Genome-Wide Analysis of Syntenic Gene Deletion in the Grasses

James C. Schnable[1], Michael Freeling[1], and Eric Lyons[2,]*

[1]Department of Plant and Microbial Biology, University of California-Berkeley

[2]iPlant Collaborative, Bio5 Institute, University of Arizona

*Corresponding author: E-mail: ericlyons@email.arizona.edu.

Accepted: 17 January 2012

Abstract

The grasses, Poaceae, are one of the largest and most successful angiosperm families. Like many radiations of flowering plants, the divergence of the major grass lineages was preceded by a whole-genome duplication (WGD), although these events are not rare for flowering plants. By combining identification of syntenic gene blocks with measures of gene pair divergence and different frequencies of ancient gene loss, we have separated the two subgenomes present in modern grasses. Reciprocal loss of duplicated genes or genomic regions has been hypothesized to reproductively isolate populations and, thus, speciation. However, in contrast to previous studies in yeast and teleost fishes, we found very little evidence of reciprocal loss of homeologous genes between the grasses, suggesting that post-WGD gene loss may not be the cause of the grass radiation. The sets of homeologous and orthologous genes and predicted locations of deleted genes identified in this study, as well as links to the CoGe comparative genomics web platform for analyzing pan-grass syntenic regions, are provided along with this paper as a resource for the grass genetics community.

Key words: polyploidy, gene loss, synteny, Poaceae, speciation.

Introduction

Evidence of ancient polyploidies, or whole-genome duplications (WGDs), are found throughout all eukaryotic lineages (Dehal and Boore 2005). These duplications, whether auto- or allopolyploidy events, instantly create copies of all genes and associated regulatory sequences contained within the nuclear genome of a species. Interestingly, ancient WGDs tend to be associated with adaptive radiations of multiple lineages, although the causality of this relationship remains controversial (Soltis et al. 2009).

Multiple explanations for the association between WGD and species radiations have been proposed. At a mechanistic level, the reciprocal loss of duplicated genes from one of the multiple subgenomes of a polyploid organism, a process known as fractionation—or in older literature as "diploidization"—could increase the speed with which hybrid incompatibly develops between populations (Lynch and Force 2000). It has also been suggested that ancient WGDs tend to be contemporaneous with major extinction events (van de Peer et al. 2009); polyploid species that survived such events would be expected to radiate into the abundant newly vacated niches left by the wave of extinctions. Finally, it may be that WGDs, by creating a new source of redundant genes suitable for co-option for novel functions or subfunctionalization, increase the potential for niche specialization (De Bodt et al. 2005) or morphological innovation (Freeling and Thomas 2006).

Following a WGD, redundant copies of many genes are removed from the genome by fractionation (Langham et al. 2004). A study of synthetic Bassica allotetraploids has reported major genomic rearrangements and deletions within as few as five generations (Osborn et al. 2003). In addition, duplicate gene deletion continues at significant levels in maize 5–12 Myr after polyploidy (Swanson-Wagner et al. 2010; Woodhouse et al. 2010; Schnable, Springer, et al. 2011); only 47% of maize–sorghum syntenic genes are still represented by genes or gene fragments at both duplicate locations within the maize genome (Woodhouse et al. 2010). A study in yeast documented ongoing loss of duplicate gene copies throughout the entire period since WGD in that lineage (Scannell et al. 2006).

The loss of duplicate genes is biased in multiple ways. The first is biased retention of both duplicate copies of certain classes of genes following WGD. These classes include genes encoding members of large multiprotein complexes, transcription factors (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Scannell et al. 2006; Freeling et al. 2007), and genes associated with large numbers of conserved noncoding regulatory elements (Schnable, Pedersen, et al. 2011). The loss of genes is also biased between duplicated regions. After first being observed in arabidopsis and maize (Thomas et al. 2006; Woodhouse et al. 2010), this bias was found to be a general property of eukaryotic WGDs (Sankoff et al. 2010). The bias in gene loss is a property of whole parental genomes in both maize and *Arabidopsis suecica* (Chang et al. 2010; Schnable, Springer, et al. 2011) and therefore may represent a useful mark for reconstructing ancestral subgenomes across organisms with ancient polyploidy.

All grass species sequenced to date share an ancient WGD tentatively dated to approximately 70 Ma (Paterson et al. 2004; Yu et al. 2005) contemporaneous with the emergence of phytoliths representing extant grass families in the fossil record (Prasad et al. 2005). Of all plant families, the grasses are represented by the most published sequenced genomes—brachypodium, maize, rice, and sorghum—representing three subfamily-level grass lineages (Goff et al. 2002; Paterson et al. 2009; Schnable et al. 2009; The International Brachypodium Initiative 2010). It is likely the grasses will retain this distinction for the foreseeable future with genome assemblies for additional grass species currently available under prepublication restrictions, in the process of being sequenced, or in the planning stages of being sequenced. Given the economic and ecological significance of the grasses and the demand for fast porting of functional information among grass species, there is a need for automated, yet accurate, tools to identify and classify orthologs and homeologs in many-to-many genomic comparisons. However, a number of known genomic events complicate the assignment of orthologous genes between grass species.

In addition to the previously mentioned ancient WGD shared by all grasses, a relatively recent WGD is found within maize, dated to 5–12 Ma, just subsequent to the divergence of this lineage from the common ancestor of sorghum and the core of its tribe (Swigoňová et al. 2004). As a result, there are two homeologous locations within the maize genome coorthologous to any single location in the genomes of rice, sorghum, and brachypodium (fig. 1). The genomic relationships created by the pregrass WGD and the second duplication in the maize lineage are summarized in figure 1. The size of the maize genome is also more than twice the next largest sequenced grass, largely as a result of multiple waves of transposon amplification in the last several million years (Baucom et al. 2009; Schnable et al. 2009). Syntenic analysis of the grasses has also detected evidence of more ancient WGD events shared by most, if not all, monocot species (Tang et al. 2010). These more ancient duplicated blocks must be identified and removed from genomic comparisons aimed at identifying duplicates from the more recent tetraploidy shared by all grasses. Finally, duplicated regions in all grasses—located on chromosomes 11 and 12 of rice and chromosomes 5 and 8 of sorghum—have a peculiar evolutionary history and have evolved in concert since the pregrass tetraploidy (Wang et al. 2011). These highly similar duplicate regions pose significant issues for some methods of automated ortholog/homeolog classification based on average sequence similarity or evolutionary distance.

