situation is obviously worse for the *Mu1* excision in which less than 0.01% of the possible diversity has been recovered from any mutable allele. The paucity of data in combination with its bias for functional revertants precludes analysis of such basic questions as whether deletions of particular length are favored or whether specific bases are inserted more frequently.

Second, the resulting number of alleles is most strongly dependent on the number of insertions. Figure 2 shows how $\log \mathcal{N}$ (from Eq. 3) depends on m and n. It is clear that the influence of n is overwhelming and that (as expected) the dependence is approximately exponential. There is little need for approximations, however, as the exact allele number is easy to compute.

Fig. 2 The log number of alleles ($\mathcal{N}$) plotted as a function of m (bases deleted) and n (bases inserted)



[1] A simple *Mathematica* package that implements these expressions is available by anonymous ftp from "kimura.stanford.edu" or by e-mail request from "magnus@fisher.stanford.edu"

Allele frequencies

We also wrote a program[2] that generates all possible alleles. This can be used to examine specific questions such as how often an in-frame stop codon will be generated under the assumption of all events having the same probability. Some observations for the excision of *Mu1* from the *Bz1* allele in *bz1::Mu1* are given in Table 3. The frequency of any particular allele can be evaluated, i.e. the sequences most likely to occur in any length class can be enumerated. Similarly, the number of alleles with any motif can be calculated. An immediate application would be to examine what fraction of *P* element excisions are expected to be precise when the template for repair lacks a *P* insert, based on the magnitude of insertion and deletion events determined thus far. Such an examination could address whether precise excision events happen by chance alone.

The program can also search for specific allele types, such as those containing "insertions" of opposite strand sequence in a "flipped" order, and determine their expected frequencies. Such inverted duplication structures are proposed to result from the formation of hairpin structures in host DNA when the *TE* excises; a staggered double-strand break is produced, the *TE* excises and the "overhang" strand is ligated to its opposite strand. Subsequent strand nicking within the hairpin during chromosome repair could, in many cases, produce an inverted duplication (Coen et al. 1989; Takasu-Ishikawa et al. 1992). Excision alleles that fit this hairpin model have been recovered, but without knowing the frequency of insertion/deletion events that fortuitously yield such alleles, it is impossible to evaluate whether there is preferential recovery of this specific class. A similar problem concerns the evaluation of molecular models of *V(D)J* recombination in immunoglobulin genes where a limited number of sequences are available; thus, knowing the "universe" of possible alleles is important in assessing whether there is preferential recovery of a specific class (Gellert 1992). Recently Meier and Lewis (1993) utilized a large empirical data set to evaluate the significance of *P* nucleotides found as base insertions during *V(D)J* recombination. Calculating the product frequency distribution exactly for the various suggested models would greatly improve the reliability of the statistical conclusions.

Discussion

Diversity of alleles

The examples presented here illustrate that a staggering allelic diversity can be created by the excision of transposons from a particular site. Even for *Ac/Ds*, which creates a limited set of excision alleles, somatic excision during the life of the plant will create a chimera in which

Table 3 Number of in-frame stop codons generated by *Mu1* excision from *bz1::Mu1*

Total number of unique alleles:	501, 837
Preserving original reading frame:	173,992
Number of these with in-frame TGA:	4,587
Number of these with in-frame TAA:	4,331
Number of these with in-frame TAG:	6,302
Number of these without stop codons:	158,996

[2] This program, written in C++, is available from the same source as stated in Footnote 1

patches of cells are expressing many different forms of the gene. Similarly programmed somatic recombination as in *V(D)J* joining creates a spectrum of products that are subjected to several levels of selection during development of the immune system. The representation of individual somatic events will depend on the timing of appearance, mitotic success and the selective conditions.

The long-term genetic consequences of the allele diversity created by *TE* excision will depend on the frequency of germinal excision events, the number of alleles produced, the number of progeny, and in plants, the extent to which the product in question is required for gametophytic function. In the case of maize, in which 10^7 pollen grains are produced per plant, even a low germinal excision frequency of 10^{-4}, typical of *Mutator*, would provide for the production of 10^3 pollen grains transmitting excision alleles (*Mutator* properties are reviewed in Walbot 1992). The pollen grains would represent, however, only a small fraction of the expected diversity of the *Mu1* excision alleles from *bz1::Mu1*. Germinal excision frequency with *Ac/Ds* and *Spm(En)* is much higher, in the range of 10^{-1} to 10^{-2}. Because far fewer different alleles (10^2–10^3) are generated by the excision of these *TEs*, each allele type should be well-represented in the pollen of a single individual.

