A Microarray Based Genomic Hybridization Method for Identification of New Genes in Plants: Case Analyses of *Arabidopsis* and *Oryza*

Chuanzhu Fan*, Maria D. Vibranovski, Ying Chen and Manyuan Long**

(*Department of Ecology and Evolution*, *The University of Chicago*, Chicago, IL 60637, USA)

Abstract

To systematically estimate the gene duplication events in closely related species, we have to use comparative genomic approaches, either through genomic sequence comparison or comparative genomic hybridization (CGH). Given the scarcity of complete genomic sequences of plant species, in the present study we adopted an array based CGH to investigate gene duplications in the genus *Arabidopsis*. Fragment genomic DNA from four species, namely *Arabidopsis thaliana*, *A. lyrata* subsp*. lyrata*, *A. lyrata* subsp*. petraea*, and *A. halleri*, was hybridized to Affymetrix (Santa Clara, CA, USA) tiling arrays that are designed from the genomic sequences of *A. thaliana*. Pairwise comparisons of signal intensity were made to infer the potential duplicated candidates along each phylogenetic branch. Ninety-four potential candidates of gene duplication along the genus were identified. Among them, the majority (69 of 94) were *A. thaliana* lineage specific. This result indicates that the array based CGH approach may be used to identify candidates of duplication in other plant genera containing closely related species, such as *Oryza,* particularly for the AA genome species. We compared the degree of gene duplication through retrotransposon between *O. sativa* and *A. thaliana* and found a strikingly higher number of chimera retroposed genes in rice*.* The higher rate of gene duplication through retroposition and other mechanisms may indicate that the grass species is able to adapt to more diverse environments.

Key words: *Arabidopsis*; comparative genomic hybridization; microarray; new genes; *Oryza*; retroposition.

Fan C, Vibranovski MD, Chen Y, Long M (2007). A microarray based genomic hybridization method for identification of new genes in plants: Case analyses of *Arabidopsis* and *Oryza*. *J. Integr. Plant Biol.* **49**(6), 915–926.

Available online at www.blackwell-synergy.com/links/toc/jipb, www.jipb.net

Received 9 Jan. 2007 Accepted 16 Feb. 2007

Supported by Fellowships from the Pew Latin American Fellows Program and Conselho Nacional de Desenvolvimento Cientfico e Tecnologico (to MDV), the USA National Science Foundation CAREER award (MCB0238168) and USA National Institutes of Health R01 grants (R01GM065429-01A1 and GM078070-01A1) to ML at the University of Chicago.

Publication of this paper is supported by the National Natural Science Foundation of China (30624808).

*Present address: Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA.

**Author for correspondence.

Tel: +1 773 702 0557;

Fax: +1 773 702 9740;

E-mail: <mlong@uchicago.edu>.

doi: 10.1111/j.1672-9072.2007.00503.x

In order to study the biological diversity of organisms, it is important to understand the mechanisms and evolutionary forces acting on the origin of novel genes. DNA duplication, exon-shuffling, retrotransposition, horizontal gene transfer, and gene origin mediated by mobile elements have been described as major molecular mechanisms involved in the creation of novel genes in different organisms (Ohno 1970; Gilbert 1978; Kimura 1983; Brosius 1991; Makalowski et al. 1994; Ochman 2001). Whole-genome duplication has been recognized as a major model for gene duplication and creation in plants (e.g. Wendel 2000). However, recent studies have demonstrated that gene duplication through transposon-mediated processes is also very common in plant genomes. For example, Wang et al. (2006) observed extensive retropositions that resulted in 1 235 primary retrogenes in the rice genome. Also in rice, DNA transposable elements carrying fragments of cellular genes ("pack-MULES") have been shown to be important in mediating gene evolution in plants by rearranging genomic sequences (Jiang et al. 2004).

A partial or completely duplicated gene is likely to become functionless owing to its sequence redundancy compared with the original copy (Kimura 1983). In contrast, a novel copy also can evolve into a functional gene (for a review, see Long et al. 2003). Direct observation and study of young genes helps us understand how the evolutionary process of the creation and fixation of functional genes occurs. Genes in their early stage of evolution present more detailed information about the mechanisms of their origination than older genes, for which these mechanisms have been normally lost over the course of the evolutionary process. One way to identify young genes is to study their phylogenetic distribution within a genus. Observation of different numbers of copies of a given gene within closely related species enables the identification of recently duplicated genes.

In order to systematically detect gene duplication events across different species, we have to rely on either large-scale experimental or comparative genomic approaches. Phylogenetic comparison of genetic signals (e.g. fluorescence *in situ* hybridization (FISH) and genomic Southern blotting) has proven to be an efficient and reliable way of identifying young protein-coding genes in *Drosophila* and mammals (Wang et al. 2004). However, these methods may not be feasible in plants owing to technical and labor-intensive issues. Comparative genomic sequence comparison is solely dependent on the availability of a complete genomic sequence for the closely related species. Unfortunately, there are only a few complete genomic sequences available for plants. Therefore, in the present study, an array based comparative genomic hybridization (CGH) was adopted to detect gene duplication events in plants.

Comparative genomic hybridization was developed to analyze the genome-wide DNA sequence number in a single experiment (Pinkel and Albertson 2005). By taking advantage of the whole-genome sequence data for model organisms, the microarray based CGH was further designed to survey variations in DNA copy number and expression (mRNA) across the entire genome, which provides an important tool for studying gene expression, DNA polymorphisms, and mutations that are caused by changes in the number of gene copies (Gilad and Borevitz 2006). Comparative genomic hybridizations can also be used to study large-scale differences between genomes of closely related species. This approach can reveal a subset of duplications relative to the genome on which the array is based. Moreover, this method tends to enable easier identification of recent duplications present in the single species or closely related monophyletic lineages.

In an effort to develop a more generally useful method to detect new gene candidates, we have adopted this technology to detect variations in the number of duplicate copies by gene gain in closely related species by hybridization with genomic DNA using the Affymetrix (Santa Clara, CA, USA) GeneChip oligo-array from model organisms (e.g. *Drosophila melanogaster*, *Arabidopsis thaliana*). The novel gene candidates then were deduced by applying the duplication events into a phylogenetics framework.

Until now, just a few plant species, such as *Oryza sativa*, *A. thaliana*, and the poplar tree have had their genomes completely sequenced, and *Oryza* and *Arabidopsis* seem to be potential genera to study recent novel genes generated by DNA duplication.

