

Grass Genomic Synteny Illuminates Plant Genome Function and Evolution

Jeffrey L. Bennetzen · Mingsheng Chen

Received: 1 October 2008 / Accepted: 21 October 2008 / Published online: 12 November 2008
© Springer Science + Business Media, LLC 2008

Abstract The genic colinearity of grass genetic maps, physical maps, and contiguous genomic sequences has been a major contributor to almost two decades of study into the structure and evolution of grass genomes. This research has led to the discovery of all of the major phenomena responsible for the rapid evolution of flowering plant genomes. These processes include polyploidy, transposable element (TE) amplification, TE-driven genome rearrangement, and DNA removal by unequal homologous recombination and illegitimate recombination. The great variety in angiosperm genome structure is largely an outcome of differences in the specificities, frequencies, and amplitudes of these common genome-altering processes. Future emphasis now needs to shift to harnessing an even broader range of studied species, and to use this phylogenomic perspective to uncover the nature and functions of the genes that are shared by particular lineages and those that set each individual species apart as a unique biological entity.

Keywords Comparative genomics · Genome evolution · Microcolinearity · Recombination · Transposable elements

Introduction

Two of the central questions in biology are what are the genetic foundations that underlie the similarities between different species or individuals within a species and what are the genetic variations responsible for the observed differences. Even from the first days of comparative genomics, it was surprising to many that humans, fruit flies, nematodes, and yeast shared a large percentage of their genes [2, 85]. Given our human-centered worldview, it was to be expected that many would be shocked by the fact that humans and our closest surviving relative, the chimpanzee, share ~98% sequence identity and an even higher similarity in gene content [81]. In plants, haplotypic differences in genome sequence within a species like maize can greatly outstrip these interspecies primate variations. Of course, not all sequence change is equally significant, and work on the evolution of maize has shown that tiny changes in regulatory loci can dramatically alter morphology and behavior [24].

The field of comparative genomics was founded on the idea that comprehensive analyses and comparison of whole genomes could uncover the essential conserved, and the importantly variable, components of any set of genomes. In plants, this comparative analysis proved to be particularly challenging for several reasons, including (1) the small number of species that were investigated, (2) their large and complex genomes, and (3) their high rate of structural rearrangement. The observation that closely related plants sometimes exhibited regions of DNA-marker colinearity [11, 38] provided a key point of constancy in these comparisons because genetic map relatedness was simple to determine with robust techniques that were not dramatically affected by genome size or the overall quality of the genetic toolkit for that species [8]. The first demonstrations of microcolinearity (also called microsytenty) by comparative

J. L. Bennetzen (✉)
Department of Genetics, University of Georgia,
Athens, GA 30602-7223, USA
e-mail: maize@uga.edu

M. Chen
State Key Laboratory of Plant Genomics, Institute of Genetics and
Developmental Biology, Chinese Academy of Sciences,
Beijing 100101, China

sequencing of orthologous chromosomal regions [17, 84] indicated that overall genomic similarity could be converted into very local analyses of the evolved structure and function of genes that were all derived from a known ancestral locus at attributable dates. Hence, both whole genome and individual gene analyses could be made in a comprehensive manner across many species, as first proposed and illustrated in the grasses.

Another enduring question that was illuminated by a comparative genome analysis strategy was the nature of the DNA in the eukaryotic nucleus. For many decades, it has been known that nuclear genomes vary dramatically in size, even between closely related species, and the mystery behind this “unexplained” or presumed “excess” DNA was termed the “C-value paradox” by Thomas [82]. Research in flowering plants, where the differences in nuclear DNA content varies more than one thousand fold, has explained this C-value variation [10]. We now know that the differences in C-value across flowering plants are very dynamic outcomes of occasional polyploidy along with great variability in transposable element (TE) amplification and processes for DNA removal. However, we do not know how often these changes (especially those caused by TEs or other small indels) generate selectable variation that can lead to changed capabilities within a species or to speciation.

This plant genome review will discuss the discovery of genomic colinearity and synteny, its biological origins, its numerous exceptions, and its uses for genome analysis. We will focus on the grasses because this is our area of greatest expertise and because this is also the source of the most comprehensive sets of data and analyses in plants. We have every reason to believe that much of what is discussed herein for the grasses will also be true in other plants, and in more distantly related eukaryotes.

Genetic map colinearity in the grasses: rules and exceptions

The crop circle

The first comparative genetic map in the grasses was a miniscule maize::sorghum comparison in a study meant to test whether restriction fragment length polymorphism (RFLP) markers generated in one species (e.g., maize) could be used to help generate a genetic map in other species (e.g., sorghum) [38]. This project indicated that maize RFLP probes could be used routinely for species as far distant as foxtail millet, a lineage that last shared a common ancestor with the maize lineage about 30 million years ago (mya) [48]. Serious grass genome comparisons were then generated by expert mapping labs, especially the

Gale and Tanksley groups [3, 4, 20]. These studies indicated a good deal of similarity in gene content and colinearity, with a low frequency of small and large exceptions. Moore et al. [67] provided a major conceptual leap when they identified a series of conserved grass genome segments and then assembled them into a comparative circle map.

