

Plant conserved non-coding sequences and paralogue evolution

Steven Lockton and Brandon S. Gaut

Department of Ecology and Evolution, University of California, Irvine, CA 92697, USA

Genome duplication is a powerful evolutionary force and is arguably most prominent in plants, where several ancient whole-genome duplication events have been documented. Models of gene evolution predict that functional divergence between duplicates (subfunctionalization) is caused by the loss of regulatory elements. Studies of conserved non-coding sequences (CNSs), which are putative regulatory elements, indicate that plants have far fewer CNSs per gene than mammals, suggesting that plants have less complex regulatory mechanisms. Furthermore, a recent study of a duplicated gene pair in maize suggests that CNSs are lost in a complementary fashion, perhaps driving subfunctionalization. If subfunctionalization is common, one expects duplicate genes to diverge in expression; recent microarray analyses in *Arabidopsis thaliana* suggest that this is the case. Plant genomes are relatively complex on a genomic level because of the prevalence of whole-genome duplication and, paradoxically, subfunctionalization after duplication can lead to relatively simple regulatory regions on a per gene basis.

Introduction

Gene duplication is a major determinant of the size and gene complement of eukaryotic genomes. Perhaps the most spectacular method of gene duplication is whole-genome duplication via polyploidization, which has a major role in the evolutionary history of eukaryotes. For example, the yeast *Saccharomyces cerevisiae* contains numerous duplicated genes and chromosomal regions that are attributed to a polyploidy event ~100 million years ago (mya) [1,2]. Similarly, the human genome retains vestiges of duplication events that occurred between 350 and 650 mya and that might be attributable to at least one polyploid event [3].

Genome duplication is particularly prominent in plants. *Arabidopsis thaliana* has experienced at least three ancient polyploid events [4–6]; rice (*Oryza sativa*) contains duplicated chromosomal regions that are attributable either to ancient segmental duplications [7] or to a paleopolyploid event [8,9]; genetic maps of maize (*Zea mays* ssp. *mays*) contain evidence of several large-scale duplication events [10]. More recently, Blanc and Wolfe [11] used EST data to investigate the number and relative age of duplicated genes in 14 plant species. Nine of these species contained evidence of ancient large-scale

duplication events, reflecting at least seven paleopolyploid events in the phylogenetic history of the sample. At least 16 polyploid events have been documented during the evolutionary history of a relatively small sample of angiosperm taxa (Figure 1).

The inescapable conclusion is that the organization and evolution of plant genomes have been shaped by many recent and ancient polyploid events. By contrast, vertebrates have probably undergone only one or two large-scale genome duplication events throughout their ~500 million year history [3,12,13]. With a few exceptions (such as amphibians [14]), extant polyploids are also rare. In mammals, for example, the only known polyploid is the tetraploid red viscacha rat of Argentina [15].

The relative frequency of genome duplication in plant and animal lineages affects their genome organization but might also have profound effects on gene function and regulation. Gene duplication has long been considered a crucial step in ‘freeing’ single-copy genes from selective constraint, enabling them to evolve new functions [12], but a pair of duplicated genes can also diverge in function as a result of changes in regulatory elements [16]. These observations raise a number of questions: What are the consequences of genome duplication for the regulatory complexity of plant genomes? Are differences in regulatory

