Plant Molecular Biology **15**: 437–448, 1990.

Nomad DNA – A model for movement and duplication of DNA sequences in plant genomes

Eran Pichersky
Biology Department, University of Michigan, Ann Arbor, MI 48109, USA

Received 20 April 1990; accepted in revised form 31 May 1990

Key words: chromosomes, DNA breaks, heteroduplex DNA, gene duplication, recombination

Introduction

Virtually all genes originate from other genes by a process of gene duplication. Since most genes encode proteins which either are essential for the survival of the cell or confer some advantage to it, duplication of a gene and subsequent divergence of one duplicate copy provide a way of keeping the old function and also acquiring a new one [38]. Thus, the mechanisms by which gene duplications arise are of great interest. Whereas a substantial body of work has accumulated concerning the mechanisms of gene duplication in animals [25], little information concerning gene duplications in plants, especially on the molecular level, is available. Plant genomes display a mode of evolution with some distinct features as compared to animal genome evolution (e.g., much lower level of conservation of synteny groups [53], 5–10 fold difference in amount of DNA even among congeneric species [52], high incidence of polyploidy [21]), and it is thus not a priori obvious that most gene duplications in plants occur by the same mechanism(s) by which the majority of gene duplications occur in animals. Although few molecular investigations of plant genes and genomes have specifically addressed the issue of gene duplications per se, a substantial amount of data derived from experiments from gene cloning, sequencing and mapping in plants has recently become available. These data are reviewed and analyzed here, and a new mechanism, the 'Nomad DNA' model, for gene movement and duplication in plants is proposed.

Models of gene duplications

At least four different mechanisms can be invoked to explain the observed gene duplications in animals and plants. These mechanisms are schematically diagrammed in Fig. 1. Different mechanisms lead to different results – e.g., linked vs. nonlinked duplicate genes, large chromosomal segments vs. short segments duplicated, etc. I first briefly describe these mechanisms and then discuss the evidence for them.

It is important to realize at the outset that most postulated mechanisms of gene duplication do not result in 'instant' duplications, that is, an additional copy is not produced within the same cell in which the initial steps leading to gene duplication have occurred. This is easily demonstrated by the mechanism of unequal crossing-over, which gives rise to tandemly linked duplications (mechanism # 1, Fig. 1A). In this case, the actual event of unequal crossing-over involving gene X changes only the distribution of the two copies of gene X, not the number of gene copies. Instead of one copy per homologous chromosome, after the crossing-over two copies are present on one chromosome and none on the other. During meiosis,

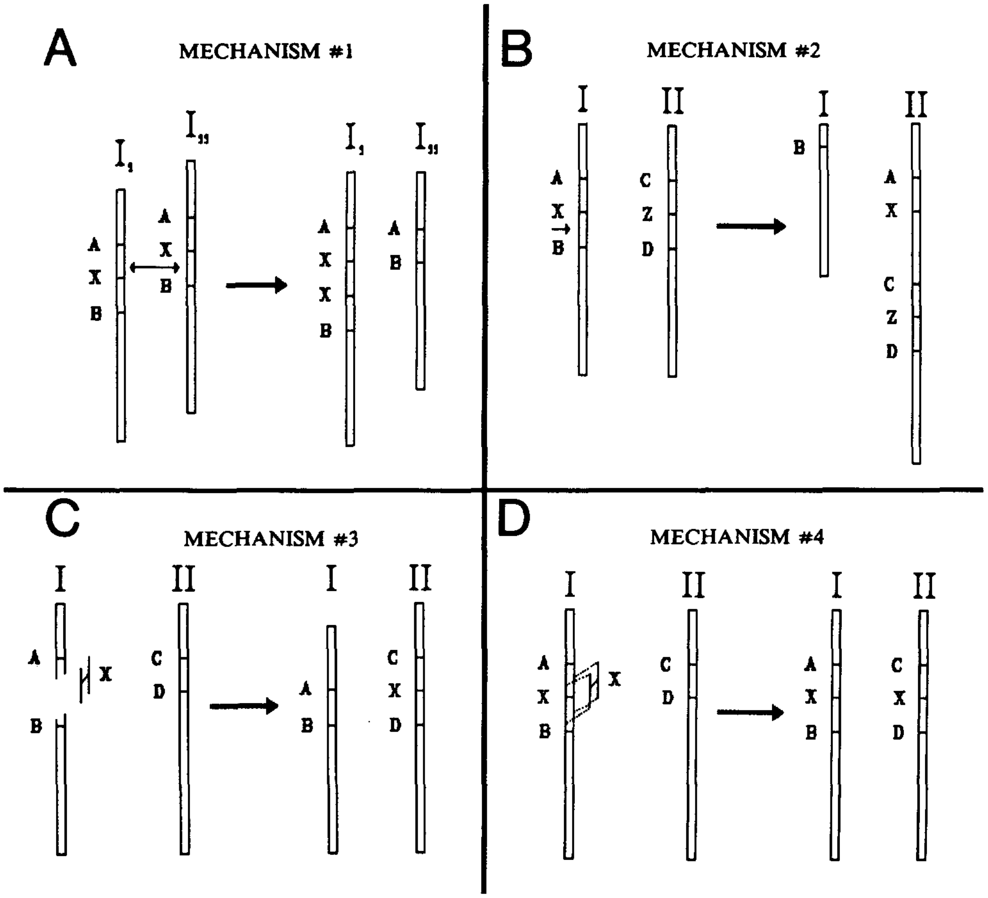

Fig. 1. Hypothetical mechanisms of gene duplications (see text).

each chromosome is segregated to a separate gamete. Following fertilization, a diploid cell may be produced with one homologous chromosome with the two copies of the X gene (tandemly-linked) and the homologous chromosome, derived from a meiosis which did not involve an unequal crossing-over, with a single X gene. In subsequent generations, a diploid cell may be produced with two homologous chromosomes each carrying two tandemly linked X genes. This result may be favored in plants because many of them have the ability to self-pollinate.

