Uneven Chromosome Contraction And Expansion In The Maize Genome

Authors

Rémy Bruggmann, Arvind K. Bharti,Heidrun Gundlach, Jinsheng Lai, Sarah Young, Ana C. Pontaroli, Fusheng Wei, Georg Haberer, Galina Fuks, Chunguang Du, Christina Raymond, **Matt C. Estep**, Renyi Liu, Jeffrey L. Bennetzen, Agnes P. Chan, Pablo D. Rabinowicz, John Quackenbush, W. Brad Barbazuk, Rod A. Wing, Bruce Birren, Chad Nusbaum, Steve Rounsley, Klaus F.X. Mayer and Joachim Messing

Abstract

Maize (Zea mays or corn), both a major food source and an important cytogenetic model, evolved from a tetraploid that arose about 4.8 million years ago (Mya). As a result, maize has extensive duplicated regions within its genome. We have sequenced the two copies of one such region, generating 7.8 Mb of sequence spanning 17.4 cM of the short arm of chromosome 1 and 6.6 Mb (25.6 cM) from the long arm of chromosome 9. Rice, which did not undergo a similar whole genome duplication event, has only one orthologous region (4.9 Mb) on the short arm of chromosome 3, and can be used as reference for the maize homoeologous regions. Alignment of the three regions allowed identification of syntenic blocks, and indicated that the maize regions have undergone differential contraction in genic and intergenic regions and expansion by the insertion of retrotransposable elements. Approximately 9% of the predicted genes in each duplicated region are completely missing in the rice genome, and almost 20% have moved to other genomic locations. Predicted genes within these regions tend to be larger in maize than in rice, primarily because of the presence of predicted genes in maize with larger introns. Interestingly, the general gene methylation patterns in the maize homoeologous regions do not appear to have changed with contraction or expansion of their chromosomes. In addition, no differences in methylation of single genes and tandemly repeated gene copies have been detected. These results, therefore, provide new insights into the diploidization of polyploid species.

Rémy Bruggmann, Arvind K. Bharti,Heidrun Gundlach, Jinsheng Lai, Sarah Young, Ana C. Pontaroli, Fusheng Wei, Georg Haberer, Galina Fuks, Chunguang Du, Christina Raymond, **Matt C. Estep**, Renyi Liu, Jeffrey L. Bennetzen, Agnes P. Chan, Pablo D. Rabinowicz, John Quackenbush, W. Brad Barbazuk, Rod A. Wing, Bruce Birren, Chad Nusbaum, Steve Rounsley, Klaus F.X. Mayer and Joachim Messing (2014) " Uneven Chromosome Contraction And Expansion In The Maize Genome" *Genome* 16: pp.1241-1251 Version of Record Available At www.genome.org

Maize, *Zea mays* L., is one of the most productive crops on earth. It is a critical source of animal feed, a staple food for many countries in Latin America and Africa, and has many industrial uses. It serves as a model species to study many basic biological processes such as recombination, transposition, meiosis, paramutation, imprinting, gene expression, and plant development. Owing to its economic and biological importance (Messing 2005), maize with its large genome of ~2400 Mb (Rayburn et al. 1993) is the next plant genome to be sequenced completely. This has been preceded by the more technically and financially feasible small genomes of *Arabidopsis* (125–155 Mb) (*Arabidopsis* Genome Initiative 2000; Hosouchi et al. 2002; Bennett et al. 2003), rice (389 Mb) (International Rice Genome Sequencing Project 2005), and the ongoing sequencing projects of *Medicago truncatula* and *Lotus japonicus* genomes, ~500 Mb each (Cannon et al. 2005; Young et al. 2005; Town 2006).

Within the cereals, conserved genetic markers can readily be found, and are largely collinear (Gale and Devos 1998). However, genome sizes vary greatly, as illustrated by the difference between the 389-Mb rice genome (International Rice Genome Sequencing Project 2005) and the 16,000-Mb hexaploid wheat genome. This discrepancy between genome size and organismal complexity has been referred to as the C-value enigma (Thomas Jr. 1971). One reason for the size variation is polyploidy, a common feature of flowering plants. Bread wheat, for example, is a hexaploid formed from three distinct genomes: A, B, and D (Sorrells et al. 2003). Ancient polyploidy events are also evident, as is the case with maize, which arose from an ancestor that underwent a whole-genome duplication (WGD), a conclusion drawn from early cytogenetic and genetic studies (McClintock 1930; Rhoades 1951). After the WGD event, chromosome breakage and fusion led to the formation of chromosomes composed of different homoeologous chromosomal segments derived from the two

parental genomes. Sequencing a sample of these homoeologous regions from different chromosomes and comparing those linked genes with orthologous regions from rice and sorghum, a species much closer to maize than rice, have permitted the alignment of genes derived from common ancestral grass chromosomes (Swigonova et al. 2004). It appears that the two progenitors of maize and the progenitor of sorghum were $n = 5$ species that diverged about 11.9 million years ago (Mya). The two maize progenitors are believed to have hybridized not later than 4.8 Mya (Swigonova et al. 2004).

Sequence analysis of collinear segments, whole-genome shotgun sequences, BAC end sequences (BES), and fully sequenced random BAC clones have also been important for our understanding of the drastic differences in genome sizes (Tikhonov et al. 1999; Meyers et al. 2001; Song et al. 2002; Messing et al. 2004; Haberer et al. 2005). Greater than two-thirds of the maize sequence obtained to date consists of nongenic repetitive DNA elements, as compared to only one-third in rice. Moreover, most of the repetitive DNA in maize is composed of LTR retrotransposons, whereas the rice genome contains a much higher proportion of DNA transposons (10.4% vs. 1.3% in maize). This difference could have a major effect on genome size, since LTR retrotransposons do not excise as part of their transposition mechanism. Instead, they use a replicative RNA-based mechanism, and thus their numbers in the genome increase with each round of transposition. Consequently, they are the major factor in the increase in genome size in maize relative to rice. Using alignments of maize LTR sequences, calculations have shown that the majority of retrotranspositions have occurred during the last 5 Myr—after the hybridization event of the two progenitors (SanMiguel et al. 1998; Swigonova et al. 2005). Therefore, in addition to genome duplication by WGD, subsequent retrotransposition has led to further expansion of the maize genome. Besides these size increases, the chromosomes have undergone breakage and fusion cycles, leading to formation of the mosaic of different homoeologous segments found today (Song et al. 2002).

Comparative analysis of orthologous regions between cereals has also allowed a higher resolution view of synteny, and its preservation or lack thereof. Although alignments of genetic markers indicate extensive synteny between grass genomes (Gale and Devos 1998), sequence analysis of orthologous chromosomal regions shows extensive gene movement during speciation. A study comparing the duplicated regions of maize with orthologous regions in rice and sorghum has shown that ~15% of all genes in any pairwise comparison are noncollinear (Lai et al. 2004b). Comparison to the nearly complete rice genome sequence shows that these genes are not simply deleted, but have moved to different genomic locations. Even within the same species, some genes may move to nonorthologous locations (Fu and Dooner 2002; Song and Messing 2003; Brunner et al. 2005). When different maize inbreds were compared with each other, it became clear that insertion of gene fragments and, perhaps, genes could create haplotype variability. Moreover, recent analysis of such haplotype variability indicated that movement of gene fragments and potentially intact genes is often caused by a new type of transposition mechanism involving a helicase function from helitron transposons (Kapitonov and Jurka 2001; Lal et al. 2003; Lai et al. 2005; Morgante et al. 2005). However, this mechanism appears to be different from one that often creates new members of a gene family. For instance, the maize zein and sorghum kafirin genes are new (relative to rice) noncollinear genes that have been differentially amplified in tandem arrays after their insertion, apparently by a non-helitron mechanism (Song et al. 2002). Interestingly, gene movement and amplification of this gene family occurred during the same time period as retrotransposition increased the genome size of maize (Song et al. 2001). The analysis of the duplicated regions of maize indicates that in nearly half of the cases, one homeologous copy of the duplicated genes was lost after tetraploidization (Ilic et al. 2003; Lai et al. 2004b; Messing et al. 2004). Thus, maize is expected to have less than double the number of genes that are present in rice, which has an estimated gene number of 37,544 (International Rice Genome Sequencing Project 2005). Based on the size of the transcriptome and the average gene number within 100 random BAC clones (Haberer et al. 2005), the number of maize genes has been estimated to range between 42,000 and 56,000, consistent with the observed frequent loss of the second copies of many duplicated genes.

