New Phytologist

Review

Research review

Gene order in plants: a slow but sure shuffle

Author for correspondence:
Todd J. Vision
Tel: +1 919 8434507
Fax: +1 919 9621625
Email: tjv@bio.unc.edu

Received: *26 April 2005*
Accepted: *30 June 2005*

Todd J. Vision

Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, USA

Key words: comparative mapping, gene duplication, gene loss, gene order, synteny.

Summary

Comparative mapping studies have revealed a great deal about the patterns of gene order and gene content evolution in plants. These findings have practical importance for leveraging genomic information from model to nonmodel plant species. However, there is much to be learned about the processes by which gene order and content evolve. The role of gene duplication and loss in the evolution of plant gene order, in particular, appears to be more important than commonly appreciated. An exciting area of current research is the study of gene order and content polymorphism within species. Some recent findings suggest that there may be a functional, and adaptive, relationship between gene order and phenotype that is mediated by the effects of gene order on transcriptional regulation.

New Phytologist (2005) **168**: 51–60

© *New Phytologist* (2005) **doi**: 10.1111/j.1469-8137.2005.01537.x

Introduction

We have come to take for granted that genes are arranged in a linear order along chromosomes. But this very basic fact raises some difficult and unanswered questions. While the linear (and higher-order) packaging of genes into chromosomes is undoubtedly critical to ensure the stable replication, recombination and segregation of genetic material during mitosis and meiosis, what rules, if any, govern the specific gene order, and how does this order evolve? Does gene order actually affect the functioning of the cell, or is it arbitrary? Is the three-dimensional arrangement of chromatin during interphase perhaps more important? If the linear order does matter, how much does variation in gene order provide raw material for adaptation and phenotypic diversification? We are regrettably ignorant about the answers to these questions. However, recent discoveries about patterns of gene order rearrangement at both microevolutionary and macroevolutionary time-scales have renewed interest in this area and suggested directions that are ripe for further investigation.

Another, very practical, motivation for understanding gene order evolution in plants is to better utilize this information for the genetic dissection of traits of agronomic importance. Simmonds (1976) estimated that there are over 100 major crops in 37 different taxonomic plant families (primarily in the angiosperms), plus a larger number of timber species. Yet only a small number of angiosperms are in the pipeline for large-scale genomic sequence (e.g. *Medicago truncatula* and tomato) or have already been sequenced (e.g. *Arabidopsis thaliana*, rice and poplar). Because of a large genome and an abundance of repetitive DNA, comprehensive genomic sequence will not be available in the foreseeable future for many of the remaining species. Furthermore, for successful map-based cloning of Mendelian and quantitative trait loci

(QTL), one would like to have a small ratio of physical to genetic map distance within a target genome; unfortunately, many crops have large genomes with relatively little recombination. For instance, 1 cM in wheat corresponds on average to approx. 5 Mb compared to only approx. 0.3 Mb in *Arabidopsis*. However, if a closely related and more tractable species can be found for which a dense map or genome sequence exists (what one might call an *informant genome*) a *comparative map* might allow one to 'borrow' molecular markers and candidate genes from the informant genome and apply them to the (less tractable) target (Ku *et al.*, 2001; Stracke *et al.*, 2002). Comparative mapping can even enable surrogate cloning, in which the gene is first isolated from the more tractable relative. Thus, the conservation of gene order among related plants has been a topic of great interest to crop geneticists. Such conservation is sometimes called synteny, though the term more properly refers only to the occurrence of two pairs of homologous genes on the same pair of chromosomes.

Despite a number of success stories, there has been some debate over whether gene order conservation in plants is sufficient to be of practically utility. Gaut (2002) has shown, by analysis of a number of relatively sparse pairwise comparative maps between grass species, that the probability of two adjacent markers being syntenic is considerably less than 100%. In the worst case (foxtail millet compared with rice) the probability was below 50%. Whether such imperfect conservation can still be useful depends largely on how, and how carefully, researchers use comparative mapping information. It needs to be recognized that naively assuming perfect conservation of order among homologous genes is a recipe for frequent disappointment. The better we understand the processes by which genomes evolve, the better we will be able to make predictions about gene content and order and, just as importantly, accurately assess how confident we are in our predictions.

The arrival of deep-time comparative maps

A number of studies have shown that recognizably conserved DNA sequences outside of protein and RNA coding genes are small and rare in comparisons between divergent plant genomes relative to the more extensive conservation of noncoding sequence seen in mammals (Lockton & Gaut, 2005). Intergenic sequence can evolve rapidly and dramatically, even varying in size substantially within a genus (Wendel *et al.*, 2002). It has been shown, at least in some systems, that this dynamism in intergenic spacer length results from rapid turnover among transposable elements, particularly retroelements (Bennetzen *et al.*, 2005). Consequently, treating the chromatin as simply a linear order of genes, what might be called the 'beads-on-a-string' approach, is a reasonable simplification, one that greatly facilitates the task of aligning distantly-related plant genomes.

For technological reasons, comparative maps were, until a few years ago, restricted to closely related species. The first dense and genome-wide comparative maps in plants were constructed using unsequenced molecular markers that needed to be experimentally genotyped in multiple species, namely gene-based restriction fragment length polymorphism (RFLP) hybridization probes. These RFLPs were used to construct comparative maps between species within the same taxonomic family, such as the cereals (Gale & Devos, 1998), nightshades (Doganlar *et al.*, 2002) and others. However, it has proven difficult to use heterologous probes for studies involving different taxonomic families because, in order for such RFLP probes to cross-hybridize, the sequences must have greater than 70–80% identity at the DNA level.

