iMap: a database-driven utility to integrate and access the genetic and physical maps of maize

Z. Fang¹,*, K. Cone²,*, H. Sanchez-Villeda¹, M. Polacco¹,³, M. McMullen¹,³, S. Schroeder¹, J. Gardiner¹, G. Davis¹, S. Havermann¹, Y. Yim¹, I. Vroh Bi¹ and E. Coe¹,³

¹Department of Agronomy, ²Division of Biological Sciences, University of Missouri, Columbia, MO 65211, USA and ³Plant Genetics Research Unit, USDA-ARS, Columbia, MO 65211, USA

Received on February 6, 2003; revised on April 21, 2003; accepted on May 8, 2003

ABSTRACT
Motivation: Because of the unique biological features, a bioinformatic platform for the integrated genetic and physical map of maize is required for storing, integrating, accessing and visualizing the underlying data.

Results: The goal of the Maize Mapping Project is to develop a fully integrated genetic and physical map for maize. To display this integrated map, we have developed iMap. iMap has three main components: a relational database (iMapDB), a map graphic browser (iMap Viewer) and a search utility (iMap Search). iMapDB is populated with current genetic and physical map data, describing relationships among genetic loci, molecular markers and bacterial artificial chromosome (BAC) contigs. The database also contains integrated information produced by applying a set of anchoring rules to assign BAC contigs to specific locations on the genetic map. The iMap Viewer and iMap Search functions are combined in the user interface to allow viewing and retrieving many types of genetic and physical map data. The iMap Viewer features side-by-side chromosome-based displays of the genetic map and associated BAC contigs. For each genetic locus, information about marker type or contig can be viewed via pop-up windows that feature links to external data resources. Searches can be conducted for genetic locus, probe or sequence accession number; search results include relevant map positions, anchored BAC contigs and links to the graphical display of relevant chromosomes. iMap can be accessed at http://www.maizemap.org

Availability: The iMap utility package is available for non-commercial use upon request from the authors.

Contact: fangz@missouri.edu

INTRODUCTION

The Maize Mapping Project (MMP, http://www.maizemap.org), a co-operative project between the University of Missouri-Columbia, the University of Georgia and the University of Arizona, is generating an integrated genetic and physical map for maize. The integrated map will serve as a framework for future studies, including genomic sequencing, functional genomics, marker-assisted selection and comparative genome analysis.

To develop the integrated map, the MMP has assembled a high-resolution genetic map for the Intermed B73/Mo17 population (IBM map) consisting of ~1850 restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR) markers (Cone et al., 2002; Lee et al., 2002; Sharopova et al., 2002). The map also includes 226 single nucleotide polymorphism (SNP) and insertion–deletion (InDel) polymorphism markers (Vroh Bi et al., manuscript in preparation). Another genetic map, IBM neighbors, was generated by extrapolating locations of loci from non-IBM maps to their nearest neighbors on the IBM map (Cone et al., 2002). The MMP continues to expand the number of loci on these maps. The physical map assembly involves HindIII fingerprinting genomic bacterial artificial chromosome (BAC) clones from three libraries of the maize inbred B73 and assembly of the BACs into contigs using FingerPrint Contig (FPC; Soderlund et al., 2000; Cone et al., 2002; Tomkins et al., 2002). The BACs have been screened with a variety of molecular markers, many of which have been mapped genetically providing anchors to integrate the physical and genetic maps (Davis et al., 1999; Sharopova et al., 2002; Yim et al., 2002).