All studies of orthologous regions described to date have investigated relatively small chromosomal intervals containing few maize genes, owing in large part to the resources available at the time of the studies. Here, we use the completed rice genome sequence and well-characterized maize BAC libraries to study much longer chromosomal regions. DNA fingerprinting was used to select BAC clones with a minimal overlap from highly redundant maize B73 BAC libraries (Yim et al. 2002) that were anchored with collinear markers from ~20 cM of the short arm of rice chromosome 3 (Buell et al. 2005). Taking advantage of the availability of the MTP (minimum tiling path) for both these contiguous regions and thereby not having to determine the next overlapping clone for a sequenced BAC, all of the 116 maize clones were shotgun-sequenced in parallel and assembled into contiguous sequences. These sequenced maize regions yielded many new transposable elements (TEs) and 479 non-TE-related gene models. We also could determine the overall maize chromosome architecture with respect to the distribution of repeat sequences and genes, and regarding the distribution of genomic DNA methylation by analyzing methylation-filtered (MF) sequences. Comparison of these large duplicated regions in maize showed that maize chromosomes have both expanded and contracted relative to the rice genome, demonstrating the unevenness of insertions and deletions of genes and TEs in closely related species.

Results and Discussion

Physical and genetic maps

Two homoeologous regions of maize, Zm1S and Zm9L, were selected for this study (Supplemental Fig. 1). Zm1S is located on the short arm of chromosome 1, spans the markers *bnlg1124* at position 29.0 and *bnlg1112* at position 46.4 of the BNL 2002 map, and is contained within a tiling path that contains 60 BAC clones (Supplemental Table A). Zm9L is located on the long arm of chromosome 9, spans the markers *bnlg619* at position 129.4 and *bnlg1129* at position 155.0 of the BNL 2002 map, and is contained within a tiling path that contains 56 BAC clones (Supplemental Table B). Based on these markers, the two homoeologous regions of the maize genome comprise a genetic distance of 17.4 and 25.6 cM, respectively, or a total of ~43 cM. If the same regions are compared to maps derived from recombinant inbred lines, the genetic distances expand dramatically as expected (Lee et al. 2002). Based on collinear markers, these regions are orthologous to rice chromosome 3 and sorghum chromosome 1 (Supplemental Table C).

Clones were subjected to shotgun sequencing, and sequences were assembled with ARACHNE and then were curated as described in Methods. All clones are listed in Supplemental Tables D (Zm1S) and E (Zm9L), with their sizes and accessions. The tiles yield a sequence of 7,822,695 and 6,560,930 bp, respectively, from these contiguous BACs. Both maize regions have dramatically higher recombination frequencies per kilobase compared to the average for the entire maize genome. The B73 genome has an estimated total physical length of ~2.4 Gb comprising about 2000 cM, or an average of ~1.2 Mb/cM (Rayburn et al. 1993; Gardiner et al. 2004). In contrast to a low recombination rate of only ~450 kb/cM in Zm1S, the Zm9L region has a high recombination frequency of ~256 kb/cM (Supplemental Table F), which is the same as the average recombination frequency in rice of 255 kb/cM (International Rice Genome Sequencing Project 2005). Still, the recombination frequency of the orthologous rice region, Os3S (200 kb/cM) (Supplemental Table C), is even higher than the genome-wide average for rice (highest in Os3 of 225 kb/cM).

Gene content

To enable any analysis of the collinear maize regions, their gene content must be determined. However, over-prediction of genes is a general problem in higher plants (Bennetzen et al. 2004), but particularly problematic in maize because of the large number of transposable elements and the ease with which these sequences can be mistaken for endogenous genes. For this reason, our methods and criteria for gene finding are chosen for their high stringency as described previously (Haberer et al. 2005). Using these methods, Zm1S contains 236 predicted genes, and Zm9L carries 243 predicted genes (Table 1). This amounts to gene densities of 1 gene per 33 kb and per 27 kb on Zm1S and Zm9L, respectively. This relatively high gene density, compared to 1 gene per 43.5 kb for the entire maize genome (Haberer et al. 2005), is consistent with the higher recombination frequency observed for Zm9L, thereby confirming a long-held hypothesis of genetic maps of eukaryotic species (Thuriaux 1977).

The number of exons per gene is similar between the two maize regions (5.5–6.8) and the rice regions (5.7). However, many genes in maize appear to be expanded in length compared to rice (average gene length: 4.2 kb for maize, 3.3 kb for rice). While the average exon length is slightly longer in rice than in maize, the reverse is true for the length of introns (Table 1). We also find a number of very large predicted genes in maize, and these large-intron genes are responsible for the larger average intron size compared to rice. The largest predicted maize gene on the chromosome 1 segment is 89.1 kb and on the chromosome 9 segment it is 64.7 kb, whereas the largest predicted gene in the comparable rice regions is only 30.4 kb. It is possible that some of these very large predicted maize genes are annotation artifacts, or mutant alleles, but we have found some cases in which maize B73 ESTs seamlessly cover the exons flanking a very large intron, suggesting that these large candidate genes are expressed and properly processed (Supplemental Fig. 2). Therefore, it appears that maize has expanded both intergenic regions and the genes themselves relative to rice.

Table 1. Statistics of gene models for the Zm1S, Zm9L, and Os3S regions

Total number of BACs	60	56	N/A
Minimal tile (bp)	7,822,695	6,560,930	4,900,000
Predicted genes	236	243	644
Predicted exons	1296	1642	3688
Average number of exons per gene	5.5	6.8	5.7
Average intron size (bp)	587	472	356
Median intron size (bp)	150	137	140
Average exon size (bp)	258	230	298
Median exon size (bp)	142	132	149
Average gene size (kb)	4.1	4.3	3.3
Median gene size (kb)	2.5	2.8	2.7
Average exon density/100 kb	16.6	25.0	75.3
Average gene density (kb per gene)	33	27	7.6
Maximum gene length (kb)	87.9	64.7	30.4
G + C content (%)			
Overall	47.2	47.2	44.3
Exons	56.0	52.7	53.3
Introns	44.7	42.5	39.1
Intergenic regions	45.6	43.3	41.6
Unassigned regions	45.5	44.9	43.2

Intergenic regions

The intergenic regions in all three chromosomes are composed of a mixture of low-copy noncoding sequences, also referred to as nonassigned intergenic sequences, and repetitive DNA elements. To separate these two categories, we used a customized and exhaustive repeat sequence library and a hierarchical repeat ontology (Methods). A list of the major repeat sequence families is presented in Table 2. To provide a genome-wide context for the results from these regions, we compared the results to reference values taken from the whole rice genome or from the 100 randomly chosen BAC clones that were previously sequenced from maize (Haberer et al. 2005).

For all regions, the observed repeat content is lower than the average genome values. This is not unexpected, given that the gene density is higher than the genome-wide average. Compared to the average for the rice genome (33% repetitive DNA), the studied Os3S region contains only 19% repetitive DNA, while the entire genome (63% for Zm1S and 59% for Zm9L). Conversely, the amount of coding space in Os3S (44%) differs from the overall rice average of 33%. Interestingly, among all 12 rice chromosomes (35%), chromosome 3 (29%) has the least amount of repeats and the highest average gene density, 1 gene per 8.7 kb, as opposed to the rice genome, 1 gene per 9.9 kb (International Rice Genome Sequencing Project 2005). The maize regions differ even more dramatically from the genome-wide coding density of 7.5%, with 12% for Zm1S and 15.8% for Zm9L. Interestingly, the nonassigned intergenic sequences are less variable, but also differ slightly from the genome-wide averages. For Os3S, the nonassigned intergenic sequences comprise ~37% of the space (compared to a genomic average of 34%), while both Zm1S and Zm9L have 25% nonassigned intergenic sequences, compared to a 22% genome-wide average.

