A GeneTrek analysis of the maize genome

Renyi Liu*, Clémentine Vitte*, Jianxin Ma*, A. Assibi Mahama†, Thanda Dhlwayo†, Michael Lee†, and Jeffrey L. Bennetzen**

*Department of Genetics, University of Georgia, Athens, GA 30602; and †Department of Agronomy, Iowa State University, Ames, IA 50011

Contributed by Jeffrey L. Bennetzen, May 7, 2007 (sent for review March 19, 2007)

Analysis of the sequences of 74 randomly selected BACs demonstrated that the maize nuclear genome contains $\approx 37,000$ candidate genes with homologues in other plant species. An additional $\approx 5,500$ predicted genes are severely truncated and probably pseudogenes. The distribution of genes is uneven, with $\approx 30\%$ of BACs containing no genes. BAC gene density varies from 0 to 7.9 per 100 kb, whereas most gene islands contain only one gene. The average number of genes per gene island is 1.7. Only 72% of these genes show collinearity with the rice genome. Particular LTR retrotransposon families (e.g., *Gyma*) are enriched on gene-free BACs, most of which do not come from pericentromeres or other large heterochromatic regions. Gene-containing BACs are relatively enriched in different families of LTR retrotransposons (e.g., *J1*). Two major bursts of LTR retrotransposon activity in the last 2 million years are responsible for the large size of the maize genome, but only the more recent of these is well represented in gene-containing BACs, suggesting that LTR retrotransposons are more efficiently removed in these domains. The results demonstrate that sample sequencing and careful annotation of a few randomly selected BACs can provide a robust description of a complex plant genome.

gene distribution | gene number | genome annotation | repetitive DNA | sample sequencing

Whole-genome sequence analysis has revolutionized the field of plant genetics. A sequenced plant genome provides the full list of genetic elements and also the context in which these elements function. The near-complete sequence of *Arabidopsis thaliana* (1) enabled the Arabidopsis 2010 project, which proposes to characterize the function of all of the genes in the genome (2). In addition, comparative analysis of the genomes of two or more species with known evolutionary relatedness is a powerful way to identify functional elements, to transfer knowledge from well studied model organisms to related plants, and to infer the mechanisms of genome evolution. For example, comparative analysis of orthologous regions from multiple cereals, including maize, rice, sorghum and wheat, has provided abundant information about the timing, nature and mechanisms of small rearrangements within those genomes (3–7).

Because of the high cost of whole-genome sequencing, it is best to choose the most cost-effective sequencing method, a choice influenced by genome size, gene content, and sequence organization. For small genomes or those with genes highly separated from most repetitive DNAs, BAC-by-BAC sequencing is the obvious approach. For the larger genomes characteristic of most flowering plants, where repeats are often intermixed with genes, gene-enrichment approaches have been suggested as a more efficient sequencing strategy (8, 9). Therefore, before deciding on an effective sequencing approach, it is appropriate to first understand the general composition and structure of the target genome.

The GeneTrek approach has been proposed as an efficient way to evaluate the general properties of any genome (10, 11). The principle is to sequence and annotate a small randomly selected portion of the genome, a strategy first used by Brenner and coworkers in the analysis of the *Fugu* genome (12). For maize, analysis of $\approx 475,000$ BAC end sequences led to the estimation that the maize genome contains $\approx 59,000$ genes and is $\approx 58\%$ repetitive DNA (13). Further, sequence analysis of 100 randomly selected

BAC clones from maize led to the prediction of 42,000–56,000 gene models and at least 66% repetitive DNA (14).

BAC sequences provide additional context information to a sample sequence analysis, thus allowing genome predictions that greatly enrich the GeneTrek approach. Here we describe a procedure for accurately predicting plant genome structure and composition with a relatively small data input and use this approach in comprehensive sequence annotation of randomly selected BACs that contain DNA from maize inbred B73. The results indicate that the maize genome contains many gene-free regions, many highly truncated gene fragments, and a nonrandom distribution of repetitive elements within different repeat-rich domains

Results

Repeat and Mobile DNA Annotation. Repetitive elements, especially LTR retrotransposons, are a major component of the maize genome (15–17). Exhaustive identification of repetitive elements is important not only for the accurate estimation of the amount of repetitive DNA but also for the estimation of gene content. Inadequate identification of repetitive elements may lead to consistent overestimation of gene number in plants because many repeats are scored as genes by gene prediction programs, and these repeats are well represented in EST databases (18, 19).