Another phenomenon which initially affects only the distribution of genes but eventually may lead to gene duplications is translocation between non-homologous chromosomes (mechanism #2, Fig. 1B) [19]. If the chromosome to which the DNA segment bearing gene X has been transferred is included in a gamete which fuses with another gamete having the normal chromosomal complement, an amphi-duplication results. In subsequent generations, through the appropriate crosses, an individual may be obtained which is homozygous for the presence of gene X on two different, non-homologous chromosomes. It should be noted that this mechanism and the previously described one could also lead, in a similar way, to deletions (which may or may not be lethal).

A third mechanism in which a gene duplication may be achieved is when a DNA fragment is removed from its internal location on a chromosome and is reintegrated at a different site, either on the same or another chromosome (mechanism

#3, Fig. 1C; see below for discussion of the details of this possible mechanism). Such an interstitial translocation is analogous to the movement of a transposable element in a conservative transposition event. Again, a gene duplication may result by the fusing of the appropriate gametes. Note that in this case, as well as in the previously described one, a broken chromosome could occur from the process and not be repaired, but this may not matter: a gamete containing a broken chromosome may not be viable but the other gametes, including the one with the extra DNA on one of its chromosomes, should not be affected.

A fourth model is analogous to the replicative transposition of transposable elements, in which a duplicated element is created and at the same time inserted at a different location in the genome (mechanism #4, Fig. 1D). Even in this case of 'instant' duplication, homozygosity of the duplication still requires additional generations. However, because plants do not have a germ line, a replicative transposition event in a somatic cell that eventually gives rise to the reproductive tissue in the flower may lead, upon selfing, to homozygosity for the duplication in one generation. It is very important to realize that this mechanism and the previous one may give the same results in the long run even if the initial steps are not identical. This follows because it is not necessary to maintain one copy in the original locus in order for duplication to occur; the gene in the original locus may be provided by the gamete from an individual in which no change in DNA (with respect to the gene in question) has occurred.

Experimental evidence for these models

Unequal crossing-over – Tandem duplication

The original event creating a tandemly repeated sequence by unequal crossing-over requires recombination outside the sequence in question. This is thought to occur at short sequences of homologous, repetitive DNA, or at stretches with fortuitous sequence similarity, during misalignment of the homologous chromosomes at meiosis [25]. The extent of misalignment determines the size of the segment which is made into tandem repeats. Once a tandem repeat which includes the gene itself is created, misalignment can bring about recombination between the first and the second blocks of homologous sequences, thus leading to further duplications. Because these homology blocks can be extensive, this event may be more frequent than the event creating the first pair of tandemly linked genes. Consistent with this postulate, plant loci which consist of tandemly linked genes generally contain more than two copies of the genes: for example, a CAB locus containing four genes has been identified in tomato [41], and a CAB locus with three tandemly linked genes is present in *Arabidopsis thaliana* [30]. An *A. thaliana* RBCS locus contains three tandemly linked genes [28], a tomato RBCS locus also contains three tandemly linked genes [49], and its counterpart in petunia contains six [10]. In some cases, the orientation of the genes may differ – there are head to head or tail-to-tail arrangements, such as in two CAB loci in tobacco, each with four genes [6], and in a CAB locus in tomato with three genes [41]. This can be explained by 'flipping' of sequences (due to recombination at inverted repeats) after the duplications occurred [25]. Clusters of genes have been found in many gene systems, although the orientation of the genes has not always been determined (e.g., actins [33]; histones [8], and additional examples reviewed in [52]). Clusters of genes are common in gene systems encoding abundant proteins (the most extensively studied plant gene systems) although gene dispersal is also found in these systems [11, 43].

Translocations and gene duplications

Plants seem to tolerate translocations much more easily than animals, and, as a consequence, translocations are common in plants [21]. Are they involved in duplications [4]? If they were, one would expect to observe long segments of duplicated DNA. In addition, duplicated sequences will tend to reside at the ends of chromosomes (terminal translocations are the most common type of translocation). In order to detect duplications of large chromosomal segments, it is necessary to examine a well-developed genetic map based on DNA markers. (Only when the DNA of two loci hybridize with each other, is it possible to determine that the loci compared are homologous; in plants, a complication is caused by the lateral transfer of genes via the genome of the photosynthetic endosymbiont [20, 55], but these genes have been separated more than a billion years from their homologous counterparts in the true nuclear lineage and thus have diverged to the extent that they do not usually cross-hybridize under conditions usually employed.) Such maps are now available for maize [22, 23], *A. thaliana* [7, 36], lettuce [29], *Brassica* sp. [32], rice [31], and tomato and several close relatives [3, 51, 53]. Examination of the available data does not support the notion that translocation is a major mechanism involved in duplications. In tomato, a single example has been discovered of several genes which are each found at two different sites (albeit on the same chromosome, chromosome 2) and are moreover arranged in the same order in both segments [52]. In maize, many syntenic groups are repeated twice, on two different chromosomes. However, whereas the tomato example may represent a true case of segmental duplication, perhaps by translocation, the situation in maize is more easily explained by ancient polyploidy followed by genome rearrangements [20, 22].

