Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus

Qinghua Wang* and Hugo K. Dooner*†‡

*The Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, NJ 08855; and †Department of Plant Biology, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved August 1, 2006 (received for review April 17, 2006)

Maize is probably the most diverse of all crop species. Unexpectedly large differences among haplotypes were first revealed in a comparison of the *bz* genomic regions of two different inbred lines, McC and B73. Retrotransposon clusters, which comprise most of the repetitive DNA in maize, varied markedly in makeup, and location relative to the genes in the region and genic sequences, later shown to be carried by two helitron transposons, also differed between the inbreds. Thus, the allelic *bz* regions of these Corn Belt inbreds shared only a minority of the total sequence. To investigate further the variation caused by retrotransposons, helitrons, and other insertions, we have analyzed the organization of the *bz* genomic region in five additional cultivars selected because of their geographic and genetic diversity: the inbreds A188, CML258, and I137TN, and the land races Coroico and NalTel. This vertical comparison has revealed the existence of several new helitrons, new retrotransposons, members of every superfamily of DNA transposons, numerous miniature elements, and novel insertions flanked at either end by TA repeats, which we call TAFTs (TA-flanked transposons). The extent of variation in the region is remarkable. In pairwise comparisons of eight *bz* haplotypes, the percentage of shared sequences ranges from 25% to 84%. Chimeric haplotypes were identified that combine retrotransposon clusters found in different haplotypes. We propose that recombination in the common gene space greatly amplifies the variability produced by the retrotransposition explosion in the maize ancestry, creating the heterogeneity in genome organization found in modern maize.

variability | helitrons | retrotransposons | corn

A comparison of ≈100 kb of sequence from the *bz* genomic regions of two maize inbred lines, McC and B73, revealed a surprising extent of noncolinearity between them (1). The lines differed in the makeup and size of retrotransposon blocks flanking genes, the pattern of interspersion of genes and retrotransposon blocks, and the presence vs. absence of four genic sequences that were later shown to be gene fragments carried by two helitron transposons, *HelA* and *HelB* (2). Overall, the two lines shared only 30% of the sequences in the region and could be aligned only at the genes they had in common (Fig. 1 *A* and *B*). McC is an inbred line carrying the *bz* region of a stock, probably of Northeastern flint origin, obtained from Barbara McClintock (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), whereas B73, the inbred chosen for the sequencing of the maize genome, traces its origin to the Iowa Stiff Stalk Synthetic (4).

Noncolinear haplotypes were also found in a comparison of the *z1C* zein genomic regions of B73 and BSSS53 (5), two Corn Belt inbreds extracted from the same population (4). The lengths of the *z1C* intervals varied by 50% because of differences in the number of zein genes and in the sizes of the retrotransposon clusters flanking them. Extensive nonhomologies from retrotransposons and helitrons were also reported at three additional loci in a comparison between the allelic regions of B73 and Mo17 (6, 7), two unrelated Corn Belt inbreds representing the heterotic pattern commonly used in the U.S. Corn Belt (8). The structure of the *bz* haplotype in Mo17 (6) differs from those in McC and B73 but is closer to the latter (Fig. 1*C*).

A survey of the variability of the *bz* genomic region in a panel of 10 public USA inbreds revealed that the size of a NotI fragment extending from *bz* to *uce2* ranged from 50 to 140 kb (1). This means that, remarkably, the size of a particular allelic region can vary by as much as 3-fold within maize. To investigate further the basis of this variability, we have carried out an analysis of BACs containing the *bz* genomic region from five additional maize haplotypes, selected to maximize genetic diversity.

Most of the North American Corn Belt germplasm is derived from mixtures of only two major USA races (9), so we have analyzed just one additional Corn Belt inbred, A188. The variability in maize genome organization uncovered recently, in retrospect, is presaged in early cytogenetic comparative studies of the highly variable races of maize. Races from different areas in the Americas were found to differ by the presence or absence of terminal and interstitial knobs, the size of the knob at a particular location, the average number of supernumerary B chromosomes, and the presence or absence of abnormal chromosome 10 (10). Therefore, to attempt to capture haplotype diversity most likely absent from North American Corn Belt lines (9), we have included in our sample four other cultivars of widely diverse geographic and genetic origin: CML258, an inbred extracted from the Mexican race Tuxpeño; I137TN, a mixed-origin inbred from South Africa; and the tropical land races Coroico, from the Amazon basin, and NalTel, from the Mexican lowland.

The vertical comparison of the eight *bz* haplotypes now available has revealed the existence of many polymorphic insertions in introns and intergenic regions. The extent of variation is unprecedented; the percentage of sequences shared by any two haplotypes ranges from 25% to 84%. Recombination in the common gene space has shuffled intergenic retrotransposon clusters, greatly amplifying the variability created by the retrotransposition explosion in the maize ancestry (11) and giving rise to the highly heterogeneous genome organization of modern maize.

Results

The results of the sequence analysis of the five new *bz* haplotypes are summarized in Fig. 1 *D–H*. As can readily be seen, each haplotype is unique. We will briefly describe each one, pointing out general similarities and differences with previously sequenced haplotypes.

A188. A188 is a mixed-origin inbred (12) that does not belong to either of the two main breeding groups producing the heterotic pattern commonly used in the U.S. Corn Belt. The structure of

Author contributions: H.K.D. designed research; Q.W. performed research; Q.W. and H.K.D. analyzed data; and H.K.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: MITE, miniature inverted-repeat transposable element; TAFT, TA-flanked transposon; TIR, terminal inverted repeat; ME, miniature element.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ493646–DQ493655).

