Genomic Duplication, Fractionation and the Origin of Regulatory Novelty

Richard J. Langham,* Justine Walsh,* Molly Dunn,[†,1] Cynthia Ko,[†,1]
Stephen A. Goff[†,1] and Michael Freeling*[,2]

*Department of Plant and Microbial Biology, University of California, Berkeley, California 94720
and [†]Torrey Mesa Research Institute, Syngenta, San Diego, California 92121

Manuscript received June 20, 2003
Accepted for publication September 17, 2003

ABSTRACT

Having diverged 50 MYA, rice remained diploid while the maize lineage became tetraploid and then fractionated by losing genes from one or the other duplicate region. We sequenced and annotated 13 maize genes (counting the duplicate gene as one gene) on one or the other of the pair of homeologous maize regions; 12 genes were present in one cluster in rice. Excellent maize-rice synteny was evident, but only after the fractionated maize regions were condensed onto a finished rice map. Excluding the gene we used to define homeologs, we found zero retention. Once retained, fractionation (loss of functioning DNA sequence) could occur within *cis*-acting gene space. We chose a retained duplicate basic leucine zipper transcription factor gene because it was well marked with big, exact phylogenetic footprints (CNSs). Detailed alignments of *lg2* and retained duplicate *lrs1* to their rice ortholog found that fractionation of conserved noncoding sequences (CNSs) was rare, as expected. Of 30 CNSs, 27 were conserved. The 3 unexpected, missing CNSs and a large insertion support subfunctionalization as a reflection of fractionation of *cis*-acting gene space and the recent evolution of *lg2*'s novel maize leaf and shoot developmental functions. In general, the principles of fractionation and consolidation work well in making sense of maize gene and genomic sequence data.

E. B. Lewis (1951) postulated a scheme by which a novel gene may arise following a gene or genome duplication. He based this hypothesis on cases of linked, diversified paralogous genes in Drosophila and maize. The Lewis scheme has much case support (*e.g.*, Ohno 1970; Li 1997; True and Carroll 2002). The theoretical expectation for an average duplicate gene pair is that one will be lost (Haldane 1933). Force *et al.* (1999) suggested a neutral scheme to explain cases in which duplicate gene retention seemed to be higher than expected. In this scheme, duplicates are retained because each partner gene loses a dispensable *cis*-acting function such that the ancestral function is now subfunctionalized into two genes. One consequence of subfunctionalization is that duplicates become fixed in a population so that they remain a potential substrate for novelty (Lynch and Force 2000). Another consequence of subfunctionalization is that the loss of a *cis*-acting function should reflect a change in DNA sequence or sequence arrangement. Indeed, the problem with testing Lewis models and subfunctionalization models experimentally is that measuring function of presumptive *cis*-acting gene space is experimentally laborious. Given the assumption that functional *cis*-acting gene space should be conserved over evolutionary time, just as are exons, then a new measure of gene function is possible: conserved noncoding sequence (CNS) patterns. CNSs are phylogenetic footprints that are so large and/or exact that only two orthologous genes from suitably diverged species are required to measure them with confidence, as has been done for mouse-human (Hardison *et al.* 1997; Dubchak *et al.* 2000; Hardison 2000) and also for maize-rice (Kaplinsky *et al.* 2002).

The terms "fractionation" and "consolidation" are used with specific meaning in this article. No matter how grand the duplication event—be it genomic, segmental, or genic—the immediate result is two DNA sequences (paralogs) where there used to be one; at this point the neutral process of fractionation begins. Fractionation is mutation leading to the loss of redundant function by any of several processes: randomization by substitution of neutral base pairs, deletion, insertion, copy over by simple sequence repeats (SSRs), and similar processes. However, fractionation applies only to situations in which duplication of a *cis*-acting unit of function has occurred, such as duplication of a gene (cistron) or duplication of a *cis*-acting part of a gene that confers some specific component of function above that of being necessary for gene function *per se*. Examples of such specific *cis*-acting function are organ specificity or late onset. Given duplication, fractionation can

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY211535, AY211534, AY180106, and AY180107.

[1]*Present address:* Syngenta Biotechnology, 3054 Cornwallis Rd., Research Triangle Park, NC 27713.

[2]*Corresponding author:* 111 Koshland Hall, Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720. E-mail: freeling@nature.berkeley.edu

cause functional loss that does not remove the full complement of *cis*-acting function. In other words, fractionation is the DNA-level cause of the loss of one of the postduplication paralogs predicted by Haldane (1933) and the mutational cause of the loss of specific *cis*-acting function (subfunctionalization) predicted by Force *et al.* (1999). Fractionation involves change in DNA sequence. Unlike function, DNA sequence can be measured in routine fashion.

Fractionation and consolidation are useful concepts when dealing with the consequences of duplicate genomes, chromosomal segments, or individual genes. For the duplicated segment, genes make useful markers. For one extreme, diagrammed in Figure 1, each gene is lost from one or the other of the two 100% syntenous (homeologous) chromosomes such that they eventually have zero sequence in common. Synteny can be seen only by consolidating (Freeling 2001) the two fractionated homeologs into a predicted ancestral sequence. This 100% fractionation example is essentially our result for the 13-gene maize duplicated region studied here, as will be shown. The alternative extreme is that each duplicate gene in the segment is retained in the duplicated ancestor, leading to 0% fractionation. In this case, consolidation is not necessary to reconstruct a likely ancestor. The concepts of fractionation and consolidation also apply intragenically. Instead of beginning with a chromosomal segment marked with a series of exon clusters (genes), one begins with a duplicated gene space marked by exons and CNSs, ideally marking *cis*-acting gene space with easily monitored islands of near-identical sequence surrounded by random, potentially functionless sequence. Thus, measuring fractionation of CNSs or disruption of a CNS pattern by insertion or deletion could provide DNA-level evidence underlying subfunctionalization (Force *et al.* 1999). We use the concepts of fractionation and consolidation at the chromosomal level and within the gene. We could not find CNSs between genes in our grass chromosomal segment, and so we could not apply fractionation and consolidation to intergenic regions. Interestingly, fractionation at the chromosomal level leads to loss of genes (fractionation of the information content of a segment over two homeologous segments) whereas fractionation within genes can lead to retention of duplicates, each encoding complementing, but partial, function.