Other examples of segmental duplications are lacking. In addition, it is known that the linkage groups of tomato and pepper, two close relatives with the same chromosome number, are radically different (a minimum of 33 chromosomal breakpoints must be postulated to explain the differences in the synteny groups), indicating that either the tomato genome or the pepper genome, or both, have undergone substantial rearrangements [53]. Yet, essentially the same number of gene duplications is found in both species [53]. Moreover, of the few duplications which are unique to one or the other species (in each case the pair of duplicate genes are not linked to each other), no correlation is found with any of the chromosomal translocations or the segmental duplication [53].

Unlinked duplications

Most molecular investigations have studied genes encoding abundant proteins, and linked gene duplications have often been found in these systems. However, many gene duplications in plants were initially discovered by the application of Mendelian genetics to elucidate the inheritance of genes encoding homologous subunits of enzymes, since the different subunits could be distinguished from each other by electrophoresis on starch or acrylamide gels after activity staining, thus providing a scorable phenotype [20]. In recent years, Southern blots using cloned genes as probes, sometimes even without knowing the identity of the proteins they encoded, have shown that many such genes are found in duplicate and sometimes multiple copies. Whereas in the investigations involving enzyme-coding genes, genetic analysis was often carried out (since it was deemed necessary to show that homologous subunits are indeed paralogous, and not the products of the same structural gene, post-translationally modified), when working directly with DNA most investigators have not usually determined linkage relationships between hybridizing fragments. Nevertheless, the genetic information obtained in the investigations of duplicate genes encoding enzymes and from DNA coding unknown products (notably in tomato and maize) have demonstrated that the majority of gene duplications in plants are not linked.

In *Clarkia*, six of the duplications of genes encoding metabolic enzymes assort independently [19, 37, 40, 47], and linkage could not be determined for the remaining two [18, 40]. In tomato, of a total of approximately 100 cDNA clones randomly picked from a cDNA library and used as probes, more than 20 hybridize to multiple unlinked loci; of these, most hybridize to two unlinked loci, but some hybridize to three or four unlinked loci [1, 52]. The percentage of unlinked duplicate gene pairs in maize (62 loci out of 217 tested, 29%) is somewhat higher than that of tomato [23]. Although, as discussed above, a large portion of these duplications may simply reflect a polyploid origin of maize (and, consistent with this hypothesis, they show segmental duplications), 18 of the 62 duplicated loci could not be determined to be part of duplicated segments of chromosomes. Moreover, the high stringency used in the hybridization experiments meant that known unlinked duplications (alcohol dehydrogenases, for example) were not recognized [23], and therefore the conclusion that 29% of loci are duplicate must be a substantial underestimate of the total number in the maize genome. In *Clarkia* and in tomato, where phylogenetic analyses were carried out, few of the duplications of genes encoding isozymes are present in related plant families, and most are restricted to one or a few closely related genera. This observation suggests that most such duplications do not persist in the long run, since, if they did, the same duplications would have been shared by large groups of taxa. Similar analyses with DNA markers have not yet been carried out.

In *A. thaliana*, only two or three duplications (two independent maps were constructed, and one duplication in each map might be the same) have been described genetically (excluding gene families such as RBCS and CAB), and the duplicate pairs are not linked to each other. In lettuce, of the 34 cDNA probes tested, one indicated an unlinked duplication, two indicated gene clusters, and 31 indicated single-copy loci. The lettuce result may indicate that lettuce has a low level of unlinked duplicate genes relative to tomato and maize, but more likely it is due to the small number of cDNA clones used. Since the cDNA clones were chosen at random, the first few clones selected are likely to represent abundant mRNAs encoding abundant proteins; indeed, the linked hybridizing fragments reported in this study [29] appear to contain CAB and RBCS genes (E. Pichersky, unpublished). The construction of the *A. thaliana* map utilized genomic DNA and not cDNA clones, and it is therefore likely that *A. thaliana* indeed contains few duplications (unless probes revealing duplicate loci were deliberately excluded [C. Sommerville, pers. commun.]). In rice, although 22% of random genomic clones were defined as 'moderate repetitive', such sequences were deliberately excluded from mapping experiments, and thus the linkage relationships among the duplicate copies are not known [31].

Thus, if one makes the distinction between genes which are found in multiple copies in all plant genomes (i.e., gene families) vs. genes which are not, but are occasionally found in duplicates only in independent, relatively recently derived lineages, the data indicate that the duplicate genes of the latter kind are almost always unlinked to each other. It should be noted, however, that if genes are linked and they encode identical proteins, then the duplication will not be detected by genetic analysis of the isozymes (see an illustrative example of a linked duplication which escaped detection by genetic analysis in [5]; and compare with [44], in which a tandem duplication was detected by genetic analysis). Also, in cases where the smallest genomic fragment the cDNA clone hybridizes to is still large enough to contain more than one gene, a gene duplication may again escape detection. Thus, although a small number of tandem duplications may have been missed, it is nevertheless clear that a substantial number, and perhaps the majority, of duplicate gene pairs in plants are not linked to each other. This conclusion is supported by recent detailed analysis of gene duplications in tomato conducted in our lab, in which a number of pairs of duplicate loci were physically isolated and probed for multiple coding sequences; in each case, none were found to contain tandem copies (E. Pichersky, unpublished).

