Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between Gossypium hirsutum and G. hirsutum × G. barbadense populations

Article in Molecular Genetics and Genomics · March 2015
Impact Factor: 2.73 · DOI: 10.1007/s00438-015-1021-y

4 authors, including:

Mingzhou Song
New Mexico State University
57 PUBLICATIONS 721 CITATIONS

Jinfa Zhang
New Mexico State University
188 PUBLICATIONS 1,737 CITATIONS

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Jinfa Zhang
Retrieved on: 23 May 2016
Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and *G. hirsutum × G. barbadense* populations

Joseph I Said · Joseph A Knapka · Mingzhou Song · Jinfa Zhang

Received: 12 December 2014 / Accepted: 24 February 2015 © Springer-Verlag Berlin Heidelberg 2015

Abstract

In cotton quantitative trait loci (QTL), studies are focused on intraspecific *Gossypium hirsutum* and interspecific *G. hirsutum × G. barbadense* populations. These two populations are commercially important for the textile industry and are evaluated for fiber quality, yield, seed quality, resistance, physiological, and morphological trait QTL. With meta-analysis data based on the vast amount of QTL studies in cotton it will be beneficial to organize the data into a functional database for the cotton community. Here we provide a tool for cotton researchers to visualize previously identified QTL and submit their own QTL to the Cotton QTLdb database. The database provides the user with the option of selecting various QTL trait types from either the *G. hirsutum* or *G. hirsutum × G. barbadense* populations. Based on the user’s QTL trait selection, graphical representations of chromosomes of the population selected are displayed in publication ready images. The database also provides users with trait information on QTL, LOD scores, and explained phenotypic variances for all QTL selected. The CottonQTLdb database provides cotton geneticist and breeders with statistical data on cotton QTL previously identified and provides a visualization tool to view QTL positions on chromosomes. Currently the database (Release 1) contains 2274 QTLs, and succeeding QTL studies will be updated regularly by the curators and members of the cotton community that contribute their data to keep the database current. The database is accessible from http://www.cottonqtldb.org.

Keywords *Gossypium hirsutum* · *Gossypium barbadense* · QTL · Fiber quality · Database · CottonQTLdb

Introduction

There are approximately 50 species in the cotton genus (*Gossypium*) found around the world in warm tropical climates (Wendel and Cronn 2002). Of the 50 species, 5 are tetraploids and assigned to the AD genome containing an A and a D subgenome each with 13 pairs of chromosomes (Wendel and Cronn 2002). The cultivated tetraploid AD genome species *G. hirsutum* (AD1) and *G. barbadense* (AD2) are grown commercially around the world for...
natural fibers. *G. hirsutum* is renowned for having a high lint yield while *G. barbadense* is known for having superior fiber strength, length, and fineness (Zhang et al. 2014a, b). In an effort to combine the high lint yield of *G. hirsutum* and the superior fiber quality traits of *G. barbadense* many studies on quantitative trait loci (QTL) have been performed using interspecific *G. hirsutum × G. barbadense* populations. The intraspecific *G. hirsutum* population is also the subject of many QTL studies for a variety of trait types since most cotton breeding programs are intraspecific in nature.

QTL are chromosomal regions which contribute cumulatively to a trait with varying percentages of phenotypic variance from each QTL. QTL for fiber quality traits have been the focus of many studies in both *G. hirsutum* and *G. hirsutum × G. barbadense* populations. With the large number of QTL studies in both populations, it is difficult to manually analyze all of the QTL collectively. QTL meta-analysis brings QTL from multiple studies together and is better suited to addressing similarities and differences in QTL placement on chromosomes between populations. Combining previous studies allows regions to emerge that are more or less rich in QTL called QTL clusters for different traits and hotspots for the same trait. This allows cotton breeders to focus their efforts only on regions with the most QTL with the highest percentages of phenotypic variance.

