COTTONGEN: A CENTRAL DATA REPOSITORY AND ANALYSIS RESOURCE FOR COTTON COMMUNITY

J. Yu
S. Jung
C.H. Cheng
T. Lee
P. Zheng
K. Buble
J. Crabb
J. Humann
H. Hough
Washington State University
Pullman, WA
D. Jones
Cotton Incorporated
Cary, NC
T. Campbell
USDA-ARS
Florence, SC
J. Udall
USDA-ARS
College Station, TX
D. Main
Washington State University
Pullman, WA

Abstract

CottonGen is a genomics, genetics and breeding database for the cotton community. It provides a comprehensive collection of data, various analysis tools, Breeding Information Management System, and links to external resources of interest to cotton researchers. CottonGen currently contains 28 (16 tetraploids and 12 diploids) annotated genome sequences; 1,520,001 genes, 112 genetic maps; 575,850 markers; 6,234 QTLs; 19,652 germplasm; metabolic pathways for 13 species (AD1-AD5, A, D, G, F, and *Gossypioides kirkii*); 25,150,265 SNP and 12,484 SSR genotype measurements; 529,050 phenotype measurements (mainly from RBTN and NCGC projects), 45,214 images (mainly of NCGC); and synteny data for 28 genomes with links to genes, mRNA, orthologs and function. Analysis and visualization tools in CottonGen include genome browser JBrowse, Synteny Viewer, MapViewer, CottonCyc, BLAST+, and the Breeding Information Management System, an online system to manage and analyze private breeding data. All the data are integrated within CottonGen and can easily be queried out through various CottonGen's search engines. This presentation will illustrate how to use various resources in CottonGen to find relevant information and perform further data mining.

Introduction

CottonGen aims to serve as the central data repository and analysis resource for the cotton research community by providing access to an integrated, comprehensive, online information system for basic, translational and applied cotton research (Yu and Main, 2015). Initiated in 2012 under the funding supports from industry, government, and academic sources, the database has superseded a cotton genome database (CottonDB) (Yu *et al.,* 2006, 2012) and a cotton marker database (CMD) (Blenda *et al*, 2006) and expanded to include annotated transcriptome and genome sequences and enhanced tools for easier data sharing, mining, visualization and retrieval of cotton research data. Another feature developed in CottonDB but adopted by CottonGen is the hosting of the website for the International Cotton Genome Initiative (ICGI). ICGI is a non-profit organization created in 2000 to increase knowledge of the structure and function of the cotton genome for the benefit of the global community. The CottonGen database is constructed using the open-source Tripal genome database toolkit (Sanderson *et al.*, 2013), which merges the power of Drupal, a popular web Content Management System, with that of Chado (Mungall *et al.*, 2007, 2011), a community-derived database schema for storage of genomic and genetic data (Yu *et al*., 2014).

Database Description

Web Interface

The CottonGen interface has been designed to provide easier access points to data and tools such as the Major Species Quick Start and Tools Quick Start featured on the homepage (Figure 1, https://www.cottongen.org/) . The Major Species Quick Start allows users to view which types of data are available for a cultivated species of interest and provides links to access these data. Similarly, Species pages under the 'Species' navigation menu provide the same information for cultivated and other species with whole genome sequence data. The Tools Quick Start is organized into genomics, genetics, breeding, and general sections; each section provides links to appropriate pages to access available data, search, tools, or general information about CottonGen. New features that can quickly familiarize users to CottonGen data and functionality include the dynamic data overview page where users can browse the current data types and numbers in CottonGen and short video tutorials. Tutorials are available for site overview, species pages, Breeding Information Management System (BIMS) and all the search pages.

