The Plant Cell, Vol. 16, 1647–1649, July 2004, www.plantcell.org

IN THIS ISSUE

Two Genomes Are Better Than One: Widespread Paleopolyploidy in Plants and Evolutionary Effects

Polyploidy, or whole genome duplication resulting in the doubling of chromosome numbers, is widespread among the flowering plants and is widely believed to play an important role in species evolution and diversification (Otto and Whitton, 2000; Wendel, 2000). It was proposed more than 50 years ago that gene duplication can result in a relaxation of stabilizing selection, allowing for the evolution of novel gene functions (Stephens, 1951). It is now suspected that almost all angiosperms are likely to be ancient polyploids that have undergone diploidization through differential mutation, elimination, and inversion in duplicated chromosomes. The majority of crop species appear to be polyploids, whether resulting from relatively recent or more ancient duplication events, and it is generally believed that polyploidy has conferred distinct advantages for the development of agronomically important traits. For example, polyploidization has been associated with increased size of harvested organs, novel gene interactions leading to new traits, and the formation of new crop species (Gepts, 2003). Although not nearly as common as in the plant kingdom, polyploidy also occurs in animals. Ohno (1970) proposed that multiple rounds of polyploidy occurred during the early evolution of vertebrates, and this idea has been revived with advances in genomics and is actively debated in the literature (reviewed in Wolfe, 2001). Thus, paleopolyploidy may be a universal feature of eukaryotes, and understanding the consequences of gene duplication is an important branch of evolutionary research. In this issue of *The Plant Cell*, **Blanc and Wolfe (pages 1667–1678)** present additional evidence of the widespread occurrence of paleopolyploidy among the flowering plants obtained from a unique approach of analyzing unigene sets of ESTs from 14 model plant species. In a companion article **(pages 1679–1691)**, the same authors investigate functional divergence in a large set of duplicated genes in Arabidopsis.

In the first article, Blanc and Wolfe analyzed unigene sets of ESTs of 14 model plant species from public databases and searched each set for paralogs or duplicated sequences. A raw set of ESTs (obtained from a single or multiple cDNA libraries) includes many redundant sequences and also multiple overlapping sequences corresponding to single genes. A unigene EST set is condensed to remove sequence redundancies and join together overlapping sequences to create a set wherein each sequence corresponds to a unique gene. Although there is a degree of error associated with this process (including sequencing errors and nonoverlapping sequences that actually correspond to a single gene), unigene sequences that share a high degree of similarity may be considered to correspond to paralogous genes (i.e., one or more copies of a gene that have arisen as a result of gene duplication events).

The 14 species analyzed were *Triticum aestivum* (wheat), *Zea mays* (maize), *Solanum lycopersicum* (tomato; formerly known as *Lycopersicon esculentum*), *S. tuberosum* (potato), *Glycine max* (soybean), *Gossypium arboretum* (diploid cotton), *G. hirsutum* (tetraploid cotton), *Medicago truncatula* (barrel medic), *Helianthus annuus* (sunflower), *Lactuca sativa* (lettuce), *Hordeum vulgare* (barley), *Mesembryanthemum crystallinum* (ice plant), *Oryza sativa* (rice), and *Arabidopsis thaliana*. For Arabidopsis and rice, sets of unigenes were assembled from EST data and analyzed in parallel with the sets of genes predicted by the complete genome sequences of these species, as one type of control analysis on the use of unigene EST sets for identifying gene paralogs. For each pair (or set) of gene paralogs identified, the timing of the duplication event was estimated based on analysis of the synonymous substitution rate (Ks, which equals the number of substitutions per synonymous site in a coding sequence). Synonymous substitutions are nucleotide changes that do not alter the amino acid encoded and are therefore assumed to accumulate in a neutral manner (not under selection) at a constant rate that is similar to the background mutation rate. Higher Ks values therefore correspond to longer periods of time since the original duplication event, and Ks values >1.0 result from the fact that a site can change, and then change again, over a long evolutionary period.

