Gene Space Dynamics During the Evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* Genomes

A. N. Massa,[1] H. Wanjugi,[2] K. R. Deal,[3] K. O'Brien,[4] F. M. You,[3] R. Maiti,[5] A. P. Chan,[5] Y. Q. Gu,[2] M. C. Luo,[3] O. D. Anderson,[2] P. D. Rabinowicz,[4,5,6] J. Dvorak,[3] and K. M. Devos*[,1]

[1]Institute of Plant Breeding, Genetics and Genomics (Department of Crop and Soil Sciences), and Department of Plant Biology, University of Georgia

[2]United States Department of Agriculture-Agricultural Research Service, Genomics and Gene Discovery Research Unit, Western Regional Research Center, Albany, California

[3]Department of Plant Sciences, University of California

[4]Institute for Genome Sciences, University of Maryland School of Medicine

[5]The J. Craig Venter Institute, Rockville, Maryland

[6]Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine

***Corresponding author:** E-mail: kdevos@uga.edu

Associate editor: Barbara Holland

Research article

Abstract

Nine different regions totaling 9.7 Mb of the 4.02 Gb *Aegilops tauschii* genome were sequenced using the Sanger sequencing technology and compared with orthologous *Brachypodium distachyon*, *Oryza sativa* (rice), and *Sorghum bicolor* (sorghum) genomic sequences. The ancestral gene content in these regions was inferred and used to estimate gene deletion and gene duplication rates along each branch of the phylogenetic tree relating the four species. The total gene number in the extant *Ae. tauschii* genome was estimated to be 36,371. The gene deletion and gene duplication rates and total gene numbers in the four genomes were used to estimate the total gene number in each node of the phylogenetic tree. The common ancestor of the Brachypodieae and Triticeae lineages was estimated to have had 28,558 genes, and the common ancestor of the Panicoideae, Ehrhartoideae, and Pooideae subfamilies was estimated to have had 27,152 or 28,350 genes, depending on the ancestral gene scenario. Relative to the Brachypodieae and Triticeae common ancestor, the gene number was reduced in *B. distachyon* by 3,026 genes and increased in *Ae. tauschii* by 7,813 genes. The sum of gene deletion and gene duplication rates, which reflects the rate of gene synteny loss, was correlated with the rate of structural chromosome rearrangements and was highest in the *Ae. tauschii* lineage and lowest in the rice lineage. The high rate of gene space evolution in the *Ae. tauschii* lineage accounts for the fact that, contrary to the expectations, the level of synteny between the phylogenetically more related *Ae. tauschii* and *B. distachyon* genomes is similar to the level of synteny between the *Ae. tauschii* genome and the genomes of the less related rice and sorghum. The ratio of gene duplication to gene deletion rates in these four grass species closely parallels both the total number of genes in a species and the overall genome size. Because the overall genome size is to a large extent a function of the repeated sequence content in a genome, we suggest that the amount and activity of repeated sequences are important factors determining the number of genes in a genome.

Key words: *Aegilops tauschii*, comparative genomics, gene space, genome evolution, genome organization.

Introduction

Sequencing of the *Oryza sativa* (rice), *Sorghum bicolor* (sorghum), *Zea mays* (maize), and *Brachypodium distachyon* genomes has generated a wealth of information about the structure and evolution of genomes in the grass family (International Rice Genome Sequencing Project 2005; Paterson et al. 2009; Schnable et al. 2009; Vogel et al. 2010). However, because the current technologies do not allow the assembly of large and complex plant genomes sequenced with a shotgun approach, our knowledge of grass genome structure and evolution is based on relatively small genomes of similar sizes; the *B. distachyon* genome is 272 Mb (Luo et al. 2009; Vogel et al. 2010), the rice genome is 389 Mb (International Rice Genome Sequencing Project 2005), and the sorghum genome is 730 Mb (Paterson et al. 2009). The maize genome, although 2.3 Gb (Schnable et al. 2009), is paleotetraploid (Gaut and Doebley 1997). Maize polyploidization is estimated to have occurred more than 4.8 Ma (Swigonova et al. 2004). In the past 3 My, the maize genome expanded by the addition of large quantities of long terminal repeat retroelements (SanMiguel and Bennetzen 1998). The size of the diploid genome ancestral to that of modern paleotetraploid maize

was therefore probably similar to that of sorghum. Numerous lineages in the *Poaceae* family have genome sizes that are much larger than those currently sequenced. For example, the genomes of diploid species in the tribe Triticeae, which includes economically important cereals such as wheat, barley, and rye, range from 3.4 to 8.8 Gb (1C) (Vogel et al. 1999; Jakob et al. 2004). No genome in this size range has been sequenced, and little is therefore known about the structure and evolution of these genomes. Because biological phenomena often become apparent in comparisons of extremes, comparative analyses of the structure and evolution of the currently sequenced genomes with those of large genomes, such as those in the tribe Triticeae, may provide new insights into grass and plant genome biology. One of the unresolved biological dilemmas is the significance of genome size variation. In grasses, for instance, genome sizes vary by more than one order of magnitude. It is generally accepted that the principal cause of this variation is the accumulation of transposable elements (TEs) of which the most abundant are long terminal repeat retroelements (Bennetzen and Kellogg 1997). An increase in TE numbers results in genome expansion, whereas their deletion by illegitimate recombination (recombination requiring little or no homology for genetic exchange) or unequal homologous recombination leads to genome contraction (Bennetzen and Kellogg 1997; Bennetzen 2005).

Genes in the larger grass genomes are often interspersed with TEs inserted into each other (SanMiguel et al. 1996; Dubcovsky et al. 2001; Wicker et al. 2001). Insertions and deletions of TEs take place at an extraordinary rate in some grass genomes. The DNA in the immediate vicinity of genes thus can be in a highly dynamic state. In Triticeae genomes, for instance, virtually the entire intergenic space is replaced in 3–4 My (Dubcovsky and Dvorak 2007). It seems unavoidable that a precipitous rate of sequence turnover in the immediate vicinity of genes would impact gene stability. It was therefore hypothesized that large grass genomes are less stable than related small genomes and that the gene space (the totality of genes and gene-related DNA) of large and complex plant genomes is evolving faster than in the related small genomes (Luo et al. 2009). The rate with which subchromosomal structural changes, which involve groups of genes, accumulate was shown to be nearly an order of magnitude faster in the Triticeae lineage than in the rice, sorghum, and *B. distachyon* lineages, including the internal branch of the grass phylogenetic tree immediately preceding the divergence of the Brachypodieae and Triticeae (Luo et al. 2009; Vogel et al. 2010). One might therefore expect that the evolution of individual genes by deletions and duplications would also be accelerated in the Triticeae genomes. Regional sequence comparisons along wheat chromosome 3B with orthologous regions in other grass genomes have suggested that this may be the case (Choulet et al. 2010). Numerous gene deletions and duplications and an overall increase in gene number were reported in chromosome 3B compared with orthologous regions in other grass genomes. Observations made on a single hexaploid wheat chromosome could potentially be confounded by factors, such as polyploidy, and genome- and chromosome-specific events (Akhunov, Akhunova, et al. 2003; Dvorak and Akhunov 2005). Extending the comparative analyses to wheat diploid relatives is therefore needed to advance the understanding of the dynamics of gene space evolution in diploid grass genomes.