In this study intraspecific *G. hirsutum* and interspecific *G. hirsutum × G. barbadense* QTL from Said et al. (2013) along with QTL from a comparative meta-QTL analysis between *G. hirsutum* and *G. hirsutum × G. barbadense* populations by Said et al. (2014), and four succeeding QTL studies containing 140 QTL were organized into a relational QTL database containing 2274 QTL. Therefore, the current Cotton QTLdb database (Release 1) includes 2274 QTL from 92 publications covering 66 different QTL trait types, and it allows the user to view individual QTL from the Said et al. (2014) comparison study between intraspecific *G. hirsutum* and interspecific *G. hirsutum × G. barbadense* populations in addition to other succeeding QTL studies data. The rationale behind the database is to provide a beneficial tool to the cotton community and to marker assisted selection (MAS) breeding programs. MAS is the process where breeders use the presence of known molecular markers which are close in proximity to known QTL to make selections. The database as a tool will allow breeders to select trait QTL of interest, display them on chromosomes, together with statistical data for each QTL, and allow the user to make an informed decision about the most important markers to be focused on in their study. Of interest to breeders are LOD scores for each QTL used which indicate the likelihood of a QTL in a region and explained phenotypic variance ($R^2$) scores for each QTL in the database. A confidence interval for each QTL is also provided to the user along with the position of the QTL determined by the original study. Users have the option of selecting specific trait QTL from either or both population, download statistical data on selected QTL, and display QTL on chromosomes using a visualization software developed in-house.

Currently, there are three operational cotton genetics databases, i.e., CottonGen (http://www.cottongen.org) (Yu et al. 2014), the Cotton Genome Database (http://www.cottondb.org) (Yu et al. 2009), and the Cotton Marker Database (CMD) (http://www.cottonmarker.org) previously called the Cotton Microsatellite Database (Blenda et al. 2006). None of the databases provide a comprehensive QTL analysis tool which is capable of comparing QTL between populations and graphically rendering QTL on a genomewide linkage map. Instead, the three databases offer a QTL search feature which displays information on individual QTL, making it difficult to conduct a comparative study between QTLs and populations. The databases are simply not specifically designed to show the user all known QTL of a specific trait on chromosomes graphically. Recently, in the cotton community of the US, the information in the two earlier databases, i.e., Cotton Marker Database and Cotton Genome Database, is being migrated to the CottonGen Database (http://www.cottongen.com). While the other two databases have not been modified or updated and are out of date, the CottonGen Database is understandably regularly curated. However, search functions (with three options—trait name, published symbol and QTL label) for QTL provided by the database are highly limited, and most of reported QTL were not curated. For example, searching the CottonGen Database with 988 QTL for 25 traits (http://www.cottongen.org/data/qtl) for ‘fiber_strength’ (as a trait name provided by the database) returned with only 20 QTL in a table format, while the recent comprehensive analysis by Said et al. (2014) included 124 and 70 QTL reported for the same trait in *G. hirsutum* and *G. hirsutum × G. barbadense* populations, respectively. Therefore, there is a need for a specialized cotton QTL database with frequent updates as a resource for the cotton community.

The new QTL search tool presented here should be very useful for the cotton community which allows researchers to view and compare the placement of known QTL on both intraspecific *G. hirsutum* and interspecific *G. hirsutum × G. barbadense* populations. We have also made our data output compatible with Biomercator (Arcade et al. 2004) software for use in meta-analysis studies. Previous meta-analysis studies by Rong et al. (2007), Lacape et al. (2010), and Said et al. (2013) quickly became obsolete as succeeding QTL studies were published. One year after Said et al. (2013) was published, the comparative meta-analysis by Said et al. (2014) provided an additional 911 QTL to the cotton community and offered them with a
useful comparison between two heavily studied populations. Only a few months after the comparative meta-analysis by Said et al. (2014), 140 additional QTL from 4 publications were published and added to the current CottonQTLdb database. The goal of this study is to keep the data from Said et al. (2014) current and useful to the cotton community by providing regular updates with QTL from succeeding QTL studies. In addition to search functions, the Cotton QTLdb database also contains a data submission service which allows researchers to add their published data to the database for use by the cotton community. Data submissions from the cotton community will ensure that new QTL will be added to the database, keeping it relevant and updated in years to come.