The comparative circular map of the grasses, also known as the crop circle, has allowed identification of the major rearrangements that differentiate grass genomes, and has provided insight into the timing of these events during grass descent from a common ancestor more than 50 mya. As shown in Fig. 1, gene order and telomere location are largely conserved at this scale, although the number of chromosomes is quite variable across species. Maize yields two concentric circles, suggestive of a whole genome polyploidy event, which has been confirmed by extensive orthologous DNA sequence analysis [80]. Two specific translocations, shown at 3 o'clock and 7 o'clock on the circle map, are shared by all investigated members of the Panicoideae subfamily but not by rice or the Triticeae. Most of the other detected rearrangements are inversions that are limited to only one or two of the species depicted (Fig. 1).

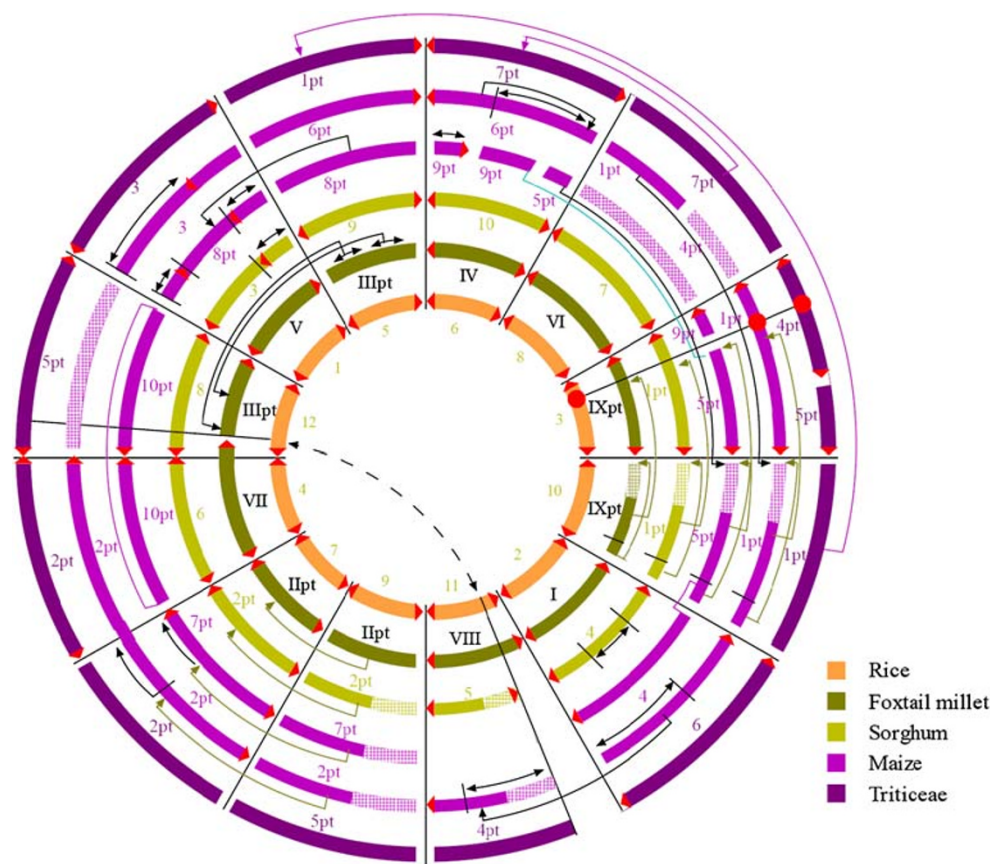
In addition, this circular map allowed the easily visualized (and thus conceptualized and transmitted) discovery that some important genes involved in domestication or other important traits appeared to be the same orthologous loci across multiple grass species (Fig. 1) [72]. This, in turn, helped encourage the use of surrogate plant chromosomes like the relatively small genome of rice to assist in the map-based cloning of genes in large-genome species like barley, wheat, or sugarcane [14, 33, 19].

Three major conclusions that were clear from the comparative circular maps were (1) the relatively low frequency of large genomic rearrangements, (2) the presence of inversions, translocations and duplications, and (3) the uneven distribution of such events, with many at boundaries near current centromere locations (Fig. 1). Among the many issues not resolved by this analysis, however, was whether the frequencies of these major rearrangements were in any way predictive or mechanistically similar to the frequencies and types of local rearrangements. More detailed physical and genetic maps would be needed to address these questions.

Physical maps, genetic maps and their comparison

Early studies by comparative genetic mapping revealed the extent of conservation of gene content and gross gene order among different grass species, but did not give many insights into the likelihood or nature of small rearrangements. In these first studies [38], it was observed that most maize RFLP probes hybridized strongly to sorghum DNA,

Fig. 1 Synteny of five crop genomes. Different color bars represent the chromosomes in different grass genomes, with their telomeres indicated by red triangles. Arrows show rearrangements relative to rice. Arrows with a single arrowhead are translocations, and those with two arrowheads are inversions. Arrows at 3 o'clock and 7 o'clock indicate rearrangements that are shared by the subfamily Panicoideae (foxtail millet, sorghum, and maize). Dotted bars indicate regions where insufficient data were available at the time of the analysis undertaken by Gale and Devos [30]. The dotted internal line indicates a duplication shared by chromosomes 11 and 12 of rice [69]. Red dots are orthologous genes controlling semi-dwarf phenotypes that are located on rice chromosome 3, wheat chromosome 4 and maize chromosome 1 [22, 73]. *pt* Part of a chromosome.



but the repetitive DNA sequences in maize usually did not hybridize to sorghum. This suggested that repetitive DNA sequences evolved much faster than genes, and that heterologous probes could thus provide some advantages over homologous probes from a repeat-rich genome.

