Building blocks for Developing Resource-Efficient Community Databases

[1]**Sook Jung**, [1]Chun-Huai Cheng, [1]Taein Lee, [1]Katheryn Buble, [1]Jodi Humann, [1]Ping Zheng, [1]Jing Yu, [1]Stephen Ficklin, [1]Dorrie Main
[1]Washington State University, Pullman, WA

Abstract

The unprecedented volume of big data being routinely generated for crop species, coupled with advanced technology enabling the use of big data in breeding, give further argument for the need to have crop community databases where relevant data is curated and integrated. Funding for such databases is, however, insufficient and intermittent, resulting in the data generated by tax-payers money being underutilized. While raising awareness of the importance of funding databases is crucial, practical solutions for efficient community database development are imperative. To meet the need for integrated database resources for various crop genomics, genetics, and breeding research communities, we have built five crop databases over the last decade using the Tripal open-source construction platform. In addition to using Tripal core modules and extension modules developed by other groups, we developed various extension modules that can be used by other groups. We describe the system and methods used for database construction, curation and analysis protocols, and the data and tools that are available in these five crop databases.

Introduction

Technological innovation in crop science, driven by "big data," allows scientists to generate and analyze extensive genomic, genetic, and breeding datasets. These datasets, ranging from whole genomes to phenotypic and genotypic data, become highly valuable when organized, annotated, and integrated with tools for browsing, querying, and analysis. Expert-managed community databases are essential for efficiently transforming data into reusable resources that drive research and application. Tripal software simplifies the creation of biological databases with minimal programming. This open-source toolkit combines Drupal and Chado, promoting data sharing and FAIR principles. Tripal is widely used, with numerous public sites and over 40 custom modules available, enhancing functionality. These modules include tools like Tripal BLAST for sequence alignment, Tripal MegaSearch for customizable queries, and TripalMap for genetic map visualization. Since 2010 we have developed six community databases using Tripal, offering integrated access to genomics, genetics, and breeding data, and analysis tools. Current databases include the Genome Database for Rosaceae, CottonGen, the Citrus Genome Database, the Pulse Crop Database, and the Genome Database for Vaccinium. The databases can be accessed at https://www.rosaceae.org/, https://www.cottongen.org/, https://www.citrusgenomedb.org/, https://www.pulsedb.org/, and https://www.vaccinium.org/, respectively.

Databases

The Rosaceae Genome Database (GDR), CottonGEN, Citrus Genome Database (CGD), Genome Database for Vaccinium (GDV), and Pulse Crop Database (PCD) serve 25 economically, nutritionally, and culturally important crops: fiber (cotton), fruit (apple, apricot, blackberry, cherry, peach, nectarine, pear, plum, raspberry, strawberry, blueberry, cranberry, orange, grapefruit, lime, lemon, tangelo, tangerine), nuts (almond), pulses (chickpea, fava bean, lentil, pea, common bean), and ornamentals (apple, cherry, rose). They are used in every U.S. state and territory and are the databases of choice of many researchers around the world. The databases contain whole genome assembly, genes/mRNAs, genetic markers, genetic maps, QTLs, MTLs (Mendelian Trait Loci), phenotype, genotype, haplotype, and publication data. Figure 1 summarizes the data type, data analysis, data integration and user interface in our databases. Users can access the data using various search pages and graphic interfaces. As shown in Figure 2, individual data pages such as gene/mRNA page and marker page, allows users to access extensively integrated data. When data is available, graphic interfaces such as MapViewer, Expression Heatmap, Synteny Viewer, and JBrowse can be accessed through the link from the gene/mRNA page and marker page. The graphic interfaces can also be accessed directly from the tool menu.

Software used for database construction

Tripal Core Module: Our crop databases were built using Tripal version 3.3. The storage backend was Chado version 1.2 installed on a PostgreSQL database (version 12.9). **Tripal Extension Modules:** Thirteen extension modules have been used in building our databases. Figure 1 shows some of the extension modules that we use. The detailed documentation can be accessed with the source code at https://tripal.readthedocs.io/en/latest/extensions.html.