The grasses, Poaceae, are particularly important for fractionation research because the common subfamilies of grass turn out to be diverged for a useful amount of time—not too much and not too little—for applying CNS analyses (Kaplinsky *et al.* 2002; Inada *et al.* 2003) and because the well-studied grass, maize, is the descendant of a tetraploidy event. The grasses are a monophyletic family. The ancestor to the major subfamilies of grass lived ~50 million years ago (MYA; Kellogg 2001) and was "diploid" in the sense that rice and maize's tribal relative, sorghum, is also diploid (Devos and Gale 2000). Maize and rice represent different subfamilies and are about as diverged as grasses can be, if the basal subfamilies are ignored. It just happened that this 50-MYA branch point in the grass lineage was recent enough so that large phylogenetic footprints—noncoding sequences that are conserved because of some function—are so strongly conserved that they are about as identical in sequence as are bits of orthologous exon. However, divergence was far enough in the past to assure that each functionless nucleotide would randomize. This leaves maize-rice CNSs as islands of conservation surrounded by unalignable randomness (Kaplinsky *et al.* 2002; Inada *et al.* 2003). While every genome evidences large-scale or whole-genome duplication in its history, as is the case for both Arabidopsis (in an ancestor to crucifers: Blanc *et al.* 2000; Paterson *et al.* 2000; Vision *et al.* 2000; Simillion *et al.* 2002) and the grass ancestor (Goff *et al.* 2002), maize is special in having a comparatively recent tetraploid ancestor. Maize has been described as a descendant of a tetraploidy event happening ~11 MYA (Gaut and Doebley 1997) relative to the maize-rice branch at ~50 MYA (Kellogg 2001) and an intratribal maize-sorghum branch at ~16 MYA (Gaut and Doebley 1997). There has been enough retention of duplicates in maize so that almost all large chromosomal regions have a clear homeologous region(s) elsewhere in the maize genome (Ahn and Tanksley 1993; Wilson *et al.* 1999; Devos and Gale 2000). Estimates of maize gene duplicate retention vary from ~70% (Ahn and Tanksley 1993) to 14% (Freeling 2001).

Kaplinsky *et al.* (2002) have shown that stringent local alignments of orthologous genes from maize and rice often uncover conserved patches of sequence in noncoding DNA (CNSs), conservations reflecting positive selection. A recent analysis of 52 additional maize-rice gene spaces (Inada *et al.* 2003) found that 73% of plant genes have at least one CNS, as strictly defined. The average was about three CNSs per grass gene. The length of these CNSs averaged ~20 bp, but occasionally could be >80 bp in length. Upstream regulatory genes were considerably more CNS-rich than were enzyme-encoding genes. To examine the concept of fractionation within a single gene space, we chose the most CNS-rich gene, with ~30 CNSs, among ~200 genes for which we found published or unpublished CNS data: maize *liguleless2* (*lg2*) and its genomic duplicate, *liguleless related sequence1* (*lrs1*).

The *lg2* gene in maize encodes a basic leucine zipper protein that is necessary to specify an exact sheath-blade transition in the maize leaf (Walsh *et al.* 1998) and to specify a timely transition from vegetative to flowering when the shoot apical meristem is founding tassel (male flower) branches (Walsh and Freeling 1999). The phenotypes of homozygotes for *lg2* deletions support these functions. The *lg2* gene maps to chromosome 3.06; homeolog *lrs1* maps to chromosome 8 near umc7 (data not shown). The deduced protein sequence encoded by these two genes is nearly identical. Our project begins with sequencing two maize bacterial artificial chromosomes (BACs), each containing a *lg2-lrs1* duplicate, and annotating them individually. We then compared each of these sequences to the orthologous rice sequence contributed by the Rice Genome Project (RGP).

MATERIALS AND METHODS

Maize BAC sequence, assembly, and annotation: Maize inbred B73 *Hin*diIII BAC library ZMMBBb filters were purchased from Clemson University Genomics Institute (CUGI) and screened by hybridization with an *lg2* cDNA probe that hybridizes to *lg2* or *lrs1*. DNA was isolated from BAC clone 249I19, called *lg2*-BAC, and from BAC clone 240N14, called *lrs1*-BAC, and the presence of an entire *lg2* or *lrs1* was confirmed using PCR. DNA from *lg2*-BAC and *lrs1*-BAC was sheared and subcloned into pBluescript shotgun libraries. Average insert size was 1.5 kb. Subclones were sequenced from both ends to approximately seven times coverage. Bases were called by Phred (Ewing *et al.* 1998). Vector and *Escherichia coli* DNA was screened using CrossMatch (Phil Green, University of Washington) and reads were assembled into contigs using Phrap version 0.990329 (Phil Green, University of Washington).

Southern hybridizations: *Genomic:* Maize B73 genomic Southerns were performed under high stringency (65° in 0.2× SSPE; 0.2% SDS). A variety of restriction enzymes were used to estimate a minimal number of fragments hybridizing to our exon probes (the probes were genomic gene space containing all or most exons of *lg2, lrs1*, and *unk4* from *lrs1*-BAC or *hypro1* from *lrs1*-BAC). Probes for the *lg2* and *lrs1* pair were used as a control for a *bona fide* retained duplicate. Gels were probed, stripped, and reprobed with the first probe to control for loss of template.

BAC: The CUGI BAC clone 222A1 was identified per methods presented above. Southern analysis using either 5′ or 3′ *lg2* exon probes, which also hybridized with *lrs1*, was used to determine that this clone lacked the 5′-most *lrs1 Hin*diIII restriction fragment, but carried the 3′-end of the *lrs1* gene. (The *lrs1* BAC is missing just the 3′-end of the gene.) Hybridizations using exon probes from genes on the *lg2*-BAC were performed at moderate stringency (65° in 0.5× SSPE; 0.2% SDS) in an effort to visualize a possible additional genomic copy downstream of *lrs1*, a region not represented on the original *lrs1*-BAC of Figure 2.

CNS discovery: Maize-rice CNSs were found using the BLAST-2-sequence parameters given in Kaplinsky *et al.* (2002) with the following additions. For CNS discovery in the *lg2* and *lrs1* gene spaces (accession nos. AY180106 and AY180107), hits that were on the opposite strands, hits that moved >50 bp from the expectation of positional conservation, and hits that were >75% mono- or dinucleotide simple sequence repeats were removed from the graphic. This was done only to aid in visualization of the underlying order of CNS sequence and position.

In our search for CNSs that might have existed between grass genes, we needed to adapt our definition of a potential hit in order not to call isolated 15/15 exact homologies that might have occurred by chance alone. We excluded any single hit below an *e*-value equivalent to 17/17 exact match and demanded a conserved cluster of two or more 15/15 hits in the same orientation. We disregarded retrotransposon gene hits and hits to mono- or dinucleotide SSRs. Under these conditions, we found zero CNSs detached or spaced away from a cluster of exons. In our (unsuccessful) search for unexpected, intergenic maize-rice CNSs or other homologies, we expanded our rice P1-derived artificial chromosome (PAC) sequence AP003287 with both adjacent chromosome 1 overlapping rice PACs AP003794 and AP003681.

Maize *lrs1* and *lg2* gene space sequence and annotation: The *lg2* and *lrs1* were contained on multiple nonoverlapping contigs, as identified in Figure 2. PCR using the BACs as templates was used to piece together the contigs. PCR products were sequenced at the University of California at Berkeley Sequencing Facility. The *lrs1*-BAC did not contain the 3′ end of *lrs1* (bp 10,327–10,850 of AY180107). This finishing sequence is from a previously identified *lrs1*-containing genomic clone obtained from a maize inbred B73 genomic library (our unpublished results). The exons of both *lg2* (AY190106) and *lrs1* (AY180106, called "lg2-like" by GenBank) were experimentally determined using a complete cDNA sequence from maize LG2-mRNA, accession no. AF036949.

