Database, 2025, baaf005
DOI: https://doi.org/10.1093/database/baaf005
Original Article

DATABASE
The Journal of Biological Databases and Curation

Building resource-efficient community databases using open-source software

Sook Jung*, Chun-Huai Cheng, Taein Lee, Katheryn Buble, Jodi Humann, Ping Zheng, Jing Yu, Dorrie Main

Department of Horticulture, Washington State University, 303c Plant Sciences Building, Pullman, WA 99164-6414, USA

*Corresponding author. Department of Horticulture, Washington State University, 303c Plant Sciences Building, Pullman, WA 99164-6414, USA.
E-mail: sook_jung@wsu.edu

Citation details: Jung, S., Cheng, C., Lee, T. *et al.* Building resource-efficient community databases using open-source software. *Database* (2025) Vol. 2025: article ID baaf005; DOI: https://doi.org/10.1093/database/baaf005

Abstract

The unprecedented volume of big data being routinely generated for nonmodel crop species, coupled with advanced technology enabling the use of big data in breeding, gives further impetus for the need to have access to crop community databases, where all relevant data are curated and integrated. Funding for such databases is, however, insufficient and intermittent, resulting in the data being underutilized. While increased awareness of the importance of funding databases is important, it is practically necessary to find a more efficient way to build a community database. 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 an open-source database platform and software. 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.

Database URL: The Genome Database for Rosaceae (GDR, www.rosaceae.org), the Genome Database for Vaccinium (GDV, www.vaccinium.org), the Citrus Genome Database (CGD, www.citrusgenomedb.org), the Pulse Crop Database (PCD, www.pulsedb.org), and CottonGen (www.cottongen.org)

Introduction

Technological innovation has propelled crop science into an era of discovery and application driven by “big data,” where scientists routinely generate and analyze larger, ever more complex, and new types of genomic, genetic, and breeding datasets for both model and nonmodel crop species. The size and diversity of this big data are unprecedented, from numerous whole genome datasets within a species, expression data across environments, to phenotypic and genotypic data for tens of thousands of individuals. When the data are organized, annotated, and integrated with other associated data, and made available with supporting tools for browsing, querying, and analysis, significant value is added for researchers. Data become a reusable, enriched critical resource that supports new and accelerated research discovery and application. Community databases managed by experts with crop-specific knowledge and computational skills are the logical medium with which to meet these challenges efficiently. Community databases developed over the last 25 years provide this critical role for many crops (MaizeGDB [1], GrainGenes [2], Solanaceae Genomics Network [3], SoyBase [4], Genome Database for Rosaceae (GDR) [5], and CottonGen [6]).

The scientific and programming expertise required to build biological databases is very costly and is a major hurdle in developing these community databases for other crops with fewer funding opportunities. In addition, the databases need to be able to adapt and accommodate new data and new data types and have appropriate and easy-to-use tools to query and visualize the new data types and associated integrated data. Databases built with customized database code are often hard to modify or extend with incoming new types of data. This makes the databases complex to manage, unable to easily accommodate “big data,” and resource-intensive to leverage for other crops.

Fortunately, the Tripal software system [7, 8] enables the development of biological databases with relatively minimal programming. Available since 2010, Tripal is an ontology-based, open-source toolkit for the construction of online biological databases [9]. Tripal combines the content management system Drupal (https://www.drupal.org) and the controlled vocabulary (CV)-driven biological relational database storage backend Chado [10]. The CV is widely used to describe data types in Chado, making it relatively easy to introduce new types of data. In addition, Tripal V3 has been redesigned to facilitate data sharing through the systematic use of CVs [8]. The use of Tripal, as a result, encourages databases to practice findable, accessible, interoperable, and reusable data principles [11].

Tripal is increasingly being used in building databases with many public genomic and genetic databases using this platform [7, 8]. There are over 30 known public sites built in Tripal and over 130 public instances worldwide tracked by

Received 7 October 2024; Revised 19 December 2024; Accepted 17 January 2025

Table 1. Tripal extension modules used in our crop databases

Type	Name	Functionality
Breeding	BIMS—https://gitlab.com/mainlabwsu/bims [15]	A module for an online breeding management system that allows breeders to manage their public or private breeding program data
Data loading/collection	MCL—https://gitlab.com/mainlabwsu/mcl [16] (Jung *et al*)	A module that enables users to upload their biological data in the accompanied templates to the Chado database schema
	Tripal Blast—https://github.com/tripal/tripal_blast	A module that provides a basic interface to allow users to utilize NCBI BLAST+ [19] in the server
	Tripal JBrowse—https://github.com/tripal/tripal_jbrowse [23]	A module that provides integration between Tripal sites and pre-existing GMOD JBrowse [18] instances
	Tripal GenBank Parser—https://gitlab.com/mainlabwsu/tripal_genbank_parser	A module that downloads data from NCBI Entrez Nucleotide database, parses the downloaded flat file, and loads the results into the Tripal/Chado database
Searching	Mainlab Chado Search—https://gitlab.com/mainlabwsu/chado_search (Jung *et al.* 2017)	A module that enables advanced search function for data stored in a Tripal/Chado database. Various search instances are provided for data types such as Gene, Marker, QTL, Map, Trait, Stock, and Organism
	Tripal MegaSearch—https://gitlab.com/mainlabwsu/tripal_megasearch [12]	A module for flexible querying and downloading biological data stored in a Tripal/Chado database
	Tripal Elasticsearch—https://github.com/tripal/tripal_elasticsearch [14]	A module that integrates Elasticsearch to the Tripal site and also allows Cross-Site Querying for searching remote Tripal contents
Visualization/display	Mainlab Tripal Data Display—https://gitlab.com/mainlabwsu/mainlab_tripal (Jung *et al.* 2017)	A module that contains a set of Drupal/PHP templates that organize and extend the default display of the biological data. Supported data types include marker, QTL, germplasm stock, map, project, haplotype block, polymorphism, image, feature, and pub
	TripalMap—https://gitlab.com/mainlabwsu/tripal_map [13]	A module that displays genetic and genomic map data stored in Chado
	Synteny Viewer—https://github.com/tripal/tripal_synview	A module that displays synteny analysis results
Analysis/annotation	Tripal Analysis BLAST—https://github.com/tripal/tripal_analysis_blast	A module that provides a method for loading XML results from the NCBI BLAST program as well as visualization of the results
	Tripal Analysis InterPro—https://github.com/tripal/tripal_analysis_interpro	A module that provides a method for loading XML results from the InterProScan program as well as visualization of the results
	Tripal Analysis Expression—https://github.com/tripal/tripal_analysis_expression [14]	A module for loading, annotating, and visualizing NCBI Biomaterials and expression data