Comparative genetic mapping between closely related grasses, such as sorghum and sugarcane, whose separate lineages diverged from each other about 8 mya, show striking map colinearity [35]. In contrast, detailed comparative genetic mapping among more distantly related species, such as maize and rice, identified numerous chromosomal rearrangements, such as telomeric fusions, nested insertions, inversions and translocations [92], although about 2/3 of these genomes appeared to still be colinear. Many of the detected rearrangements were confirmed by comparative physical mapping, such as (from a rice perspective) the fusion of rice chromosomes 3 and 10 and chromosomes 7 and 9 into single chromosomes in the Panicoideae lineage [88]. In addition, comparative physical mapping also uncovered the ancient grass genome duplication shared by maize [88], wheat [77], and other grasses [71].

Comparative physical mapping between sorghum and rice revealed different genome components with very different degrees of microcolinearity. In euchromatic regions, where most meiotic recombination occurs, greater microcolinearity was observed; however, less microcolinearity was observed in

recombination-poor heterochromatin, such as pericentromeric regions [12]. This phenomenon was also apparent in comparison of homoeologous chromosomal regions in rice derived from the ancient duplication at the origin of the grasses, where little colinearity was retained in pericentromeric regions [83]. In addition, the heterochromatic regions of sorghum have been preferentially expanded relative to rice, as compared to euchromatic regions [51]. Future detailed studies of microcolinearity in heterochromatin are needed to uncover the dynamics and mechanisms for macro- and micro-rearrangements in these crossover-deficient parts of grass genomes [61, 63].

Microcolinearity

Across the grasses (and a bit beyond)

Even from the start, comparisons of genomic sequence in orthologous regions of different grass species examined a very large time frame, such as rice versus sorghum [17] or rice versus various Triticeae [28, 37, 26], all comparisons where the investigated species last shared a common ancestor ~50 mya. In this time frame, the sequences between genes appeared to be completely different, although very tiny “conserved non-coding sequences”

(CNS) were later discovered [45, 36]. Even introns of orthologous genes, although largely consistent in location across all flowering plants, contained obvious conserved sequences only at the boundaries needed to specify appropriate RNA processing. Hence, the general conclusion could be reached that anything still conserved after 50 million years of grass genome divergence was likely to have an important function.

Gene content and order, on the other hand, were mostly conserved on segments of a few dozen to a few hundred kb even after 50 million years of independent grass genome evolution. Comparisons to rice have been particularly useful in this regard because (1) it is evolutionarily quite distant from the other important grasses like maize, wheat, barley, and sorghum [48], (2) it has a relatively small genome (~400 Mb) with a high gene density, (3) its genome was an early target for comprehensive sequence analysis [41], and (4) it has proven to be more stable vis-à-vis small local rearrangements than other grasses like maize, sorghum, wheat or barley [9].

In the most comprehensive comparisons to date, between rice and two panicoid grasses, sorghum and maize, the frequency of gene movement over the last fifty million years was calculated as at least 5%, and possibly as high as 25%, between sorghum and rice [53]. This number does not include the gain or loss of tandemly repeated gene copies, a very common phenomenon in all grass lineages investigated. Most of the genic rearrangements in maize compared to either rice or sorghum are apparent gene losses on one of two maize homoeologues [40, 53], an expected outcome of the polyploidization event about five mya that gave rise to the *Zea* lineage [80]. However, too little data yet exist to identify possible subtle patterns in types of rearrangement. Moreover, rearrangements involving genes are likely to be under selective pressure, so the events currently observed in any species are a combined outcome of those events that have occurred, minus those that were subsequently removed by chance or by selection against some specific changes.

In more distant comparisons, with longer ancestral divergence times, colinearity across orthologous regions appears to be much more rare than within the grasses. In the rice flatsedge, *Cyperus iria*, the near-adjacent *Sh2* and *A1* homologues appear to be conserved in order and orientation, but one of the two genes in between in the grasses is missing in the sedge (A. Pontaroli and J. Bennetzen, unpub. obs.). However, this is the only comparison that has been done to the grasses in this ~110-million-years-of-divergence window [13]. Similarly, *Musa* (e.g., banana) genomes show some colinearity with the grasses after >115 million years of divergence from their last shared ancestor, but more than 50% of the annotated genes were non-colinear in a comparison to rice [56]. With even more distant comparison

to the eudicots, >220 million years of independent descent, only rare segments of genic colinearity are observed at either full genome or local genome scales [58].

The most frequent type of structural change in all investigated angiosperm nuclear genomes has been observed to be the differential insertion and subsequent instability of transposable elements (TEs). In large-genome species like maize and barley, most of the DNA between genes is comprised of TEs, especially long terminal repeat (LTR) retrotransposons [78, 86, 89, 75]. These elements transpose by reverse transcription of an RNA transcript and insertion of the resultant DNA, so transposition does not involve excision. Because LTR retrotransposons make up more than 50% of most or all large flowering plant genomes and their high content varies somewhat proportionally with angiosperm genome size, it is clear that these TEs are the most important factor responsible for genome size variation in flowering plants [10]. Because these TEs (and all other unselected DNAs) are fragmented and removed so rapidly by accumulated small deletions (see below), all of the insertions appear to be very recent, usually within the last 2–6 million years [87]. This accounts for the near-complete lack of homology of the intergenic regions in orthologous genome segments with grass lineages that last shared a common ancestor more than 50 mya.

