National Science Foundation-Sponsored Workshop Report. Maize Genome Sequencing Project[1]

Jeffrey L. Bennetzen, Vicki L. Chandler*, and Patrick Schnable

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907 (J.L.B.); Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721 (V.L.C.); and Department of Agronomy, Iowa State University, Ames, Iowa 50011 (P.S.)

In response to a mandate from the maize (*Zea mays*) genetics community, a National Science Foundation-sponsored workshop was held in St. Louis on July 2, 2001, to discuss technical approaches for a Maize Genome Sequencing Project. This workshop included academic, governmental, and industrial scientists with expertise in analysis of human, animal, plant, and microbial genomes as well as observers from federal funding agencies and those representing U.S. corn growers. The participants of the St. Louis workshop were in unanimous agreement that sequencing all maize genes and placing them on a cross-referenced physical-genetic map was an extremely worthy, feasible, and timely goal that can be achieved at a reasonable cost with existing technologies.

Maize is one of the most important economic crops in the United States. It is also the best-studied and most tractable genetic system among the cereals, making it the premier model system for studying this important group of crops. A serious limitation to continued advances in both basic and applied research in maize is the lack of a comprehensive understanding of gene content and gene organization within the maize genome. Maize gene sequencing and functional analysis will help elucidate the molecular basis of agronomically important traits and thereby facilitate improvements in maize and other crop species. These agronomic improvements will have enormous impacts on mankind through improving human health, increasing energy production, and protecting our environment. The production of novel compounds in plants, including industrial feed stocks, biofuels, and medicinal compounds will increase the demand for corn and thereby directly benefit the agricultural community. The production of nutritionally enhanced foods that are safer and less allergenic than the foods we eat today will directly benefit consumers.

The maize genome is approximately the same size, and at least as complex, as that of the previously sequenced human genome. Various technical approaches for sequencing the maize genome were discussed. All prior genome projects have employed either a minimal tiling path or whole genome shotgun sequencing followed by computer assembly approaches. The highly repetitive nature of the maize genome (large numbers of dispersed highly similar repeats) raised concerns regarding whether the data resulting from a whole genome shotgun sequencing project could be properly assembled into a complete genome sequence with existing bioinformatic tools. This concern, coupled with the high cost of a whole genome shotgun sequencing project on a genome the size of maize, has led the maize community to develop a third option whereby several cutting-edge technologies could be employed to identify, sequence, and assemble all of the genes of the complex maize genome. Such an approach would focus sequencing resources on the genic regions while minimizing the sequencing of the large repetitive component of the maize genome. This gene-enriched sequencing will provide a paradigm for the efficient and cost-effective sequencing of other large, complex genomes of plants and animals that would otherwise be prohibitively expensive to solve by whole genome sequencing. A majority of the participants concluded that a Maize Genome Sequence Project that focused on the gene-rich, low-copy fraction of the genome would be most appropriate. A minority of the participants felt that a full genome shotgun sequence would be the best sequencing approach. A third approach of sequencing gene-rich bacteria artificial chromosomes (BACs) received support from two participants.

We invite the community of plant biologists to read the report below and offer your comments, views and suggestions. Please send your remarks via e-mail to the following address (MaizeGSP@aspb.org) and they will be posted at the *Plant Physiology* web site (http://www.aspb.org). We are looking forward to a discussion of the important issues raised by this report.

WORKSHOP SUMMARY

In response to a mandate that arose from electronic communications within the maize genetics community and at formal sessions convened to address this

[1] This workshop was supported by a grant from the National Science Foundation (DBI–0126620 to V.L.C.).

* Corresponding author; e-mail chandler@ag.arizona.edu; fax 520–621–7186.

www.plantphysiol.org/cgi/doi/10.1104/pp.010817.

issue during two international meetings in early 2001, a National Science Foundation-sponsored workshop was held in St. Louis on July 2, 2001, to discuss technical approaches for sequencing the maize genome (see list of participants and full report at http://www.agron.missouri.edu/cooperators.html). The following questions were addressed at the workshop:

Should the Maize Genome Be Sequenced Now?