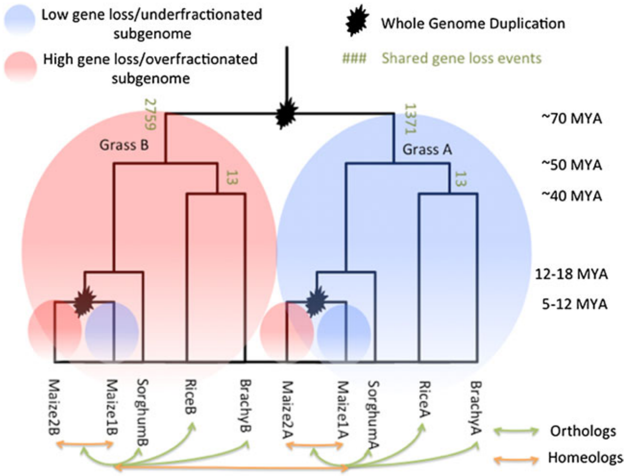

Fig. 1.—Subgenome relationships in the grasses. Relationships of a genomic location in the grasses, taking into account both the ancient WGD in the ancestor of all sequenced grass species and the more recent genome duplication in the maize lineage. Each duplication creates separate homeologous low gene loss (underfractionated) and high gene loss (overfractionated) subgenomes. Two loci are orthologous if the branch point where they diverged represents a speciation event (no mark) or homeologous if the branch point where they diverged is a WGD (marked with a starburst). Branch lengths not to scale.

Genomes and Genomic Regions

Whole-genome duplication: Abbreviation WGD. The duplication of an entire genome. WGDs generate polyploid organisms. May be subclassified as auto- or allo- denoting a single parental genome or multiple parental genome origin, respectively.

Diploid: Denotes that a genome has two homologous copies of each chromosome.

Polyploid: Denotes that a genome has more than two homologous copies of each chromosomes.

Subgenome: The constituent genomes within a polyploid species, each of which is derived from the entire genome of a parent or ancestral species and prior to fractionation, contained all the genes found throughout the clade within which the polyploid species falls.

Syntenic region: Two or more homologous genomic regions descended from a common ancestral genomic region. Syntenic regions are evidence by homologous genes arranged in a collinear order.
Fractionation: The loss of one or the other duplicated gene copy following a WGD. (near synonym: diploidization)
Fractionation bias: The uneven distribution of gene deletions between duplicated genomic regions following WGD.
Underfractionated: The copy of a duplicate chromosomal region from which fewer genes were lost.
Overfractionated: The copy of a duplicate chromosomal region from which more genes were lost.

Evolutionary Relationships and Types
Homolog: Of common ancestry. Homologous genes and genomic regions are derived from a common ancestral gene or genomic region.
Orthologs: Homologous genes or genomic regions derived from the divergence of lineages.
Paralog: Homologous genes or genomic regions derived from their duplication within a lineage.
Homeologs: The subset of paralogs created by WGD. (synonyms: ohnolog; syntenic paralog)
Pan-grass gene: A gene present in the ancestral preduplicated genome of the grasses remaining at its ancestral position. Pan-grass genes are detected though comparison of syntenic region within and among grass genomes.
Ancestral gene: A gene hypothesized to be present in the ancestral genome at its current extant location. Ancestral genes are defined by their conserved genomic position in multiple lineages or subgenomes.

Many previously published methodologies for ortholog identification use some variation of best BLAST hit. In order to identify WGD events, the evolutionary distances of homologous gene pairs are often calculated using synonymous mutation or 4DTV values, and the histogram of values interrogated for distinct peaks (Tuskan et al. 2006; Barker et al. 2008).

A number of tools do incorporate identification of syntenic blocks as discussed: (Soderlund et al. 2011). In comparisons between multiple flowering plant species, all with extensive histories of WGD, it is necessary to distinguish between more recent and more ancient syntenic blocks (Tang et al. 2011).

In this paper, we demonstrate a method for ortholog/homeolog classification based on the identification of syntenic blocks of genes in inter- and intraspecies genomic comparisons followed by the calculation of aggregate divergence data for all gene pairs within the block. Our method permits the subsequent identification of high-confidence gene loss/transposition events that are crucial for the study of genome evolution following polyploidy. In addition, this method permits the identification of two subgenomes shared by all sequenced grasses—a low gene loss underfractionated subgenome (GrassA) and a high gene loss overfractionated subgenome (GrassB)—as previously demonstrated for the much younger maize tetraploidy (Schnable, Springer, et al. 2011). We use this method to identify orthologs and homeologs between four grass species with published genome sequences and reconstruct the ancestral subgenomes comprising each grass' modern genome. We assign gene loss events to nodes on the grass phylogenetic tree and search for reciprocally lost duplicated genes which might have contributed to reproductive isolation during the radiation of the major grass lineages.

Materials and Methods

Generating Lists of Syntenic Orthologs/Homeologs

Lists of syntenic gene pairs were initially generated for all pairwise comparisons—including self–self comparisons—using the SynMap utility of CoGe (Lyons et al. 2008) with the parameters described in supplementary table S3 (Supplementary Material online)of this paper. Individual stretches of syntenic genes were merged into larger syntenic blocks using the method described in (Yang 2007).

Synonymous substitution rates between individual syntenic gene pairs were calculated within the SynMap utility for aligned coding sequences of gene pairs guided by the alignment of the translated coding sequences of gene pairs by nwalign (http://pypi.python.org/pypi/nwalign/). Synonymous substitution rates for these aligned sequences were calculated by a customized version of CODEML (Alexandrov et al. 2009).

Syntenic blocks containing 12 or more gene pairs were assigned to an evolutionary event, whether speciation (orthologous) or WGD (homeologous), based on a unified synonymous substitution rate (K_s) for genes contained within the block. This unified synonymous substitution rate is defined as the average synonymous substitution rate among gene pairs contained within the syntenic block after discarding the most diverged two-thirds of genes contained within the syntenic block. The calculation of synonymous substitution rates is very sensitive to errors in gene model annotation or sequence alignment, and examining only the lowest one-third of K_s values provides sufficient data set to differentiate sequence blocks while eliminating any distortion from the very high substitution rates calculated between incorrectly aligned coding sequences. Grass genomes also include a class of high third base pair position GC content genes that generate unreliable synonymous substitution rate calculations (Alexandrov et al. 2009).

These calculations produced two fully distinct peaks for synonymous substitution rates of syntenic gene blocks for interspecies comparisons: one corresponding to orthologous syntenic blocks created by speciation and the other to homeologous syntenic blocks resulting from the pregrass tetraploidy. Intraspecies comparisons identified a single fully distinct peak of homeologous syntenic blocks resulting from the the pregrass duplication in sorghum, rice, and brachypodium, and the more recent maize lineage-specific tetraploidy within maize (supplementary fig. S4, Supplementary Material online).

Joining Pairs into Orthologous Blocks and Identifying Lost Orthologs

Homeologous and orthologous pairs of genes defined by inter- and intraspecies comparison were merged using in-house python scripts to produce lists of pan-grass syntenic genes. When no ortholog of a syntenic group of genes was identified in a species, a predicted orthologous location was identified using the first orthologously conserved genes within that genome up and downstream of the missing gene. If these conversed genes were separated by more than 1 MB or were located on different chromosomes, the group of genes was considered to have no syntenic coverage in the missing species.

