The Maize Genome Sequencing Project

Vicki L. Chandler* and Volker Brendel

Plant Sciences Department, University of Arizona, Tucson, Arizona 85721 (V.L.C.); and Iowa State University, Department of Zoology and Genetics, 2112 Molecular Biology Building, Ames, Iowa 50011–3260 (V.B.)

On September 20, 2002, the National Science Foundation (NSF) announced the launch of the Maize Genome Sequencing Project. The momentum for this endeavor has been building within the maize (*Zea mays*) genetics and larger plant science community for several years. Reasons for launching a concerted effort at this time are at least 4-fold. First, advances in DNA sequencing technology now allow faster sequencing at a lower cost than in the past. Second, new high-resolution, high-throughput DNA fingerprinting methods should yield a minimum clone set colinear with the genetic map of the maize genome. Third, promising approaches to preparing fractions of the maize genome enriched for genes have been developed. Fourth, comparative analyses of maize with rice (*Oryza sativa*) or Arabidopsis suggest that the genome sequences of these two species will not be sufficient to understand the precise details of maize gene content and expression. This *Update* reviews the project goals and the expected deliverables deriving from the two funded consortia.

WHY SEQUENCE MAIZE?

The cereals, including maize and rice, account for 70% of food production worldwide, and in addition, maize is an economically important crop in the United States. Maize 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, as well as other monocots. Although cereals are of economic importance and a greater understanding of their genes will have great impact, much interesting biology can also be learned from these species. For example, the recent diversification of the grasses makes them an ideal collective system for dissecting genetic control of morphological and genomic diversity (for review, see Kellogg, 2001). Comparative analyses of several cereal genomes, including maize, rice, sorghum (*Sorghum bicolor*), wheat (*Triticum aestivum*), and barley (*Hordeum vulgare*), have shown extensive conservation of gene content and order at the level of the overall genetic map (Gale and Devos, 1998). However, local rearrangements often interfere with microsynteny, providing evidence for the differentiation of grass genomes (Tikhonov et al., 1999; Keller and Feuillet, 2000; Dubcovsky et al., 2001; Fu and Dooner, 2002; Li and Gill, 2002; Song et al., 2002). These rearrangements include tandem gene duplications, small inversions, and translocations of one or a few genes between chromosomes. Thus, rice is likely to be too diverged to serve as a resource for efficient map-based cloning of maize traits. However, once the structure of the maize genome is better understood in relation to rice, reciprocal comparative studies will be possible among grasses (for review, see Freeling, 2001). Having the maize sequence is also likely to benefit rice genome annotation, as illustrated by comparisons of the mouse and human genomes (Gregory et al., 2002).

Our current picture of the maize genome is largely derived from data generated by projects previously funded by the NSF Plant Genome Research Program. Two deep coverage bacterial artificial chromosome (BAC) libraries (Cone et al., 2002; Tomkins et al., 2002) have been produced, and an integrated genetic/physical map using a high-resolution agarose fingerprinting method (Cone et al., 2002; http://www.genome.arizona.edu/fpc/maize/) is being generated. At present, maize sequence data comprise expressed sequence tags (ESTs), genome survey sequences (GSSs) from transposon-tagged sites, random clone insert sequences, a few BAC clone end-sequences (BAC-ends), and sample sequences of genome subclones selected for hypomethylation or the presence of long open reading frames (ORFs). A handful of maize BAC clones have also been sequenced completely. Comparison of predicted maize ORFs from these sequence data with the Arabidopsis proteome suggests that maize-specific or highly diverged proteins contribute to a maize proteome that is anticipated to be about 50,000 proteins, or about twice the size of that of Arabidopsis (Brendel et al., 2002). More than 30% of the EST-derived ORFs and more than 70% of GSS-derived ORFs of maize do not match any Arabidopsis proteins. Although it is likely a number of these maize ORFs will be in rice, it will be interesting to determine the number of genes different between maize and rice. The upcoming sequence resources should expand significantly our knowledge of the gene space of flowering plants and additionally allow elucidation of possible differences in gene content between the monocot and dicot lineages.

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

www.plantphysiol.org/cgi/doi/10.1104/pp.015594.