In this paper, we compared the patterns and rates of evolutionary events in the diploid *Aegilops tauschii* genome with those in the small rice, sorghum, and *B. distachyon* genomes. The genome size of *Ae. tauschii* was estimated to be 4,020 Mb (Arumuganathan and Earle 1991), which places it at the lower end of the genome-size spectrum in Triticeae. *Aegilops tauschii* was used previously in the study of the rates of structural changes in grasses (Luo et al. 2009; Vogel et al. 2010). It is also the diploid source of the wheat D genome (Kihara 1944; McFadden and Sears 1946) and understanding its biology is important for advancing the biology of polyploid wheat. Of the three small genomes, *B. distachyon* is phylogenetically most closely related to the tribe Triticeae, and its genome sequence has been anticipated to facilitate wheat structural and functional genomics (Vogel et al. 2010). The comparison of the sequenced regions of hexaploid wheat chromosome 3B with orthologous regions in *B. distachyon* chromosome Bd2 and rice chromosome Os1 led to the conclusion that the synteny level between wheat and *B. distachyon* is about the same as between wheat and rice (Choulet et al. 2010). This observation is also reinvestigated here by employing genomic comparisons involving the diploid *Ae. tauschii*.

Synteny of nearly 10 Mb of *Ae. tauschii* genomic sequence, generated from nine contiguous and randomly selected regions across the genome, with orthologous regions in *B. distachyon*, rice, and sorghum was used here to reconstruct the gene content in these regions in the common ancestor of the BEP and PACCAD clades. The BEP clade includes the subfamilies Bambusoideae, Ehrhartoideae, and Pooideae (Barker et al. 2001). *Aegilops tauschii* and *B. distachyon* belong to the subfamily Pooideae (Triticeae and Brachypodieae tribes, respectively), and rice belongs to the subfamily Ehrhartoideae. The PACCAD clade is formed by the subfamilies Panicoideae, which includes sorghum and maize, Aristidoideae, Centhothecoideae, Chloridoideae, Arundinoideae, and Danthoideae (Barker et al. 2001). The estimated divergence ages for the Triticeae and Brachypodiaeae are 32–39 Ma (Bossolini et al. 2007; Vogel et al. 2010), for the Pooideae and Ehrhartoideae 40–54 Ma (Vogel et al. 2010), and for the BEP and PACCAD clades 45–60 Ma (Vogel et al. 2010). The inferred ancestral gene content in the nine investigated genomic regions and the time in million years along the branches of the phylogenetic tree are used to estimate the rates of gene deletions and gene duplications in the *Ae. tauschii*, rice, sorghum, and *B. distachyon* lineages, in the branch that represents the common ancestor to the *B. distachyon* and *Ae. tauschii* lineages immediately before their divergence (referred to as internode 2), and in the branch that represents the ancestor

to the Ehrhartoideae and Pooideae immediately before the split of those two lineages (referred to as internode 1). The rate estimates, in turn, are used to reconstruct the evolutionary history of the four genomes. These numerical analyses show, for the first time, that gene duplication rates, the gene number in the genome, and overall genome size are intimately related in the grass family.

Materials and Methods

Selection of Contigs and BAC Sequencing

Aegilops tauschii contigs of bacterial artificial chromosome (BAC) clones with an estimated minimum length of 0.8 Mb were selected from 11,656 contigs assembled from fingerprinted BAC clones (Luo et al. 2003). Map information was used to select clones from different chromosomes and from different regions along the telomere–centromere axis that differed in the physical distance from the centromere and recombination rate (supplementary table S1, Supplementary Material online). Except for these criteria, selection was random. BAC clones representing a minimum tiling path were sequenced on an ABI 3730xl using Sanger technology. Sequences were assembled using the Celera Assembler (Myers et al. 2000). Where possible, scaffolds within each BAC were ordered manually using the overlap between neighboring BACs as guidance. More detail on the clone selection, sequencing, and scaffold reordering is given in the supplementary text S1, Supplementary Material online. BAC sequences are available from GenBank (http://www.ncbi.nlm.nih.gov/genbank/, last accessed 04/14/2011, accession numbers: AC237682–AC237794 and AC241723–AC241798, supplementary table S2, Supplementary Material online).

Gene Annotation

Gaps between scaffolds were closed by adding 50 Ns to generate a contiguous sequence for each BAC. The BAC sequences were run through the DAWGPAWS pipeline, which does ab initio as well as homology-based identification of genes and TEs (Estill and Bennetzen 2009). The DAWGPAWS output file was visualized in Apollo (Lewis et al. 2002), and gene models were annotated manually. A predicted gene was considered a true gene if 1) it had homology at the protein level (E value $\leq e^{-10}$) to expressed sequence tags and/or an annotated gene in one of the sequenced grass genomes and 2) the corresponding *Ae. tauschii* protein covered at least 2/3 of the putative orthologous rice, maize, or sorghum protein. The second criterion was disregarded if the gene was adjacent to a gap in the *Ae. tauschii* sequence or was located close to the end of a BAC clone.

Comparative Analyses

The genes identified in *Ae. tauschii* were used as queries in Blast searches at the protein level (E value $\leq e^{-10}$) to identify orthologous regions in *B. distachyon*, *O. sativa*, and *S. bicolor* genomes (See supplementary text S1, Supplementary Material online, for information on the assembly and annotation version used). Information on orthologous/paralogous relationships between genes and syntenic relationships between grass genomes was obtained from Gramene (www.gramene.org, last accessed 04/14/2011).

To assess the rate with which the *Ae. tauschii*, *B. distachyon*, rice, and sorghum syntenic regions were evolving, we determined whether orthologs of genes that were present in the syntenic regions in any of the latter three species and absent in *Ae. tauschii* were found in conserved positions across species boundaries using the comparative tools available in Gramene (www.gramene.org). Because gene annotation is not perfect and often includes TEs, only genes that had orthologs in other monocots were included in this analysis. More detailed information on the comparative methodologies can be found in the supplementary text S1, Supplementary Material online.

Calculating the Gene Deletion and Gene Duplication Rates and Total Gene Numbers

The gene deletion rate (m) was estimated for each lineage from the number of deleted genes relative to the reconstructed ancestral state using equation (1),

$$m = -\frac{\ln n_{(t)} - \ln g_{(0)}}{t}, \quad (1)$$

where $g_{(0)}$ is the number of genes in the ancestral genome and $n_{(t)} = g_{(0)} - M$, with M the number of genes deleted during time t across the nine regions in a specific genome.

The gene duplication rate k was computed using equation (2),

$$k = \frac{\ln g_{(t)} - \ln g_{(0)}}{t} + m, \quad (2)$$

where $g_{(0)}$ is the ancestral gene number, $g_{(t)}$ is the observed (or inferred) number of genes at time t across the nine regions, and m is the deletion rate.

The total initial number of genes per genome, $G_{(0)\text{total}}$, at the time of lineage divergence was estimated using the total number of genes at time t ($G_{(t)\text{total}}$) and the length of the lineage divergence in million years (t), using equation (3).

$$G_{(0)\text{total}} = \frac{G_{(t)\text{total}}}{e^{(k-m)t}}. \quad (3)$$

More information on the derivation of the equations can be found in supplementary text S1, Supplementary Material online.

