Mega-change gene copies are produced by reverse transcription of precursor RNA with replaced introns.

Drew Schwartz
Indiana University, Bloomington, IN USA and Hebrew University of Jerusalem, Israel

Present address

This report describes the presence of copy genes in plants where the sequence similarity to their parental genes is almost completely limited to limited to the exons. Evidence is presented that indicates that they arose via a novel processing of transcribed precursor RNA during which the introns are replaced rather than excised, followed by reverse transcription and incorporation into the chromosomes. Intron replacement may play an important role in evolution as it results in megagene changes.

In the course of evolution new genes arise by polyploidy resulting in the duplication of the entire genome, or by duplication of chromosomal sectors by rearrangements such as translocations, inversions and other transpositions followed by chromosomal recombination. With the function carried out by one gene copy mutations could accumulate in the duplicate. New functional retrogenes (1,2) can result from reverse transcription of messenger RNA (mRNA) into complimentary DNA (cDNA) that is then incorporated into the genome. In this report evidence for another mechanism for production of gene copies is proposed that for the most part occurs only under conditions of stress when reverse transcription is most prevalent (3,4). The mechanism involves an alternate processing of the precursor RNA transcript such that instead of excision of the introns and the joining of the exons to produce mRNA, the introns are replaced by other sequences. The altered precursor RNA could undergo reverse transcription and incorporation of the DNA into the genome resulting in the formation of new gene sequences. These can be recognized as copies when the parental gene sequences are used as queries in BLASTs, as they would have sequences similar to the exons of the parental genes. The Maize ZmGDB-BLAST Service and NCBI discontinuous megablast were employed in this study.

Eleven loci of maize with identified exon/intron patterns, were examined in a computer search for second copy genes. The novel RNA processing was detected in six of these genes, using their entire sequence as the query in the ZmGDB BLAST. In each case almost all of the introns had been replaced by other non-similar DNA sequences. The similarity in sequence was almost entirely limited to the exons. For the most part, with few exceptions, the borders of the exons in both the parental and second copy genes are very close in position, within a few bases. These six maize genes are A1, Adh1, Ch1, Lg1, P1 and Sh1. The five genes for which such no copy was detected were Su1, Lem1, Bz1 and B1. The A1* copy gene was reported by Bernhard et al. (5), and following their nomenclature the copies will be identified by an asterisk. In the case of the P1 gene two copies, P1-A* and P1-B*, were detected that differ in the sequences that replaced the introns of the parental gene.

The preservation of exon sequence similarity does not ensure that these sequences will subsequently continue to serve as exons in the copy gene. The segments of the gene sequence that serve as exons and introns may be influenced by the replaced intron-like sequences and
hence the copy gene may acquire a function that is completely different from that of the parental function.

Six of the seven copy genes are located on chromosomes other than those of their parental genes. A1* moved from chromosome 3 to 8; Adh1* from 1 to 4; Ch1* from 1 to 5; Lg1* from 2 to 10. P1-A* and P1-B* moved from chromosome 1 to 3 and 9, respectively. The sources of most of the sequences that replaced the introns are not known. Every replacement intron in the seven copies was used as query in the searches of the entire maize genome using maize ZmGDB and NCBI BLASTs, and with only some short sequence exceptions, no hits were detected. Evidence to be presented suggests that the replacement sequences may be formed by the joining of very short sequences.

**Parental gene / copy gene comparisons**

The copy gene sequences with strong similarity to the parental gene exons will be referred to as exonlike, and the sequences that replace the introns will be referred to as c-introns. The accession numbers the parental genes and the Bacterial Artificial Chromosomes (BACs) in which the copy genes are located and their position are shown in Table 1.

**A1**

The exon/intron borders of the parental A1 gene (6) are at positions very close to those of the exonlike/c- intron borders in the A1* copy gene. There are four exons and thus six exon/intron borders. Changes in positions of exon-like/c-intron borders are designated as plus or minus (see fig.1). Those cases where the complete intron is not replaced and a portion of the parental intron bordering the exon-like sequence is retained, will be referred to as plus. Minus refers to those cases where the replacement extends into the the bordering exon-like sequence. Borders two, three and four are unchanged; border #1 = plus 5, #5 = plus 7 and #6 = plus 4. Alignments indicate no significant sequence similarities between the first two introns and c-intron replacement sequences. The third c-intron of A1* shows signs of incomplete replacement. Sequence alignments indicate that it contains a 36 base segment identical to a portion of the A1 third intron, except for two single and one double nucleotide polymorphism. BLAST searches using the replacement sequences as query did not turn up any other matches in the maize genome.