Is it possible that most gene duplications occurred by unequal crossing-over, resulting in linked genes, and the duplicate genes were later separated (perhaps via mechanisms #3 or #4)? If this is the case, then the subset of gene duplications deemed to be of recent origin (as determined, for example, from phylogenetic distribution) should exhibit a larger percentage of linked duplications than the entire data base of duplications. At present, this hypothesis cannot be tested because a systematic analysis of the time of origin of gene duplications based on their phylogenetic distribution has seldom been done. However, the observation that the majority of duplicate genes in plants are not linked to each other does contradict another prediction derived from the hypothesis that the majority of duplications occur by unequal crossing-over. If duplications originally occur by unequal crossing-over and are only later dispersed, any such dispersal may lead to additional duplications through the fusion of gametes with different chromosomes with respect to the position of the genes in question, as described above (one kind of gamete, the most common one, will contain the set of chromosomes including the one with the tandem duplication, the other rare gamete will include a set of chromosomes in which one copy of the gene has moved to another locus.) We would then expect to observe individuals with a tandem duplication *and* an additional copy elsewhere in the genome. This is indeed often the case in gene families [10, 43], indicating that when gene clusters preexist, this prediction is valid. However, as discussed above, this situation (i.e., a tandemly linked pair of genes and an additional copy elsewhere) is not often found in the duplications of genes which are not members of gene families, suggesting that in these cases a tandem duplication did not exist prior to the movement of the gene.

The notions that the movement of a gene from one locus to another is the sole requirement for the creation of an unlinked duplication in a sexually reproducing organism, and that this process of gene duplication is thus unrelated to whether the original locus contained two copies or just one, are somewhat counterintuitive; the assumption is, generally, that a tandem duplication is a necessary prelude to dispersal. However, as discussed previously, there is no absolute requirement for the copy of the gene in the original locus to be maintained in the gamete where the gene has inserted into a new locus, because the copy of the gene in the original locus will be provided in the other gamete. Thus, because of the finding that many gene duplications in plants are not linked, the true cause of these gene duplications must be sought in the mechanism of movement of DNA segments from one region in the genome to another. How do sequences in the plant genome move from one site to another?

Movement of DNA

Movement of DNA between organelles

Recently, it was shown that an intron of a nuclear gene in *Lycopersicon esculentum* (tomato) contains a 133 bp DNA fragment derived from the coding region of the chloroplast gene *psbG* (a second fragment derived from the first one is found elsewhere in the same intron, and will not be discussed further here; an intact copy of the *psbG* is still present in the chloroplast genome) [42]. The exact sequence of the 11 nucleotides at the 3′ end of the inserting chloroplast sequence is also found at the 5′ border of the insertion. The *psbG* segment is also found in the intron of the homologous gene in each of the *Lycopersicon* species examined but not in species from related genera (e.g. *Solanum*, *Petunia*, *Nicotiana*), suggesting that the original transposition event (chloroplast to nucleus) occurred since the divergence of *Lycopersicon* from other genera in the family Solanaceae, but before radiation of species in that genus. Additional chloroplast sequences integrated into the tomato nuclear genome (E. Pichersky, in preparation) also display similar direct repeats.

Features of the *psbG* and of the additional insertions do not match those of sequence insertion mechanisms involving transposable elements. Transposons, which are well-documented in plants, generate direct duplications of *target* DNA [12]. The *psbG* and the other insertions are flanked by direct repeats, but in these cases the copy of the sequence which appears duplicated is part of both donor and target sequences. In addition, transposons have inverted repeats at their ends; the chloroplast inserts do not.

A mechanism involving homologous recombination has been hypothesized to explain the inserting [42]. In this model, one end of the insertion element with the homology to the target site formed a heteroduplex with one staggered end of the target DNA, while at the other broken end of the target DNA, the recessed end was filled in and was then blunt-end ligated to the free end of the inserting element (see details below, and Figs. 2, 3). This model implies that the chloroplast DNA insertions occurred because linear DNA fragments from the chloroplast genome (released from a broken chloroplast?) were present in the nucleus, perhaps during meiosis or mitosis when the nuclear envelope breaks down and chromosomes contain many gaps.

A short DNA sequence found both on the mitochondrial and nuclear genomes of *Nicotiana* has recently been reported [2]. Examination of the published sequences reveals a similar situation to the *psbG* insertion: the five nucleotides at one end of the sequence present in both compartments is repeated on the other end of the mitochondrial sequence only, suggesting that a similar mechanism to that of the *psbG* insertion was responsible for the integration of this nuclear sequence into the mitochondrial chromosome.

Copies of chloroplast sequences have also been found in the mitochondrial genome of many plants [48]. Analysis of these sequences and the integration sites also failed to detect any similarities with transposable-element type insertion [14]. It was concluded that the only plausible mechanism must have been homologous recombination of very short sequences [14]. In support of this conclusion, some short direct repeats were found at the ends of one insertion examined in detail.

The 'Nomad DNA' hypothesis: a general mechanism of movement of linear DNA fragments in the genome

The 'traditional' view of the movement of DNA segments from one locus to another postulates an interaction between two non-homologous loci in which sequences are exchanged. This is thought to occur through homologous recombination at short repetitive sequences [25]. However, whereas a similar explanation is advanced for unequal crossing-over, where a single cross-over event is required, insertion of a sequence into the middle of a chromosome at a non-homologous locus will require two cross-over events on either side of the sequence. There is no evidence for this kind of orderly interaction between different chromosomes, or different loci of the same chromosome, in plant genomes (as determined, for example, from the frequency of gene conversion among unlinked duplicate genes [52]). Alternatively, it has been suggested that dispersal occurs as a two-step process – first, a sequence is excised and forms a circle through intralocus recombination at flanking repeats, then the circle integrates into another locus again through recombination at a repeat unit [25].