A total of 7.56 Mb of repeats were identified on the 74 randomly selected BACs by comparing BAC sequences to the TIGR maize repeat database (20) and an in-house maize retrotransposon database (P. SanMiguel, personal communication). Intact LTR retrotransposons were then identified by a structure-based search, leading to the characterization of an additional 710 kb of LTR retrotransposons. Hence, mobile and/or repetitive DNA accounted for at least 8.3 Mb or 68% of the BAC sequences (Table 1).

Gene Annotation. When the FGENESH gene prediction program was applied to the 74 BACs, 2,137 gene models were predicted. Some models [1,790 (84%)] were parts of identified repeats (mostly LTR retrotransposons) and were removed from further analysis. Based on stringent criteria (i.e., either significant homology ($e \leq 10^{-10}$) to other species and/or colinearity with rice genes), 216 gene models were classified as verified gene candidates. After comparing to rice or *Arabidopsis* genes, 28 of these gene models were identified as gene fragments. This left 188 annotated genes over 74 BACs, with an average gene density of 1 gene per 65 kb (Table 1). The remaining 131 gene models were classified as “hypothetical proteins,” many of which will eventually be shown to be parts of heretofore undiscovered transposons (18). If these gene numbers are extrapolated to the whole maize genome, with an estimated size of 2,400 Mb, then maize contains 37,000 (verified) to

Author contributions: R.L. and C.V. contributed equally to this work; R.L., C.V., J.M., and J.L.B. designed research; R.L., C.V., J.M., and A.A.M. performed research; R.L., C.V., J.M., A.A.M., T.D., M.L., and J.L.B. analyzed data; and R.L., C.V., M.L., and J.L.B. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: TSD, target site duplication; My, million years; RJM, repeat junction marker; TE, transposable element; IBM, Internated B73 $\times$ Mo17.

†To whom correspondence should be addressed. E-mail: maize@uga.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0704258104/DC1.

© 2007 by The National Academy of Sciences of the USA

Fig. 2. Graphic representation of BAC AC147789. Total size of the BAC is shown in parentheses. For LTR retrotransposons, family names are shown on top of the element, and the number after the dash represents each copy. LTR divergence is represented on top of the name. Locations of the primers used for RJMs are represented as small arrows.

family (SI Table 6). The “top” elements are usually younger (based on comparison of the LTR sequences) than the elements in which they have inserted, although there are six exceptions (see Fig. 2 and SI Figs. 4–8).

Generation of Unique Markers from Repeated Sequences and Mapping of Gene-Free BACs. Gene-free genomic fragments contain mainly repetitive sequences. Hence, they may be difficult to map genetically, because they do not harbor unique sequences that can be used as molecular markers. However, the insertion of a particular LTR retrotransposon within another can be used as a unique marker, because the combination of both the nature of the two elements and the site of insertion may be unique in the genome (11). Of 29 LTR repeat junction markers (RJMs) tested, 11 (37.9%) led to the amplification of a single band of expected size in B73 maize and seven (24.1%) to the amplification of a single band of expected size, plus one or two additional faint bands (SI Table 7; see SI Table 8 for details). Five of these 18 RJMs were polymorphic between B73 and Mo17 and were used to position these BACs on the maize genetic map. These five polymorphic markers are from BACs AC147789 (marker A6), AC148081 (markers B3 and B4), and AC147809 (markers F2 and F5).

Evidence for linkage was found for all markers, leading to the positioning of BAC AC147789 on chromosome 1, bin 1.06, between markers umc1972 (15.4 cM) and asg58 (43.1 cM); BAC AC148081 on chromosome 7, bin 7.02, between markers phi034/cyp6 (0 cM) and umc1983 (18.9 cM); and BAC AC147809 on chromosome 6, bins 6.01–6.02, between markers umc1006 (4.6) and bnlg1867 (4.9 cM) (SI Table 9). In cases where two markers were available for the same BAC (i.e., markers B3 and B4 on AC148081 and markers F2 and F5 on AC147809, respectively), these two markers were mapped within 0.5 cM, confirming the correlation between the bands observed and the markers designed (data not shown).

BAC fingerprint data (www.genome.arizona.edu/fpc/WebAG-CoL/maize/WebFPC) were used to physically locate all six BACs. This step also confirmed the RJM mapping results and helped estimate the position of the remaining three BACs for which no RJM marker could be found: AC148172 in bin 4.05, AC148161 in bin 5.03, and ac148159 in bin 5.05 (**SI Table 9**).