We currently lack a vocabulary to precisely describe the degree of conservation of genic content and colinearity between any two species, much less across multiple species, although a gene-pair conservation terminology is currently in development (L. Feng and J. Bennetzen, unpub. res.). However, it is clear that some lineages are very unstable (e.g., pearl millet, sorghum, maize) and others are much more stable (e.g., rice and foxtail millet) at the level of compared genetic maps and/or microcolinearity [23, 75, 9, 40]. We do not yet know the reasons for these differences, nor whether high conservation at one scale (e.g., genetic map) in any way correlates with high conservation at other scales (e.g., physical map or microcolinearity). It is clear, though, that certain types of gene rearrangement are rare (e.g., movement of a gene to a wholly different chromosome) while others are relatively common (tandem duplication, deletion or inversion of small genic segments).

Analysis of microcolinearity and gene content conservation at long time frames has the advantage of the accumulation of multiple events for analysis, but this is more than counterbalanced by three negative aspects of concentrating on such ancient rearrangements. First, natural selection has had a great deal of time to remove any events that had even a minor organismal disadvantage, so one only observes certain classes of tolerated or advantageous events that might not be proportional to the true spectrum of *de*

novo rearrangements. Second, the components of the genome responsible for the rearrangement have had ample time to decay into a state where they are invisible to current annotation approaches. And, third, individual events may be buried underneath second, third or more layers of events at the same location. For all of these reasons, investigations of orthologous regions in closely related lineages are justified, and are expected to be “there to discover” because of the relatively high rate of local chromosomal rearrangement in the grasses.

Colinearity dynamics within a 0–15 million year window of grass genome evolution

Orthologous sequence comparisons across short time frames has the potential to reveal both the rate and the mechanisms for disruption of colinearity. In a sequence comparison of the *adh1*-orthologous regions of maize and sorghum, two species that last shared a common ancestor about 12 mya [80], a 212-kb maize sequence was found to be largely collinear with a 66-kb sorghum sequence [84]. The more than three-fold size difference is mainly due to nested LTR retrotransposon insertions in the maize genome [78, 84]. In the original annotation, orthologs of nine maize genes were detected in the sorghum region in perfect colinear order; however, three additional genes in this sorghum segment were not found in the maize *adh1* region. In subsequent analyses, one of the “missing” maize genes was found to be located in the *adh1*-homoeologous region of maize [40]. This has now turned out to be a routine situation in the maize genome, where two maize segments represent each sorghum region due to a polyploidy event in the *Zea* lineage within the last few million years [80]. Gene deletion (usually of only one homoeologous copy) subsequent to polyploidization has now reduced the originally doubled copy number of genes ($2\times$) to less than $1.5\times$ [53]. The other two non-colinear genes in the *adh1*-orthologous regions of sorghum are found elsewhere in the genomes of maize and other grasses and are hypothesized to have been caused by the insertion of two unlinked genes, either as two subsequent events or by a single event involving three chromatids. In dramatic contrast, a comparison of the *adh1*-orthologous regions between sorghum and sugarcane, both gene colinearity and strong homology of non-coding regions were observed [42], indicating greater stability in these lineages over this shorter (~8 million year) time frame of divergence.

In at least some genomes, polyploidization is followed by extensive genomic change resulting in the silencing and elimination of duplicated genes [1]. In grasses, polyploidy has been a recurrent theme, with many lineages exhibiting full genome duplications over the last few million years. Local sequence comparisons in these species, such as maize

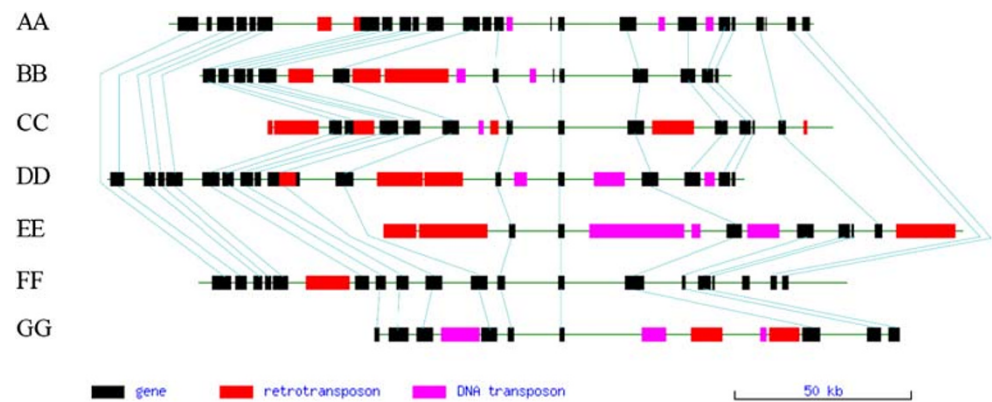
[40, 53, 64], wheat [27, 46, 90, 34, 16, 15] and sugarcane [42], have revealed interesting features of gene and genome evolution in recent polyploids. LTR retrotransposon amplification and altered regulation (e.g., silencing) or loss of duplicated genes are repeated themes. Inactivation and eventual elimination of duplicated genes can be mediated by altered epigenetic regulation, deletions, TE insertions, and/or point mutations causing premature stop codons.