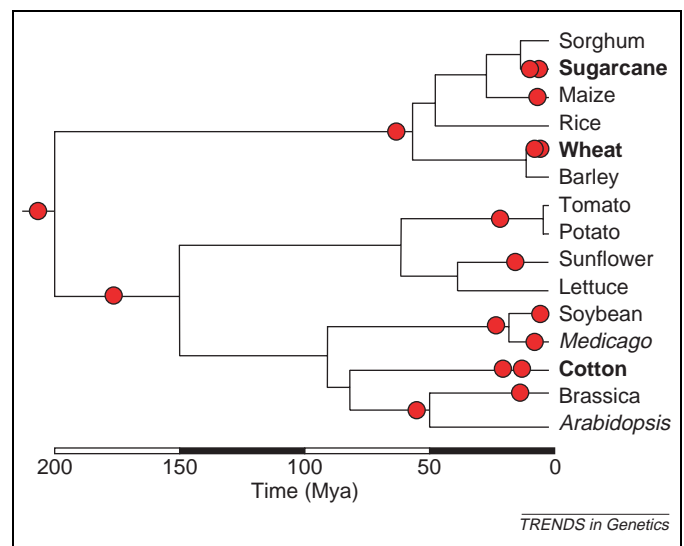


Figure 1. The phylogeny of a representative sample of angiosperms. Polyploid events represented by red circles and extant polyploid species are indicated in bold. The data used in this figure was obtained from Paterson *et al.* [9] and from Blanc and Wolfe [11], but note that dates for polyploid events are sometimes inconsistent in the two papers. For example, Blanc and Wolfe [11] predict a polyploid event in the *Arabidopsis* lineage ~25 million years ago (mya).

Corresponding author: Brandon S. Gaut (bgaut@uci.edu).

Available online 25 November 2004

motifs – more specifically conserved non-coding sequences (CNSs) – evident between plant and animal genomes? How does genetic duplication affect regulatory elements, and what are the consequences for the evolution of promoter regions between paralogous genes? We address these questions in this article.

The complexity of plant and animal CNSs

CNSs are short stretches of non-coding DNA that have been preserved between species. Such conservation might be indicative of selective constraint and hence function. CNSs are found predominantly in upstream regulatory regions and are enriched for sequences that perform regulatory functions [17,18]. Indeed, CNSs have a functional role in gene expression [19], and it is thought that they are assembly points for large, multi-protein complexes that perform gene-regulatory functions [18]. In the past, CNSs were identified on a single DNA sequence using pattern-recognition programs, with limited success [20]. These putative regulatory elements are now usually detected using computational tools that exploit the power of cross-species comparative genomics [20,21]. Another type of non-coding DNA, conserved non-genic sequences (CNGs) are similar to CNSs in that they are conserved through evolution. However, CNGs do not cluster near genes but can have regulatory functions via long-range chromosomal interactions [22].

Recent surveys of plant and animal genomes have shown remarkable differences in the size and quantity of their CNSs. A suitable comparison is among human, mouse, rice and maize genomes. Mice and humans diverged ~75 mya [23]; maize and rice diverged from their common ancestor 50–70 mya [24,25]. The synonymous nucleotide-substitution rate of plant and mammalian nuclear genes is also similar, at $\sim 6.5 \times 10^{-9}$ and 4.0×10^{-9} substitutions per site per year, respectively [26,27]. Because the two plants and the two mammals diverged at roughly the same time, and evolved at roughly the same rate, it is reasonable, as a first approximation, to compare their CNS patterns. In maize and rice, there is, on average, three CNSs per gene but CNSs could not be detected in 27% of genes studied [17]. Conversely, mammalian genes possess an average of 17.7 CNSs [18], and all of the mammalian genes studied to date possess CNSs [17]. Plant CNSs are smaller in size. The average length of CNSs in maize and rice orthologues is <12 bp [26], and in a sample of seven plant genes none possessed a CNS that was >60 bp. Twenty two CNSs between 60- and 99-bp long and 12 CNSs that were ≥ 100 bp were found in six mammalian genes [18].

It is important to note that our comparison of two mammals with two cereals might not hold for plants and animals as a whole. Furthermore, some differences between mammals and grasses could be due to sampling; in mammals, there is far more sequence data available to enable detection of CNSs. Another important caveat is that methods of CNS detection differ among studies; however, methodology alone does not appear to explain the differences between plants and animals. When the same methods were applied to cereals and mammals, substantial differences remained [18].