Positions of the centromeres and knobs located on maize chromosomes 1, 4, 5, 6, and 7 were compared with the position of the six BACs (SI Table 9). Results of this comparison reveal that AC147789 is located $>5 \mu\text{m}$ from the centromere and knob 1S described in traditional maize varieties (27), AC148172 is near ($<1 \mu\text{m}$) a centromere, AC148161 is located far ($\approx 9 \mu\text{m}$) from any centromere, AC148159 is located $\approx 2 \mu\text{m}$ from knob 5L, and

AC147809 appears to be located near knob 6L1, described in traditional maize varieties and Mexican teosintes (27) (the mapping interval spans the knob region). For BAC AC148081, the Inter-mated B73 $\times$ Mo17 (IBM) mapping interval spans a centromere, but the interval is quite large and marker phi034/cyp6 (chromosome 7S) is tightly linked to the BAC. Therefore, it appears that BAC AC148081 is located on chromosome 7S, at least 3 μ m from the centromere.

Hence, of the six gene-free BACs analyzed, one (AC148172) is located in a pericentromeric region, one (AC148159) is located close to a B73 knob, and one (AC147809) is located close to a knob that is present in traditional maize varieties and Mexican teosintes (27) but not visible in B73 by *in situ* hybridization (28), whereas the three other BACs are from chromosomal locations that are distant to known heterochromatic regions.

Discussion

The Structure of the Maize Genome. To sequence large plant genomes in a cost-effective manner, it is appropriate to first evaluate genome composition and structure, then choose the best sequencing strategy. Our results demonstrate that the annotation of a small set of randomly selected BACs is an effective way to evaluate the key properties of a large and complex plant genome, such as total gene number, amount of repetitive DNA, and gene distribution.

Because all large plant genomes contain numerous transposable element (TEs) that may be difficult to identify and differentiate from genes, it is challenging to obtain an accurate estimation of gene number (18). Even in the relatively simple rice genome, the estimation of gene number from draft whole-genome sequence and finished individual chromosomes has varied from $\approx 32,000$ to $\approx 70,000$ (29–33). The estimation of gene number in maize is particularly challenging, because the complete genome sequence is not available, and the majority of the genome consists of nested LTR retrotransposons (15). The lower boundary of our estimation of maize gene number (37,000) is similar to the gene number estimated in the nearly completed rice genome sequence ($\approx 32,000$) (34). This is consistent with the fact that, even though maize has a fairly recent tetraploid history (35), it is approaching a diploid status because 50–90% of the duplicated copies of genes have been deleted at least partially in one of the homoeologous regions (4, 6, 36). In a recent study, Haberer *et al.* (14) predicted 42,000–56,000 genes in the maize genome and concluded that 22% of their 100 randomly chosen maize BACs were missing genes. The differences in our analyses, for instance the $\approx 30\%$ of maize BACs that we found to be lacking in verified genes, is an outcome of the more conservative gene annotation in the current study. We found that

versity of Arizona, and Rutgers University (www.broad.mit.edu/annotation/plants/maize/randomclones.html) (14). We selected only those 80 BACs with sizes >80 kb for annotation. Six of these BACs were excluded later, because their predicted size changed dramatically upon subsequent assemblies recorded in GenBank, leaving 74 BACs covering 12.2 Mb (0.5%) of the maize genome for the analysis.

Annotation of LTR Retrotransposons and Other Repeats. Known repeats were identified by comparing BAC sequences with the TIGR maize repeat database (20) and an in-house database of maize repeats with cross-match (www.phrap.org/phredphrap-consed.html). LTR retrotransposons were manually identified by structural features such as pairs of LTRs, TSDs, a primer binding site and a polypurine tract.

Gene Annotation and Distribution Analysis. BAC sequences were first subject to *ab initio* gene prediction with FGENESH (www.softberry.com) using the matrix for monocot plants. Gene models that were integral parts of known repeats were removed from further analysis. The remaining gene models were used as query sequences to search the National Center for Biotechnology Information nonredundant protein database, the *Arabidopsis* protein database (<ftp://ftp.tigr.org/pub/data/a.thaliana>, version 5.0), all BAC sequences from the international rice genome sequencing group (21), and the TIGR plant gene indices (56). All BLAST hits were manually evaluated to determine whether a gene model is likely to be a real gene based on *e* value, alignment with the query, hit annotation, and number and evolutionary range of species with sequences that exhibited homology.

