An Integrated Web Resource for Cotton

Alan R. Gingle,* Hongyu Yang, Peng W. Chee, O. Lloyd May, Junkang Rong, Daryl T. Bowman, Edward L. Lubbers, J. LaDon Day, and Andrew H. Paterson

ABSTRACT

“The Cotton Diversity Database” (http://cotton.agtec.uga.edu) is a Web resource for cotton (Gossypium spp.) phenotypic and genomic data. A primary goal for this resource is to provide both a useful management tool for breeders and other applied scientists and a research tool for genetic and genomic scientists. The resource contains four interface suites that include displays for each of the available phenotypic or genomic data types. These display suites are accessible via the genotype portal, a search interface that allows users to begin with a cotton accession and obtain all available data. The phenotypic data displays include graphical views of overall cultivar performance with means and between group standard deviations indicated in an easy-to-interpret graphical manner for common trial measures such as lint yield, micronaire, etc. The genomic data displays include interactive graphical views of genetic map and diversity data types. Genetic map data is displayed in both traditional linear and two-dimensional comparative dot plot formats. Genetic diversity data is displayed in an interactive tree-based format showing degrees of similarity among genotypes. The data are stored in Oracle relational database (RDBMS) schemas containing tables and views for data storage, auto-calculated statistics and display parameters. The searchable RDBMS provides flexibility for a wide range of query and search options as well as integration paths amongst the various data types.

World cotton commerce of about $20 billion annually is made possible by the unusual ability of certain members of the Gossypium genus to produce single-seeded seedborne epidermal trichomes (lint fibers) that can reach as much as 5 to 6 cm in length. In the USA alone, cotton is typically grown on about 5 million hectares per year, more than all crops except maize (Zea mays L.), wheat (Triticum aestivum L.), or soybean [Glycine max (L.) Merr.] (USDA-NASS, 2002). The value of cotton fiber grown in the USA is typically about $6 billion yr⁻¹, while cottonseed oil and meal add another $500 million yr⁻¹. Cotton fiber exports account for about $4 billion yr⁻¹ of the U.S. trade surplus (Meyer et al., 2005) and business revenue stimulated by this crop in the U.S. economy is estimated at about $120 billion (National Cotton Council, www.cotton.org/econ/world).

The past decade of genomics research has yielded a cotton molecular map comprised of 3347 loci detected by 2669 probes (Rong et al., 2004). The cotton map has been linked to at least 295 QTLs, influencing 26 traits related to plant growth, development, and morphology (Wright et al., 1999; Jiang et al., 2000), reproductive biology (Lan et al., 1999); fiber yield and quality (Jiang et al., 1998; Paterson et al., 2002; Chee et al., 2005a, 2005b; Draye et al., 2005), disease resistance (Wright et al., 1998) and the preservation of productivity and quality under drought stress (Saranga et al., 2001, 2004). At least four high quality BAC libraries exist, for G. barbadense L., G. raimondii Ulbr. (A.H. Paterson, unpublished, 2004), and two strains of G. hirsutum (Tomkins et al., 2001). Genomic tools have been employed toward development of strategies for genetic improvement (El-Zik and Thaxton, 2001; Xia et al., 2001), analyses of genetic diversity (Van Esbroeck et al., 1998; Brubaker and Wendel, 2001; McCarty and Percy, 2001; Bowman and Gutierrez, 2003; Lubbers et al., 2004), and many other areas.

Research efforts focused on analyses of genetic diversity and the development of genomic resources to establish the context for diversity studies are particularly important for cotton (Van Esbroeck et al., 1998; Benedict and Altman, 2001; Lubbers et al., 2004; Van Beel et al., 2005). The US cotton gene pool is genetically impoverished as a result of at least four genetic bottlenecks. All cottons cultivated in the USA are tetraploid, thought to have arisen in the New World about 1 to 2 million years ago as a result of an unusual hybridization event between an invasive A-genome diploid genotype and an indigenous D-genome diploid. Polyploid cottons are thought to be monophyletic, representing the first genetic bottleneck. A second bottleneck was associated with the domestication of cotton from a small subset of the wild genotypes. A third bottleneck was imposed by human sampling of tetraploid cotton genotypes from their center of diversity in Mexico and Central America, and spread northward into the USA, and also to China, India, Egypt, Australia, and other countries. Finally, a fourth bottleneck results from the recent widespread deployment of transgenes in only a small number of genotypes.