Materials and methods

The database was written in Python, an open source programming language. The Python code was used to present and interpret data, synthesize chromosome maps, and interact between the web portal and the database. All data and tools are hosted on an Amazon cloud server and managed by the Cotton Breeding and Genetics research group at New Mexico State University. The database was designed to operate optimally using Google Chrome or Mozilla Firefox browsers and was not designed to be used with Macintosh browsers such as Safari. The database was also not designed to be used on iPads or other mobile devices such as smart phones and may result in a distorted view of the links and graphics on the site.

Data curated by Said et al. (2014) and four succeeding QTL studies were used to create a comparison between intraspecific G. hirsutum and interspecific G. hirsutum × G. barbadense populations with a total of 2274 QTL for 66 different QTL trait types from 92 publications listed in Supplementary Table 1a for intraspecific G. hirsutum populations (An et al. 2010; Cai et al. 2013; Chen et al. 2008, 2009; Fang et al. 2014; Ge et al. 2008; Gore et al. 2014; Guo et al. 2008; Gutierrez et al. 2010; Hu et al. 2008; Jia et al. 2011; Feng et al. 2009; Kong et al. 2011; Kumar et al. 2012; Li et al. 2006, 2008, 2010, 2012; QingZhi et al. 2013; Liu et al. 2010, 2012a, b, 2013a, b, c; Lopez-Lavalle et al. 2012; Mei et al. 2014, 2013; Zhiyuan et al. 2013; Nusurat et al. 2012; Qin et al. 2008, 2009; Romano et al. 2009; Shen et al. 2005, 2006a, b; Sun et al. 2012; Ulloa et al. 2009; Wang et al. 2006, 2007a, b, c, 2009, 2010, 2011, 2012a, b, c; Wu et al. 2009; Xu et al. 2010; Yang et al. 2007, 2009; Yao et al. 2010; Yin et al. 2002; Zhang et al. 2005, 2006, 2009, 2010, 2011a, b, c, 2012, 2013, 2014a, b; Zhao et al. 2014), and Supplementary Table 1b for interspecific G. hirsutum x G. barbadense populations (Chee et al. 2005a, b, b; Draye et al. 2005; Fang et al. 2013; Gutierrez et al. 2011; Jiang et al. 1998, 2000; Lacape et al. 2005, 2010, 2013; Mei et al. 2004; Paterson et al. 2003; Saranga et al. 2004; Shen et al. 2010, 2006a, b; Su et al. 2013; Ulloa et al. 2009, 2011, 2013; Wang et al. 2008, 2012a, b, c, 2013, 2014; Wright et al. 1999, 1998; Yang et al. 2008; Yu et al. 2012, 2013a, b; Zhang et al. 2014a, b). The study by Said et al. (2014) and the succeeding studies gathered QTL data from the 92 publications with each QTL being placed according to the chromosome location in cM from the publication. QTL data was collected in Biomercator V3 excel format with ten columns of data for each QTL. The map name, QTL name, chromosome number, trait type, LOD score, phenotypic variance ($R^2$), single interval mapping (SIM), position in centimorgans (cM), and flanking left and right confidence intervals (CI) were recorded for each QTL. The first column denoting the map name is up to the discretion of the user as different maps will be named differently by users. The seventh column denoting SIM is used by Biomercator V3 to determine if a QTL was declared using single interval mapping (SIM), which is rarely used and almost always has a value of ‘N’ standing for ‘No’. The user is able to select specific traits and populations and download QTL data for each population type in excel format. The comma separated value (csv) file is compatible with Biomercator V3 if the user wishes to use the meta-analysis functions of the program outside of the database.

