8-2005

Toward Sequencing the Sorghum Genome: A U.S. National Science Foundation-Sponsored Workshop Report

Stephen Kresovich
*University of South Carolina - Columbia, sk@sc.edu*

Brad Barbazuk

Joseph A. Bedell

Andrew Borrell

C. Robin Buell

*See next page for additional authors*

Follow this and additional works at: [https://scholarcommons.sc.edu/biol_facpub](https://scholarcommons.sc.edu/biol_facpub)

Part of the *Biology Commons*

Publication Info


DOI: 10.1104/pp.105.065136

© Plant Physiology, 2005, American Society of Plant Biologists


This Article is brought to you by the Biological Sciences, Department of at Scholar Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.
Update on Sequencing the Sorghum Genome

Toward Sequencing the Sorghum Genome. A U.S. National Science Foundation-Sponsored Workshop Report

Sorghum Genomics Planning Workshop Participants

Members of the worldwide sorghum (Sorghum spp.) community, including private sector and international scientists as well as community representatives from closely related crops such as sugarcane (Saccharum spp.) and maize (Zea mays), met in St. Louis, Missouri, on November 9, 2004, to lay the groundwork for future advances in sorghum genomics and, in particular, to coordinate plans for sequencing of the sorghum genome. Key developments that made this workshop timely included advances in knowledge of the sorghum genome that provide for the development of a genetically anchored physical map to guide sequence assembly and annotation, the growing role of the sorghum genome as a nucleation point for comparative genomics of diverse tropical grasses including many leading crops, and the need for dramatically increased sorghum production to sustain human populations in many regions where its inherent abiotic stress tolerance makes it an essential staple. This report reviews current knowledge of the sorghum genome, a community-endorsed schema for integrating this knowledge into a finished sequence, and early plans for translating the sequence into sustained advances to benefit a worldwide group of stakeholders.

WHAT ARE SOME OF THE UNIQUE CONTRIBUTIONS TO BIOLOGY AND AGRICULTURE THAT WILL RESULT FROM SEQUENCING OF THE SORGHUM GENOME?

Sorghum (Sorghum bicolor L. Moench) is one of the world’s leading cereal crops, providing food, feed, fiber, fuel, and chemical/biofuels feedstocks across a range of environments and production systems. Worldwide, sorghum is the fifth most important cereal crop (http://apps.fao.org/default.jsp). Its remarkable ability to produce a crop under adverse conditions, in particular with much less water than most other grain crops, makes sorghum an important “failsafe” source of food, feed, fiber, and fuel in the global agroecosystem. For example, in arid countries of northeast Africa such as Sudan, sorghum contributes about 39% of the calories in the human diet (http://www.fao.org, 1999 statistics). Increased demand for limited fresh water supplies, coupled with global climatic trends and expanding populations, suggests that “dryland” crops such as sorghum will be of increasing importance.

As a model organism for tropical grasses that carry out “C4” photosynthesis, sorghum is a logical complement to the C3 grass Oryza (rice), the first monocot...
WHAT IS THE NATURE AND ORGANIZATION OF THE SORGHUM GENOME?

Estimates of the physical size of the sorghum genome range from 700 Mb based on Cot analysis (Peterson et al., 2002) to 772 Mb based on flow cytometry (Arumuganathan and Earle, 1991). This makes the sorghum genome about 60% larger than that of rice, but only about one-fourth the size of the genomes of maize or human. GC content is estimated at 37.7% (Peterson et al., 2002). Because sorghum is a predominantly self-pollinated plant, most genotypes are homozygous, including each of the three genotypes for which extensive genetic maps, bacterial artificial chromosome (BAC) resources, and physical maps have been constructed.

DNA renaturation kinetic analysis (Peterson et al., 2002) shows the sorghum genome to be comprised of about 16% foldback DNA, 15% highly repetitive DNA (with individual families occurring at an average of 5,200 copies per genome), 41% middle-repetitive DNA (average 72 copies), and 24% low-copy DNA. About 4% of the DNA remained single stranded at very high Cot values and is assumed to have been damaged (thus, the other percentages are slight underestimates).