Therefore, there appears to be profound differences between mammalian and grass CNSs. Assuming that CNSs reflect regulatory complexity, there are two possible reasons for these differences. First, the increased amount of gene and genome duplication in grasses could affect the evolution of regulatory regions. This possibility will be discussed in detail in the next section. Second, relatively simple regulatory regions in plants could reflect differences in developmental and organismal complexity. No matter how one defines organismal complexity (e.g. the number of cell lines, tissue types, organs and organ-systems or connections between cells and tissues), mammals are more complex than plants. Mammals are also strongly developmentally regulated, with terminally differentiated cells, as opposed to plants, which have totipotent cells. Finally, there are relatively few reports of alternative splicing in plants [28], suggesting that individual plant genes might not have as many functions as those in mammals. Therefore, it is possible that mammalian genes require more regulatory motifs to regulate development and to function properly, for example, to down-regulate genes in differentiated cell lines [17]. In this context, it is interesting to note that mammals have less genome duplication, greater organismal complexity and more CNSs per gene compared with plants.

CNSs in plant paralogues

In addition to organismal complexity, gene duplication can also contribute to the substantial differences between CNSs in cereal and mammalian genomes. Gene duplication is an important driving force for generating evolutionary novelty. Ohno's classical model of gene duplication [12] states that a gene under tight functional constraint is 'freed' from selection pressures once a duplication event creates a redundant copy. This liberated gene copy has two potential fates: it can acquire a new adaptive function via mutation ('neofunctionalization') or can be lost by accumulating deleterious mutations, eventually becoming a pseudogene ('nonfunctionalization'). Force and colleagues proposed a further mechanism: their 'duplication-degeneration-complementation' (DDC) model argues that complementary, deleterious mutations in regulatory elements between a duplicate gene pair partitions ancestral gene function, creating two divergent genes possessing different 'sub-functions' of the ancestral gene [16].

In theory, subfunctionalization partitions regulatory functions between two daughter genes. Because there is more genome duplication in grasses compared with mammals, there is also more opportunity for subfunctionalization to occur, potentially resulting in a greater rate of CNS loss per gene. If this is true, there should be more genes in plant genomes and more genes per gene family, but fewer CNSs per gene. Conversely, in mammals there has been little genome duplication, and therefore relatively little possibility for subfunctionalization. A comparison of the number of genes in gene families among several sequenced genomes supports this viewpoint (Figure 2). *Arabidopsis* and rice have the highest proportion of genes that are members of gene families.