‡To whom correspondence should be addressed. E-mail: dooner@waksman.rutgers.edu.

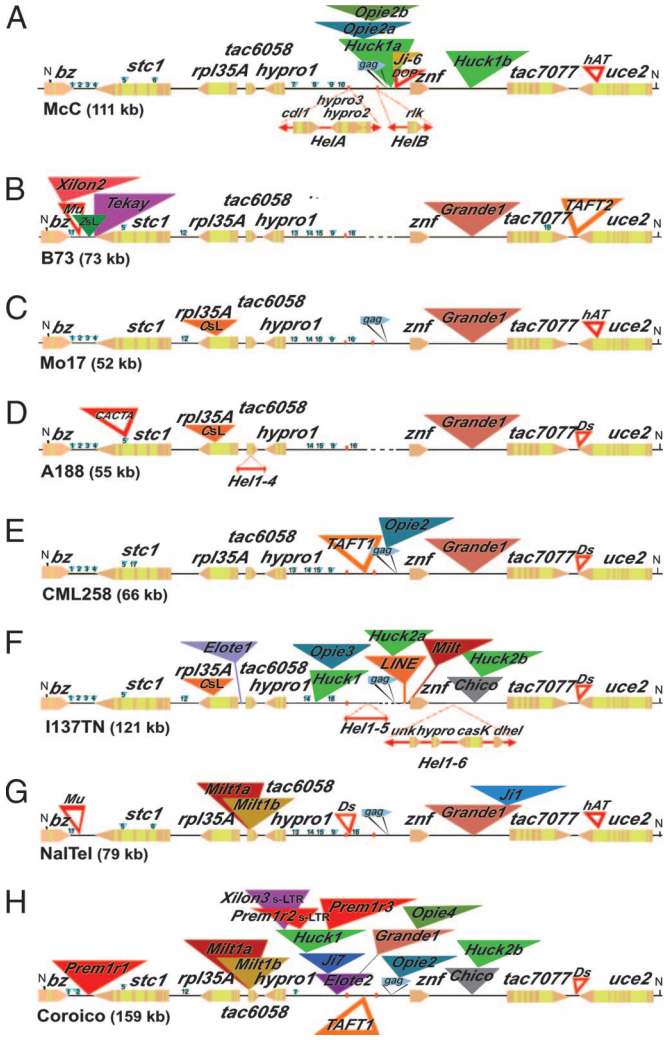

Fig. 1. Organization of eight *bz* haplotypes. Each haplotype is identified by the name of the genetic line, followed by the size of the cloned NotI fragment, in parentheses. The locations of the NotI sites at the proximal and distal ends are marked by Ns on the left and right, respectively. Genes are shown as pentagons pointing in the direction of transcription; exons are in bronze and introns in yellow. There are eight genes in the region: *bz*, *stc1*, *rpl35A*, *tac6058*, *hypro1*, *znf*, *tac7077*, and *uce2* (21). The same symbols are used for gene fragments carried by helitrons (*Hels*), which are represented as bidirectional arrows below the line for each haplotype. The vacant sites for *HelA* and *HelB* in each haplotype are provided as reference points and marked with short vertical strokes. Dashed lines represent deletions. Retrotransposons are indicated by solid triangles of different colors. DNA transposons and TAFTs, which are probably also DNA transposons, are indicated by open triangles in red and orange, respectively. Small insertions are indicated in light blue and are numbered as indicated in Table 3. Only the genes have been drawn to scale.

the 55-kb *bz* haplotype of A188 (Fig. 1*D*) closely resembles that of Mo17 (Fig. 1*C*) throughout its length. Both are short and relatively devoid of retrotransposons compared with the rest. The two haplotypes share a single 12.4-kb *Grande1* element between the *znf* and *tac7077* genes and a 0.68-kb *Cin1* solo-LTR in the large third intron of the *rpl35A* gene. However, the *Hopscotch* gag-pol fragment located between *hypro1* and *znf* is missing in A188, most likely from a deletion event, because sequences adjacent to *Hopscotch* in Mo17, including the *HelB* vacant site, are also missing in A188. The two haplotypes share the *HelA* vacant site and many of the same small insertions. A188 has an extra miniature inverted-repeat transposable element (MITE) in the fifth intron of the *stc1* gene, a *Tourist* element into which has inserted a 7.8-kb CACTA transposon. This transposon is probably inactive; it encodes a complete TNPD protein, but only a fragment of TNPA. In addition, A188 has a new 469-bp helitron inserted in an intron of the *tac6058* gene, which we have termed *Hel1-4* (see *Discussion*). This helitron shares homology at its 5′ and 3′ ends with *HelA* of McC, but, unlike previously described maize helitrons (2, 7, 13, 14), has no coding capability. Last, a 513-bp *hAT* DNA transposon separates *tac7077* from *uce2*.

CML258. CML258 is an inbred of the tropical/subtropical (TS) group (12) developed at the International Maize and Wheat Improvement Center (CIMMYT), Mexico City, Mexico. Its *bz* haplotype (Fig. 1*E*) resembles that of Mo17 in general organization. It shares with Mo17, and with the other Corn Belt inbreds, the single *Grande1* retrotransposon; both *HelA* and *HelB* vacant sites, including the *Hopscotch* gag-pol fragment; and 8 of 10 small insertions. It differs from Mo17 in having an *Opie2* LTR retrotransposon inserted in *Hopscotch*, two MITEs inserted in intron 5 of *stc1*, and the 0.5-kb *hAT* element between *tac7077* and *uce2*. However, the main difference is in the occurrence of a novel type of insertion, which we call TAFTs (TA-flanked transposons), in the hypervariable *hypro1-znf* intergenic region.