Building on a rich history of genetics research supported by a wide range of sources, recent National Science Foundation (NSF)-funded activities have significantly advanced current knowledge of the sorghum genome. High-density maps of one intraspecific S. bicolor (Klein et al., 2000; Menz et al., 2002) and one interspecific S. bicolor × Sorghum propinquum (Chittenden et al., 1994; Bowers et al., 2003) cross provide about 2,600 sequence-tagged sites (based on low-copy probes that have been sequenced), 2,454 amplified fragment length polymorphisms, and approximately 1,375 sequence-scanned (based on sequences of genetically anchored BAC clones) loci. More than 800 markers mapped in sorghum are derived from other taxa (hence serve as comparative anchors), and additional sorghum markers have been mapped directly in other taxa or can be plotted based on sequence similarity. The two maps share one common parent (S. bicolor BTx623) and are essentially colinear (F. Feltus, G. Hart, K. Schertz, A. Casa, S. Kresovich, P. Klein, P. Brown, and A.H. Paterson, unpublished data). Recent cytological characterization of the individual sorghum chromosomes has provided a generally accepted numbering system (Kim et al., 2005).

The small size of the sorghum genome facilitates its use as a tropical grass model. While the maize and sugarcane genomes are similar in size to the human genome, the sorghum genome is approximately 75% smaller, variously estimated at 690 to 760 Mb. BAC libraries are available for BTx623 (about 12× coverage from HindIII and 8× from BamHI), S. propinquum (13–14× coverage from EcoRI [approximately 7×] and HindIII [approximately 7×]) and IS6320C (approximately 9× coverage from HindIII). A total of 69,545 agarose-based fingerprints from BTx623 BACs are also
anchored with 139,434 hybridization loci from 5,147 probes (about 2,000 of which are genetically mapped). In parallel, 40,957 agarose-based fingerprints from *S. propinquum* are anchored with 148,758 hybridization loci from 5,683 probes (2,000 genetically mapped). Each of these has been assembled into WebFPC-accessible physical maps (http://www.stardaddy.uga.edu/fpc/bicolor/WebAGCoL/WebFPC/ and http://www.stardaddy.uga.edu/fpc/propinquum/WebAGCoL/WebFPC/). Additional resources include 20,000 high information content fingerprint (HICF) fingerprints (from genetically mapped contigs) and six-dimensional BAC pools (5× deep) from BTx623, and 10,000 HICF fingerprints and six-dimensional BAC pools (5× deep) from IS3620C. Targeted HICF of additional contig-terminal BACs is in progress to fill gaps. About 456 *S. propinquum* and 303 *S. bicolor* BAC contigs (41% of BACs, 80% of single-copy loci) appear to be well anchored to euchromatic regions, with the percentage of the genome attributable to euchromatin likely to rise appreciably with additional anchoring. The finding that 41% of BACs are already anchored to euchromatin while only 24% of the sorghum genomic DNA is single or low copy (with an overall kinetic complexity of 1.64 × 10^6; Peterson et al., 2002) suggests that euchromatin includes a mixture of low-copy and repetitive DNA.

A detailed report of the Sorghum Genomics Planning Workshop, sponsored by NSF, is available as supplemental data. The goals of the workshop were to (1) obtain a status report on the development and accessibility of sorghum genome research information, technologies, and infrastructure; (2) identify future priorities and needs for sorghum genomics research; (3) better organize the sorghum community; and (4) foster sorghum improvement.

Prior to the meeting, a survey was conducted to establish priority needs of the broader user community. The findings of the survey, obtained from the input of approximately 60 respondents among 140 members of the international sorghum (and closely related sugarcane) communities polled, are attached to the report. Topics included in the workshop were those identified as key issues by the user community. In summary, these can be classified into four focus areas, mapping, sequencing, germplasm, and database/bioinformatics, each of which is addressed below.