With finite resources and many important goals for the life sciences, the question must be asked whether the maize genome should be sequenced and, if so, should it be sequenced now. Recent completion of human, animal, and plant genome sequences have demonstrated that genomic sequencing is the most comprehensive route to gene discovery and the first step toward identifying the function of every gene. The completion of the Arabidopsis genome sequence (The Arabidopsis Genome Initiative, 2000) and the ongoing sequencing of the rice (*Oryza sativa*) genome have been exceptionally informative as to gene content in these species, which has strengthened and will continue to strengthen research programs throughout plant biology. It is clear Arabidopsis will continue as the most important model species for many aspects of basic plant biology. Current data indicate, however, that a comprehensive understanding of gene function in crop plants such as maize that are only distantly related to Arabidopsis will require thorough investigations of their own genomes to identify all of their genes and to determine the functions of those genes.

Comparative analyses have demonstrated that genome rearrangements at the local chromosome level can be incredibly frequent. For example, Arabidopsis shows essentially no colinearity with two very important model monocot plants, rice and maize (Bennetzen et al., 1998; Devos et al., 1999). Comparative analyses of several grass genomes, including maize, rice, sorghum (*Sorghum bicolor*), wheat (*Triticum aestivum*), and barley (*Hordeum vulgare*), have revealed extensive conservation of gene content and order at the level of the overall genetic map (Gale and Devos, 1998). However, as many as 15,000 local rearrangements differentiate the maize and rice genomes (Tikhonov et al., 1999; Dubcovsky et al., 2001). Thus, current results suggest that rice will often be a too distant model to facilitate rapid map-based cloning in maize and other important cereals such as wheat, barley, sorghum, and oats (*Avena sativa*). Moreover, there may be cereal genes missing from Arabidopsis and vice versa. If all maize genes were identified and ordered on the physical-genetic map, comparisons to the completely sequenced rice and Arabidopsis genomes would be much more informative. Access to such informed comparisons would facilitate the identification of common genes having known ancestral relationships. This would allow the knowledge gained from gene function in any of these species to be used in the study of other cereal and grass species.

All other genome projects, with the exception of the human and mouse genome sequencing projects, have involved model organisms with relatively small genome sizes. These previous projects have employed either a minimal tiling path or whole genome shotgun sequencing followed by computer assembly approaches. Several researchers within the maize community have developed a third option whereby several cutting-edge technologies can be employed to identify, sequence, and assemble all of the genes of the complex maize genome. Such an approach will focus sequencing resources on the genic regions (approximately 50,000 genes representing 10%–15% of the maize genome) while minimizing the sequencing of the large repetitive component of the maize genome. This gene-enriched sequencing will provide a paradigm for the efficient and cost-effective sequencing of other large, complex genomes of plants and animals that would otherwise be prohibitively expensive to solve by whole genome sequencing.

Analysis of this important model species and crop plant will greatly enhance our understanding of plant development, gene regulation, stress tolerance, transposable element function, genome evolution, and important agronomic traits. A strong and vibrant network of academic and industrial researchers has produced numerous tools, such as physical and genetic maps, forward and reverse genetics, and plant transformation, all of which facilitate functional characterization of maize genes.

A serious limitation to continued advances in both basic and applied research in maize is the lack of a comprehensive understanding of gene content and gene organization within the maize genome. The elucidation of all the maize genes and their placement on a cross-referenced physical-genetic map would broadly empower the entire maize community, leading to a rapid increase in the ability of academic and industrial researchers to understand the functions of maize genes, e.g. by associating phenotypes with the specific genes responsible for those phenotypes. The resulting data will impact not only research on maize, but also all aspects of basic plant biology. Maize gene sequencing and functional analysis will help elucidate the molecular basis of agronomically important traits and thereby facilitate improvements in maize and other crop species (Bennetzen and Freeling, 1993; Freeling, 2001). These agronomic improvements will have enormous impacts on humankind through improving human health, increasing energy production, and protecting our environment. Examples of these kind of agronomic improvements include creating plants that are more resistant to drought, salt, pathogens, and other stresses. Access to such crops will reduce the use of environmentally toxic chemicals. It will also be possible to produce novel compounds in plants, including industrial

feedstocks, biofuels, and medicinal compounds that will increase the demand for corn and thereby directly benefit the agricultural community. The knowledge and understanding that arises from the Maize Genome Sequencing Project will allow scientists to develop nutritionally enhanced foods that are safer and less allergenic than the foods we eat today, thus directly benefiting consumers.