Both these mechanisms (double cross-over and the deletion-reintegration of circles) are based on the Holliday [24] and Meselson and Radding [34] models of recombination, which postulate single-stranded breaks and single-strand ex-

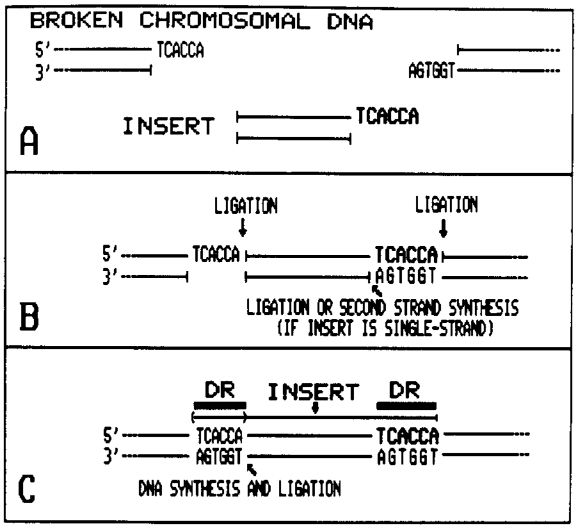

Fig. 2. A model for integration of a linear DNA molecule into a broken chromosomal site with 3′ single-strand overhang (see text for details). DR, direct repeat, denoted by heavy line. Thin line above segment indicates extent of the insert. Thin line in parentheses (——) indicates sequence which constitutes the 5′ DR; it has sequence identity with the 3′ end of the insert, but it is generated by the filling-in of the broken target DNA and is thus not derived from the insert DNA. The 6 bp sequence illustrated is purely artitrary, as is its length.

changes. However, recent results strongly support a recombination model involving double-stranded breaks, where each end point contains a large tract of single-stranded DNA [50]. Consistent with this latter observation and the evidence derived from analysis of sequences which have moved from one compartment in the plant to another, I suggest that many gene duplications in plants have their beginning in the integration of DNA fragments into double-stranded breaks. This model is illustrated in Figs. 2 and 3. The DNA fragments being plugged into their new locations may themselves be either double-stranded (the result of chromosomal breakage elsewhere in the genome) or single-stranded segments which 'peeled off' the DNA helix elsewhere due to single-stranded nicks [27]. A double-stranded insert which broke off from its original location will conform with Model #3 in Fig. 1, while a single-stranded 'peel-off' which leaves one strand in the original site will fit Model #4 (there is also evidence that single-strand DNA is quickly converted to double-strand DNA in plant cells [15]).

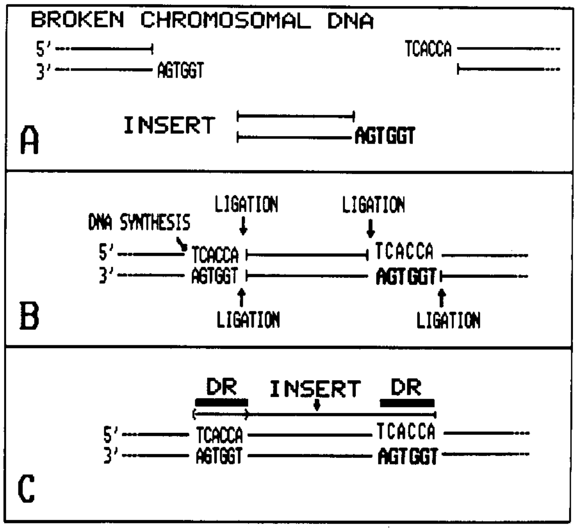

Fig. 3. A model for integration of linear DNA molecule into a broken chromosomal site with 5′ single-strand overhang (see text for details).

Based on the analyses of the integration junctures cited above, I propose a general mechanism in which insert DNA is fitted into double-stranded breaks because it forms a heteroduplex at one end due to fortuitous sequence similarities. Work with animal cells has shown that complementarity of even one to four nucleotides is enough to produce a ligatable junction, and in some cases no heteroduplex DNA is required at all to covalently link the ends [45, 39, 54]. In the model, I assume that ends which can form heteroduplex are preferentially ligated, but because it is less likely that the insert will have sequence similarity to the target DNA at both ends, it is assumed here that the other end is blunt-end ligated. Depending on the type of overlap, 3′ (Fig. 2) or 5′ (Fig. 3), fill-end and then ligation reactions occur (or vice versa). For example, if the target DNA has 3′ protruding end and the insert was single-stranded, the insert can be converted to double-strand DNA after one ligation reaction at the site of the heteroduplex formation (Fig. 2). In some instances, it appears that ligation is required to occur first between a blunt DNA end and a single-stranded DNA end (Fig. 2B), because the fill-in reaction should not be able to proceed for lack of a 5′ primer for the DNA polymerase enzyme. However, it has been demonstrated that eukaryotic cells contain enzymatic activity capable of converting such 5′ recessed ends to blunt ends without DNA degradation, possibly by using non-covalently bound DNA ends to serve as primers prior to ligation [39, 54]. Because a heteroduplex is formed only on one end, while the other end is filled-in and then ligated, the characteristic direct repeat, whose basic unit is originally present in both the insert and target DNA, is created (Figs. 2, 3). The observed 'direct repeat' in fact flanks the insert only on one end, while at the other end this sequence is actually part of the insert itself. It should be noted that if the non-heteroduplexed single-stranded protruding end of the target DNA is degraded, no direct repeats will be evident.

This model, which I call the 'Nomad DNA' mechanism, makes the prediction that shorter fragments are more likely to be successfully integrated, because in short fragments both ends are close to each other, and therefore given that one end has formed a heteroduplex with one end of the chromosomal breakpoints, the other is nearby too. (If the chromosomal break occurred during mitosis or meiosis, the chromosomal endpoints will be held in place next to each other by the protein scaffolding of the condensed chromosomes.) The chance of successful integration will decrease proportionally to the length of the inserting element also because longer fragments are likely to be enzymatically degraded. Movement of DNA through a circular intermediate or through double cross-over events will not show this predicted relationship between the size of the fragment and the frequency of movement. This model also suggests that fragments smaller than the minimum size of a functional gene will move around the genome much more frequently than gene-size fragments; of course, duplications of such fragments are not likely to be fixed in the population since it appears they confer no selective advantage. Thus, it is likely that by examining only the number of duplications of functional genes, the true frequency of duplications of DNA fragments in the plant genome is greatly underestimated. In support of the 'Nomad DNA' integration model, one study found a 2–3 fold enhancement of integration of exogenously added DNA in the plant nucleus compared with circular DNA [46], although other studies have failed to repeat this observation [26].