When a predicted orthologous region was identified, a three step process was used to confirm the absence of a syntenic ortholog. First, all annotated genes within the predicted orthologous region were compared using LASTZ (Harris 2007) with all members of the group of syntenically conserved genes in other species. Any gene with sequence similarity to the existing group of conserved syntenic genes was considered a conserved ortholog and added to the syntenic group. If no gene within the predicted region was hit, the sequence of the entire predicted region was extracted and compared with the existing group of conserved syntenic genes using LASTZ with default settings. Any hit with a score of 3,000 or greater within the region was considered an unannotated conserved gene or gene fragment. Gaps with no syntenic matches to either annotated genes or unannotated sequences were further subdivided between those where a gap of 50 or more N_s were present at the predicted location and those were there were no annotated gaps within the predicted location.

If the same gene was included in multiple syntenic groupings, the group with fewer identified orthologous and homeologous genes was removed from our comparison. Syntenic groups were three or more genes not classified as local duplicates of each other were all identified as orthologs within the same species were also removed from the data set. These predominately consisted of sequences that were treated as separate genes in some species but merged into single gene in others.

Putative homeologous gene pairs identified only in a single species where neither copy of the gene was sorted with evidence of syntenic orthologs in any other grass species were omitted from our analysis.

Local duplicate genes were defined as a series of homologous genes interrupted by now less than 20 intervening genes (40 genes in maize, given the greater gene density of the maize working gene set). Homology was defined using the same parameters used by SynMap.

Assignment of Regions as Over/Underfractionated

Seventeen pairs of large contiguous homeologous regions were manually defined using a rice–rice syntenic dotplot. Regions with distinct elbows, as seen in the comparison of rice chromosomes 8 and 9, were split into multiple segments. For each homeologous pair of regions, the number of pan-grass syntenic genes present in one region without any evidence of conserved homeologs in the other was extracted. In three pairs of regions, including the recombination prone end of rice chromosomes 11 and 12, the difference in pan-grass homeologs retained at syntenic locations was less than 10%. These regions were excluded from further analyses. In the remained 14 cases, it was possible to assign one region to the overfractionated pan-grass subgenome and the other to the underfractionated pan-grass subgenome. As the mechanism of fractionation has previously been shown to be almost entirely single gene deletions (Woodhouse et al. 2010), *P* values were calculated using a binomial approach with a null hypothesis that gene deletion was equally likely in both homeologous regions.

Region Loss Methods

The sorghum genome was scanned for cases where 40 or more sequential genes lacked identified sytenic orthologs from the same maize subgenome. Cases where overlapping gaps in the coverage of both maize subgenomes were discarded as these likely represent regions where sorghum-specific insertions and rearrangements have made it impossible to detect synteny. The remaining 16 apparent deletions were classified based on the average number of maize genes each sorghum gene within the region served as the best hit for. Based on 1,000 permutations of random sets of sorghum genes, we determined a median region averages 1.996 best hits of maize genes to each sorghum gene within the region, with 95% confidence bounds between 1.175 and 4.025 best hits of maize genes to each sorghum gene in the region (supplementary fig. S9, Supplementary Material online). Three putative deletions fell outside of this confidence interval and were manually investigated using the CoGe toolkit.

Visualizing Fractionation Bias

In figure 5 and supplementary figure S8 (Supplementary Material online), biased gene content between duplicate regions is computed using the number of pan-grass genes located between neighboring homeologous gene pairs in two syntenic regions. The number of intervening gene pairs

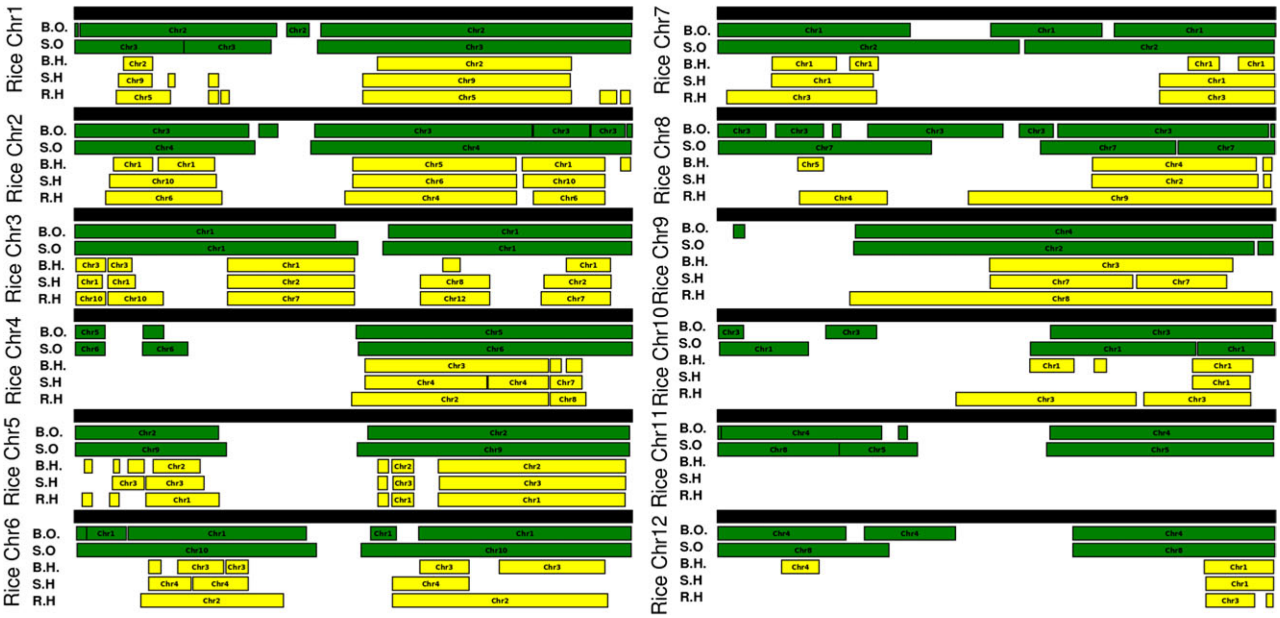

Fig. 2.—Orthologous and homeologous coverage of the rice genome by syntenic regions in the sorghum, brachypodium, and rice genomes. Orthologous syntenic regions are marked in green and homeologous ones are marked in yellow. Coverage is scaled by gene counts, not nucleotides, which will tend to accentuate the gene rich chromosome arms and deemphasize the gene poor pericentromeric regions. B.O. = Brachypodium orthologous region. S.O = Sorghum orthologous region. R.H. = Rice homeologous region. B.H. = Brachypodium homeologous region. S.H. = Sorghum homeologous region.

is averaged across a sliding window of 30 homeologous gene pairs. Homeologous pairs separated by greater than or equal to eight pan-grass genes are omitted from this analysis as previous work has shown that these likely represent small translocations (Woodhouse et al. 2010).

Results

Identification of Syntenic Gene Sets and Lost Genes

Syntenic gene sets were generated using SynMap (Lyons et al. 2008), and both inter- and intraspecies comparisons between all sequenced grasses (for details, see Materials and Methods). Our primary data set consisted of 16,923 orthologous gene groups where genes or predicted locations could be identified in the three grass species which have remained diploid since the radiation of the major grass lineages (supplementary data set S1, Supplementary Material online). Figure 2 shows orthologous regions of the sorghum and brachypodium genomes and homeologous regions of the sorghum, rice, and brachypodium genomes aligned to the 12 chromosomes of the modern rice genome. Similar displays are possible using either the sorghum or brachypodium genomes as reference genomes (supplementary fig. S1 and S2, Supplementary Material online).