The *Arabidopsis* genus consists of eight species, all of which are indigenous to Europe, with two species extending into northern and eastern Asia, and North America, extending into the central US (Figure 1). The time at which *A. thaliana* diverged from other species was estimated to be approximately 5 million years ago (MYA) and sequence divergence is estimated within 5%–10% across eight species.

The *Oryza* genus comprises approximately 23 species with worldwide distribution and is represented by 10 genome groups (Khush 1997). Although phylogenetic relationships within the genus are not totally solved, a recent analysis based on the intron sequences of one of the subgroups, the AA genome group, estimated that these species diverged approximately 2 MYA (Zhu and Ge 2005; Figure 2), which provides an ideal species system for the identification of young gene duplications using CGH.

Although rice is a valuable food resource all over the world and has an important impact on the economy, *A. thaliana*, as a plant model organism, has been more extensively and intensely studied through different biological approaches. The genus has

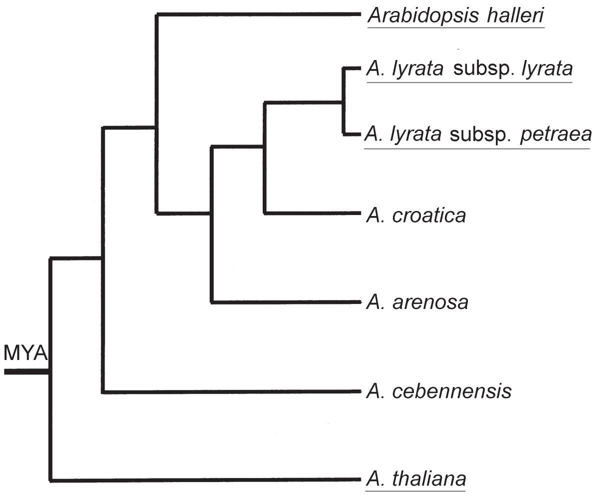

Figure 1. Phylogenetic relationships within *Arabidopsis*.

The species underlined were chosen for microarray hybridization. MYA, million years ago.

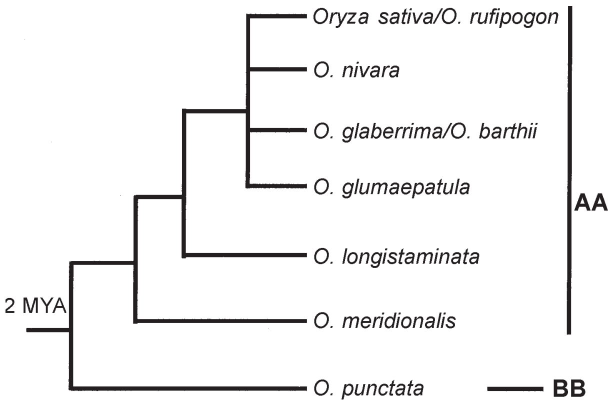

Figure 2. The phylogenetic tree of AA genome *Oryza* species rooted by the BB genome species *O. puctata*. MYA, million years ago.

been recently defined and delimited by molecular phylogeny combined with cytological, morphological, and ecological profiles. Moreover, the *Arabidopsis* species have some significantly different physiological and morphological characteristics, which should be companied with functional gene divergence leading to new gene origination. For example, the shift between selfing and outcrossing is a profound evolutionary transition, which would cause effects the extent of genetic diversity accompanied by gene gain and loss. In the *Arabidopsis* genus, the *A. thaliana* and *A. suecica* species are selfing, whereas others are self-incompatible.

As a consequence of this intense research, the *A. thaliana* genome has been completely sequenced before *O. sativa* and, therefore, contains a larger set of well-annotated genes (The Arabidopsis Initiative 2000; International Rice Genome Sequencing Project 2005). In addition, large-scale genetic tools based on genomic sequence, such as microarray, are more developed for *A. thaliana* than *O. sativa*. For example, Affymetrix has two different types of *A. thaliana* gene chip arrays, including a second-generation array (ATH1) and one high-density tiled array designed for the identification of novel transcripts and genes. In contrast, the only *O. sativa* microarray available from Affymetrix is a first-generation gene chip array (for more information, see http://www.affymetrix.com/products/arrays/specific/).

For the reasons detailed above, we decided to use microarray based CGH to search for potential candidates of duplicated genes in the *Arabidopsis* species. The results from the *Arabidopsis* CGH study suggested that the method could be used to identify candidates of duplication in other plant genera containing closely related species, such as *Oryza,* particularly for the AA genome species.

Results

Determination of the threshold value for duplication events

We were attempting to detect differences in gene copy numbers in a comparative analysis of different species using DNA microarray hybridization with fractionated genomic DNA. The first important technical question was to find the cut-off intensity difference or ratio to detect duplicate copies from the analysis. A conventional practice is to assume a ratio of 2 (the ratio of the intensity from a duplicate *vs* single copy sequence) as a criterion to identify DNA duplication. Using the Affymetrix GeneChip System, our group conducted a pilot experiment that detected the intensity ratio for a *Drosophila* line (z[1]w[118];Dp(1;2)w[+]70h; Bloomington Drosophila Stock Center, Bloomington, Indiana, USA) that has a known duplicate region involving 31 genes in regions 003A07-08/003C02-03 and 031A03 compared with wild-type Canton S (Bloomington Drosophila Stock Center; J Emerson et al., unpubl. data, 2003). This experiment showed that the ratio of 2 is too conservative a criterion. Instead, the ratio distribution for the duplicates peaks is approximately 1.3–1.5. Therefore, we applied this empirical observation as a basis to detect new gene duplicates in different species. So far, we have characterized a number of candidate genes using this approach in terms of their gene structure, expression, and evolutionary genetics (Fan and Long 2006).

Selection of duplicate gene candidates

In order to find putative duplicate genes (among four *Arabidopsis* species) in a particular species or branch of the *Arabidopsis* genus (Figure 1), we used Affymetrix tiling array platforms to hybridize entire genomes. These arrays contain adjacent probes from an entire genome and are spotted for hybridization. For example, having a tiling array of a given species, it is possible to hybridize DNA from species in the same genus to study evolutionary features.