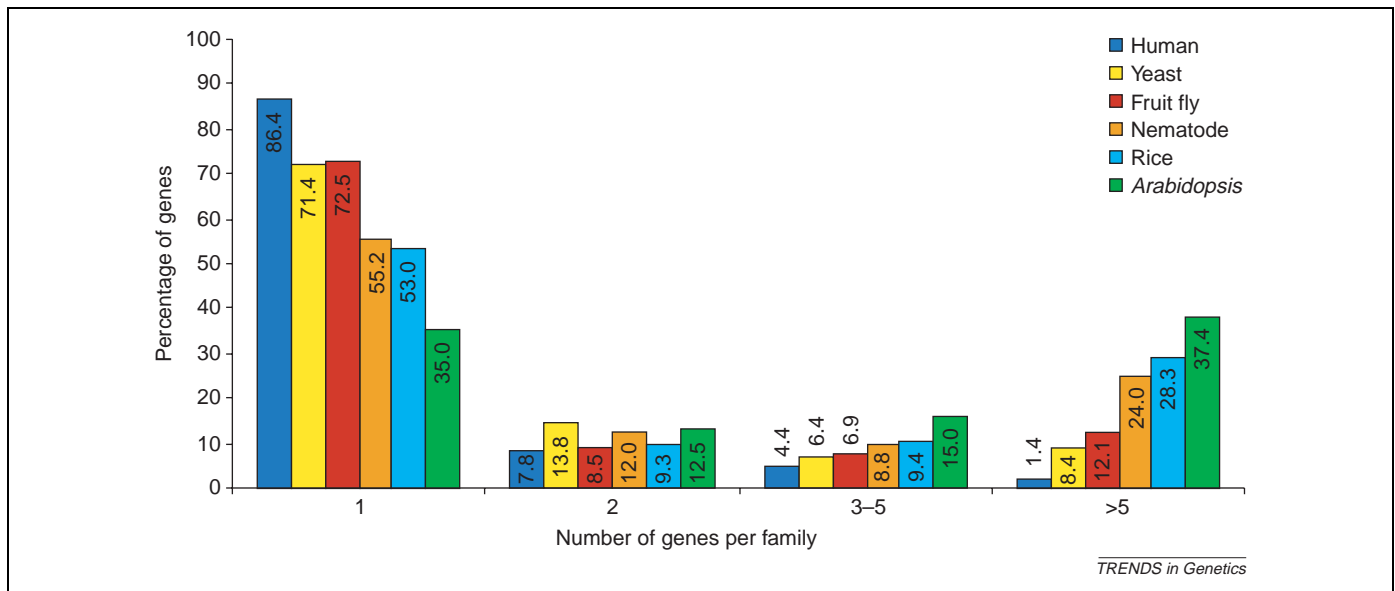


Figure 2. The percentage of genes present as either singletons or in gene families in the genomes of humans, yeast, nematode, rice, *Drosophila* and *Arabidopsis*. Data for yeast, *Drosophila*, nematode and *Arabidopsis* were taken from Ref. [29], based on the definition that gene family members yield a BLASTp E-value $< 1e-20$ and align over 80% or more of the protein length. The rice data are unpublished (C. Rizzon and B.S. Gaut, unpublished) but are based on a gene-family definition that produces nearly identical results to the definition used in Ref. [29] (data not shown). The human data are from reference Ref. [30]. We report results based on the least stringent gene-family definition employed in Ref. [30], but it is unclear whether this definition is less or more stringent than that used Ref. [29].

Only 35% of *Arabidopsis* genes are singletons (single-copy) and $>37\%$ are members of a gene family that consists of five or more members [29]. However, in the human genome, singletons represent $>77\%$ of genes, with only 0.4% of genes in large (more than five genes) multigene families [30]. The proportion of human [30] and rice genes in gene families was defined using different criteria from those of *Arabidopsis*, *Drosophila*, yeast and nematode [29] (Figure 2). Nonetheless, the available evidence suggests that plants contain more genetic duplicates than other organisms [31] probably as a result of polyploidy and mechanisms such as tandem duplication. Paradoxically, plants could be relatively simple at the level of individual gene regulation for the same reason.

Comparative genomics has traditionally looked at CNSs only between orthologous genes, but these arguments suggest that it is important to study CNS patterns between paralogues. Accordingly, Langham and colleagues sought evidence of CNS subfunctionalization between maize paralogues [32]. They coined two terms to describe the patterns they found (Figure 3). The first, 'fractionation', describes the loss of functioning DNA sequence by mutation between duplicated regions. These regions are lost in a fashion that is similar to that described in the DDC model, and hence fractionation is the process by which subfunctionalization occurs. However, fractionation is not restricted to single genes. The scale of fractionated regions can range from chromosomal segments, to genes on a chromosome, to CNSs within a gene. The second term, 'consolidation', does not describe any evolutionarily event, but is an intellectual process that aims to reconstruct the pre-duplication state of a putative ancestral DNA molecule. If fractionation is the complementary loss of sequence after a duplication event, then all of the original pre-duplicated DNA should be present between the two copies. Consolidation, as the term

suggests, reunites the two fractionated sequences and, in effect, reverses fractionation. Once a fractionated sequence is consolidated (for example, between two maize duplicates), the underlying synteny with a non-duplicated relative (e.g. rice) becomes apparent [33], facilitating the identification of genes and CNSs by comparative genomics.

Langham and colleagues sequenced two maize bacterial artificial chromosomes (BACs) containing *liguleless2* (*lg2*; a well-characterized gene encoding a leucine zipper protein, on chromosome 3), and its duplicate