It is a challenge to store, integrate, access and visualize an integrated genetic and physical map. At the time we developed our map utility, several Internet tools were available that could be used to display integrated maps; however, they did not provide all the features needed for
the maize map. Two tools, the Map Viewer developed at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html) and the ENSEMBL viewer developed at the Sanger Institute (ENSEMBL, http://www.ensembl.org), were developed to accompany genome sequencing efforts and both rely heavily on sequence information for organizing the map displays (Hubbard et al., 2002; Wheeler et al., 2002). A third map viewer, employed by the Arabidopsis Information Resource (http://www.arabidopsis.org), displays the sequence, genetic and physical maps separately by transmitting GIF images via the Internet (Huala et al., 2001). One limitation of this tool is that it does not offer a direct link for an associated marker or sequence on these maps. Another interesting software is the Genome Information displayed Orderly Tool (GIOT), developed by the Rice Genome Program, Japan (Sakata et al., 2000, http://rgp.dna.affrc.go.jp). GIOT displays an aligned genetic and physical map by reading data from a succinct, easily updated tabular file and provides both gross and fine map presentations. Detailed information about a marker is provided in a pop-up window. However, this tool displays ambiguous contigs on the physical map and separately presents expressed sequence tag (EST) map and linkage map (generated by RFLPs and QTLs). As we developed our integrated map tool, two other map viewers were developed. WebChrom at the Arizona Genome Institute (http://www.genome.arizona.edu/fpc/maize/WebChrom) and Genetic & Comparative Maps tool at Gramene (http://www.gramene.org) display genetic and physical maps and the markers that are common to the two maps.

For the unique features of maize genetic and physical maps, we developed a bioinformatic platform, iMap, tailored for the maize data. iMap had to meet the following requirements: (i) to operate in a database environment for comprehensive data management; (ii) to integrate dynamically the genetic and physical maps based on the contig-anchoring rules; (iii) to display side-by-side genetic map and associated anchored BAC contigs; and (iv) to allow searches for map locations by genetic locus, probe or sequence accession number. Currently, iMap integrates the physical map data with two genetic maps, IBM and IBM neighbors, and can be viewed at http://www.maizemap.org.

**SYSTEM AND METHODS**

**Hardware and software**

The system uses two servers: a database server and a WWW server. The database server with the Sybase Adaptive Server Enterprise Version 11.9.2 runs on a Dell Linux platform. The WWW server (Apache 1.3.1) operates on a Unix platform. Several program modules are implemented in the WWW server to optimize the operation of the system, graphic display and search capability. The modules, which run in a web browser, are written in Java Applet (version 1.4) and

![Fig. 1. The iMap structure.](https://academic.oup.com/bioinformatics/article-abstract/19/16/2105/242511)
INTEGRATING GENETIC AND PHYSICAL MAP DATA

Integrating the genetic and physical map data is a challenge that requires processing different types of information to define relationships among the entities. Figure 2 presents a flowchart showing the types of data obtained from genetic and physical mapping experiments and steps involved in processing the data to produce a final integrated map. Genetic maps provide information on Marker–Locus relationships. Physical mapping experiments provide data on Marker–BAC relationships and BAC–Contig assemblies; from these data, Marker–Contig relationships can be extrapolated. Integrating the genetic and physical data requires anchoring of each contig to its proper genetic locus followed by presentation of the details for the markers and contigs corresponding to each genetic locus.

Anchoring contig to genetic locus

The assignment of a contig to a genetic locus is a major challenge in making the integrated genetic and physical map. Although it is obvious that one contig should have one genetic position, several types of conflicts can arise that make Contig–Locus assignments problematic (Cone et al., 2002). To minimize internal conflicts, a set of guidelines was developed for anchoring contigs to genetic loci (Fig. 3). Initially, a table is generated containing all the Contig–Locus relationships. Although the table contains some conflicting relationships, it serves as the framework for making unambiguous anchoring assignments of contigs to genetic loci. The first step in the anchoring process is to make all the assignments for which no conflicts exist, e.g. strict one contig–one locus associations, where one contig is associated with one locus and vice versa. The second anchoring step makes assignments for contigs that are associated with two or more linked loci, which are within five loci or 10 cM of their neighbors on the genetic map. Loci with same coordinates are considered as one position. After these steps, the Contig–Locus association table is reset by removing the anchored contigs and the loci to which they are anchored from the general pool of data to be compared. Then, assignments are made for the new ‘unique’ contigs that are now associated with loci that do not associate with any other contig. This last step is reiterated two more times, re-setting the Contig–Locus table after each iterative anchoring round.