In the present study, we individually hybridized DNA from four *Arabidopsis* species (*A. thaliana*, *A. halleri* and two subspecies from *A. lyrata*, namely *A. lyrata* subsp. *lyrata* and *A. lyrata* subsp. *petraea*) against a tiling array of the genome of *A. thaliana* to identify possible cases of gene duplication between species. Duplication of a given genomic region in a given species, in this case of genes, would appear as an intensification of its hybridization signal compared with the hybridization signal of other species (see details below).

In order to obtain a higher-quality data set of *A. thaliana* genes to be probed, decreasing false positives due to spurious hybridizations, we first filtered our set of genes. First, the 27 922 genes contained in the *A. thaliana* tiling array were filtered for

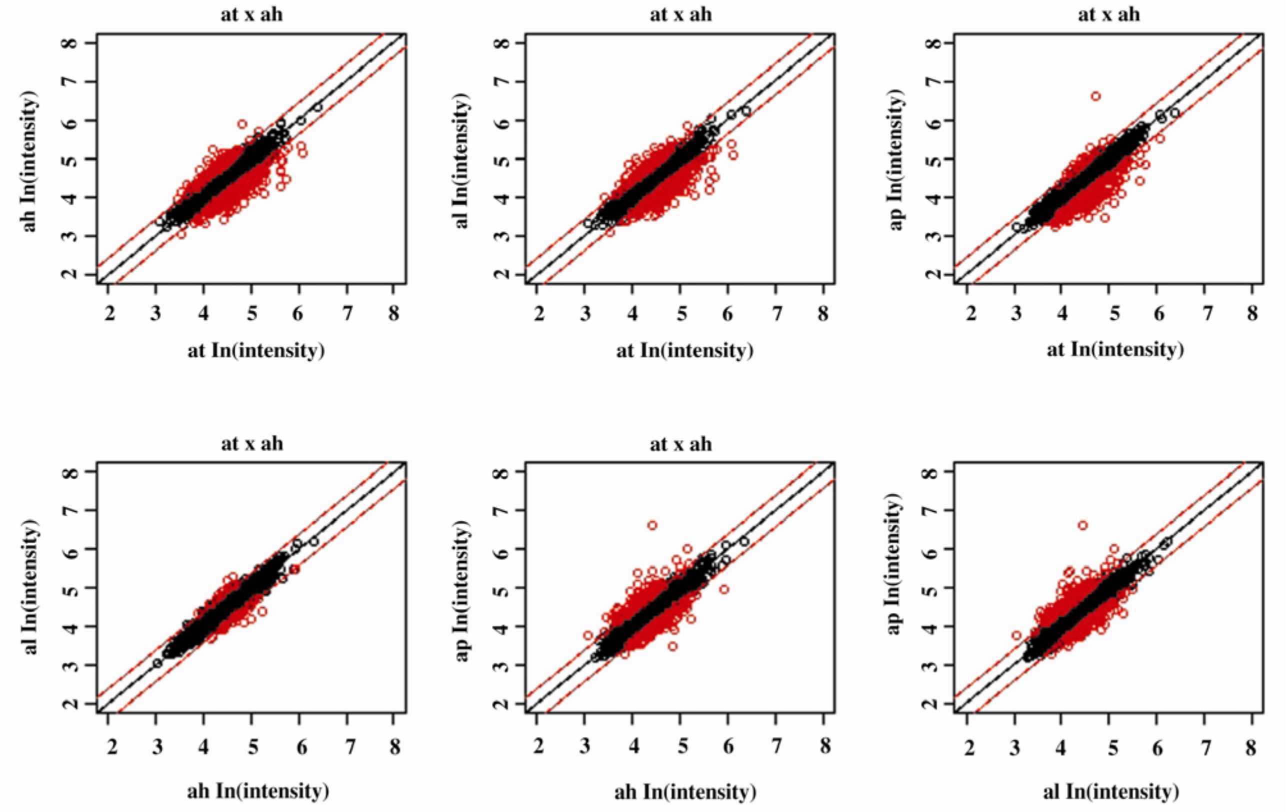

Figure 3. Plot of species pair-wise comparison of natural log (ln) intensity produced by hybridization against *Arabidopsis thaliana* tiling array.

at, *A. thaliana*; ah, *A. halleri*; al, *A. lyrata* subsp. *lyrata*; ap, *A. lyrata* subsp. *petraea*. The expected 1.5-fold difference in intensity is shown as red lines; genes presenting 1.5-fold difference are indicated in red.

transposable elements by looking for words in their description that indicated the presence of these elements, yielding 25 802 genes (see **Materials and Methods**). Second, because simultaneous hybridization of multiple probes from the same gene can more reliably indicate that the gene is also present in another genome, genes that were represented by only one probe (*n* = 477) were also discarded, leading to a filtered dataset of 25 325 genes.

The resulting mean hybridization intensities from three replicates for each of the 25 325 genes were then plotted for pairs of organisms (i.e. data for all species were plotted against each other, including *A. thaliana* self-hybridization signals; Figure 3). We found 1 206 genes exhibiting a 1.5-fold or higher ratio of hybridization intensities in at least one of the six pairwise comparisons, which is the fold ratio used as the threshold for the detection of DNA duplication. In order to assign species- and group-specific duplication of *A. thaliana*, *A. halleri* or *A. lyrata* on these data, we searched the data in Figure 3 for a pattern of hybridization intensities between species described in Table1. For example, gene duplications specific to *A. halleri* were those presenting an intensity ratio ⩾1.5 compared with the *A. lyrata* subspecies and *A. thaliana*. This criterion requires at least three independent 1.5-fold or higher ratio intensities between hybridization pairwise comparisons, which decreased the number of candidates to 94 duplicated genes distributed among the *Arabidopsis* genus (Table 1).

Although a 1.5-fold intensity ratio has been shown to be enough to detect differences in expression and DNA duplication (see above), we further selected genes that exhibited statistically significant differences ($P \leqslant 0.05$ and $Q \leqslant 0.05$), yielding 61 and 31 candidates, respectively (see Table 1 and Table 2). We believe that these 31 genes are the most reliable candidates because they correspond to the genes with a less than 5% false discovery rate. However, many of the remaining 63 candidates could be analyzed and confirmed in the future. For example, if one of these candidates is a product of partial duplication (e.g. exon duplication), it may exhibit a 1.5-fold intensity difference but may not necessarily be statistically significant owing to a high variation of the hybridization intensity between probes.