If gene duplications and gene deletions occur at random, then the age distribution profile of gene duplication events (the number of pairs of duplicates in a genome having a certain Ks value plotted against the corresponding Ks value) is expected to show an L-shape: there is an initial peak of duplicates corresponding to the youngest age classes of duplicated genes, which falls off as members of the older age classes are eliminated as a result of gene loss. Against this background, a single duplication of an entire genome, or large portion of the genome, in a narrow time frame, may be revealed as a single large secondary peak with a Ks value corresponding to the time of the duplication event (see Blanc and Wolfe, 2004a, Figure 1). Larger Ks values (e.g., >0.75) are associated with increasingly large error (Li, 1997). To minimize the associated error while retaining a reasonably sized data set, Blanc and Wolfe used only Ks values <2.0 for this analysis. The resulting age distributions of Ks values for the 14 species showed the expected L-shapes, with the largest proportion of duplicates having the smallest Ks values rapidly dropping off to smaller and smaller proportions having larger Ks values. This shape is interpreted to indicate that

IN THIS ISSUE

duplications occur in the present (initial peak of low Ks values) and that a large fraction of duplicates are eliminated from the genome over time (steep drop off in representation of higher Ks values).

Blanc and Wolfe's analysis of unigene sets from 14 species revealed several limitations to this approach. For example, the method was only able to detect putative large-scale duplication events occurring in a relatively narrow time window centered around 30 million years (Myr) ago. They were unable to detect very recent polyploidy events (i.e., those occurring within ~5 Myr) that are known to have occurred in the cotton and wheat genomes and are suspected in sunflower. Because the analysis omitted Ks values > 2.0, it also necessarily missed more ancient large-scale duplication events in Arabidopsis, which have been detected by various groups and are estimated to have occurred more than 100 Myr ago (Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003). Given these limitations, the authors still found evidence of polyploidy occurring within the past 30 Myr in nine of the 14 species analyzed. These species encompass a wide range of angiosperm families, providing support for the idea that polyploidy has occurred in the majority of flowering plants.

The analysis of Arabidopsis data revealed some interesting and surprising findings that underscore the complicated evolutionary history of this seemingly simple and uncomplicated genome. A prominent secondary peak in the Ks age distribution profile was apparent, centered around Ks 0.75 to 0.8. Using a similar method based on synonymous substitution rates, Lynch and Conery (2000) also detected a secondary peak in age distribution of Ks values centered around Ks ~0.8. These authors used an estimated rate of substitution of 6.1 per silent site per billion years (average of values reported in Li, 1997; Lynch, 1997) to date this event to ~65 Myr ago, whereas Blanc and Wolfe used an estimated rate of 15 silent substitutions per billion years (estimated for dicots in Koch et al., 2000) to arrive at a date of ~26 Myr ago. Blanc et al. (2003) used the same approach and concluded that a polyploidy event occurred in Arabidopsis during the early emergence of the crucifer family between 24 and 40 Myr ago, much more recently than has been estimated previously. For example, Simillion et al. (2002) concluded that the Arabidopsis genome has undergone three rounds of large-scale duplication or polyploidization and estimated that the most recent polyploidy event occurred ~75 Myr ago, whereas Vision et al. (2000) estimated that there have been at least four large-scale duplication events that occurred 100 to 200 Myr ago.

Interestingly, another broad secondary peak was observed centered around Ks 0.35 that corresponds to a similar peak that was observed when Ks data were plotted for a subset of Arabidopsis duplicate genes that are known from the whole genome analysis to be arranged in tandem array. This peak was puzzling to the authors because it is difficult to explain a large number of tandem duplicates spread throughout the genome having formed in a single narrow time frame. The authors favor the explanation that this peak may have resulted from a lack of younger aged tandem duplicates, which may be eliminated preferentially during recombination because of their high degree of sequence similarity. However, this does not explain the lack of a similar peak associated with the age distribution profile of the rice data set, and it is estimated that tandemly arrayed genes make up a similarly high proportion of the predicted gene coding sequences in these two species (17% in Arabidopsis [Zhang and Gaut, 2003] and 15% in rice [van de Poele et al. 2003]). The authors hypothesize that the difference between the age distributions of tandemly duplicated genes in Arabidopis and rice is because of a recent increase in the rate of DNA deletion in Arabidopsis, but a definitive explanation awaits further investigations into the origin and evolution of tandemly arrayed genes.