A χ^2 test was used to test the null hypothesis that genes are uniformly distributed in the maize genome. Because one requirement of a χ^2 test is that cell (group) frequency must be at least 5 (57), each BAC was randomly assigned to one of the 10 groups (a randomized block design) and then the expected and observed gene number in each group was used to calculate the χ^2 value.

Gene Fragment Identification. Each gene model was used as a query to search *Arabidopsis* proteins (TIGR annotation version 5.0) and rice proteins derived from full length cDNAs. The best match from either *Arabidopsis* or rice was used as query to search maize BAC DNA sequences to ensure that homologs were not shortened by annotation errors. Each alignment was manually evaluated. Maize DNA sequences upstream and downstream of the alignment were also checked to make sure that any fragment identified was not caused by sequence gaps or BAC boundaries. A gene fragment was defined by the subjective criteria that the maize gene model could cover only part ($\leq 70\%$) of the homologs in *Arabidopsis* or rice, and the alignment between the query and hit would have a high identity that was maintained at both ends of the alignment. Hence, many gene fragments and other pseudogenes would be missed by these criteria.

Colinearity with Rice. Gene models from a single BAC were used as queries to search against rice BAC sequences. If two or more gene models had significant (BLAST *e* value $\leq 10^{-5}$) matches on the same rice BAC, they were judged as having colinearity with rice. Because the randomly selected maize BAC sequences were in draft stage and usually consisted of unordered pieces, it was not possible to compare the order and orientation of some genes.

Choice and Annotation of Gene-Free BACs. From the 21 BACs for which no verified gene was annotated, six were chosen for further characterization. These BAC sequences are present in GenBank under accession nos. AC147789, AC148081, AC148159, AC148172, AC148161, and AC147809 (SI Table 2). Sequence analysis was based on the sequences available in GenBank in February 2005. For

all BACs except AC148161 (which contains three contigs), the sequence is represented in one contig.

Annotation of the six BACs sequences was performed in several steps, each step allowing the characterization of a certain type of repeat, such as known TEs, unknown LTR retrotransposons, tandem repeats, and previously undescribed repeats (see SI Text for details of these steps). After each step, the annotated regions were removed from the sequence to allow the characterization of other repeats in which the first annotated elements may have inserted. For each characterized TE, special care was given to detect the presence of the TSD.

Primers Design and PCR. Primers were chosen to flank the junction between two LTR retrotransposons, based on annotation of the BAC sequences from maize line B73. The TSD sequence of the inserted LTR retrotransposon was chosen to be in the middle of the PCR product, with a product size ranging from 500 to 600 bp. Primers were designed by using the Primer3 software (<http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.www.cgi>), with a GC clamp of 2. Primer sequences are listed in SI Table 10.

DNAs were extracted from 2-week-old plants from maize lines B73 and Mo17 using the CTAB method described by M. Frohlich (<http://fgp.bio.psu.edu/fgp/methods/method.CTAB.html>) without conducting the CsCl purification steps. PCRs were performed in a 20- μ l total volume, containing 100 ng of DNA, 2 μ l of $10\times$ Taq polymerase buffer (Roche, Indianapolis, IN), 0.16 μ l of 25 mM dNTPs (Roche), 3 μ l of 4 μ M solutions of each primers (Invitrogen), 0.16 μ l of 5 units/ μ l of Taq (Roche), completed with nuclease-free water (Fisher Scientific, Pittsburgh, PA). They were carried out with the following steps: denaturation at 94°C for 3 min, followed by 35 touchdown cycles with denaturation at 94°C for 30 sec, annealing with temperature decreasing from 62°C to 55°C in 30 sec, and extension of 1 min at 72°C, followed by a final extension step of 3 min at 72°C. In each cycle, the increase of the temperature between the annealing and the extension steps was carried out with a ramp of 1°C per sec.

Estimation of LTR Retrotransposon Insertion Dates. The insertion date of each LTR retrotransposon harboring two LTRs was estimated following the method described in ref. 22. The divergence between LTRs was computed by MEGA version 3.0 (58), using the Kimura 2 parameter distance (59) that corrects for both homoplasy and differences in the rates of transition and transversion. The substitution rate used was 1.3×10^{-8} substitution/site per year, as suggested by Ma and Bennetzen (60).

Comparison of Gene-Free and Gene-Containing Regions. Six gene-containing BACs, GenBank accession nos. AF123535 (15), AF448416 (24), AY555142 (7), AY664413, AY664414, and AY664415 (23), were chosen to compare to the six gene-free BACs in this study. LTR retrotransposons copy number and insertion dates were recalculated to avoid differences caused by different analytical procedures.