Table 1 shows the six types of species- or group-specific duplications that can be basically classified as the two different kinds of duplication that can be detected using our method. One type is non-*A. thaliana* species-specific duplication, comprising genes exhibiting a higher hybridization signal in *A. halleri* and/or *A. lyrata* DNA compared with the *A. thaliana* self-hybridization signal. These genes are present in *A. thaliana* because they are spotted in the *A. thaliana* tiling array, but they probably have higher numbers of copies in other species in the genus. Genes with a higher signal in *A. thaliana* hybridization compose the second type of duplication. These genes exist in *A. thaliana* and either do not exist in other species or exist in other species in the genus but are duplicated only in *A. thaliana*.

Features of the candidates

The 94 candidates listed by locus identification number can be found in Table 2, which contains, for each gene, the duplication type, number of probes, a brief annotation, chromosomal location and the number of orthologous genes found in *O. sativa*, a plant outgroup that diverged from *Arabidopsis* 150–200 MYA. One can notice that 25 of the 94 candidates are described as pseudogenes. However, the *A. thaliana*-specific group is the only set enriched with them (23 of the 69). Among those that are not described as pseudogenes, 18 are defined as hypothetical proteins or similar. Again, 15 belong to the *A. thaliana*-specific group. Owing to the enrichment of pseudogenes and hypothetical proteins, this group seems to be enriched with possible non-functional genes. This and other characteristics discussed below that are more prominent in this group make *A. thaliana*-specific candidates a set of genes more prone to having the lowest rate of experimental confirmation (see **Discussion**).

Another interesting feature is the significant enrichment of chromosome 4 genes in our set of candidates. Approximately 25% of the 94 candidates (*n* = 24) are localized on the fourth *A. thaliana* chromosome ($\chi^2 = 5.7$, df = 1, $P = 0.017$ compared with the filtered dataset of 25 325 genes). This enrichment seems to be present in all duplication-specific groups and in the more reliable dataset of 31 candidates.

Duplications inside the *Arabidopsis* genus may include specific genes from the group or may consist of ancient genes. In order to verify the frequency of ancient genes among our candidates, we searched for homologous genes present in the genome of *O. sativa*, a plant outgroup that diverged from *Arabidopsis* 150–200 MYA. Table 2 lists the results of the search for new *Arabidopsis* genes in the rice genome. Our analysis was based on amino acid alignment between the predicted gene products of our candidates against the peptide in *O. sativa* (see **Materials and Methods**). Because 25 of our candidates are pseudogenes, the alignment search was restricted to 69 candidates with amino acid sequence. Ten of the 69 candidates had at least one hit against the *O. sativa* sequences. Interestingly, seven (of 23) of those were non-*A. thaliana*-specific duplications, showing that the method is probably less sensitive for the selection of duplication in ancient genes in the species used as a platform for the hybridization.

Discussion

Gene duplication is profound phenomenon in plant genome evolution. By taking the CGH approach, we were able to detect the strength of gene duplication in *Arabidopsis*. The rate and pattern of gene duplication can only be determined by looking for young duplications that evolved recently at the single species or lineage level. We have found a handful of gene duplication events in *A. thaliana* and fewer duplication candidates in *A. halleri* and *A. lyrata* owing to a coordinated decrease in hybridization strength caused by sequence divergence between species. The candidates suggested by the CGH approach provide a path to study gene origination and evolution

Table 1. Criteria for the selection of species- or group-specific gene duplicate candidates based on a 1.5-fold difference of intensity in hybridizations against the *Arabidopsis thaliana* tiling array

Pairwise comparisons						Duplication type	Candidates	
at × ah	at × al	at × ap	ah × al	ah × ap	al × ap		All	Sig.
+	+	+	NA	NA	NA	at-specific	69	18
–	NA	NA	+	+	NA	ah-specific	5	2
NA	–	NA	–	NA	+	al-specific	7	2
NA	NA	–	NA	–	–	ap-specific	12	9
NA	–	–	–	–	NA	al, ap-specific	0	0
–	–	–	NA	NA	NA	ah, al, ap-specific	1	0

at, *A. thaliana*; ah, *A. halleri*; al, *A. lyrata* subsp. *lyrata*; ap, *A. lyrata* subsp. *petraea*; +, higher intensity signal ratio (⩾1.5-fold) in the first species of the comparison; –, higher intensity signal ratio (⩾1.5-fold) in the second species of the comparison; NA, not applied; Sig., significant at $P \leqslant 0.05$ and $Q \leqslant 0.05$.