Drupal reports [7]. Since Tripal is an open-source and modular software, site developers can create their own extension modules to share with other Tripal users. Over 40 extension modules are available on the Tripal website (https://tripal.info/) for use by database developers. Some examples include the Tripal BLAST (https://github.com/tripal/tripal_blast/), which provides an intuitive user interface to the National Center for Biotechnology Information (NCBI) Blast+ tool; Tripal MegaSearch module, which allows interactive and customizable query and download [12]; the TripalMap module, which displays genetic map data [13]; the Tripal Analysis Expression module for loading, annotating, and visualizing NCBI Biomaterials and expression data [14]; and the Breeding Information Management System (BIMS) that allows breeders to store, manage, archive, and analyze their public or private breeding program data [15].

Since 2010 we have built six community databases (https://www.nrsp10.org) for specialty crops and cotton using Tripal. Our goal is to provide an integrated web-based community database resource providing access to publicly available genomics, genetics, and breeding data and data-mining tools in the most efficient way. This report describes the system and software that we used and the procedures for data curation, analysis, and data upload as well as the data and functionality of our databases. Our five current databases are the GDR [5], CottonGen [17], the Citrus Genome Database (CGD), the Pulse Crop Database (PCD), and the Genome Database for Vaccinium (GDV). 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/.

Hardware

While our databases could be hosted on a single server, they are hosted using a split architecture due to best practices, the scope of crop databases, and the additional websites hosted in our program. The application server, which hosts Tripal and Drupal, has 2 × 8-core Intel Xeon E5-2620v4 Processors, 128GB RAM, and 10TB usable storage. The database server, which hosts PostgreSQL, has 2 × 6-core Intel Xeon Gold 6128 processors, 8TB storage, and 256GB RAM. Ten additional databases and websites are hosted on these servers using the same split configuration.

As a starting point, a single site could be hosted using Intel Xeon E-series processors, 2–3TB storage, and 256GB RAM. Depending on client usage, performance needs, and data stored, it may be necessary to scale up from this minimum recommendation.

Software used for database construction

Tripal core module

Our crop databases were built using Tripal version 3.3. The Tripal Web Services module was enabled to allow data exchange with other Tripal sites. For data visualization, Tripal v2 legacy modules were enabled, in favor of the Tripal DS

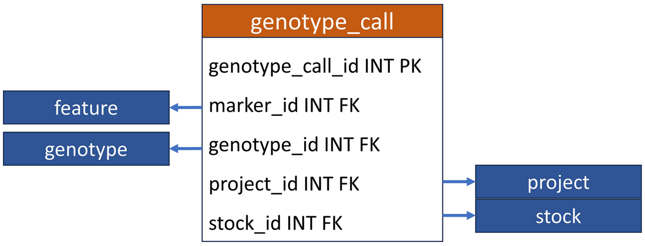


Figure 1. Schematic diagram of how genotype data are stored in Chado. Feature, genotype, project, and stock are existing tables of Chado and genotype_call is a custom table.

module. The storage backend was Chado version 1.2, installed on a PostgreSQL database (version 12.9). These modules are available at https://tripal.org.

Tripal extension modules

Thirteen extension modules have been used in building our databases. Table 1 describes the type, name, and brief functionalities of each module. The detailed documentation can be accessed with the source code at https://tripal.readthedocs.io/en/latest/extensions.html.

Non-Tripal software

Non-Tripal software such as JBrowse 1.16.4 [18], BLAST+ 2.12.0 [19], and Pathway Tools 25.0 [20] have been implemented.

Database schema and data loader

Tripal uses Chado as a backend schema. Genome and transcriptome data are stored in Chado following the documentation provided in Mungall and Emmert [10] and the documentation on the Generic Model Organism Database (GMOD) website. Genome and transcriptome data are loaded using the Open Biological and Biomedical Ontologies loader, Fast Adaptive Shrinkage Threshold Algorithm loader, and Generic Feature Format loader that are provided by the Tripal core module and the loader provided by the Tripal Analysis InterPro module. The synteny data are loaded by the loader provided with the synteny module. The expression data and gene sequence data from NCBI are loaded using the Tripal Analysis Expression module and the Tripal GenBank Parser module.

The genetic data, as well as genomic and variation data, are stored using various modules in Chado as previously described [21, 22], except the Single Nucleotide Polymorphism (SNP) genotype data, which are stored using the genotype_call table similar to the table described in Sanderson *et al.* [23]. Figure 1 describes how SNP genotype data are currently stored in Chado. This is an update from figure 6 of Jung *et al.* [22]. All the genetic data, as well as phenotype and genotype data, are loaded by the Mainlab Chado Loader (MCL).