Database description

Site navigation

The database homepage (CottonQTLdb, Release 1.0) contains information about the projects which helped launched the Cotton QTL database and references. From the homepage or any other page the user may select from five links found in the upper right corner of the screen as seen in Fig. 1. The second link ‘Trait Descriptions’ takes the user to a table which contains a complete list of the traits found in the database along with descriptions of each trait. The third link ‘Data Sources’ takes the user to a complete list of publications which have contributed QTL data to the database. The fourth link ‘QTL Search’ is the main tool of the database which takes the user to a screen that gives the user an option to load their own QTL data in excel format, or continue to the trait selection screen. From the trait selection screen, specific traits and populations can be selected. Following the user’s selection of traits by population, the QTL which meet the user’s criteria are displayed in tabular format displaying the QTL name, chromosome number, LOD score, $R^2$, position, and confidence intervals (Fig. 2). By clicking the ‘Display Linear Chromosome Diagram’ the
user can view the selected QTL on chromosomes graphically. Figure 3 outlines the steps from the ‘QTL Search’ option needed to navigate the site. The final link ‘Data Submission’ contains information necessary for the user to submit their QTL data to the database and contact information of the curators.

Chromosome images

Images of chromosomes are generated using an in-house software written in Python which links the QTL file containing QTL positions and statistical data with the [“Guazuncho2” (G. hirsutum) × “VH8-4602” (G. barbadense)] map file. Figure 4 shows chromosome c1 from the interspecific G. hirsutum × G. barbadense population with QTL displayed to the left as red lines with QTL names, the population type, positions and confidence intervals displayed in cM along the map. The length of the line indicates the confidence interval with the top of the line placed on the cM position of the left confidence interval and the bottom of the line at the right confidence interval on the chromosome map. Confidence intervals are also displayed next to the QTL name along with the exact position of the QTL according to its publication of origin. Users
can benefit from these confidence intervals as they denote the certainty of the position of each QTL. With narrower confidence intervals denoting higher certainty of placement than broader ones, users can make informed decisions on which QTL are most important in the region. To the right of the chromosome, the molecular markers from the [“Guzuncho2” (G. hirsutum) × “VH8-4602” (G. barbadense)] map are displayed. The images generated by the database provide publication ready diagrams for users conducting their own QTL investigations. Users can save chromosome images generated by the database by right clicking the image and saving it as a jpeg file.

The CottonQTLdb provides an excel csv file of QTL selected by the user for use with other QTL software available outside of the database. Meta-analysis can be performed using Biomercator V3 or a higher version with the excel csv output of the database with the previously mentioned 10 QTL file columns. Biomercator V3 software was used by the Said et al. (2013, 2014) meta-analysis studies to detect mQTL regions and calculate Akaike information criterion (AIC) scores for each mQTL. For this reason, the QTL data was preserved in Biomercator V3 format, and may be used with Biomercator V3 or higher. Biomercator V3 provides a meta-analysis feature which uses AIC scores to indicate the likelihood that specific QTL are part of a meta-QTL (mQTL) region. The program presents the best four AIC scores and models for the QTL and chromosome selected as images. The program indicates different mQTL regions in different colors for the user to assess the positions of mQTL. Figure 5 illustrates an mQTL prediction by Biomercator V3 which may be generated by users given the QTL excel file provided by the Cotton QTLdb database.

Data submission

The cotton QTL database accepts data curated from either published papers or data in the process of publication. Data submitted should follow the format of the database with the 10 Biomercator V3 compatible columns in excel csv format as described. Users wishing to submit their data should email the curators of the database with an attachment containing the data in csv format along with a brief description of the nature of the study. The population used in the study, any relevant references, and trait descriptions of QTL if not already described in the database should be included in a cover page with submissions. Users may submit populations other than intraspecific G. hirsutum and interspecific G. hirsutum × G. barbadense populations thus expanding the population types found in the database. More information on database submissions and curator contacts can be found following the data submission link at http://www.cottonqtldb.org. The QTL data submitted should either be in the process of publication or already published for consideration to maintain the integrity of the database. Following an approval process the curators will contact the user and add the data to the database if applicable.

If users do not want to make their data publicly available they may use the informal submission tool to privately view
their data alongside selections in the database. Informal submissions of private data are only seen by the user and not officially added to the database and require no approval process by the curators. Informal submission data can be uploaded for use only by the user by using the ‘browse’ button on the main cotton QTL web services page. The curators are not responsible for data submitted informally and claim no responsibility for the results of that data.