Table 2. Searching for *Arabidopsis* new duplicate candidates in the rice genome

Locus_ID	Duplt. in	Chr.	No. probes	No. hits in rice	Annotation
AT1G24062	at	1	3	0	Encodes a defensin-like (DEFL) family protein
AT1G28450*	at	1	3	0	MADS-box family protein, similar to MADS-box protein GI:2160701 from (*Pinus radiata*)
AT1G34280	at	1	5	0	Expressed protein
AT1G37999**	at	1	11	NA	Pseudogene, hypothetical protein, contains similarity to replication proteins from (*A. thaliana*)
AT1G42290	at	1	4	NA	Pseudogene, similar to putative helicase, similar to putative helicase GI:4585936 from (*A. thaliana*); *P*= 5.4e-143.
AT1G42300	at	1	2	NA	Pseudogene, hypothetical protein
AT1G42580*	at	1	3	0	Hypothetical protein, contains similarity to hypothetical proteins
AT1G44990	at	1	6	0	Expressed protein
AT1G60290*	at	1	2	NA	Pseudogene, similar to chalcone–flavonone isomerase (EC 5.5.1.6; chalcone isomerase) in Radish, *P*= 1.7e-11.
AT1G63070*	at	1	3	0	Pentatricopeptide (PPR) repeat-containing protein, contains Pfam profile PF01535: PPR repeat
AT1G63535*	at	1	3	0	Encodes a DEFL family protein
AT1G66050*	at	1	3	3	Zinc finger (C3HC4-type RING finger) family protein, contains zinc finger, C3HC4 type (RING finger), signature, PROSITE:PS00518
AT2G04600*	at	2	4	0	Expressed protein
AT2G05084	at	2	2	0	Hypothetical protein
AT2G05635**	at	2	9	0	Hypothetical protein
AT2G06940	at	2	5	NA	Pseudogene, similar to putative pectin methylesterase, blastp match of 41% identity and 3.5e-39 *P*-value to GP
AT2G06983	at	2	4	0	Encodes a member of a family of small, secreted, cysteine rich proteins with sequence similarity to SCR (S locus cysteine-rich protein).
AT2G07390	at	2	2	0	Hypothetical protein
AT2G07395*	at	2	4	NA	Pseudogene, hypothetical protein
AT2G07640	at	2	2	0	D2,D4-dienoyl-CoA reductase-related, contains similarity to peroxisomal D2,D4-dienoyl-CoA reductase (*Mus musculus*) GI:5031508
AT2G10234*	at	2	7	NA	Pseudogene, hypothetical protein
AT2G10560**	at	2	2	0	Expressed protein
AT2G12660	at	2	4	NA	Pseudogene, xyloglucan endotransglycosylase, blastp match of 82% identity and 5.9e-44 *P*-value to GP
AT2G12690	at	2	2	NA	Pseudogene, hypothetical protein
AT2G16140**	at	2	25	0	Expressed protein, contains similarity to hypothetical proteins
AT2G23850	at	2	2	NA	Pseudogene, similar to Uricase II (EC 1.7.3.3; rate oxidase; nodule-specific uricase) in Kidney bean, French bean), *P* = 2.0e-27.
AT2G36040*	at	2	3	0	Expressed protein; expression supported by MPSS
AT3G16750**	at	3	17	0	Expressed protein; expression supported by MPSS
AT3G23320**	at	3	15	0	Hypothetical protein
AT3G29105*	at	3	5	NA	Pseudogene, similar to non-specific lipid transfer protein GB:AAB47967 in *Hordeum vulgare*, *P* = 8.4e-18
AT3G30750*	at	3	4	0	Expressed protein
AT3G31370*	at	3	6	0	Expressed protein, contains similarity to hypothetical proteins
AT3G32022	at	3	5	NA	Pseudogene, helicase, blastp match of 37% identity and 5.2e-207 P value to GP
AT3G32120	at	3	2	0	Hypothetical protein
AT3G32130**	at	3	13	0	Hypothetical protein
AT3G33075*	at	3	2	NA	Pseudogene, hypothetical protein
AT3G33081*	at	3	2	NA	Pseudogene, hypothetical protein
AT3G33572**	at	3	2	0	Hypothetical protein

Table 2. (continued)

Locus_ID	Duplt. in	Chr.	No. probes	No. hits in rice	Annotation
AT3G43760**	at	3	16	0	Expressed protein
AT3G45790	at	3	3	0	Protein kinase-related, contains eukaryotic protein kinase domain, INTERPRO: IPR000719
AT3G46320*	at	3	2	10	Histone H4, nearly identical to histone H4 (*A. thaliana*) GI:166740
AT3G57110	at	3	3	0	Hypothetical protein
AT4G01640*	at	4	2	0	Expressed protein
AT4G03175*	at	4	2	0	Protein kinase family protein, contains similarity to Swiss-Prot:P51566 protein kinase AFC1 (*A. thaliana*)
AT4G03292**	at	4	10	0	Expressed protein; expression supported by MPSS
AT4G03890**	at	4	20	0	Hypothetical protein, contains Pfam profile PF03384: *Drosophila* protein of unknown function, DUF287
AT4G04660**	at	4	15	NA	Pseudogene, hypothetical protein, similar to zinc finger protein (*A. thaliana*) GI:976277
AT4G05570	at	4	2	0	Expressed protein
AT4G06518	at	4	7	NA	Pseudogene, hypothetical protein
AT4G06592	at	4	3	NA	Pseudogene, hypothetical protein
AT4G07452	at	4	3	0	Hypothetical protein
AT4G07630	at	4	2	NA	Pseudogene, hypothetical protein, contains Pfam profile PF03078: ATHILA ORF-1 family
AT4G07650**	at	4	3	NA	Pseudogene, hypothetical protein
AT4G08033*	at	4	6	NA	Pseudogene, hypothetical protein
AT4G08071**	at	4	11	NA	Pseudogene, hypothetical protein
AT4G09290**	at	4	2	0	Expressed protein
AT4G09470	at	4	4	0	Expressed protein
AT4G10660	at	4	2	0	Hypothetical protein
AT5G22960	at	5	2	0	Serine carboxypeptidase S10 family protein, similar to serine carboxypeptidase III (Precursor; SP:P37891; *Oryza sativa*)
AT5G25000	at	5	4	0	Hypothetical protein
AT5G25920**	at	5	32	0	Expressed protein
AT5G28350*	at	5	2	2	Expressed protein
AT5G29624*	at	5	12	0	DC1 domain-containing protein, contains Pfam PF03107: DC1 domain
AT5G29629**	at	5	11	NA	Pseudogene, hypothetical protein
AT5G33431*	at	5	5	NA	Pseudogene, hypothetical protein, temporary automated functional assignment
AT5G46500**	at	5	6	0	Expressed protein
AT5G48595	at	5	2	0	Encodes a DEFL family protein.
AT5G51620	at	5	3	0	Expressed protein
AT5G52690*	at	5	7	0	Heavy metal-associated domain-containing protein, contains Pfam profile PF00403: Heavy metal-associated domain
AT1G50140**	ah	1	112	0	AAA-type ATPase family protein, contains Pfam domain, PF00004: ATPase, AAA family
AT2G27402*	ah	2	7	0	Expressed protein
AT3G32160	ah	3	3	0	Expressed protein
AT4G15110**	ah	4	69	1	Cytochrome P450 97B3, putative (CYP97B3), identical to cytochrome P450 97B3 (SP:O23365; *A. thaliana*)
AT5G34830*	ah	5	16	0	Expressed protein; expression supported by MPSS
AT5G54330*	al,ap,ah	5	8	0	Hypothetical protein, contains Pfam profile PF03478: Protein of unknown function (DUF295)
AT1G29450	al	1	4	0	Auxin-responsive protein, putative, similar to auxin-induced protein 6B (SP:P33083; *Glycine max*)

Table 2. (continued)