CHALLENGES OF SEQUENCING THE MAIZE GENOME

The maize genome represents a significant new challenge for sequencing. At 2,500 million bp, the maize genome is about 20 times larger than that of Arabidopsis, about six times larger than that of rice, and about the same size as the human genome. However, its organization is more complex than the other genomes sequenced to date. The genes of maize compose only about 20% of the genome and are organized into islands of variable size that are scattered throughout a sea of highly conserved, high-copy retrotransposons and other repetitive sequences (San Miguel et al., 1996). Given this complex organization, how should the maize genome be sequenced? On July 2, 2001, NSF sponsored (DBI–0126620) a workshop in St. Louis to discuss technical approaches for a Maize Genome Sequencing Project. This workshop included academic, governmental, and industrial scientists with expertise in genome analysis as well as observers representing federal funding agencies (NSF, Department of Energy, National Institutes of Health, and U.S. Department of Agriculture) and U.S. corn growers' associations. All participants at the workshop agreed that genome sequencing and the placement of maize genes on a cross-referenced physical-genetic map was a feasible, worthy and timely goal that should be given a high priority. The workshop report (Bennetzen et al., 2001) described several strategies that could be used to focus sequencing efforts on gene-rich regions within complex genomes such as maize. The cost of identifying most of the maize genes and placing them on the integrated physical and genetic map was estimated at approximately $52 million.

One area highlighted in the fiscal year 2002 NSF Plant Genome Research Program Solicitation (NSF 01-158; http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf01158) was large-scale DNA sequencing of specific regions (e.g. gene-rich regions) or clones of large plant genomes. As a result of this competition, two projects were awarded $10.2 million over 2 years to begin to sequence the maize genome. One project (DBI-0211851: http://www.fastlane.nsf.gov/servlet/showaward?award=0211851) led by Joachim Messing of Rutgers University includes Rod Wing and Cari Soderlund of the University of Arizona, Francis Quetier of Genoscope (Evry, France), Hans-Werner Mewes and Klaus Mayer of the Munich Information Centre for Protein Sequences, and Michele Morgante of DuPont (Wilmington, DE) as a technical consultant. The other project (DBI–0221536: http://www.fastlane.nsf.gov/servlet/showaward?award=0221536), led by Karel Schubert of the Danforth Center (St. Louis), includes Roger Beachy also of the Danforth Center, Cathy Whitelaw and John Quackenbush of The Institute for Genomic Research (TIGR; Rockville, MD), Nathan Lakey of Orion Genomics (St. Louis), and Jeff Bennetzen of Purdue University (West Lafayette).

The project led by Joachim Messing will deliver a high-resolution, sequence-ready map of the maize genome. This map will integrate 450,000 fluorescent-based BAC clone fingerprint reads, 450,000 end sequences from 225,000 BACs, and 10× shotgun sequence of about 140 BACs seeded from about 10 points throughout the genome (totaling about 20 Mb of sequence or 1% of the genome). The project led by Karel Schubert will evaluate two gene enrichment technologies ("methylation-filtering" [Rabinowicz et al., 1999] and "high C_ot selection"). One million end reads from 250,000 clones from each of the methylation-filtered libraries and the high C_ot libraries will be assembled into contigs, annotated, and placed on the maize and rice genome maps. Together the two projects will yield the resources listed as the highest priorities in the maize genome sequencing workshop report. The integrated outcomes will be a maize sequence resource that will allow analysis of the overall architecture of the genome, including the size and distribution of the gene islands, the gene densities within these, and the range of gene structures. In addition, these resources will provide a minimal clone set that is colinear with the genetic map, providing the foundation for future large-scale sequencing of the maize genome, and the proof-of-concept for new methods to rapidly and selectively enrich for the genes of any large, complex plant genome in a cost-effective manner.

A meeting was held at NSF on September 18, 2002, to coordinate the two projects. Discussions at this meeting included how data would be shared between projects and communicated to the public. A single advisory committee will be formed to advise both projects, and frequent conference calls are planned both within and among the consortia. In addition, a single Web site from which all project data and related maize resources can be accessed was proposed.

FINDING PROJECT DATA

It will be important for the project data to be rapidly and readily accessible to maize researchers and the broader scientific community. The plans for release of each of the major deliverables are as follows. (a) All sequence data, including trace files, will be automatically deposited with GenBank within 24 h, or at most within a week, after production. Thus, public access to the data will be achieved with the shortest possible turn-around time. (b) BAC-end sequences from the Messing project will be deposited in dbGSS (http://www.ncbi.nlm.nih.gov/dbGSS/index.html). (c) A minimum tiling path of the BACs will be derived and displayed with the fingerprint contig (FPC) software (Soderlund et al., 2000) at Arizona Genomics Institute (Tucson; http://www.genome.arizona.edu/fpc/maize/). (d) Sequenced BACs will be initially deposited in the high throughput genomic sequences division of GenBank (http://www.ncbi.nlm.nih.gov/HTGS/). (e) Subsequent annotation of a total of 20 Mb of distinct BAC sequences will be conducted at Munich Information Centre for Protein Sequences. (f) Trimmed single-pass end sequences derived from methylation-filtered and high C_0t clones will be submitted to the high throughput genomic sequences division of GenBank. (g) Methylation-filtered and high C_0t clones will be available through the Arizona Genomics Institute (http://www.genome.arizona.edu/fpc/maize/).