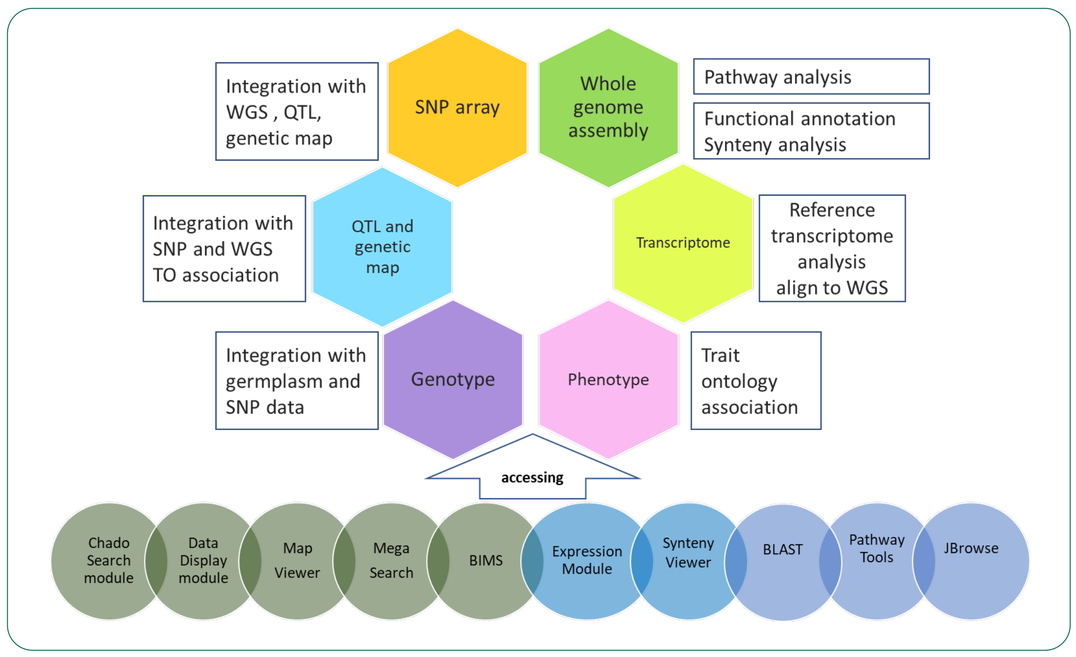

Figure 1. Data Types, Analysis, Integration, and User Interface. Green circles: Tripal modules developed by our group; blue circles: Tripal modules developed by other groups; purple circles: non-Tripal software