Locus_ID	Duplt. in	Chr.	No. probes	No. hits in rice	Annotation
AT1G63270*	al	1	15	1	ABC transporter family protein
AT1G63280*	al	1	4	0	Serpin-related/serine protease inhibitor-related, similar to protein zx in *Hordeum vulgare* subsp.*vulgare* and serpin *Triticum aestivum*
AT2G16010	al	2	4	0	Hypothetical protein
AT4G01330**	al	4	42	0	Protein kinase family protein, contains protein kinase domain, Pfam:PF00069; contains serine/threonine protein kinase domain
AT4G37920**	al	4	38	1	Similar to expressed protein (*A. thaliana*; TAIR:At1g36320.1); similar to P0552C05.25 in *O. sativa* cv. *japonica*
AT5G39863	al	5	6	NA	Pseudogene, similar to receptor-like kinase, blastp match of 65% identity and 4.1e-20 *P*-value to GP
AT1G12725**	ap	1	11	0	Expressed protein
AT1G21330*	ap	1	13	0	Similar to hypothetical protein (*Arabidopsis thaliana*; TAIR:At2g40680.1); similar to orf315 (*Beta vulgaris* subsp. *vulgaris*)
AT1G71410**	ap	1	88	1	Protein kinase family protein, contains protein kinase domain, Pfam:PF00069
AT2G04420**	ap	2	12	0	Expressed protein
AT3G12070**	ap	3	37	2	Similar to geranylgeranyl transferase type II β-subunit SP:P53611 in humans (GI: 1552549)
AT3G44490	ap	3	2	0	Histone deacetylase-related/HD-related, similar to SP
AT4G06646**	ap	4	22	NA	Pseudogene, similar to OJ1081_B12.13, blastp match of 33% identity and 9.8e-51 P-value to GP
AT4G15290**	ap	4	46	0	Cellulose synthase family protein, similar to *Zea mays* cellulose synthase-5 (gi: 9622882), -4 (gi:9622880)
AT4G35610**	ap	4	15	0	Zinc finger (C2H2 type) family protein, contains Pfam domain PF00096: Zinc finger, C2H2 type
AT4G35640**	ap	4	45	1	Encodes a cytosolic serine *O*-acetyltransferase involved in sulfur assimilation and cysteine biosynthesis.
AT4G35650*	ap	4	27	3	Isocitrate dehydrogenase, highly similarity to NAD$^+$-dependent isocitrate dehydrogenase subunit 1 (*A. thaliana*)
AT5G47630**	ap	5	24	0	Acyl carrier family protein/ACP family protein, there is a homologue in *A. thaliana*

at, *A. thaliana*; ah, *A. halleri*; al, *A. lyrata* subsp. *lyrata*; ap, *A. lyrata* subsp. *petraea*. *Gene candidates with *P* values = 0.05 in species pairwise *T*-test; **gene candidates with *P* values = 0.05 and *Q* values = 0.05 in species pairwise *T*-tests (where the *Q* value represents the false discovery rate).

at the species level. We will confirm the duplicated candidates and investigate their evolutionary features in future studies.

The pattern and amount of gene duplication, as determined through retroposition, between *Arabidopsis* and *Oryza* are markedly different. Zhang et al. (2005) recently identified 69 retrosequences in *A. thaliana*, of which more than one-third were found to be pseudogenes, with the remaining having unknown functionality. In contrast, Wang et al. (2006) show that there are extensive retropositions that resulted in over 1 000 identified primary retrogenes in the rice genome. Substitution analyses revealed that the vast majority of these retrogenes are subject to purifying selection against mutations on protein sequences, suggesting with expression and age evidence that they are likely functional retrogenes. Strikingly, Wang et al. (2006) further found that 42% of these retrosequences have recruited new exons from flanking regions, generating a large number of chimerical genes, and some of these chimerical structures are conserved in sorghum and maize. However,

Table 3. Comparison of retroposed genes found in *Arabidopsis thaliana* and *Oryza sativa*

Species	Intact retroposed genes	Retroposed genes with chimeric structure
A. thaliana	69	0
O. sativa	898 (of 1 235 in total)	380

none of retroposed genes in *Arabidopsis* was found to have a chimeric gene structure (Table 3), which may be worthy of further analysis of gene structure using a similar method to that used in rice (Wang et al. 2006). The higher rate of gene duplication through retroposition and other mechanisms may demonstrate that the grass species is able to adapt to more diverse environments.

Previous analysis of chimeric gene structure in the potato genome revealed that the chimera is a common gene structure for organelle-derived nuclear genes (Long et al. 1996; Timmis et al. 2004). Drea et al. (2006) observed that the OEP16 family, which encodes a channel protein of the outer membrane of chloroplasts in *Arabidopsis*, duplicated a new copy, OEP16-S, which acquired a new exon, suggesting that exon shuffling plays a role in the functional divergence of this gene family. Domon and Steinmetz (1994) found that the anther-specific gene in sunflower encoding a proline- and glycine-rich polypeptide with a signal peptide shares important sequence stretches in the 5' coding and upstream regions with another anther-specific gene, suggesting an origination mechanism of exon shuffling. These observations indicate that chimeric genes should not be only restricted to the grass species (Wang et al. 2006). However, the marked contrast between rice and *Arabidopsis* in the retroposed new genes suggests different mechanisms in *Arabidopsis* that may have generated new chimeric genes; for example, non-allelic or non- homologous recombination (Arguello et al. 2006), gene fusion (Nakamura et al. 2007), and other mechanisms at the DNA level. These predictions should be further tested using the new gene candidates identified in the present study.

Advances during the genomic era allowed an exponential increase in DNA sequence data. However, basically only model organisms or representative species for a given group were subjects of sequencing interest. With the exception of a few genera, such as *Drosophila* and *Caenorhabditis*, DNA sequences from different species of the same genus are mostly products of specific research interests in a particular gene or for phylogenetic analyses. As the main genomes were sequenced, experimental tools, such as microarray, became available, making many types of large-scale analysis possible. The method of using microarray to find duplication between closely related species allows the identification of many genetic differences within the same genus without requiring genome sequence information from all the species studied.

The work described herein shows that, using this method, we were able to select candidate duplicate genes within the *Arabidopsis* genus. Moreover, the method can be applied to other plant genera with closely related species and with at least one species with genomic data and a microarray platform. With all the effort applied in the past years to sequence the *O. sativa* genome and to annotate all its genes, rice is a potential genus in which to search for recently duplicated genes using the technique described herein.