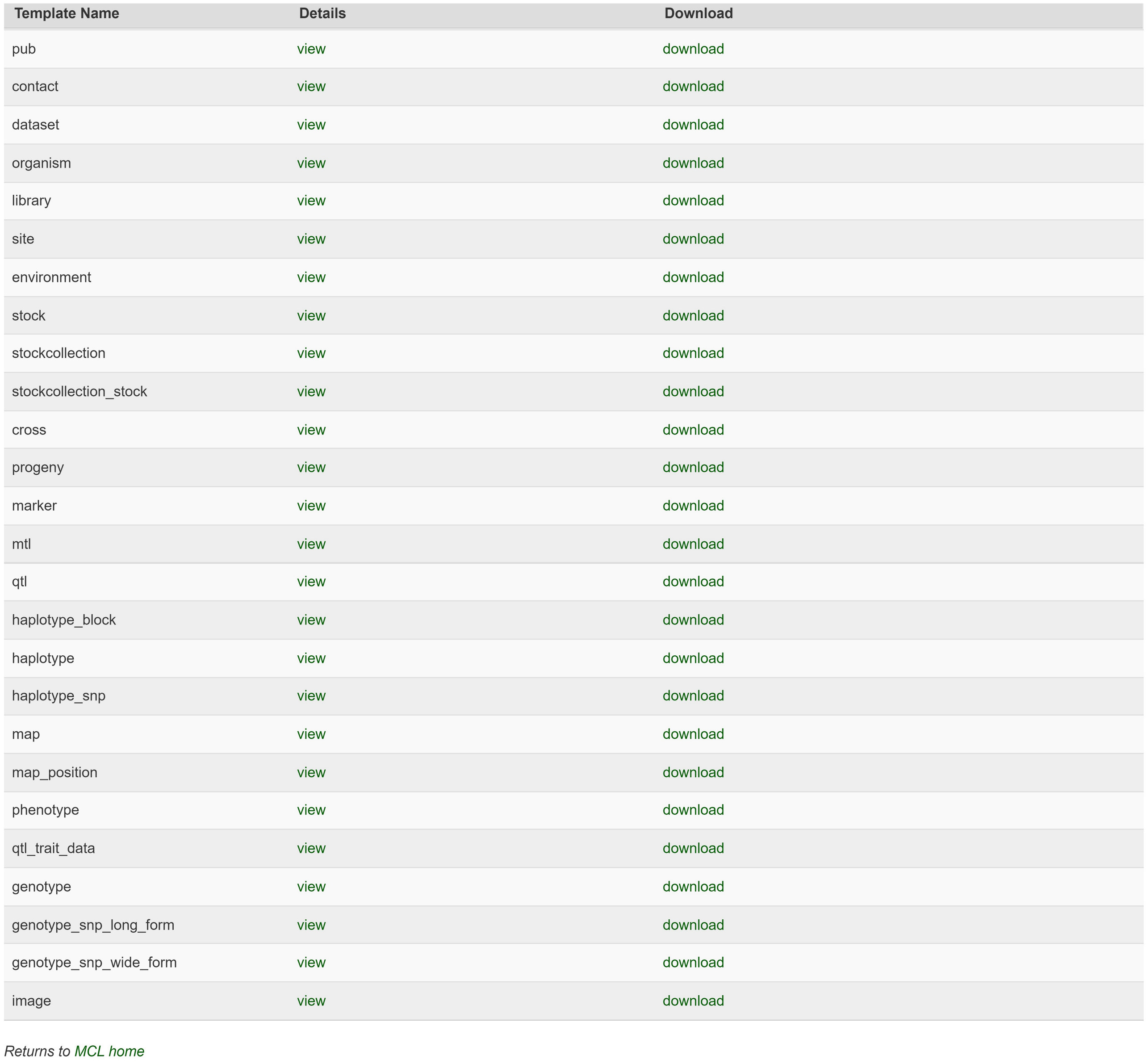
Data curation and analysis

Whole genome data

We obtain peer-reviewed whole genome data through either publication or direct submission by researchers. The assembly and annotation data are made available to view, download, and search in JBrowse, BLAST, PathwayCyc, Synteny Viewer, Sequence Search, and Gene/Transcript Search pages. Additional analyses provided by our team on the assemblies include computational annotation of predicted genes with the assignment of InterPro protein domains [24] and Gene Ontology (GO) terms [25], as well as homology to known proteins. In detail, the sequences are functionally characterized by pairwise comparison using the BLASTX algorithm against the Swiss-Prot protein database. Information on the top 10 matches with an *E*-value of ≤1E-06 is recorded and stored in our databases together with the sequences. InterPro domains and Gene Ontology assignments are made using InterProScan. The top BLASTP matches for the predicted proteins with the Swiss-Prot and TrEMBL databases are also provided as a downloadable Excel file. Our team also performs synteny analysis to find conserved syntenic regions among the closely related genomes using MCScanX [26] with default settings. PathwayCyc analyses are performed for important genomes using the PathwayTools software ([20]; http://bioinformatics.ai.sri.com/ptools/). The genome assembly and associated annotation file, along with GO terms predicted using InterProScan (https://www.ebi.ac.uk/interpro/download/) and Enzyme Commission terms predicted with DeepEC [27] from predicted proteins, are used to generate the input files for PathwayTools using an in-house Perl script. The PathwayCyc databases for each genome are then built with PathwayTools using the default settings. The built databases are checked for consistency before the final overview is generated and then displayed on our databases.

Transcriptome data

RNA-seq and dbEST datasets are used to create a reference transcriptome (RefTrans) for each genus or crop and provide putative functions identified by homology to known proteins. The RNA-seq sequences from peer-reviewed publications are downloaded from the NCBI Short Read Archive [28] and subjected to quality control using Trimmomatic (v0.32, default parameters [29]) and custom Perl scripts. The remaining RNA-seq reads are assembled de novo with Trinity (v2.6.6 [30]) using default assembly parameters and a minimum coding length of 200 bases. Quality control of the Expressed Sequence Tags (ESTs) includes vector sequence screening (UniVec_Core, ftp://ftp.ncbi.nih.gov/pub/UniVec/) using cross_match [31], removal of tRNA/rRNA/snRNA sequences identified using tblastx [32], and Poly-A tail trimming. The filtered ESTs are assembled using the CAP3 program (P −90 [33]). Bowtie (v2.3.3) [34] is applied to multimap the RNA-seq reads and ESTs back to the assembled contigs and singlets. The contigs and singlets are clustered into genes using CD-HIT (v4.6.4 [35]) and Corset (v1.0.7; [36]) with default parameters. The longest isoforms greater than 500 nt are selected to represent each Corset cluster and create the RefTrans sequences. The functional characterization of the RefTrans sequences, the homology with other genes, and the assignment of InterPro domains and GO terms are done by the same procedure as described above. Unigene sets are constructed with ESTs as described in Jung *et al.* [37] when RNA-seq sequences are not available. RefTrans are mapped to an appropriate whole genome assembly using the alignment tool “BLAT.” Alignments with an alignment length of 97% and 97% identity are preserved.