RESULTS

Fractionation and consolidation of a 13-gene segment of grass chromosome: We chose maize inbred B73 BACs containing *lg2* (GenBank accession no. AY211535; cDNA sequence was available) and its genomic duplicate, *lrs1* (AY211534); these were sequenced to seven times coverage and assembled into contigs as described in materials and methods. The resulting maize contigs were individually assessed by BLASTx (Altschule *et al.* 1990) and GeneMark.hmm (Lukashin and Borodovsky 1998) using the *Caenorhabditis elegans* model. The results are diagrammed on the *lg2*-BAC and *lrs1*-BAC tracks of Figure 2. Most maize contigs appeared to be composed entirely of transposons. These are not plotted in Figure 1. A total of 8 genes were called from *lrs1*-BAC contigs and a total of 5 from *lg2*-BAC contigs, which sums to the *lg2/lrs1* pair plus 11 *unpaired* genes. Independently, the contigs, whether or not we annotated genes on them, were virtually assembled onto a single rice PAC (AP003287). Of the 12 genes we annotated in maize (counting *lg2* and *lrs1* as one gene), 11 were present in a cluster and annotated by the RGP (rice track of Figure 1). The rice annotation did not lead to new genes being discovered in maize. In addition to the 11 maize genes anchored to rice, our maize-rice comparison found an exon-like homology in the RGP PAC at 71,700–72,198: a *ferredoxin* gene missed by RGP and by ourselves. This brings the total count of genes in both maize and rice to 12. All of these genes are ~85% (75–89%) identical in sequence over at least 75% of the exon (see supplemental Figure 1 available at http://www.genetics.org/supplemental/). Of the 5 known rice genes, all were in maize. Of the 7 rice genes with experimental evidence for existence, but unknown function (*unk*s), 6 were found in maize; *unk3* was missing. Of the 9 hypothetical rice genes (*hypro*s), maize carried *hypro1*. Since hypothetical genes are just that, we do not count a missing "*hypro*" as an unexpected event. Were it not for the exceptional *unk3*, which is unexpectedly missing in both maize homeologs, consolidation of the fraction-

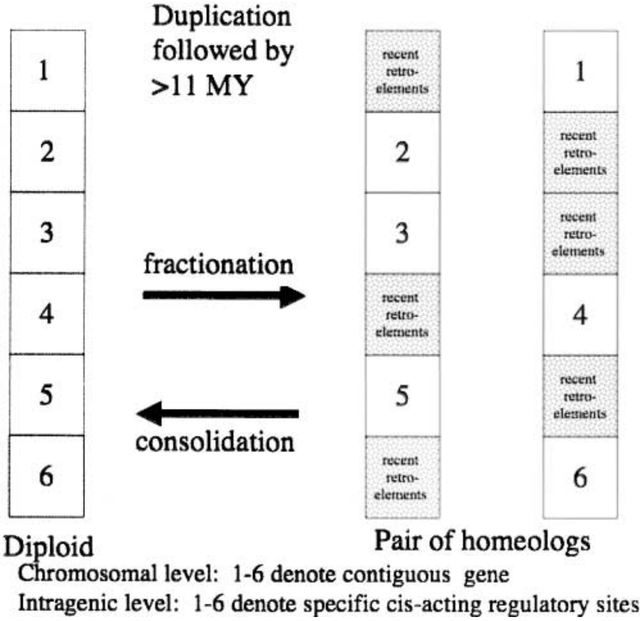


FIGURE 1.—One extreme outcome of duplication followed by fractionation. In this case, the two resulting homeologs have been 100% fractionated. They share zero sequences in common. Only by consolidating the homeologs into a putative ancestor can the perfect synteny of these sequences be evidenced (FREELING 2001). Consolidation is a mental exercise, not a mechanism. In reality, some duplicates are retained, meaning that fractionation is <100%. These principles apply to genes on a chromosome, *cis*-acting sites that may act from a distance, and specific *cis*-acting regulatory sites that may exist within a gene's space.

ated *lg2*-BAC and the fractionated *lrs1*-BAC would yield perfect synteny for this region of grass chromosome.

An exceptional maize *rab7A*-related gene fragment was supported by convincing BLASTx hits, but was not present in the rice PAC; 11 of 12 maize genes were present. *Rab* genes are part of multigene families of small GTPases. There are several *rabA* genes in rice (data not shown).

The single maize genes we found orthologous to the rice genes on PAC AP003287 gave exon identities from 75 to 89%. This degree of conservation is consistent with recent, post-tetraploidy function (see supplemental Figure 1 available at http://www.genetics.org/supplemental/), 11 MY of no selection being adequate to greatly degenerate identities. As for function today, we have genetic and expression evidence for maize *lg2* function only.

Where did the 11 fractionated maize genes go? An unequivocal answer to this question for this or any individual maize case study is simply not possible because the maize genome is not sequenced and because use of stringent Southern hybridization data is flawed. When an expected genomic fragment is *not* found using Southerns, then chances are high (not proved) that the gene is indeed gone. The flaw is, when Southerns do find one or more fragments in the genome in addition to the gene in question, interpretation is equivocal; perhaps there are paralogs, or particularly conserved gene regions among paralogs, or simply spurious hybridizations, or a fragment of a gene that is functionally dead. For this reason, our best attempt to address this "where did the genes go?" question requires some explanation.

Consider the extreme alternative: every newly duplicated gene or gene cluster in maize has a reasonably high probability of having moved physically to another unlinked location over the last 11 MY. Such an explanation simply could not explain our data because a near-complete ancestral genome—11 of 12 genes—was left behind at the expected locus on the homeologs (Figure 2). Only selection could account for this, and selection for this one complete function would not exist if other unlinked copies also function. Thus, logic alone leads to a likely conclusion: the missing genes are functionally inert. However, this is an argument, not a proof. The genes on the *lrs1*-BAC surround the genes on the *lg2*-BAC. The easiest way to account for this with a single chromosomal aberration would be to evoke an inversion or short-range movement that would place the "missing" genes on the *lg2*-BAC on the other side (to the left in Figure 2) of the *lrs1*-BAC; this rearrangement would be within the *lrs1* chromosome. To test for this local movement coincidence, we went back to the CUGI B73 BAC library and found an additional *lrs1*-specific BAC, clone 222A1, that meets with *lrs1*-BAC at *lrs1*, which together span 230 kbp. This new ∼100-kbp BAC was grown and isolated as were the original BACs, and Southern analysis was used to determine that the clone lacked the 5′-most *lrs1 Hin*dIII restriction fragment (leaving ∼7.3 kb of *lrs1* at one end). This positions the new BAC as running to the left of the *lrs1* BAC diagrammed in Figure 2, adding ∼93 kbp of chromosomal sequence, as diagrammed in Figure 3. As a positive control, the *hypro1* and *chitinaseB* probes were also used to hybridize to the *lg2*-BAC from which they came. The autoradiographic results of this hybridization experiment are shown in Figure 3: the missing genes are not within this extra 93 kbp to the left of *lrs1*, but are detected, as expected, on the *lg2*-BAC (Figure 3, right lanes). We did not address rearrangements that might have placed the missing genes at a distance >93 kbp, so these data also are only supportive, not conclusive.