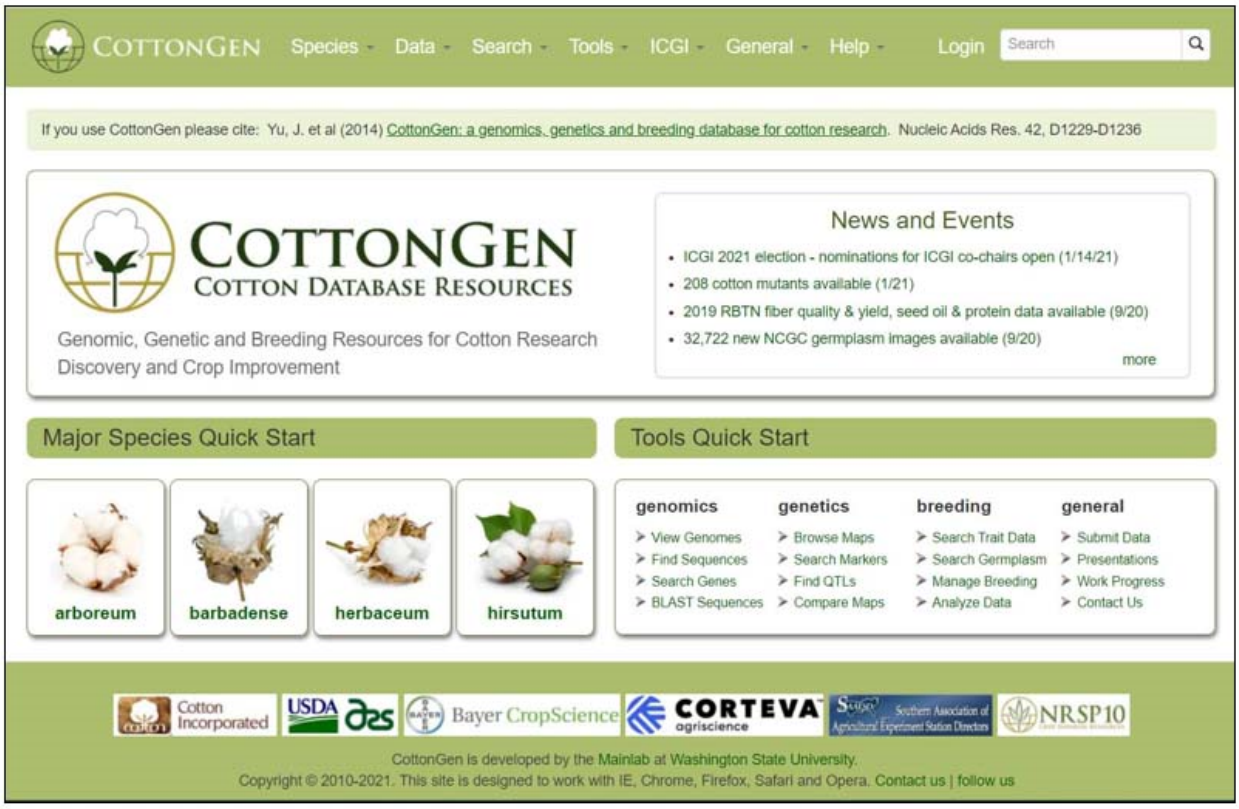

Figure 1. CottonGen Home Page (www.cottongen.org)

Data Available

Data available in CottonGen:

- Whole genome assemblies and annotations of 8 diploid and 5 tetraploid species (some with multiple assemblies done by different research groups). They are: *G. arboreum (3), G. raimondii (3), G. kirkii (1); G. australe (1), G. herbaceum (1), G. longicalyx (1), G. thurberi (1), G. turneri (1), and 5 tetraploid species: G. hirsutum (9), G barbadense (4), G. tomentosum (1), G. mustelinum (1), G. darwinii (1)*;
- 1,138,797 genes and 1,678,307 mRNAs from whole genome assemblies and parsed from NCBI nucleotide sequences;
- 214,180 RefTrans for *G. hirsutum, G. barbadense, G. arboreum, G. raimondii* by CottonGen Team;
- Total 575,852 genetic markers including 459,825 SNPs, 101,049 SSRs;
- 115 genetic maps with 130,088 Loci from 110 genetic maps, 2 consensus maps, 2 bin maps, and 1 silico map;
- 6,497 QTLs and 208 mutants, including 4,013 quality traits, 1,192 agronomical trait, 273 biotic stress traits,

and 189 biochemical traits;
- 85 species, they are: the 4 cultivated species, 53 wild species, and 28 cross or lab made diploid, tetraploid, and hexaploid hybrids;
- 19,640 germplasm including data from collection and sub-collections from US-NCGC, US-GRIN, China, and Uzbekistan;
- 524,709 phenotypic scores from the US regional breeders' tests; the trait evaluations from US, Uzbekistan, and China germplasm collections; and the data collected from various QTL studies;
- 45,214 images, 44,999 from NCGC digital characterizations project;
- 16,382 publications, including journal articles, and conference proceedings, patents, book chapters and theses;
- 645 colleagues contact information and 419 ICGI memberships

Tools Available

CottonGen Tools include: a) BIMS, the Breeding Information Management System, is a secure and comprehensive online breeding management system developed for the generic Tripal Database Platform which allows breeders to store, manage, archive and analyze their private breeding program; b) NCBI BLAST, basic local alignment search tool for rapid sequence comparison; c) CottonCyc, Cyc databases are constructed using PathwayTools and using gene models from the specified whole genome assemblies; d) JBrowse, a visualization tool for genome browse; e) MapViewer, a graphical tool for viewing and comparing genetic maps developed by the MainLab at Washington State University for Tripal databases; f) Primer3, a widely used program for designing PCR primers (PCR = Polymerase Chain Reaction); g) Sequence Retrieval, a tool to download nucleotide sequences including chromosomes, scaffolds, genes, mRNAs, transcript coding sequences, RefTrans contigs and unigene contigs; and h) SyntenyViewer, a tool to view conserved syntenic regions among publicly available cotton genomes were analyzed by CottonGen.