Template Name	Details	Download
pub	view	download
contact	view	download
dataset	view	download
organism	view	download
library	view	download
site	view	download
environment	view	download
stock	view	download
stockcollection	view	download
stockcollection_stock	view	download
cross	view	download
progeny	view	download
marker	view	download
mtl	view	download
qtl	view	download
haplotype_block	view	download
haplotype	view	download
haplotype_snp	view	download
map	view	download
map_position	view	download
phenotype	view	download
qtl_trait_data	view	download
genotype	view	download
genotype_snp_long_form	view	download
genotype_snp_wide_form	view	download
image	view	download

Returns to MCL home

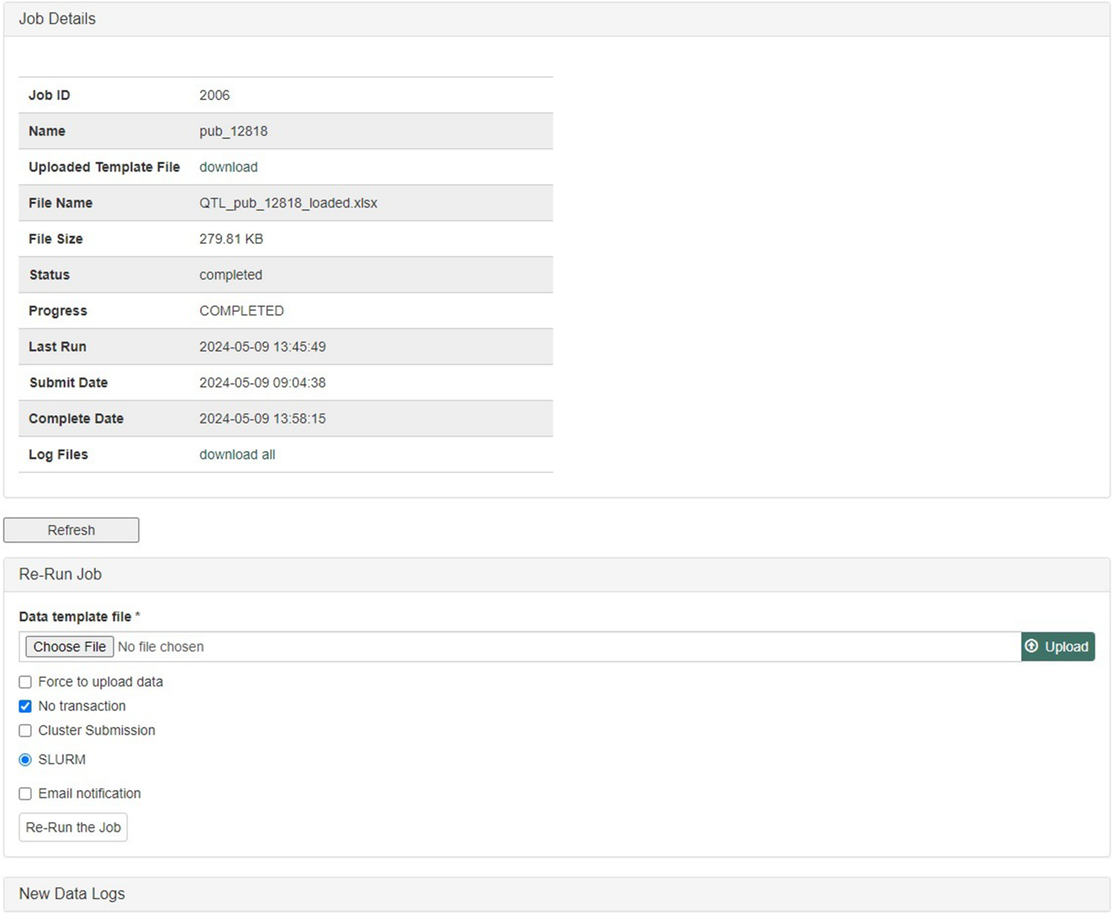
Figure 2. Templates available in the MCL interface. Curators can view and download individual templates or all the templates in one file.

NCBI gene sequences and expression data

Our database team periodically downloads crop sequences from the NCBI nr database [28] using the Tripal GenBank Parser module described in Table 1. All sequences are then parsed for gene, mRNA, coding sequence, 5′ UTR, and 3′UTR features and imported into our databases. Similar to predicted genes from whole genome sequences, genes parsed from NCBI are further annotated by homology to genes in other species, InterPro protein domains, and GO terms. Biomaterial data from NCBI, including expression data available from publications, are imported into our databases using the Tripal Analysis Expression module.

Genetic map, marker, QTL, phenotype, and genotype data

The database team monitors new peer-reviewed publications on molecular markers, genetic maps, Quantitative Trait Loci (QTLs), phenotype, and genotype data to curate and integrate with other genetic and genomic data. The data in the publication are entered into the templates that accompany the MCL and loaded through the web-based interface of the MCL extension module. Various data templates can be accessed in the curator interface once the MCL module is implemented. In the current version of MCL, templates are available for each of the following data types: db (database), cv (controlled vocabulary), property, pub, library, trait, contact, dataset, image, descriptor, site, stock, cross, progeny, marker, Mendelian Trait Locus (MTL), QTL, map, map position, phenotype, genotype, and haplotype (Fig. 2). Once curators enter the data in the templates, they can upload the data using the web interface. The uploading page shows the status of all the submitted uploading jobs and provides a link to a page, where curators can view the details of each job being uploaded (Fig. 3).


Job Details

Job ID	2006
Name	pub_12818
Uploaded Template File	download
File Name	QTL_pub_12818_loaded.xlsx
File Size	279.81 KB
Status	completed
Progress	COMPLETED
Last Run	2024-05-09 13:45:49
Submit Date	2024-05-09 09:04:38
Complete Date	2024-05-09 13:58:15
Log Files	download all

Data template file *

□ Force to upload data
☑ No transaction
□ Cluster Submission
● SLURM
□ Email notification

Figure 3. Uploading Job Detail page in the MCL interface. After submitting data in templates, curators can check the progress and any errors.

Genetic map data include mapped positions of molecular markers, QTLs, and heritable phenotypic markers, along with information on mapping populations and publications. Marker annotations encompass marker aliases, source germplasm, source descriptions, primer sequences, polymerase chain reaction conditions, literature references, and map positions where available. For SNPs, marker details also include SNP array names, SNP array IDs, dbSNP IDs, alleles, flanking sequences, and probes. QTLs and MTLs data in our databases are curated with aliases, curator-assigned QTL labels, published symbols, trait names, taxa, trait descriptions, screening methods, map positions, associated markers, statistical values, datasets, contact information, and references. Trait names and abbreviations are standardized by curators across all trait loci data entered into the databases to ensure integration with the Trait Ontology (TO)[38], facilitating data integration across databases, organisms, and data types. When terms are not available from TO, we develop new terms for submission.