One of the striking differences observed for Os3S is that the region contains a higher than average quantity of DNA transposons (by a factor of 1.44). This is largely due to the insertion of miniature inverted-repeat-transposable elements (MITES) (Wessler et al. 1995), which are known to be preferentially associated with genes (Zhang et al. 2000). Conversely, Os3S harbors

Table 2. Repeat elements genome-wide, and in the three regions of rice and maize

Classification	Os	Zm[a]	Os3S	Zm1S	Zm9L
Class I: Retroelement[b]	**61.9**	**95.2**	**47.4**	**95.2**	**95.5**
LTR retrotransposon	**59.50**	**94.40**	**45.40**	**93.55**	**94.20**
Ty1/*copia*	8.37	32.92	5.86	48.07	46.14
Ty3/*gypsy*	37.89	52.71	29.20	39.77	43.07
TRIM	0.11	0.01	0.08	0.03	0.03
Other LTR	13.12	8.75	10.27	5.68	4.95
Non-LTR retrotransposon	**1.76**	**0.39**	**1.76**	**0.92**	**1.10**
LINE	1.13	0.39	0.75	0.92	1.10
SINE	0.63	0.00	1.01	0.00	0.00
Unclassified retroelement	0.62	0.38	0.23	0.72	0.17
Class II: DNA transposon	**34.5**	**2.6**	**50.0**	**2.7**	**3.0**
DNA transposon superfamily	**17.15**	**1.54**	**20.05**	**1.31**	**1.51**
CACTA superfamily	7.98	0.89	2.97	0.43	0.49
hAT superfamily	1.11	0.16	2.35	0.16	0.26
Mutator superfamily	1.99	0.02	2.16	0.05	0.02
Tc1/Mariner superfamily	0.45	0.01	1.20	0.03	0.03
PIF/Harbinger	0.03	0.17	0.10	0.13	0.24
Other transposon family	5.59	0.29	11.27	0.50	0.27
MITE	**14.14**	**0.42**	**25.53**	**0.43**	**0.61**
Stowaway	2.56	0.05	4.30	0.06	0.10
Tourist	2.58	0.07	5.07	0.08	0.10
Other MITE	8.99	0.29	16.16	0.28	0.41
Unclassified DNA transposon	3.21	0.59	4.36	0.95	1.04
Helitron	**0.14**	**0.01**	**0.04**	**0.00**	**0.06**
Simple sequence	0.94	0.44	0.35	0.08	0.13
High copy number gene	0.15	0.02	0.10	0.02	0.01
Other	2.40	1.80	2.17	2.03	1.37
Total as percent of genome					
Repetitive DNA	33	70	19	63	59
Coding space	33	8	44	12	16
Unassigned space	34	22	37	25	25

[a]100 random BACs.
[b]Percent of total repetitive DNA.
Summary data for each category is in bold.

a lower than average level of retrotransposable elements by a factor of 0.76.

The distribution of repeat elements correlates well with the higher than average gene density in the maize chromosome regions studied. These features are preserved even in the segment that underwent the largest expansion (Zm1S). In this respect, it is interesting to note that these regions are located at the distal ends of their respective chromosomes (Fig. 1). It has previously been proposed that regions close to the telomeric regions of the wheat genome have increased gene density (Akhunov et al. 2003). Furthermore, analysis of maize pachytene chromosomes suggests that gene density in euchromatin is fourfold higher than in heterochromatin and that expressed sequence tag (EST) markers mostly map to distal euchromatin (Anderson et al. 2006). As in all other complex eukaryotic genomes studied, genome-wide analyses also show a strong correlation between gene density and recombination rates in maize (Anderson et al. 2006).

Alignment of syntenic regions

When all three regions are aligned based on conserved sequences, the Zm1S region becomes the limiting factor despite being the longest sequence of ~7.8 Mb (Fig. 1). This region has experienced the highest degree of relative expansion, which can be expressed as the size of the maize region compared to the rice reference region. Based on distinct junctions of nonhomoeologous sequences separating sequence segments of collinear genes, one can divide the aligned sequences into segments of homoeologous sequences. Individual segments of Zm1S have expansion factors between 1.9 for segment A and 4.2 for segments B1 and B2 (Fig. 1). The expansion rates of Zm9L vary from 0.8 (segments B1 and B2) to almost no expansion (segments A and C) to 1.8 (segment D). Consequently, degrees of expansion between Zm1S and Zm9L differ by as much as a factor of 5.25 (segments B1 and B2). If, after the WGD event, maize had retained all of its duplicated regions, the relative sizes of the present-day maize and rice genomes would suggest a genome-wide threefold expansion rate for all homoeologous regions in maize.

While segments B1 and B2 of Zm1S expanded beyond the average, all others are far below the expected expansion rates. It is not clear whether this was caused by a lower-than-average frequency of transposon amplification in all regions but B1 or B2, or by a higher frequency of the small deletions (most caused by illegitimate recombination) that have been found to be responsible for genome shrinkage in rice and *Arabidopsis* (Devos et al. 2002; Ma and Bennetzen 2004; Ma et al. 2004), or by a combination of the two.

Another marked feature of Zm1S is an insertion of ~600 kb of nonconserved DNA between segments A and B1. Smaller nonconserved junctions are also found at the same orthologous position in rice and the other maize chromosome. Such junctions can also be found at position 4.1 Mb on Zm1S, at 1.7 Mb on Os3S, and at 3.5 Mb on Zm9L. On Zm9L, a region of 1.8 Mb has undergone an inversion relative to rice and the other maize chromosome. Sequences on Zm1S covering the inversion of Zm9L may extend beyond the sequenced region that would encompass segment C. From this comparison, it appears that all three chromosomes have junctions in the same position that interrupt synteny blocks. In one case the junction appears to have grown relatively large, in another it is the site of an inversion. Such conserved junctions have been observed for another orthologous region of sorghum, maize, and two rice subspecies and hypothesized as potential hotspots for chromosome changes (Song et al. 2002). One possibility is that sites for chromosome breakage and fusion have been conserved through speciation. Such hotspots are reminiscent of fragile sites in human chromosomes that can undergo chromosome breakage and recombination, a process that is tightly regulated to avoid deleterious mutations (Schwartz et al. 2005).

Contraction and expansion of synteny blocks

To study the underlying causes for the unequal degrees of local expansion in more detail, we analyzed the contribution of genic, repetitive, and the nonassigned (nongenic and nonrepetitive) sequences within the individual segments. Interestingly, the ap-