For example, locus umc1976 associates with contig ctg14 and locus bnlg1247 associates with contigs ctg14 and ctg3931. At the first round of iteration, ctg14 is anchored to locus umc1976 because no other contig associates with umc1976. At the second round, after removing ctg14 and umc1976 from the Contig–Locus association, locus bnlg1247 becomes ‘unique’ because it only associates with ctg3931. Therefore, ctg3931 is anchored to bnlg1247. Anchoring confidence is highest in the first two steps of anchoring when strict one contig–one locus and one contig-linked loci associations are made. Confidence is lower for iterative rounds of anchoring.
Markers types and physical objects

<table>
<thead>
<tr>
<th>Marker types</th>
<th>Physical objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>D cDNA</td>
<td>B Associated with a BAC hit</td>
</tr>
<tr>
<td>F FISH</td>
<td>b Associated with a BAC hit not assembled into a contig</td>
</tr>
<tr>
<td>I InDel</td>
<td>G Associated with any numbers of contig</td>
</tr>
<tr>
<td>N SNP</td>
<td>C Associated with one contig</td>
</tr>
<tr>
<td>O overgo</td>
<td>c Associated with two contigs</td>
</tr>
<tr>
<td>R RFLP</td>
<td>m Associated with three and more contigs</td>
</tr>
<tr>
<td>S SSR</td>
<td>a Contig(s) that are anchored on a locus</td>
</tr>
</tbody>
</table>

Marker types:
- cDNA, complementary DNA or expressed sequence tag (EST); FISH, fluorescent in situ hybridization; InDel, insertion/deletion polymorphism; SNP, single nucleotide polymorphism; overgo, oligonucleotide hybridization probe; RFLP, restriction fragment length polymorphism; SSR, simple sequence repeat; BAC, bacterial artificial chromosome.

Denoting relationships among locus, marker and BAC contig

Once a contig is anchored to a genetic locus, then a data string is generated to describe the relationships among the locus, marker and BAC contig, so that these details can be presented in the integrated map display. Each character in the data string represents one marker type or physical object that associates with a given locus. Table 1 shows the possible marker types and physical objects, with corresponding abbreviations for each. In the data string, the character ‘|’ is used as a separator. The left side of character ‘|’ presents marker types, physical objects and whether this locus has an anchored contig; the right side presents which marker type(s) were used to detect the physical objects. For example, locus umc1354 on IBM chromosome 1 has a string SDOBbCa[S]. In this case, ‘SDOB-bCa’ on the left side of ‘|’ means that umc1354 is identified by marker types SSR (S), cDNA (D), and overgo (O); and this locus is associated with physical object BAC (B), singleton BAC (b) that is not assembled into a contig, and one contig (C) that is anchored on the integrated map (a). The letter ‘S’ on the right side of ‘|’ means that the marker type SSR was used to identify BAC(s) and contig. This string has multiple purposes in the application and will be described more in the implementation section.

IMPLEMENTATION

iMapDB

iMapDB was designed and developed to fulfill several functions: (i) to hold genetic map information; (ii) to serve as a repository for physical map data from FPC; (iii) to integrate genetic and physical map data via common markers that have been genetically mapped and detect BAC contigs; (iv) to provide the assembled physical map information based on the contig anchoring rules; and (v) to supply an integrated map resource for data retrieval.

iMapDB includes two levels, raw data tables and viewing tables (Fig. 1). The data model for the raw data tables incorporates a variety of data types that describe the genetic and physical map components and the relationships among them. These data types are incorporated as objects that are linked to the genetic information (i.e. Locus_Object and Genetic_Map_Info) and as objects that are associated with the physical map information (i.e. BACs and Contigs). The relationships are defined by common markers (probes) which associate with both the genetic loci and BAC contigs (in object Physical_Genetic_Rel). Diagrams of the schema are available at http://www.maizemap.org

The viewing tables serve as a data access resource for the iMap Viewer and iMap Search by providing the integrated map information in specialized formats, a format required by the iMap Viewer. The iMap Viewer and iMap Search retrieve data from these pre-calculated tables in order to optimize the system performance.