Our data set included a significant number of predicted locations for orthologous genes where no orthologous gene was identified. These "missing" data points could be divided into three categories.

1. Recent pseudogenes or missed gene annotations: Cases where no annotated gene model matched the genes conserved in other grass species, but sequence homologous to the genes found in other species was identified at the predicted orthologous location (fig. 3*A*).
2. Gaps in sequence: Cases where no sequence similar to the missing gene was identified, but the predicted orthologous location included a gap in the pseudomolecule assembly, raising the possibility that the region containing the missing gene was not sequenced or assembled (fig. 3*B*).
3. True deletions: Cases where no gaps or unannotated homologous sequence were present (fig. 3*C*).

The higher frequency of gaps in the sorghum genome assembly meant that only a small number of high confidence gene loss events were identified in this lineage (fig. 4*A*). However, because sorghum diverged before the split of the rice and brachypodium lineages, it is possible these genes were inserted into their present location because the rice/brachypodium lineage diverged from sorghum. Brachypodium contains more of both class #1 missing genes—no gene model but syntenic homologs sequence—($P < 0.0001$, chi-square test, degree of freedom [df] = 2) and class #3 missing genes—high confidence gene losses—($P < 0.0001$, chi-square test, df = 2) (fig. 4*A*).

The rice orthologs of deleted brachypodium genes are not significantly enriched in any of the rice GOSlim annotations (Ouyang et al. 2007) relative to other syntenically

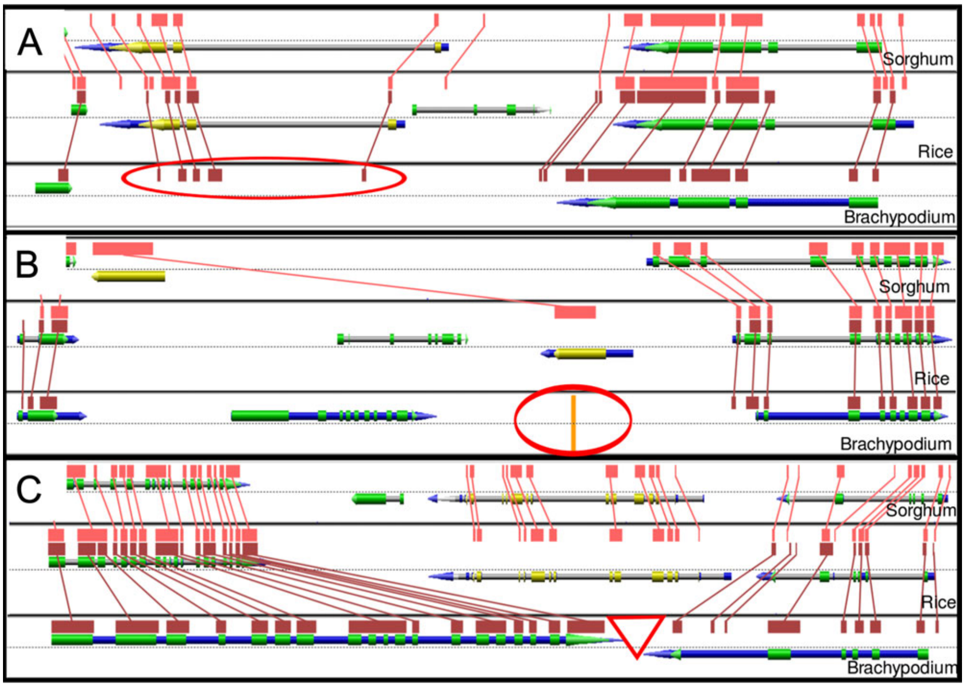

Fig. 3.—Three categories of potential gene loss identified by syntenic analysis. In each case, a gene conserved in sorghum (top panel) and rice (middle panel) is shown along with the syntenically predicted orthologous location in the brachypodium genome (bottom panel). Each panel represents a genomic region with the dashed line separating the top and bottom strands of DNA. Gene models are composite arrows with gray representing the extent of the gene, blue the mRNA, and green/yellow protein-coding sequence. (*A*) No gene model corresponding to the conserved gene in rice (Os12g42550) and sorghum (Sb08g022000) was annotated in brachypodium, however, unannotated sequence present at the predicted orthologous location in brachypodium (marked with a red circle) is similar to the coding sequence of the annotated rice and sorghum genes. (*B*) Neither an annotated gene nor unannotated sequence in brachypodium corresponds to the syntenically conserved gene in rice (Os07g43700) and sorghum (Sb02g040190). However, a gap in the brachypodium genome assembly (orange bar marked with the red circle) raises the possibility that the brachypodium ortholog of these genes was simply not captured during the whole genome shotgun sequencing of the brachypodium genome or not correctly assembled into the pseudomolecule. (*C*) A high confidence gene deletion. The example gene Sb04g035110/Os02g54120 has a predicted orthologous location (red triangle) which does not contain an orthologous gene, unannotated homologous sequence, or a gap in the pseudomolecule assembly.

conserved rice genes (supplementary table S1, Supplementary Material online). Lost genes were compared with the population of syntenically retained genes rather than all rice genes because we found that genes retained syntenically in rice and at least one other species were enriched in 72 of 94 terms in the rice GOSlim vocabulary. The mobile fraction of grass transcriptomes is largely uncharacterized (Schnable and Freeling 2011), and in rice, this is reflected by the fact that 64% of rice genes with syntenic orthologs in other species have at least one GOslim annotation, whereas only 24% of nonsyntenic rice genes do.

No Evidence of Segmental Deletions in Maize

To identify large posttetraploidy deletions from maize segments, the subgenomes of maize (i.e., maize1 and maize2) were aligned to the orthologous regions of the sorghum genome (supplementary fig. S6, Supplementary Material online). After discarded sorghum regions absent from both maize subgenomes—presumably representing clusters of gene insertions into sorghum or regions without sufficient conservation of synteny to identify orthology—16 regions of 40 or more sorghum genes were identified that were orthologous to a only one syntenic region in maize (supplementary table S5, Supplementary Material online).