Community Resources and Activities

Community resources mainly include a) 'Cotton Trait Ontology' (a set of standardized and structured vocabularies for cotton traits, www.cottongen.org/data/trait_ontology) aims to provide a central location with controlled vocabulary for phenotypic traits to improve discussion and collaboration among research groups within the cotton community. The terms in the Cotton Trait Ontology were developed from trait evaluation data within five germplasm collections from four countries and from QTL-trait association data obtained from over one hundred peer-reviewed publications. The vocabulary was established in 2016 by CottonGen with input from Drs. Lori Hinze, Richard Percy, and Russell Kohel (USDA-ARS, College Station, TX) and thereafter is continue the incremental validation along with the new data imported to CottonGen.; b) 'Community Projects' (www.cottongen.org/data/community_projects) is a platform to host large community projects' information and data including summary, objectives, participants, protocols, developed resources, publications, etc. c) 'Community Archives' (www.cottongen.org/data/community_archives) is used as the repository of community's historical information, keynote presentations, etc. currently the archive contains files of "The 70th Anniversary of The Cotton Improvement Conference 1948-2018" – presented by Dr. McCarty, "Some Perspectives from 50 Years of Cotton Breeding" – presented by Dr. Bourland, and "Annual Cotton Genetics Research Awards" history and Recipients.

Community Activities include BIMS Workshops hosted at 2016, 2017, and 2019 Beltwide Cotton Conferences; the ICGI Officers biennially Elections hold on each odd-year; and ICGI biennially Research Conferences hosted on each even-year.

Concluding Remarks and Future Direction

CottonGen is the consolidated cotton genomics, genetics and breeding database for the cotton community. It aims to provide a central repository for public and private cotton genomics, genetics and breeding data; to provide data mining opportunities via intuitive online tools; and to provide opportunities for data sharing and communication in the cotton community. It is constructed using the open-source Tripal genome database toolkit, which merges the power of Drupal, a popular web Content Management System with that of Chado, a community-derived database schema for storage of genomic and genetic data. Data types in CottonGen include maps and markers, whole genome assemblies and annotations, gene and sequences with analyzed data, taxonomic and germplasm data and publication data. CottonGen maintains online resources for ICGI, a non-profit organization created as a global affinity group

with common goals and interests. From its released on 1 March 2012 to 31 December 2020, CottonGen quarterly visits increased from 2,310 by 1,348 unique users from 37 countries (2013 Q1) to 11,458 visits by 5,693 unique users from 114 countries (2020 Q4).

CottonGen will continue to integrate genomic/genetic/breeding data, further improve search tools and the development of BIMS, adding more genome data and PAN genome data, hosting more CottonGen Training Workshops.

Funding

Cotton Incorporated; the USDA-ARS Crop Germplasm Research Unit at College Station, TX; Southern Association of Agricultural Experiment Station Directors; Bayer CropScience; CORTEVA agriscience and NRSP10. Components of the infrastructure for CottonGen were created under funding for Tripal development for other databases (USDA NIFA [2009-51181-06036, 2009-51181-05808]). As these databases all use the same underlying Tripal infrastructure, source code was shared amongst all these databases. That code is freely available on the Tripal website at http://tripal.info. Funding for open access charge: CottonGen Grant.

Acknowledgements

We acknowledge with thanks our funding sources, the cotton research community providing data and feedback and the Tripal community of developers for developing and sharing Tripal modules and code, the AgBioData Consortium and US Land Grant Universities for support.

References

Blenda, A., J. Scheffler, B,. Scheffler, B., *et al.*. 2006. CMD: a cotton microsatellite database resource for Gossypium genomics, BMC Genomics. 7,132.

Mungall, C.J. and D.B. Emmert. The FlyBase Consortium. 2007. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics, 23(13), i337-46. doi:10.1093/bioinformatics/btm189

Mungall, C.J., C. Batchelor, K. Eilbeck. 2011. Evolution of the Sequence Ontology terms and relationships. J. Biomed. Inform, 44, 87-93.

Sanderson, L.A., S.P. Ficklin, C.H. Cheng, *et al*. 2013. Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. Database (Oxford), 2013,bat075.

Yu, J., L.L.Hinze, , J.Z. Yu, and R.J. Kohel. 2006. CottonDB.org - New website for cotton genome database. Proceedings of International Cotton Genome Initiative Research Conference, September 18-20, 2006, Brasilia, Brazil https://www.ars.usda.gov/research/publications/publication/?seqNo115=197886.

Yu, J., and D. Main. 2015. Role of Bioinformatics Tools and Databases in Cotton Research. P303-338. *In* D.D. Fang and R.G. Percy (eds.) Cotton. 2nd Edition, Agronomy Monograph 57.

Yu, J., R. Kohel, L. Hinze, J.Z. Yu, J. Frelichowski, S. Ficklin, D. Main, and R.G. Percy. 2012. CottonDB. Proceedings of the International Plant and Animal Genome Conference XX: January 14-18, 2012, San Diego, CA, USA. https://pag.confex.com/pag/xx/webprogram/Paper1715.html.

Yu, J., S. Jung, C.H. Cheng, F.P Ficklin, T. Lee, P. Zheng, D. Jone, R.G. Percy, D. Main. 2014. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res., 42(D1), D1229–D1236. First published on November 06, 2013, 10.1093/nar/gkt1064.