Our databases also house phenotypic and genotypic evaluation data sourced from publications and/or GRIN [39]. Phenotype data obtained from GRIN or publications are transferred to MCL templates. Details such as associated datasets, phenotypic descriptors, accessions, and phenotypic values are

Table 2. The production values of the crops that our databases serve, usage of the databases

Database	Crops served	Average value of production (2016–22) $ billion	2019–23 Usage by community Number of citations

GDR, www.rosaceae.org	Almond, apple, apricot, blackberry, cherry, nectarine, peach, pear, plum, raspberry, rose, and strawberry	14.137	Users = 173 429
			Avg. number of countries = 157.4
			Visits = 484 134
			Pages = 5 730 926
			Citations = 1166

CGD, www.citrusgenomedb.org	Grapefruit, lemon, lime, orange, tangelo, and tangerine/mandarin	3.407	Users = 57 386
			Avg. number of countries = 143.6
			Visits = 96 400
			Pages = 678 272
			Citations = 132

(continued)

Table 2. (Continued)

Database	Crops served	Average value of production (2016–22) $ billion	2019–23 Usage by community Number of citations

PCD (formerly Cool Season Food Legume Database), www.pulsedb.org	Beans, chickpea, lentil, and pea	1.454	Users = 40 765
			Avg. number of countries = 125
			Visits = 55 634
			Pages = 296 185
			Citations = 54

CottonGen, www.cottongen.org	Cotton	6.345	Users = 129 400
			Avg. number of countries = 121.2
			Visits = 271 460
			Pages = 2 021 507
			Citations = 603

GDV, www.vaccinium.org	Blueberry and cranberry	1.112	Users = 27 650
			Avg. number of countries = 96.2
			Visits = 51 391
			Pages = 638 451
			Citations = 70

Total	25 crops	26.445	Users = 428 630
			Avg. number of countries = 128.7
			Visits = 959 019
			Pages = 9 365 341
			Citations = 2025

entered into templates and loaded using the MCL loader. Accession names are standardized and compared with those in our databases. Genotype data, such as SNP genotypes, Simple Sequence Repeat genotypes, and haplotype data, are similarly transferred to templates, with marker and accession names standardized before loading into our databases.

Databases

The GDR, CottonGEN, CGD, GDV, and 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, and tangerine),

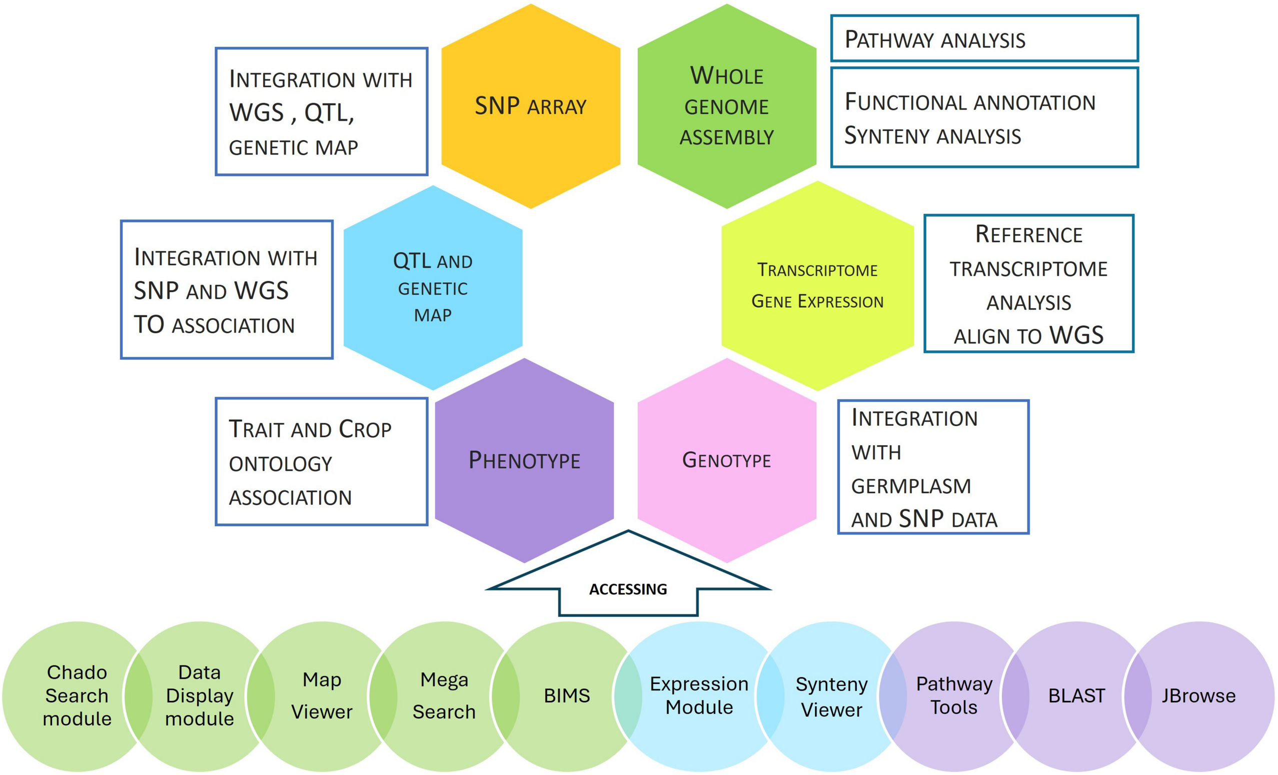
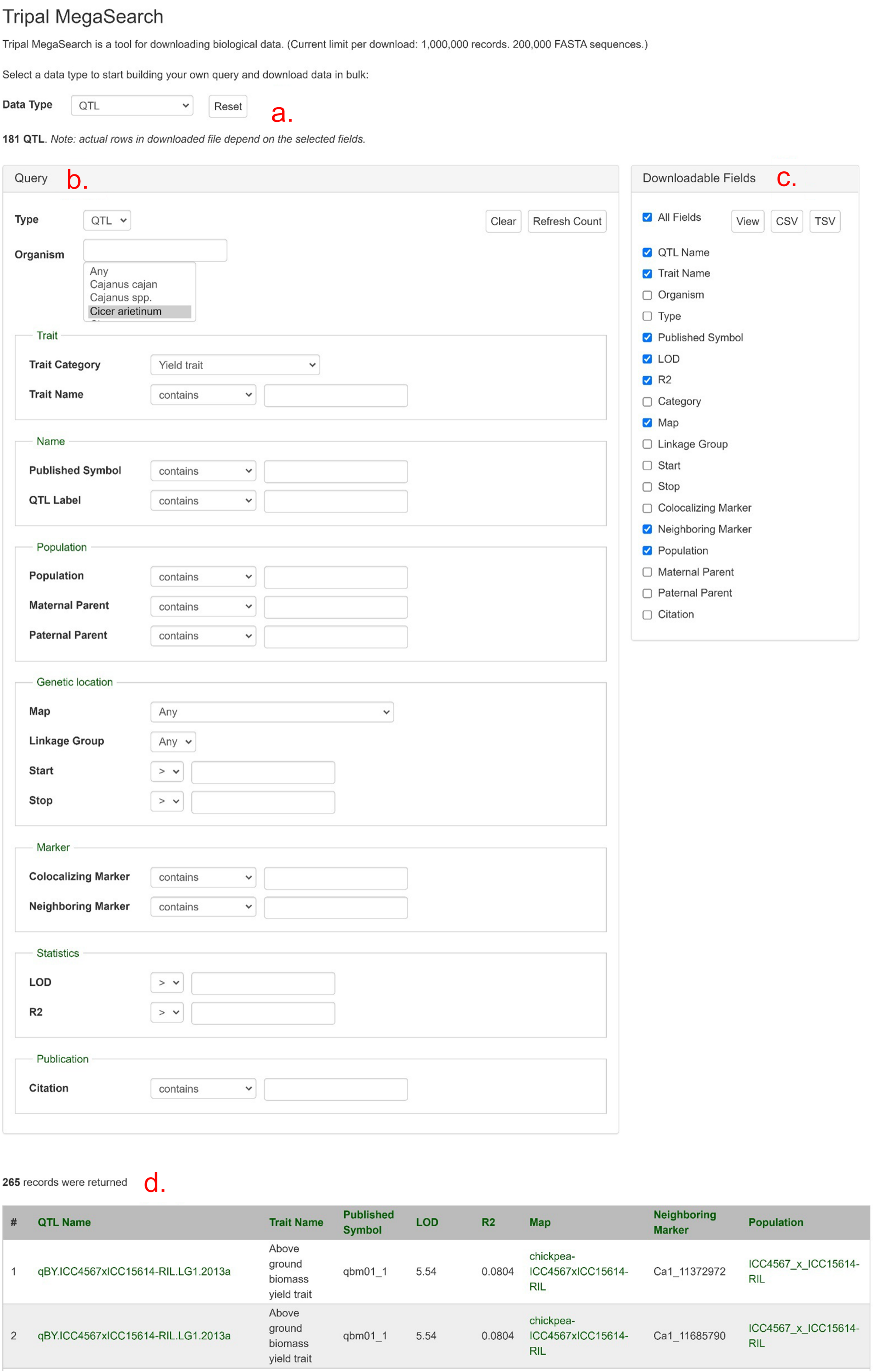