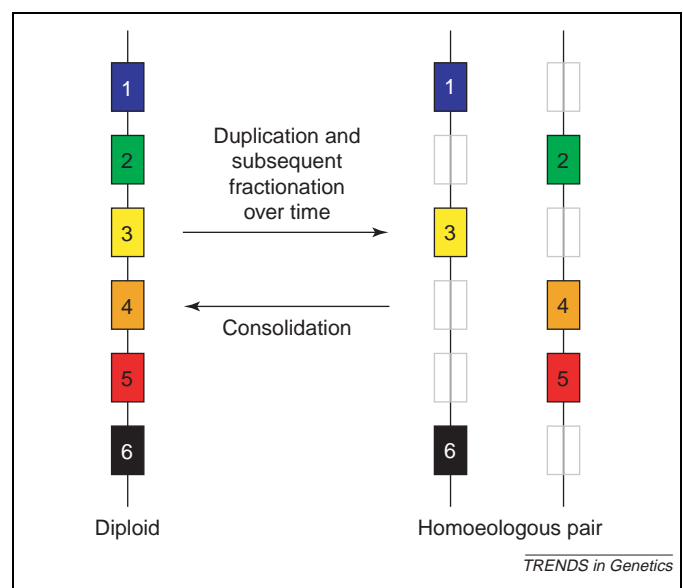


Figure 3. Diagram depicting the result of fractionation and the concept of consolidation. Coloured boxes indicate functional genomic components, which could be genes on a chromosome or *cis*-acting regulatory elements within a gene. White boxes indicate lost functional elements. In this hypothetical example, 100% of the functional genomic components are fractionated, but at least one copy of each component was retained after fractionation.

liguleless related sequence1 (lrs1; on chromosome 8). The two BACs contained homoeologous regions possessing 13 genes in total. After two of these fractionated regions were consolidated, there was excellent synteny with their homologous rice chromosomal region. Apart from the duplicate pair, *lg2* and *lrs1* (which were used to identify the homoeologous region), no maize genes were found to be retained between the two sequences. At the chromosomal level, and at the gene level in the two homoeologous regions, 'fractionation' was almost complete. We should note that such fractionation of genes is relatively common. In *Arabidopsis*, for example, only ~23% of genes that are duplicated in large-scale duplication events are retained as pairs. For the remaining 77% of genes, one of the two copies has been lost [4].

More importantly, Langham and coworkers also examined the pattern of fractionation at the level of CNSs within the *lg2-lrs1* duplicate gene pair. To date, *lg2* is the most CNS-rich grass gene that has been identified, with ~30 CNSs. Comparing both maize paralogues, *lg2* and *lrs1*, with their single-copy orthologues in rice, they found that *lg2* had lost one CNS, whereas *lrs1* had lost two of the 30 CNSs. They proposed that these three CNSs have fractionated since the most recent polyploidy event in maize 11 mya [34], and that subfunctionalization is ongoing between *lg2* and *lrs1*. Additional studies of CNS evolution between paralogues will provide important insights into the regulatory fate of duplicated genes.

Fractionation: functional biases and the evolution of gene expression

Given the prevalence of genome duplication in plants, fractionation has a crucial role in shaping the functional complement of plant genomes. However, surprisingly little is known about the patterns and processes of fractionation. For example, some studies indicate that genome rearrangement occurs rapidly after polyploidization, suggesting that gene loss is initially rapid but eventually slows [35,36]. Nonetheless, we do not know how widely this applies or the long-term rates of gene loss in plant genomes. It is also important to identify the genes that are lost during fractionation. Is the process of fractionation biased toward particular genes and gene functions? Finally, the DDC model predicts that fractionation occurs between regulatory regions of duplicated genes. If this is true, then the patterns of gene expression of duplicated genes should also diverge. Is there any evidence that this is true?

In a remarkable *tour de force*, Blanc and Wolfe [37] have answered some of these questions using *Arabidopsis* as a model system. They first asked whether there was any functional bias regarding the retention and loss of duplicated genes. To address this question, they used the Gene Ontology (GO) database [38] (<http://www.geneontology.org/>), the Munich Information Centre for Protein Sequences (MIPS; <http://mips.gsf.de/>) and the *Arabidopsis thaliana* database (MatDB; <http://mips.gsf.de/proj/thal/db/index.html>) [39] to classify genes in duplicated chromosomal regions into functional categories. Given this information, they asked whether genes with particular functions were preferentially retained as a gene pair. With

respect to the most recent polyploid event in *Arabidopsis*, they found that retained genes are not evenly distributed. Genes that encode proteins involved in signal transduction and transcription have been retained more often than is expected by chance. Conversely, the majority of genes involved in DNA repair, defence and apoptosis, among others, tend to be single copy; their post-polyploid paralogue has been lost. Overall, Blanc and Wolfe [37] discovered that genes encoding proteins located in the nucleus and plasma membrane are preferentially retained, and that genes encoding proteins in organelles are preferentially lost.