As mentioned previously, there are two distinct types of duplication that can be potentially detected by our method: (i) non-*A. thaliana* species-specific duplication; and (ii) *A. thaliana*-specific duplication. The detection of the second type of duplication is limited by the degree of sequence divergence between different species from the same genus, leading to artificially different hybridization intensities. A higher intensity in the *A. thaliana* signal may be due simply to optimized base-pairing of identical sequences, compared with significantly poorer hybridization of the DNA from other species (owing to a high sequence divergence), which would lead to the detection of a false positive candidate of duplication. Here, a false positive specifically refers to a gene that does exist in other species and is not truly duplicated in *A. thaliana*.

To confirm the duplicated candidates, we blasted the 69 *A. thaliana*-specific candidates to the genomic sequences of *A. thaliana* and found that the majority of the candidates (62 of 69) have at least two paralogs in the species *A. thaliana*. Because we used a high stringent criteria (>100 hitting score and <0.000 1 *E* value), it is possible that a handful of these *A. thaliana*-specific candidates diverged less than 5 MYA. We further blasted non-*A. thaliana* candidates to the *A. thaliana* genes. The numbers of homologous genes found was different for *A. thaliana* duplicate candidates and for non-*A. thaliana* duplicate candidates. As indicated in Table 4 the number of candidates that are specific duplications in *A. thaliana* contains many more hits (using nucleotide sequences) than the non-*A. thaliana*-specific duplicate candidates. However, when we performed a peptide search, the non-*A. thaliana* group had many more hits. This is probably because this group does not exhibit as many close paralogs as the *A. thaliana* candidates.

Nonetheless, our approach of requiring at least a 1.5-fold hybridization signal difference in at least three different comparisons minimizes the effect of sequence divergence because it requires that the divergence is consistently present between

Table 4. Blast search duplicated candidates to *Arabidopsis thaliana* genomic data

Specific duplication	No. genes	Analysis	No. genes with peptide sequence	No. genes with hits	No. hits
A. thaliana	69	Pep/blastp	46	19	66
A. thaliana	69	Nuc/blastn	–	21	54
Non-*A. thaliana*	25	Pep/blastp	23	12	44
Non-*A. thaliana*	25	Nuc/blastn	–	3	3

A. thaliana and the other three species. Filtering genes that contain only one probe also reduces the effect of sequence divergence. Homologous genes with higher probe density will only present no hybridization to any of their own probes in cross-species comparisons if the sequence divergence is high. So, within the *A. thaliana* candidates, the most reliable candidates would be those chosen based on a greater numbers of probes.

Another problem of the method we used is that recently duplicated genes in *A. thaliana* are less prone to detection. Tiling array probes are, in general, unique sequences from the genome. Recently duplicated genes tend to bear a lower number of probes, probably because they contain less unique regions. Our filter that select genes with more than one probe possibly discarded *A. thaliana* genes recently duplicated in *A. thaliana*.

As shown by our Blast analysis using *O. sativa* as the outgroup, it seems that *A. thaliana*-specific duplications are less enriched with older genes than other species-specific duplications. This pattern may be the product of the experimental design itself. The group of *A. thaliana* candidates included genes that are species specific, whereas non-*A. thaliana* candidates are genes that are also present in *A. thaliana* (second and first types of duplication, respectively).

Based on the previously explained limitations of the present study, we believe that the method described herein should only be used to select possible candidates for duplication within a genus. This approach should not be used to estimate the ratio of duplication because not all types of duplication have the same probability for detection. In conclusion, duplicate candidates should be confirmed experimentally, especially the *A. thaliana*-specific genes that are prone to be false positive candidates.

Our approach to identify complete gene duplication is a preliminary way to search for genomic duplication. Partial gene duplications, such as exon duplication, are possible between species and may be involved in important evolutionary features. In addition, tiling arrays offer the possibility to search through non-homologous regions along the genome that can constitute different sources of DNA duplication between closely related species.

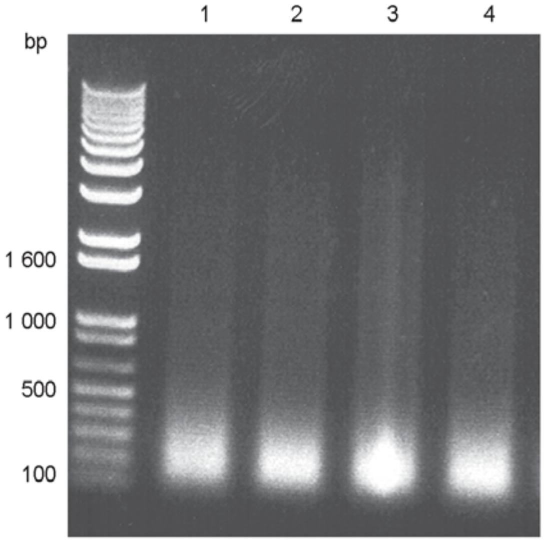

Figure 4. Gel picture of products of fluorescent random labeling of genomic DNA.

Lane 1, *Arabidopsis thaliana*; lane 2, *A. lyrata* subsp. *petraea*; lane 3, *A. lyrata* subsp. *lyrata*; lane 4, *A. halleri*.

Materials and Methods

Species sampling and genomics of DNA extraction

Based on phylogeny of *Arabidopsis*, four species, namely *Arabidopsis thaliana*, *A. lyrata* subsp. *petraea*, *A. lyrata* subsp. *lyrata*, and *A. halleri*, were chosen for comparative genomic hybridization. Genomic DNA was extracted from the leaf tissue of a single plant using a Qiagen (Germantown, MD, USA) purification kit and according to the manufacturer's instructions.

Fluorescent random labeling of genomic DNA using the Invitrogen (Carlsbad, CA, USA) BioPrime labeling system

Approximately 300 ng genomic DNA was added, on ice, to 60 μL of 2.5 × random primer solution and made up to a final volume of 132 μL with distilled water. The mixture was denatured by incubation at 99 °C for 10 min in polymerase chain reaction (PCR) block and immediately put on ice for at least 5 min. Then, 15 μL of 10 × dNTP solution (with biotin-dCTP) and 3 μL Klenow were added to the denatured DNA mixture. The reaction was incubated in PCR block at 25 °C for 16 h and the reaction was terminated by the addition of 15 μL stop solution. The incubation resulted in small biotinylated oligos of approximately 100 bp. Labeled DNA was precipitated by the addition of 20 μL of 3 mol/L sodium actate and 400 μL cold 100% EtOH. This solution was incubated on ice for 2 h and centrifuged at 15 000*g* for 10 min, followed by a wash with 70% EtOH. The pelleted DNA was dried and resuspended in 105 μL of deionized distilled H_2O (Figure 4).