Transmission of excision alleles will also depend on the role of the affected gene during the gametophytic phase of the life cycle. As pollen (and embryo sacs) are haploid, gametophytes will be non-functional if they have a deficiency in any locus required for mitosis, cell growth, gamete differentiation, intermediary metabolism or fertilization. Thus, the alternation of generation in plants provides a mechanism for eliminating deleterious alleles in the gametophyte. As competition among pollen is often severe, even mildly defective alleles of many loci may be effectively eliminated prior to fertilization.

Testing excision and recombination models

The enormous number of possible alleles also has direct, practical implications when it comes to using observed data to test various models of transposition. Specific alleles are expected to be very rare, and sample sizes will have to be large in order to detect them. It is important to realize, however, that the required sample size is only increased for testing hypotheses that are directly affected by the size of the universe of events. For example, testing whether in-frame alleles appear at a higher than expected frequency does not require larger sample sizes just because the number of different in-frame alleles is large, whereas testing whether all alleles appear with the same probability does.

At the present time, there is insufficient data to test most models of DNA breakage and repair, but these data can be generated. As shown here, it is relatively straightforward to derive explicit expectations for simple null-models, and expectations for more complicated models can computed. It is our conviction that such expectations will prove useful in critically examining mechanistic explanations for excision and repair processes

Acknowledgements We thank Mike Cummings, Sally Otto, Michael Lieber and members of the Walbot lab for comments on the manuscript. This work was supported by a grant from the National Institutes of Health (NIH GM49681) to V.W.

References

Blackman RK, Gelbart WM (1989) The transposable element *hobo* of *Drosophila melanogaster*. In: Berg DE, Howe MM (eds) Mobile DNA. American Society of Microbiology, Washington, D.C., pp 523–529

Britt AB, Walbot V (1991) Germinal and somatic products of *Mu-1* excision from the *Bronze-1* gene of *Zea mays*. Mol Gen Genet 227:267–276

Coen ES, Robbins TP, Almeida J, Hudson A, Carpenter R (1989) Consequences and mechanism of transposition in *Antirrhinum majus*. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D.C., pp 413–436

Doseff A, Martienssen R, Sundaresan V (1991) Somatic excision of the *Mu1* transposable element of maize. Nucleic Acids Res 19:579–581

Engles WR, Johnson-Schlitz DM, Eggleston WB, Sved J (1990) High-frequency *P* element loss in *Drosophila* is homolog dependent. Cell 62:515–525

Fedoroff NV (1989) Maize transposable elements. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D.C., pp 375–411

Finnegan DJ (1992) Transposable elements. Curr Opinion Genet Dev 2:861–867

Gellert M (1992) Molecular analysis of *V(D)J* recombination. Annu Rev Genet 26:425–446

Levy AA, Walbot V (1991) Molecular analysis of the loss of somatic instability in the *bz2::mu1* allele of maize. Mol Gen Genet 229:147–151

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

Meier JT, Lewis SM (1993) *P* nucleotides in *V(D)J* recombination: a fine-structure analysis. Mol Cell Biol 13:1078–1092

Menssen A, Höhmann S, Martin W, Schnable PS, Peterson PA, Saedler H, Gierl A (1990) The En/Spm transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in the *A2* gene. EMBO J 9:3051–3057

Moerman DG, Waterston RH (1989) Mobile elements in *Caenorhabditis elegans* and other nematodes. In: Berg DE, Howe MM (eds) Mobile DNA. American Society of Microbiology, Washington, D.C., pp 537–556

Plasterk RH (1991) The origin of footprints of the *Tc1* transposon of *Caenorhabditis elegans*. EMBO J 10:1919–1925

Raboy V, Kim H.-Y, Schiefelbein JW, Nelson OE Jr (1989) Deletions in a *dSpm* insert in a maize *bronze-1* allele alter RNA processing and gene expression. Genetics 122:695–703

Schwarz-Sommer Z, Gierl A, Cuypers H, Peterson PA, Saedler H (1985) Plant transposable elements generate the DNA sequence diversity needed in evolution. EMBO J 4:591–597

Sullivan TD, Schiefelbein JW Jr, Nelson OE Jr (1989) Tissue-specific effects of maize *Bronze* gene promoter mutations induced by *Ds1* insertion and excision. Dev Genet 10:412–424

Takasu-Ishikawa E, Yoshihara M, Hotta Y (1992) Extra sequences found at *P* element excision sites in *Drosophila melanogaster*. Mol Gen Genet 232:17–23

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

Weil CF, Wessler SR (1990) The effects of plant transposable element insertion on transcription initiation and RNA processing. Annu Rev Plant Physiol Plant Mol Biol 41:527–552