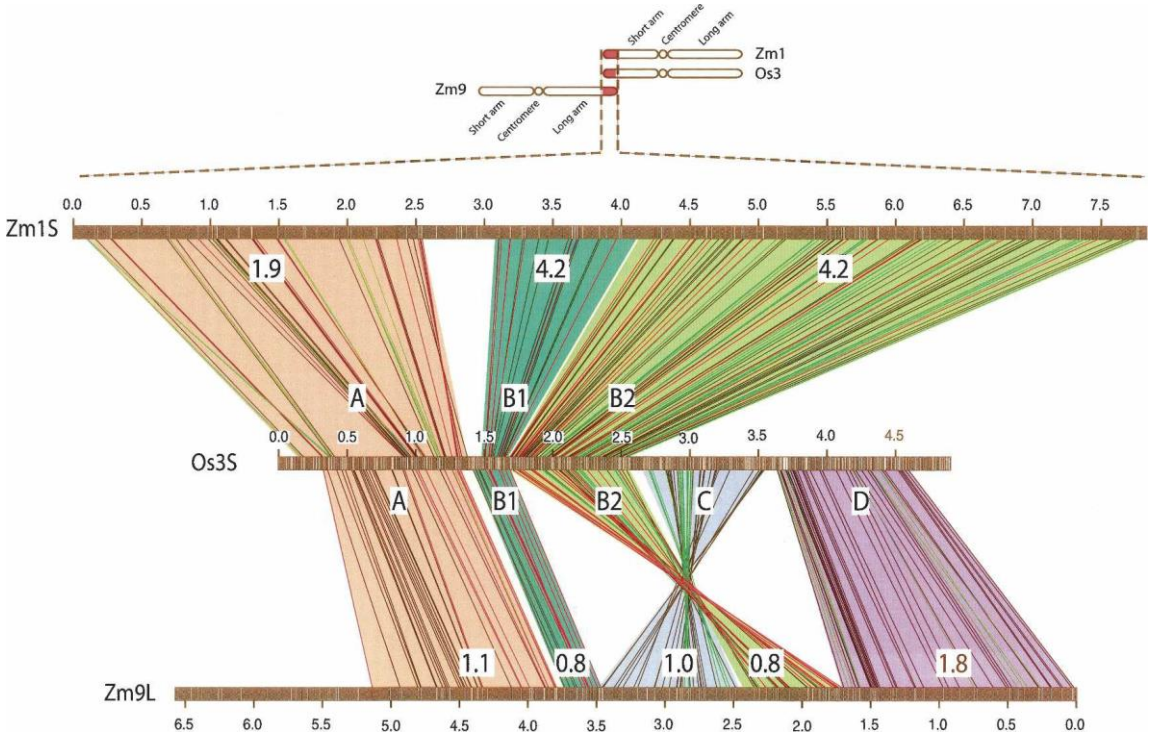

Figure 1. Alignments of chromosomal regions from Zm1S, Os3S, and Zm9L. The three horizontal lines with scales in million basepairs (Mb) represent the chromosomal regions from Zm1S, Os3S, and Zm9L. On these horizontal lines, the positions of the genes are marked with a white vertical bar. *Above* those lines, a schematic diagram shows the alignment of these regions with respect to their chromosomal locations. Vertical lines connect syntenic genes between Zm1S and Os3S, and Zm9L and Os3S, respectively. Line color indicates syntenic arrangement of genes between Zm1S–Os3S–Zm9L (red), syntenic arrangement of tandem genes (green), and syntenic arrangements of genes either between Zm1S–Os3S or between Zm9L–Os3S (black). Syntenic blocks are labeled A, B1, B2, C, and D. Syntenic blocks are interrupted by regions without corresponding genes. The expansion grade of each syntenic block with respect to rice is shown as fold expansion.

parent lack of expansion on Zm9L is largely due to a reduction of genic space by a factor of 0.4–0.5 for segments A, B1, B2, and C and 0.7 for segment D (Table 3). The same is true for the nonassigned genomic space containing intergenic and regulatory regions, which are also reduced in almost all segments, although to a more moderate degree except in segment D.

A different situation occurs for the Zm1S region. Although segment A experienced an overall expansion, the genic space is considerably reduced by a factor of 0.4. In addition, the nonassigned space has a moderate expansion factor of 1.3. For the drastically expanded segments B1 and B2, we observed relative expansion factors of 1.2 and 1.4 for the genic space, and 2.3 and 2.7 for the unassigned space (Table 3). Interestingly the expansion factor of genes between random maize BACs versus random rice regions is 1.4 (Table 3). Based on 37,544 gene models in rice, the number of gene models in maize would be 37,544 × 1.4 = 52,562, which falls within the range proposed earlier (Haberer et al. 2005).

Further analysis of repeat elements in different segments illustrates the uneven target site preference of retrotransposition (Table 3). The Zm1S and Zm9L regions have very different distributions of repeat elements. Segments of Zm9L are quite variable and range between 34.6% and 61.2% in repeat content (overall, 58.8%). In comparison, Zm1S shows less variability and has a local repeat concentration that ranges between 62% and 65.5% (overall, 63.1%). The B1 segments of both maize regions have the largest increase in retroelement content relative to rice, with Ty1/*copia*-like elements increasing substantially more than the Ty3/*gypsy*-like elements. The increase in Zm1S is nearly a magnitude higher than in Zm9L. B1 is marked on one side by a rather large junction of nonhomoeologous sequences that occurs in both Zm1S and Zm9L. On the other side, an inversion event is specific to Zm9L, and perhaps is involved in the halted expansion of Zm9L relative to Zm1S. Once the entire maize genome is sequenced, it will be possible to assess whether there is a general correlation between local genome expansion and the presence of inversions or other chromosomal rearrangements.

Analysis of these two large regions does not reveal evidence of large gene islands separated by retrotransposon blocks. As previously reported, most gene islands are small (one to two genes) (Bennetzen et al. 2005) and vary between the different homoeologous regions. A picture is emerging in which different chromosomal regions evolve into a mosaic of syntenic blocks with differential expansion caused by the contraction of genic and intergenic space in combination with the addition of different combinations of repeat elements.

Gene loss after WGD to form maize

Studying the sequence of these large contiguous regions provides an opportunity to study microcollinearity within the syntenic blocks at a much higher resolution than possible with genetic or physical maps. The surprising result is the extent of noncollinearity found between rice and maize, and between the two maize chromosomes (Fig. 2A). Out of 644 predicted rice genes (see Methods), only 270 (42%) are collinear with one of the maize

Table 3. Size and expansion factors for the different syntenic blocks

Syntenic block	A	B1	B2	C	D
Size in megabases					
Zm1S	2.34	1.13	3.50		
Os3S	1.22	0.27	0.84	1.01	0.96
Zm9L	1.35	0.23	0.65	0.98	1.68
Expansion factor for Zm1S versus Os3S with respect to:					
Size in megabases	1.9	4.2	4.2		
Genic space	0.4	1.4			
Unassigned sequences	1.3	2.3	2.7		
Repeat elements	6.5	28.0	13.0		
LTR retrotransposons	11	521	22		
Ty3/*gypsy*-like elements	9	380	14		
Ty1/*copia*-like elements	22	833	142		
Expansion factor for Zm9L versus Os3S with respect to:					
Size in megabases	1.1	0.8	0.8	1.0	1.8
Genic space	0.4	0.5	0.4	0.4	0.7
Unassigned sequences	0.7	0.7	0.7	0.7	1.1
Repeat elements	3.5	3.1	1.6	2.7	7.1
LTR retrotransponsons	6	55	2.8	5	43
Ty3/*gypsy*-like elements	5	54	1.8	4	33
Ty1/*copia*-like elements	11	79	18	27	126
Expansion factor for 100 random maize BACs versus random rice regions with respect to:					
Size in megabases	6.1				
Genic space	1.4				
Unassigned sequences	4.1				
Repeat elements	13.0				
LTR retrotransponsons	20.0				
Ty3/*gypsy*-like elements	18.0				
Ty1/*copia*-like elements	51.0				

chromosomes. The number increases only slightly to 306 (48%) when less stringent microsynteny criteria are used. Because a large number of BAC clones are still at phase 1 (unordered contigs due to physical gaps) stage and not finished (Supplemental Tables D and E), the order of contigs within BAC clones is not always resolved (see also Methods). This relaxed stringency would also allow for small-scale rearrangements within syntenic blocks.

When comparing both maize regions with rice, the number of genes collinear in all three regions is <50%. This is not unexpected because, as has been noted previously, one copy of a duplicated gene has been frequently lost after the WGD event. Of the 270 genes in rice that are present in either Zm1S or Zm9L, only 298 genes (55%) of the possible 540 (one copy in each maize region) were detected. This number increases to 346 (64%) with the relaxed stringency described above. This is a lower percentage than previously observed for the analysis of five different duplicated regions of the maize genome (Lai et al. 2004b). Within those regions, 17 out of 24 collinear genes were detected. However, collinearity was more narrowly defined in the earlier study because sorghum could be used as an independent reference for determining orthologous genes in rice. Unfortunately, the sorghum chromosome 1 region orthologous to the regions analyzed here is not yet available as a minimum tiling path.

An even greater loss of the second gene copies of the genes in the duplicated genome (90%) occurred after the hybridization of the two fungal species that formed baker's yeast (Kellis et al. 2004). These studies illustrate that comparative analysis of diploid and polyploid examples of closely related species provide important clues about the diploidization process that commonly follows polyploidization (Wolfe 2001).

Significant gene movement in rice and maize

A large number of the genes located in all three chromosomal regions are noncollinear, 64 for Zm1S (27%), 69 for Zm9L (28%), and 338 for Os3S (52%). However, the presence of a gene on one chromosome and its absence on the other do not mean that such a gene was lost. For genes present in the maize regions and absent in rice Os3S, a simple test against the near-complete rice genome sequence can reveal whether the gene is present elsewhere in the rice genome. The extra maize genes were compared against all genes from rice (TIGR 3 release and IRGSP; BLASTP $E < 1e\text{-}20$). For 43 noncollinear genes (out of 64; 67%) from Zm1S and 47 noncollinear genes (out of 69; 68%) from Zm9L, homologous genes have been detected elsewhere in the rice genome. Still, 21 and 22 genes from Zm1S and Zm9L, respectively, had no homologs in rice (Fig. 2A). Additional comparisons of these remaining genes against EST and tentative consensus collections (TCs) from other plant species revealed an additional six and three genes, respectively, that have a homologous counterpart elsewhere, adding further evidence for these being real genes in maize.