iMapDB is updated periodically as new mapping data are obtained. To make updating the database easy and reduce the need for manual commands, a graphical interface for updating was created, written in Visual Basic with implementation of Perl and SQL scripts. On a development server, the database is updated by clicking a button on the interface to automatically drop the existing database, create a new database, load the genetic map data from MaizeDB and MMP-LIMS, split and parse the physical data from FPC file, and generate the integrated genetic and physical map information. Once verified, the new database is moved to production.

iMap Viewer

The iMap Viewer features side-by-side displays, by chromosome, of the genetic map and associated anchored BAC contigs (Fig. 4). It was developed by adapting software GIOT and new features were made for displaying the integrated genetic and physical map. Improvements include display of a physical map with unambiguously anchored contigs, pull-down menus that let a user highlight loci identified by a particular genetic marker type or physical object type, and dynamic operation in a database environment.

The genetic map is displayed on the left (Fig. 4A), with locus names listed from top to bottom, based on their map coordinates (cM). Locus names in plain type represent on-frame loci for which order has been firmly established. Locus names in italics represent off-frame loci for which order is unknown. On the physical map (Fig. 4A, right), anchored contigs associated with the genetic loci are displayed as circles in two vertical ranks corresponding to their association with ‘on frame’ or ‘off frame’ loci. If multiple, linked loci associate with the same anchored contig, a vertical line is drawn through multiple circles to indicate that these represent the same contig and a pair of red lines is
**Fig. 4.** The iMap graphical display and linked contig information. (A) Display of IBM chromosome 3 in iMap Viewer. Left side presents genetic map; all loci identified by marker type SSR were highlighted in blue by selecting SSR from pull-down menu button ‘Highlight loci identified by marker type’ in the top. Right side shows physical map as circles corresponding to anchored BAC contigs. Central areas show which marker types are associated with the highlighted genetic locus and whether this locus associates with a contig. (B) Contig information in a pop-up window. Clicking on any marker type or contig in the middle of iMap Viewer causes a pop-up window to appear with detailed information and links to external data resources. (C) Details of contig number 163 in WebFPC.

Drawn horizontally to the map ruler to indicate how many cM this contig covers. The contig number and its associated locus name appear when the mouse is positioned over a contig circle.

Several highlighting and display features in the iMap Viewer utilize the data string (discussed earlier) denoting the relationship of each locus to markers and physical object types. First, the string allows highlighting of all genetic loci identified by a specific marker type or physical object type. When a type is selected from one of two pull-down menus on the interface (Fig. 4A, gray area), all genetic loci of the selected type will be highlighted in blue based on matches with the character in the string. Second, the string allows display of all marker and physical object types associated with a given locus when a locus name on the genetic map or a circle (contig) on
the physical map is clicked; these object types are shown in the central areas labeled Marker Type and Physical Object (Fig. 4A). Last, the string indicates which marker(s) provide anchors between the genetic and physical maps; marker(s) that only associate with the genetic locus are colored in blue and marker(s) that also associate with BAC contig(s) are labeled in magenta.

The iMap Viewer allows users to search for a locus on the displayed map and show more details. To search the current chromosome map, the exact locus name is entered in the search box and a message will appear at the upper right corner of the genetic map indicating whether the locus is found. If found, the locus name and the circle(s) corresponding to its anchored contig will be highlighted in magenta, and marker and physical object types will appear in Marker Type and Physical Object areas. For each locus, details of genetic marker and contig are displayed in pop-up windows (Fig. 4B) by clicking on a Marker Type or Physical Object. Details on genetic markers are retrieved via links to external databases. Details on BAC contigs can be viewed via a link to WebFPC (Fig. 4C).