Growing concern about genetic vulnerability of the cotton gene pool to a wide range of biotic and abiotic hazards is exemplified by recent investigation of trends in yield improvement. A “Blue Ribbon Committee” of public and private scientists convened by the National Cotton Council (Helms, 2000) determined that indeed, U.S. cotton yields had peaked in 1992 and by 1998 had reached a disturbing rate of decline of about 16.8 kg ha⁻¹ yr⁻¹ (3.3% annual rate). Accompanying this yield decline, year-to-year variations in yield were almost four
times greater in the period from 1980 to 1998 than in 1960 to 1979. This increased volatility in yield translates directly into higher risk for the grower.

The over-exploitation of a few genetic backgrounds has been implicated as the principal contributor to the erosion of the cotton genetic base. Breeding programs have repeatedly employed a few genetic backgrounds to generate new cultivars, resulting in a narrow genetic base for further cultivar enhancement (May et al., 1995). Additionally, growers have planted large areas to highly related cultivars resulting in a high level of field genetic uniformity. For example, Van Esbroeck et al. (1998) reported that by 1995 field uniformity (mean coefficient of parentage weighted by hectarage) had reached as high as 38% in the midsouth U.S. production area. Such a lack of genetic diversity may have contributed to reduced yields that have been associated with a reduction in the plasticity of crops to respond to pathogens, infestations, climactic changes, and agriculture practices (Walsh, 1981; Van Esbroeck et al., 1998; Manifesto et al., 2001; Lubbers et al., 2004). Also, others have noted the importance of natural outcrossing (Percy and Wendel, 1990; Wang et al., 1995) and intentional introgression that built on the wealth of variability in the wild species leading to better agronomic and resistance properties during cotton’s early domestication (Altman, 1993; Benedict and Altman, 2001).

A solution to the problem of genetic vulnerability requires integration of research, education, and extension activities, gaining the partnership of stakeholders to take steps that can be implemented quickly with discernible results, and elicit scientific and behavioral changes that outlive the funding period. We are working toward this goal by three major activities:

Creation of a Web-based resource to provide objective information about relatedness of genotypes, as a management tool for producers to reduce field genetic uniformity, and a research tool for scientists to glean new information useful for crop improvement.

Development of new interspecific gene combinations useful to improve elite cotton genotypes, simplified genomic tools needed to support use of these genes, and permanent genetic stocks and molecular tools that will improve understanding of the entire complement of genes that determine the productivity and quality of cotton.

Outreach activities to engage stakeholders and educate them in the usage of these Web-accessible data and resources.

In this article we describe an integrated Web database resource for phenotypic, molecular marker, genetic mapping, and genetic diversity data types that has been developed as one component of a USDA IFAFS project (PI: A. Paterson) and in partnership with collaborating labs, providing data and advice. Our approaches to database and interface design are described along with key features of each interface suite. Special focus is given to database and interface features that enhance integration of the varied data types without sacrificing the breeder or researcher’s ability to obtain answers to a wide range of focused queries. Currently, the database contains a limited number of accessions; however, we are hopeful that the combination of genetic diversity and performance data will prove useful to breeders and genomics researchers alike. Thus far trial performance data have been contributed in a number of formats that have been parsed and imported into the database. Genetic map and diversity data were contributed by the collaborating labs associated with the underlying IFAFS project.

**DESCRIPTION**

The Cotton Diversity Database is entered via its main page (Fig. 1A) that contains general information about genetic vulnerability and diversity, two themes of the underlying IFAFS project that led to its development. Sidebar (Fig. 1B) and horizontal menus allow users to navigate to project-related information, data or informatic tool downloads, other cotton resources, and our interface suites for phenotypic, genetic diversity, genetic map and comparative map data. The search and display interfaces are illustrated in Fig. 2 through 6. They are linked to search forms that allow accession name, performance rank, and genetic diversity based searches that can extend throughout the range of data types contained within the resource’s database. Each accesses an Oracle RDBMS back-end that contains schemas for its associated data types. MySQL compatible versions have been developed for some of these interfaces (Yang et al., 2005; Yang and Gingle, 2005) and more are in development. The packages are available via the Downloads page and come with database schema information and creation scripts for those who wish to implement their own database instances. The search and display functions of each interface suite are described below.