Some evidence suggests that specific alterations recur in independent polyploidizations in wheat [27, 46] and *Brassica napus* [59]. However, most eventually fixed changes do not occur instantly in post-polyploid genome rearrangements, at least not in maize. In *adh1*-homoeologous regions, for instance, fragments of partially deleted genes remain, indicating the incomplete status of removal several million years after polyploidy, and showing that these gene losses are primarily by the accumulation of multiple small deletions [40]. Another example of reasonably stable polyploid gene copies comes from a comparative study of the *adh1*-orthologous regions of maize, sorghum and sugarcane [42]. The two sugarcane homoeologous haplotypes show perfect genic colinearity. In addition, two maize homoeologous regions yielded the same gene content, order and orientation as in sugarcane. Our data on comparative analysis in the *Oryza* genus also reveals excellent stability of polyploid genomes formed less than two million years ago (Chen et al., unpub. res.).

The *Oryza* genus contains about 24 species that belongs to ten different genome types [31]. A project, entitled the *Oryza* MAP Alignment Project (OMAP), was launched to build a framework for comparative biology in the *Oryza* genus [93]. Representative species, ranging from closely related species/subspecies, such as those with AA genomes, which diverged from their common ancestor less than a million years ago, to more distantly related species, such as *O. brachyantha* and *O. granulata*, whose ancestors diverged about 10 mya, were chosen for bacterial artificial chromosome (BAC) library construction, BAC end sequencing, and physical map construction [5, 49]. The initial analyses revealed excellent gene colinearity both in their physical maps [50] and in sequence comparisons [96]. Genome size variations in the *Oryza* genus were found to be mainly caused by lineage-specific amplifications of LTR retrotransposons [74, 6]. Our systematic comparative analysis of the sequence of the *MONOCLUM1*-orthologous regions across the *Oryza* genus not only revealed high gene colinearity but also identified new genes that appear to have originated *de novo* in the AA genomes (Fig. 2 and Chen et al., unpub. res.), which highlights the advantage of multiple species comparisons.

Intraspecific local sequence comparisons have also identified interesting features of grass genome structure

Fig. 2 Microcolinearity in the *MONOCULMI*-orthologous regions across the *Oryza* genus. Black boxes represent genes. Red boxes indicate retrotransposons. Fuchsia boxes symbolize DNA transposons. Orthologous genes are connected by lines.



and evolution. A detailed sequence comparison of the *bronze* region of maize inbred lines McC and B73 found that LTR retrotransposon clusters differed one hundred percent in location relative to the genes in the *bronze* region between these two lines [29]. This suggests an amazingly rapid process both for TE insertion and for removal of ancestral TEs. In addition, the first annotation of these two regions suggested that the genes themselves differed between these lines in this region. An apparent four-gene cluster was detected in McC but not in the orthologous position in B73 [29]. Later, these sequences were found to be comprised of four gene fragments within *Helitrons*, a new type of eukaryotic transposon [44, 54, 52, 68]. This phenomenon resembles Pack-MULEs, a type of TE first named and comprehensively described in rice, that also capture and mobilize gene fragments [43]. Although neither *Helitrons* nor Pack-MULEs usually mobilize intact genes, they do commonly acquire more than one gene fragment in the same element. When transcribed, these internal fragments are often fused (via intron processing) into transcripts that could encode novel protein products [43, 68]. This process of exon shuffling, first proposed by Gilbert [32] as the reason for the existence of introns, could be creating new genes in plants at an amazing rate. The maize nuclear genome, for instance, has more than 4,000 *Helitrons* that contain inserted gene fragments [68] (L. Yang and J. Bennetzen, unpub. res.). However, there is not yet a proven case of any of these *Helitron*- or Pack-MULE-generated “new” genes having actually acquired a genetic function essential to its host. Given the rapid rate of unselected DNA loss from plant genomes (see below), it is unlikely that conversion of these chimeric gene candidates into true genes will occur commonly, but even rates as low as one in a million would be significant. Other than the standard route of gene duplication, which primarily creates subfunctionalized or (rarely) mildly modified new gene functions (reviewed in [39]), there is no known aggressive process for the generation of new genes. Perhaps Pack-MULEs and *Helitrons* will eventually be proven to provide this process. At the very least, we expect to see the

discovery of more cases of TE components being co-opted for organismal functions in plants, as in the recent identification of transcription factors in *Arabidopsis* derived from the *Mutator* transposase [57].

Even if TE-vectored gene fragments are rarely if ever true genes with a selected host function, they certainly are a complication to genome annotation. Even without internal gene fragments, low-copy-number TEs are often misannotated as genes, giving rise to as much as two-fold overestimations of gene numbers [7]. This type of overestimation in gene number can play particular havoc with assessment of genic colinearity, as evidenced by studies in rice showing hundreds of gene differences between different races of *O. sativa* that were later shown to all be explained by mis-annotated TEs [9]. Hence, many early publications showing numerous genic exceptions to microcolinearity are incorrect because of this routine annotation error.

Sequence comparisons in closely related haplotypes in *Arabidopsis*, in rice and in wheat have demonstrated that unequal homologous recombination and illegitimate recombination are the major forces that remove DNA from flowering plant genomes [16, 21, 60, 90, 91]. These activities can remove >100 Mb of DNA from a plant genome in just one million years [62], but the rate of removal appears to be much faster in some angiosperms than in others [87]. Most of the removed DNA is derived from TEs, but other intergenic DNA and extra gene copies are also removed by these processes [60].