The 'family barrier' was overcome by the advent of large expressed sequence tag (EST) collections. These are sequences derived more or less randomly from large collections of expressed transcripts, and so provide a sample of protein coding genes; they provide a convenient and useful pool of markers with which to construct a dense linkage map (Rudd, 2003). There are now dozens of species that have > 5000 ESTs in GenBank; in many of these species, the ESTs are being mapped using an assortment of different experimental techniques. Importantly, putative homology between ESTs mapped in different species can be determined *in silico* (i.e. computationally). Computational techniques permit detection of statistically significant similarity between protein-coding sequences even when their amino acid identity is as low as 30%. For many proteins, homologs can be detected that diverged as far back as the common ancestor of eukaryotes and prokaryotes, much less the common ancestor of land plants. Thus, mapped ESTs (or otherwise sequenced protein-coding sequences) can serve as *anchors,* or putatively homologous markers, in comparative maps between highly divergent species. In fact, the large gene families that are often revealed by analysis of sequenced markers present a new kind of challenge: one must sift through the pile of irrelevant (and sometimes spurious) homologies to find those few that inform the comparative map. This has helped motivate the development of increasingly sophisticated computational techniques for identifying chromosomal segments sharing common ancestry from different kinds of map data (Calabrese *et al.*, 2003; Simillion *et al.*, 2004; Hampson *et al.*, 2005).

Another development that has helped to revolutionize our understanding of the evolution of gene order in plants has been the wide adoption of large insert cloning vectors (e.g. bacterial artificial chromosomes, or BACs), which allow cloning of well over 100 kb of contiguous sequence. As with ESTs, libraries of BAC clones are being developed for a phylogenetically diverse set of plant species. While ESTs placed on a genome-wide linkage map allow one to explore macrosynteny (gene order conservation on a coarse scale), the sequencing or physical mapping of BACs from homologous regions allows one to explore microsynteny, in which even very small rearrangements are detectable. Finally, the handful of publicly available large-scale plant genome sequences that have been

completed and are underway allow both genome-wide and high-resolution comparisons of gene order. All told, the democratization of genomic technology to more plant species through the increasing availability of EST sequences, genetic maps, large-insert clone sequences, and whole-genome sequences has allowed the characterization of gene order conservation and rearrangement over much wider phylogenetic distances, and at much finer resolution, then was possible a decade ago.

The conundrum of novel genes

In addition to changes in gene order, distantly related plant species can differ dramatically in gene content. At the simplest level, genes may be present in one species that have no detectable homolog in another. Allen (2002) identified 154 genes that do not appear to be present in *Arabidopsis* but are known from other plant genomes. These may represent actual deletions of the genes. Alternatively, some may be present in the genome but still absent from the available sequence data, or may be unrecognized due to rapid sequence evolution.

Novel genes have also been identified that appear to represent lineage-specific acquisitions; Graham *et al.* (2004) recently reported over 2500 EST-derived gene predictions that lack any clear homologs outside of the legumes. Gains or losses of entire families of related genes may also be identified. Over 600 potentially legume-specific families were identified in the same study by Graham *et al.* (2004). The authors propose that many of these genes and gene families are functionally related to the ability of legumes to enter into a symbiotic relationship with nitrogen-fixing bacteria. One caveat, though, is that EST sequencing in the legumes has been heavily biased toward root-derived cDNA libraries precisely because of the interest in the biology of nitrogen fixation. If comparable attention to roots were paid in other plant families, homologs to some of these novel genes and gene families might be found.

Apparently novel lineage-specific genes may be annotation artifacts. For example, early analyses of the draft rice sequence predicted 40 000–60 000 genes, up to two times the number predicted in *Arabidopsis* (Goff *et al.*, 2002; Yu *et al.*, 2002). Suspiciously, about half of the predicted rice genes had no homolog in *Arabidopsis* while over 80% of *Arabidopsis* genes had a homolog in rice. Bennetzen and colleagues (2004) argue that the vast majority of these putatively novel rice genes are fragments of long-terminal repeat retrotransposons, and do not belong in the gene count. They advocate using comparative sequence analysis to flag predicted genes that have no homologs in related species. This unfortunately results in a strong bias against the discovery of lineage-specific genes, but the problem of gene overprediction is arguably the greater one in complex plant genomes.

There are several possible biological explanations for a novel gene, assuming it is not simply an annotation artifact. As suggested above, it might represent a protein undergoing extremely fast amino acid substitution. Thus, its homologs, though present in other species, are simply not recognized (Domazet-Loso & Tautz, 2003). More exotic possibilities can be considered, however. One is that the gene was introduced via horizontal transfer from another organism. Although there is evidence for horizontal transfer from the genomes of other organisms into land plants (Aoki, 2004), such events appear to be very rare, and therefore are unlikely to explain the acquisition of a large number of novel genes. Other processes might include the construction of a new gene from noncoding sequence or, more plausibly, from the parts or wholes of two or more pre-existing genes (Long *et al.*, 2003). Chimeric genes do occur, though they would not necessarily lack detectable homology with their ancestral genes. A recent report suggests that the rate at which chimeras are generated in eukaryotes is low, roughly one-tenth the rate of gene duplication (Conant & Wagner, 2005), although no plant genomes themselves have been analysed in this way. The mechanisms by which chimeras arise are in need of more study. Recently, Jiang *et al.* (2004) reported the existence of a new class of DNA transposons in rice called Pack-MULEs that are able to capture and concatenate pieces of DNA from multiple genes. However, the concatenated pieces tend to be small, unlike those found in many functional chimeras.