Pseudogenes

Some animal genomes contain large numbers of non-active copies of genes. These copies were derived by reverse transcription of mRNA and integration of the copy DNA into the chromosome, and are thus termed 'pseudogenes' (strictly speaking, a gene which had been functional but has recently been disabled by a mutation, such as deletion, insertion, or base substitution, is termed 'defective' or simply 'mutant' gene). Could the unlinked duplications found in plants be the result of pseudogene integration? Since pseudogenes are derived from mRNA, their distinctive marks are the lack of promoter and introns, and the presence of a poly(A) sequence at the 3′ end. Only a few reports of the isolation and characterization of plant pseudogenes have appeared [13]. We have examined the genomes of *Lycopersicon* spp. and related genera for the presence of pseudogenes by performing Southern blots using poly(A) sequence as a probe (Fig. 4). The results indicate that *Lycopersicon* species have only a few DNA fragments containing poly(A) tracts of 25 nucleotides or longer. In contrast, petunia and tobacco appear to contain many such fragments. It is not known whether these poly(A)-rich fragments indeed contain pseudogenes; on the other hand, it is possible to conclude that the *Lycopersicon* spp. contain at most only a few pseudogenes, so that the level of gene duplication seen in tomato is not a reflection of the presence of pseudogenes. In support of this conclusion, in all cases examined in tomato, both copies of duplicated genes contained introns (e.g.,[17, 49]).

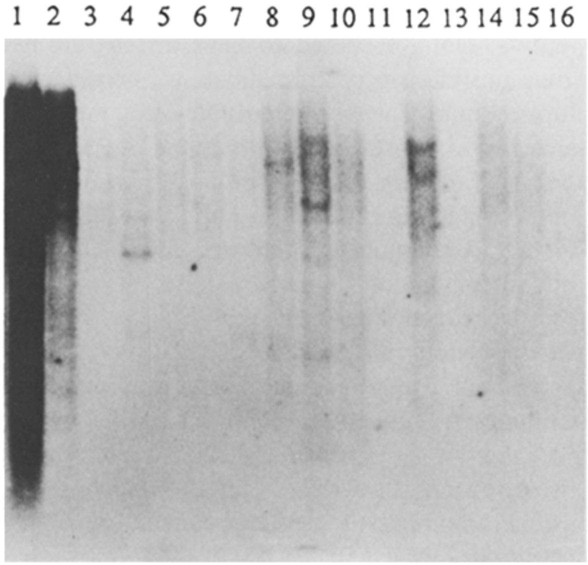

Fig. 4. Southern blots of DNA from species of the Solanaceae family, using poly(dA) sequence as a probe. Lane 1, *Nicotiana tabacum*; 2, *Petunia hybrida*; 3, *Datura meteloides*; 4, *Capsiccum annum* (pepper); 5 to 9, plants of the genus *Solanum*; 10 to 16, plants of the genus *Lycopersicon* (*L. esculentum* [tomato] is in lane 16).

Duplications and genome size

DNA content of plant genomes varies dramatically, even within a genus [52]. Thus, it is often stated that most of the difference is made of 'repetitive' DNA assumed to have arisen through some mechanism of 'amplification', or repeated duplications. If we assume the average size of a gene is 5 kb (a common estimate is 1–2 kb, but this does not take into account 1–3 kb for promoter and 3′ regions and 1–2 kb for introns), the 7×10^4 kb genome of *A. thaliana* [35] can accommodate no more than 14000 genes, and the 7×10^5 kb genome of tomato [52] can contain approximately 140000 genes. The correct number of genes in a plant genome is still unknown, but estimates range to more than 60000 [16]. *A. thaliana* has few repeats [35]. However, this is also true for tomato; only 5% of the nuclear DNA sequences are estimated to consist of highly repetitive DNA [56]. The usually duplicated genes (i.e. gene families) are present as gene families in *A. thaliana* as well, and only in somewhat reduced numbers. For example, *A. thaliana* has three tandemly linked copies of PSII Type I CAB genes [30] whereas tomato has eight [41]; *A. thaliana* has four RBCS genes [28] vs. five in tomato [49]. However, as discussed above, *A. thaliana* almost completely lacks non-linked duplications; RFLP mapping in *A. thaliana* has revealed only 2–3 cases of unlinked duplications. Nevertheless since only about 20% of the tomato nuclear genes are duplicated, the 10-fold difference in DNA content between tomato and *A. thaliana* cannot possibly be explained by a difference in the number of copies of mildly duplicated sequences. It is likely that the remaining difference is accounted for by shortening/lengthening of introns and 5′ and 3′ regions. Thus, differences in genome size may partially reflect the level of sequence duplication, but other factors may also be involved.

Conclusions

It is argued here that there are two main pathways for gene duplications in plants. Genes encoding abundant proteins are most often found in clusters, and the tandemly linked duplication in these clusters are created by unequal crossing-over events. In contrast, genes encoding non-abundant proteins are seldom found in duplicate copies, and, when they are, the two copies are almost always unlinked to each other. It is hypothesized that these apparent random, sporadic duplications are the result of the process of breakage and reintegration elsewhere in the genome of linear DNA fragments. This process of movement of DNA fragments may also result in the occasional duplication of a gene which is already part of a gene cluster. The reintegration of linear DNA segments is postulated to occur by heteroduplex formation of short sequences at one end, and ligation without heteroduplex formation at the other end, and this mechanism is distinct from the process of transposon integration.