**Phenotypic Data**

The performance trial interface (Fig. 2) provides an interactive display of cotton performance data for traits that relate to lint fiber elongation, fineness, length, strength, and yield. Currently, data for 567 cotton accessions from trials conducted at the University of Georgia (Day et al., 2001, 2002, 2003, 2004, 2005), Mississippi State University (Creech et al., 1998–2001; Wallace et al., 2004, 2005) and North Carolina State (Bowman, 2004, 2005) are stored in our database. Also, performance trial data for the 2005 growing season are currently being added. Descriptive information on the underlying database schema is available via the site’s download page. Accessions are grouped first by state and then by trial year. Within a trial year, a further division is made by other categories according to a range of criteria that can be selected via the menu (Fig. 2A). For example, menu selections include a list of all, the top 25 performers or the best with respect to any of the stored performance traits. The listed accessions link to a data display that includes overall performance (Fig. 2B) and performance vs. location data (Fig. 2C) displayed in a color-coded bar chart format with both the averages.
and between group standard deviations indicated. The data are plotted relative to the minimum and maximum performers, indicated by the lower and upper scale numbers respectively (Fig. 2B, 2C). A tab for general trial information that shows the investigators, locations, and years of the trial, (Fig. 2D) is provided as well. Performance data for a particular accession is also available through the genotype portal (Fig. 6A).

**Genetic Map**

Our genetic map interface for cotton (Fig. 3) provides an interactive display of the genetic map data (Rong et al., 2004) that is stored in the associated database schema. It allows one to select and browse the cotton linkage groups or chromosomes (Fig. 3A). For example, a typical inquiry would proceed in the following fashion. After linkage group or chromosome selection, the genetic map is displayed with probes organized alongside the map based on their locations. Clicking at the location generates a list box containing the probe names that are at the selected location. For locations with large numbers of probes, a scroll bar appears in the box. For queries that are focused on a particular map object, the probes page provides an alternative entry point. Molecular probe sequence can be obtained with a mouse click (Fig. 3B) as well as via the molecular probe search form (Fig. 3C). Also, the appropriate link to GenBank is included in the probe information form (Fig. 3B) whenever GenBank records are available. Overview information about probes and number of associated linkage groups is also shown on this page. A downloadable install package has been developed for a version of this interface, which is MySQL compatible (Yang et al., 2005), and it is available via the site’s download page. The package includes a database schema description and creation script along with the other required files and instructions.

**Comparative Map**

The OxfordGrid display (Fig. 4A) provides a comparative overview of genetic map data and is being applied to display comparative views of the cotton genetic maps developed by Rong et al. (2004) for the D and AD genomes. It is composed of a matrix of cells with each representing a pairwise comparison of map data for two linkage groups or chromosomes. These are arranged along the axes with one forming grid columns and the other grid rows. Letters or numbers, for compactness, represent linkage groups or chromosomes and a legend...
appears adjacent to the grid on the right (Fig. 4A). The dots inside each cell represent pairs of map locations that are associated with a common molecular marker. The degree and pattern of synteny/colinearity between the two groups is manifested in the cell’s dot density and structure. Individual linkage group pairs can be viewed by a mouse click over the selected grid cell (Fig. 4A). This results in a close-up view of the selected linkage groups (Fig. 4B) that is interactive in the following way. The associated map locations are indicated by red-colored dots with each being a link to detailed linkage group (Fig. 4B and 4C) and probe information for the selected locations. Also, detailed display instructions are available. A downloadable install package has been developed for a version of this interface, which is MySQL compatible (Yang and Gingle, 2005), and it is available via the site’s download page. The package includes a database schema description and creation script along with the other required files and instructions.

**Genetic Diversity**

The genetic diversity interface suite (Fig. 5, 6D) provides an interactive display of molecular marker–RFLP locus based genetic diversity data for a range of *G. hirsutum* accessions with a *G. barbadense* outgroup (Lubbers et al., 2004). Our genetic diversity display generates phylogenetic tree or cladogram views of the entire accession set (Fig. 5A) and subtree views are available with a mouse click over the desired subtree root (Fig. 5B). Also, user adjustable real-time tree cutting and associated clustering is available with clusters distinguished by color-coded taxonomic nodes.