The consequences of gene duplication and loss

More subtle changes in gene content can be generated by lineage-specific expansions and contractions of a gene family that is shared between species (Shiu *et al.*, 2004). If turnover (i.e. duplication and loss) within a gene family is happening sufficiently rapidly (Durbin *et al.*, 2000), it may not be possible to identify one-to-one correspondence between the members of the family in two different species. This phenomenon poses a challenge to current methods for comparative mapping between distantly related species (Sankoff, 1999). Thus, an important area for future methodological research is to consider phylogenetic relationships among potential anchor genes in the analysis of comparative map data. In fact, much of the disturbance to synteny in plants appears to be a secondary result of gene duplication. However, not all gene duplications are created equal. In particular, they differ in their effect on gene order depending on their chromosomal context.

Gene duplications may be classified as: tandem, when a single gene is duplicated and the two daughters occur adjacent to one another in the genome; dispersed, when a single gene is duplicated but one of the daughters moves to a novel position; and segmental, when a contiguous tract containing multiple genes is duplicated (Remington *et al.*, 2004). Special cases of segmental duplication include duplication of one or a limited number of chromosomes (a class of aneuploidy), as appears to have happened in the recent history of rice (Wang *et al.*, 2005), or of an entire genome (polyploidy), as appears to have

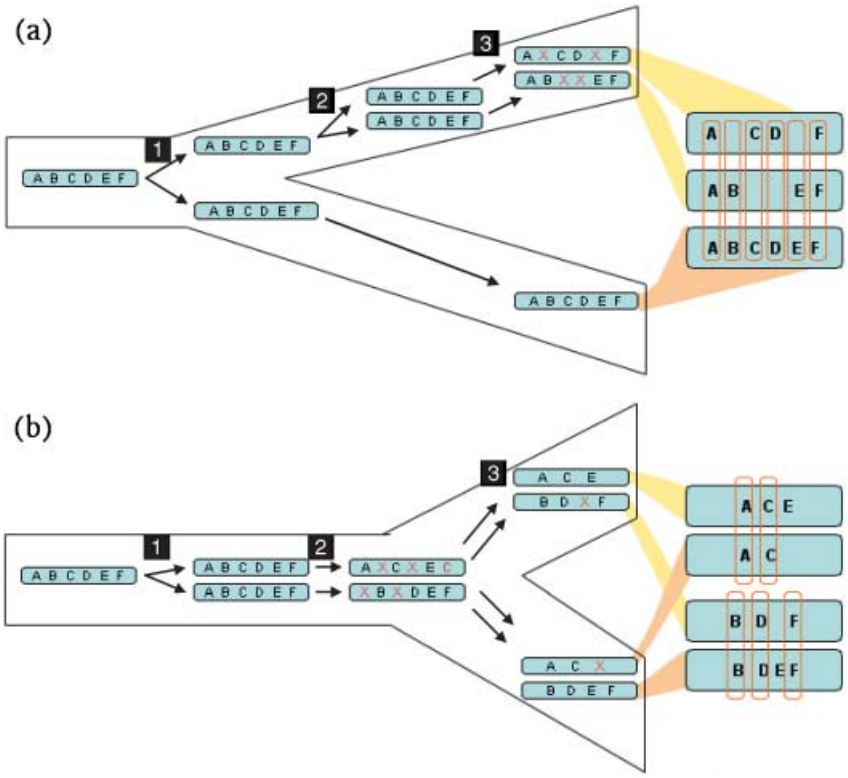

Fig. 1 Two different consequences of segmental duplication followed by gene loss. (a) A network of synteny resulting from segmental duplication after the divergence of the species being compared. Letters within chromosome segments represent genes. The segments evolve within the organismal phylogeny at left; the comparative map is shown at right. 1, speciation; 2, segmental duplication; 3, reciprocal gene loss. (b) Nonsyntenic markers resulting from differential resolution of a duplicated gene in two lineages that diverged after a segmental duplication. 1, segmental duplication; 2, incomplete reciprocal gene loss; 3, differential resolution of duplicate gene E leads to discordance between this gene and the overall pattern of synteny.

occurred several times in the history of Arabidopsis (Simillion *et al.*, 2002).

A tandem duplication does not appreciably disturb gene order as long as both daughter genes occur at the same position as their common ancestor, a dispersed duplication results in the insertion of a gene into a novel position, and a segmental duplication preserves the local order within the duplicated tract but not the identity of the neighbors outside of the tract. Obviously, chromosomal and genome duplications are special in this respect, since the duplicated tracts do not have neighbors. However, two daughter chromosomes can subsequently be affected by large-scale rearrangements that place their constituent genes in the proximity of different neighbors.

Dispersed and segmental duplications can both lead to much greater disturbances of gene order when coupled with gene loss. In the case of a dispersed duplication, if the lost copy were the ancestral one, then it would be impossible to discern that the gene has not simply been transposed from one genomic location to another. Segmental duplication results in a different pattern. If one member of each segmentally duplicated gene pair is lost, and the segment in which the gene is retained differs among the gene pairs, then a distinctive pattern of rearrangement results in which the neighbors of each gene are a only a subset of the neighbors in the ancestral segment.