Conversely, we addressed the same question for the 338 genes present in rice Os3S, but lacking orthologous counterparts on both Zm1S and Zm9L. In the absence of the entire maize genome sequence, we took advantage of the complete collection of maize genomic survey sequences (GSS containing high C_0t [HC], MF, BES, and whole-genome shotgun sequences) and expressed sequences (EST and TCs) and asked whether the extra Os3S genes are homologous to any of the maize-derived GSSs or ESTs/TCs. Out of the 338 extra rice genes, 220 (65%) were assigned to maize GSSs or ESTs/TCs. Because the current maize GSSs and ESTs/TCs do not cover the complete gene set of maize, this can be considered an underestimate of the number of rice genes with a homolog elsewhere in the maize genome. In fact, for 51 out of the 118 rice genes without identified maize counterparts, homologs could be detected in ESTs/TCs collections from other plants. Nevertheless, our results indicate that maize and rice contain additional (or different) genes without homologous counterparts in the other species at levels of 9%–10% (Zm1S + Zm9L vs. rice) and 18% (Os3S vs. maize). In support of this finding, a recent comparison of maize unigenes or cDNAs with the rice genome indicates that 22% of the maize genes are not present in rice (Lai et al. 2004a). Their presence in other grass

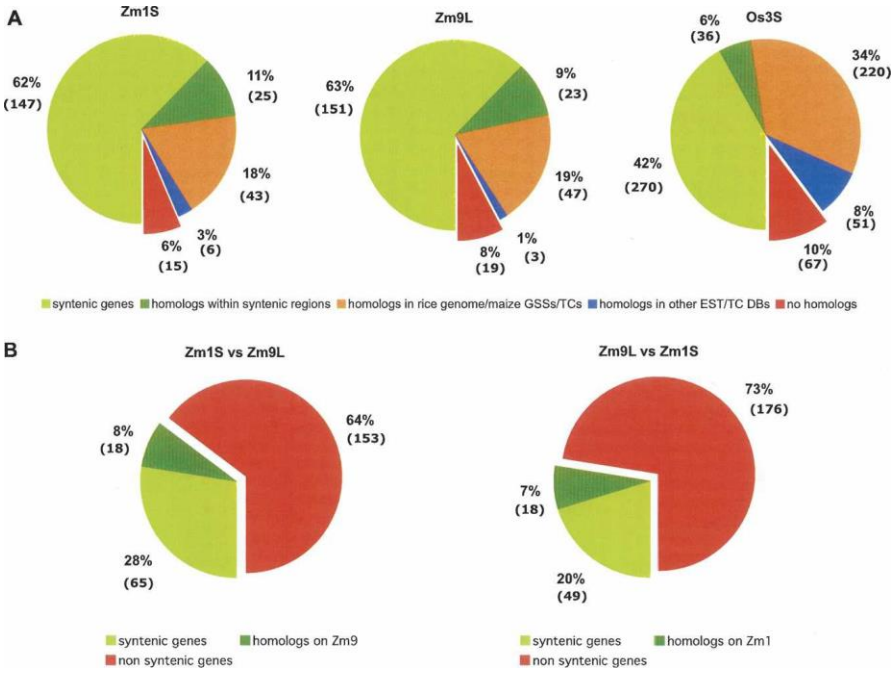

Figure 2. Corresponding, syntenic genes between the Zm1S, Os3S, and Zm9L chromosomal regions. Gene models for all three regions were determined as described in the text. A graphic representation is given for genes (*A*) that are syntenic, have homologs within the syntenic regions, have homologs in the rice genome or maize GSSs/EST collections, have a homologous counterpart in other plant EST collections, or are species-specific; and (*B*) that are syntenic between both maize regions and have a homolog or no homolog on the other maize chromosome. Each fraction is labeled with the number of genes and the percentage of the total within its sample.

species suggests that speciation has resulted in substantial gene loss or death.

Duplicated genes in maize

If maize experienced a high degree of gene loss after the WGD event, how many of the genes remain as duplicated copies? We analyzed both maize regions by bidirectional best BLASTP hits (BBHs) of their derived protein sequences (Fig. 2B). The number of duplicated genes with both copies retained is remarkably low, with 28% and 20% for Zm1S and Zm9L, respectively. Applying less stringent synteny criteria (e.g., that collinearity does not need to be preserved, but only a BBH needs to be present), the percentage increases to 35% and 28%, respectively.

In summary, 150 genes of both maize regions are related to each other, while 153 and 176 genes from Zm1S and Zm9L, respectively, represent either a single copy of a duplicated gene or a noncollinear gene as described above. Even with the inclusion of genes conserved on both maize chromosomes, an estimate of the number of genes lost and gained relative to rice is uncertain. Recently an additional comparison against a syntenic region from sorghum has shown that both copies of a gene derived from an ancestral chromosome can be lost relative to a copy conserved in its position in sorghum and rice (Lai et al. 2004b). Thus, it is clear that having the additional data point from the sorghum syntenic region would be instrumental in reconstructing the evolutionary history of duplication, retention, and syntenic gene order. Still, the differences between collinear and noncollinear genes are striking and indicate that both gene death and gene mobility have been frequent in the maize lineage post-polyploidization.

DNA methylation of the maize regions

An important component in the analysis of maize genome organization is the distribution of methylated nucleotides. DNA methylation and de-methylation have been correlated with tissue-specific activation of gene expression and open and closed chromatin formations (Spena et al. 1983; Bennetzen et al. 1994; Lund et al. 1995a,b). Methylation has also been shown to silence the activity of transposable elements (Chomet et al. 1987).

To gain insights into the distribution of methylation within the large contiguous sequenced regions of the maize genome, we mapped genomic survey sequences (GSSs) derived from a methylation-filtered (MF) library (Rabinowicz et al. 1999; Palmer et al. 2003; Whitelaw et al. 2003) to the Zm1S and Zm9L regions. The DNA fragments are 1–2 kb in size and highly enriched for sequences that have nonmethylated bases in the maize nuclear genome. As the maize genome contains large quantities of highly conserved sequences (e.g., transposable elements), widely distributed across the genome, we applied stringent criteria to anchor MF-GSS clones to their specific maize region of origin. For nonrepetitive regions, this strategy allows for high specificity to identify regions within the Zm1S and Zm9L regions that are represented among MF sequences. It should be noted that the MF libraries as well as the BAC clones were derived from the same maize inbred line, B73 (Palmer et al. 2003; Whitelaw et al. 2003), and the sequencing error rate of the GSS has been reported to be 2.3×10^{-3} or lower (Fu et al. 2004). In contrast, MF sequence clones containing conserved repetitive sequence can often not be unambiguously anchored to a specific region, as many highly similar copies may be present throughout the genome. In addition to MF GSSs, we mapped another collection of sequences derived from high C_0t (HC) enrichment, which has been demonstrated to be enriched for low-copy-number sequences (Whitelaw et al. 2003; Yuan et al. 2003).