Figure 4. Data types, analysis, integration, and user interface. The first five circles: Tripal modules developed by our group; the next two circles: Tripal modules developed by other groups; the last three circles: non-Tripal software.



Tripal MegaSearch

Tripal MegaSearch is a tool for downloading biological data. (Current limit per download: 1,000,000 records. 200,000 FASTA sequences.)

Select a data type to start building your own query and download data in bulk:

Data Type QTL Reset

181 QTL. *Note: actual rows in downloaded file depend on the selected fields.*

Query

Type	QTL	
Organism	Any Cajanus cajan Cajanus spp. Cicer arietinum	

Trait

Trait Category	Yield trait	
Trait Name	contains	

Name

Published Symbol	contains	
QTL Label	contains	

Population

Population	contains	
Maternal Parent	contains	
Paternal Parent	contains	

Genetic location

Map	Any	
Linkage Group	Any	
Start	>	
Stop	>	

Marker

Colocalizing Marker	contains	
Neighboring Marker	contains	

Statistics

LOD	>	
R2	>	

Publication

Citation	contains	

Downloadable Fields

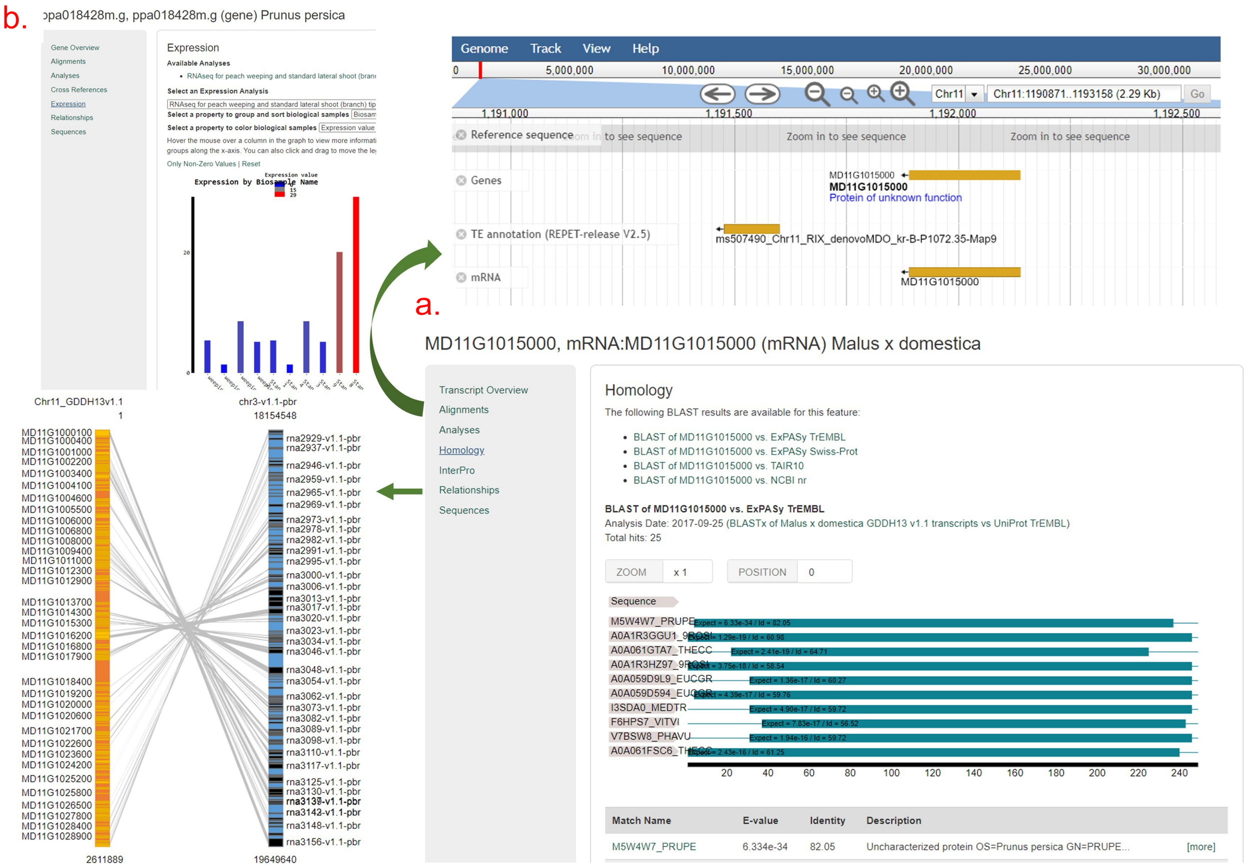
☑ All Fields

☑ QTL Name
☑ Trait Name
□ Organism
□ Type
☑ Published Symbol
☑ LOD
☑ R2
□ Category
☑ Map
□ Linkage Group
□ Start
□ Stop
□ Colocalizing Marker
☑ Neighboring Marker
☑ Population
□ Maternal Parent
□ Paternal Parent
□ Citation

265 records were returned

#	QTL Name	Trait Name	Published Symbol	LOD	R2	Map	Neighboring Marker	Population
1	qBY.ICC4567xICC15614-RIL.LG1.2013a	Above ground biomass yield trait	qbm01_1	5.54	0.0804	chickpea-ICC4567xICC15614-RIL	Ca1_11372972	ICC4567_x_ICC15614-RIL