How segmental duplication and gene loss reveals itself in a comparative map depends on whether the duplication event occurred (1) along one of the two lineages being compared or (2) in their common ancestor. In the first case (Fig. 1a), the genes within the unduplicated genome are distributed within two different segments in the duplicated genome, though some genes may remain in all three segments. Those genes that do remain will be collinear and have conserved transcriptional orientation. This pattern has been dubbed a 'network of synteny' (Ku *et al.*, 2000; Mayer *et al.*, 2001). In the second case (Fig. 1b), the two species being compared diverged after the genome duplication event, in which case genes still present in duplicate at the time of divergence may be lost from different segments in the two species. Thus, although there would be a one-to-one correspondence between the homologous segments in the two species, there might be a number of genes that are discordant with the overall pattern.

A recent analysis of the distribution of synonymous nucleotide divergence among duplicated genes in a variety of EST datasets by Blanc & Wolfe (2004b) strongly suggests that large segmental duplications (aneuploidy or polyploidy events) have occurred within the last few tens of millions of years in many different flowering plant lineages. Between each polyploidy event, one might anticipate that hundreds to thousands of individual gene duplications would become fixed, judging by the average rate of such fixations in eukaryotic genomes (Lynch & Conery, 2003). Nonetheless, polyploidy events still appear to play a more important role in shuffling gene order simply because of the much larger number of genes

that are duplicated during each event. Maere *et al.* (2005) estimate that the three genome duplications in the lineage leading to *Arabidopsis* have been responsible for *c.* 59% of the gene duplications that have been retained in during the last 350 Myr. If the vast majority of genes revert to single copy afterward, as appears to have been the case in *Arabidopsis* (Maere *et al.*, 2005), then a gene has only a very small chance of retaining both of the immediate neighbors that flanked it before the polyploidy event. Incidentally, Maere *et al.*'s analysis assumes a constant background rate of gene duplication and a fixed distribution of synonymous sequence divergence rates among those duplicates, assumptions brought into doubt by Blanc & Wolfe's (2004b) analysis of tandem duplications in *Arabidopsis*. This violation, however, is unlikely to greatly influence Maere *et al.*'s results.

The predominant role of segmental duplication followed by gene loss in shuffling gene order between distantly related plant genomes is supported by a burgeoning number of micosynteny studies, some of the earliest of which were between tomato and *Arabidopsis*, which diverged early in the history of eudicots (Ku *et al.*, 2000) and rice and *Arabidopsis*, which diverged even earlier (Mayer *et al.*, 2001). In the latter study of a 333-kb genomic sequence from rice, five syntenic segments were identified in *Arabidopsis*. Assuming that there have been three polyploidy events in the *Arabidopsis* lineage since its divergence with rice, then three unidentified syntenic segments must remain to be identified in *Arabidopsis*. Of the 56 annotated rice protein-coding genes, 22 had homologs present on one or more of the *Arabidopsis* segments; the largest number for any single *Arabidopsis* segment being eight. Only one pair of homologs showed evidence of having an inverted order between segments, and the majority of genes had conserved transcriptional orientation.

A different prediction of the segmental duplication-gene loss model is that single genes discordant with the overall pattern of synteny will be particularly common in comparisons between two species whose common ancestor had recently undergone polyploidy (Fig. 1b). Livingstone *et al.* (1999) noted that, in comparative maps among species of the same ploidy, both within the Poaceae and the Solanaceae, 20–40% of markers were discordant. In both groups, there is evidence for polyploidy in the common ancestor having occurred shortly before the divergence of the species being compared (Blanc & Wolfe, 2004b; Paterson *et al.*, 2004). In fact, Paterson *et al.* (2004) found that rice markers present on a comparative map with sorghum, but at a nonsyntenic locus, were twice as likely to be found within the duplicated region of the rice genome. Thus, independent loss of segmentally duplicated genes along two lineages that diverged following a polyploidy event (as illustrated in Fig. 1b) does appear to explain some, if not all, cases of synteny violation between such species. The results of Paterson and colleagues also suggest that, while some gene losses may occur within the first few generations of polyploid establishment, at least in allopolyploids (Kashkush *et al.*, 2002), a substantial fraction of the duplicated genes are resolved only on longer evolutionary time-scales. This phenomenon may also help to explain Gaut's (2002) findings of unusually frequent violations of synteny in the grasses, cited above.

One of the more surprising findings to come out of the recent studies on the complex history of gene duplication and loss in *Arabidopsis* is that the retention of different classes of duplicate genes is highly nonrandom. While the overall rate of gene loss following small-scale (tandem and dispersed) and large-scale (segmental) duplication in *Arabidopsis* is not grossly different, there is evidence that some polyploidy events had higher retention rates than others (when corrected for the time since duplication) (Maere *et al.*, 2005). The products of the genes retained from large-scale segmental duplications are greatly enriched in transcription factors, signal transduction proteins and other classes of regulatory proteins, while those retained by small-scale duplication are greatly enriched in genes coding for proteins involved in secondary metabolism, or implicated in responses to biotic and abiotic stresses (Blanc & Wolfe, 2004a; Maere *et al.*, 2005). This functional bias persisted over multiple rounds of polyploidization and gene loss in *Arabidopsis* (Seoighe & Gehring, 2004; Maere *et al.*, 2005). There is considerable variation among gene families in the extent to which proliferation involved small-scale vs large-scale duplication (Cannon *et al.*, 2004; Remington *et al.*, 2004).