Results

Aegilops tauschii BAC Contigs

A total of 90 *Ae. tauschii* BAC clones equivalent to 9.7 Mb and representing a minimum tiling path across nine contigs with a length between 0.61 and 1.91 Mb were sequenced. Four of the contigs originated from chromosome 4D (1 from the short arm and 3 from the long arm), three from chromosome 2D (1 from the short arm and two from the long arm), and one each from the short arms of chromosomes 3D and 6D. Each BAC sequence assembled into between 1 and 22

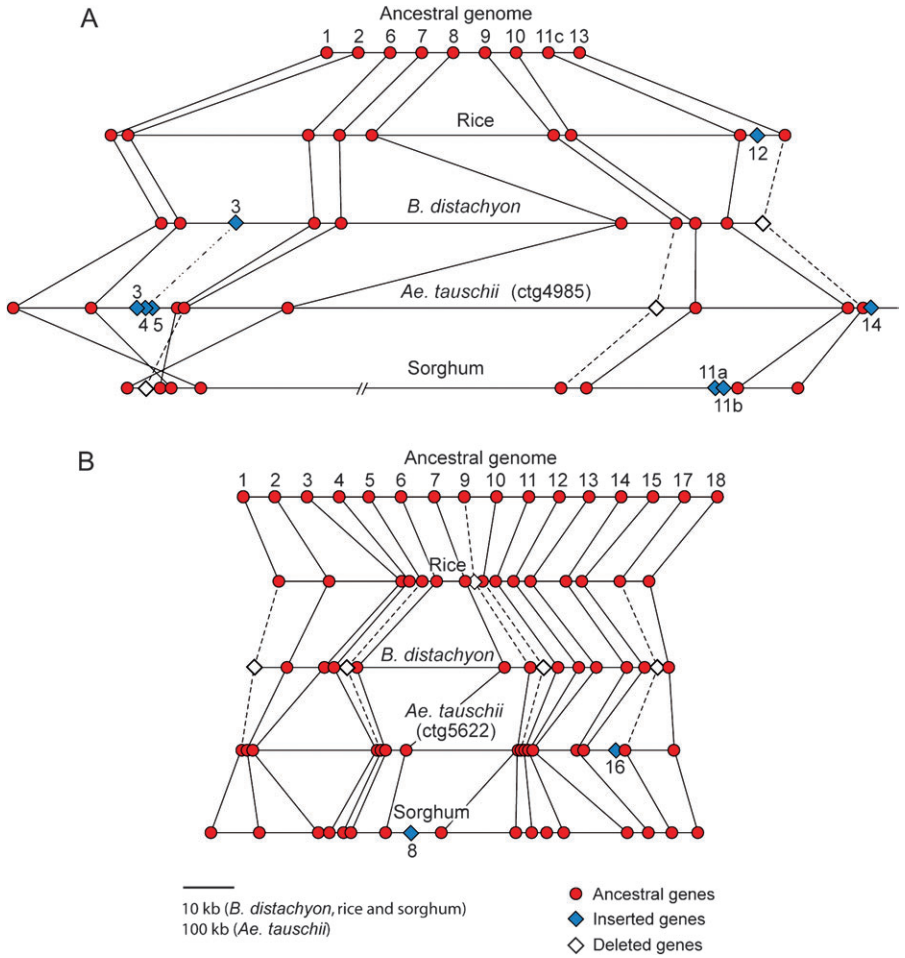

FIG. 1. Structural organization of two genomic regions in four current-day grasses and their ancestor. (*A*) ctg4985. (*B*) ctg5622. Ancestral genes are represented by red circles, deleted genes by white diamonds, and inserted genes by blue diamonds. Orthologous genes are connected by full lines or, in cases of gene deletions or insertions, by dashed lines.

scaffolds. The average and median number of scaffolds per BAC were 4.5 and 3.0, respectively. Sixty-four percent of the scaffolds were manually ordered and oriented using scaffold information from overlapping BAC clones. Some 30% of the gaps between scaffolds could be closed using overlapping BAC sequence.

A total of 90 genes were annotated in the 9,695 kb of *Ae. tauschii* sequence. Although the BAC sequences contained gaps, we are confident that few, if any, genes were missed. The *Ae. tauschii* BAC clones were sequenced to a depth of 8–10X. Data simulations in wheat, maize, and rice have indicated that 3X coverage is sufficient to identify >95% of all genes (Ramakrishna W, Bennetzen JL, personal communication.). Furthermore, none of the gaps that were covered by sequence from overlapping BAC clones contained genes. In addition to the full-length genes, a total of 40 pseudogenes that were truncated or contained stop codons or frameshift mutations were observed in the nine *Ae. tauschii* contigs, giving a ratio of pseudogenes/genes of 1/2.2. A similar ratio of pseudogenes to genes (1/2.9) was reported for wheat chromosome 3B (Choulet et al. 2010). The number of pseudogenes per contig ranged from 1 in ctg3466 and ctg5080 to 12 in ctg6274. Forty-five percent of the pseudogenes originated through tandem duplications, and 8% were fragments of chloroplast genes. The origin of the remaining pseudogenes is unknown, but presumably they were paralogous copies of genes elsewhere in the genome. The number of pseudogenes per contig (excluding chloroplast gene fragments) correlated with the number of full-length gene insertions (Pearson's $r = 0.75$; $P = 0.019$). The correlation remained significant even when chloroplast gene fragments were included.

A list of the contigs, including their length, map position, BAC clone members, BAC clone characteristics, annotated genes and pseudogenes, and their positions in the contigs are summarized in supplementary table S3, Supplementary Material online.

Ancestral Gene Numbers

The 90 annotated genes were used for comparative analysis with *B. distachyon*, rice, and sorghum. The colinearity data across the nine syntenic regions in four species provided information on the putative gene content of each region in a species ancestral to these lineages and the subsequent evolution of these regions (fig. 1; supplementary table S4,

Supplementary Material online). Genes that were present in colinear positions in sorghum and in at least one BEP clade species (*Ae. tauschii*, *B. distachyon*, and rice) were considered to have retained their ancestral position. A gene was assumed to represent an "insertion" if it was annotated in a target region in one species and the presumed orthologs were located in noncolinear positions relative to this species but in colinear positions in at least one BEP and one PACCAD clade species. In the reconstruction of the evolutionary events that shaped each region, we thus considered two events: 1) insertion of genes into a region (this includes tandem duplications) and 2) deletion of genes from a region.

Based on the colinearity data, it was possible to decide on the ancestral state for 108 of the 111 genes encountered in the four genomes (supplementary table S4, Supplementary Material online). Seven genes were located in noncolinear positions in all four species and hence provided no information regarding their ancestral location. Nevertheless, the noncolinear locations suggested that they were insertions. For genes that were colinear only in species belonging to the BEP clade, two evolutionary scenarios are possible. The colinear location of the gene in the BEP species represents the ancestral gene location and the gene absence in the orthologous sorghum region is the derived state, or the gene was absent from the ancestral region and was inserted in a common ancestor to the Pooideae and Ehrhartoideae species. To distinguish between the two scenarios, we examined the location of the orthologous genes in maize and/or foxtail millet (Schnable et al. 2009; www.gramene.org; www.phytozome.net [last accessed 04/14/2011]). For four genes (ctg3466—Gene 15; ctg6274—Genes 2, 10, and 11; supplementary table S4, Supplementary Material online), colinearity was conserved between the gene locations in maize and/or foxtail millet and at least one member of the BEP clade, indicating that those genes had been deleted from the ancestral position in sorghum. For three other genes (ctg6274—Genes 4, 12; ctg4985—Gene 7), the maize and foxtail millet orthologs were either colinear with the sorghum location or their syntenic relationship was unclear. No conclusions could be drawn regarding the ancestral origin of those genes. One of those genes, Gene 12 on ctg6274, was colinear only between *Ae. tauschii* and rice. This suggests that regardless of whether this gene was inserted in the region in a common ancestor to *Ae. tauschii*, *B. distachyon*, and rice or was present in this region in the ancestral grass genome, it was deleted in the *B. distachyon* lineage. Gene 2 on ctg4985 was colinear only between the Pooideae species *Ae. tauschii* and *B. distachyon* and was present in different noncolinear positions in rice and sorghum. However, the identification of an ortholog in maize in a position that was colinear to the rice location suggested that the rice and maize genes most likely have retained their ancestral position. This is the only example we found of a gene that had been inserted into the common ancestor of *Ae. tauschii* and *B. distachyon* after its divergence from the rice lineage. Two genes (ctg3466—Gene 6a; ctg5748—Gene 3) were present in colinear positions in rice and sorghum and absent from the orthologous *Ae. tauschii* and *B. distachyon* lineages, and these genes are assumed to have been deleted in the ancestor of the *Ae. tauschii* and *B. distachyon* lineages.