Figure 3 gives a graphical overview of the density distribution for MF and HC clones in the two maize Zm1S and Zm9L regions. Both filtered libraries cover a similar percentage of the genic (CF: 32%; MF: 30%) as well as of the complete genomic space (CF: 12.4%; MF: 16.6%). The cumulative coverage of CF and MF for the genic space was 49%, indicating overlapping and complementary specificities for the two methods. About 91.5% of all genes were detected by at least one GSS tag. A previous comparison of CF and MF with a set of 78 full-length cDNAs also showed that 95% of the cDNAs are tagged by at least one sequence read (Springer et al. 2004). However, total nucleotide coverage of the full-length cDNAs approached 75%, although this obviously included only exons. In a separate study, 152 predicted genes from early-published maize BAC sequences were analyzed and exhibited a similarly high level of coverage, across both exons and introns (W.B. Barbazuk, unpubl.). Here, we find a much

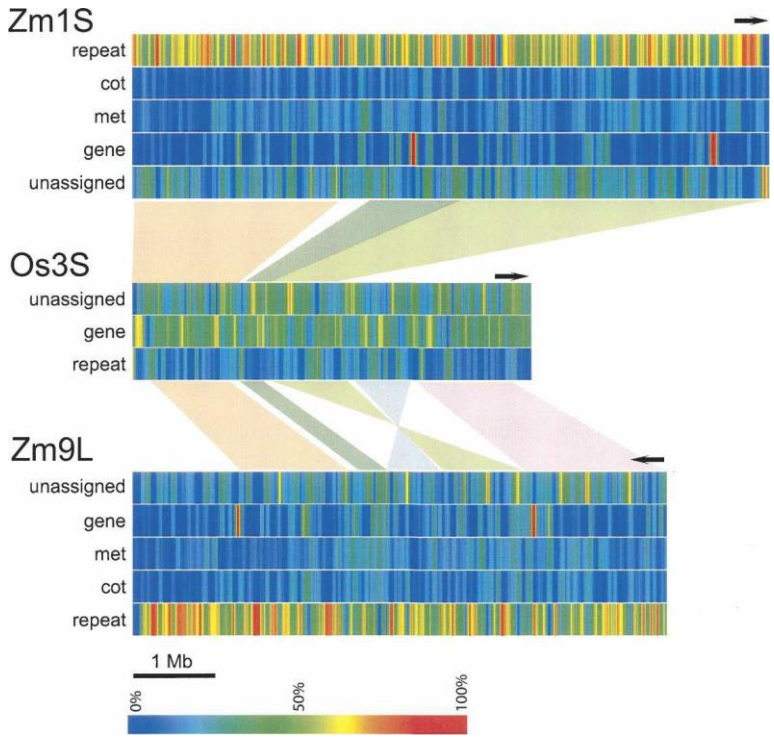

Figure 3. Position and density of sequence features of the maize regions Zm1S and Zm9L, and rice chromosomal region Os3S. Color-coded densities of several genomic features (repeats, coverage by filtrated GSSs and genes) are shown for Zm1S, Zm9L, and Os3S. Blue represents lowest (0%), green medium (50%), and red highest (100%) density of the respective feature. Relative density has been determined within a sliding window of 50 kb and in steps of 1000 bp. Locations of GSSs represent alignment positions by BLASTN. Note that the four high-density regions within the maize gene bars indicate four very large genes (>50 kb gene size) that were predicted by our annotation.

lower coverage across introns, which contributes to the lower overall coverage mentioned above. Interestingly, the coverage of introns (expressed as the percentage of sequence length) is lowest in the genes with the largest introns (Supplemental Table G). The difference between the data presented here and previous reports is likely explained by the varying abundance of genes with large introns present in each data set. For the 152-gene set, introns had an overall mean of 410 bp and a median of just 141 bp, and the largest intron was just over 7 kb. In contrast, the values for the Zm1S and Zm9L regions presented here were 587 bp and 472 bp (mean), 150 bp and 137 bp (median), and 23 kb and 62 kb (maximum intron size). This difference is perhaps not surprising since the first maize BAC sequences were not chosen at random, but rather by probing with known genetic markers, and they were consequently biased in terms of gene density. For instance, a comparison of 117 nonrandom published BAC sequences with 100 randomly selected BAC sequences demonstrated that the nonrandom BACs had a lower repeat content than the random BACs (52% vs. 66%) (Haberer et al. 2005). Moreover, the 100 random regions exhibited a combined coverage of CF and MF for all predicted genes of 51% (Haberer et al. 2005), which is close to the findings observed here. Therefore, the difference between the three data sets, the 78 full-length cDNAs, the 152 predicted genes from earlier sequenced BACs, and the 644 predicted genes from this study is that coverage of genes by CF and MF is reduced in genes with larger introns. It also has been shown that untranslated transcribed regions (UTRs) and promoter regions tend to have a lower coverage by a GSS tag (Haberer et al. 2005). Therefore, larger introns and the regulatory flanking regions appear to have a higher frequency of methylated bases and short repeat sequences than the exons of genes. For instance, regulatory elements can be separated from the coding regions by methylated retroelements in the *B1* and *Tb1* loci (Stam et al. 2002; Clark et al. 2004), and the large intron of the *P1* gene is methylated both in the normal allele and its epiallele *P-pr* throughout development (Das and Messing 1994). Still, the high percentage of tagged genes, with the HC and MF libraries each comprising only 450 Mb of single sequence reads compared to a total genome size of 2.4 Gb, underpins the value of both sequence collections for a rapid exploration of the genic space in maize when their physical map position is not required.

To analyze the potential epigenetic influences on the regulation of duplicated genes, we were particularly interested in similarities and differences of the methylation patterns for both tandemly repeated genes and conserved syntenic genes between the Zm1S and Zm9L regions. Coverage of tandemly repeated genes by MF clones was quite similar to the average coverage of all 479 genes. In particular, 78% of all genes and 75% of tandem genes had at least one matching MF clone.

Genes conserved in both syntenic maize segments revealed a significantly higher coverage by MF clones (0.42x of the genic space vs. 0.30x for all genes), suggesting a higher degree of hypomethylation for this class of genes. Interestingly, the fraction of these conserved genes with no MF coverage (3.8%; 3 out of 78) is significantly lower compared to the expectation for all genes [$P < 0.0001$; $P(X :\leq 3) = B_{n,p}(78,\ 3,\ X :\leq 3)$]. However, given the small sample size and the incomplete sequence space currently covered by MF clones, it is not yet clear whether these observations indicate a functional characteristic of conserved syntenic genes.

Interestingly, the expansion that occurred in different intervals of Zm1S and Zm9L relative to Os3S did not create a differential methylation pattern. Hypomethylation seems to be present even in regions where an increased expansion by retrotransposition occurred. Future analysis of other duplicated regions of the maize genome will help to refine the correlation of methylation and chromosome architecture.

Conclusions

We have obtained the first large contiguous sequences of maize, spanning >43 cM of genetic distance, by an economical sequencing strategy. The sequence data enabled an analysis of the changes that occurred after the progenitor of maize was formed by hybridization of two closely related species. By using the known genome sequence of rice as a reference, important insights into genome evolution have been obtained. An unforeseen

result is the enormous variability of chromosomal expansion in different syntenic blocks. Our data suggest that the ancestral chromosomes contained seeding sites for inversions and insertions within segmental duplicated regions, including sites for differential growth by insertions and deletions. It appears that the C-value enigma results from a composite of chromosome contraction and expansion. In addition, the death and birth of new genes is pronounced. Approximately a tenth of the predicted maize genes lack a homologous counterpart in rice, a phenomenon unseen in closely related mammalian genomes (International Human Genome Sequencing Consortium 2004). Furthermore, the mobility of genes is quite high. Nearly a third of the predicted genes in both species seem to have moved to new positions. Upon hybridization of its two progenitors, maize has lost one copy of more than half of its duplicated genes, indicating that an increase in gene numbers by diversification of protein functional characteristics and/or regulatory properties is limited. The genic space in maize has, in part, also enlarged because of an increase in intron sizes.

Methods

DNA sequencing, sequence assembly, and data deposition

The tiling path for both maize chromosomes was selected at the Arizona Genomics Institute. All BACs were validated by cross-checking their HindIII profiles with pre-existing agarose fingerprints. Overlap between clones is quite variable because they were not yet optimized with high-resolution fingerprinting (Nelson et al. 2005). The average overlap amounts to 15% for the tile of chromosome 1 and 21% for chromosome 9 (Supplemental Table F), which would extrapolate to ~17,200 clones for all FPCs, currently ~90% of the B73 genome (http://www.genome.arizona.edu/fpc/maize/).

In total, 116 BAC clones were chosen for sequencing. Sixty clones for maize chromosome 1 were sequenced at The Institute of Genomic Research, while Zm9L was sequenced with 40 clones at the Broad Institute, and 16 at the Plant Genome Initiative at Rutgers (Supplemental Tables D and E). Randomly sheared BAC DNA in a very narrow size range of 4 kb was used to construct shotgun libraries in the pOTWI vector for Zm9L, while the pCR4TOPO vector (Invitrogen) and 6–8-kb inserts were used for Zm1S BACs. Inserts were sequenced from both ends using universal primers (Vieira and Messing 1982), ABI 3730 capillary sequencers, and the ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction kit (Applied BioSystems). Sequence assembly was carried out using ARACHNE (Batzoglou et al. 2002; Jaffe et al. 2003). This resulted in 50 phase 1 and 10 phase 2 BACs (ordered contigs with a few sequencing gaps) from Zm1S and 28 phase 1 and 28 phase 2 BACs from Zm9L (Supplemental Tables D and E). Overlaps between neighboring BACs were determined using BLASTN, while resultant pseudomolecules were constructed after careful inspection and verification of each overlap both manually as well as by TPF processor (http://www.ncbi.nlm.nih.gov). The Zm1S pseudomolecule has 15 physical gaps, while the Zm9L pseudomolecule contains 10 physical gaps.