What Is the Nature and Organization of the Maize Genome?

The first topic of the workshop was a discussion of our current knowledge of the organization of the maize genome and its comparison with other closely related species. This discussion was informed by experiments that included phylogenetic analyses (Kellogg, 2001), $C_o t$ analyses, the complete genome sequence of approximately 150 maize genes (see GenBank), the complete sequence of several BAC clones spanning known maize genes and similar regions in other cereals, and shotgun sequencing of approximately 5000 random maize genomic clones.

The Poaceae, which include most of the major food crops, are well separated from dicots. Their phylogeny has been extensively studied and is well understood (Kellogg, 2001). Evolutionary divergence of the grass family occurred primarily within the past 50 to 60 million years. Maize and sorghum diverged about 20 million years ago, whereas rice and maize diverged about 50 million years ago. The approximately 430-Mb genome of rice is one of the smaller genomes of the Poaceae. In contrast, the maize genome is about the size of the human genome at 2,500 Mb; the barley genome is 4,900 Mb; and the hexaploid wheat genome is 16,000 Mb. Comparative mapping with conserved molecular markers has revealed long stretches of conserved gene order with some major exceptions (Gale and Devos, 1998). Importantly, however, comparisons of BACs that contain a common conserved gene derived from these different species indicate that local rearrangements are very common (Tikhonov et al., 1999; Dubcovsky et al., 2001). Many of these rearrangements are tandem gene duplications and small inversions within the same chromosomal regions. However, others include translocations of one or several genes between chromosomes. On average, similar regions of the rice and maize genome have approximately one to two rearrangements per BAC, suggesting that there are more than 15,000 rearrangements between these two genomes. Most of the maize genome is repetitive; genes are estimated to represent at most 15% of the genome. The bulk of the genome is made up of retrotransposons (approximately 65%–70%; Bennetzen, 1996). These elements transpose via an RNA intermediate and are primarily responsible for the large differences in genome size among the different grass species. Most of this repetitive DNA in maize is represented by a few retrotransposon elements that have amplified to large numbers (10,000–30,000 each) relatively recently in evolutionary time (SanMiguel et al., 1996). Several classes are very homogenous, including many members that exhibit >90% sequence identity, some with 97% or greater sequence identity. The retrotransposons are dispersed throughout the genome. Inverted repeat transposable elements that move via a DNA intermediate, satellite DNA, and centromere and telomere-specific repeats represent the remainder of the non-genic fraction of the genome. The inverted repeat transposable elements tend to be relatively low copy (tens to hundreds of copies of each family), dispersed, and closely associated with genes. Thus, the picture of the maize genome is one of genes (often associated with inverted repeat transposons within introns or nearby), interspersed with large islands of nested retrotransposons inserted between genes (Bennetzen et al., 1998). Active genes tend to be relatively under-methylated as compared with the repetitive sequences in the intergenic regions (Bennetzen et al., 1994).

Typically, maize genes contain small introns, such that the transcribed region spans on average approximately 4 to 5 kb. The handful of fully sequenced BACs contain anywhere from two to 16 genes. The BACs that have been sequenced to date were characterized because they contained known genes or were tightly linked with specific genetic traits. Thus it is likely that there are regions of the genome with an even lower gene density. The general conclusion was that not enough maize BACs have been characterized to determine an average gene density per BAC for the genome as a whole. There are currently funded plant genome projects sequencing maize BACs, and it was suggested that a few randomly selected BACs should be targeted for sequencing immediately.

What Is the Status of the Maize Physical and Genetic Maps?