Using known maize genes as auxiliary templates, all gene-enriched sequences will be assembled into contigs at TIGR using programs similar to the TIGR tools currently used for EST assembly in the production of TIGR gene indices (Liang et al., 2000). There will be no restrictions on the public use of these gene assemblies. The FPC maps and BAC-end sequences will serve as anchors for tying the TIGR assemblies to the maize genome. Both projects will map their sequence assemblies to the rice genome for comparative analysis. As the results of these analyses become available, they will be disseminated through a single Web site (http://www.maizegenome.org). This site will include a BLAST server with links from sequences to the individual project sites and a general description of the overall program with current progress toward the sequencing and mapping goals. This site will also provide a forum for community input and feedback.

From experience with Arabidopsis (and virtually all eukaryotic genome projects), annotation of gene structure and functional characterization of gene products will be a time-consuming endeavor that will continue well beyond the initial release of sequence data and their preliminary characterization. However, the informatics plan for this project provides the community with immediate access to all the primary data. Thus, researchers can look forward to a rich data-mining resource for their particular gene or genes of interest from the very beginning of the project.

The U.S. Department of Agriculture (USDA) is funding complementary efforts to integrate cereal genome sequence data such as these into a larger context. A next-generation maize genetics/genomics database (MGDb, http://www.mgdb.org/) under development by Volker Brendel (Iowa State University, in collaboration with the USDA Agricultural Research Service [ARS]) will provide comprehensive access to maize genetic and genomic data (including the new maize genome sequence data from these NSF projects). Gramene (http://www.gramene.org/) is a new data resource for comparative genome analysis in the grasses (Lincoln Stein, Cold Spring Harbor Laboratory; Sam Cartinhour, USDA/ARS, Ithaca; Susan McCouch, Cornell University), which links genome sequence and map data from rice to maps (physical and genetic) and DNA sequence of other cereals (including maize). The convergence of all these efforts will give plant biologists an unprecedented detailed view of plant genome content and organization within the next few years. In addition, it will provide an important public resource for growers and breeders.

Within the coming year, we should have a fuller picture of the architecture of the maize genome that includes the approximate size range and distribution of the gene islands, and an idea of the structure of a typical gene. In addition, we should have the first substantial public maize gene sequence collection that includes information about the promoter, upstream, and downstream sequences. This resource will set the stage for a large-scale sequencing effort. It is essential that the community be engaged in this endeavor from the very beginning. Feedback from the community will be critical for the maintenance, update, and dissemination of the project outcomes. The success of the Maize Genome Sequencing Project will very much depend on the extent to which the community becomes involved.

Received October 2, 2002; returned for revision October 4, 2002; accepted October 7, 2002.

LITERATURE CITED

Bennetzen J, Chandler V, Schnable P (2001) National Science Foundation-sponsored workshop report: Maize Genome Sequencing Project. Plant Physiol **127:** 1572–1578

Brendel V, Kurtz S, Walbot V (2002) Comparative genomics of Arabidopsis and maize: prospects and limitations. Genome Biol Rev **3:** 1005.1–1005.6

Cone K, McMullen M, Bi IV, Davis G, Yim Y-S, Gardiner J, Polacco M, Sanchez-Villeda H, Fang Z, Schroeder S et al. (2002) Genetic, physical and informatics resources for maize: on the road to an integrated map. Plant Physiol **130:** 1598–1605

Dubcovsky J, Ramakrishna W, San Miguel 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

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

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

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA et al. (2002) A physical map of the mouse genome. Nature **418:** 743–750

Keller B, Feuillet C (2000) Colinearity and gene density in grass genomes. Trends Plant Sci **5:** 246–251

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

Li W, Gill BS (2002) The colinearity of the Sh2/A1 orthologous region in rice, sorghum and maize is interrupted and accompanied by genome expansion in the triticeae. Genetics **160:** 1153–1162

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J (2000) An optimized protocol for analysis of EST sequences. Nucleic Acids Res **28:** 3657–3665

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

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

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res **10:** 1772–1787

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

Tikhonov AP, San Miguel 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

Tomkins JP, Davis G, Main D, Yim Y, Duru N, Musket T, Goicoechea JL, Frisch DA, Coe EH Jr, Wing RA (2002) Construction and characterization of a deep-coverage bacterial artificial chromosome library for maize. Crop Sci **42:** 928–933