Several recent studies have accentuated the fact that not all genomic regions evolve at the same rate. Disease resistance gene clusters are known to be unstable even in map position [55], and to also undergo high rates of unequal recombination [76], including some recombination events that are delimited to specific sites that can optimize novel pathogen recognition specificities [70]. Ribosomal RNA gene clusters also appear to vary in map position even in close relatives [25]. Perhaps most surprising, the composition and arrangement of sequences in centromeres have been found to be hyper-variable, primarily by the

process of unequal homologous recombination [61, 63, 65]. This rapid rearrangement by recombination in a region that is deficient in crossovers suggests a very tight control over the outcomes of recombination, especially a powerful bias toward non-crossover, intrastrand and/or sister chromatid outcomes [61]. This core centromeric instability has been argued to yield centromeres that have the potential to out-compete other centromeres for choice as the germinal nucleus in egg development [66].

In summary, local sequence comparisons of closely related grass genomes and of intraspecific haplotypes have begun to reveal the major mechanisms driving genome evolution. These include gene and genome duplication, gene silencing and eventual deletion of duplicated genes subsequent to polyploidization, transposable element amplification, gene movement mediated by transposition of mobile elements, unequal homologous recombination, and illegitimate recombination. All of these processes are quite variable even when comparing closely related species, so their differences in levels of activity (and, possibly, specificity) are responsible for the very different genomes found in flowering plants.

The past, present, and future of plant genome comparisons

Perhaps the most valuable insight gained from comparative genomic analyses in rice and related grasses has been the astounding instability of genome structure against a fairly conserved set of biological functions. As mentioned above, at a local genome level, two maize plants are often more different from each other than a human is from a chimpanzee, or even from a macaque. The grasses and other angiosperms obviously insulate their gene functions from the great majority of this genome change, in manners that we do not now understand at even the most minimal level.

As shown in *Drosophila*, pursuit of full genome analyses in several species within a dense phylogenetic framework can be exceptionally productive [18, 79]. In plants, the *Oryza* genus provides such a unique opportunity to investigate various aspects of gene and genome evolution with the availability of a robust phylogenetic framework [31, 97], rich genomic resources [5, 49], and a near-perfect reference genome [41]. The ongoing sequence comparisons in the *Oryza* genus will provide dramatic and lineage-oriented insights into the creation of new genes, the evolution of gene structure and function, conserved non-coding sequences, the evolutionary dynamics of duplicated gene in polyploid species, centromere drive and a wealth of other issues.

As maize genome sequencing nears completion of its first draft, whole genome comparison of maize and rice will

provide an unprecedented opportunity to study grass genome function and evolution. Because maize is derived from a fairly recent tetraploid [80], identifying the homoeologous segments and subsequent comparisons of these segments will illustrate how genome duplication has shaped the maize genome, and reveal the evolutionary fate of this type of duplicated gene [47, 94]. Because all grass genomes are derived from a shared paleopolyploid [71, 83, 95], identification and comparison of two sets of homoeologous chromosomal segments in rice and four sets of homoeologous chromosomal segments in maize will reveal common and lineage-specific patterns of conservation [77], suggest mechanisms for gene movement [40, 53], and possibly identify signatures of cases where these movements led to significant biological outcomes.

The exciting next few years of grass genome comparative genomics, with great emphasis on the *Oryzae* and on maize and its relatives (e.g., sorghum and sugarcane), will provide a framework for the next generation of plant genome analyses. At the technical level, comparative genome analysis on a few model species like rice, maize, sorghum, and *Brachypodium* has opened up avenues to the highly leveraged study of any other grass. No single species is more enriched for “interesting” genes than any other species, but the traditional tractability of studying these interesting genes was centered on the model species with excellent molecular, physiological, biochemical, cell biological and genetic toolkits. Because of comparative genomics, this historical limitation no longer holds true.

With highly conserved gene content across the grasses, small-genome surrogates (or, even better, those surrogates with sequenced genomes) can be used to provide facile access to any shared grass gene. Moreover, the discovery of novel genes or modified gene functions that make each species unique can now be performed by simple EST analysis or trait mapping. Once these candidate genes for family- or genus- or species-specific gene functions are identified, they can now be easily isolated and tested for the ability to condition novel biological function by introduction into easily-transformed model species.

Despite, perhaps because of, the many important discoveries that have been made over the last 15–20 years of plant comparative genomics, we have more questions to answer now than we did at the outset. Because of the continued extraordinary increases in throughput and decreases in cost of nucleic acid sequence analyses, many more plant species will be investigated with a much broader (and better-conceived) set of phylogenetic justifications. Genetic maps, physical maps and EST analyses are all needed for hundreds or thousands of plant species to identify shared and novel traits. Every one of these genes can be tested for function in a few model species (by forward genetic, reverse genetic and transgenic technologies), so the

orthologues, paralogues and “new” genes can also be compared and “uncovered” in a conducive genetic background or backgrounds. With such torrents of data on the horizon, better tools for sorting the gold from the grit will be needed. We have no doubt that the plant science community is up to this task, and that rice will continue its exceptional comparative genomic contributions to this ongoing golden age of plant biology.

Acknowledgements The writing of this manuscript was supported by grants from the National Science Foundation of the USA (#0501814 to JLB) and the Chinese Academy of Sciences (#KSCX2-YW-N-028 to MC).