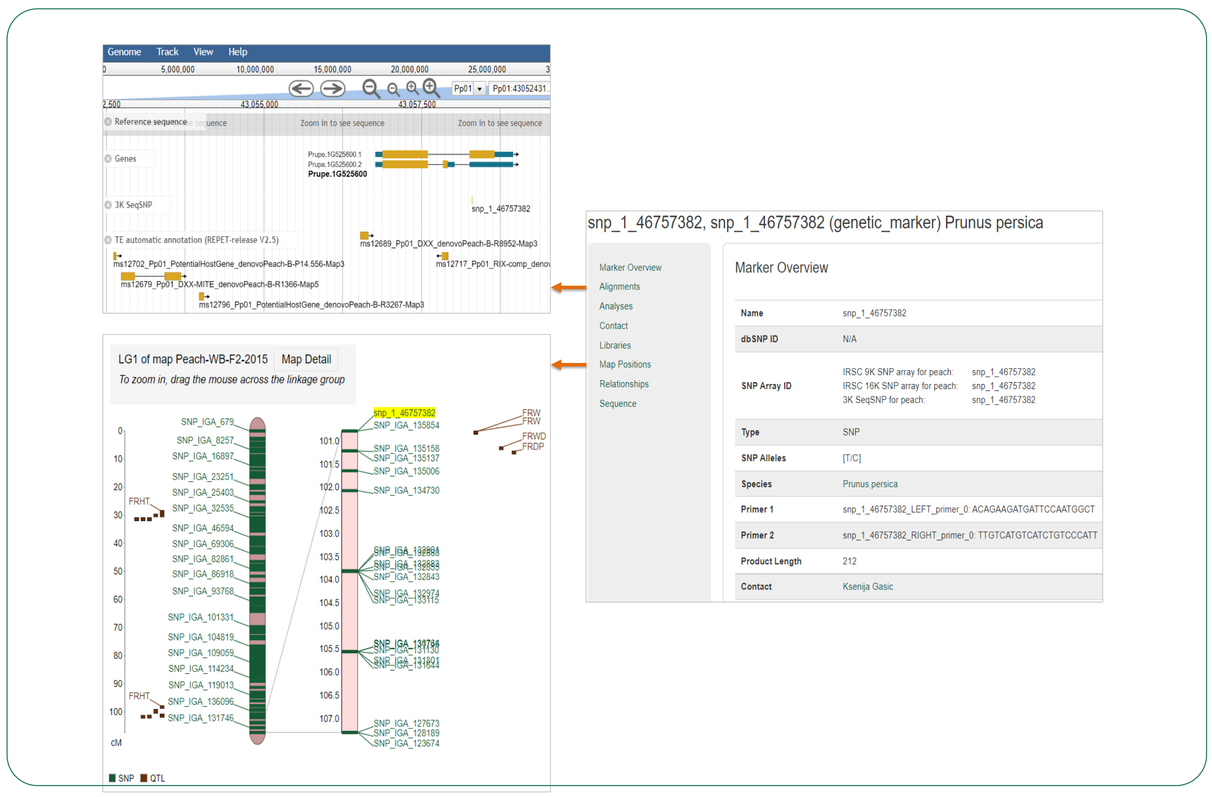

Figure 2. An example Marker page. JBrowse is hyperlinked from the Alignments section and MapViewer page is hyperlinked from the Map Positions section.

Data Curation, Analysis, and Integration

Whole Genome Data: Whole genome data is made accessible via JBrowse, BLAST, PathwayCyc, Synteny Viewer, Sequence Search, and Gene/Transcript Search pages. Our analyses include computational annotation of predicted genes, InterPro protein domains, GO terms, and homology to known proteins using BLASTX against Swiss-Prot. Synteny analysis is performed with MCScanX. PathwayCyc analyses are conducted using PathwayTools. **Transcriptome Data**: RNA-Seq and dbEST datasets create reference transcriptomes (RefTrans) for each genus or crop. RNA-Seq from NCBI SRA are used in the pipeline utilizing many software such as Trimmomatic, Trinity, CAP3, Bowtie, CH-HIT and Corset. Functional characterization is done as previously described. **Genetic Map, Marker, QTL, GWAS, Phenotype, and Genotype Data:** We curate and integrate data on molecular markers, genetic maps, QTLs, GWAS, phenotypes, and genotypes. Data from publications are entered into Mainlab Chado Loader (MCL) templates and uploaded via a web interface, standardizing marker and accession names before database integration. **Data integration** is achieved through analyses and manual curation. Genomic features and genetic markers and QTLs are integrated using shared markers, synteny analysis, and sequence alignments. Phenotype and genotype data are integrated using Trait Ontology and shared germplasm data. Figure 1 summarizes the data type, data analysis, data integration and user interface in our databases.

Conclusion and Future Directions

Building and maintaining our crop databases using the open-source Tripal genome database toolkit saved significant time and effort, allowing more time to be spent on data analysis and curation. Data curation, analysis and integration that keeps up to date with new publications is the key in usefulness of a crop database. This makes using an efficient database system crucial, especially for more orphan crops with limited funding. The control vocabulary-driven Chado also enables the database to accommodate new data types, which further reduces the potential cost for restructuring the database schema and interfaces. We plan to add more types of data in our databases such as gene annotation data in collaboration with community researchers and other crop databases.

Funding and Acknowledgements

Federal Funding
NIFA
NRSP10 CROP DATABASE RESOURCES
USDA ARS Agricultural Research Service
NSF
US Land Grant Universities

Industry Funding
Cotton Incorporated
USA Dry Pea & Lentil Council
Northern Pulse Growers Association

Community Support
AgBioDatabase
GMOD
Tripal
Rosaceae, Cotton, Vaccinium, Citrus and Pulse Researchers