Even given the flaws of genomic Southern searches for missing genes, we went hunting for two: *unk4* and *hypro1*. The results are available from supplemental Figure 2 at http://www.genetics.org/supplemental/. *unk4* exon sequence was used to probe B73 Southern blots, which were then washed under stringent conditions; no second fragment evidencing a potentially duplicate gene was found using several enzymes. This negative result constitutes strong but not conclusive support for loss or randomization of the *lrs1*-homeolog of the *unk4*

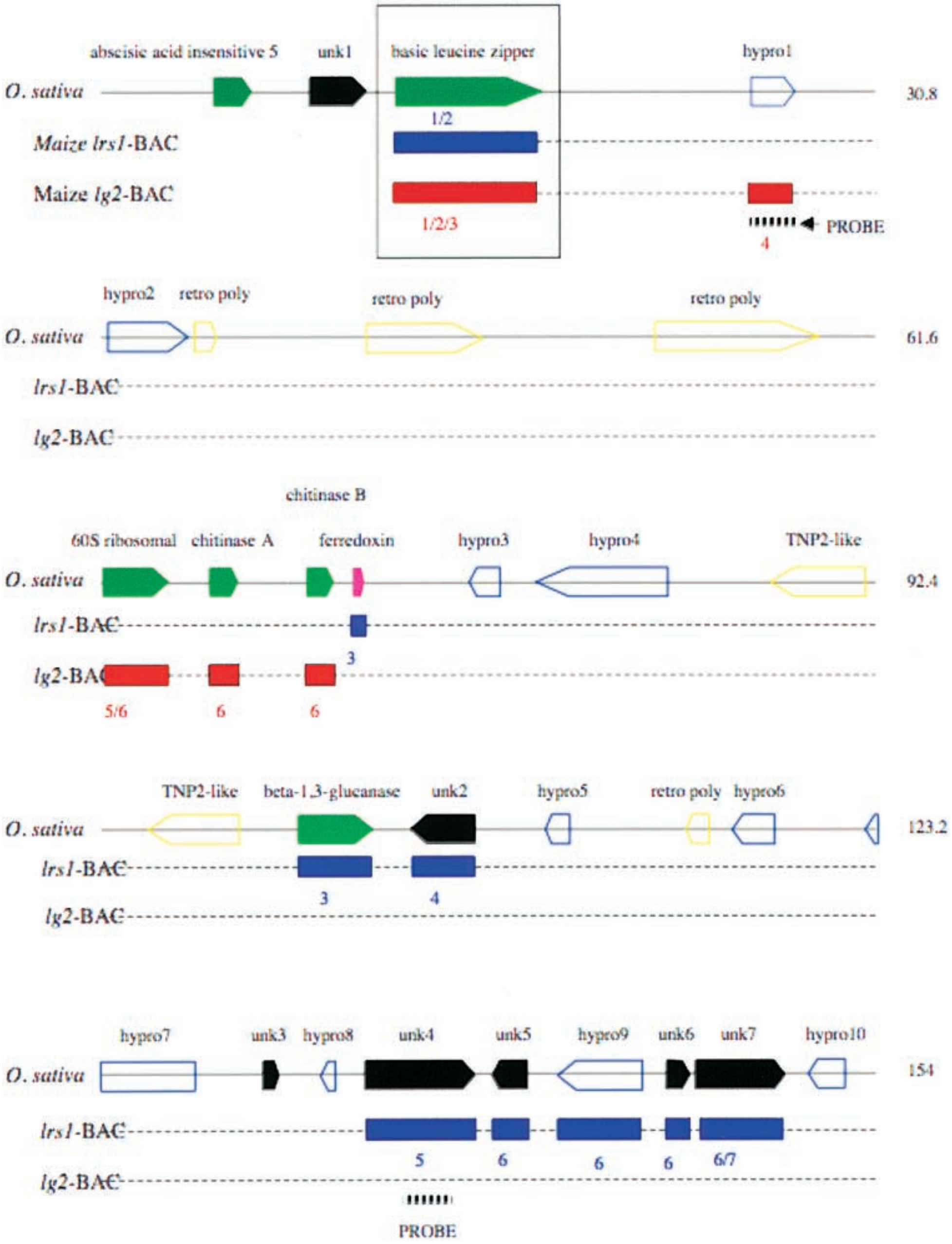


Figure 2.—A 13-gene region of the maize/rice genome. Virtual alignment of an annotated RGP PAC sequence (AP003287) with two maize BAC contig collections is shown: maize *lrs1*-BAC (AY211534, seven contigs with blue numbers from 1 to 7 from left to right, covering 53,425 bp) and maize *lg2*-BAC (AY211535, six contigs with red numbers from 1 to 6 from left to right, covering 23,962 bp). The two homeologous regions in maize were identified because of the retention of the *lrs1/lg2* gene pair; this retained duplicate is enclosed in a box. The top line is rice chromosome drawn approximately to scale. The middle line anchors genes that we annotated on the maize *lrs1*-BAC, and the bottom line anchors genes annotated on the maize *lg2*-BAC. Except for the magenta rice gene, a *ferredoxin* gene annotated only because of synteny, all rice genes were annotated by RGP; those with experimental support are solid green or black and those that are known parts of transposons or are hypothetical only are represented as outlines. Maize contigs containing transposons only, which constituted most of the BAC sequences, were ignored after they failed to hybridize to any portion of the rice PAC. Those that did hybridize reflected an annotated gene. The blue and red numbers relate these gene-carrying contigs to the order of contigs as they appear in GenBank.

gene. We found a second hybridizing fragment when *hypro1* was used as probe, and the *lg2/lrs1* duplicate was found as a control, as expected. Taken at face value, this positive result supports retention of this gene. However, since false positives are expected, the *hypro1* result is difficult to interpret correctly.

A search for phylogenetic footprint markers in maize intergenic space: Were grasses like mammals, then we could hope to find maize-rice CNSs (large, exact phylogenetic footprints) between genes, somehow acting over several to many kilobases to affect activity of a region of a chromosome (Loots *et al.* 2002 and references therein). A significant percentage of CNSs found between humans and mouse are located between genes, and a significant fraction of these are structured as matrix attachment regions (Glazko *et al.* 2003). A putative scaffold attachment region has been shown to be a phylogenetic footprint between sorghum, a close maize relative, and rice (Avramova *et al.* 1998).

Given these successes, we used the sequence libraries comprising our two maize BACs as queries and “blasted” these onto the known orthologous rice PAC and the two adjacent RGP rice PAC clones as subjects. We used “find CNS” bl2seq conditions (Kaplinsky *et al.* 2002) known to find all hits of *e*-value equal to or greater than a 15/15 exact nucleotide match. We looked for a pattern of two or more small hits or one large hit in any area of a contig not containing exons or not farther than a few kilobases from called exons. While most of our 11 genes certainly had CNSs very close or within 2.5 kb