In order to confirm that the entire pseudomolecule mapped to the same location on the physical and genetic maps, we mapped several new SSRs (Supplemental Table H). Four BACs, distributed across the 56 BAC Zm9L megacontig, were chosen for detection of SSRs (Castelo et al. 2002). A total of 165 SSRs were found in the 603 kb of insert DNA in these four BACs. The SSR frequency in these four BACs ranges from 19 to 32 SSRs per 100 kb with an average of 27 SSRs per 100 kb (Supplemental Table I).

In order to detect polymorphism, primer pairs were designed from five SSR-flanking regions per BAC. For the genetic mapping, two RIL (recombinant inbred line) mapping populations from Ben Burr (Brookhaven National Laboratory, NY) were used (Burr et al. 1988), to increase the chance of finding useful polymorphism. First, RIL mapping population 2 (CO159 × Tx303) was used to detect polymorphism through PCR amplification. Out of the 20 primer pairs, PCR amplification was observed in 14 of them (two of them had no product in one of the parents). Therefore, only 12 were taken into consideration, and polymorphism was observed in only three of the 12 (polymorphism frequency of 25%) that involved only two out of the four BACs. Therefore RIL mapping population 1 (T232 × CM37) was used to detect polymorphism in the remaining two BACs. The success rate for PCR amplification was exactly the same (60%) as earlier. However, the degree of detected polymorphism was observed to decrease slightly from 25% to 17% (Supplemental Table B), involving only one of the two BACs. Hence, of the four BACs, only three could be mapped, and, as expected, all three BACs (ZMMBBc0051H21, c0536M23, and c0320B12), as well as the test marker (umc1505), were mapped to chromosome 9 bin 07.

Sequences of genetic markers from maize, rice, and sorghum were downloaded from the community Web sites, and used to determine their exact positions in the maize and rice sequences at the PGIR. After extensive curation and editing, files of the pseudomolecules were sent to the Munich Information Center for Protein Sequences (MIPS) for analysis.

Sequence analysis

For sequence annotation, we used the same data processing pipeline described previously (Haberer et al. 2005). In short, repeat elements were detected by openRepeatMasker 3.1 using a customized library (http://www.repeatmasker.org/) for plant repeat elements and cutoff values of 100 for hit length and 255 for hit score. The underlying library for plant repeat elements was compiled from different publicly available resources and classified by a hierarchical repeat ontology. The clustered nonredundant set comprised 5294 sequences with a total size of 12.6 Mb. Repeat masked sequences were analyzed for their coding potential by applying extrinsic (homology-based) and intrinsic (ab initio gene prediction methods) criteria and methods. Genes were detected by applying FGeneSH++ (Salamov and Solovyev 2000) and GenemarkHMM (Lukashin and Borodovsky 1998) using monocot and maize matrices, respectively. In addition, BLAST homology searches of the respective BAC sequences against EST assemblies (BLASTN) and protein sequences (SWISS-PROT and UniProt; BLASTP) were carried out (Altschul et al. 1990). EST collections included assemblies of *Arabidopsis thaliana*, *Medicago truncatula*, *Triticum aestivum*, *Sorghum bicolor*, *Hordeum vulgare*, *Saccharum officinalis*, *Oryza sativa*, and *Zea mays* (The TIGR Gene Index Database at http://www.tigr.org/tdb/tgi). We used the GenomeThreader program for spliced alignments of the EST and TC sequences (Gremme et al. 2005). GenomeThreader has been especially designed for gene structure prediction in organisms containing long introns and uses a spliced alignment strategy. All gene models underwent manual curation and adaptation to supporting experimental evidence (e.g., EST, TC, and protein homologies), if necessary. The annotations including supporting evidence for each BAC can be accessed online or downloaded in the Apollo-compatible GameXML format from the MIPS maize database (http://mips.gsf.de/proj/plant/jsf/maize/index.jsp). In addition, the sequences of the pseudomolecules Zm1S (7.82 Mb) and Zm9L (6.56 Mb) are available in FASTA format on the same Web page in the download area.

The gene annotation of the syntenic part of rice was based on the TIGR rice assembly (version 3). Sequences were masked for repetitive elements applying identical methods as used for both maize regions. For gene prediction, TIGR (v3) gene models as well as complementary monocot ESTs/TCs collections were mapped to the genomic sequences. Potential transposable elements were identified by BLASTN comparisons against the MIPS plant repeat library (for cutoff parameters, see above). Identified repetitive elements and transposable elements within rice were excluded from the analysis of syntenic relationships.

In order to determine the syntenic relation between predicted genes/proteins, the bidirectional best BLASTP hits (BBHs; *e*-value < 1e-20) of the proteins of each chromosomal region was determined. The syntenic pairs were determined manually once with rice as central reference, where genes from both maize chromosomes were compared to rice, and between the two maize chromosomes. The micro-collinear gene clusters were merged into macro-collinear blocks.

The criterion for homologous genes considered as tandem duplicated genes was a BLASTP *e*-value < 1e-20 and a maximal distance of one nonhomologous gene inserted between the homologous genes.

Coverage of filtered clones

The MF and HC sequence reads available at TIGR (http://www.tigr.org/tdb/tgi/maize/) were used to determine the coverage of genes by the filtered sequence reads. All filtered sequence reads were compared against the two pseudomolecules by BLASTN sequence comparison. To anchor a clone to a genomic location, an alignment length of at least 90% of the clone length and a minimal sequence identity of 98% over the alignment length were required. Genomic/exonic/intronic coverage was determined on a nucleotide basis and was normalized to the length of the respective segment.

Acknowledgments

This work was supported by the National Science Foundation Plant Genome grants 0211851 (PI: J.M., The Plant Genome Initiative at Rutgers) and 0221536 (PI: K. Schubert, Donald Danforth Plant Science Center). Work at MIPS was, in part, supported by the GABI program of the German Ministry for Education and Research (BMBF).

References

Akhunov, E.D., Goodyear, A.W., Geng, S., Qi, L.L., Echalier, B., Gill, B.S., Miftahudin, Gustafson, J.P., Lazo, G., Chao, S., et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res.* **13:** 753–763.

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

Anderson, L.K., Lai, A., Stack, S.M., Rizzon, C., and Gaut, B.S. 2006. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome Res.* **16:** 115–122.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12:** 177–189.

Bennett, M.D., Leitch, I.J., Price, H.J., and Johnston, J.S. 2003. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately 25% larger than the *Arabidopsis* Genome Initiative estimate of approximately 125 Mb. *Ann. Bot. (Lond.)* **91:** 547–557.

Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E., and SanMiguel, P. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. *Genome* **37:** 565–576.

Bennetzen, J.L., Coleman, C., Liu, R., Ma, J., and Ramakrishna, W. 2004. Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7:** 732–736.

Bennetzen, J.L., Ma, J., and Devos, K.M. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond.)* **95:** 127–132.

Brunner, S., Fengler, K., Morgante, M., Tingey, S., and Rafalski, A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17:** 343–360.

Buell, C.R., Yuan, Q., Ouyang, S., Liu, J., Zhu, W., Wang, A., Maiti, R., Haas, B., Wortman, J., Pertea, M., et al. 2005. Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. *Genome Res.* **15:** 1284–1291.

Burr, B., Burr, F.A., Thompson, K.H., Albertson, M.C., and Stuber, C.W. 1988. Gene mapping with recombinant inbreds in maize. *Genetics* **118:** 519–526.

Cannon, S.B., Crow, J.A., Heuer, M.L., Wang, X., Cannon, E.K., Dwan, C., Lamblin, A.F., Vasdewani, J., Mudge, J., Cook, A., et al. 2005. Databases and information integration for the *Medicago truncatula* genome and transcriptome. *Plant Physiol.* **138:** 38–46.