TAFTs are flanked on either side by TA microsatellites with as many as 50 copies of the repeat. The *bz* haplotypes that lack TAFTs have three TA repeats at the corresponding position. The *TAFT1* element in the CML258 *bz* region is 2.2 kb in length and exists in several copies in the maize genome. It possesses imperfect terminal inverted repeats of ≈40 bp and internal sequences that, although homologous to several maize ESTs, do not appear to have any coding capability. Curiously, *TAFT1* shares terminal inverted repeats (TIRs) with other larger TA-flanked sequences (GenBank accession nos. AF466931 and AF488416) that are predicted to encode proteins with homology to the putative transposase of the maize element *Jittery* (15). This latter observation suggests that TAFTs may belong to the *Mutator* transposon superfamily.

I137TN. I137TN is a mixed origin inbred (12) developed in South Africa. Its *bz* haplotype, the second largest so far sequenced, has a very distinct organization (Fig. 1*F*). It shares with previously described haplotypes only the *Cin1* solo LTR in *rpl35A*, the *Hopscotch* gag-pol fragment, and the 0.5-kb *hAT* element between *tac7077* and *uce2*. It has a new 8-kb *copia*-like LTR retrotransposon, which we have called *Elote1* (young corn, in Nahuatl), between *rpl35A* and *tac6058* and five different large insertions in the *hypro1-znf* intergenic region. From proximal to distal side, these are a 22.7-kb *Huck1-Opie3* double-decker retrocluster, which is missing the terminal 13 bp of the *Huck1* 3′ LTR; a 1-kb helitron, *Hel1-5*, which has homology with several maize ESTs, but no coding capability; the *Hopscotch* gag-pol fragment; an 18-kb insertion consisting of a 12-kb *Huck2* LTR retrotransposon piggybacked on a novel 6-kb non-LTR retroelement; and a 3.6-kb *Milt* LTR retrotransposon. The 6-kb retroelement has typical features of mammalian LINE elements (16): it is 5′ truncated, encodes an incomplete reverse transcriptase, ends in a polyA tract, and is flanked by a 15-bp target site duplication (TSD). Between *znf* and *tac7077* are inserted a 7.7-kb helitron, *Hel1-6*, and a 19-kb double-decker retrocluster. Like other compound helitrons, *Hel1-6* carries several different incomplete genes, which would encode protein fragments with homology to unknown and hypothetical proteins from rice, a casein kinase, and a DNA helicase. The 19-kb retrocluster has a 13.7-kb *Huck2* element inserted into a novel 5.3-kb LTR retrotransposon, which we have named *Chico* for its short LTRs (304 bp) and to contrast it with the larger *Grande1* found in the same intergenic region of other *bz* haplotypes (Fig. 1 *B–E*).

Table 1. Homologous sequences shared by the eight different *bz* haplotypes, expressed as a percentage of the average length of the two *bz* haplotypes being compared

	McC	B73	Mo17	A188	CML258	I137TN	Nal-Tel	Coroico
McC	100	—	—	—	—	—	—	—
B73	39	100	—	—	—	—	—	—
Mo17	42	69	100	—	—	—	—	—
A188	36	69	81	100	—	—	—	—
CML258	51	62	84	75	100	—	—	—
I137TN	56	30	38	35	46	100	—	—
Nal-Tel	36	62	76	64	66	37	100	—
Coroico	52	25	31	27	43	49	40	100

NalTel. NalTel is a small-eared land race from the east coast of Mexico (9). Like other open-pollinated land races (17, 18), it contains a mixture of genotypes. The characterized BAC carries the majority *bz* allele detected among 10 clones sequenced from a pool of five individuals (Table 2, which is published as supporting information on the PNAS web site). The NalTel haplotype (Fig. 1*G*) shares features with several other haplotypes, mostly B73. On the proximal side, it has the same unique MITE and *Mutator*-like element as B73, although the *Mutator* element does not carry a *Xilon2* retrotransposon. On the distal side, NalTel shares with McC and Mo17 a 0.9-kb *hAT* element in the last intron of *uce2*. The central sequence of the NalTel *bz* haplotype is a mosaic. The *stc1* allele of NalTel resembles the one in McC in having the same MITE insertions in introns 3 and 5. The *hypro1-znf* intergenic region resembles that of Mo17 in all five small insertions, the *HelA* and *HelB* vacant sites, and the *Hopscotch* gag-pol fragment. However, it has an extra 0.9-kb *Ds*-like *hAT* element immediately adjacent to the *HelA* vacant site. The *znf-tac7077* intergenic region has the same *Grande1* retrotransposon as several other haplotypes, but the element in NalTel is differentiated from the others by the gain of a 9.49-kb *Ji1* retrotransposon. *Ji1* has intact *gag* and *pol* ORFs and, based on the perfect identity of its LTRs, must have inserted very recently (19, 20). The main difference between NalTel and the previous haplotypes lies in the insertion of two *Milt1* elements, a 5.2-kb *Milt1a* and a 9.3-kb *Milt1b*, in *tac6058*, within 190 bp of each other. Both insertions fall in an exon of *tac6058*, a gene whose transcript is probably noncoding RNA (21).