Bernhard et al. (5) havereported that the A1* copy gene as well as A1 are also present in teosinte but that only A1 exists in rice. The present study found that both are also present in sorghum.

**Adh1**

Adh1 (6) has 10 exons. The Adh1* copy gene underwent exon skipping (7) and has retained only six exon-like sequences. Sequences that are similar to the exons 3, 5, 6 and 8 are absent in Adh1*. The positions of the 10 borders of Adh1* relative to Adh1 are as follows: borders one, three, eight and ten are unchanged; border #2=plus 5, #4 = minus 43, #5 = plus 5, #6 = plus 8, #7 = minus 8, #9 = plus 4, and #10=plus 3. There is no similarity between the introns and c-introns that are of unknown origin with two exceptions. The 660 base second c-intron has a 23 base sequence with exact identity to a segment in chromosome #6 of maize. The third c-intron, 873 base long, contains a 335 base segment that is part of a mobile element distributed throughout the maize chromosomes.
Ch1*

Ch1 (8) and Ch1* have four exon and exon-like sequences. The positions of the six borders of Ch1* relative to Ch1 are: #1 = plus 7, #2 = plus 5, #3 = plus 10, #4 = minus 22, #5 = plus 3 and #6 = plus 7. There is no similarity in the first two intron and c-intron sequences. The 93 base third intron sequence of Ch1 is much smaller than the 327 base c-intron of which 268 bases are very similar to the sequence of a mobile element present on all of the 10 maize chromosomes.

Lg1*

Lg1 (9) and Lg1* have three exons and exon-like sequences. The four border positions of Lg1* relative to Lg1 are: #1 = plus 7, #2 = unchanged, #3 = plus 7 and #4 = plus 39. The 349 base third exon-like sequence of Lg1* is interrupted by the insertion of a 58 base sequence of an unknown source. There are five segments of the 1894 base Lg1 first intron that were not replaced. They total 510 bases and are in the same order in the two gene copies. Thus, the intron replacement may not result from a single event and may involve the accumulation of multiple short segments.

P1*A

P1 (10) and P1-A* have three exons and exon-like sequences and four borders. The first and third borders of P1 and the P1-A* copy gene are unchanged. Border #2 = plus 4, and border #4 = plus 4. The third exon of P1 is 744 bases in length but the match in P1-A* exon-like is limited to only the first 75 bases. The first P1-A* c-intron sequence has no similarity with the first P1 intron, nor with any other maize sequences. The 2184 base second P1-A* c-intron sequence shows only two short regions of similarity with other maize sequences.

P1*B

P1-B* has three exon-like sequences. The four borders of P1-B* relative to P1 are: #1 = plus 4, #2 = minus 3, #3 = plus 4 and #4 = plus 8. There is no sequence similarity between the first intron of P1 and that of its c-intron replacement in P1-B*, nor any similarity to other sequences in the maize genome. The 3932 base second c-intron sequence of P1-B* provides evidence for incomplete replacement as it has two short sequences, 51 and 83 bases long, that match closely to sequences in the second P1 intron, as well as a 188 base segment that matches a sequence in a mobile element that is widely distributed throughout the maize genome. The c-introns of P1-A* and P1-B* are completely different.

Sh1

Sh1(11) has 15 exons, but seven, numbers 1, 2, 5, 9, 13, 14 and 15 do not have matches in the Sh1* copy. The positions of the Sh1* borders relative to Sh1 are as follows: #5 1 and 2 (between exons 3 and 4) are minus 5 and minus 3 respectively, #s 3 and 4 (between exons 6 and 7) are minus 39 and plus 4, #s 5 and 6 (between exons 7 and 8) are plus 4 and minus 121, #s 7 and 8 (between exons 10 and 11) are unchanged, borders #9 and 10 (between exons 11
and 12) are minus 110 and unchanged. Intron 13 is 436 bases in length. A 134 base sequence in the center of this intron has been maintained in the Sh1* copy at its original position. The c-introns replacements are of unknown origin except for the 687 baselength second c-intron that is situated between exon-like sequences that match Sh1 exons four and six. It has four segments, 101, 171, 60 and 79 bases long, that are similar to sequences in the genome.

Discussion

The mechanism involved in the intron replacement is not known. The data suggests that it may occur piecemeal and that the replacement of individual introns may involve multiple short sequences from different sources that are then joined together. Note the second c-intron of P1-B*, the third c-intron of Ch1 the Adh1* third c-intron and the Sh1 second c-intron. In addition, the incomplete replacements observed in Lg1*, P1-B* and A1* support a piecemeal replacement mechanism. The intron replacement could possibly involve mRNA and result from insertions into the sites that were previously occupied by introns, as occurs with intron homing (12), first described in yeast. However, two observations render this unlikely: the similarity in lengths of many of the intron/c-intron pairs and incomplete replacements.