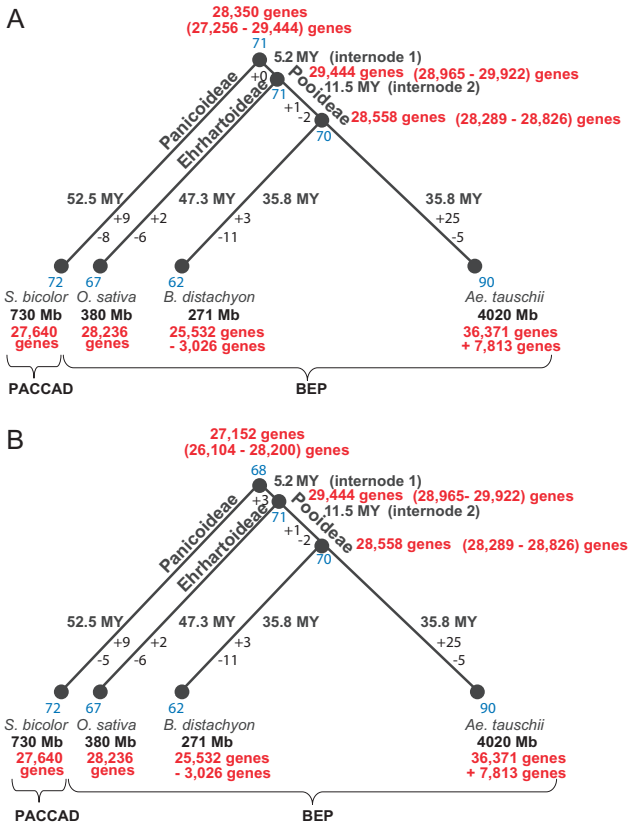

Fig. 2. Phylogenetic tree showing relationships among the analyzed species and dynamics of gene content changes. The number of gene deletions (numbers preceded by − sign) and insertions (numbers preceded by + sign) that occurred in each branch of the phylogenetic tree are given along the branches. The nodes of the phylogenetic tree that were analyzed are indicated with closed circles. The total numbers of annotated (sorghum, rice, and *B. distachyon*), inferred (*Ae. tauschii*), or computed genes (ancestor of Brachypodieae and Triticeae, ancestor of Ehrhartoideae and Pooideae, and ancestor or BEP and PACCAD clades) are given in red. The range for the computed genes is listed in parenthesis. The numbers of genes identified or inferred in the nine investigated regions are in blue. The length of branches based on divergence times are given in million years. (*A*) A tree based on an ancestral gene number of 71; (*B*) A tree based on an ancestral gene number of 68.

Of the 111 genes annotated in the investigated regions, 46 were syntenic in all four species analyzed, 25 were syntenic in at least one BEP and one PACCAD clade species, and 3 were syntenic in at least two BEP clade species. Depending on whether the three genes that are present only in BEP clade species are either ancestral or were inserted into a common ancestor to *Ae. tauschii*, *B. distachon*, and rice, there are 71 or 68 genes, respectively, that represent the ancestral state of the gene space in the nine genomic regions (fig. 2). The remaining 40 (in case of 71 ancestral genes) or 43 genes (in case of 68 ancestral genes) were inserted into their current locations. A total of 25 insertions and 5 deletions differentiated the gene content of the nine regions in *Ae. tauschii* from that of the

Table 1. Number of Genes Deleted and Duplicated Relative to the Ancestral State of 71 or 68 Genes Over the Past 45–60 My.

	Ancestral Gene Number = 71		Ancestral Gene Number = 68	
Lineage	No. of Inserted and Duplicated Genes	No. of Deleted Genes	No. of Inserted and Duplicated Genes	No. of Deleted Genes
Brachypodium distachyon	4[b]	13[a]	7[b]	13[a]
Oryza sativa	2[b]	6[a]	5[b]	6[a]
Sorghum bicolor	9[b]	8[a]	9[b]	5[a]
Aegilops tauschii	26[a]	7[a]	29[a]	7[a]

NOTE.—Numbers in columns sharing the same letter are not significantly different at the 5% probability level (2 × 2 contingency table and Fisher's exact test).

presumed *Ae. tauschii*/*B. distachyon* ancestor (fig. 2). In *B. distachyon*, 3 insertions and 11 deletions took place over the last 32–39 (a mean of 35.8) My. Overall, the number of evolutionary events that occurred during 45–60 (mean of 52.5) My of evolution was significantly lower in the three smaller grass genomes than in *Ae. tauschii*, which is entirely due to the higher number of gene insertions in the *Ae. tauschii* genome compared with the *B. distachyon*, rice, and sorghum genomes (table 1). The number of gene deletions did not significantly differ among the four species. In *Ae. tauschii*, gene insertions and tandem gene duplications significantly outnumbered gene deletions, whereas the ratio was reversed in *B. distachyon* ($P = 0.0002$) and rice ($P = 0.007$) compared with that in *Ae. tauschii*. In these species, gene deletions surpassed gene insertions and tandem duplications (table 1). In sorghum, the ratio of insertions and tandem duplications to deletions did not differ significantly from that in *Ae. tauschii* ($P = 0.1$).

Gene Insertion and Duplication Rates and Gene Space Evolution

The number of deleted and inserted genes can be used to estimate the gene deletion rate (m) and gene duplication rate (k) for each lineage. As discussed above, there were either 71 or 68 genes ($g_{(0)}$) in the ancestor of the BEP and PACCAD clades. If the ancestral gene number was 71, the ancestor of the Ehrhartoideae and Pooideae had the same number of genes as the ancestor to the BEP and PACCAD clades, and the sorghum lineage lost eight genes during 52.5 My of evolution (fig. 2*Ae*). If the ancestral gene number was 68, then the ancestor of the Ehrhartoideae and Pooideae gained three genes before the divergence of the Pooideae and Ehrhartoideae, and the sorghum lineage lost five genes during its subsequent evolution to current-day sorghum (fig. 2*B*). Our synteny data also showed that the Pooideae lineage gained one gene and lost two genes before the divergence of the Triticeae and Brachypodieae (fig. 2). The number of genes gained and lost in all lineages assuming either 71 or 68 ancestral genes, and the corresponding k and m values are listed in table 2. Because no genes were deleted or inserted in the nine regions after the divergence of the PACCAD and BEP clades and before the divergence of the Pooideae and Ehrhartoideae under the 71 ancestral gene scenario, k and m were zero for this branch of the phylogenetic tree

Table 2. Information About the Gene Space in Four Current-Day Species and Their Ancestors.