A physical map is essential for genome analysis, independent of whether a minimal tiling path, whole shotgun sequencing, or gene enrichment strategies are pursued. One of the major uses of a genome sequence is the ability to perform efficient map-based cloning of genes and to associate candidate genes with important biological or agronomic traits. A key aspect of this approach is a well-integrated physical and genetic map. The strategy and progress of a National Science Foundation-funded Plant Genome grant to achieve this goal was discussed. A maize physical map is being generated by isolating approximately 400,000 BAC clones using three different restriction digestions (approximately 26-fold coverage), fingerprinting the BAC clones via the electrophoresis of *Hin*dIII restriction digestion fragments through agarose gels, analyzing the resulting gel images using IMAGE software, and then using

fingerprinted contigs (FPC) software to assemble the contigs. The goals are to place 14,600 markers on the map, resulting in one marker every 171 kb, and to place 4,800 markers anchored to the genetic map, resulting in one anchored marker every 520 kb. The BAC libraries have been prepared. Currently, fingerprint and band calling data are being generated and entered into FPC. The statistics are posted to a public web page and updated regularly. As of June, 2001, data have been generated and entered into FPC for 78,827 BAC clones, representing 4× coverage. As expected, as more data are entered the number of contigs is dropping. It is estimated that all data will be generated by spring of 2002 and then manual editing will begin. With the human map it took about a year to reduce 7,000 contigs to 700. Less time may be required for maize. The human genome physical map was edited continuously as data were added to a final density of 15× coverage. The plan with maize is to add most of the data prior to manual editing and to do 26× coverage. Thus, current projections are that the physical map will be complete by spring of 2003. New improvements in FPC were also discussed, as was progress toward completing the rice physical map. Potential problems associated with features of maize (genome duplications and the number of highly similar retrotransposons) that might impact contig assembly were discussed.

There was extensive discussion about whether BAC end sequencing would be valuable. The highly repetitive nature of the maize genome indicates that only approximately 10% of the BAC end sequences would be in low-copy sequences; consequently these ends would not be particularly informative in terms of gene identification. Could such sequences aid contig assembly? Several participants felt that the BAC end sequences, combined with the fingerprints, and expressed sequence tag anchoring would allow contigs to be built efficiently and provide the community with immediate access to a sample sequencing of virtually every region of the genome. This could provide the foundation for a complete sequence of the genome or allow one to pick gene rich contigs to sequence. Other participants were more enthusiastic about fingerprinting and BAC end sequencing of clones that were derived from libraries constructed with restriction enzymes with methylation sensitive enzymes like *Sal*I. Current estimates are that this approach would yield BAC end sequences that were anchored in genes or gene-flanking DNA between 25% and 75% of the time. By a solid majority, the participants felt that some BAC end sequencing should begin immediately, perhaps involving up to $4 million in funding commitment. There was absolute agreement that a well-integrated physical-genetic map was essential for the maize genome sequencing initiative, independent of the sequencing strategy chosen.

What Is the Best Strategy for Sequencing the Maize Genome?

Various technical approaches for sequencing the maize genome were discussed. These included whole genome shotgun sequencing and several methods for enriching for genic regions followed by low-cost shotgun sequencing. The participants were in agreement that several of these could be successfully used to sequence the maize genome, but there was significant discussion as to the relative efficiency of each and the synergy between approaches.

The highly repetitive nature of the maize genome and the observation that many of the abundant repeats are highly conserved, raised concerns regarding whether the data resulting from a whole genome shotgun sequencing project could be properly assembled into a complete genome sequence with existing bioinformatic tools. This concern, coupled with the high cost of a whole genome shotgun sequencing project on a genome the size of maize, led a majority of the participants to conclude that a project that focused on the gene-rich, low-copy fraction of the genome would be more appropriate. The latter approach also has the potential of providing useful gene sequence information to the community at the earliest possible date. A minority of the participants felt that a full genome shotgun sequence would be the best sequencing approach.

There was general agreement among participants that various technologies exist for selectively sequencing the gene-rich fraction of the maize genome. These technologies include using the differences in DNA methylation between genes and the highly repetitive component of the genome as a way to enrich for genes in a shotgun DNA sequencing project (Rabinowicz et al., 1999). Alternatively, DNA fragments can be selected that contain open-reading frames or that have high $C_o t$ values to enrich for genes. There was also discussion of the fact that maize mobile elements that transpose via a DNA intermediate are preferentially associated within or near genes. Thus, strategies to identify sequences adjacent to DNA transposons should also enrich for genes. After extensive discussion, there was general agreement that it is not yet clear which technologies will be the most appropriate for efficiently isolating genes, recovering large numbers of genes, and efficiently placing these genes on the physical-genetic map. It is also understood that there may not be a single method that is optimal for achieving all of these goals as each method has both strengths and biases.