**Pedigree Data**

The pedigree information display (Fig. 6B) contains an accession’s name, pedigree (Calhoun et al., 1997), source information and a legend, describing notation. The data are based on the report by Calhoun et al. (1997), which also contains a range of additional cultivar and accession data types where applicable (e.g., experimental designation, plant variety protection number, Crop Science registration number, year of release, originator, and reference). We have implemented links to GRIN (www.ars-grin.gov/npgs/) where additional information is available. Also, descriptive information
on the underlying database schema is available via our site's download page.

**Search Interfaces and Menus**

Search features and interfaces complement the various data displays by offering rapid accession, performance trait or molecular probe based searches for focused data views. These include the performance trait search menu, the mapped probe search form, and the genotype portal. The performance trait search menu (Fig. 2A) facilitates accession selections that begin with trial state and year. Once a particular study is chosen, accessions can be selected based on their performance ranking in any of the range of measured fiber traits. The mapped probe search form (Fig. 3C) facilitates probe selection based on name or number of map locations. The genotype portal (Fig. 6A) facilitates searches based on accession name and available data types. Its form allows users to obtain lists of accession names, beginning with a particular letter or containing a particular text string, and then to filter these lists by requiring the presence of pedigree, performance, and/or diversity data. The resulting accession list contains links that launch displays or interfaces for each of the available data types. The form's pedigree link launches the pedigree information display (Fig. 6B). Its performance link launches the performance trial display (Fig. 6C) described above. The genetic diversity-related links launch the phylogenetic tree interface (Fig. 6D).

**Sample Database Searches**

The database's search and query features are best illustrated through examples, and two examples are described below. One illustrates a search beginning with a molecular marker or probe and seeking general probe information and genetic map context including information on possible duplication and synteny of the context region. The second illustrates a search beginning with a cotton accession of interest and seeking to identify genetically distant cultivars that have acceptable values of desired performance traits. For the molecular marker-based search, an investigator interested in a particular mapped probe can begin by selecting it from Probes by Letter menu option of the IntegratedMap Display for Cotton (Fig. 3A). Once selected, the genetic map display can be launched to display any of the probe's map locations.
and GenBank record are available, the last via a link to NCBI. At this point, an investigator may wish to determine if the probe maps to a duplicated or collinear region in one of the cotton genomes. They can obtain this information by selecting the genome and location of interest in the OxfordGrid display (Fig. 4A). For the accession-based search, an investigator interested in identifying accessions that are genetically distant from a particular accession can begin with the Genotype Portal search tool (Fig. 6A) and first determine if genetic diversity data are available. If so, a link will allow them to launch the TreeView interface (Fig. 5, 6D) with the selected accession highlighted. The investigator can then navigate the phylogenetic tree to locate genetically distant accessions. If trial performance data are available, the performance of the distantly related accession(s) can be obtained for the range of performance traits. Pedigrees can also be obtained when available.

**Approaches to Data Integration**

Our approach to data integration relies on links between the distinct data types and their associated graphical user interfaces. These exist for genetic diversity, pedigree, and performance trial data types as illustrated in Fig. 6. Cotton accessions form the basis for most of the links with both performance (Fig. 6C) and genetic diversity (Fig. 6D) data keyed to it. The associated graphical user interfaces play a major role in data integration through synergies that we have built into them to maximize the potential of the links between the various genetic and performance data sets. These synergies have taken the form of one interface establishing a broader context for the data displayed in another. An example is the genetic diversity tree display that can provide a diversity context for performance variations within vs. between diversity tree clusters (Fig. 2, 5). A second example is the OxfordGrid interface that provides a comparative context for genetic map data, highlighting regions of colinearity or synteny between genomes (Fig. 4). Maximizing the potential of these data type links through interface synergies increases the reach of queries into the store of information. For example, user queries associated with an accession can lead to infor-

---

**Fig. 4.** The OxfordGrid interface is employed to generate comparative pairwise views of the cotton AD and D genomes. (A) It opens with a dot-plot style view of the cotton genetic maps. A horizontal menu above the dot-plot allows the selection of either inter- or intragenomic pairwise views that are based on molecular markers common to both genomes and locations. (B) A mouse click over a selected cell launches an enlarged view of the chosen combination of linkage groups and chromosomes and (C) in this display, a mouse click over a probe location pair launches detailed views of the associated cotton genetic map regions.
Information on both neighboring and distantly related accessions (Fig. 6D).