Lineage	Genome Size (Mb)	Time[a] (t)	No. of Genes $g_{(t)}$	Inserted and Duplicated Genes	Deleted Genes	k (Gene^{-1} My^{-1})	m (Gene^{-1} My^{-1})	k/m Ratio	$G_{(t)\text{total}}$	$G_{(0)\text{total}}$
Ancestral gene number = 71 (fig. 2*A*)										
Brachypodium distachyon	271	35.8	62	3	11	0.001385	0.004775	0.290	25,532	28,826[b]
Aegilops tauschii	4020	35.8	90	25	5	0.009090	0.002070	4.391	36,371	28,289[b]
Internode 2	?	11.5	70	1	2	0.001251	0.002485	0.504	28,558[c]	28,965[d]
Oryza sativa	380	47.3	67	2	6	0.000641	0.001867	0.343	28,236	29,922[d]
Internode 1	?	5.2	71	0	0	0.000000	0.000000	—	29,444[e]	29,444[f]
Sorghum bicolor	730	52.5	72	9	8	0.002543	0.002277	1.117	27,640	27,256[f]
Ancestral gene number = 68 (fig. 2*B*)										
B. distachyon	271	35.8	62	3	11	0.001385	0.004775	0.290	25,532	28,826[b]
Ae. tauschii	4020	35.8	90	25	5	0.009090	0.002070	4.391	36,371	28,289[b]
Internode 2	?	11.5	70	1	2	0.001251	0.002485	0.504	28,558[c]	28,965[d]
O. sativa	380	47.3	67	2	6	0.000641	0.001867	0.343	28,236	29,922[d]
Internode 1	?	5.2	71	3	0	0.008302	0.000000	—	29,444[e]	28,200[f]
S. bicolor	730	52.5	72	9	5	0.002543	0.001455	1.748	27,640	26,104[f]

NOTE.— t is the divergence time in million years, $g_{(t)}$ is the total number of genes in the investigated regions, k is the duplication rate, m is the deletion rate, $G_{(t)\text{total}}$ is the total number of genes in the genome at time t, and $G_{(0)\text{total}}$ is the estimated total number of genes in the ancestral genome.

[a] From Vogel et al. (2010).

[b] In the ancestor to the Brachypodieae and Triticeae.

[c] An estimate generated by averaging the $G_{(0)\text{total}}$ of *B. distachyon* and *Ae. tauschii*.

[d] In the ancestor to the Ehrhartoideae and Pooideae.

[e] An estimate generated by averaging the $G_{(0)\text{total}}$ of *O. sativa* and Internode 2.

[f] In the BEP clade and PACCAD clade ancestor.

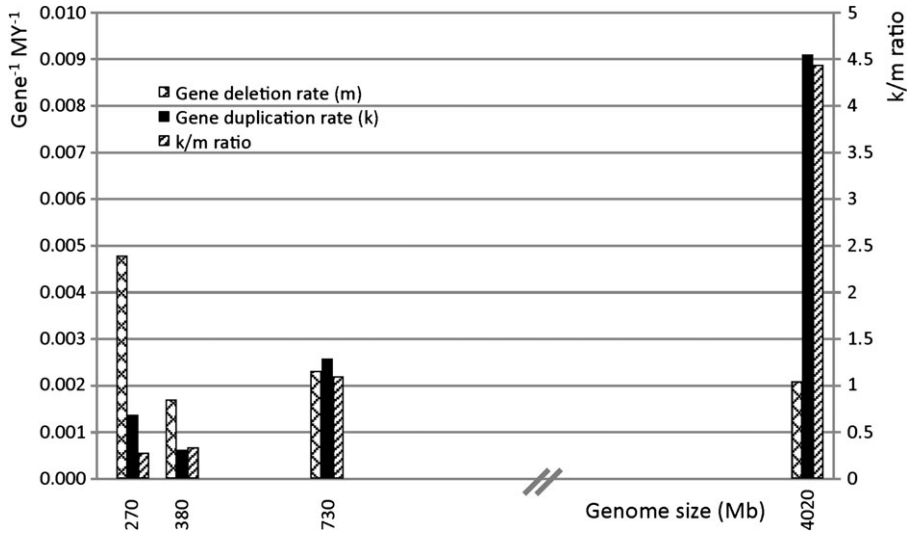

Fig. 3. Relationships between rates of gene space evolution and overall genome size. The left-hand *y* axis is the gene duplication rate (*k*) and gene deletion rate (*m*) in gene^{-1} My^{-1}. The right-hand *y* axis is the *k*/*m* ratio.

(fig. 2*Ae*, table 2). In the smallest genomes, *B. distachyon* (270 Mb) and rice (380 Mb), the deletion rate is greater than the duplication rate (fig. 3 and table 2) suggesting that the gene space of those species is contracting, particularly for the genome of *B. distachyon*. The rates are similar in the larger sorghum (730 Mb) genome and reversed in the large *Ae. tauschii* (4,020 Mb) genome suggesting that the *Ae. tauschii* gene space is expanding.

The estimates of deletion and duplication rates for each lineage can be used to estimate the total initial number of genes per genome ($G_{(0)\text{total}}$) at the time of lineage divergence, based on the total number of genes annotated in rice, sorghum, and *B. distachyon* and predicted in *Ae. tauschii* ($G_{(t)\text{total}}$) and the divergence time *t* (table 2). The total number of genes for *B. distachyon* and rice was obtained from Vogel et al. (2010) and $G_{(t)\text{total}}$ for sorghum was from Paterson et al. (2009). The total number of genes in the *Ae. tauschii* genome was estimated to be 36,371, based on the 90 genes annotated in the nine investigated regions spanning 9.7 Mb, and a total genome size of 3,920 Mb after the removal of 2.6% for satellite DNA (Li et al. 2004). This gene number is comparable to the ~37,000 genes predicted for the average of the three hexaploid wheat genomes based on partial sequence data from hexaploid wheat (Bennetzen JL, San Miguel P, Devos KM, unpublished data) but smaller than the ~41,000 protein-coding genes predicted for the wheat B genome on the basis of chromosome 3B data (Choulet et al. 2010). The total number of genes $G_{(t)\text{total}}$ for internode 2 was estimated by averaging the total initial gene number $G_{(0)\text{total}}$ computed for the *B. distachyon* lineage and for the *Ae. tauschii* lineage (table 2). Similarly, the average of $G_{(0)\text{total}}$ for internode 2 and for the rice lineage provided $G_{(t)\text{total}}$ for internode 1, and the gene number $G_{(t)\text{total}}$ in the ancestor of the BEP and PACCAD clades was obtained by averaging the $G_{(0)\text{total}}$ values from internode 1 and the sorghum lineage. The $G_{(0)\text{total}}$ and $G_{(t)\text{total}}$ estimates for all lineages are summarized in table 2.

The total gene numbers in the extant sorghum and rice genomes are within or close to the range of the inferred gene number in the common ancestor of the BEP and PACCAD clades and the common ancestor of the Ehrhartoideae and Pooideae, respectively (fig. 2). However, significant changes in gene numbers were observed in the *B. distachyon* and *Ae. tauschii* lineages following their divergence from a common ancestor (fig. 2). Compared with the mean initial gene number $G_{(0)\text{total}}$ for the ancestor of the Brachypodieae and Triticeae (28,558 genes), the *B. distachyon* genome has contracted by 3,026 genes (10.6% of the initial number of genes), whereas the *Ae. tauschii* genome has expanded by 7,813 genes (27.4% of the initial number of genes). The gene numbers changed little along internodes 1 and 2 (fig. 2). Total gene numbers in the extant *B. distachyon*, rice, sorghum, and *Ae. tauschii* genomes are correlated with the overall genome size of those species (Pearson's $r = 0.98$, $P = 0.022$).