In the companion article, Blanc and Wolfe assessed patterns of functional divergence of duplicated gene pairs in Arabidopsis by analyzing functional genomics data available for a large set of paralogous genes. The data set for this analysis included only those duplicates (based on their Ks values and genomic position data) formed during the most recent polyploidy event and excluded tandem duplicates and more ancient duplicates. Interestingly, they found evidence of a nonrandom pattern of elimination of duplicated genes. The duplicates of genes encoding proteins associated with signal transduction and those associated with transcription functions appear to have been preferentially retained, whereas those associated with DNA repair have been preferentially lost (i.e., are underrepresented among paralogs). Interestingly, Birchler et al. (2001) reported that regulatory genes associated with signal transduction and transcription tend to be dosage dependent and, therefore, also primary determinants of quantitative traits. This observation might help to explain why these classes of genes are selected to be maintained in the genome while other duplicates are deleted.

Blanc and Wolfe's analysis of gene expression data showed substantial divergence in patterns of gene expression among duplicate genes: more than one-half of the gene pairs estimated to have formed in the most recent polyploidy event showed significantly different patterns of gene expression. These results support prevailing models of the evolution of duplicated genes that hold that selection can be relaxed after gene duplication (because of functional redundancy), allowing for sequence divergence that may result in new tissue specificities or other functions (e.g., Ohno, 1970; Force et al., 1999).

Finally, the authors found evidence that duplicates resulting from polyploidy can diverge concertedly (i.e., there may be preferential retention or loss of an entire set of genes having interrelated functions, such as those encoding components of a particular biochemical pathway). Here, they looked for associations in the expression patterns between pairs of duplicates, such that the intrapair correlations were low (i.e., the paralogs were highly divergent) but the interpair correlations were high (i.e., duplicate members of several different genes diverged from their paralogs in the same direction). They found 37 concerted divergence associations of this type, involving 30 distinct pairs of duplicate genes. They note that there is very little experimental information available regarding

IN THIS ISSUE

the functions of most of the genes identified in these putative coevolving clusters, and the idea that such genes constitute interacting networks is based on the concept that genes having highly similar expression patterns are likely to be involved in the same biological pathway. A few cases of possibly relevant functional associations between genes undergoing concerted divergence were noted based on putative gene functions reported in the literature, including a set of genes whose function might be linked to senescence and a set linked to membrane function.

The authors argue that simultaneous duplication of all genes gives to polyploidy a singular evolutionary role with respect to other types of gene duplication mechanisms (i.e., duplications resulting from unequal crossing over or transposon activity) because it may allow the concerted evolution of entire pathways toward new functions or patterns of expression. These two reports by Blanc and Wolfe provide evidence that polyploidy is indeed widespread among angiosperms and further support the notion that a major ramification of polyploidy is functional divergence of duplicated genes, which may occur in concerted fashion among groups of interrelated genes after a polypoidy event.

Nancy A. Eckardt
News and Reviews Editor
neckardt@aspb.org

REFERENCES

Birchler, J.A., Bhadra, U., Bhadra, M.P., and Auger, D.L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: Implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev. Biol. **234,** 275–288.

Blanc, G., Hokamp, K., and Wolfe, K.H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. **13,** 137–144.

Blanc, G., and Wolfe, K.H. (2004a). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16,** 1667–1678.

Blanc, G., and Wolfe, K.H. (2004b). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16,** 1679–1691.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422,** 433–438.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151,** 1531–1545.

Gepts, P. (2003). Ten thousand years of crop evolution. In Plants, Genes, and Crop Biotechnology, M.J. Chrispeels and D.E. Sadava, eds (Sudbury, MA: Jones and Bartlett), pp. 328–359.

Koch, M.A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol. Biol. Evol. **17,** 1483–1498.

Li, W.H. (1997). Molecular Evolution. (Sunderland, MA: Sinauer Associates).

Lynch, M. (1997). Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. Mol. Biol. Evol. **14,** 914–925.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151–1155.

Ohno, S. (1970). Evolution by Gene Duplication. (London: George Allen and Unwin).

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. Annu. Rev. Genet. **34,** 401–437.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **99,** 13627–13632.

Stephens, S.G. (1951). Possible significance of duplication in evolution. Adv. Genet. **4,** 247–265.

Vandepoele, K., Simillion, C., and Van de Peer, Y. (2003). Evidence that rice and other cereals are ancient aneuploids. Plant Cell **15,** 2192–2202.

Vision, T.J., Brown, D., and Tanksley, S.D. (2000). The origins of genomic duplications in *Arabidopsis*. Science **290,** 2114–2117.

Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol. Biol. **42,** 225–249.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2,** 333–341.

Zhang, L., and Gaut, B.S. (2003). Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome Res. **13,** 2533–2540.