Addressing the issue of subfunctionalization between duplicated genes, Blanc and Wolfe [37] next asked whether gene expression diverged between the genes that are retained as duplicates, using a dataset obtained from 62 Affymetrix chip analyses from a series of environmental conditions and tissues. They examined the expression profiles of 1137 'recently' duplicated genes (from a relatively recent paleopolyploid event) and 420 'older genes' (from one or more relatively ancient paleopolyploid events). To measure the divergence of gene-expression pattern between duplicated genes, they calculated the Pearson correlation coefficient (r). In theory, paralogous genes will have a correlation coefficient of 1.0 immediately after a polyploid event. As expression profiles diverge, the r -value decreases.

First they measured r for all duplicated gene pairs. Next, they measured r for 10 000 randomly chosen, non-duplicated genes to determine a statistically significant cut-off point. At this cut-off point, a pair of duplicated genes was considered significantly divergent in their gene expression patterns. For the 10 000 randomly chosen gene pairs, 95% of the correlation coefficients were <0.52 ; therefore, any pair of genes with an expression profile of $r < 0.52$ was considered to have a diverged expression pattern. Using this criterion, Blanc and Wolfe discovered that 57% of the recently duplicated genes and 73% of genes from the older polyploid events have diverged in expression [37]. Thus, the majority of duplicated gene pairs retained from paleopolyploid events have diverged in expression pattern, suggesting that fractionation of regulatory regions can occur.

These observations extend a previous study of polyploid cotton that documented rapid subfunctionalization of homoeologous genes [40]. For example, 25% (10 of 40) of the assayed cotton gene pairs had one copy that was either down regulated or silenced in ovule tissue. In cotton, epigenetic mechanisms can contribute to divergence in gene-expression patterns. However, one cannot determine whether expression divergence between *Arabidopsis* duplicates is due to epigenetic phenomena or DNA sequence evolution of *cis*-regulatory regions. Further study of CNSs in paralogues, combined with expression data, will be insightful.

Changes in expression patterns are likely to affect genes within the same biochemical or regulatory network. To extend their observations to networks, Blanc and Wolfe [37] focused on 248 recently duplicated genes with divergent expression patterns. They speculated that many of the 248 recently duplicated genes must be

members of regulatory networks. If the expression pattern of one paralogue has changed relative to its duplicate, the expression pattern of its interacting genes should change in a correlated fashion. The ultimate evolutionary result would be two parallel gene networks with different expression patterns. To identify such an event, the authors identified duplicate genes with highly diverged ($r \ll 0.1$) expression patterns and also identified genes with correlated ($r > 0.7$) patterns of expression relative to one of the two duplicates. Genes with high correlation coefficients are likely to be co-regulated and, hence, are more likely to belong to the same gene network or biological pathway. Using these criteria, Blanc and Wolfe identified 37 examples of concerted divergence of expression involving 30 pairs of genes. These genes could be organized into six parallel networks, one of which contained 13 genes. Thus, Blanc and Wolfe demonstrated that divergent expression between duplicated genes is not always an isolated event; it can affect entire networks of genes, with potentially dramatic evolutionary consequences [37].

Future directions

It is not an exaggeration to say that 100% of plants are either polyploid or have an evolutionary history of paleopolyploidy. However, the timing and phylogenetic placement of these events (Figure 1), will continue to be a focus of genetic research. The knowledge gained will have important practical applications. In the cereals, for example, there is substantial interest in isolating genes for agronomic traits from 'small genome' crops such as rice and extending knowledge to 'large genome' crops, for example, maize and wheat. To improve the chances for this approach to succeed, the timing and extent of genome duplications, in addition to their effects on genome organization, need to be elucidated.