Arabidopsis whole-genome tiling array and hybridization

We applied the forward strand Affymetrix *Arabidopsis* whole-genome tiling arrays, which contains approximately 6.4 approximate million 25-nt oligonucleotide probes, consisting of

3.2 million perfect match (PM) probes that perfectly match genomic sequence and 3.2 million mismatch (MM) probes bearing a different 13th nucleotide. One array covers approximately 97% of the forward strand of all five chromosomes. Each 35-bp genomic region is represented by a 25-nt feature chosen for optimal hybridization characteristics with 35-bp window.

The biotin-labeled target DNA fragments (approximately 100–150 bp) from random labeling methods were hybridized onto forward-strand whole-genomic tiling array following the standard Affymetrix protocol.

Microarray analysis

All statistical analyses were performed in R (http://r-project.org). After individual "cel." files with raw intensity data were read into R, a matrix corresponding to the original 2 560 × 2 560 features was recreated. Intensities were log transformed and normalized using quantile methods. The mean log intensity of a 51 × 51 feature sliding-window was calculated at each coordinate. Only data from the unique 1 683 620 and corresponding MM features were used for spatial correction. This matrix of window means was subtracted from the original matrix of log intensities, yielding the spatial-corrected feature intensity for each unique feature (Borevitz et al. 2003).

T-test statistic and false discovery rate

Six pair-wise *T*-tests were performed between 4 *Arabidopsis* species: the *T* value was calculated based on the mean feature intensity of three replicates from each species. The false discovery rate (FDR) was calculated using Bioconductir's *Q* value, R package (http://faculty.washington.edu/~jstorey/qvalue/). The method provides an estimated *Q* value for each *T*-test, which represents the minimum FDR (Storey and Tibshirani 2003). Our reliable data set was composed of 31 genes presenting *Q* values $\leqslant$0.05 in pair-wise intensity comparisons between different species.

Computational analysis

A. thaliana gene and peptide sequences from The *Arabidopsis* Information Resource (TAIR), release 6.0, were used in our analyses (http://www.arabidopsis.org/). Genes corresponding to transposable elements were identified by searching their description for the words "retro", "transcriptase", or "transposase". In order to identify orthologous genes between *A. thaliana* and *O. sativa*, peptide sequences of the two organisms were aligned with Blast. *Oryza sativa* L. subsp. *indica* peptide sequences were downloaded from the BGI-RISe Database (http://rise.genomics.org.cn/rice/index2.jsp, accessed December 2006). We then selected the alignments presenting at least 50% of protein identity and covering at least 75% of both sequences.

Acknowledgements

The authors thank Dr Hedibert F. Lopes (University of Chicago, Graduate School of Business, Chicago, USA) for his discussions regarding the statistical methods, Noboru Jo Sakabe (IQ-USP/LICR-Sao Paulo Branch, Sao Paulo, Brazil) for critical reading of the manuscript. The authors also thank those people who provided seed samples of *Arabidopsis*: ABRC Stock Center (The Ohio State University, Columbus, OH 43210, USA), Justin Borevitz (Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois 60637, USA), Daphne Preuss (Department of Molecular Genetics and Cell Biology, The University of Chicago, Chicago, Illinois 60637, USA), Joy Bergelson (Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois 60637, USA), and Mark MacNair (Department of Biological Sciences, University of Exeter, Exeter, EX4 6EZ, UK).

References

Arguello JR, Chen Y, Yang S, Wang W, Long M (2006). Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet.* **2**, e77.

Borevitz JO, Liang D, Plouffe D, Chang H, Zhu T, Weigel D et al. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**, 513–523.

Brosius J (1991). Retroposons—seeds of evolution. *Science* **251**, 753.

Domon C, Steinmetz A (1994). Exon shuffling in anther-specific genes from sunflower. *Mol. Gen. Genet.* **244**, 312–317.

Drea SC, Lao NT, Wolfe KH, Kavanagh TA (2006). Gene duplication, exon gain and neofunctionalization of OEP16-related genes in land plants. *Plant J.* **46**, 723–735.

Fan C, Long M (2007). A new retroposed gene in *Drosophila* heterochromatin detected by microarray-based comparative genomic hybridization. *J. Mol. Evol*. **64**, 272–283.

Gilad Y, Borevitz J (2006). Using DNA microarrays to study natural variation. *Curr. Opin. Genet. Dev.* **16**, 553–558.

Gilbert W (1978). Why genes in pieces? *Nature* **271**, 44.

Hughes AL (1994). The evolution of functionally novel proteins after gene duplication. *Proc. R Soc. Lond B* **256**, 119–124.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* **436**, 793–800.

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

Khush GS (1997). Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35**, 25–34.

Kimura M (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Long M, Betran E, Thornton K, Wang W (2003). The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875.

Long MY, DeSouza SJ, Rosenberg C, Gilbert W (1996). Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc. Natl. Acad. Sci. USA* **93**, 7727–7731.

Makalowski W, Mitchell GA, Labuda D (1994). *Alu* sequences in the coding regions of mRNA: A source of protein variability. *Trends Genet.* **10**, 188–193.

Nakamura Y, Itoh T, Martin W (2007). Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. *Mol. Biol. Evol.* **24**,110–121.

Ochman H (2001). Lateral and oblique gene transfer. *Curr. Opin. Genet. Dev.* **11**, 616–619.

Ohno S (1970). *Evolution by Gene Duplication.* Springer, Berlin.

Pinkel D, Albertson DG (2005). Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* **37** (Suppl.), S11–S17.

Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445.

The Arabidopsis Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135.

Wang W, Yu H, Long M (2004). Duplication–degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat. Genet.* **36**, 523–527.

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z et al. (2006). High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **8**, 1791–1802.

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

Zhang Y, Wu Y, Liu Y, Han B (2005). Computational identification of 69 retroposons in Arabidopsis. *Plant Physiol.* **138**, 935–948.

Zhu Q, Ge S (2005). Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytol.* **167**, 249–265.

(Handling editor: Hong Ma)