To test whether these 16 regions were indeed single copy—as opposed to one syntenic region simply not being detectable using our approach—all annotated genes in maize were compared with sorghum genes within the candidate regions (supplementary fig. S9, Supplementary Material online). For regions that were, in fact, deleted from the maize genome, sorghum genes within the region should be the "best" match to fewer maize genes, whereas regions without detectable synteny, as a result of rearrangement or misassembly, should show no difference in this metric. The average sorghum gene was found to be the best BLAST hit of 1.996 maize genes from the B73_refgen2 working gene set. Using random permutations of sorghum genes, it was determined that in intervals of at least 40 genes, the average number of best BLAST hits from maize genes per sorghum gene was between 1.175 and 4.025 genes 95% of the time

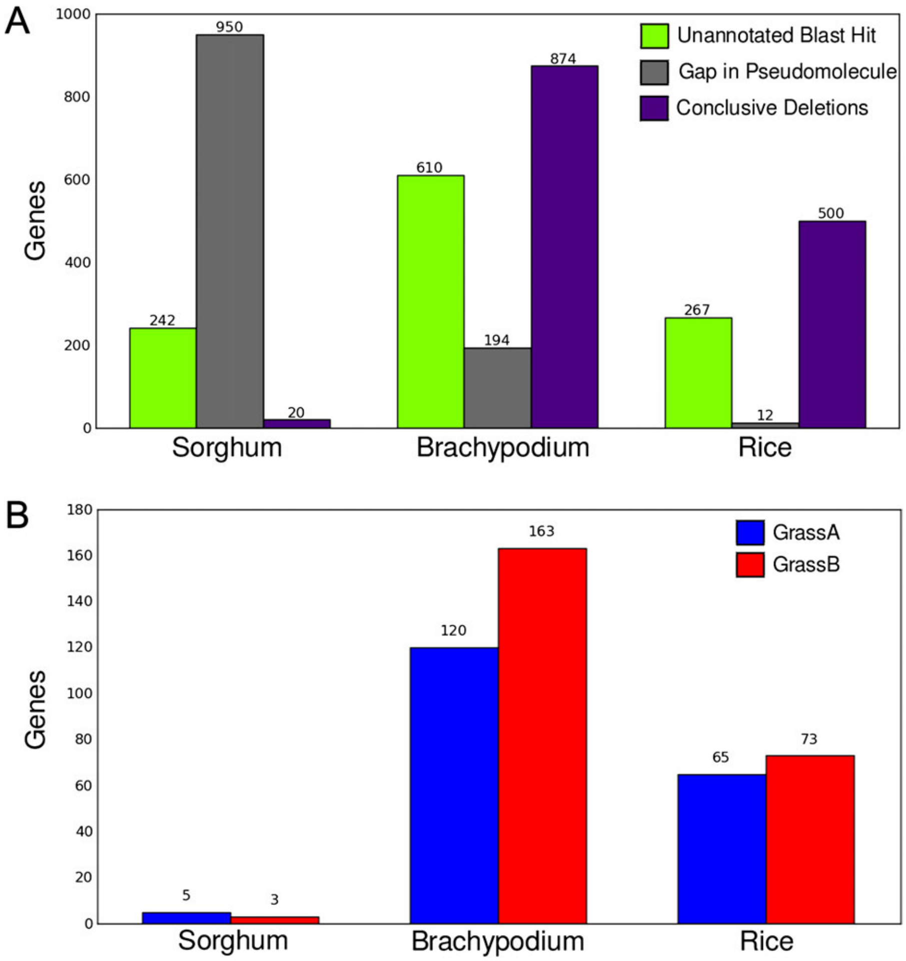

Fig. 4.—Frequency of gene loss events in multiple lineages and subgenomes. Rates of gene loss between species and subgenomes. (*A*) Genome-wide counts of the three types of gene loss described in figure 3 for the sorghum, brachypodium, and rice genomes. (*B*) Counts of only conclusive deletions located in regions assigned to the GrassA or GrassB subgenomes. Only gene deletions where the homeologous duplicate is still retained by the species were counted in this analysis.

(see Materials and Methods). Thirteen of the 16 sorghum regions with putative segmental maize deletions were within these bounds; the remaining three regions had an average number of maize best blast hits below the lower bound of this confidence interval (supplementary table S5, Supplementary Material online). These three regions were manually checked (supplementary fig. S7, Supplementary Material online). Two were found to have an additional syntenic region that was missed by computational approaches (supplementary table S5, Supplementary Material online), leaving only one potential segmental deletion spanning 56 genes in sorghum. Of these genes, 32 had syntenic matches in rice for which GOSlim annotations were available (supplementary table S6, Supplementary Material online). Although there was no significant enrichment in GOSlim annotations, we note that only the most extreme enrichments will be significant with such small data sets.

Relative to the other grasses, maize has experienced a much higher rate of gene loss. This is expected given that maize underwent a second, more recent, paleopolyploidy and is experiencing ongoing fractionation of duplicate gene pairs (Woodhouse et al. 2010; Schnable, Springer, et al. 2011). Given that current assemblies of the maize genome exhibits high levels of presence absence variation in gene content (Springer et al. 2009; Swanson-Wagner et al. 2010) and current versions of the maize genome omit at least 300 genes found in the reference inbred B73 (Lai et al. 2010), we omitted maize from our subsequent analyses of gene loss following the pregrass WGD.

Fractionation Bias between Homeologs and Subgenome Reconstruction

Given the recent reports that biased fractionation was a property of whole genomes in maize and *A. suecica*, it

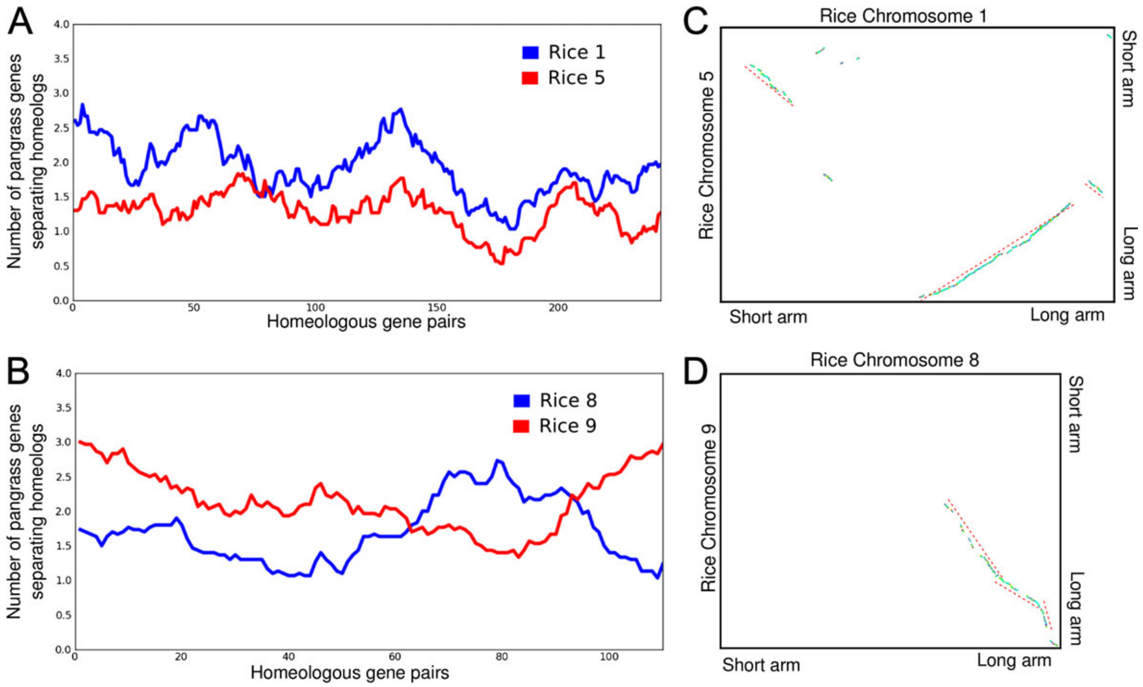

Fig. 5.—Rates of orthologous gene retention across large syntenic segments. Bias in gene content between homeologous rice chromosomes. (*A*) Running average of pan-grass genes between the homeologous regions of rice chromsomes 1 and 5. (*B*) Running average of pan-grass genes between the homeologous regions of rice chromsomes 8 and 9. *C* and *D* are dotplots showing syntenic regions identified between pairs of rice chromosomes. These dotplots are scaled using gene content rather than total nucleotides so the slope of syntenic diagonals represents a crude measure of fractionation bias. (*C*) Comparison of rice chromosomes 1 and 5. (*D*) Comparison of rice chromosomes 8 and 9.