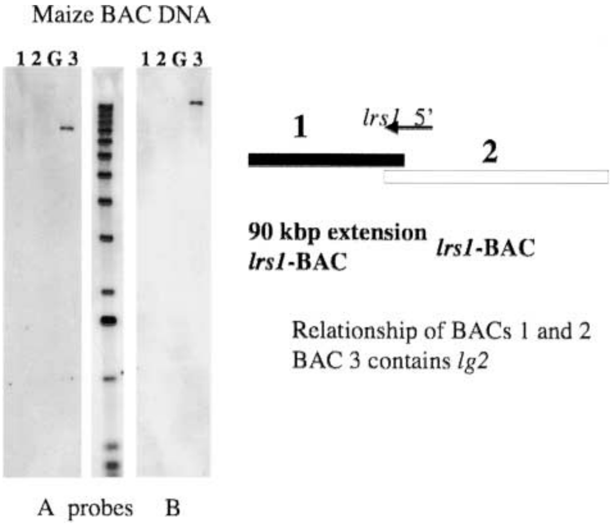


Figure 3.—Two autoradiograms of the same Southern blot in which DNA was restricted with *Bam*HI, separated, blotted, and probed with an exon-rich gene fragment present on the *lg2*-BAC (lane 3, 249I19) but fractionated from the *lrs1*-BAC (lane 2, 240N14) and an additional BAC (lane 1, 222A1) that extends the *lrs1*-BAC-marked chromosome ~90 kbp 3′ of the *lrs1* gene. The relationship of *lrs1*-BACs 1 and 2 is diagrammed; scale is approximate. The 3′-end of the *lrs1*-BAC (2), as diagrammed, includes the 3′ of *lrs1* added during finishing. (A) Hybridization was to the entire coding sequence of *hypro1*, this being the closest gene to *lg2* on the *lg2*-BAC. (B) The blot in A was stripped and reprobed with a probe covering the entire coding sequence of *chitinaseB*; this is potentially the furthest gene from *lg2* on the *lg2*-BAC. Both blots were hybridized at moderate stringency (65° in 0.5× SSPE; 0.2% SDS) in an effort to visualize a possible genomic duplicate. The lanes carried either BAC (1, 2, and 3 as represented by the diagram) or B73 whole-genome (G) DNA. Note that neither of the *lrs1*-containing BACs (1 and 2) carried these *lg2* region genes. However, the *lg2*-BAC did carry these genes (lane 3), as expected. Were this blot exposed longer, one or two "bands" would be seen in the B73 whole-genome DNA lanes, denoted G. The central DNA ladder is for sizing.

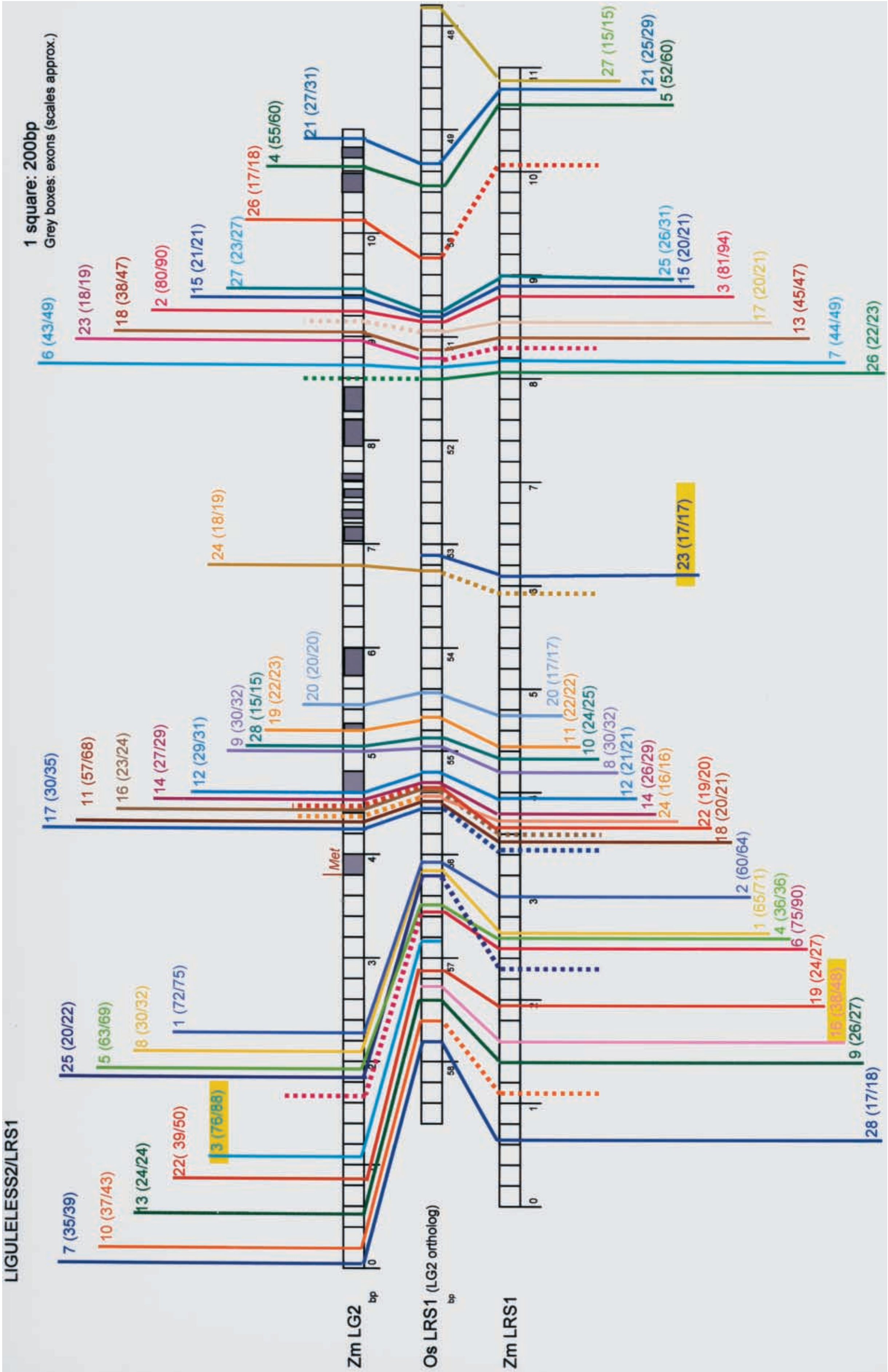
of exons, there were no distant or intergenic CNSs. Therefore, we have no intragenic markers to check for fractionation of intergenic regions following the tetraploidy event.

Fractionation of sequence markers within a single gene's space: We chose our BACs because we knew that *lg2* and *lrs1* were retained duplicate genes and because we knew that *lg2/lrs1* was particularly CNS rich even among upstream regulatory grass genes, with ~30 individual CNSs identified (Inada *et al.* 2003 for *lrs1*). We predicted that these CNSs would serve as gene-space markers with which to test intrageneic fractionation following the tetraploidy event. Exons cannot fractionate without inactivating gene function. However, some *cis*-acting sequences might fractionate if they were not essential to basal gene function (Force *et al.* 1999; Lynch and Force 2000). Using Kaplinsky *et al.*'s (2002) computational definition of grass CNSs, we found 30 CNSs, including 7 CNSs >40 bp long. These 30 CNSs are identified by the colored, "parallel" lines of Figure 4. CNSs composed of >75% mono- or dinucleotide SSRs were discarded, and alignments that change strand (that were inverted) were also excluded for Figure 4.

It is important to note that the solid lines of Figure 4 denote CNSs that are derived from two independent maize-rice pairwise blasts: *lg2*-rice ortholog (Figure 4, top pair) and *lrs1*-rice ortholog (Figure 4, bottom pair). The two results of these two alignments are plotted onto the single rice ortholog gene space (Figure 4, center sequence) that they have in common. The overall result is that almost all of the intragenic *lg2/lrs1* CNSs have been "conserved," not fractionated, over the 11 MY following the tetraploidy event.