The group recommended that it be left up to proposal applicants and reviewers to determine which approaches are most appropriate for reaching the ultimate goal of sequencing and mapping all maize genes. Information that should be required in proposals include an estimation of what percentage of the maize genes and what parts of the genes will be

sequenced via the proposed approach and the cost of doing so on a per gene basis. Applicants should provide data on average sequencing accuracy (phred score) and length. Applicants should also indicate how their method will further the gene mapping goal and at what cost on a per gene basis. These assessments should be made using publicly available data to allow review by independent scientists, and the methods used should be described in sufficient detail such that reviewers can make comparisons among methods. Given the likelihood that it will be necessary to use more than one approach to achieve the overall project goals, applicants should also describe how their proposed strategy complements other approaches that may be proposed. Applicants should indicate whether and how their strategy could serve as a model for the sequencing of other large, complex plant and animal genomes. Finally, applicants should provide a list of deliverables and a detailed timetable for distribution of their results and biological materials, as relevant.

Which Genotype Should Be Sequenced?

The participants were in unanimous agreement that the inbred line B73 should be the primary focus of the sequencing project. This inbred line was the source of the BAC libraries that are being used to develop a framework physical map and of many of the public expressed sequence tags. Moreover, this public inbred is a good representative of commercially important maize germplasm. It was recognized that sequencing other genotypes would provide single-nucleotide polymorphisms, certainly a worthwhile goal. However, the fact that most maize inbreds have around one nucleotide sequence polymorphism every 70 bp, thus confusing sequence assembly efforts, emphasizes the importance of primarily using only one DNA source.

How Should Maize Genome Sequence Data Be Disseminated?

The participants agreed that it is essential that data resulting from the Maize Genome Sequencing Project be freely available to the entire research community in the shortest time possible. Applicants should describe their data release policies. Alternatively, a uniform set of data release policies should be described by the appropriate funding agencies, and the participants at this meeting thought that immediate release of sequencing traces was a reasonable strategy. Minimum standards for quality of data should be established. Applicants should agree not to seek "reach-through" protection of intellectual property generated as a result of the maize genome sequencing project. This group recommends that a willingness to promptly submit gene sequence and mapping data to the public database GenBank should be an essential criterion for all project participants. The group discussed the need for the timely development of a database that integrates all of the components of the Maize Genome Sequencing Project with each other and with other existing pertinent data sets and databases. It was agreed that it will be necessary to provide project funding specifically for the development of such a database.

What Would It Cost to Sequence the Maize Genome? and How Long Would It Take?

The participants concurred that the goal of sequencing all of the genes in the maize genome and placing these on the integrated physical and genetic map could be pursued by a combination of technologies that would cost about $52 million. The breakdown of estimated costs would be:

- Library construction and evaluation, $3 million
- BAC-end sequencing, $4 million
- 10-Fold redundant sequencing of the gene-rich and low-copy-number regions, $34 million
- Locating all of the genes on an integrated physical-genetic map, $8 million
- Establishing a comprehensive database system, $3 million.

Given this full set of resources, the genes in the maize genome could be sequenced in 2 years from the initiation of the project. Year 3 of the project would involve mapping all of the identified genes to the integrated physical-genetic map of maize, completion of annotation, and integration of maize data with that from other species. For the gene-enrichment approach, a minimum of $10 to $15 million would be required in the first year, for library generation and evaluation, low-redundancy sequencing to compare different gene-enrichment techniques, development of cost-effective mapping strategies, and initiation of the database system. The bulk of the funding would be required in year 2 when the majority of the DNA sequencing would occur. The participants determined that sufficient DNA sequencing capacity is available in the United States for completing the sequencing activity in 12 to 18 months. The informatics components could also be completed within 2 to 3 years, especially with tight coordination with other developing databases.