During evolution the processes of fractionation and divergence were important, both as sources of novel gene function and as sources of functional redundancy. Although we have some partial answers on important issues from *Arabidopsis*, much more work is required in several areas. First, the functional biases that occur during the fractionation process need to be understood. Are the functional biases in *Arabidopsis* consistent with other plant taxa? Undoubtedly, bioinformatic analysis of the rice genome will help to provide the answer to this question. Second, what is the pattern of sequence evolution in genes that are retained as duplicates? Some studies have hinted that genes in the duplicated pair evolve at different rates [2,37,41] with occasional signatures of positive selection [42]. If positive selection is common, neofunctionalization could be widespread. Third, although there are hints that neofunctionalization can occur, there are not many documented examples of functional recruitment after gene duplication [43,44]. At some point, bioinformatic and comparative analyses will need to be superseded by enzymatic and chemical assays to prove that duplicated genes have diverged in function. Until this occurs on a large scale, we can only glean inferences about neofunctionalization indirectly.

We also need to understand the evolution of CNSs and the evolution of *cis*-regulatory regions. Thus far, few plant genes have been examined for CNSs. Although the sample is sufficient to make some broad conclusions – that cereal and mammal CNSs differ substantially – far more work is required. Until these characterizations are made, we cannot know if the distribution of CNSs follows a general pattern in plants or differs substantially between monocots (e.g. the cereals) and dicots (e.g. the Brassicaceae). Furthermore, with the exception of Langham *et al.* [32], there is little information about the effect of gene duplication on CNSs and the evolution of paralogues. Once characterized, CNSs need to be related to patterns of gene expression and sequence divergence. These issues have not been addressed in plants, but they have in yeast. For example, Papp and colleagues [45] observed a significant negative relationship between the age of duplicated gene pairs (as measured by synonymous substitution rates) and the number of regulatory motifs shared between duplicated yeast genes. Although this pattern is consistent with subfunctionalization, Zhang *et al.* [46] found that differences in the number of regulatory motifs between paralogues correlate only weakly with patterns of gene expression. Is this also true for plant systems? Even if it is not, eventually, CNSs will need to be characterized functionally, either by transgenic 'promoter bashing' or perhaps by hybrid analysis [47].

A potent source of gene duplication has been almost completely ignored in the comparative genomics literature: tandemly arrayed genes. Most studies involving duplicated genes have focused on larger-scale polyploid, aneuploid or segmental duplication events; however, tandemly arrayed genes are probably as potent a source for neofunctionalization, subfunctionalization and CNS evolution as chromosomal duplicates. In *Arabidopsis*, for example, tandemly arrayed genes comprise up to 18% of all genes [48], whereas genes duplicated by polyploidy events represent <27% of genes [49]. There have been several isolated examples of functional and evolutionary analysis of a single tandem array (e.g. Clauss *et al.* [50]), but a broader focus on the generation and evolution of tandem arrays is merited. Altogether, gene duplication – whether via polyploidy or tandem arrays – is a prominent evolutionary process in plant genomes. Continued study of the evolution of the regulation, and regulatory regions, of duplicated genes will provide fundamental insights into the processes governing functional and taxonomic divergence.

Acknowledgements

We are grateful for the comments of two anonymous reviewers. This work is supported by NSF grants DEB-0316157 and DBI-0321467 to B.S.G.

References

- 1 Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713
- 2 Kellis, M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624
- 3 McLysaght, A. *et al.* (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31, 200–204