might be possible to use fractionation bias as a marker to reconstruct ancestral genomes in ancient polyploid species such as the grasses. However, the lack of a suitable outgroup for the grasses creates new issues for quantifying fractionation bias. Between one and three quarters of the genes in arabidopsis have transposed to new locations since the divergence of the arabidopsis and papaya lineages ~70 Ma (Freeling et al. 2008). As the pregrass tetraploidy is estimated to also be approximately the same age (Paterson et al. 2004), any study of fractionation bias must first account for the mobile portion of grass genomes.

To compensate for recently inserted genes, we considered only genes orthologously conserved in sorghum and either rice or brachypodium to represent fractionated genes conserved in their ancestral locations. As sorghum and the rice–brachypodium lineage diverged ~50 Ma (The International Brachypodium Initiative 2010), this comparison allows us to filter out genes inserted during 70% of the length of time since the pregrass duplication. Excluding one duplicated region on rice chromosomes 11 and 12 that shows evidence of concerted evolution in multiple grass lineages (supplementary fig. S8*B*, Supplementary Material online) (Wang et al. 2011), 14 of 16 homeologous regions showed at least a 10% bias in the pan-grass retained genes without homeologs (supplementary table S2, Supplementary Material online). Biased retention of genes was consistent across all of rice chromosomes 1 and 5 (fig. 5*A*), which are homeologous across their entire length (Salse et al. 2009) (fig. 5*C*) and are representative of most homeologous regions within the rice genome. The second largest homeologous region (shared by rice chromosomes 2 and 4) displayed a similar pattern (supplementary fig. S8*A*, Supplementary Material online).

Bias in the number pan-grass genes with no homeolog between duplicate syntenic regions was used as a marker to assign duplicated regions to one of two subgenomes. Region which included more ancient syntenic genes without duplicates in the homeologous grass genomic region were assigned to the subgenome Grass A (underfractionated subgenome), whereas the homeologous region with fewer ancient syntenic genes remaining after homeologous duplicates were excluded was assigned the subgenome Grass B (overfractionated subgenome) (fig. 1 and supplementary fig. S3, Supplementary Material online).

Identification of an Ancient Homeologous Recombination Event

The rice homeologous regions were scanned for locations where the direction of biased gene retention switched between homeologs in order to identify ancient recombination events. One such switch was identified between rice

Table 1
Five High Confidence Reciprocal Gene Losses

Sorghum gene(s)	Rice gene(s)	Brachy gene(s)	Annotation	Link
Sb04g033870 Sb10g007450	Os06g11410	Bradi3g58300	Cyclin	http://genomevolution.org/r/2qwp
Sb02g011380 Sb07g024380	Os08g44300	Bradi4g38330/40	Calcineurin-like phosphoesterase	http://genomevolution.org/r/2qwq
Sb06g020480/90 Sb04g024990	Os02g38350	Bradi5g13690	Regulator of chromosome condensation domain containing	http://genomevolution.org/r/2qwr
Sb06g017750 Sb04g023130	Os04g36670	Unannotated sequence	OsArgos: Arabidopsis ortholog regulates organ size	http://genomevolution.org/r/2qws
Sb04g003550	Os06g48350	Bradi3g03850	Translation Initiation Factor 5	http://genomevolution.org/r/2qwt

chromosomes 8 and 9 (fig. 5*B* and *D*). Both the proximal and distal ends of the homeologous region contain more pan-grass syntenically retained genes on chromosome 9, however, in the central portion of the homeologous region, more pan-grass syntenically retained genes are found on chromosome 8. The changes in content are only visible when comparing homeologous regions and not when comparing orthologous regions between species (supplementary fig. S5, Supplementary Material online). This indicates the change, likely an ancient homeologous recombination event, occurred prior to divergence of the sorghum and rice lineages. Interestingly, one of the two boundaries between the central and the flanking portions of the region subsequently served as an inversion breakpoint in sorghum (supplementary fig. S5*B*, Supplementary Material online).

Ongoing Gene Loss from Homeologous Gene Pairs

Some homeologous duplicate genes are retained in only some of the grass species examined (fig. 4*B*). As with the total number of high confidence gene losses, the brachypodium genome includes the greatest number of these lost homeologous duplicates. Genes located on Grass B (underfractionated regions) are significantly more likely to be lost from the genome of brachypodium than duplicate copies of the same set of genes located on Grass A (overfractionated regions) ($P = 0.0062$, binomial test). The small bias in the same direction observed for homeologous genes lost from the rice genome is not statistically significant ($P = 0.2757$, binomial test). Only eight high confidence losses of homeologous genes were observed in sorghum. This likely is a result of the number of gaps in the sorghum pseudomolecules (fig. 4*A*) and not due to a lower overall rate of gene loss in this lineage.

Reciprocal Homeologous Gene Loss

By including interspecies comparisons of the grasses, it was possible to identify reciprocally lost homeologous genes between rice, sorghum, and brachypodium. For this analysis, gene sets were excluded if they contained missing genes that fall into class #2 predicted locations which include gaps in the pseudomolecules or contained genes not located in the Grass A (underfractionated) or Grass B (overfractionated) subgenomes.

The remaining data set contained 1,345 genes groups represented by retained duplicate genes from the pregrass WGD. In 1,111 cases—82.6% of the total—both duplicate gene copies were retained in all three of sorghum, rice, and brachypodium. In another 222 cases—16.5%—one gene copy was retained in all three species, whereas the other copy had been lost from the genomes of either one or two species. Genes copies located on the Grass B subgenome were marginally more likely to be the copy lost in one or more lineages—121 cases—however, this bias was not statistically significant. These lost genes were not found to be significantly enriched in any annotation using rice GO-Slim terms.

In the remaining 12 cases, each copy of the gene was deleted in at least one lineage. However, in seven of these cases, both copies of the gene were lost from the same species, suggesting these genes function in some nonessential role, making them unlikely candidates to drive hybrid incompatibility (supplementary table S2, Supplementary Material online). The final five cases (0.4% of all retained duplicated genes; 0.12% of single copy ancestral genes located within these duplicate regions) represent the only credible candidates for reproductive barriers resulting from reciprocal gene loss following WGD in the grasses and are summarized in table 1.