**Discussion**

The CottonQTLdb database is a specialized resource to the cotton community providing a QTL visualization and comparison tool previously unavailable. The database home screen provides users with the QTL by trait type search option, submission options, and resource links as seen in Fig. 1. The resource links include general information,
on cotton and the database, QTL trait descriptions, data sources used in the database, and an option to contact the curators. The versatility, comprehensiveness and simplicity of the Cotton QTLdb database make it an ideal resource for cotton geneticists and breeders alike. Our software presents users with the most recent and comprehensive QTL data gathered spanning 16 years of research since 1998. The interface for selecting QTL on the QTL search page provides QTL data and graphical representations of each QTL along the genome displayed on each chromosome. Users may select either intraspecific *G. hirsutum* or interspecific *G. hirsutum × G. barbadense* population or both populations using the search feature and upload their own data in conjunction which offers the user a lot of versatility and analysis options. The software is an excellent analysis tool for researchers conducting their own QTL studies.

The database currently contains QTL from both intraspecific *G. hirsutum* and interspecific *G. hirsutum × G. barbadense* populations. The user has the ability to select specific traits for one or both the populations for comparative studies involving QTL placement. This feature may be used to compare differences in QTL placement, QTL clusters, or QTL hotspots between populations for comparative studies. Using the number of QTL, their placement, and relative LOD and $R^2$, the two populations can be compared for their QTL content and placement. It was shown in previous meta-analysis studies (Rong et al. 2007; Lacape et al. 2010; Said et al. 2013, 2014) that QTL are not evenly distributed over the tetraploid cotton genome. Using this data provides hints to likely positions of QTL in future studies around QTL clusters and hotspots. Breeders can also compare the two populations for QTL placement and determine which anchoring markers are of commercial interest in each population. With the LOD scores, $R^2$, and confidence interval data provided for each QTL, breeders can determine which QTL are most beneficial to focus on in MAS breeding programs or further mapping studies. Regions with QTL having higher $R^2$ scores in narrower confidence intervals are more likely to produce better MAS results with lower $R^2$ scores and less certain QTL placement with wide confidence intervals.

Due to the high volume of QTL in the database, users have the option of only selecting traits of interest in specific populations while performing a QTL search. The QTL search tool provides the user with a complete list of all traits available for each of the two populations. The user may then select one or more traits from either or both populations for comparison. QTL comparisons are made possible by selecting either specific traits to compare within a population or selecting the same trait in both populations to view similarities and differences. This feature allows researchers to better compare QTL results from various studies.

The cotton QTL database provides a valuable new resource and tool to the cotton community as it contains most updated cotton QTL than any database available. For the first time, the cotton community will have a QTL database which is designed to elucidate similarities and differences in QTL trait type and placement between the two most heavily researched cotton populations, i.e., *G. hirsutum* and *G. hirsutum × G. barbadense*. Since the database is not focused on a specific trait type, it offers a resource to a wide array of cotton scientists with diverse interests. The user friendly interface and resources make the cotton QTL database a practical tool for both geneticists and breeders alike.

**Fig. 5** Biomercator V3 generated image of chromosome c3 in the *G. hirsutum* population with lint percentage QTL mQTL regions

---

The versatility, comprehensiveness and simplicity of the Cotton QTLdb database make it an ideal resource for cotton geneticists and breeders alike. Our software presents users with the most recent and comprehensive QTL data gathered spanning 16 years of research since 1998. The interface for selecting QTL on the QTL search page provides QTL data and graphical representations of each QTL along the genome displayed on each chromosome. Users may select either intraspecific *G. hirsutum* or interspecific *G. hirsutum × G. barbadense* population or both populations using the search feature and upload their own data in conjunction which offers the user a lot of versatility and analysis options. The software is an excellent analysis tool for researchers conducting their own QTL studies.