It is essential to effectively rule out the alternative possibility that the changes observed in the copy genes arise from mutations in the DNA of duplicated chromosomal genes that had been transferred to other sites in the genome. This is unlikely since only the parental gene would be required for full enzymatic activity in the organism, and the exon sequences of the copy genes would not be under selective pressure. Also the high rate of mutations in the introns required to eliminate similarity is unlikely in view of the infrequent mutations observed in the retained intron sequences of introns with incomplete replacement. Furthermore, it is highly unlikely that in every case the transposition would involve only that segment of the gene DNA extending from the start of the first exon to the end of the last exon, and hence the flanking sequences of the parental and copy genes should show some similarity. 2000 bases of DNA flanking the first and last exons of each pair of the parental and copy genes were compared and no similarity was observed except for three exceptions. In these three cases the sequences of the copy genes that match with the parental genes include short neighboring but not adjoining segments that terminate in what may be promoters as indicated by a TATA box sequence. The matching sequence of the fourth exon/exon-like terminates at position 3287 of A1. Beyond that point, the next 160 bases of A1 don’t match with A1* but the matching resumes for a 74 base stretch terminating with the possible promoter TATA box sequence of TATATAT. A similar situation exists with Lg1. The matching sequence of the third exon/exon-like terminates at position 6575. There is no similarity for the next 57 bases of Lg1, 6575 to 6633, but beyond this point the sequences of Lg1 and Lg1* match once again for 377 bases up to Lg1 base 7011, terminating at the possible promoter site indicated by the TATAATA sequence. P1-B* presents a similar picture in the flanking region of the first exon. There is a TATATTAT sequence at the start of a 164 base segment in P1-B* that matches with a segment of P1. The next 194 bases do not match, but the similarity in P1 and P1-B* sequences resumes at the first exon. Since RNA transcription is initiated at the promoter site, the finding that the matching sequences of these three copy and parental genes terminate at the TATA box strongly supports the proposed model for the precursor RNA origin of these copy genes with replacement introns.

Searches conducted in other organisms have shown that such gene copies with similar exons sequences and replaced introns are not limited to the grains and also occur in other plants species, such as Arabidopsis (ex. catalase, glutamine synthetase and alpha tubulin) and Mesotaenium caldarium cellulose synthetase. Intron replacement may have evolved only in the plant kingdom since it had not been detected in a search carried out in Drosophila and in the
mouse where alternative splicing is prevalent, but further searches in other organisms is required.

The alternate novel mechanism for new gene formation proposed in this report may play a major role in fast evolution as it could result in mega-changes in gene sequences. It is proposed to occur only or most frequently during periods of stress since it requires reverse transcription from RNA to DNA for incorporation into a chromosome.

References

Table 1

Positions and accession numbers of the maize BACs containing the copy gene Sequences.

<table>
<thead>
<tr>
<th>gene</th>
<th>accession #</th>
<th>number position</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1*</td>
<td>AC 198491</td>
<td>(49579-50898)</td>
</tr>
<tr>
<td>Adh1*</td>
<td>AC 213880</td>
<td>(20347-23197)</td>
</tr>
<tr>
<td>Ch1*</td>
<td>AC 214428</td>
<td>(66879-65519)</td>
</tr>
<tr>
<td>Lg1*</td>
<td>AC 195335</td>
<td>(157771-154907)</td>
</tr>
<tr>
<td>P1-A*</td>
<td>AC 182604</td>
<td>(83632 -81024)</td>
</tr>
<tr>
<td>P1-B*</td>
<td>AC 212695</td>
<td>(96266-100449)</td>
</tr>
<tr>
<td>Sh1*</td>
<td>AC194021</td>
<td>(175190-178550)</td>
</tr>
</tbody>
</table>
The positions indicated extend from the start of the first exon to the end of the last exon-like sequence.
Fig. 1

Comparisons of the three border positions in parental and copy gene pairs described in the text. The upper line of each pair represents the parental gene sequence. The exons and exon-like sequences are depicted as lines of 'equal' signs (=). The intron is shown as a line of single dashes (-->), and the c-intron by the line of dots. Plus refers to cases where the sequence match extends beyond the exon border into the intron. Retention of up to three of the intron bases will be counted as unchanged borders since they are also likely to occur by chance in the c-intron sequence. Minus refers to those cases where the exon/exon-like match does not extend all the way to the 3-exon border. A mismatch of a single exon base at the border will be considered as an unchanged border as it could also result from a single nucleotide polymorphism.