Discussion

Ancient Subgenomes and Hidden Evolutionary Events

Bias in gene loss between homeologous regions has been studied and confirmed for a wide range of species (Thomas et al. 2006; Sankoff et al. 2010; Woodhouse et al. 2010). However, it only recently has been demonstrated that this bias is likely a property of the whole parental genomes of a tetraploid rather than of individual duplicated segments (Chang et al. 2010; Schnable, Springer, et al. 2011). As such, biased gene loss represents a powerful mark for reconstructing paleogenomes in ancient tetraploid species, even, or especially, in the absence of useful outgroups. In this study, we assigned nearly all duplicated regions in grass genomes derived from an ancient tetraploidy into low gene loss and high gene loss subgenomes, Grass A and Grass B, respectively. In rice, over- and underfractionated regions are often colocalized on the same chromosomes (supplementary fig. S3, Supplementary Material online), meaning modern chromosomes are a chimera of subgenomes. Because reconstructions of paleochromosomes usually assume homeologous regions located on the same modern chromosome derive from the same ancestral chromosome, published reconstructions of grass ancestral protochromosomes (Salse et al. 2009) should be reexamined.

We identified a case in rice (chromosomes 8 and 9) and sorghum (chromosomes 2 and 7) where over- and underfractionated regions are colocalized on the same chromosomes (supplementary fig. S4, Supplementary Material online). Interestingly, this unique event is only apparent when comparing homeologous syntenic regions within a species and not orthologous syntenic regions between species. Such a pattern may occur by one of two processes: 1) fractionation bias is not constant along a chromosome or 2) homeologous regions were exchanged between chromosomes through homeologous recombination. It has been previously reported that biased gene loss is consistent across entire ancestral chromosomes in maize and entire parental genomes in *A. suecica* (Chang et al. 2010; Schnable, Springer, et al. 2011) providing evidence that fractionation bias does not change across a chromosome. Additionally, one end of this apparently exchanged region later served as an inversion breakpoint on sorghum chromosome 7, which is consistent with current models regarding the reuse of chromosome breakpoints (Larkin et al. 2009).

Biased fractionation is likely a result of genome dominance (Schnable, Springer, et al. 2011), a phenomena observed in numerous allotetraploid species where genes from one parental genome tend to show higher expression in wide hybrids or allopolyploids than homeologous genes from originating from the other parental species (Buggs et al. 2010; Chang et al. 2010; Flagel and Wendel 2010). Given that genome dominance appears to be linked to qualitative differences between parental genomes rather than mode of inheritance (paternal vs. maternal) (Flagel and Wendel 2010), the bias we observed may be evidence that the pregrass duplication resulted from allopolyploidy.

Incomplete Coverage of the Pregrass Tetraploidy

Only 65.7% of the rice genome has an identified homeologous region from the pregrass tetraploidy (Yu et al. 2005). Deletion of large genomic regions has been observed in newly synthesized polyploids (Gaeta et al. 2007) so it might be argued that our analyses, which exclude all genes without identified homeologous regions, exclude a major category of fractionating gene loss. However, in an analysis of the several million year old maize tetraploidy, almost no evidence was found for large segmental deletions from either subgenome. The largest gaps in the syntenic coverage of the sorghum genome by maize (supplementary fig. S6, Supplementary Material online) were shared by both maize subgenomes and particularly centered around centromeres. This finding is consistent with a previous report that there was no evidence of large deletions (greater than or equal to 4 sequential genes) during fractionation in maize (Woodhouse et al. 2010). Therefore, the incomplete coverage of the sorghum, rice, and brachypodium genomes by duplicated segments from the pregrass WGD likely results from duplications where the syntenic signal has sunk below the limits of detectability as the result of ongoing fractionation, gene insertion, chromosomal rearrangements, and genome assembly errors.

An unduplicated outgroup sequence will aid in the identification of these highly fractionated and rearranged regions for all grasses. Although large deletions are common in the early generations of a newly tetraploid species, large scale deletions will almost always include one or more dose-sensitive genes and are expected to be selected against in subsequent generations, allowing paleopolyploids to retain near complete subgenomes at the level of whole regions, even as individual genes are lost by fractionation (Xiong et al. 2011).

Ancient and Ongoing Gene Loss

To enable the study of biased gene loss following WGD in the grasses, it was necessary to develop accurate methods of identifying genes which truly have been deleted from their ancestral location. We found that the rate of gene loss in the rice and brachypodium lineages has been significantly different. The rate of syntenic gene loss in the brachypodium lineage has been 75–115% higher than in rice since the divergence of those two lineages. The direction of this difference, although not the absolute rates of gene loss, is consistent with a study of genomic regions in sequenced grass orthologous to nine sequenced contigs from *Aegilops tauschii* (Massa et al. 2011). If the increased rate of gene loss observed in brachypodium is explained by the same evolutionary pressures for a small genome size that resulted in the brachypodium genome being only half the size of the rice genome, the fact that genes located on Grass B were significantly more likely to be lost in brachypodium than their homeologous duplicates on Grass A suggests that even after tens of millions of years, Grass B genes remain the more expendable member of a gene pair. The increased levels of unannotated syntenic blast hits in brachypodium may represent gene fragments generated by the ongoing deletion of genes via the same short deletion mechanism shown to remove genes in maize (Woodhouse et al. 2010). An alternative explanation is that these syntenic blast hits represent real genes missed during the annotation of the brachypodium genome. However, even if these genes are not counted as losses, Grass subgenome B has lost more genes in brachypodium (fig. 4*B*).

The rate of ongoing fractionation in the grasses may be higher than we are measuring. Grass B gene copies are more prone to fractionation overall. A study in yeast reported that in the later stages of fractionation, the same copy of individual gene pair tends to be lost independently in multiple lineages (Scannell et al. 2006). Both of these pieces of data would tend to suggest that a significant number of duplicate pairs may have independently been lost in multiple lineages following the major grass lineage radiation. Although independent deletions of the same gene copy would not create reproductive barriers, it is important to consider their existence when measuring the rate of fractionation in the grasses.