There are a few possible explanations for nonrandom patterns of retention among functional classes of genes. One is that polyploidy allows the simultaneous coevolutionary divergence of many genes within the same tightly interconnected developmental or regulatory pathway. Conversely, genes retained after small-scale duplication are those that can contribute to an adaptive phenotype without requiring compensating changes at other loci (such as many enzymes involved in secondary metabolism, for example). There is evidence suggesting coevolutionary divergence of whole pathways following polyploidy from the work of Blanc & Wolfe (2004a). The expression profile of a gene may be defined as the relative abundance of its mRNA in different cells, under different conditions, and at different stages of development. The authors searched for sets of genes retained from the most recent polyploidy event in *Arabidopsis* for which the two copies had only weakly correlated expression profiles (Pearson $r < 0.1$), but for which there was a high correlation in the expression profiles between the copies from different pairs ($r = 0.7$). Several sets of gene pairs were discovered that satisfied this criterion, the largest involving 13 pairs. However, the functional significance in these cases of putative concerted divergence has yet to be determined.

An alternative explanation for the biased retention of certain classes of genes following small-scale vs segmental duplication would be the presence of long-range *cis*-regulatory elements. Those genes that rely on distantly placed transcription

factor binding sites for their appropriate regulation would be less likely to be retained following small-scale duplication for the simple reason that the necessary sequence would not be copied intact. Genes that are larger, on average, would have a similar bias. Studies in *C. elegans* (Katju & Lynch, 2003) and *Arabidopsis* (Moore & Purugganan, 2003) both suggest that duplications may often be truncated or include only a limited amount of intergenic spacer sequence. Whatever the reason for the functional biases among retained genes, it reinforces the idea that the kinds of evolutionary change enabled by polyploidy may be qualitatively different from that enabled by the accumulation of many small-scale duplication events. An exciting area of future work is to determine what lineage-specific phenotypic innovations were enabled by each of the ancient large-scale duplications.

The pattern of segmental duplication followed by gene loss also leads to methodological challenges in comparative mapping. The problem is that multiple segments in one genome may map to multiple segments in another, and that the number of anchors in common between any pair of segments is only a fraction of the number of genes present in the common ancestor. This makes syntenic segments harder to detect, especially when maps are sparse, and has limited the utility of *Arabidopsis*, in particular, for comparative mapping in the eudicots. Segments that descended from a common ancestor may only be revealed by comparison with a third segment from the same or a different species that is syntenic to them both (Simillion *et al.*, 2002; Vandepoele *et al.*, 2002). Computational methods for comparative mapping have begun to address this issue by directly or indirectly reconstructing the gene order in the ancestor and thereby improving the power to detect highly divergent syntenic segments (Blanc *et al.*, 2003; Bowers *et al.*, 2003; Huan *et al.*, 2003; Langham *et al.*, 2004; Simillion *et al.*, 2004). As an example of the power of this approach, Simillion *et al.* (2004) report that over 25% of the *Arabidopsis* genome is segmentally duplicated with five or more copies (suggestive of at least three rounds of segmental duplication), compared with c. 8% of the genome that fell into this class when only pairwise alignments between contemporary syntenic segments were considered.

Other mechanisms of gene order evolution

While duplication does appear to be a major driver of gene order evolution, it is not the only way that changes in gene order can happen. Chromosomal rearrangements such as translocations, large inversions, and fusion/fission events have been known as engines of karyotypic evolution in eukaryotes for many years (Levin, 2002; Eichler & Sankoff, 2003). However, unless they are accompanied by additional rearrangements, which they sometimes may be (Livingstone *et al.*, 1999), they only affect local gene order at the breakpoints of the rearrangement. Estimates from different plant comparative maps suggest that these types of rearrangements cumulatively occur at an average rate of 1–10 per Myr, though the rate clearly varies with the species being compared and the methodology used (Kellogg & Bennetzen, 2004). Such rearrangements do have a notable effect on patterns of macrosynteny: the more rearrangements, the smaller the syntenic blocks (Nadeau & Sankoff, 1998); in fact, the size distribution of syntenic blocks, together with estimates of divergence time, is what is used to derive these estimates. Such macrorearrangements occur too slowly to affect the positions of most genes relative to their near neighbors.

An additional class of gene order rearrangement is small-scale inversion involving only a few genes. Such inversions could be happening at a much faster rate than the macrorearrangements discussed above and therefore be responsible for considerable churning of local gene order. It has been suggested that small (several kilobase-sized) inversions are an important factor in yeast gene order evolution (Seoighe *et al.*, 2000). However, comparisons of syntenic segments at the sequence level have shown that small-scale inversions, at least as evidenced by the occurrence of occasional violations of gene order colinearity and flips of transcriptional orientation between syntenic segments (Mayer *et al.*, 2001), are comparatively rare in plants.

Gene order polymorphism: are we at the tip of the iceberg?

While patterns of single nucleotide polymorphisms are being characterized within small regions in several different plant systems, and karyotypic polymorphisms in plants have been studied for many years (Levin, 2002), only recently has much attention been given to the possibility of polymorphism in gene content and order at an intermediate scale. This is somewhat surprising. Since local gene order differences can be observed between species, there must have been a point in time when they were polymorphic within species, and it is thus reasonable to suppose that some polymorphism in gene order is present within contemporary species as well. However, because of the difficulty of observing such polymorphisms, only a few scattered cases have been reported (Ziolkowski *et al.*, 2003).