- 4 Vision, T.J. *et al.* (2000) The origins of genomic duplications in *Arabidopsis*. *Science* 290, 2114–2117
- 5 Simillion, C. *et al.* (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13627–13632
- 6 Bowers, J.E. *et al.* (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438
- 7 Vandepoele, K. *et al.* (2003) Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15, 2192–2202
- 8 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100
- 9 Paterson, A.H. *et al.* (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9903–9908
- 10 Hampson, S. *et al.* (2003) LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res.* 13, 999–1010
- 11 Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678
- 12 Ohno, S. (1970) *Evolution by gene duplication*, Springer-Verlag
- 13 Postlethwait, J.H. *et al.* (1998) Vertebrate genome evolution and the zebrafish gene map. *Nat. Genet.* 18, 345–349
- 14 Soltis, D.E. and Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* 14, 348–352
- 15 Gallardo, M.H. *et al.* (1999) Discovery of tetraploidy in a mammal. *Nature* 401, 341
- 16 Force, A. *et al.* (1999) Preservation of duplicate genes by complementary degenerative mutations. *Genetics* 151, 1531–1545
- 17 Inada, D.C. *et al.* (2003) Conserved noncoding sequences in the grasses. *Genome Res.* 13, 2030–2041
- 18 Kaplinsky, N.J. *et al.* (2002) Utility and distribution of conserved noncoding sequences in the grasses. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6147–6151
- 19 Frazer, K.A. *et al.* (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* 14, 367–372
- 20 Loots, G.G. *et al.* (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12, 832–839
- 21 Dubchak, I. *et al.* (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* 10, 1304–1306
- 22 Dermitzakis, E.T. *et al.* (2004) Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* 14, 852–859
- 23 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
- 24 Wolfe, K.H. *et al.* (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 86, 6201–6205
- 25 Kellogg, E.A. (2001) Evolutionary history of the grasses. *Plant Physiol.* 125, 1198–1205
- 26 Guo, H. and Moose, S.P. (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* 15, 1143–1158
- 27 Li, W.-H. and Graur, D. (1991) *Fundamentals of Molecular Evolution*, Sinauer Associates
- 28 Dralyuk, I. *et al.* (2000) ASDB: database of alternatively spliced genes. *Nucleic Acids Res.* 28, 296–297
- 29 *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
- 30 Li, W.H. *et al.* (2001) Evolutionary analyses of the human genome. *Nature* 409, 847–849
- 31 Szathmary, E. *et al.* (2001) Molecular biology and evolution. Can genes explain biological complexity? *Science* 292, 1315–1316
- 32 Langham, R.J. *et al.* (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945
- 33 Freeling, M. (2001) Grasses as a single genetic system: reassessment 2001. *Plant Physiol.* 125, 1191–1197
- 34 Gaut, B.S. and Doebley, J.F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6809–6814
- 35 Song, K. *et al.* (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implication for polyploid evolution. *Proc. Natl. Acad. Sci. U. S. A.* 92, 7719–7723
- 36 Ozkan, H. *et al.* (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* 13, 1735–1747
- 37 Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by Polyploidy during *Arabidopsis* evolution. *Plant Cell* 16, 1679–1691
- 38 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
- 39 Schoof, H. *et al.* (2002) MIPS *Arabidopsis thaliana* database (MAtdB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.* 30, 91–93
- 40 Adams, K.L. *et al.* (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4649–4654
- 41 Zhang, L. *et al.* (2002) Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 19, 1464–1473
- 42 Carginale, V. *et al.* (2004) Adaptive evolution and functional divergence of pepsin gene family. *Gene* 333, 81–90
- 43 Ober, D. and Hartmann, T. (1999) Homospermidine synthase, the first pathway-specific enzyme of pyrrolizidine alkaloid biosynthesis, evolved from deoxyhyppusine synthase. *Proc. Natl. Acad. Sci. U. S. A.* 96, 14777–14782
- 44 Tropf, S. *et al.* (1994) Evidence that stilbene synthases have developed from chalcone synthases several times in the course of evolution. *J. Mol. Evol.* 38, 610–618
- 45 Papp, B. *et al.* (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet.* 19, 417–422
- 46 Zhang, Z. *et al.* (2004) How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet.* 20, 403–407
- 47 Wittkopp, P.J. *et al.* (2004) Evolutionary changes in *cis* and trans gene regulation. *Nature* 430, 85–88
- 48 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
- 49 Blanc, G. *et al.* (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13, 137–144
- 50 Clauss, M.J. and Mitchell-Olds, T. (2004) Functional divergence in tandemly duplicated *Arabidopsis thaliana* trypsin inhibitor genes. *Genetics* 166, 1419–1436