Based on data from teleosts and yeast, the Wolfe laboratory has presented the hypothesis that genome duplications may sometimes drive speciation by increasing the speed at which reproductive barriers form. Even a small number of reciprocally lost loci between separate populations could result in hybrid offspring being unlikely to possess a full complement of essential genes (Lynch and Force 2000). The grasses, a diverse and highly successful clade whose origin is associated with genome duplication seemed an likely candidate for reciprocal gene loss–driven speciation. However, the frequency of reciprocally lost genes we observed was strikingly lower than that found in studies of WGD in other lineages. In polyploid yeast, 4–7% of ancestral loci examined were identified as homeologs, which had been reciprocally lost between different species (Scannell et al. 2006). A study in the ray-finned fishes (teleosts) reported that 8% of single-copy genes between zebrafish and Tetraodon where in fact reciprocally lost homeologs (Sémon and Wolfe 2007). Our own identification of only five putative reciprocally lost homeologs in the grasses out of thousands of gene pairs and single copy syntenic genes examined is strikingly different. One possible explanation for the difference we observe is that the teleosts and and yeast WGDs represent autopolyploidies, and, in the absence of genome dominance differentiated between two parental subgenomes, the early fractionation of gene pairs was more stochastic in these lineages, resulting in greater numbers of RGL events. This agrees with the observation in yeast that early gene losses were equally likely to remove either copy of a duplicate gene pair (Scannell et al. 2006). In plants, although the majority of polyploidy events are predicted to be autopolyploidies (Ramsey and Schemske 1998), the majority of named polyploid species arise through allopolyploidy (Mallet 2007). The impact of various forms on polyploidy on speciation and evolutionary success has been well reviewed (Rieseberg and Willis 2007; Soltis and Soltis 2009).

It may be tempting dismiss these findings as a result of the young age of the pregrass tetraploidy relative to the yeast and teostate duplications. However, the hypothesis that reciprocal loss of duplicated genes enables increased rates of speciation requires that these gene deletions occur contemporaneously with speciation, and this was, in fact, found to be the case in yeast (Scannell et al. 2006). A small number of grass species diverged prior to the split of the most recent common ancestor of the maize–sorghum and rice–brachypodium lineages (Grass Phylogeny Working Group 2001), and these lineages may hold more examples of reciprocal gene losses. However, the vast majority of grass species diverged contemporaneous with or following the maize–sorghum rice–brachypodium split (Grass Phylogeny Working Group 2001). Given the lack of evidence for significant levels of reciprocal gene loss from this point onward, we conclude that reciprocal gene loss of duplicate genes resulting from WGD was probably not responsible for the radiation of the primary grass lineages. This contrasts with individual reports that the reciprocal loss of duplicates genes resulting from individual dispersed duplications create hybrid incompatibility in arabidopsis and rice (Bikard et al. 2009; Mizuta et al. 2010).

Concluding Remarks

Having multiple whole-genome sequences for several clades of organisms provides a rich data set for studying the evolution of genomes. Angiosperm genomes, in general, are remarkable for having repeated WGD events that permeate their lineages. In particular, the grass lineage combines these two facets: several grass genomes are currently available with several more arriving soon, and a WGD event occurred prior to their radiation. We show that by classifying the evolutionary history of sets of genes and identifying the subgenomes comprising modern grass genomes provides an opportunity to understand the evolution of individual genomes and the grass lineage as a whole. Importantly, the ongoing process of fractionation remains biased in the grasses preferentially and consistently targeting one subgenome for gene loss, and that unlike previously studies in yeast and teleosts, reciprocal gene loss of duplicated genes is not likely to be the driving force of the grass radiation.

Supplementary Material

Supplementary figures S1–S10, tables S1–S10, and data set are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the iPlant Collaboration (DBI: 0735191) to E.L., a National Science Foundation grant (DBI 0701871) to M.F., and a Chang-Lin Tien Graduate Fellowship to J.C.S.

Literature Cited

Alexandrov NN, et al. 2009. Insights into corn genes derived from large-scale cDNA sequencing. Plant Mol Biol. 69:179–194.

Barker MS, et al. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol. 25:2445–2455.

Baucom RS, et al. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet. 5:e1000732.

Bikard D, et al. 2009. Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science 323: 623–626.

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

Buggs RJ, et al. 2010. Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol Ecol. 19:132–146.

Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11:R125.

De Bodt S, Maere S, van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 20:591–597.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Flagel LE, Wendel JF. 2010. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol. 186:184–193.

Freeling M, et al. 2008. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Res. 18: 1924–1937.

Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC. 2007. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis. Plant Cell 19:1441–1457.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805–814.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell Online 19:3403–3417.

Goff SA, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296:92–100.

Grass Phylogeny Working Group. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Ann Mo Bot Gard. 88: 373–457.

Harris RS. 2007. Improved pairwise alignment of genomic data. [cited 2012 Jan 31]. Available from: http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf.

The International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463:763–768.

Lai J, et al. 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet. 42:1027–1030.

Langham RJ, et al. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166:935–945.

Larkin DM, et al. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res. 19:770–777.

Lynch M, Force AG. 2000. The origin of interspecific genomic incompatibility via gene duplication. Am Nat. 156:590–605.

Lyons E, Pedersen B, Kane J, Freeling M. 2008. The Value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. Trop Plant Biol. 1:181–190.

Mallet J. 2007. Hybrid speciation. Nature 446:279–283.

Massa AN, et al. 2011. Gene space dynamics during the evolution of *Aegilops tauschii, Brachypodium distachyon, Oryza sativa, and Sorghum bicolor* genomes. Mol Biol Evol. 28(9):2537–2547.

Mizuta Y, Harushima Y, Kurata N. 2010. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proc Natl Acad Sci U S A. 107:20417–20422.

Osborn TC, et al. 2003. Understanding mechanisms of novel gene expression in polyploids. Trends Genet. 19:141–147.

Ouyang S, et al. 2007. The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res. 35:D883–D887.

Paterson AH, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A. 101: 9903–9908.

Prasad V, Strömberg CAE, Alimohammadian H, Sahni A. 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science 310:1177–1180.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu Rev Ecol Syst. 29:467–501.

Rieseberg LH, Willis JH. 2007. Plant speciation. Science 317:910–914.

Salse J, et al. 2009. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc Natl Acad Sci U S A. 106:14908–14913.

Sankoff D, Zheng C, Zhu Q. 2010. The collapse of gene complement following whole genome duplication. BMC Genomics 11:313.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341–345.

Schnable JC, Freeling M. 2011. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PLoS One. 6:e17855.

Schnable JC, Pedersen BS, Subramaniam S, Freeling M. 2011. Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses. Front Plant Sci. 2:2.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 108:4069–4074.

Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.

Sémon M, Wolfe KH. 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. Trends Genet. 23:108–112.

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

Soderlund C, Bomhoff M, Nelson WM. 2011. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. 39:e68.

Soltis DE, et al. 2009. Polyploidy and angiosperm diversification. Am J Bot. 96:336–348.

Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol. 60:561–588.

Springer NM, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5:e1000734.

Swanson-Wagner RA, et al. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20:1689–1699.

Swigoňová Z, et al. 2004. Close split of sorghum and maize genome progenitors. Genome Res. 14:1916–1923.

Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A. 107:472–477.

Tang H, et al. 2011. Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12:102.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.

Tuskan GA, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604.

van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10:725–732.

Wang X, Tang H, Paterson AH. 2011. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. Plant Cell 23(1):27–37.

Woodhouse MR, et al. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol. 8:e1000409.

Xiong Z, Gaeta RT, Pires JC. 2011. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc Natl Acad Sci U S A. 108(19):7908–7913.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yu J, et al. 2005. The genomes of Oryza sativa: a history of duplications. PLoS Biol. 3:e38.

Associate editor: Yves van de Peer