Variation in copy number between the gene-free and -containing BACs was determined for the 10 most-abundant LTR retrotransposons. For each element, two methods were used: (i) a *P* value was estimated by comparing the observed copy number in the gene-free regions to a binomial distribution with *n* trials (*n* being the total number of copies found in both gene-free and -rich regions for this element), and a probability of a copy to be in a gene-free region estimated by $P = N_1/N_1 + N_2$, with *N*₁ and *N*₂ the total number of copies found for all elements in the gene-free and gene-rich regions, respectively; (ii) a *P* value was estimated by using a nonparametric test. For each element, a theoretical distribution of the copy number in gene-free region was built by using 999 simulations in which *N*_{GF} copies were randomly picked without replacement from the total number of copies observed in both gene-rich and gene-free regions (*N*_{GF} + *N*_{GR}), with *N*_{GF} and *N*_{GR} the observed number of copies

in gene-free and -rich regions, respectively. The *P* value was estimated by comparing the observed number of copies in the gene-free regions to this distribution. A Bonferroni correction was used to correct for the 10 individual tests that were made (one for each of the 10 most-abundant elements) out of the same sample.

BAC Mapping and Comparison to Centromere and Knob Locations. BACs with polymorphic markers were genetically mapped in a subset of 94 inbred lines from the IBM maize mapping population (61) using MapMaker 2.0. In addition, all BACs were mapped by using BAC fingerprint data retrieved from the Arizona Genomic Institute Web site (www.genome.arizona.edu/fpc/WebAGCoL/maize/WebFPC). These data were then used to locate the corresponding bin, using the location of these markers on the IBM2 2004 neighbors map.

To precisely position the six BACs on the genome, positions of the markers on the IBM2 2004 map were converted to physical locations, using the Morgan2McClintock translator (62) (available at www.lawrencelab.org/Morgan2McClintock). This allowed prediction of the physical position of the markers as a fraction of the distance on the arm from the centromere. Similarly, knob-containing bins were physically mapped on the chromosomes, allowing the estimation of the distance between the six BACs and these heterochromatic regions.

We thank P. SanMiguel for providing an unpublished maize retrotransposon data set, C. Lawrence for advice on locating maize knobs, L. Yang for Helitron analysis, and H. Zheng and J. Estill for help with statistical analyses. This research was supported by National Science Foundation Grant DBI-0501814 (to J.L.B.).