A number of recent studies, primarily in maize, have attempted to characterize patterns of polymorphism in small-scale genome structure, precipitated by the discovery of dramatic divergence in gene and other sequence content between alleles at the *bronze* locus of maize (Fu & Dooner, 2002). Brunner *et al.* (2005), in the most extensive study to date, sequenced several 100 kb of two alleles from four arbitrarily chosen genomic loci plus the aforementioned *bronze* locus. Genes, or gene fragments, unique to one of the alleles were commonplace; at one locus, unique genes were nearly three times more abundant than shared genes. The unique genes were physically clustered and generally found scattered in nonsyntenic positions within the rice genome. To what extent

these polymorphisms reflect the kind of variation that would evolve into fixed changes in gene order, as opposed to simply reflecting an accumulation of random and nonfunctional genomic fragments within intergenic spacer regions, remains to be determined. Fu & Dooner (2002) have hypothesized that complementation of unshared genes could be one of the factors contributing to heterosis, or the elevated performance of hybrids between inbred maize lines. However, empirical studies of the phenotypic consequences of this class of polymorphism are in their infancy (Song & Messing, 2003). The clearest functional links to date come not from large-scale comparative sequencing studies, but rather involve disease resistance polymorphisms that are known to result from the presence or absence of particular genes (Tian *et al.*, 2002; Scherrer *et al.*, 2005).

Genome structural polymorphism has also been studied in wheat by Jan Dvorak and colleagues (Akhunov *et al.*, 2003a,b; Dvorak *et al.*, 2004). His group has shown that perturbations of microsynteny among the three genomes of hexaploid wheat are largely because of small duplications and deletions that occurred in wheat's diploid ancestors before hybridization. Many of the duplications and deletions involved genes. Interestingly, the rates at which duplicated genes are inserted and deleted genes are lost has been found to be correlated with the rate of recombination along the chromosome. This may result from a mechanistic link between recombination and insertion/deletion events, or it may reflect the sensitivity of natural selection on both deleterious and advantageous polymorphisms to the local recombinational environment (Gordo & Charlesworth, 2001). Whatever the cause, this bias would lead to a faster erosion of synteny in high recombination regions. A large number of genotypes have been surveyed in these studies, revealing that deletion/duplication polymorphisms have a range of frequencies, though most are close to fixation.

Evidence for nonrandom gene order

It is of obvious practical utility to know the extent of gene order conservation between crops and sequenced model organisms. But does gene order itself affect function at the cellular and organismal level, and can it contribute to variation in fitness? There are a number of ways in which it could, and some intriguing lines of evidence that gene order is, at the very least, not random.

There are a few special instances in which gene order, or at least gene neighborhood, clearly matters. One case is the self-incompatibility locus in the mustard family Brassicaceae, where essentially complete linkage between genes controlling recognition and display of pollen genotype is required for proper functioning (Kusaba *et al.*, 2001). A similar pressure exists for very tight linkage between sex determination loci on the X and Y chromosomes (Charlesworth, 2002).

More generally, in a number of eukaryotes, including *Arabidopsis*, it has been observed that the transcriptional profiles of neighboring genes are more similar to one another than would be expected by chance (Williams & Bowles, 2004). It is not yet clear whether this is a result of selection on gene order polymorphism that favors transcriptional similarity of neighboring genes, or whether it is simply a side-effect of regional transcriptional controls. There is some evidence for a selective explanation, though not in plants, from a recent study by Singer *et al.* (2005) reporting that clusters of coexpressed genes tend to contain fewer macrorearrangements than expected by chance in comparisons between the mouse and human genomes. This is consistent with the idea that clusters of coexpressed genes are preferentially conserved intact. Yet, a testable mechanistic hypothesis for the adaptive significance of these clusters has yet to be proposed.

There have also been reports of clusters of functionally related genes in the genomes of eukaryotes. Lee & Sonnhammer, 2003) analysed *Arabidopsis*, as well as other completed genomes, and found genes associated with several biochemical pathways to be more closely spaced in the genome than expected by chance. However, the clusters were very loose, consisting of a few genes interspersed with others and extending over many megabases. Such loose clusters of functionally related genes may well be the result of the many gene duplications in *Arabidopsis* that, while not tandem, still tend to be found on the same chromosome and closer to each other than random loci (Vision *et al.*, 2000). If that is true, then it might be more useful to study the mechanisms by which dispersed gene duplications arise than the functional relationships among the genes in such dispersed clusters.

Gene order polymorphism within a species may also have indirect consequences, though these are even more speculative. Imagine two diploid genotypes that differ from each other in the location of a critical gene. One-sixteenth of the F_2 hybrids from a cross between these genotypes would lack both copies of the gene. If there are many such polymorphisms segregating, then gene order differences could contribute to substantial reductions in hybrid fitness (Lynch & Conery, 2001). This is similar to the idea put forward by Fu & Dooner (2002), mentioned above, that the complementation of deficiencies in crosses between inbred lines might contribute to heterosis in maize.

To make further progress in this area, it would be very useful to overcome the technical challenge of identifying intraspecific polymorphisms systematically and genome-wide, since physically mapping or sequencing large numbers of allelic BAC clones is not currently feasible. Having a map of such polymorphisms among a wide sample genotypes would allow researchers to explore the phenotypic consequences of such polymorphisms at the cellular and organismal levels using tools such as QTL and association mapping. It will be very interesting to discover to what extent quantitative variation in nature is underlain by segregation for alleles that differ in gene content.