The iMap Viewer operates dynamically in a database environment. Java servlet and the JConnect database driver are used for communication between iMap Viewer applet and iMapDB database. Once data are retrieved from the database and delivered to the applet, the iMap Viewer controls the drawing of the maps and displays the information.

iMap Search

The search utility assists users in finding where genes of interest are located and in selecting a chromosome to view. The iMap Search, written in Perl/CGI/DBI, allows searches in iMapDB by genetic locus, probe or sequence accession number (GenBank/EMBL/DDBJ). Wildcard searching is available if the exact name of the entity is not known. When only one record is found, the result is shown directly in the iMap Viewer with the retrieved locus highlighted in magenta. When two or more records are found, the search results are presented as text reports that include relevant map positions of locus, probe, sequence and any BAC contig associated with those positions. The text reports, which can be downloaded to the users desktop, offer hotlinks to data resources (MaizeDB and GenBank) and to the iMap Viewer showing relevant chromosomes with the selected locus highlighted.

DISCUSSION

The iMap utility is designed for comprehensive data management, analysis, integration, search and visualization of the marker-based integrated map. The iMap database is not only populated with data from the genetic map and BAC fingerprinting, but also holds the integrated map information obtained by applying contig-anchoring rules (Figs 2 and 3). The iMap Viewer presents side-by-side chromosomal views of the genetic map and associated BAC contigs (Fig. 4) and offers details on genetic and physical map objects via pop-up windows. The iMap Search generates reports of relevant genetic locus, marker or sequence-based map positions and any BAC contig associated with those positions.

An advantage of iMap over other map viewers currently available for grass genomes is the anchoring algorithm that is an essential component of the iMapDB and is used to facilitate the least ambiguous integration of genetic and physical mapping data. WebChrom and Gramene offer alternative ways to view genetic and physical mapping data, but because they do not use the iMap anchoring rules, they do not offer the same integrated data set that iMap does. Specifically, WebChrom uses a ‘majority rule’ in assigning contigs to loci; therefore the anchoring pattern is not the same as in iMap. Currently, Gramene does not apply any rules to the physical map data and displays all contigs associated with a marker. Thus, the integrated map information in Gramene is also different from that presented by iMap. As the integrated map evolves through manual editing efforts now underway, we expect that the integrated map data presented by iMap, WebChrom and Gramene will coalesce.

iMap can be adapted to other genomics efforts where the main resources are genetic and physical map data. iMapDB schema incorporates attributes common to all genetic and physical mapping efforts. The main modifications needed for applying iMap to other genomes would include: adding new marker types, i.e. AFLP or CAPS; altering the Java source code to display these marker types in iMap Viewer and to link to the data resources; and selecting a suitable database driver. Because of this facile adaptability, iMap may be the integrated map tool of choice for other plant genome projects.

A future application of iMap will be in comparative genome analysis. Used together with cMap (Fang et al., 2003, http://www.agron.missouri.edu/cMapDB/cMap.html) that allows comparison of maps based on locus, marker and sequence accession, it will be useful for evaluating syntenic relationships between species. Currently, iMap permits sequence searches via sequence accession number, but an enhancement is planned that would allow users to enter a DNA sequence, BLAST against any public maize sequence, and obtain BLAST scores and maize map locations for related maize genes. These data could be linked to iMap such that users entering a rice or sorghum sequence could learn which maize sequences are related and where they map and further check syntenic relationships between species using the cMap viewer at MaizeDB (http://www.agron.missouri.edu/cMapDB/cMap.html) which displays comparative genetic maps for maize, rice and sorghum (Fang et al., 2003). This approach could reveal syntenic and possible orthologous relationships and would complement the resources available from genomic efforts in rice and sorghum (Klein et al., 2000; Draye et al., 2001; Chen et al., 2002).
ACKNOWLEDGEMENTS

We thank our collaborators Cari Soderlund, Fred Engler and Rod Wing at the University of Arizona and Andrew Paterson and John Bowers at the University of Georgia. We are grateful to the Rice Genome Program, Tsukuba, Japan, for providing the source code for GIOT. Names are necessary to report factually on available data; however, neither the USDA nor the University of Missouri guarantees nor warrants the standard of the product, and the use of the name implies no approval of the product to the exclusion of others that may also be suitable. This research was funded by the National Science Foundation Plant Genome Grant DBI9872655.

REFERENCES