Gene Space Evolution and Synteny Erosion

Of the 90 genes annotated in the *Ae. tauschii* genome across the nine regions, 54 (60%), 58 (64%), and 57 genes (63%) were found in colinear positions in *B. distachyon*, rice, and sorghum, respectively (supplementary table S4, Supplementary Material online) Nine genes were shared by *Ae. tauschii* and rice and were absent in *B. distachyon*. Five genes were shared by *Ae. tauschii* and *B. distachyon* and were absent in rice. Despite the increasing phylogenetic distance between the species pairs *Ae. tauschii*–*B. distachyon*, *Ae. tauschii*–rice, and *Ae. tauschii*–sorghum, the extent of colinearity between the genomes within each of the three pairs was similar. More than 75% of the genes annotated in the target regions in the three small genome species were colinear with one another and with *Ae. tauschii*.

A casual look at figure 3 suggests that the total rate with which gene synteny is being eroded by the combined effects of gene deletions (gene deletion rate) and insertions of duplicated gene copies (gene duplication rate) in a lineage does

Table 3. Comparison of Rates of Subchromosomal and Total Structural Changes and Single Gene Deletion and Insertion Rates.

Lineage	Rate of Subchromosomal Structural Changes (change My^{-1})	Rate of All Structural Changes (change My^{-1})	$k + m$ (gene My^{-1})
Brachypodium distachyon	0.14	0.34	0.006161
Aegilops tauschii	1.01	1.15	0.011160
Internode 2	0.09	0.13	0.003736
Oryza sativa	0.08	0.08	0.002507
Sorghum bicolor[a]	0.09	0.09	0.004821
S. bicolor[b]	0.09	0.09	0.003998
r^a	0.934 ($P = 0.020$)	0.966 ($P = 0.008$)	
r^b	0.942 ($P = 0.016$)	0.981 ($P = 0.003$)	

NOTE.—Rates of subchromosomal structural changes, principally inversions and translocations, are given in changes per million years (My) (Vogel et al. 2010). The total rate of single gene deletions and duplications ($k + m$) in the *B. distachyon*, rice, sorghum, *Ae. tauschii* lineages, and internode 2 of the phylogenetic tree is calculated from table 2.

[a] For 71 gene scenario.

[b] For 68 gene scenario.

not parallel the overall genome size or parallels it only weakly. To test the relationship quantitatively, Pearson's correlation coefficient r was computed for the rate of synteny loss in a lineage due to the accumulation of gene deletions and duplications, $k + m$, and the total number of genes $G_{(t)\text{total}}$. The correlation was not statistically significant ($r = 0.78$, $P = 0.12$ and $r = 0.74$, $P = 0.09$, for the 71 and 68 ancestral genes, respectively). The lineage of *B. distachyon*, which has the smallest genome, shows the second highest $k + m$ rate, after that of *Ae. tauschii* (fig. 3). The cause of this lack of relationship is the idiosyncratic nature of the gene deletion rate (m), which does not correlate with the total gene number $G_{(t)\text{total}}$ ($r = -0.54$, $P = 0.35$ and $r = -0.42$, $P = 0.48$, for the 71 and 68 ancestral genes, respectively). However, the duplication rate k correlated with the total gene number $G_{(t)\text{total}}$ ($r = 0.93$, $P = 0.022$ and $r = 0.77$, $P = 0.071$, for 71 and 68 ancestral genes scenarios, respectively). $G_{(t)\text{total}}$ also correlated with the k/r ratio ($r = 0.95$, $P = 0.011$ and $r = 0.92$, $P = 0.028$, for the 71 and 68 ancestral gene scenarios, respectively) (fig. 3). The reduced level of colinearity (determined by $k + m$) between the *Ae. tauschii* genome and the other three grass genomes is thus mainly due to the high number of duplications (k value) that have taken place in the *Ae. tauschii* gene space. As speculated in the introduction, there also appears to be a relationship between the rate of gene synteny erosion as measured by the $k + m$ rate along each lineage and the rate with which each lineage has accumulated subchromosomal ($r = 0.93$, $P = 0.020$ and $r = 0.94$, $P = 0.016$, for the 71 and 68 ancestral gene scenario, respectively) and total structural changes ($r = 0.97$, $P = 0.008$ and $r = 0.98$, $P = 0.003$, for the 71 and 68 ancestral gene scenario, respectively) (table 3).

Discussion

Gene Space Evolution in Relation to Genome Size

Detailed sequence comparisons of orthologous BAC clones of grass species differing by as much as 6-fold in genome size had previously shown that some genomes are more stable than others (Chen et al. 1997; Foote et al. 1997; Bennetzen and Ma 2003; Ilic et al. 2003; Bossolini et al. 2007). The small rice genome, for example, often had a gene content that most closely resembled that of the grass ancestor, whereas the 2,500 Mb maize genome had more gene rearrangements (Bennetzen and Ma 2003; Ilic et al. 2003). Genome studies in maize are complicated by the paleotetraploid origin of the maize genome, and the effects of genome size and relaxed purifying selection constraints on duplicated gene pairs are intertwined. Similar concerns may be raised about comparative studies using polyploid wheat, although wheat chromosomes have spent much shorter time at the polyploid level than maize chromosomes, and inferences on genome evolution made in wheat will be less affected by these confounding factors than those made in maize.

Our study is the first comparative analysis that includes comprehensive data on gene-space evolution in a large (4,020 Mb) diploid plant genome. Of the 90 genes annotated in the 9.7 Mb of *Ae. tauschii* genomic sequence originating from nine different regions in four chromosomes, approximately two-thirds of the genes were present in their ancestral locations. In the smaller genomes, the percentage of annotated genes with conserved ancestral positions varied from 88% in sorghum to 90% in *B. distachyon* and rice. In *Ae. tauschii*, at least 27% of the annotated genes had been inserted into the studied genomic regions during the past 35.8 My. The number of insertions that occurred into the syntenic *B. distachyon*, rice, and sorghum regions during similar or longer time frames was significantly lower (5%, 3%, and 12.5%, respectively). These superficial observations leave little doubt that gene space in smaller grass genomes is more stable than that in the large *Ae. tauschii* genome. This has already been shown for the rates of accumulation of structural chromosome changes (Luo et al. 2009; Vogel et al. 2010), which were shown here to correlate highly with the level of gene synteny erosion (determined by the $k + m$ values). This correlation suggests that structural chromosome evolution by chromosome inversions and translocations, and gene space evolution by single gene duplications and deletions is coupled and may have common causes.

The quantification of gene duplication and gene deletion rates in the four terminal lineages and two internodes (fig. 2) provided insights into the mechanisms of grass genome evolution going far beyond these observations. Similar to the processes determining the size of an entire genome (Devos et al. 2002; Ma and Bennetzen 2004), the size of a gene space results from two opposing processes, one removing genes from the genome and the other inserting them into the genome, principally by gene duplication. The assumption that gene duplication is the driving force behind the insertions is supported by the fact that 70% of the insertions in rice, sorghum, and *B. distachyon* in our study were the result of either tandem duplications of ancestral genes already present in the region or of dispersed duplications of genes elsewhere in the genome. The

computation of the gene deletion and gene duplication rates was based on the number of genes located in a 9.7 Mb sample of the *Ae. tauschii* genome, which is about 0.25% of the genome. Similar percentages of the rice, *B. distachyon*, and sorghum genomes were used to estimate gene deletion and gene duplication rates for those genomes. The rate estimates were then used to compute the total number of genes $G_{(0)\text{total}}$ that were present in the ancestral genomes. For each node on the phylogenetic tree, two independent estimates were obtained (fig. 2). In all cases, the independent estimates were within a narrow range, showing that our sampling and computations based on it were representative for the evolutionary lineages.