To evaluate the data of Figure 4 in an informed way, it is important to calculate the expectation for conservation of neutral sequences when the common ancestor lived 11 MYA, the approximate time of the maize tetraploidy event. Any 15-bp positionally conserved sequence comparing maize and rice (ancestor 50 MYA) has about four chances in a million of being carried over in the absence of selection (Kaplinsky *et al.* 2002; Inada *et al.* 2003). For maize-maize (ancestor 11 MYA), the expectation of carryover is 13%. (Both calculations assume a uniform 7×10^{-9} base substitutions/neutral base pair/year.) Therefore, the lower *e*-value CNSs of Figure 4 (like *lg2*CNS27, 23/27) have a calculated 15% chance of existing on both homeologs of maize without selection, and more significant CNSs (like *lg2*CNS1, 72/75) have a much higher expectation for conservation whether selected for or not because they can degrade in significance and still be called CNSs. The maize duplication event occurred so recently that there has not been enough time to randomize functionless sequences by base substitution alone. Therefore, it is safest to assume that "conservation" of CNSs between the maize homeologous genes *lg2* and *lrs1* is expected whether these CNSs function or not. For us to maximize our chance of seeing this expected conservation (carryover),

Figure 4.—Two independent blast comparisons plotted coordinately: maize gene *lg2* with its rice ortholog and maize gene *lrs1* with the same rice ortholog. The purple alignment line of lg2CNS17 represents how sequences near 5′ exons align. Exons were identified in genomic DNA for all three genes using the complete cDNA of LG2-mRNA (AF036949) and then masked. Bl2seq conditions were modified from those of Kaplinsky *et al.* (2002). The BLAST result is represented by the solid, multicolored lines connecting the maize *lg2* (Zm LG2) and rice *lrs1* (*Os LRS1*) gene diagrams. The identity match is indicated parenthetically. A color connects CNSs that are essentially the same. A dotted line reflects a maize *lg2-lrs1* CNS retention, but just below the 15/15 cutoff. A yellow highlight denotes those rare CNSs that are fractionated. The insertion into *lg2* promoter is noted.


LIGULELESS2/LRS1
1 square: 200bp
Grey boxes: exons (scales approx.)
Zm LG2
bp
Os LRS1 (LG2 ortholog)
bp
Zm LRS1
Met
0
1
2
3
4
5
6
7
8
9
10
48
49
50
51
52
53
54
55
56
57
58
11
7 (35/39)
10 (37/43)
13 (24/24)
22(39/50)
3 (76/88)
25 (20/22)
5 (63/69)
8 (30/32)
1 (72/75)
17 (30/35)
11 (57/68)
16 (23/24)
14 (27/29)
12 (29/31)
9 (30/32)
28 (15/15)
19 (22/23)
20 (20/20)
24 (18/19)
6 (43/49)
23 (18/19)
18 (38/47)
2 (80/90)
15 (21/21)
27 (23/27)
26 (17/18)
4 (55/60)
21 (27/31)
28 (17/18)
9 (26/27)
16 (38/48)
19 (24/27)
6 (75/90)
4 (36/36)
1 (65/71)
2 (60/64)
18 (20/21)
22 (19/20)
24 (16/16)
14 (26/29)
12 (21/21)
8 (30/32)
10 (24/25)
11 (22/22)
20 (17/17)
23 (17/17)
26 (22/23)
7 (44/49)
13 (45/47)
17 (20/21)
3 (81/94)
15 (20/21)
25 (26/31)
5 (52/60)
21 (25/29)
27 (15/15)

we manually inspected each sequence when the original data indicated that a CNS had fractionated. The dotted lines of Figure 4 denote that manual inspection did indeed find an alignable, but degraded, sequence. This situation is exemplified by the dotted extension of yellow *lg2*CNS24 (18/19) that does exist in the expected position in *lrs1*. So, if base substitution were the only fractionation mechanism, the general expectation is that maize-maize CNS patterns will appear to be conserved, but are actually carryover from the common ancestor. Of course, base substitution is not the only mechanism of fractionation, as will be discussed.

On a background of expected conservation, the exceptions are outstanding. Figure 4 denotes four exceptions to conservation: a 1.4-kbp insertion in the promoter region of *lg2* is visible because it moves almost all 5′ *lg2* CNSs upstream. Additionally, three CNSs are fractionated, as denoted by yellow highlighting. These three are the large (76/88) promoter *lg2*CNS3 that is not present in maize *lrs1* and the smaller *lrs1*CNSs—CNS16 (38/48) and CNS23 (17/17)—that are not present in *lg2*. "Not present" in this case means that no amount of imagination could find an alignment anywhere within the gene space. As with the fractionation results involving genes on our chromosomes (Figure 2), fractionation within the *lg2/lrs1* gene behaves like a qualitative character: a CNS is either retained or fractionated.

Observation of Figure 4 reveals that the well-studied maize *lg2* gene is the more unique and divergent gene of the *lg2-lrs1* pair. The insertion in *lg2* is particularly striking, as is the loss of an 88-bp promoter CNS. The divergence evidenced in CNS pattern, combined with genetic studies in rice and maize, which will be discussed, support the hypothesis that the LG2 regulatory function in maize is newly evolved, a case in support of the Lewis scheme for the evolution of novelty. Since the maiz*e lrs1* gene has lost two CNSs, and the *lg2* has lost one, the involvement of subfunctionalization becomes a reasonable hypothesis.

DISCUSSION

The concepts of fractionation and consolidation have worked well in our efforts to reconstruct the evolutionary history of a 13-gene segment of a grass chromosome and also to reconstruct the evolution of a regulatory gene with a novel function. We attempted to find markers between the genes of our maize BACs by looking for CNSs with the rice orthologous chromosome. Unlike the situation in mammals, all big phylogenetic footprints were associated closely with exons, so we could not test fractionation of any sort of long-range, *cis*-regulatory function.

Our annotation of maize BAC contigs for gene content identified 13 genes, and 12 of these were present in one chromosomal region of rice. We found one new gene in the region by synteny only; this gene eluded annotation by ourselves in maize and also by the Rice Genome Project. In general, virtual assembly of maize BAC contigs using a bit of finished rice genome worked efficiently. Except for the gene we chose as a retained duplicate, *lg2/lrs1*, none of the other 12 "experimental" maize genes we found in this single grass region were retained in the homeologous maize BACs (Figure 2) or in an adjacent *lrs1*-BAC (Figure 3). Hybridization evidence strongly supports loss for one of these "missing genes" and it can be reasonably argued that the average missing gene must be lost or randomized, not moved elsewhere in the genome. Even so, our data support only the contention that genes fractionated from our region are actually missing from the genome. However, this case of 0% retention (not counting the *lg2/lrs1* gene pair that was a given) is not unequivocal in itself, and it would be wrong to generalize from it.