Coroico. Coroico is the predominant race in the Amazon basin and surrounding lowlands (9). Three different *bz* alleles were identified in a sample of five individuals (Table 2), but the polymorphisms are upstream of the NotI site and do not serve to identify the allele present in the analyzed BAC. The Coroico *bz* haplotype contains the most complex pattern of retroelement insertions of the eight analyzed to date (Fig. 1*H*). Its structure is a mosaic of common and unique insertions. Coroico shares insertions with several other haplotypes: with NalTel, the two *Milt1* LTR retroelements in *tac6058*; with CML258, the TAFT element and the *Hopscotch*/*Opie2* complex in the *hypro1-znf* intergenic region; and with I137TN, the *Chico* LTR retroelement between *znf* and *tac7077*. Interestingly, none of these insertions occur in any of the four North American inbreds examined. In addition, Coroico has several unique insertions. A 7.8-kb *gypsy* retroelement related to *PREM1* sits between *bz* and *stc1*. Like the *PREM1* element in the McC haplotype (22), its LTRs share only the 3′ terminal part (1.6 of 3.9 kb) with the original *PREM1-E* element (23). We have called this group of elements *PREM1r* for *PREM1 r*elated. A large (66.5-kb), branched, and multilayered retrotransposon tree is inserted between *hypro1* and *znf*. This tree has a two-branch *Elote* element at its base. The 37.5-kb proximal branch contains a *Ji7* element in the first layer, a *Huck1* element in the second, and three different elements in the third. These are a 3.6-kb *Xilon3* solo LTR, which has homology to *Xilon1* only in the first 1.4 kb; a 3.5-kb *PREM1r2* solo LTR, which shares only the 3′ terminal 1.5 kb with other members of the family; and a 9.1-kb *PREM1-r3* element, whose LTRs similarly share only the 3′ part with other *PREM1* LTRs. The 23.1-kb distal branch of the tree contains a *Grande1* retrotransposon and an element of the *Opie* family. Last, as in I137TN, a 13.7-kb *Huck* element is inserted within the *Chico* retroelement in the *znf-tac7077* intergenic region. The pattern of small insertions in Coroico is also a mosaic but appears to be closest to that of I137TN. These two haplotypes share the small insertions 1, 2, 5, and 12 (Fig. 1). However, the only other identifiable small insertion is the *Stowaway* element located immediately distal to *hypro1* (7 in Fig. 1), which is found only in McC.

Discussion

Diversity of Maize Haplotypes. A cursory inspection of Fig. 1 is enough to reveal the exceptional diversity of haplotypes found in modern maize. We use the term "haplotype" in its classical genetic context, i.e., to refer to a set of very closely linked alleles that tend to be inherited as a unit because they are not easily separable by recombination. The term is appropriate here, because the genetic distance between *bz* and *tac7077*, genes at opposite ends of the analyzed fragment, is just 1 cM (L. He and H.K.D., unpublished data). Insertion polymorphisms occur in the introns and untranslated regions of several genes and in every intergenic region. The insertions include helitrons; LTR and LINE retrotransposons; members of the *hAT*, *CACTA*, and *Mutator* superfamilies of DNA transposons; numerous MITEs and other small insertions; and a new class of transposons that we have termed TAFTs. They range in size from a 58-bp MITE to a 66.5-kb branched retrotransposon tree. Insertion nests are found in seven of the eight *bz* haplotypes, Mo17 being the only exception. Nests of LTR retrotransposons (24) are the most common, but other combinations of insertions also occur, such as LTR retrotransposons into non-LTR retrotransposons (I137TN) or into DNA elements (B73).

Table 1 lists the percentage of shared homologous sequences at allelic locations in pairwise comparisons of the eight *bz* haplotypes. The percentage ranges from a low of 25% in a B73xCoroico hybrid to a high of 85% in a Mo17xCML258 hybrid. Haplotypes of the three North American Corn Belt inbred lines are more closely related to each other and to the two tropical lowland entries (CML258 and NalTel) than they are to haplotypes of other tropical and subtropical origins (I137TN and Coroico), in agreement with findings from previous large-scale isozyme and simple sequence repeat studies (12, 25). In fact, the five haplotypes share the same *Grande1* retrotransposon in the *znf-tac7077* intergenic region; the only difference among them is the recent acquisition of a *Ji1* retrotransposon by *Grande1* in NalTel.

On average, any two of the eight *bz* haplotypes share only 50% of their sequences. Similarly, a comparison of ≈0.3–0.4 Mb per locus for each of three loci in B73 and Mo17 found that, on

average, only 50% of the sequence at each locus was shared between the two maize inbreds (6). If this level of dissimilarity extended to other regions of the genome, two homologues taken at random from the pool of eight in Table 1 would pair through only half of their length. This could lead to dramatic differences in estimates of genetic distances in different heterozygotes, but it does not. Although maize geneticists work with many different lines, estimates of map distances for particular genetic intervals, while variable (26, 27), are not wildly different. The reason is that retrotransposons, the largest contributors to sequence heterogeneity, are recombinationally inert (22, 28). However, short map distances between very closely linked genes could be affected by local structural heterozygosity. For example, recombination between the adjacent *bz* and *stc1* genes may be lower in a heterozygote between McC and B73 haplotypes, which contains a heterozygous 26-kb retrotransposon block in the *bz-stc1* interval, than in a heterozygote between McC and A188 haplotypes, which does not. Recent data indicate this is the case (29).