Conclusion

Comparative mapping has historically been a largely descriptive enterprise. Yet, the macroevolutionary forces involved in gene order evolution, namely different modes of gene duplication and loss, are now sufficiently well characterized that comparative mapping is also poised to become a predictive enterprise. This will have enormous practical benefit to map-based studies of orphan plant species. In order to truly understand the significance of gene order evolution, however, increased attention will need to be paid to the molecular processes that generate the polymorphism and the microevolutionary processes that govern its fate. This necessitates studies not only of patterns of polymorphism but also the potential for such polymorphisms to affect the phenotype and be acted upon by natural selection. Advances in experimental methodology that facilitate the study of this new form of variation are likely to open up the field in coming years. The time is ripe to understand better the gene order shuffle and what it means to the organism.

Acknowledgements

This paper benefited from the helpful comments of two anonymous reviewers, and was supported by a grant from the National Science Foundation (DBI-0227314).

References

Akhunov ED, Akhunova AR, Linkiewicz AM, Dubcovsky J, Hummel D, Lazo G, Chao S, Anderson OD, David J, Qi L, Echalier B, Gill BS, Miftahudin Gustafson JP, La Rota M, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NL, Wennerlind EJ, Nduati V, Anderson JA, Sidhu D, Gill KS, McGuire PE, Qualset CO, Dvorak J. 2003a. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proceedings of the National Academy of Sciences, USA* **100**: 10836–10841.

Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalier B, Gill BS, Miftahudin Gustafson JP, Lazo G, Chao S, Anderson OD, Linkiewicz AM, Dubcovsky J, La Rota M, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NL, Gonzalez-Hernandez JL, Anderson JA, Choi DW, Close TJ, Dilbirligi M, Gill KS, Walker-Simmons MK, Steber C, McGuire PE, Qualset CO, Dvorak J. 2003b. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Research* **13**: 753–763.

Allen KD. 2002. Assaying gene content in *Arabidopsis. Proceedings of the National Academy of Sciences, USA* **99**: 9568–9572.

Aoki S. 2004. Resurrection of an ancestral gene: functional and evolutionary analyses of the *Ngrol* genes transferred from *Agrobacterium* to *Nicotiana*. *Journal of Plant Research* **117**: 329–337.

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W. 2004. Consistent over-estimation of gene number in complex plant genomes. *Current Opinions in Plant Biology* **7**: 732–736.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Annals of Botany (London)* **95**: 127–132.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research* **13**: 137–144.

Blanc G, Wolfe KH. 2004a. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**: 1679–1691.

Blanc G, Wolfe KH. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**: 1667–1678.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**: 343–360.

Calabrese PP, Chakravarty S, Vision TJ. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19**: i74–80.

Cannon SB, Mitra A, Andrew B, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana. BMC Plant Biology* **4**: 10.

Charlesworth D. 2002. Plant sex determination and sex chromosomes. *Heredity* **88**: 94–101.

Conant GC, Wagner A. 2005. The rarity of gene shuffling in conserved genes. *Genome Biology* **6**: R50.

Doganlar S, Frary A, Daunay MC, Lester RN, Tanksley SD. 2002. A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the solanaceae. *Genetics* **161**: 1697–1711.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila. Genome Research* **13**: 2213–2219.

Durbin ML, McCaig B, Clegg MT. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Molecular Biology* **42**: 79–92.

Dvorak J, Yang ZL, You FM, Luo MC. 2004. Deletion polymorphism in wheat chromosome regions with contrasting recombination rates. *Genetics* **168**: 1665–1675.

Eichler EE, Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**: 793–797.

Fu H, Dooner HK. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences, USA* **99**: 9573–9578.

Gale MD, Devos KM. 1998. Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences, USA* **95**: 1971–1974.

Gaut BS. 2002. Evolutionary dynamics of the grasses. *New Phytologist* **154**: 15–28.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100.

Gordo I, Charlesworth B. 2001. Genetic linkage and molecular evolution. *Current Biology* **11**: R684–R686.

Graham MA, Silverstein KA, Cannon SB, VandenBosch KA. 2004. Computational identification and characterization of novel genes from legumes. *Plant Physiology* **135**: 1179–1197.

Hampson SE, Gaut BS, Baldi P. 2005. Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* **21**: 1339–1348.

Huan J, Prins J, Wang W, Vision TJ. 2003. Reconstruction of ancestral gene order following segmental duplication and gene loss. *Proceedings of the 2nd IEEE Computational Systems Bioinformatics Conference (CSB 2003)* 484–485. http://csdl2.computer.org/comp/proceedings/csb/2003/2000/00/20000484.pdf

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**: 569–573.

Kashkush K, Feldman MW, Levy A. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160.

Katju V, Lynch M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* **165**: 1793–1803.

Kellogg EA, Bennetzen JL. 2004. The evolution of nuclear genome structure in seed plants. *American Journal of Botany* **91**: 1709–1725.

Ku HM, Vision T, Liu J, Tanksley SD. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of the National Academy of Sciences, USA* **97**: 9121–9126.

Ku HM, Liu J, Doganlar S, Tanksley SD. 2001. Exploitation of Arabidopsis-tomato synteny to construct a high-resolution map of the ovate containing region in tomato chromosome 2. *Genome* **44**: 470–475.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME. 2001. Self-incompatibility in the genus *Arabidopsis*: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. *Plant Cell* **13**: 627–643.