Zero percent duplicate retention for maize is obviously too low. On the other hand, the estimate of 70% retention is probably too high. The Ahn and Tanksley (1993) estimate is based on Southern hybridization data, and there are many situations in which a fractionated gene might be evidenced as a false positive using such hybridization data: the existence of paralogs (duplications) that precede the duplication event, the existence of some very conserved regions among otherwise distant paralogs, the existence of spurious hybridizations of high GC or SSR regions, the existence of dead gene fragments, and the like. Two other case studies do not use genomic hybridization data, and each also yields a low estimate of retained duplicates for maize. The first involves identification and mapping of a gene family in both rice and maize (Sentoku *et al.* 1999). In this study, seven *knox* class I homeobox genes were identified by homeobox sequence and map position in rice and then related to sequence and map positions of the most homologous (orthologous) genes in maize: only one of these seven was retained as a duplicate in maize (*rs1/gn1*), which computes to ∼14% retention (Freeling 2001). In a case study similar to our own in that both homeologs from maize were used in the comparison, Ilic *et al.* (2003 and data shared with us before publication) found that of eight genes present in the reconstructed maize ancestor, only one was retained as a duplicate; this computes to 12.5% retention. These workers found one result that is of special interest: partially fractionated genes. That is, cases in which one of the retained pair of genes was an obvious pseudogene with lowered nucleotide identity. This result wreaks havoc with attempts to measure anything meaningful using positive results of genomic hybridization experiments. On the basis of these three case studies only, we conclude that maize is well along on the path toward diploidy, as predicted by Haldane (1933). So far, case studies found 14, 12.5, and 0% (this study) retention. More case studies are needed.

Given that fractionation of recently duplicated maize chromosomes has been extensive, future research should be particularly careful not to misinterpret nonsyntenic results when both homeologs are not included in the study. A case in point is the recently published comparison of gene content from two different genotypes in homologous regions of maize chromosome 9 near *bz1* (Fu and Dooner 2002). In their study, one cultivar had genes missing in the region as compared to the other. Fu and Dooner did not sequence the homeologous *bz1* region where these missing genes might have been. This omission leaves open the possibility that multiple editions of post-tetraploidy fractionation in maize might have continued into modern times into the teosinte (wild maize) gene pool from which the various races of maize were selected by humans. In other words, multiple fractionation outcomes from the maize tetraploidy event may have generated diverse gene contents and may conceivably be generating maize diversity even today. In any case, our results do not detract from Fu and Dooner's (2002) novel hypothesis that maize heterosis might be explained at the level of gene content. In general, given the high gene fractionation expectation for maize, any random stretch of maize chromosome is expected to be incomplete. Both or all homeologs must be sequenced, and the results condensed into a putative ancestor before suggesting that maize genes have unexpectedly gone missing.

In mammals, there is solid evidence for the existence of CNSs that act on more than one gene and often from a distance >10 kbp (Loots *et al.* 2002; Glazko *et al.* 2003). We found no such intergenic CNSs in our maize BACs. The only convincing CNSs or patterns of phylogenetic footprints between either homeologous maize chromosome segment individually and its orthologous rice chromosome occurred between exons of the same gene or within a few kilobases of them. Therefore, we have no chromosomal regional markers to use to test for fractionation following the tetraploidy event. This result would be a surprise if we thought that CNS research results found in mouse-human comparisons would tend to hold up in grasses. In fact, while ~35% of noncoding gene space in mammals is CNS, only ~2% of grass gene space is conserved (Inada *et al.* 2003). Indeed, 27% of grass genes have no CNSs at all. Perhaps the sort of gene regulation that explains CNS function is utilized far more intensively in higher animals than in higher plants (Inada *et al.* 2003).

The ~30 CNSs characterizing both maize *lg2*/rice and maize *lrs1*/rice (Figure 4) provided an adequate amount of marker detail for this gene's space. This TGA1a-type basic leucine zipper transcription factor gene is not the average gene. Rather, it is the most CNS rich of any grass gene we have measured (Inada *et al.* 2003; our unpublished results). The general conclusion that CNS position and sequence tend to be retained in both homeologs is an inescapable deduction from Figure 4. The calculations reported in the results support the conclusion that the rate of neutral base-pair substitution traditionally assigned to the grass family is not high enough to randomize nucleotide sequence over only 11 MY. So, if base substitution were the only mechanism of fractionation, then our general result of maize-maize sequence conservation is not surprising. It is clear that other mechanisms of fractionation operated in the maize lineage during the last 11 MY. The mechanism that completely fractionated our 13-gene grass chromosomal segment in <11 MY, for example, appears to have been deletion. Copy over by simple sequence repeats is another possible fractionation mechanism, as is any mechanism involving transposon insertion or excision (*e.g.*, "scrambling"; Kloeckener-Gruissem and Freeling 1995). Base substitution is not the only mechanism for fractionation. If a mechanism has an average target greater than a single base pair, then there is the complication that fractionation would remove a linked group of elements. For example, the average deletion in maize could be in the kilobase range; if so, it would make an inefficient intragenic fractionation mechanism because removing a neutral CNS would tend to remove an essential CNS or exon, and lethality would result. To the extent that a mutation mechanism acts on small (1–300 bp) targets, it would mediate intragenic fractionation as well as chromosomal fractionation. At present, we have no quantitative estimates of expected rates of any sort of mutation in the grasses except base substitution, but that does not imply that base substitution is the primary mutational agent.

Intragenic fractionation did occur in the percentage range: two shorter CNSs are missing from *lrs1* and one particularly significant CNS is missing from maize *lg2* (Figure 4). These CNSs are fractionated even when we manually looked for any remnant of alignable sequence anywhere in the gene space. The fractionated CNSs were unexpected only because we do not know rates of any sort of mutational mechanism except base substitution; they appear to have been deleted or copied over, not randomized 1 bp at a time.

The *lg2* insertion is a gross change of gene content (Figure 4). The nucleotides within this 1.4-kb insertion are not structured like known transposons and do not exist anywhere in the rice genome; there is inadequate maize or sorghum (a tribal relative) sequence to hope to find an origin for this post-tetraploidy-inserted sequence (our unpublished results). We predict that this insertion, and perhaps the rare CNSs that were fractionated, conditioned a change of expression of this transcription factor that somehow evolved into a new, specific leaf function. In other words, the *lg2* gene and the specific LG2 functions appear to have evolved recently, and after the tetraploidy event, in general support of the Lewis (1951) scheme.

Were *lg2* truly a novel gene that evolved just a few million years ago, it would explain the peculiar distribution of known *liguleless* genes in maize and rice. The absence of a ligule in a grass plant is readily observed because of an upright, leaves-up stature. In maize, this phenotype is saturated: of 27 independently screened liguleless mutants, 18 are *lg1* alleles, 9 are *lg2* alleles, and 0 map elsewhere (Harper and Freeling 1996; D. Braun and J. Walsh, unpublished results). All *liguleless* mutants reported in rice map to 4L, which is where rice *lg1* is located (Kaplinsky *et al.* 2002). There are no *liguleless* mutants mapping to 8L in maize, the site of *lrs1*, and there are also no reports of a *liguleless* mutant mapping near *lrs1* (or away from *lg1* in rice 4L) in rice (our unpublished observations). In conclusion, the evidence is strong that *lg2* transcription factor regulation has been modified and co-opted (True and Carroll 2002) to a new, ligule-specific function at some point or points after tetraploidy.