Insertions. ***Retrotransposons.*** As in other regions of the maize genome (e.g., refs. 22, 24, and 30), LTR retrotransposons in the different *bz* haplotypes can occur as single elements, nested clusters, or solo LTRs. The time of insertion of the first two classes can be estimated from a comparison of the sequence of the two LTRs (19). In agreement with previous analyses, the elements at the top of the clusters in McC, I137TN, NalTel, and Coroico are generally younger than the ones in the bottom (Table 3, which is published as supporting information on the PNAS web site). Interestingly, solo LTRs, which range in size from 683 bp in *Zeon1* to 3,571 bp in *Xilon3*, occur either singly (B73, Mo17, A188, and I137TN) or at the top of a branch (Coroico) but not at the base of any branch. The same is true for the few solo LTRs in the *adh1* locus (24). In contrast, in barley, where solo LTRs are relatively much more common than in maize (31), several examples of nested solo LTRs have been reported at the *Rar* locus (32). This difference could be due simply to chance, because the number of solo LTRs described so far in maize is low, or it could indicate that the intrachromosomal recombination events that produce solo LTRs have occurred only recently in maize, subsequent to the burst in transposition that created the retroclusters and that presumably peaked ≈1.5 million years ago (19, 33). Restriction of solo LTR formation in maize to only very recent evolutionary times would also help to explain why the relative ratio of intact LTR elements to solo LTRs is higher in maize than other plants (31, 32, 34, 35).

Thirty different LTR retrotransposon insertion sites can be counted among the eight haplotypes in Fig. 1. Of them, nine sites had been identified earlier in McC or B73 (1) and one in Mo17 (6), and 20 are new. Among the newly identified sites, three are occupied by retrotransposons with novel LTRs not previously found in the sequence database (*Chico*, *Elote1*, and *Elote2*). Other elements, like the *Prem1*-related (*Prem1r*) elements of Coroico, have LTRs with considerably less than 50% sequence identity to other LTRs, the suggested criterion for grouping retrotransposons into families (11), but were still assigned to the *Prem1* family because of high identity to the *Prem1* LTR over a long stretch of sequence. The relative ease with which new LTRs have been identified in carefully annotated sequence comparisons between two inbreds (1, 5, 6), coupled with the large amount of LTR retrotransposon polymorphisms uncovered in germplasm of geographically diverse origin, supports the prediction (11) that there may be thousands of different families of low- to middle-copy number LTR retrotransposons in maize.

Non-LTR retrotransposons comprise a very low fraction (0.2%) of the maize genome (36), and the *bz* genomic region is no exception. A single non-LTR retrotransposon was identified in >700 kb of combined sequence, the LINE element inserted proximal to *znf* in I137TN. The subsequent insertion of the LTR retrotransposon *Huck2* into this LINE allows us to estimate that the latter inserted in the I137TN *bz* genomic region at least 0.3 million years ago.

Helitrons. Helitrons are a novel class of transposable elements discovered recently by computational analysis of the genome sequences of *Caenorhabditis elegans*, *Arabidopsis*, and rice (37). All of the maize helitrons identified so far are large (>1 kb) and contain gene fragments (2, 7, 13, 14). Although they differ greatly in internal sequences, maize helitrons share substantial sequence homology at their ends. The high sequence conservation of the 3′ terminal 30 nucleotides allows the grouping of maize helitrons into two major clades or superfamilies, *Hel1* and *Hel2* (38). The new helitrons identified in this study belong to the *Hel1* superfamily, hence their designation as *Hel1-x* elements (where *x* is a number identifying the specific element). One of the two helitron elements in I137TN, *Hel1-6*, is of the same general type previously described in maize, large and carrying gene fragments. However, a different type of helitron, much smaller and without gene fragments, can be readily identified by resequencing the same genomic region in multiple lines. *Hel1-4* and *-5* are two such helitrons. *Hel1-4* is a 469-bp helitron inserted in the *tac6058* gene of A188, and *Hel1-5* is a 994-bp helitron inserted in the *hypro1-znf* intergenic region of I137TN. Neither has any coding capability, so they are akin to the simple *AthE1* repeats of *Arabidopsis* (39) and the *Helitrony* elements of *C. elegans* (37). BLASTN analysis reveals that sequences closely related to both *Hel1-4* and *-5* are scattered in the genome, indicating these elements are also present in multiple copies.

Among the sequenced *bz* haplotypes, helitrons are considerably less abundant than LTR retrotransposons. Helitrons are present only in McC, I137TN, and A188, where they make up 7.9%, 5.4%, and 0.8%, respectively, of the total sequence. The overall helitron content of the maize genome has been estimated to be <0.01% (36), but this may be an underestimate because helitrons lack the obvious structural features of most eukaryotic transposons and are harder to detect by computational searches. On the other hand, they can be readily detected in vertical sequence comparisons, so one can expect more helitrons of the agenic shorter type to be uncovered in future sequence comparisons of different maize lines.

Class II DNA transposons. Members of every major superfamily of excisive DNA transposons can be found in the small panel of *bz* haplotypes examined.

A 2.5-kb defective *MuDR* derivative (40) with an apparently complete *MURB* gene, but only a fragment of the *MURA* transposase gene, is present in the *bz-stc1* intergenic region of NalTel and B73. The elements have become differentiated, though, by the insertion of a *Xilon2* LTR retrotransposon in B73, but not in NalTel.

