Resources and Opportunities

Access to the Maize Genome: An Integrated Physical and Genetic Map[1]

Edward Coe*, Karen Cone, Michael McMullen, Su-Shing Chen, Georgia Davis, Jack Gardiner, Emmanuel Liscum, Mary Polacco, Andrew Paterson, Hector Sanchez-Villeda, Cari Soderlund, and Rod Wing

Department of Agronomy (E.C., G.D., J.G., M.M., M.P., H.S.-V.), Division of Biological Sciences (K.C., E.L.), and Department of Computer Engineering/Computer Science (S.C.), University of Missouri, Columbia, Missouri; Agricultural Research Service, United States Department of Agriculture, Columbia, Missouri (E.C., M.M., M.P.); Clemson University, Clemson, South Carolina (C.S., R.W.); and University of Georgia, Athens, Georgia (A.P.)

STRATEGIES TO ACCESS A COMPLEX CROP GENOME

Crop plant research is poised to make revolutionary strides including the following: cloning target genes based on their function and/or their position in the genome; documenting all genes and their interplay; defining and exploring all the existing genetic diversity in a species; and using functional information and syntenic relationships of genes in closely related species to extrapolate gene function in crop plants. The challenge, however, is to develop a set of comprehensive and systematic resources to facilitate these research endeavors. Genomic resources in maize (*Zea mays*) will undergird sequencing of the maize genome and will complement and contribute to research in the cereals, other grasses, and other crop plants.

For maize, developing genomics tools means facing the daunting realities of size and complexity. At approximately 2,500 megabases, the maize genome is comparable in size to that of humans, and the complexity is likely to be greater because of the abundance of multiple families of repetitive elements. Thus, gaining access to the maize genome is best tackled by following Goethe's advice to seek entry to the whole by going to its parts:

Willst du ins Unendliche schreiten,

Geh nur im Endlichen nach allen Seiten.

—Johann Wolfgang von Goethe

To efficiently take a genome apart and put it back together requires a combination of genetic and physical mapping. In principle, genetic mapping serves to subdivide and order the genome by crossing over and recombining parts, whereas physical mapping allows ordering of genomic fragments (parts) by determining the overlap among them. Accordingly, the components of the public effort to develop resources to access the complete maize genome are to produce a high-resolution genetic map densely populated with markers; to produce, fingerprint, and assemble a deep-coverage library of bacterial artificial chromosomes (BACs) into physical map segments; and through molecular markers to integrate the genetic and physical maps. The tools that are being built in maize will provide a scaffold upon which to hang the sequence and the gene constitution of the genome and will link the sequence to the collected efforts of the maize genetics community over the past century.

THE GENETIC MAP

A core resource essential for developing an integrated genetic/physical map for maize is a densely marked, high-resolution genetic map. We have constructed a >1850 marker map for the intermated B73/Mo17 (IBM) population (Davis et al., 2001). The parents of the population, B73 and Mo17, represent the two major heterotic groups of U.S. maize germplasm. The IBM population consists of 304 recombinant inbred lines that underwent four rounds of random mating at the F_2 stage. The additional meioses result in a 3-fold expansion of the genetic map (Liu et al., 1996). The combination of a large number of lines and the map expansion generate a map resource with approximately 17 times the order resolution power of the prior maize map standard (UMC 98 genetic map, Davis et al., 1999). The IBM map is populated with >1,000 RFLP and >850 simple sequence repeat (SSR) markers.

In addition to the map per se, numerous related resources are provided to the maize researcher. Seeds of the IBM lines can be obtained from the Maize Genetics Cooperation Stock Center (Urbana, IL). A collateral resource resulting from map development is numerous additional genetic markers. These markers include novel single copy RFLP probes, markers for Mutator insertion sites, and new SSR primer sets. Links to obtaining clones for hybridization probes, screening images for RFLP probes, primer and

screening information for 1,800 maize SSR loci, and mapscore data for the IBM population are available from the Maize Genome Database (MaizeDB, http://www.agron.missouri.edu/). Twelve hundred SSR loci were developed by our group (Sharopova et al., 2001). Comparative maps for SSR loci not on the IBM map have also been constructed and displayed.

We have undertaken a number of initiatives promoting the use of the IBM map as a community resource. A subset of 94 IBM lines has been identified for general community use. We distribute DNA of the 94 lines in microtitre plate format for individual investigators to map their genes much as radiation hybrid panel DNA is available for mammalian gene mapping. We have implemented Web entry for submission of map score data, from which the resulting locus positions are returned to the investigator. These features are intended to provide the maize community at large with resources to contribute to generating comprehensive information on maize gene map position.