References

- Adams KL, Wendel JF. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 2005;8:135–41.
- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. The genome sequence of *Drosophila melanogaster*. *Science* 2000;287:2185–95.
- Ahn S, Anderson JA, Sorrells ME, Tanksley SD. Homoeologous relationships of rice, wheat and maize chromosomes. *Mol Gen Genet* 1993;241:483–90.
- Ahn S, Tanksley SD. Comparative linkage maps of the rice and maize genomes. *Proc Natl Acad Sci USA* 1993;90:7980–4.
- Ammiraju JS, Luo M, Goicoechea JL, Wang W, Kudrna D, et al. The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Res* 2006;16:140–7.
- Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, et al. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *Plant J* 2007;52:342–51.
- Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W. Consistent over-estimation of gene number in complex plant genomes. *Curr Opin Plant Biol* 2004;7:732–6.
- Bennetzen JL, Freeling M. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet* 1993;9:259–61.
- Bennetzen JL, Ma J. The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Curr Opin Plant Biol* 2003;6:128–33.
- Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. *Ann Bot (Lond)* 2005;95:127–32.
- Bonierbale MW, Plaisted RL, Tanksley SD. RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics* 1988;120:1095–103.
- Bowers JE, Arias MA, Asher R, Avise JA, Ball RT, et al. Comparative physical mapping links conservation of micro-synteny to chromosome structure and recombination in grasses. *Proc Natl Acad Sci USA* 2005;102:13206–11.
- Bremer K. Gondwanan evolution of the grass alliance of families (Poales). *Evolution* 2002;56:1374–87.
- Bruggeman R, Rostoks N, Kudrna D, Kilian A, Han F, Chen J, et al. The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. *Proc Natl Acad Sci USA* 2002;99:9328–33.
- Chantret N, Salse J, Sabot F, Bellec A, Laubin B, et al. Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. *J Mol Evol* 2008;66:138–50.
- Chantret N, Salse J, Sabot F, Rahman S, Bellec A, et al. Molecular basis of evolutionary events that shaped the *hardness* locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* 2005;17:1033–45.
- Chen M, SanMiguel P, Bennetzen JL. Sequence organization and conservation in *sh2/al*-homologous regions of sorghum and rice. *Genetics* 1998;148:435–43.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 2007;450:203–18.
- Cunff LL, Garsmeur O, Raboin LM, Pauquet J, Telismart H, et al. Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (*bru1*) in highly polyploid sugarcane ($2n \sim 12 \times \sim 115$). *Genetics* 2008;180:649–60.
- Devos KM, Atkinson MD, Chinoy CN, Harcourt RL, Koebner RMD, Liu CJ, et al. Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor Appl Genet* 1993;85:673–80.
- Devos KM, Brown JK, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res* 2002;12:1075–9.
- Devos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theor Appl Genet* 1995;91:282–8.
- Devos KM, Pittaway TS, Reynolds A, Gale MD. Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. *Theor Appl Genet* 2000;100:190–8.
- Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. *Cell* 2006;127:1309–21.
- Dubcovsky J, Dvorak J. Ribosomal RNA multigene loci: nomads of the Triticeae genomes. *Genetics* 1995;140:1367–77.
- Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, et al. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol* 2001;125:1342–53.
- Feldman M, Liu B, Segal G, Abbo S, Levy AA, Vega JM. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics* 1997;147:1381–7.
- Feuillet C, Keller B. High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl Acad Sci USA* 1999;96:8265–70.
- Fu H, Dooner HK. Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA* 2002;99:9573–8.
- Gale MD, Devos KM. Plant comparative genetics after 10 years. *Science* 1998;282:656–9.
- Ge S, Sang T, Lu BR, Hong DY. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc Natl Acad Sci USA* 1999;96:14400–5.
- Gilbert W. Why genes in pieces? *Nature* 1978;271:501.
- Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, et al. Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature* 2006;439:749–52.
- Gu YQ, Coleman-Derr D, Kong X, Anderson OD. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. *Plant Physiol* 2004;135:459–70.
- Guimaraes CT, Sills GR, Sobral BW. Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. *Proc Natl Acad Sci USA* 1997;94:14261–6.
- Guo H, Moose SP. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* 2003;15:1143–58.