A 7.8-kb CACTA element (41) is inserted within a *Tourist* MITE in the fifth intron of the A188 *stc1* gene. This element is not transpositionally active in A188 (data not shown), although it may be mobile in other lines, because it has perfect 13-bp TIRs that differ from those of *Spm/En* in only two positions.

Nonautonomous *hAT* elements (41) with no evident coding capacity are present at three locations: in the *hypro1-znf* intergenic region of NalTel; in the *tac7077-uce2* intergenic region of I137TN and Coroico; and in the last intron of *uce2* in McC, Mo17, and NalTel. All elements are small (0.5–0.9 kb) and end in 11-bp TIRs. The TIRs of *Ds* and the first two elements are >70% identical, so these elements may behave like *Ds*, but their mobility in response to *Ac* has not been tested.

Small insertions. MITEs (42) and other small sequences without obvious TIRs are present in every *bz* haplotype examined. We will refer to them collectively as miniature elements (MEs) based on their small size. Seventeen different MEs were identified as insertion polymorphisms in one or another *bz* haplotype (Table 4, which is published as supporting information on the PNAS web site).

These sequences are clearly insertions, because, in addition to being polymorphic, they are present at other sites in the genome. Most of them (14/19) have typical MITE features. All of the MEs in the *bz* genomic region inserted in either introns or intergenic spaces, in contrast to those at three other loci, where half of the MITEs inserted in transposons or LTR retrotransposons (6).

The junctions of MEs with host DNA, like those of LTR retrotransposons, appear to be very stable components of the maize genome and help to define relationships among haplotypes. For example, one finds the same combination of four MEs (numbers 1-2-3-4) in the *bz-stc1* intergenic region of McC, Mo17, A188, CML258, and I137TN, indicating that in those inbreds, this region had a common origin. Similarly, the same *Tourist* element (number 5) is present in the fifth intron of all of the *stc1* alleles but one, again pointing to a common origin. In the only exception, *stc1-Mo17*, the *Tourist* insertion site has been deleted. Interestingly, in some haplotypes, one particular ME may be missing from a combination of adjacent MEs. For example, ME 13 from the combination 13-14-15-9 is the only one of the group missing in A188. Alternatively, a MITE that is part of a group in one haplotype may be present singly in another (compare ME 7 in McC and Coroico). The latter two observations could suggest a temporal order of ME arrival, occasional loss of MEs by excision, or shuffling of MEs by rare recombination between haplotypes.

TAFTs. A new type of insertion flanked by long stretches of TA repeats, hence named TAFTs for TA-flanked transposons, was discovered during the course of this investigation. TAFTs have the following properties. (*i*) They are flanked by TA microsatellites with as many as 50 copies of the TA repeat on either side of the insertion. (*ii*) The corresponding vacant site in haplotypes that lack TAFTs consists generally of a few (three or four) copies of the TA repeat, but there are exceptions. (*iii*) The elements identified so far possess imperfect TIRs of ≈40 bp and are relatively large (>2 kb). (*iv*) Related copies are found at other locations in the maize genome, where they are also flanked by TA repeats.

The *TAFT1* element in CML258 and Coroico is 2.2-kb long and is flanked on either side by multiple copies of TA. The vacant site in the other haplotypes has three copies of the TA dinucleotide. *TAFT1* is 85% identical to oligo-TA flanked sequences of similar length that occur in B73 locus 9009 (GenBank accession no. AY664415) and in BAC c573L14 (GenBank accession no. AY555143). Although *TAFT1* does not appear to encode any protein, its TIRs share homology with those of other larger TA-flanked sequences in the maize genome that have the capability of encoding two proteins: one with homology to JITA, the putative *Jittery* transposase (15), and one of unknown function. Such putative TAFTs are >10 kb long and occur in both the *wx* (GenBank accession no. AF488416) and *rp1* (GenBank accession no. AF466931) loci of inbred B73. Because JITA is related to MURA, the *MuDR* transposase (40), TAFTs may be unusual members of the *Mutator* superfamily. Hereafter, we will refer to the ends upstream and downstream of the putative transposase as 5′ and 3′, respectively. A 13-kb oligo TA-flanked sequence with a similar type of organization also occurs in rice BAC AC136501. BLASTX analysis indicates this sequence has the potential to encode proteins with amino acid homology to both ORFs in the putative TAFTs found at the maize *wx* and *rp1* loci. These observations suggest that TAFT elements are not restricted to maize.

The *TAFT2* element was discovered in the *tac7077-uce2* intergenic region of the B73 *bz* haplotype upon reanalysis of that sequence (1). It is 2.5-kb long and has imperfect 5′ and 3′ TIRs that are 78% and 40% identical, respectively, with those of *TAFT1*. The number of TA repeats in the vacant site of the other haplotypes varies from 4 to 37. Like *TAFT1*, *TAFT2* does not appear to encode any protein but shares termini with TA-flanked sequences that do. Its TIRs are 90% identical to sequences found at the ends of a 4.1-kb TA-flanked sequence close to the *tb1* locus of inbred B73 (GenBank accession no. AY325816). Furthermore, *TAFT2* and the putative *TAFT* in *tb1* are 78% identical over the 5′ terminal 530 bp, pointing to a common origin. The latter element encodes a fragment of a putative metal-transporting ATPase (gene 2 in ref. 43), suggesting that, like Pack-MULEs (44) and helitrons (45), TAFTs may be able to capture fragments of genes and mobilize them in the genome.