THE PHYSICAL MAP

As a first step in producing a physical map, we constructed three genomic DNA libraries in BACs using DNA from the inbred line B73. B73 was selected because it is one of the parents of the genetic mapping population; thus, markers mapped on the IBM population and used to screen the BAC libraries could provide anchors for connecting the genetic and physical maps. To ensure deep coverage and minimal gaps in sequence representation in the libraries, libraries were made using three different restriction enzymes, and each library contained sufficient numbers of clones to provide severalfold coverage of the haploid genome. Details of the three libraries are shown in Table I. Together, the libraries represent 27-fold genome coverage.

To assemble the BACs into contigs, the clones are fingerprinted by digesting with *Hin*dIII. The fingerprinting involves fractionating the fragments on agarose gels, scanning the gel images, digitizing them with IMAGE software (Sanger Center, UK), and an-

alyzing for contig formation using Fingerprint Contig (FPC) software (Soderlund et al., 2000). Contigs are generated automatically with a cutoff value of 10^{-12}. Once all fingerprint data have been collected, the contigs will be edited manually to ensure accuracy and consistency.

Accuracy and speed of contig assembly can be enhanced by screening the BAC libraries with molecular markers. Resulting BAC addresses provide important anchor points for assembling individual BACs into contigs. A variety of markers is being used to screen the BAC libraries.

Core RFLP Markers

The 90 core RFLP markers that serve as bin landmarks on the genetic map have been used for BAC screening. The majority (79) of the markers could be used directly to hybridize to BAC filter sets. The rest of the markers contained repetitive sequences that precluded obtaining definitive addresses, pointing to the need to develop single-copy overgo probes from these markers.

Maize ESTs

In a partnership of our group with DuPont (Wilmington, DE) and Incyte Genomics (Palo Alto, CA), a unigene set of approximately 10,000 maize ESTs is being used to screen clones from the *Hin*dIII and *Eco*RI BAC libraries. The unigene set was generated by DuPont, using the publicly available maize ESTs to seed their proprietary EST collection to define unigene clusters containing the longest possible sequences for each gene in the set. At Incyte Genomics, the unigene sequences were masked for repetitive elements, overgo probes designed for each gene, and the probes used to hybridize to filters containing BAC DNA.

Sorghum Markers

Overgo probes derived from sorghum genomic and cDNA clones are being used to screen the BACs. These probes were designed from sequences that hybridize across several cereal genomes, including maize.

Amplified Fragment-Length Polymorphism (AFLP) and Miniature Inverted-Repeat Transposable Element (MITE) Markers

An additional approach for screening the BACs is based on a strategy recently applied to constructing an integrated map for sorghum (Klein et al., 2000). In this method, BAC pools are screened with multiple-site PCR markers, AFLPs, and MITEs to assign markers to specific BACs.

Table I. *Descriptions of maize B73 BAC libraries*

Enzyme	No. of Clones	Average Insert Size	Haploid Equivalents
		kb	
*Hin*dIII[a]	230,400	137	12.5
*Eco*RI[b]	110,592	162	7.1
*Mbo*I[c]	110,592	167	7.4

[a] *Hin*dIII library constructed at CUGI. Filter sets (5×) and clones are available from CUGI (http://www.genome.clemson.edu/orders/). [b] *Eco*RI library constructed at Children's Research Hospital of Oakland (CHORI). Filter sets (5×) and clones are available from CHORI or from CUGI. [c] *Mbo*I library constructed at CHORI, in collaboration with Jo Messing. Filters and clones are available from CHORI (http://www.chori.org/bacpac/home.htm).

MaizeDB serves as the clearinghouse for information about all markers and the BACs they detect. BAC address information for each marker is then incorporated via FPC into the growing BAC contig assemblies. A Java applet (WebFPC) has been created to display contigs on the web at http://www.genome.clemson.edu/projects/maize/fpc/. BAC contigs are updated monthly, and the data can be searched by individual BAC clone, marker, or contig.