Acknowledgements

I wish to thank Drs. L.D. Gottlieb, M. MacGrath, and S.D. Tanksley for critical review of the manuscript. Supported by grants from USDA (88-37262-3878) and NSF (DMB-8819210).

References

1. Bernatzky R, Tanksley SD: Majority of random cDNA clones correspond to single loci in the tomato genome. Mol Gen Genet 203: 8–14 (1986).
2. Bernatzky R, Mau SL, Clark E: A nuclear sequence associated with self-incompatibility in *Nicotiana alata* has homology with mitochondrial DNA. Theor Appl Genet 77: 320–324 (1989).
3. Bonierbale MW, Plaised RL, Tanksley SD: RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics 120: 1095–1103 (1988).
4. Burnham C: Discussions in cytogenetics. Burgess Publishing Co, Minneapolis, MN (1962).
5. Cannon RE, Scandalios JG: Two cDNAs encode nearly identical Cu/Zn superoxide dismutase proteins in maize. Mol Gen Genet 219: 1–8 (1989).
6. Castresana C, Stanelone R, Malik VS, Cashmore AR: Molecular characterization of two clusters of genes encoding Type I CAB polypeptides of PSII in *Nicotiana plumbaginifolia*. Plant Mol Biol 10: 117–126 (1987).
7. Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM: Restriction fragment length polymor-

phism linkage map for *Arabidopsis thaliana*. Proc Natl Acad Sci USA 85: 6856–6860 (1988).

8. Charebet N, Philipps G, Gigot C: Organization of the histone H3 and H4 multigene family in maize and related species. Mol Gen Genet 219: 404–412 (1989).
9. Cheng WY, Scott NS: A contiguous sequence in spinach nuclear DNA is homologous to three separated sequences in chloroplast DNA. Theor Appl Genet 77: 625–633 (1989).
10. Dean C, van der Elzen P, Tamaki S, Dunsmuir P, Bedbrook J: Linkage and homology analysis divide the eight genes for the small subunit of petunia ribulose bisphosphate carboxylase into three gene families. Proc Natl Acad Sci USA 82: 4964–4968 (1985).
11. Dean C, Pichersky E, Dunsmuir P: Structure, evolution and regulation of RbcS genes in higher plants. Ann Rev Plant Physiol Plant Mol Biol 40: 415–439 (1989).
12. Doring HP, Starlinger P: Molecular genetics of transposable elements in plants. Ann Rev Genet 20: 175–200 (1986).
13. Drouin G, Dover GA: A plant processed pseudogene. Nature 328: 557–558 (1987).
14. Fejes E, Masters BS, McCarthy DM, Hauswirth WW: Sequence and transcriptional analysis of a chloroplast insert in the mitochondrial genome of *Zea mays*. Curr Genet 13: 509–515 (1988).
15. Furner IJ, Higgins ES, Berrington AW: Single-stranded DNA transforms plant protoplasts. Mol Gen Genet 220: 65–68 (1989).
16. Goldberg RB: Plants: Novel developmental processes. Science 240: 1460–1467 (1988).
17. Gottesman S, Squires C, Pichersky E, Carrington M, Hobbs M, Mattick JS, Dalrymple B, Kuramitsu H, Shiroza T, Foster T, Clark WP, Ross B, Squires C, Maurizi MR: Conservation of the regulatory subunit of the Clp ATP-dependent protease in prokaryotes and eukaryotes. Proc Natl Acad Sci USA 87: 3513–1317 (1990).
18. Gottlieb LD: Gene duplication and fixed heterozygosity for alcohol dehydrogenase in the diploid plant *Clarkia franciscana*. Proc Natl Acad Sci USA 71: 1816–1818 (1974).
19. Gottlieb LD: Evidence for duplication and divergence of the structural gene for phosphoglucose isomerase in diploid species of *Clarkia*. Genetics 86: 289–306 (1977).
20. Gottlieb LD: Conservation and duplication of isozymes in plants. Science 216: 373–380 (1982).
21. Grant V: Plant Speciation, 2nd ed. Columbia University Press, New York (1981).
22. Helentjaris T, Webber D, Wright S: Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics 118: 353–363 (1988).
23. Helentjaris T: A genetic linkage map for maize based on RFLPs. Trends Biol Sci 3: 217–221 (1987).
24. Holliday R: A mechanism for gene conversion in fungi. Genet Res 5: 282–304 (1964).
25. Jeffreys AJ, Harris S: Process of gene duplication. Nature 296: 9–10 (1982).
26. Köhler F, Cardon G, Pöhlman M, Gill R, Schieder O: Enhancement of transformation rates in higher plants by low-dose irradiation: Are DNA repair systems involved in the incorporation of exogenous DNA into the plant genome? Plant Mol Biol 12: 189–199 (1989).
27. Koukolikova Z, Albright L, Hohn B: The mechanism of T-DNA transfer from *Agrobacterium tumefaciens* to the plant cell. In: Holn Th, Schell J (eds) Plant DNA Infectious Agents, pp. 110–148. Springer, Berlin Heidelberg New York (1987).
28. Krebbers E, Seurinck J, Herdies L, Cashmore AR, Timko MP: Four genes in two diverged subfamilies encode the ribulose-1,5-bisphosphate carboxylase small subunit polypeptides of *Arabidopsis thaliana*. Plant Mol Biol 11: 745–759 (1988).
29. Landry BS, Kesseli RV, Farrar B, Michelmore RW: A genetic map of lettuce (*Lactuca sativa* L) with restriction fragment length polymorphism, isozyme, disease resistance and morphological markers. Genetics 116: 331–337 (1987).
30. Leutwiler LS, Meyerowitz EM, Tobin EM: Structure and expression of three light-harvesting chlorophyll *a*/*b*-binding protein genes in *Arabidopsis thaliana*. Nucl Acids Res 14: 4051–4076 (1986).
31. McCouch SR, Kochert G, Yu ZH, Wang ZY, Kush GS, Coffman WR, Tanksley SD: Molecular mapping of rice chromosomes. Theor Appl Genet 76: 815–829 (1988).
32. McGrath JM, Quiros CF, Harada JJ, Landry BS: Mapping *Brassica oleracea* chromosomes with monosomic alien addition lines: identification and stability of alien chromosomes. Genome, in press (1990).
33. McLean M, Baird WmV, Gerats AGM, Meagher RB: Determination of copy number and linkage relationships among five actin gene subfamilies in *Petunia hybrida*. Plant Mol Biol 11: 663–672 (1988).
34. Meselson MS, Radding CM: A general model for genetic recombination. Proc Natl Acad Sci USA 72: 358–361 (1975).
35. Meyerowitz EM, Pruitt RE: *Arabidopsis thaliana* and plant molecular genetics. Science 229: 1214–1224 (1985).
36. Nam H-G, Giraudat J, den Boer B, Moonan F, Loos WDB, Hauge BM, Goodman HM: Restriction fragment length polymorphism linkage map of *Arabidopsis thaliana*. Plant Cell 1: 699–705 (1989).
37. Odrzykoski IJ, Gottlieb LD: Duplications of genes encoding 6-phosphogluconate dehydrogenase in *Clarkia* (Onagraceae) and their phylogenetic implications. Syst Bot 9: 479–489 (1984).
38. Ohno S: Evolution by Gene Duplications. Springer-Verlag, New York (1970).
39. Pfeiffer P, Vielmetter W: Joining of nonhomologous