An interim Sorghum Genomics Steering Committee was charged with the development of the “white paper” outlining key priorities for sorghum genomics for the next 5 to 10 years, as well as devising a mechanism for, and carrying out, an election process to maintain and enhance community activities and impact.

While a natural long-term goal is a high-quality assembled sequence that is finished to Bermuda standards, this is likely to be accomplished in stages that build on one another.

The white paper (see supplemental data) details a three-stage strategy, also noting that aspects of the three stages are proceeding to some degree in parallel.

### Stage 1: Gene Space Characterization

The sorghum gene space is presently represented by approximately 200,000 expressed sequence tags (ESTs) that have been clustered into approximately 22,000 uniscripts, representing more than 20 diverse libraries from several genotypes. Genome annotation will benefit from additional EST sequencing, emphasizing full-length clones. This also presents an opportunity to sample both physiological and genetic (single-nucleotide polymorphism) diversity by drawing these ESTs from diverse genotypes.

About 500,000 methyl-filtered (MF) reads that provide estimated 1× coverage of the MF-estimated gene space (Bedell et al., 2005) have been assembled into contigs (SAMIs; http://magi.plantgenomics.iastate.edu/). Another reduced-representation strategy, Cot-based cloning and sequencing (CBCS), was first demonstrated in sorghum in 2001 (as noted in GenBank accessions AZ921847–AZ923007) and further detailed subsequently (Peterson et al., 2002). This method offers the potential to further enhance gene space coverage beyond that offered by ESTs and MF, in a complementary manner as demonstrated for maize. Sequencing of the low-copy DNA to similar levels of coverage, by MF- and CBCS-based methods, is viewed as a logical intermediate step toward efficiently capturing the sequence complexity of sorghum. Primary sequencing of genomic DNA should be focused on the inbred genotype BTx623 (see below), to foster sequence assembly.

### Stage 2: Gold-Standard Physical Map

Most genomic resources for sorghum have been developed using a U.S. inbred, BTx623, which was selected as a focal point for genomic sequencing and which enjoys about 20× genome coverage by two sets of BACs cloned using two different restriction enzymes. However, a gold-standard physical map will necessarily integrate data from multiple genotypes to help resolve genomic instabilities or other genotype-specific or species-specific features that interfere with cloning and/or sequencing. As such, our integrated physical map will comprise detailed alignment among BTx623, *S. propinquum*, and *S. bicolor* accession IS3620C. This will not only provide for filling gaps, but also advance application to studies of unique aspects of plant biology for which these diverse genotypes represent botanical models.

### Stage 3: Finished Sequencing

Sorghum is a relatively complex genome, but with a smaller overall genome size and less repetitive DNA than many of its relatives, such as maize and sugarcane. While the physical map will provide the means to conduct BAC-by-BAC finished sequencing of a minimum tiling path, ongoing technological and computational improvements may offer compelling efficiencies to whole-genome shotgun (WGS)-based approaches, or more probably to hybrid approaches.
that integrate aspects of BAC-based and WGS approaches. A high-quality genetically oriented physical map provides a robust guide for assembly by either BAC-based or WGS approaches, and the community remains open to considering a range of options for completion of the sequence based on economics and the state of the art as funding becomes available. In addition, it is noted that the relationship between the physical map and the sequence may be iterative; for example, targeted physical anchoring of WGS contigs may expedite assembly and closure.

Database Resources/Bioinformatics

While existing Web-based resources focus on comparative structural and evolutionary genomics (http://cggc.agtec.uga.edu/), functional genomics of the transcriptome (http://fungen.botany.uga.edu/), and genomics of the unique abiotic stress responses of sorghum (http://sorgblast2.tamu.edu/), the community recognizes a growing need to develop a unified sorghum database much like Maize GDB or rice-centric Gramene. This database may be centralized or federated but should maintain critical links to the individual groups' databases, thus taking advantage of the respective strengths of individual groups in annotation and curation of data that they have firsthand knowledge of and that is of primary importance to them. In addition, the establishment of a “Sorghum Portal,” with links to relevant Web resources, is recommended. Finally, data also need to be accessible in formats compatible with usage by scientists in regions where Internet access remains unavailable or too slow to efficiently download genomic data sets.