2	qBY.ICC4567xICC15614-RIL.LG1.2013a	Above ground biomass yield trait	qbm01_1	5.54	0.0804	chickpea-ICC4567xICC15614-RIL	Ca1_11685790	ICC4567_x_ICC15614-RIL

Figure 5. An example MegaSearch page. (a) Data type section where users can choose data type. (b) Query section that provides a query form that allows users to perform complex queries using various metadata as filters. (c) Downloadable Fields section where users can choose data fields to view and download. (d) The result table where users can see the data with chosen fields with hyperlinks.



b. ppa018428m.g, ppa018428m.g (gene) Prunus persica

Gene Overview
Alignments
Analyses
Cross References
Expression
Relationships
Sequences

Expression

Available Analyses

- RNAseq for peach weeping and standard lateral shoot (branc

Select an Expression Analysis RNAseq for peach weeping and standard lateral shoot (branch) tip

Select a property to group and sort biological samples Biosam

Select a property to color biological samples Expression value

Hover the mouse over a column in the graph to view more informati groups along the x-axis. You can also click and drag to move the le

Only Non-Zero Values | Reset

MD11G1015000, mRNA:MD11G1015000 (mRNA) Malus x domestica

Transcript Overview
Alignments
Analyses
Homology
InterPro
Relationships
Sequences

Homology

The following BLAST results are available for this feature:

- BLAST of MD11G1015000 vs. ExPASy TrEMBL
- BLAST of MD11G1015000 vs. ExPASy Swiss-Prot
- BLAST of MD11G1015000 vs. TAIR10
- BLAST of MD11G1015000 vs. NCBI nr

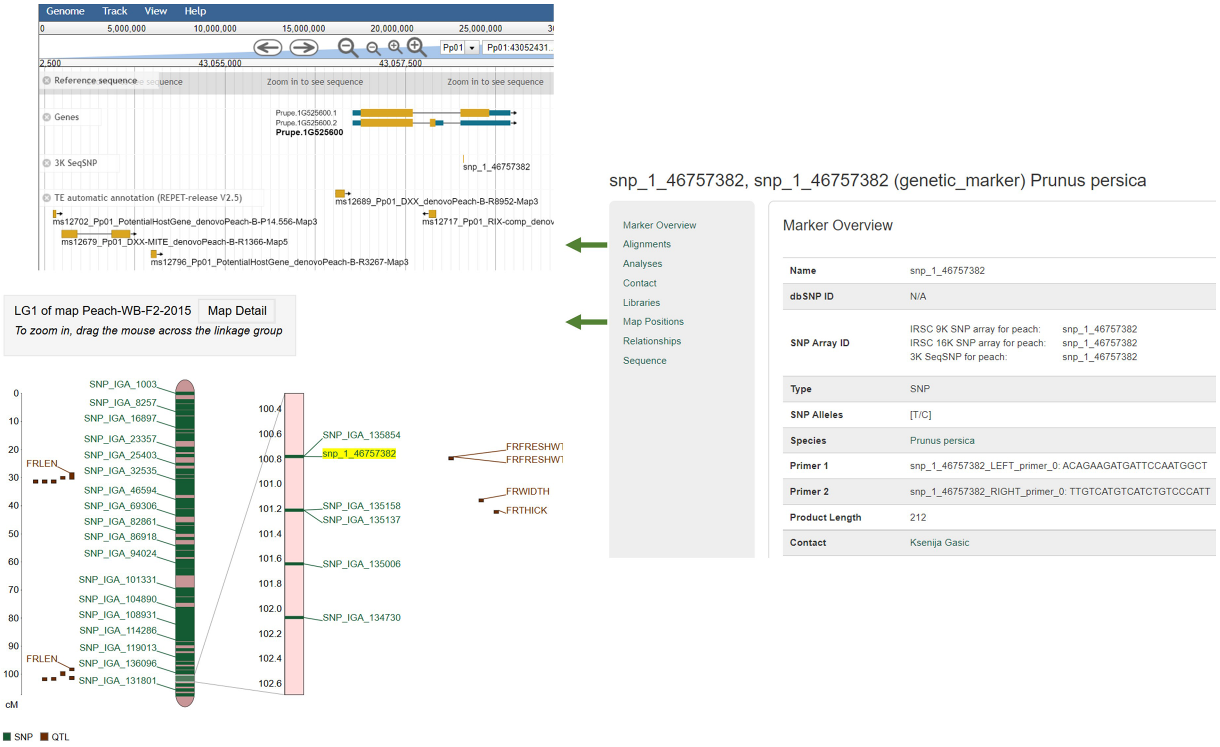
BLAST of MD11G1015000 vs. ExPASy TrEMBL
Analysis Date: 2017-09-25 (BLASTx of Malus x domestica GDDH13 v1.1 transcripts vs UniProt TrEMBL)
Total hits: 25