DNA double strand breaks *in vitro.* Nucl Acids Res 16: 907–924 (1988).

40. Pichersky E, Gottlieb LD: Evidence for duplication of the structural genes coding plastid and cytosolic isozymes of triose phosphate isomerase in diploid species of *Clarkia.* Genetics 105: 421–436 (1983).
41. Pichersky E, Bernatzky R, Tanksley SD, Breidenbach W, Kausch A, Cashmore AR: Isolation, molecular characterization, and genetic mapping of two clusters of genes encoding chlorophyll *a/b* binding proteins in *Lycopersicon esculentum* (tomato). Gene 40: 247–258 (1985).
42. Pichersky E, Tanksley SD: Chloroplast DNA sequences integrated into an intron of a tomato nuclear gene. Mol Gen Genet 215: 65–68 (1988).
43. Pichersky E, Brock TG, Nguyen D, Hoffman NE, Piechulla B, Tanksley SD, Green BR: A new member of the CAB gene family: Structure, expression and chromosomal location of *Cab*-8, the tomato gene encoding the Type III chlorophyll *a/b*-binding polypeptide of photosystem I. Plant Mol Biol 12: 257–270 (1989).
44. Roose M, Gottlieb LD: Alcohol dehydrogenase in the diploid plant *Stephonomeria exigua* (Compositae): gene duplication, mode of inheritance, and linkage. Genetics 95: 171–186 (1980).
45. Roth DB, Wilson JH: Nonhomologous recombination in mammalian cells: Role for short sequence homologies in the joining reaction. Mol Cell Biol 6: 4295–4304 (1986).
46. Shillito RD, Saul M, Paszkowski J, Potrykus I: High efficiency direct gene transfer to plants. Bio/technology 3: 1099–1103 (1985).
47. Soltis PA, Soltis DE, Gottlieb LD: Phosphoglucose mutase gene duplications in *Clarkia* (Onagraceae) and their phylogenetic implications. Evolution 41: 667–671 (1987).
48. Stern D, Palmer JD: Extensive and widespread homologies between mitochondrial and chloroplast DNA in plants. Proc Natl Acad Sci USA 81: 1946–1950 (1984).
49. Sugita M, Manzara T, Pichersky E, Cashmore AR, Gruissem W: Genome organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. Mol Gen Genet 209: 247–256 (1987).
50. Sun H, Treco DG, Schultes NP, Szostak JW: Double-strand breaks at an initiation site for meiotic gene conversion. Nature 338: 87–90 (1989).
51. Tanksley D, Miller J, Paterson A, Bernatzky R: Molecular mapping of plant chromosomes. In: Gustafson JP, Appels RA (eds) Chromosome Structure and Function, pp 157–173. Plenum Press, New York (1987).
52. Tanksley SD, Pichersky E: Organization and evolution of sequences in the plant nuclear genome. In: Gottlieb LD, Jain S (eds): Plant Evolutionary Biology, pp 55–83. Chapman and Hill Inc, London (1988).
53. Tanksley SD, Bernatzky R, Lapitan NL, Prince JP: Conservation of gene repertoire but not gene order in pepper and tomato. Proc Natl Acad Sci USA 85: 6419–6423 (1988).
54. Thode S, Schafer A, Pfeiffer P, Vielmetter W: A novel pathway of DNA end-to-end joining. Cell 60: 921–928 (1990).
55. Weeden NF: Genetic and biochemical implications of the endosymbiotic origin of the chloroplast. J Mol Evol 17: 133–139 (1981).
56. Zamir D, Tanksley SD: Tomato genome is comprised of fast-evolving, low copy-number sequences. Mol Gen Genet 213: 254–261 (1988).