Applications to Benefit Worldwide Stakeholders

Much of the value of a sorghum sequence would be realized through better understanding of the levels and patterns of diversity in extant germplasm, which can contribute both to functional analysis of specific sorghum genes and to deterministic improvement of sorghum for specific needs and environments. Extensive ex situ sorghum germplasm collections exist within the U.S. National Plant Germplasm System and the International Crops Research Institute for the Semi-Arid Tropics. An informal meeting of breeders and geneticists at Cornell in September 2004 laid the groundwork for development of “core” sorghum panels (including wild species, landraces, and elite genotypes), which will capture as much genetic diversity as possible while minimizing redundancies. These panels are expected to be suitable for association genetics-based approaches that explore relationships between phenotypes and haplotype variation at candidate genes directly, or for markers distributed either throughout the genome (genome-wide association tests) or closely linked to a target locus. Plans were also outlined for development of about 25 recombinant inbred line populations initially needed to foster joint quantitative trait loci-association genetics studies. The germplasm planning group met again on February 21, 2005, to further advance a recommendation to be circulated to the broader sorghum community for feedback.

Africa is recognized as the center of origin and diversity for sorghum, the source of about 50% of the accessions in the world collections, and the location of much in situ diversity lacking from existing collections. Sorghum is thus an attractive vehicle for engagement of the African scientific community in genomics and its applications, in particular regarding documentation and analysis of in situ diversity that is presently inaccessible to Western scientists. However, the problems of limited communications infrastructure, slow Internet connectivity, and lack of information technology support will present special challenges. Periodic data dissemination in CD format, together with the development of simplified database structures that use off-the-shelf software but are cross-compatible with more sophisticated structures such as the planned centralized database (above), will be essential. Implementation of these new capabilities will benefit from the engagement of institutions such as the Consultative Group on International Agricultural Research Centers, the International Sorghum and Millet Collaborative Research Support Program, the Rockefeller and Syngenta foundations, and others.

To sustain coordination and communication, the initial Sorghum Genomics Executive Committee will coordinate the identification and election of its successors, to comprise a 15-member committee with at least five and two seats filled by non-U.S. and private sector members, respectively, to serve 3-year terms on a staggered basis (five new members per year). The elected committee will establish guidelines for future committee actions; collect, collate, and disseminate information; organize consortium meetings; and serve as an advocacy group for each country and area of research, with emphasis on cohesiveness of the community and importance of the crop.

SYNTHESIS

As a model organism for tropical grasses that carry out C₄ photosynthesis, sorghum is a logical complement to the C₃ grass Oryza, the first monocot plant with a near-completely sequenced genome. The relatively small genome of sorghum is likely to be appreciably less complex to assemble than the larger and more repetitive genomes of other major C₄ crops, such as maize and sugarcane. Detailed physical maps provide a foundation upon which to overlay sequence assemblies, linking them to a rich history of genetics and genomics research based on inclusion of genetically mapped sequence-tagged sites in the physical map. Sequencing of sorghum will fill a key gap in plant biogeography in view of its African origin, permitting phylogenetic triangulation of key events in cereal evolution and in particular leading to new
insights into parallel but independent domestication of the cereals. Its remarkable ability to produce a crop under adverse conditions, in particular with much less water than most other grain crops, makes sorghum an important failsafe source of food, feed, fiber, and fuel in the global agroecosystem with special relevance to Africa. Moreover, sorghum genome analysis offers novel learning opportunities relevant to weed biology as well as to improvement of a wide range of other forage, turf, and biomass crops. The infrastructure to link a finished sequence to a rich history of genetics research, toward resolution of a wide range of challenges facing a worldwide set of stakeholders, is largely in place.

ACKNOWLEDGMENT

The U.S. National Science Foundation supported the Sorghum Genomics Workshop (under award no. DBI–0456171).

Received May 3, 2005; revised May 3, 2005; accepted June 5, 2005; published August 11, 2005.

LITERATURE CITED