Little is known about the exact phylogenetic branch point in the tribe Andropogoneae where this tetraploidy event/*lg2* origin took place. Recent sequence data (Mathews *et al.* 2002) have been subjected to Bayesean analysis of base substitution frequencies from several genes in many tribes constructed from a 50% node consensus of 95,000 trees. The genus Elionuris, classified in the Rottboelliinae subtribe by Clayton and Renvoize (1986), grouped with maize and Tripsacum (known to be a genus very close to Zea). Other Rottboelliinae grouped in various places in the tree and not together. The maize/Tripsacum lineage tetraploidy branch point is not placed. Only now are the leaves and ligules of these tropical grasses being observed by collectors in search of evidence of a LG2 function. Mapping the *lg2* insertion and *lg2/lrs1* fractionated CNSs in Rottboelliinae species onto the tribal phylogenetic tree should be illuminating. Since we have an anti-LG2 antibody that localizes maize LG2 to the adaxial layers of the primordial ligule (J. Walsh and M. Freeling, unpublished results), mapping this immunolocalization as a character onto the tribal phylogenetic tree should also be illuminating.

The function of the *lg2* insertion, the fractionated CNSs, or any CNS is not known. We use CNSs as conserved markers only. The CNS/insertion fractionation data of Figure 4 provide a logically sound starting point for analyses to find these functions using conventional molecular genetics. Given these functions, it should be possible to reconstruct the regulatory history of an ancestral basic leucine zipper transcription factor gene as it evolved to specify the Liguleless2 function. Given the complexity of the *lg2/lrs1* grass gene, with its many CNS markers, mapping insertions and deletions and finding fractionated CNSs permits a computational parsing of gene space that would be difficult or impossible to do in any other way.

We thank Damon Lisch and an anonymous reviewer for thoughtful comments, Randall Tyers for help finishing the *lg2* and *lrs1* sequences, and Nancy Nelson for expert editorial help. Funding was provided by the University of California, Berkeley-Syngenta strategic alliance to M.F. and National Institutes of Health grant 2RO1-GM42610 to M.F.

LITERATURE CITED

Ahn, S., and S. D. Tanksley, 1993 Comparative linkage maps of rice and maize genomes. Proc. Natl. Acad. Sci. USA **90:** 7980–7984.

Altschule, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tools. J. Mol. Biol. **215:** 403–410.

Avramova, Z., A. Tikhonov, M. Chen and J. L. Bennetzen, 1998 Matrix attachment regions and the structural colinearity in the genomes of two grass species. Nucleic Acids Res. **26:** 761–767.

Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12:** 1093–1101.

Clayton, W. D., and S. Renvoize, 1986 *Genera Graminum: Grasses of the World.* Kew Bulletin Additional Series XIII, Royal Botanical Gardens, Kew, Her Majesty's Stationary Office, London.

Devos, K. M., and M. D. Gale, 2000 Genome relationships: the grass model in current research. Plant Cell **12:** 637–646.

Dubchak, I., M. Brudno, G. G. Loots, L. Pachter, C. Mayor *et al.*, 2000 Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. **10:** 1304–1306.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Force, A. M., M. Lynch, F. B. Pickett, A. Amores and Y.-L. Yan, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.

Freeling, M., 2001 Grasses as a single genetic system: reassessment 2001. Plant Physiol. **125:** 1191–1197.

Fu, H., and H. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. USA **99:** 9573–9578.

Gaut, B. S., and J. F. Doebley, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94:** 6809–6814.

Glazko, G. V., E. V. Koonin, I. B. Rogozine and S. A. Shabalina, 2003 A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet. **19:** 119–124.

Goff, S. A., D. Ricke, T.-H. Lan, G. Presting, R. Wang *et al.*, 2002 A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296:** 92–100.

Haldane, J. B. S., 1933 The part played by recurrent mutation in evolution. Am. Nat. **67:** 5–19.

Hardison, R. C., 2000 Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. **16:** 369–372.

Hardison, R. C., J. Oeltjen and W. Miller, 1997 Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res **7:** 959–966.

Harper, L., and M. Freeling, 1996 Interactions of *liguleless1* and *liguleless2* function during ligule induction in maize. Genetics **144:** 1871–1882.

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

Inada, C. D., A. Bashir, C. Lee, B. C. Thomas, C. Ko *et al.*, 2003 Conserved noncoding sequences in the grasses. Genome Res. **13:** 2030–2041.

Kaplinsky, N. J., D. M. Braun, J. Penterman, S. A. Goff and M. Freeling, 2002 Utility and distribution of conserved noncoding sequences in the grasses. Proc. Natl. Acad. Sci. USA **99:** 6147–6151.

Kellogg, E. A., 2001 Evolutionary history of the grasses. Plant Physiol. **125:** 1198–1205.

Kloeckener-Gruissem, B., and M. Freeling, 1995 Transposon-induced promoter scrambling: a mechanism for the evolution of new alleles. Proc. Natl. Acad. Sci. USA **92:** 1836–1840.

Lewis, E. B., 1951 Pseudoallelism and gene evolution. Cold Spring Harbor Symp. Quant. Biol. **16:** 159–174.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Loots, G. G., I. Ovcharenki, L. Pachter, I. Dubchak and E. Rubin, 2002 rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. **12:** 832–839.

Lukashin, A. V., and M. Borodovsky, 1998 GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26:** 1107–1115.

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473.

Mathews, S., R. E. Spanger, R. J. Mason-Gamer and E. A. Kellogg, 2002 Phylogeny of Andropogoneae inferred from phytochrome B, GBSSI and NDHF. Int. J. Plant Sci. **163:** 441–450.

Ohno, S., 1970 *Evolution by Gene Duplication*, Springer-Verlag, Berlin.

Paterson, A. H, J. E. Bowers, X. D. Burow, C. G Elsik, C.-X. Jiang *et al.*, 2000 Comparative genomics of plant chromosomes. Plant Cell **12:** 1523–1539.

Sentoku, M., S. Yatuka, N. Kurata, Y. Ito, H. Kitano *et al.*, 1999 Regional expression of the rice Kn1-type homeobox gene family during embryo, shoot and flower development. Plant Cell **11:** 1651–1664.

Simillion, C., K. Vandepoele, M. C. Van Montagu, M. Zabeau and Y. Van de Peer, 2002 The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **99:** 13627–13632.

True, J. R., and S. B. Carroll, 2002 Gene co-option in physiological and morphological evolution. Annu. Rev. Cell Dev. Biol. **18:** 53–90.

Vision, T. D., D. B. Brown and S. D. Tanksley, 2000 The origins of genomic duplications in *Arabidopsis*. Science **290:** 2114–2117.

Walsh, J., and M. Freeling, 1999 The *liguleless2* gene in maize functions during the transition from the vegetative to the reproductive shoot apex. Plant J. **19:** 489–495.

Walsh, J., C. A. Waters and M. Freeling 1998 The maize gene liguleless2 encodes a basic leucine zipper protein involved in the establishment of the blade-sheath boundary. Genes Dev. **12:** 208–218.

Wilson, W. A., S. E. Harrington, W. L. Woodman, M. Lee, M. E. Sorrells *et al.*, 1999 Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. Genetics **153:** 453–473.

Communicating editor: V. Sundaresan