**Database and System Overview**

This resource’s Web front-end makes extensive use of Web interfaces that are based on dynamic image maps to achieve both a wide range of responsive graphic features and compatibility with a broad range of browser types and versions. These have been developed in either C#/ASP.Net or Java/JSP environments and are hosted on either IIS or Apache-Tomcat Web/Application servers within a multi-tiered architecture consisting of Web, application, and database architectures. The resource’s back-end is an Oracle RDBMS that hosts schemas developed for performance trial and genetic or genomic data types. Advanced Oracle features like materialized views (Dawes and Thomas, 2002) and table partitioning (Thomas and Bryla, 2002) are employed to achieve rapid query performance when large table sizes are involved.

To complement the Web accessible data search and display tools, we have developed MySQL compatible versions of our genetic map related Web interfaces (Yang et al., 2005; Yang and Gingle, 2005) and this effort is being extended to our other informatic tools. These installation packages are freely available via the download page (http://cotton.agtec.uga.edu/Cotton/downloads.aspx; verified 21 June 2006).

Each data type and associated Web display has distinct requirements for schema structure and query performance. For example, the schema for performance trial and pedigree data contains seven tables. Six are related to pedigree and trial data for each of the commonly measured cotton fiber and yield traits as well as test plot, general trial, and author citation information. The remaining table contains display parameters such as column titles and units. In addition, four database views maintain online performance statistics like between group standard deviations as described above and shown in Fig. 2B. Other examples are genetic diversity and map data. Genetic diversity data are contained in a schema that includes tables for primary RFLP-molecular probe analysis, genetic distance matrix, and phylogenetic tree structure data. These facilitate many of the interdata type links and

---

Fig. 5. The suite of phylogenetic tree display screens. In this resource, it is used to display genetic diversity data in an interactive fashion. (A) The initial display shows the complete phylogenetic tree associated with molecular marker analysis based genetic diversity data. (B) Subtree display can be generated by a mouse click over the selected node as indicated by the dashed arrows. Also, an interactive tree-cutting feature provides a color-coded indication of clusters associated with the user selectable depth for tree cutting.
dynamic genetic diversity tree cutting and clustering that is described above and shown in Fig. 5B. This schema is under development and will be released with an upcoming package for the associated interface tool. Genetic map data are contained in a schema that includes eight tables for organism information and map data for molecular probes, mapped polymorphisms and anchored BAC contigs; though, the schema is not currently populated with the latter two data types for cotton. In addition, synteny data is stored to enable the comparative map views of our OxfordGrid interface (Fig. 4). The genetic map and synteny related schemas are included with their respective interface installation packages, available via the download page.

**FUTURE TRENDS**

In addition to ongoing database and Web interface improvements, we plan data access and software availability enhancements. With respect to data access, we have recently established Web services for some of the existing cotton data types and plan to expand these to facilitate a more seamless access to our data resources.

Also, we are exploring the development of a Web interface to streamline performance data entry. With respect to software availability, we are continuing the development of downloadable installation packages for our informatic tools. At the time of this writing, a MySQL compatible package is being developed for our genetic diversity interface (Fig. 5) and it will be made available, including database schema information and creation scripts, via the site’s download page. These packages will enable other research teams to use our data analysis and presentation tools with their data sets. Also, “Reducing the Genetic Vulnerability of Cotton” (PI: A. Paterson, University of Georgia) has joined with “Comparative Evolutionary Genomics of Cotton” (PI: J. Wendel, Iowa State) to establish a Web portal (http://gossypium.info) that will provide a common entry port for these and other participating Cotton resources.

**ACKNOWLEDGMENTS**

The authors wish to thank the collaborating labs for providing data, access to Web resources, and advice. We wish to note that this effort is supported as one component of an
REFERENCES