The estimates of gene deletion rates among the four lineages over 52.5 My of evolution were similar, ranging from 1.7×10^{-3} to 3.9×10^{-3} per gene per My, although the absence of statistical differences should not be taken as meaning that there are no real differences among the lineages. The deletion rate in the *B. distachyon* and *Ae. tauschii* lineages in the period following their divergence from a common ancestor is 4.8×10^{-3} and 2.1×10^{-3} per gene per My, respectively (table 2). A deletion rate of 3.2×10^{-3} per gene per My for the *Ae. tauschii* and *Triticum urartu* lineages following their divergence 2.7 Ma estimated on the basis of analysis of 3,195 genes is also within this range (Dvorak and Akhunov 2005).

A different situation was encountered for gene duplication rates. The *Ae. tauschii* gene duplication rate of 6.8×10^{-3} per gene per My over the entire 52.5 My was significantly greater than gene duplication rates in the remaining three lineages ($0.57–2.5 \times 10^{-3}$ per gene per My) over the same time period. If we consider only the duplications that occurred in the *Ae. tauschii* and *B. distachyon* genomes after their divergence from a common ancestor, the difference in duplication rates becomes even more pronounced with *k* values in *Ae. tauschii* and *B. distachyon* being 9.1×10^{-3} and 1.4×10^{-3} per gene per My, respectively (table 2). It should be noted that the gene duplication rate in *Ae. tauschii* is almost certainly higher than 9.1×10^{-3} because the region also contained 40 pseudogenes (ratio of pseudogenes/genes = 1/2.2), which originated through gene duplications. Smaller grass genomes, on the other hand, appear to contain fewer pseudogenes (pseudogene/gene ratio is $\leq$1/20) (Paterson et al. 2009; Thibaud-Nissen et al. 2009; Vogel et al. 2010), which is consistent with a lower duplication rate in those genomes. We did not include pseudogenes in our analysis because they had not been systematically annotated in the *B. distachyon*, rice, and sorghum genomes. Our data suggest that the expansion of genome size in the *Ae. tauschii* lineage over the past 32–39 My was accompanied by an expansion of gene space.

The overall genome size expansion in the *Ae. tauschii* lineage almost certainly reached the size characteristic of modern Triticeae genomes before the Triticeae radiation about 10 Ma because all Triticeae species have genomes larger than 3.3 Gb (Dvorak 2009). Genome size expansion during the radiation of the Triticeae was relatively minor, being equivalent to a factor of 2 at most (Dvorak 2009) rather than an order of magnitude that characterized the initial stages of evolution in the Triticeae lineage. Slowing down of the overall genome expansion in the recent past of the Triticeae lineage may have been paralleled by a leveling off of the gene duplication rate, which was estimated for interchromosomal gene duplications to be 2.9×10^{-3} per locus per My during the past 2.7 My (Dvorak and Akhunov 2005). Of the 25 duplicated genes in *Ae. tauschii*, 4 (16%) were tandem duplications. Even if the rate reported by Dvorak and Akhunov is increased by 16%, the duplication rate of 9.1×10^{-3} per gene per My observed here is still higher. An independent line of evidence suggesting an increase in gene number in the Triticeae evolutionary lineage comes from the wheat chromosome 3B study (Choulet et al. 2010). Although no rate quantification was attempted in that study, the estimate of 41,000 genes for the B genome is indicative of genome size expansion given that the initial number of genes $G_{(0)\text{total}}$ in the gene space of the genome ancestral to the Triticeae and Brachypodieae, inferred from our analysis, was in the range 28,289–28,826. The larger number of genes estimated for the B genome might be due to the larger size of the B genome relative to the D genome.

The estimates of $G_{(0)\text{total}}$ for *B. distachyon* and *Ae. tauschii* suggest that gene numbers remained more or less constant in the lineage until the divergence of the *B. distachyon* and *Ae. tauschii* lineages 35.8 Ma. The common ancestor of the two lineages had about 28,558 genes (fig. 2 and table 2). The gene number then increased by 7,813 genes (about 27%) to 36,371 genes in the *Ae. tauschii* lineage but contracted by 3,026 genes (about 11%) to 25,532 genes in the *B. distachyon* lineage. The measure that appears to reflect gene space evolution most faithfully is the *k*/*m* ratio, which differs by an order of magnitude between these two Pooideae lineages. The value of the *k*/*m* ratio in the four species analyzed is correlated to both gene space size and overall genome size.

What may cause these dramatic shifts in gene duplication and gene deletion rates that are seen during grass evolution? The logical culprits are TEs. Variation in genome size among related diploid species is believed to be principally caused by the accumulation or loss of TEs. Some TEs have been shown to propagate gene fragments during transposition (Kapitonov and Jurka 2001; Jiang et al. 2004; Lai et al. 2005), which may facilitate the evolution of new genes (Morgante et al. 2005; Morgante 2006). A closer look at the duplicated sequences propagated by *MULEs* or *Helitrons*, however, showed that most of them were pseudogenes (Juretic et al. 2005; Lai et al. 2005; Morgante et al. 2005). Nevertheless, there are examples of complete genes being transposed by TEs in Triticeae genomes and acquisition of novel exons and novel transcriptional regulation (Akhunov et al. 2007). Because the number of pseudogenes in an *Ae. tauschii* contig is correlated with the number of inserted genes into that contig, both gene categories likely result from the same duplication/insertion mechanism. Repair of double-stranded DNA breaks introduced into DNA by the

presence of some TEs (Wicker et al. 2010) may be the actual common mechanism that accounts for the positive relationship between the number of duplicated genes and the overall genome size observed here.

The fact that the magnitude of gene duplication rate and magnitude of gene deletion rate in individual grass lineages are not correlated suggests that different mechanisms govern the two processes. Studies in *Arabidopsis*, rice, and wheat have shown that DNA can be removed through both illegitimate recombination and unequal homologous recombination (Devos et al. 2002; Wicker et al. 2003; Ma et al. 2004). Although those studies were carried out on the repeated DNA fraction, these mechanisms, in particular illegitimate recombination, can also lead to the removal of genes (Devos et al. 2002). Gene removal through illegitimate recombination is largely independent of the repeat content of a genome, and hence of genome size. In the four species analyzed, there was no significant correlation between the gene deletion rate and genome size.