One can only speculate at this time on the mechanism of transposition of TAFT elements. TAFTs appear to insert into chromosomal sites that contain short stretches of oligo TA. Because TAFTs are flanked on either side by TA microsatellites with as many as 50 copies of the repeat, the transposition process appears to lead to the amplification of an oligo TA already present at the insertion site. One possibility would be if the TAFT transposase is site-specific and introduces staggered cuts at preexisting oligo TA sites. The short single-stranded gaps at the end of the transposon, which are repaired by host functions to generate the characteristic target site duplications, would be repaired imprecisely because of the internally repetitive nature of the target site. Stuttering of the DNA polymerase at the single-stranded oligo TA gap would lead to the synthesis of a longer stretch of TAs than needed to repair the gap and thus to the formation of TA bubbles at either end of the transposon. Repair DNA synthesis would subsequently fill in the bubbles, producing TA homoduplexes of variable length next to the transposon.

Origin of the Haplotype Variation. The haplotypic diversity in maize is staggering. Because retrotransposons do not excise, the retrotransposon–host DNA junctions remain fixed through time, unless they are deleted by chromosomal breaks caused by subsequent unrelated events. At the *bz* locus, there is evidence that such deletion events did occur, because a *Hopscotch* gal-pol fragment not flanked by LTRs is present in diverse *bz* haplotypes, suggesting an ancient origin. But most retrotranspositions have left their signature in the maize genome, enabling us today to establish relationships among haplotypes based on a comparison of their retrotransposon makeup. Thus, a comparison of the *znf-tac7077* intergenic region reveals three haplotype groups: one group has a *Grande1* element at the same site (B73, Mo17, A188, CML258, and NalTel); a second group has a *Chico-Huck2* combination (I137TN and Coroico); and a third group has a single *Huck1* element (McC). Similarly, a comparison of the long third intron of *rpl35A* identifies two groups, based on the presence (Mo17, A188, and I137TN) vs. absence (the rest) of a *Cin1* solo LTR. And a comparison of *tac6058* separates out haplotypes by the presence (NalTel and Coroico) vs. absence (the rest) of two neighboring *Milt* elements. It is clear from this simple comparison that, in terms of its retrotransposon makeup for these three segments in the *bz* genomic region, inbred I137TN, for example, is a chimera of three clades. It groups with Mo17 and A188 on the basis of *rpl35A*; with all haplotypes, except NalTel and Coroico, on the basis of *tac6058*; and only with Coroico on the basis of the *znf-tac7077* intergenic region. Thus, the retroelements in different parts of the same haplotype have been shuffled extensively by recombination, which occurs at least 2 orders of magnitude more frequently in genes than in retrotransposons (22).

Because recombination in intergenic regions is very low, even in the absence of retrotransposons (29), it may be possible to treat these regions largely as indivisible blocks. Fig. 1 supports this view. The only exception, the *hypro1-znf* intergenic regions of CML258 and Coroico, which share the distal, but not the proximal, side could have been reshuffled by recombination within the *TAFT1* element in the center. Analysis of other

haplotypes will reveal whether this block view of maize genome organization is valid.

The retrotransposon explosion that produced this diversity of insertions is estimated to have taken place ≈1 million to 1.5 million years ago (19, 20, 46), much earlier than the domestication of maize from teosinte, which has been dated to 9,000 years before present (47). Thus, the haplotype diversity uncovered here, like the high sequence diversity at select loci (48, 49), was already present in the wild progenitor of maize at the time of domestication. Although maize went through a domestication bottleneck, it is still estimated to possess ≈75% of the allelic diversity present in teosinte (48), except at domestication genes (50–52). Because teosinte can also be expected to be more polymorphic than maize for large insertions, we have begun to examine *bz* haplotype variability in different teosinte accessions.

The present study raises an intriguing question: how many different haplotypes are possible at the average unselected locus in maize? Among the eight *bz* haplotypes, at least three different blocks of retrotransposon–host DNA junctions were identified in four of the seven intergenic regions. Assuming no other polymorphic intergenic regions and just five possible blocks of retrotransposon–host DNA junctions per polymorphic region, there would be 625 *bz* haplotypes if all of the regions recombined with each other. This latter assumption is reasonable, given that linkage disequilibrium in maize generally decays within genes (49, 53). Extrapolating to the whole genome under the very conservative assumption of 100 freely recombining regions, there would be 7.8×10^{69} possible combinations of retrotransposon blocks in maize! It would be surprising if at least some of this variability did not lead to phenotypic differences.

Materials and Methods

BAC Isolation. NotI BAC clones of the *bz* genomic region from different maize inbreds and land races were isolated as described (54).

DNA Sequencing, Assembly, and Analysis. The BAC clones were sequenced by the shotgun-sequencing strategy, assembled, and analyzed as described (22). Retrotransposons were annotated following suggested conventions (11). Solo LTRs are distinguished from truncated elements with only one LTR (not found in any *bz* haplotype yet) by the retention of the 5-bp target site duplication. Small-insertion polymorphisms were identified by comparisons among the different *bz* haplotypes using MegAlign from DNAStar (Madison, WI). These sequences were then used to query the maize sequence databases and named according to previously described insertions or given new names, if not previously identified. Most of these small insertions (14/19) had typical MITE features (Table 4).

We thank Major Goodman for advice in choosing maize lines for analysis, Brandon Gaut for helpful discussions, Limei He for help with the ABI3700 sequencer, Chunguang Du for assistance with the calculation of retrotransposon insertion dates, and members of the Dooner laboratory for comments on the manuscript. This research was supported by National Science Foundation Grant DBI 03-20683.