1. The Arabidopsis Genome Initiative (2000) *Nature* 408:796–815.
2. Ausubel FM (2002) *Plant Physiol* 129:394–437.
3. Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL (2002) *Genetics* 162:1389–1400.
4. Ilic K, SanMiguel PJ, Bennetzen JL (2003) *Proc Natl Acad Sci USA* 100:12265–12270.
5. Bennetzen JL, Ma J (2003) *Curr Opin Plant Biol* 6:128–133.
6. Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, Llaça V, Tanyolac B, Park YJ, Jeong Y, Bennetzen JL, et al. (2004) *Genome Res* 14:1924–1931.
7. Ma J, SanMiguel P, Lai J, Messing J, Bennetzen JL (2005) *Genetics* 170:1209–1220.
8. Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, et al. (2003) *Science* 302:2118–2120.
9. Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) *Science* 302:2115–2117.
10. Bennetzen JL (2003) *Proc Tenth Int Wheat Genet Symp* 1:215–220.
11. Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) *Proc Natl Acad Sci USA* 102:19243–19248.
12. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S (1993) *Nature* 366:265–268.
13. Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KFX, et al. (2004) *Proc Natl Acad Sci USA* 101:14349–14354.
14. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Birren B, et al. (2005) *Plant Physiol* 139:1612–1624.
15. 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.
16. Meyers BC, Tingley SV, Morgante M (2001) *Genome Res* 11:1660–1676.
17. SanMiguel P, Bennetzen JL (1998) *Ann Bot* 82:37–44.
18. Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) *Curr Opin Plant Biol* 7:732–736.
19. Sabot F, Guyot R, Wicker T, Chantret N, Laubin B, Chalhoub B, Leroy P, Sourdille P, Bernard M (2005) *Mol Genet Genomics* 274:119–130.
20. Ouyang S, Buell CR (2004) *Nucleic Acids Res* 32:D360–D363.
21. Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, Antonio BA, Baba T, et al. (2005) *Nature* 436:793–800.
22. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) *Nat Genet* 20:43–45.
23. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) *Plant Cell* 17:343–360.
24. Fu H, Dooner HK (2002) *Proc Natl Acad Sci USA* 99:9573–9578.
25. Ananiev EV, Phillips RL, Rines HW (1998) *Genetics* 149:2025–2037.
26. Ma J, Devos KM, Bennetzen JL (2004) *Genome Res* 14:860–869.
27. McClintock B, Kato Y, Blumenschein A (1981) *Chromosome Constitution of Races of Maize* (Colegio de Postgraduados, Chapingo, Mexico).
28. Kato A, Lamb JC, Birchler JA (2004) *Proc Natl Acad Sci USA* 101:13554–13559.
29. Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. (2002) *Science* 296:92–100.
30. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, et al. (2002) *Science* 296:79–92.
31. Feng Q, Zhang YJ, Hao P, Wang SY, Fu G, Huang YC, Li Y, Zhu JJ, Liu YL, Hu X, et al. (2002) *Nature* 420:316–320.
32. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu JZ, Niimura Y, Cheng ZK, Nagamura Y, et al. (2007) *Nature* 420:312–316.
33. Yu Y, Rambo T, Currie J, Sasaki C, Kim HR, Collura K, Thompson S, Simmons J, Yang TJ, Nah G, et al. (2003) *Science* 300:1566–1569.
34. Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskewich R, et al. (2007) *Genome Res* 17:175–183.
35. Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaça V, Bennetzen JL, Messing J (2004) *Genome Res* 14:1916–1923.
36. Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M (2004) *Genetics* 166:935–945.
37. Ramakrishna W, Emberton J, Ogden M, SanMiguel P, Bennetzen JL (2002) *Plant Cell* 14:3213–3223.
38. Lai J, Li Y, Messing J, Dooner HK (2005) *Proc Natl Acad Sci USA* 102:9068–9073.
39. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) *Nat Genet* 37:997–1002.
40. Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan LL, Shiloff BA, Bennetzen JL (2001) *Plant Physiol* 125:1342–1353.
41. Avramova Z, Tikhonov A, SanMiguel P, Jin YK, Liu CN, Woo SS, Wing RA, Bennetzen JL (1996) *Plant J* 10:1163–1168.
42. Devos KM, Beales J, Nagamura Y, Sasaki T (1999) *Genome Res* 9:825–829.
43. Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y (2002) *Genome Res* 12:1792–1801.
44. Liu H, Sachidanandam R, Stein L (2001) *Genome Res* 11:2020–2026.
45. Bennetzen JL, SanMiguel P, Chen MS, Tikhonov A, Francki M, Avramova Z (1998) *Proc Natl Acad Sci USA* 95:1975–1978.
46. Brunner S, Pea G, Rafalski A (2005) *Plant J* 43:799–810.
47. Vitte C, Bennetzen JL (2006) *Proc Natl Acad Sci USA* 103:17638–17643.
48. Devos KM, Brown JKM, Bennetzen JL (2002) *Genome Res* 12:1075–1079.
49. Ma J, Bennetzen JL (2006) *Proc Natl Acad Sci USA* 103:383–388.
50. Peacock WJ, Dennis ES, Rhoades MM, Pryor AJ (1981) *Proc Natl Acad Sci USA* 78:4490–4494.
51. Dennis ES, Peacock WJ (1984) *J Mol Evol* 20:341–350.
52. Viotti A, Privitera E, Sala E, Pogna N (1985) *Theor Appl Genet* 70:234–239.
53. Ananiev EV, Phillips RL, Rines HW (1998) *Proc Natl Acad Sci USA* 95:10785–10790.
54. Dawe RK, Hiatt EN (2004) *Chromosome Res* 12:655–669.
55. Yim YS, Davis GL, Duru NA, Musket TA, Linton EW, Messing JW, McMullen MD, Soderlund CA, Polacco ML, Gardiner JM, et al. (2002) *Plant Physiol* 130:1686–1696.
56. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) *Nucleic Acids Res* 29:159–164.
57. Bhattacharyya GK, Johnson RA (1977) *Statistical Concepts and Methods* (Wiley, New York).
58. Kumar S, Tamura K, Nei M (2004) *Brief Bioinform* 5:150–163.
59. Kimura M (1980) *J Mol Evol* 16:111–120.
60. Ma J, Bennetzen JL (2004) *Proc Natl Acad Sci USA* 101:12404–12410.
61. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A (2002) *Plant Mol Biol* 48:453–461.
62. Lawrence CJ, Seigfried TE, Bass HW, Anderson LK (2006) *Genetics* 172:2007–2009.