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**: 935–945.

Lee JM, Sonnhammer EL. 2003. Genomic clustering analysis of pathways in eukaryotes. *Genome Research* **13**: 875–882.

Levin DA. 2002. *The role of chromosomal change in plant evolution*. Oxford, UK: Oxford University Press.

Livingstone KD, Lackney VK, Blauth JR, van Wijk R, Jahn MK. 1999. Genome mapping in Capsicum and the evolution of genome structure in the Solanaceae. *Genetics* **152**: 1183–1202.

Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogue evolution. *Trend in Genetics* **21**: 60–65.

Long M, Deutsch M, Wang W, Betran E, Brunet FG, Zhang J. 2003. Origin of new genes: evidence from experimental and computational analyses. *Genetica* **18**: 171–182.

Lynch M, Conery J. 2001. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* **3**: 35–44.

Maere S, De Bodt S, Raes J, Casneuf T, van Montagu M, Kuiper M, van de Peer Y. 2005. Modeling gene and genome duplication in the eukaryotes. *Proceedings of the National Academy of Sciences, USA* **102**: 5454–5459.

Mayer K, Murphy G, Tarchini R, Wambutt R, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, Lemcke K, Haase D, Hall CR, van Dodeweerd AM, Tingey SV, Mewes HW, Bevan MW, Bancroft I. 2001. Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of Arabidopsis thaliana. *Genome Research* **11**: 1167–1174.

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences, USA* **100**: 15682–15687.

Nadeau JH, Sankoff D. 1998. Counting on comparative maps. *Trends in Genetics* **14**: 495–501.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences, USA* **101**: 9903–9908.

Remington DL, Vision TJ, Guilfolye TJ, Reed JA. 2004. Contrasting modes of diversification in the Aux/IAA and ARF gene families. *Plant Physiology* **135**: 1738–1752.

Rudd S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science* **8**: 321–329.

Sankoff D. 1999. Genome rearrangement with gene families. *Bioinformatics* **15**: 909–917.

Scherrer B, Isidore E, Klein P, Kim JS, Bellec A, Chalhoub B, Keller B, Feuillet C. 2005. Large intraspecific haplotype variability at the *rph7*locus results from rapid and recent divergence in the barley genome. *Plant Cell* **17**: 361–374.

Seoighe C, Gehring C. 2004. Genome duplication led to a highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends in Genetics* **20**: 461–464.

Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Huizar L, Davis RW, Scherer S, Tait E, Shaw DJ, Harris D, Murphy L, Oliver K, Taylor K, Rajandream MA, Barrell BG, Wolfe KH. 2000. Prevalence of small inversions in yeast gene order evolution. *Proceedings of the National Academy of Sciences, USA* **97**: 14433–14437.

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH. 2004. Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* **16**: 1220–1234.

Simillion C, Vandepoele K, van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences, USA* **99**: 13627–13632.

Simillion C, Vandepoele K, Saeys Y, van de Peer Y. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Research* **14**: 1095–1106.

Simmonds NW. 1976. *Evolution of crop plants*. London, UK: Longman.

Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH. 2005. Clusters of coexpressed genes in mammalian genomes are conserved by natural selection. *Molecular Biology and Evolution* **22**: 767–775.

Song R, Messing J. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proceedings of the National Academy of Sciences, USA* **100**: 9055–9060.

Stracke S, Kistner C, Yoshida S, Mulder L, Sato S, Kaneko T, Tabata S, Sandal N, Stougaard J, Szczyglowski K, Parniske M. 2002. A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* **417**: 959–962.

Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002. Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **99**: 11525–11530.

Vandepoele K, Simillion C, Van de Peer Y. 2002. Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends in Genetics* **18**: 606–608.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290**: 2114–2117.

Wang X, Shi X, Hao B, Ge S, Luo J. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytologist* **165**: 937–946.

Wendel JF, Cronn RC, Johnston JS, Price HJ. 2002. Feast and famine in plant genomes. *Genetica* **115**: 37–47.

Williams EJ, Bowles DJ. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Research* **14**: 1060–1067.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92.

Ziolkowski PA, Blanc G, Sadowski J. 2003. Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome. *Nucleic Acids Research* **31**: 1339–1350.

About *New Phytologist*

- *New Phytologist* is owned by a non-profit-making **charitable trust** dedicated to the promotion of plant science, facilitating projects from symposia to open access for our Tansley reviews. Complete information is available at **www.newphytologist.org**.
- Regular papers, Letters, Research reviews, Rapid reports and both Modelling/Theory and Methods papers are encouraged. We are committed to rapid processing, from online submission through to publication 'as-ready' via *OnlineEarly* – the 2004 average submission to decision time was just 30 days. Online-only colour is **free**, and essential print colour costs will be met if necessary. We also provide 25 offprints as well as a PDF for each article.
- For online summaries and ToC alerts, go to the website and click on 'Journal online'. You can take out a **personal subscription** to the journal for a fraction of the institutional price. Rates start at £109 in Europe/$202 in the USA & Canada for the online edition (click on 'Subscribe' at the website).
- If you have any questions, do get in touch with Central Office (**newphytol@lancaster.ac.uk**; tel +44 1524 594691) or, for a local contact in North America, the US Office (**newphytol@ornl.gov**; tel +1 865 576 5261).