The database currently contains QTL from both intraspecific *G. hirsutum* and interspecific *G. hirsutum × G. barbadense* populations. The user has the ability to select specific traits for one or both the populations for comparative studies involving QTL placement. This feature may be used to compare differences in QTL placement, QTL clusters, or QTL hotspots between populations for comparative studies. Using the number of QTL, their placement, and relative LOD and $R^2$, the two populations can be compared for their QTL content and placement. It was shown in previous meta-analysis studies (Rong et al. 2007; Lacape et al. 2010; Said et al. 2013, 2014) that QTL are not evenly distributed over the tetraploid cotton genome. Using this data provides hints to likely positions of QTL in future studies around QTL clusters and hotspots. Breeders can also compare the two populations for QTL placement and determine which anchoring markers are of commercial interest in each population. With the LOD scores, $R^2$, and confidence interval data provided for each QTL, breeders can determine which QTL are most beneficial to focus on in MAS breeding programs or further mapping studies. Regions with QTL having higher $R^2$ scores in narrower confidence intervals are more likely to produce better MAS results with lower $R^2$ scores and less certain QTL placement with wide confidence intervals.

Due to the high volume of QTL in the database, users have the option of only selecting traits of interest in specific populations while performing a QTL search. The QTL search tool provides the user with a complete list of all traits available for each of the two populations. The user may then select one or more traits from either or both populations for comparison. QTL comparisons are made possible by selecting either specific traits to compare within a population or selecting the same trait in both populations to view similarities and differences. This feature allows researchers to better compare QTL results from various studies.

The cotton QTL database provides a valuable new resource and tool to the cotton community as it contains most updated cotton QTL than any database available. For the first time, the cotton community will have a QTL database which is designed to elucidate similarities and differences in QTL trait type and placement between the two most heavily researched cotton populations, i.e., *G. hirsutum* and *G. hirsutum × G. barbadense*. Since the database is not focused on a specific trait type, it offers a resource to a wide array of cotton scientists with diverse interests. The user friendly interface and resources make the cotton QTL database a practical tool for both geneticists and breeders alike.
Future developments of the database may include more cotton species and populations, more trait types, and additional QTL. Since the database is open for independent submissions by users, the number of populations, traits, and QTL within the database can expand as research progresses. Currently, only the two most heavily studied commercially interesting *G. hirsutum* and *G. barbadense* populations are featured in the database which is likely to change with future updates. Future expansions may include other intraspecific or interspecific populations which can be used for comparative studies. Currently, there are 66 trait types listed in the QTL database; however, future studies may expand on these trait types as high-throughput phenotyping becomes a reality. With the support of the cotton community, the database should progress alongside the most current genome research being published to provide a useful tool to the cotton community as a whole. Potentially when the AD1 and AD2 genomes are sequenced, the database can include a link to correlate a QTL with its physical location containing potential genes on the genomes. Future expansions may also include a phylogenetic comparison feature for QTL between populations using sequence data.

In addition to its inclusion of all published QTL in the database, the cottonqtldb.org has new functions that are missing in other existing cotton genome databases. These new functions include QTL sources, QTL graphic presentation and its comparison between mapping populations, and submission of users’ own QTL data. Since the two other currently publicly accessible cotton genome databases (http://www.cottonmarker.org, http://www.cottondb.org) in the US are being consolidated or migrated into the CottonGen Database (http://www.cottongen.org), http://www.cottongen.org/external_resources. In the future, the search functions and other features with frequent updates of reported QTL in http://www.cottonqtldb.org can be incorporated into the CottonGen Database if desired by the community with resources available to support the integration.

Acknowledgments This study was funded in part by New Mexico Agricultural Experiment Station.

Conflict of interest The authors declare that they have no conflict of interests.

Ethnical standard This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References


in upland cottonseeds with two different genetic systems across environments. Euphytica 188:453–463


Liu H, Quampah A, Chen J, Li J, Huang Z, He Q, Zhu S, Shi C (2013a) QTL mapping based on different genetic systems for essential amino acid contents in cottonseeds in different environments. PLoS One 8:e57531


Yao JB, Zhang YS, Chen W, Luan MB, Feng ZD, Guo XM (2010) Genetic dissection of the introgresed genomic components from Gossypium barbadense L. that contribute to improved fiber quality in Gossypium hirsutum L. Mol Breed 32:1–16