1. Fu H, Dooner HK (2002) *Proc Natl Acad Sci USA* 99:9573–9578.
2. Lai J, Li Y, Messing J, Dooner HK (2005) *Proc Natl Acad Sci USA* 102:9068–9073.
3. Ralston EJ, English J, Dooner HK (1988) *Genetics* 119:185–197.
4. Gerdes JT, Behr CF, Coors JG, Tracy WF (1993) *Compilation of North American Maize Breeding Germplasm* (Crop Sci Soc of Am, Madison, WI).
5. Song R, Messing J (2003) *Proc Natl Acad Sci USA* 100:9055–9060.
6. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) *Plant Cell* 17:343–360.
7. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) *Nat Genet* 37:997–1002.
8. Hallauer AR, Russell WA, Lamkey KR (1988) in *Corn and Corn Improvement*, eds Sprague GF, Dudley JW (Am Soc Agron, Madison, WI), pp 463–564.
9. Goodman MM, Brown WL (1988) in *Corn and Corn Improvement*, eds Sprague GF, Dudley JW (Am Soc Agron, Madison, WI), pp 33–79.
10. McClintock B, Kato TA, Blumenschein A (1981) *Chromosome Constitution of Races of Maize* (Colegio de Postgraduados, Chapingo, Mexico).
11. SanMiguel P, Bennetzen JL (1998) *Ann Bot* 82:37–44.
12. Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) *Genetics* 165:2117–2128.
13. Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) *Plant Cell* 15:381–391.
14. Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK (2005) *Plant Mol Biol* 57:115–127.
15. Xu Z, Yan X, Maurais S, Fu H, O'Brien DG, Mottinger J, Dooner HK (2004) *Plant Cell* 16:1105–1114.
16. Moran JV, Gilbert N (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (Am Soc Microbiol Press, Washington, DC), pp 836–869.
17. Goodman MM, Stuber CW (1983) *Maydica* 28:169–187.
18. Doebley JF, Goodman MM, Stuber CW (1985) *Am J Bot* 75:629–639.
19. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) *Nat Genet* 20:43–45.
20. Ma J, Bennetzen JL (2004) *Proc Natl Acad Sci USA* 101:12404–12410.
21. Fu H, Park W, Yan X, Zheng Z, Shen B, Dooner HK (2001) *Proc Natl Acad Sci USA* 98:8903–8908.
22. Fu H, Zheng Z, Dooner HK (2002) *Proc Natl Acad Sci USA* 99:1082–1087.
23. Turcich MP, Mascarenhas JP (1994) *Sex Plant Reprod* 7:2–11.
24. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, *et al.* (1996) *Science* 274:765–768.
25. Stuber CW, Goodman MM (1983) *Biochem Genet* 21:667–689.
26. Williams CG, Goodman MM, Stuber CW (1995) *Genetics* 141:1573–1581.
27. Timmermans MC, Das OP, Bradeen JM, Messing J (1997) *Genetics* 146:1101–1113.
28. Yao H, Schnable PS (2005) *Genetics* 170:1929–1944.
29. He L, Dooner HK (2006) in *48th Ann Maize Genet Conf Abstracts*, Vol 48, p 133.
30. Song R, Llaca V, Linton E, Messing J (2001) *Genome Res* 11:1817–1825.
31. Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) *Plant Cell* 11:1769–1784.
32. Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) *Genome Res* 10:908–915.
33. Bennetzen JL, Ma J, Devos KM (2005) *Ann Bot (London)* 95:127–132.
34. Devos KM, Brown JKM, Bennetzen JL (2002) *Genome Res* 12:1075–1079.
35. Vitte C, Panaud O (2003) *Mol Biol Evol* 20:528–540.
36. Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, *et al.* (2004) *Proc Natl Acad Sci USA* 101:14349–14354.
37. Kapitonov VV, Jurka J (2001) *Proc Natl Acad Sci USA* 98:8714–8719.
38. Dooner HK, Lal SK, Hannah LC (2006) *Maize Genetics Coop Newslet* 81, www.agron.missouri.edu/mnl/81.
39. Surzycki SA, Belknap WR (1999) *J Mol Evol* 48:684–691.
40. Walbot V, Rudenko GN (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (Am Soc Microbiol Press, Washington, DC), pp 533–564.
41. Kunze R, Weil CF (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (Am Soc Microbiol Press, Washington, DC), pp 565–610.
42. Feschotte C, Zhang Y, Wessler SR (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (Am Soc Microbiol Press, Washington, DC), pp 1147–1158.
43. Clark RM, Linton E, Messing J, Doebley JF (2004) *Proc Natl Acad Sci USA* 101:700–707.
44. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) *Nature* 431:569–573.
45. Lal SK, Hannah LC (2005) *Proc Natl Acad Sci USA* 102:9993–9994.
46. Clark RM, Tavare S, Doebley J (2005) *Mol Biol Evol* 22:2304–2312.
47. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J (2002) *Proc Natl Acad Sci USA* 99:6080–6084.
48. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998) *Proc Natl Acad Sci USA* 95:4441–4446.
49. Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) *Proc Natl Acad Sci USA* 98:9161–9166.
50. Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) *Nature* 398:236–239.
51. Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES, IV (2002) *Proc Natl Acad Sci USA* 99:12959–12962.
52. Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD (2005) *Plant Cell* 17:2859–2872.
53. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES, IV (2001) *Proc Natl Acad Sci USA* 98:11479–11484.
54. Fu H, Dooner HK (2000) *Genome Res* 10:866–873.