Alternative approaches and timelines were also discussed. If the same approach described above were pursued over a 4-year timeframe, it is anticipated that costs would increase by about $4 million as a consequence of loss of efficiencies in library construction, gene sequencing, and mapping. If funding were available earlier, the sequence could be completed in less than 3 years, thereby providing some cost savings and an earlier availability of the sequence to a scientific community that could use the information now. If a full genome shotgun approach were to be undertaken, a 10-fold redundant sequence

data set would cost about $135 million at today's rates, or a total of about $150 million with the added database, BAC end, and mapping activities. Some participants felt lower ("draft") coverage (approximately four to 5 times redundancy) would identify most of the genes. Using such an approach, a full genome shotgun sequence could generate a 5× draft for about $70 million in 1 to 2 years. A similar gene-enriched "draft" approach could generate a 5× draft for about $25 million. However it should be noted that, despite the enthusiasm for "draft" sequences in the animal genome community, drafts provide a data set that cannot be assembled even within many genes and may require more than $100 million in additional expenditures to complete the sequence.

Who Should Participate in the Maize Genome Sequencing Project?

Because of the immediate impact a maize genome sequencing project would have on basic and mission-oriented research, the Maize Genome Sequencing Project should be administered through the Interagency Program, involving the National Science Foundation, U.S. Department of Agriculture, Department of Energy, and National Institutes of Health. Given its recent experience as the lead agency in the Arabidopsis Genome Project, and its oversight of several plant genome projects producing results with direct impact on the Maize Genome Sequencing Project, the National Science Foundation is the logical choice to be the lead agency. Additional funds should be sought from the U.S. Congress to support this initiative. Proposals should be solicited from the entire community of qualified scientists and U.S. institutions, including existing DNA sequencing centers, federal laboratories, and private organizations. Involvement of international collaborators and industry should be encouraged, as long as a policy of rapid data release to the public without reach-through intellectual property is strictly followed.

WORKSHOP CONCLUSIONS

Sequencing the genes of maize via a new kind of genome project will provide an exceptional opportunity to develop and validate the most advanced technologies for the analysis of complex genomes. The proposed project will serve as a beacon for future efforts to sequence all of the genes of other animals and plants with complex genomes with equal rapidity, efficiency, and low cost. The released information will be an unequalled resource enabling all plant scientists to study and to understand plant genomes and will serve as the foundation for future crop improvement efforts. The community and technologies are available to pursue a Maize Genome Sequencing Project now, and the attendees at this meeting encourage all efforts to initiate this Project at the earliest possible date.

RECOMMENDED TIMELINE

Fiscal Year 2002

Competition to identify groups/centers to carry out the library construction, sequencing, and database development. Implementation of BAC end sequencing.

Fiscal Year 2003

Sequencing and technology development with the goal of generating several hundred Mb of sequence using multiple approaches. Development of efficient strategies to locate sequences to the physical map. Initiate development of the database.

Fiscal Year 2004

Intensive sequencing of the gene-rich component of the maize genome using the best combination of approaches identified in the previous year. Goal is to complete sequencing the genespace of maize at 10× coverage. Continued mapping of identified genes and continued development of the database.

Fiscal Year 2005

Complete mapping of all identified genes and complete the development of the maize genome database.

Received September 6, 2001; accepted September 6, 2001.

LITERATURE CITED

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

Bennetzen JL (1996) The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol **4:** 347–353

Bennetzen JL, Freeling M (1993) Grasses as a single genetic system: genome composition, colinearity and compatibility. Trends Genet **9:** 259–261

Bennetzen JL, SanMiguel P, Chen M, Tikhonov A, Francki M, Avramova Z (1998) Grass genomes. Proc Natl Acad Sci USA **95:** 1975–1978

Bennetzen JL, Schrick K, Springer PS, Brown WE, San Miguel P (1994) Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome **37:** 565–576

Devos KM, Beales J, Nagamura Y, Sasaki T (1999) Arabidopsis-rice: will colinearity allow gene prediction across the eudicot-monocot divide? Genome Res **9:** 825–829

Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol **125:** 1342–1353

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

Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci USA **95:** 1971–1974

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol **125:** 1198–1205

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet **23:** 305–308

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci USA **96:** 7409–7414