INTEGRATING THE PHYSICAL AND GENETIC MAPS

Current estimates indicate that at least 50% of the maize genome consists of complex arrays of retrotransposon-like elements and that the majority of these repetitive elements represent a small number of related families (San Miguel et al., 1996). There are no prior attempts to assemble a physical map for any eukaryotic organism with this structure. The size of the maize genome, 2,500 Mb, also contributes to the difficulty of assembly. With the first of 83,000 of the 450,000 BAC fingerprints completed, FPC assembled the BACs into 13,000 contigs and 9,300 singletons (Clemson University Genomics Institute [CUGI] Web site, August 2001). Assuming 450,000 BACs, 150 kb in length, representing $27\times$ in coverage, and 80% overlap; fingerprinting alone is expected to result in approximately 2,000 apparent contigs (Lander and Waterman, 1988). Complexity in assembly caused by repetitive elements may cause that number to increase.

It is clear the task at hand is to place large numbers of genetically mapped anchor points against the BACs to both coalesce contigs and order contigs on the chromosome framework. To tie the assembled BAC contigs to the genetic map, the BACs must be screened with genetically mapped markers. The 90 core RFLP markers define the bin boundaries on the genetic map and set a framework for the integrated map. Many of the AFLP and MITE markers detected by screening the BAC pools are polymorphic in the IBM population and serve as genetic anchors. In addition, the sorghum markers are useful anchoring tools because they have all been mapped in sorghum, most have been mapped in maize, and many cross-hybridize to DNA from other cereals such as rice (*Oryza sativa*), sugarcane (*Saccharum officinarum*), and *Pennisetum glaucum*. Thus, these markers will not only help connect the BAC contigs to the maize map, but because of the colinearity of cereal genomes, they will also facilitate creation of comparative maps.

Of the 10,000 ESTs that will be used to screen the BAC libraries, we anticipate anchoring 1,000 contigs by EST and SSR markers currently on the IBM map. Many of these locations were determined by mapping SSRs that were derived from the public EST sequences. We are initiating development of 2,000 single-nucleotide polymorphism markers derived from the ESTs. Once these global anchoring approaches draw the majority of contigs to the genetic framework, directed sequencing of BACs and BAC ends will be used to derive the source sequence for development of additional single-nucleotide polymorphism markers to attempt to bring the physical map to closure.

At MaizeDB, a BLAST server has been implemented to allow users to compare sequences of interest to all public maize sequences (including those in the unigene set) to return a map location if known. The output of such a query is a bin location, BLAST score, and database links for more information, e.g. MaizeDB link to map details; CUGI link to BAC contigs; GenBank; and ZmDB (*Zea mays* database at Iowa State University) links for sequence and clone information.

Maps can be viewed at MaizeDB in three ways. For any single map, chromosome-specific views are available. For simultaneously looking at two genetic maps, a comparative map viewer has been developed by enhancing GIOT software obtained from the Rice Genome Project (Japan). This viewer shows side-by-side comparisons of similar regions of selected versions of maize maps and can be expanded to include comparisons of maize to sorghum and rice. For comparing the maize genetic map to the developing physical map—as marker data are obtained for the BAC contigs—the GIOT viewer can be adapted to allow queries by locus, marker name, or linkage group to display the associated physical contigs.

The genomic resources being developed for maize will serve as a skeleton upon which to hang the sequence and the gene constitution of the genome, and will link the sequence to a century of accumulated genetic studies by members of the maize genetics community. Genomic and genetic knowledge in other cereals, other grasses, and other plant species will complement the resources in maize and will be enhanced in concert with them.

MaizeDB can be accessed at http://www.agron.missouri.edu/, where maps, probes, primers, screening images, and lab protocols are presented.

LITERATURE CITED

Davis G, McMullen M, Baysdorfer C, Musket T, Grant D, Staebell MS, Xu G, Polacco M, Koster L, Melia-

Hancock S et al. (1999) A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged sites (ESTs) in a 1736-locus map. Genetics **152:** 1137–1172

Davis G, Musket T, Melia-Hancock S, Duru N, Sharopova N, Schultz L, McMullen M, Sanchez H, Schroeder S, Garcia AA et al. (2001) The Intermated B73 x Mo17 Genetic Map: A Community Resource. Maize Genetics Conference Abstracts **43:** W15, P62

Klein P, Klein RR, Cartinhour S, Ulanch P, Dong J, Obert J, Morishige DT, Schlueter S, Childs K, Ale M et al. (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res **10:** 789–807

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics **2:** 231–239

Liu S, Kowalski SP, Lan TH, Feldmann IA, Paterson AH (1996) Genome-wide high-resolution mapping by recurrent intermating using *Arabidopsis thaliana* as a model. Genetics **142:** 247–258

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

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T et al. (2001) Development and mapping of SSR markers for maize. Plant Mol Biol (in press)

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