37. Han F, Kilian A, Chen JP, Kudrna D, Steffenson B, Yamamoto K, et al. Sequence analysis of a rice BAC covering the syntenous barley *Rpg1* region. *Genome* 1999;42:1071–6.
38. Hulbert SH, Richter TE, Axtell JD, Bennetzen JL. Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes. *Proc Natl Acad Sci USA* 1990;87:4251–5.
39. Hurler M. Gene duplication: the genomic trade in spare parts. *PLoS Biol* 2004;2:E206.
40. Ilic K, SanMiguel PJ, Bennetzen JL. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc Natl Acad Sci USA* 2003;100:12265–70.
41. IRGSP. The map-based sequence of the rice genome. *Nature* 2005;436:793–800.
42. Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann JC, Arruda P, et al. Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J* 2007;50:574–85.
43. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 2004;431:569–73.
44. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* 2001;98:8714–9.
45. Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. Utility and distribution of conserved noncoding sequences in the grasses. *Proc Natl Acad Sci USA* 2002;99:6147–51.
46. Kashkush K, Feldman M, Levy AA. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 2002;160:1651–9.
47. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 2004;428:617–24.
48. Kellogg EA. Evolutionary history of the grasses. *Plant Physiol* 2001;125:1198–205.
49. Kim H, Hurwitz B, Yu Y, Collura K, Gill N, SanMiguel P, et al. Construction, alignment and analysis of 12 framework physical maps that represent the 10 genome types of the genus *Oryza*. *Genome Biol* 2008;9:R45.
50. Kim H, SanMiguel P, Nelson W, Collura K, Wissotski M, Walling JG, et al. Comparative physical mapping between *Oryza sativa* (AA genome type) and *O. punctata* (BB genome type). *Genetics* 2007;176:379–90.
51. Kim JS, Islam-Faridi MN, Klein PE, Stelly DM, Price HJ, Klein RR, et al. Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. *Genetics* 2005;171:1963–76.
52. Lai J, Li Y, Messing J, Dooner HK. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci USA* 2005;102:9068–73.
53. Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, Llaca V, et al. Gene loss and movement in the maize genome. *Genome Res* 2004;14:1924–31.
54. Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC. The maize genome contains a *helitron* insertion. *Plant Cell* 2003;15:381–91.
55. Leister D, Kurth J, Laurie DA, Yano M, Sasaki T, Devos K, et al. Rapid reorganization of resistance gene homologues in cereal genomes. *Proc Natl Acad Sci USA* 1998;95:370–5.
56. Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, et al. Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. *BMC Genomics* 2008;9:58.
57. Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H. Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. *Science* 2007;318:1302–5.
58. Liu H, Sachidanandam R, Stein L. Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. *Genome Res* 2001;11:2020–6.
59. Lukens LN, Pires JC, Leon E, Vogelzang R, Oslach L, Osborn T. Patterns of sequence loss and cytosine methylation within a population of newly resynthesized *Brassica napus* allopolyploids. *Plant Physiol* 2006;140:336–48.
60. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* 2004;101:12404–10.
61. Ma J, Bennetzen JL. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci USA* 2006;103:383–8.
62. Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* 2004;14:860–9.
63. Ma J, Jackson SA. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res* 2006;16:251–9.
64. Ma J, SanMiguel P, Lai J, Messing J, Bennetzen JL. DNA rearrangement in orthologous *orp* regions of the maize, rice and sorghum genomes. *Genetics* 2005;170:1209–20.
65. Ma J, Wing RA, Bennetzen JL, Jackson SA. Plant centromere organization: a dynamic structure with conserved functions. *Trends Genet* 2007;23:134–9.
66. Malik HS, Henikoff S. Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev* 2002;12:711–8.
67. Moore G, Devos KM, Wang Z, Gale MD. Cereal genome evolution, Grasses, line up and form a circle. *Curr Biol* 1995;5:737–9.
68. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by *helitron*-like transposons generate intraspecies diversity in maize. *Nat Genet* 2005;37:997–1002.
69. Nagamura Y. Conservation of duplicated segments between rice chromosome 11 and chromosome 12. *Breeding Science* 1995;45:373–6.
70. Nagy ED, Bennetzen JL. Pathogen corruption and site-directed recombination at a plant disease resistance gene cluster. *Genome Res*. 2008; doi:10.1101/gr.078766.108.
71. Paterson AH, Bowers JE, Chapman BA. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* 2004;101:9903–8.
72. Paterson AH, Lin YR, Li Z, Schertz KF, Doebley JF, Pinson SR, et al. Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* 1995;269:1714–8.
73. Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, et al. ‘Green revolution’ genes encode mutant gibberellin response modulators. *Nature* 1999;400:256–61.
74. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* 2006;16:1262–9.
75. Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, et al. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* 2002;162:1389–400.
76. Richter TE, Pryor TJ, Bennetzen JL, Hulbert SH. New rust resistance specificities associated with recombination in the *Rpl* complex in maize. *Genetics* 1995;141:373–81.
77. Salse J, Bolot S, Throude M, Jouffé V, Piegu B, Quraishi UM, et al. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 2008;20:11–24.
78. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 1996;274:765–8.

79. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 2007;450:219–32.
80. Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, et al. Close split of sorghum and maize genome progenitors. *Genome Res* 2004;14:1916–23.
81. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005;437:69–87.
82. Thomas CA Jr. The genetic organization of chromosomes. *Annu Rev Genet* 1971;5:237–56.
83. Tian C, Xiong Y, Liu T, Sun S, Chen L, Chen M. Evidence for an ancient whole-genome duplication event in rice and other cereals. *Acta Genetica Sinica* 2005;32:519–27.
84. Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc Natl Acad Sci USA* 1999;96:7409–14.
85. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. The sequence of the human genome. *Science* 2001;291:1304–51.
86. Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, et al. Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. *Plant Cell* 1999;11:1769–84.
87. Vitte C, Bennetzen JL. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc Natl Acad Sci USA* 2006;103:17638–43.
88. Wei F, Coe E, Nelson W, Bharti AK, Engler F, et al. Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* 2007;3:e123.
89. Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J* 2001;26:307–16.
90. Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu ZD, Dubcovsky J, et al. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell* 2003;15:1186–97.
91. Wicker T, Yahiaoui N, Keller B. Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. *Plant J* 2007;51:631–41.
92. Wilson WA, Harrington SE, Woodman WL, Lee M, Sorrells ME, McCouch SR. Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. *Genetics* 1999;153:453–73.
93. Wing RA, Ammiraju JS, Luo M, Kim H, Yu Y, et al. The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol Biol* 2005;59:53–62.
94. Wu Y, Zhu Z, Ma L, Chen M. The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. *Mol Biol Evol* 2008;25:1003–6.
95. Yu J, Wang J, Lin W, Li S, Li H, et al. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* 2005;3:e38.
96. Zhang S, Gu YQ, Singh J, Coleman-Derr D, Brar DS, Jiang N, et al. New insights into *Oryza* genome evolution: high gene colinearity and differential retrotransposon amplification. *Plant Mol Biol* 2007;64:589–600.
97. Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, et al. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol* 2008;9:R49.