Castelo, A.T., Martins, W., and Gao, G.R. 2002. TROLL—Tandem Repeat Occurrence Locator. *Bioinformatics* **18:** 634–636.

Chomet, P.S., Wessler, S., and Dellaporta, S.L. 1987. Inactivation of the maize transposable element Activator (Ac) is associated with its DNA modification. *EMBO J.* **6:** 295–302.

Clark, R.M., Linton, E., Messing, J., and Doebley, J.F. 2004. Pattern of diversity in the genomic region near the maize domestication gene *tb1*. *Proc. Natl. Acad. Sci.* **101:** 700–707.

Das, O.P. and Messing, J. 1994. Variegated phenotype and developmental methylation changes of a maize allele originating from epimutation. *Genetics* **136:** 1121–1141.

Devos, K.M., Brown, J.K., and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12:** 1075–1079.

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

Fu, Y., Hsia, A.P., Guo, L., and Schnable, P.S. 2004. Types and frequencies of sequencing errors in methyl-filtered and high c_0t maize genome survey sequences. *Plant Physiol.* **135:** 2040–2045.

Gale, M.D. and Devos, K.M. 1998. Comparative genetics in the grasses. *Proc. Natl. Acad. Sci.* **95:** 1971–1974.

Gardiner, J., Schroeder, S., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Morgante, M., Landewe, T., Fengler, K., Useche, F., Hanafey, M., et al. 2004. Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiol.* **134:** 1317–1326.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47:** 965–978.

Haberer, G., Young, S., Bharti, A.K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R.A., Rounsley, S., Birren, B., et al. 2005. Structure and architecture of the maize genome. *Plant Physiol.* **139:** 1612–1624.

Hosouchi, T., Kumekawa, N., Tsuruoka, H., and Kotani, H. 2002. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. *DNA Res.* **9:** 117–121.

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

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

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

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13:** 91–96.

Kapitonov, V.V. and Jurka, J. 2001. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci.* **98:** 8714–8719.

Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.

Lai, J., Dey, N., Kim, C.S., Bharti, A.K., Rudd, S., Mayer, K.F., Larkins, B.A., Becraft, P., and Messing, J. 2004a. Characterization of the maize endosperm transcriptome and its comparison to the rice genome. *Genome Res.* **14:** 1932–1937.

Lai, J., Ma, J., Swigonova, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.J., Jeong, O.Y., Bennetzen, J.L., et al. 2004b. Gene loss and movement in the maize genome. *Genome Res.* **14:** 1924–1931.

Lai, J., Li, Y., Messing, J., and Dooner, H.K. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci.* **102:** 9068–9073.

Lal, S.K., Giroux, M.J., Brendel, V., Vallejos, C.E., and Hannah, L.C. 2003. The maize genome contains a helitron insertion. *Plant Cell* **15:** 381–391.

Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D., and Hallauer, A. 2002. Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. *Plant Mol. Biol.* **48:** 453–461.

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

Lund, G., Messing, J., and Viotti, A. 1995a. Endosperm-specific demethylation and activation of specific alleles of *a*-tubulin genes of *Zea mays* L. *Mol. Gen. Genet.* **246:** 716–722.

Lund, G., Prem Das, O., and Messing, J. 1995b. Tissue-specific DNase I-sensitive sites of the maize P gene and their changes upon epimutation. *Plant J.* **7:** 797–807.

Ma, J. and Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* **101:** 12404–12410.

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

McClintock, B. 1930. A cytological demonstration of the location of an interchange between two non-homologous chromosomes of *Zea mays*. *Proc. Natl. Acad. Sci.* **16:** 791–796.

Messing, J. 2005. The maize genome. *Maydica* **50:** 377–386.

Messing, J., Bharti, A.K., Karlowski, W.M., Gundlach, H., Kim, H.R., Yu, Y., Wei, F., Fuks, G., Soderlund, C.A., Mayer, K.F., et al. 2004. Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci.* **101:** 14349–14354.

Meyers, B.C., Tingey, S.V., and Morgante, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11:** 1660–1676.

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

Nelson, W.M., Bharti, A.K., Butler, E., Wei, F., Fuks, G., Kim, H.-R., Wing, R.A., Messing, J., and Soderlund, C. 2005. Whole-genome validation of high-information-content fingerprinting. *Plant Physiol.* **139:** 27–38.

Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., de la Bastide, M., Martienssen, R.A., and McCombie, W.R. 2003. Maize genome sequencing by methylation filtration. *Science* **302:** 2115–2117.

Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R., and Martienssen, R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23:** 305–308.

Rayburn, A.L., Biradar, D.P., Bullock, D.G., and McMurphy, L.M. 1993. Nuclear DNA content in F1 hybrids of maize. *Heredity* **70:** 294–300.

Rhoades, M.M. 1951. Duplicate genes in maize. *Am. Nat.* **85:** 105–110.

Salamov, A.A. and Solovyev, V.V. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10:** 516–522.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.

Schwartz, M., Zlotorynski, E., Goldberg, M., Ozeri, E., Rahat, A., le Sage, C., Chen, B.P., Chen, D.J., Agami, R., and Kerem, B. 2005. Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. *Genes & Dev.* **19:** 2715–2726.

Song, R. and Messing, J. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci.* **100:** 9055–9060.

Song, R., Llaca, V., Linton, E., and Messing, J. 2001. Sequence, regulation, and evolution of the maize 22-kD *a* zein gene family. *Genome Res.* **11:** 1817–1825.

Song, R., Llaca, V., and Messing, J. 2002. Mosaic organization of orthologous sequences in grass genomes. *Genome Res.* **12:** 1549–1555.

Sorrells, M.E., La Rota, M., Bermudez-Kandianis, C.E., Greene, R.A., Kantety, R., Munkvold, J.D., Miftahudin, Mahmoud, A., Ma, X., Gustafson, P.J., et al. 2003. Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.* **13:** 1818–1827.

Spena, A., Viotti, A., and Pirrotta, V. 1983. Two adjacent genomic zein sequences: Structure, organization and tissue-specific restriction pattern. *J. Mol. Biol.* **169:** 799–811.

Springer, N.M., Xu, X., and Barbazuk, W.B. 2004. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol.* **136:** 3023–3033.

Stam, M., Belele, C., Ramakrishna, W., Dorweiler, J.E., Bennetzen, J.L., and Chandler, V.L. 2002. The regulatory regions required for B' paramutation and expression are located far upstream of the maize b1 transcribed sequences. *Genetics* **162:** 917–930.

Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res.* **14:** 1916–1923.

Swigonova, Z., Bennetzen, J.L., and Messing, J. 2005. Structure and evolution of the r/b chromosomal regions in rice, maize and sorghum. *Genetics* **169:** 891–906.

Thomas Jr., C.A. 1971. The genetic organization of chromosomes. *Annu. Rev. Genet.* **5:** 237–256.

Thuriaux, P. 1977. Is recombination confined to structural genes on the eukaryotic genome? *Nature* **268:** 460–462.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl. Acad. Sci.* **96:** 7409–7414.

Town, C.D. 2006. Annotating the genome of *Medicago truncatula*. *Curr. Opin. Plant Biol.* **9:** 122–127.

Vieira, J. and Messing, J. 1982. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19:** 259–268.

Wessler, S.R., Bureau, T.E., and White, S.E. 1995. LTR-retrotransposons and MITEs: Important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5:** 814–821.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302:** 2118–2120.

Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2:** 333–341.

Yim, Y.S., Davis, G.L., Duru, N.A., Musket, T.A., Linton, E.W., Messing, J.W., McMullen, M.D., Soderlund, C.A., Polacco, M.L., Gardiner, J.M., et al. 2002. Characterization of three maize bacterial artificial chromosome libraries toward anchoring of the physical map to the genetic map using high-density bacterial artificial chromosome filter hybridization. *Plant Physiol.* **130:** 1686–1696.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A., and Tabata, S. 2005. Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Physiol.* **137:** 1174–1181.

Yuan, Y., SanMiguel, P.J., and Bennetzen, J.L. 2003. High-C_0t sequence analysis of the maize genome. *Plant J.* **34:** 249–255.

Zhang, Q., Arbuckle, J., and Wessler, S.R. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. *Proc. Natl. Acad. Sci.* **97:** 1160–1165.