Brachypodium–Triticeae Synteny and the Utility of *B. distachyon* in Triticeae Genomics

The faster rate of gene evolution in the large *Ae. tauschii* genome compared with the smaller grass genomes, an observation that has also been made for gross chromosomal rearrangements (Luo et al. 2009), brings into question whether or not *B. distachyon* would be a better genomic model for Triticeae than rice. In our study, nine genes were shared by *Ae. tauschii* and rice and were absent in *B. distachyon*, and five genes were shared by *Ae. tauschii* and *B. distachyon* and were absent in rice. Overall, 60%, 64%, and 63% of the *Ae. tauschii* genes were in colinear positions in *B. distachyon*, rice, and sorghum, respectively. This level of colinearity is similar to that uncovered in a comparison of the *Ae. tauschii* linkage map with the sorghum, *B. distachyon*, and rice genome sequences in which approximately 64%, 65%, and 66% of the loci on the *Ae. tauschii* genetic map were colinear with genes along the sorghum, *B. distachyon*, and rice pseudomolecules, respectively (Luo et al. 2009; Vogel et al. 2010). Hence, for comparative genomic applications in the Triticeae, *B. distachyon* and rice are about equally useful as models. Our confirmation of the equivalent synteny involving four chromosomes of diploid *Ae. tauschii* with *B. distachyon*, rice, and sorghum of a similar observation made in wheat chromosome 3B (Choulet et al. 2010) shows that this surprising fact is not caused by polyploidy or restricted to a single chromosome. Because synteny is a function of gene deletion and gene duplication rates and because those rates are not constant but differ among and along lineages, the rate of synteny erosion is lineage dependent and corresponds poorly to phylogenetic distance. The principal asset of *B. distachyon* over rice, in addition to easier cultivation, is its greater sequence similarity with the Triticeae genomes, which does correlate with phylogenetic distance (Vogel et al. 2010).

Conclusion

A comprehensive comparison of 9.7 Mb of sequence data from the 4.02 Gb *Ae. tauschii* genome with the genome sequences of rice, sorghum, and *B. distachyon* and the quantification of gene deletion and gene duplication rates in different branches of the phylogenetic tree relating those four species suggest that gene numbers can both increase and decrease over evolutionary times and that the rate of this process is lineage dependent. The primary determinant of the overall gene number in a genome appears to be the gene duplication rate, which is related to overall genome size. Gene duplications are an important mechanism for the evolution of new genes, which may facilitate plant adaptation. The hypothesized relationship between genome size and gene space size is based on four grass genomes, which is an inadequate sample for the formulation of an evolutionary theory. If future analyses of a larger sample of plant genomes substantiate this hypothesis, the relationship between overall genome size, gene space size, and TE content may constitute a foundation for understanding the role of TEs and genome size in plant evolution and contribute to the clarification of one of the major biological questions that remains unanswered.

Supplementary Material

Supplementary tables S1–S4 and text S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the Joint Technology Center J. Craig Venter Institute (JCVI) for carrying out sequencing and BAC assembly, and the JCVI Informatics group for data management and data submission to public archives. We thank Xiangyang Xu (University of Georgia, Athens) for mapping contigs on the wheat deletion lines and Richard Howitt (University of California, Davis) for checking the mathematics. This work was supported by the National Science Foundation Plant Genome Research Program Cooperative Agreements DBI-0077766 and DBI-0638558.

References

Akhunov ED, Akhunova AR, Dvorak J. 2007. Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. *Mol Biol Evol*. 24:539–550.

Akhunov ED, Akhunova AR, Linkiewicz AM, et al. (31 co-authors). 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proc Natl Acad Sci U S A*. 100:10836–10841.

Arumuganathan K, Earle ED. 1991. Nuclear DNA content of some important plant species. *Plant Mol Biol Rep*. 9:208–218.

Barker NP, Clark LG, Davis JI, et al. (13 co-authors). 2001. Phylogeny and subfamilial classification of the grasses (*Poaceae*). *Ann Mo Bot Gard*. 88:373–457.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev*. 15:621–627.

Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell*. 9:1509–1514.

Bennetzen JL, Ma J. 2003. The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Curr Opin Plant Biol*. 6:128–133.

Bossolini E, Wicker T, Knobel PA, Keller B. 2007. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J*. 49:704–717.

Chen M, SanMiguel P, de Oliveira AC, Woo SS, Zhang H, Wing RA, Bennetzen JL. 1997. Microcolinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes. *Proc Natl Acad Sci U S A*. 94:3431–3435.

Choulet F, Wicker T, Rustenholz C, et al. (24 co-authors). 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22:1686–1701.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res*. 12:1075–1079.

Dubcovsky J, Dvorak J. 2007. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866.

Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan LL, Shiloff BA, Bennetzen JL. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol*. 125:1342–1353.

Dvorak J. 2009. Triticeae genome structure and evolution. In: Feuillet C, Muehlbauer GJ, editors. Genetics and genomics of Triticeae. Heidelberg: Springer-Verlag. p 685–711.

Dvorak J, Akhunov ED. 2005. Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. *Genetics* 171:323–332.

Estill JC, Bennetzen JL. 2009. The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods* 5:8–18.

Foote T, Roberts M, Kurata N, Sasaki T, Moore G. 1997. Detailed comparative mapping of cereal chromosome regions corresponding to the *Ph1* locus in wheat. *Genetics* 147:801–807.

Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci U S A*. 94:6809–6814.

Ilic K, SanMiguel PJ, Bennetzen JL. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc Natl Acad Sci U S A*. 100: 12265–12270.

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

Jakob SS, Meister A, Blattner FR. 2004. The considerable genome size variation of *Hordeum* species (*Poaceae*) is linked to phylogeny, life form, ecology, and speciation rates. *Mol Biol Evol*. 21:860–869.

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

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE. 2005. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res*. 15:1292–1297.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A*. 98:8714–8719.

Kihara H. 1944. Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare* (Japanese). *Agric Hort (Tokyo)*. 19:13–14.

Lai JS, Li YB, Messing J, Dooner HK. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci U S A*. 102:9068–9073.

Lewis SE, Searle S, Harris N, et al. (19 co-authors). 2002. Apollo: a sequence annotation editor. *Genome Biol*. 3:research0082. 1–research0082.14.

Li W, Zhang P, Fellers JP, Friebe B, Gill BS. 2004. Sequence composition, organization and evolution of the core Triticeae genomes. *Plant J*. 40:500–511.

Luo MC, Deal KR, Akhunov ED, et al. (32 co-authors). 2009. Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc Natl Acad Sci U S A*. 106:15780–15785.

Luo MC, Thomas C, You FM, et al. (10 co-authors). 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* 82:378–389.

Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res*. 14:860–869.

Ma JX, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A*. 101:12404–12410.

McFadden ES, Sears ER. 1946. The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J Hered*. 37:81–89.

Morgante M. 2006. Plant genome organisation and diversity: the year of the junk! *Curr Opin Biotechnol*. 17:168–173.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet*. 37:997–1002.

Myers EW, Sutton GG, Delcher AL, et al. (29 co-authors). 2000. A whole-genome assembly of *Drosophila*. *Science*. 287:2196–2204.

Paterson AH, Bowers JE, Bruggmann R, et al. (45 co-authors). 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–556.

SanMiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann Bot*. 82:37–44.

SanMiguel P, Tikhonov A, Jin Y-K, et al. (11 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765–768.

Schnable PS, Ware D, Fulton RS, et al. (157 co-authors). 2009. The B73 maize genome: complexity, diversity and dynamics. *Science* 326:1112–1115.

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. On the tetraploid origin of the maize genome. *Comp Funct Genomics* 5:281–284.

Thibaud-Nissen F, Ouyang S, Buell CR. 2009. Identification and characterization of pseudogenes in the rice gene complement. *BMC Genomics*. 10:317–329.

Vogel JP, Garvin DF, Mockler TC, et al. (135 co-authors). 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768.

Vogel KP, Arumuganathan R, Jensen KB. 1999. Nuclear DNA content of perennial grasses of the Triticeae. *Crop Sci*. 39:661–667.

Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res*. 20:1229–1237.

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J*. 26:307–316.

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu Z-D, Dubcovsky J, Keller B. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell* 15:1186–1197.