ZOOM x 1 POSITION 0

Sequence
M5W4W7_PRUPE
Expect = 6.33e-34 / Id = 82.05
A0A1R3GGU1_9ROSI
Expect = 1.29e-19 / Id = 60.98
A0A061GTA7_THECC
Expect = 2.41e-19 / Id = 64.71
A0A1R3HZ97_9ROSI
Expect = 3.75e-18 / Id = 58.54
A0A059D9L9_EUCGR
Expect = 1.36e-17 / Id = 60.27
A0A059D594_EUCGR
Expect = 4.39e-17 / Id = 59.76
I3SDA0_MEDTR
Expect = 4.90e-17 / Id = 59.72
F6HPS7_VITVI
Expect = 7.83e-17 / Id = 56.52
V7BSW8_PHAVU
Expect = 1.94e-16 / Id = 59.72
A0A061FSC6_THECC
Expect = 2.43e-16 / Id = 61.25
20
40
60
80
100
120
140
160
180
200
220
240

Match Name	E-value	Identity	Description	
M5W4W7_PRUPE	6.334e-34	82.05	Uncharacterized protein OS=Prunus persica GN=PRUPE...	[more]

Figure 6. An example gene/mRNA page. (a) Homology section of an mRNA page and JBrowse and Synteny Viewer page that are hyperlinked from Alignments and Relationships sections. (b) Expression section of a gene page that shows a snippet from the Expression Heatmap Viewer.



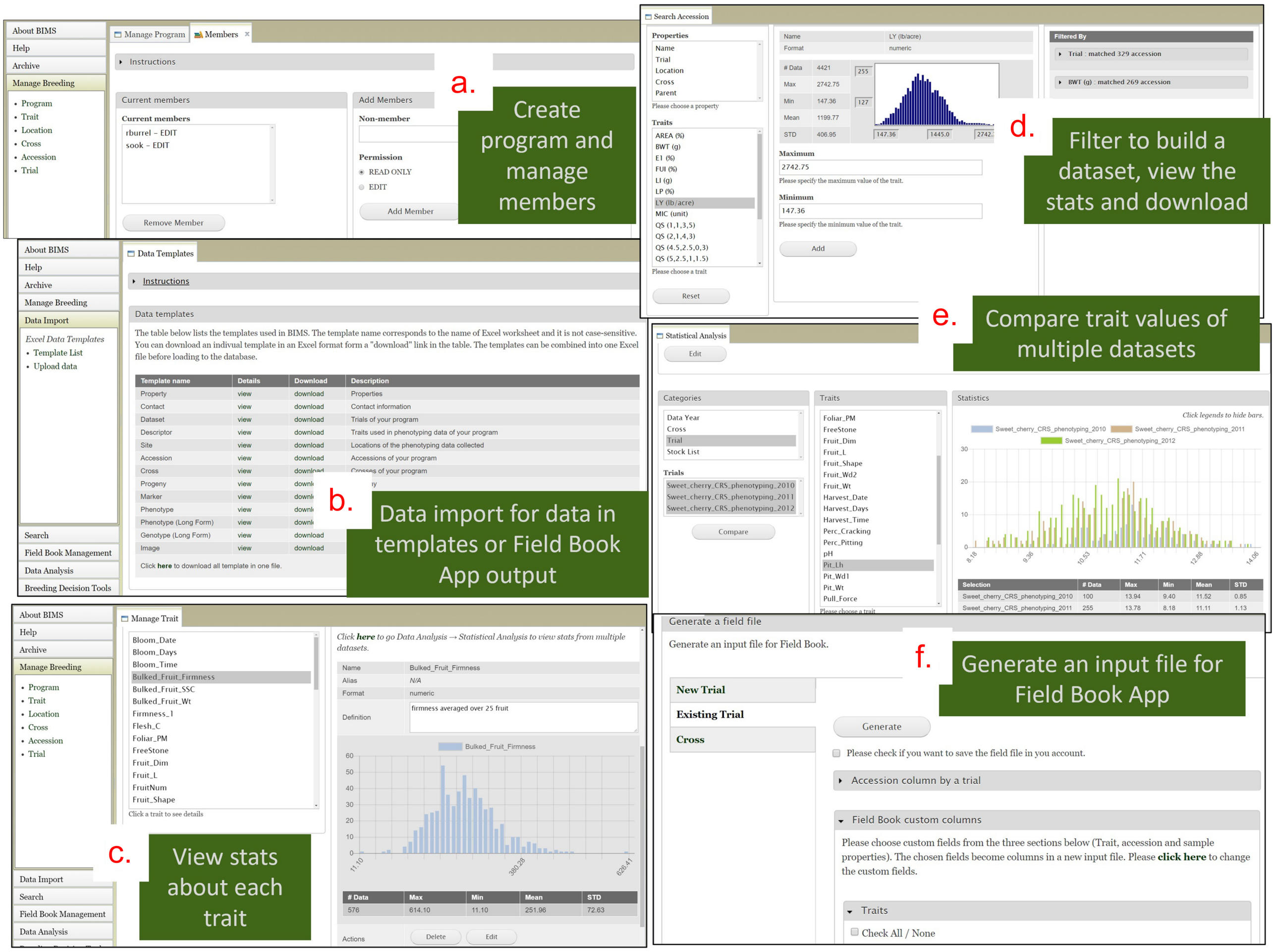
snp_1_46757382, snp_1_46757382 (genetic_marker) Prunus persica

Marker Overview
Alignments
Analyses
Contact
Libraries
Map Positions
Relationships
Sequence

Marker Overview

Name	snp_1_46757382
dbSNP ID	N/A
SNP Array ID	IRSC 9K SNP array for peach: snp_1_46757382 IRSC 16K SNP array for peach: snp_1_46757382 3K SeqSNP for peach: snp_1_46757382
Type	SNP
SNP Alleles	[T/C]
Species	Prunus persica
Primer 1	snp_1_46757382_LEFT_primer_0: ACAGAAGATGATTCCAATGGCT
Primer 2	snp_1_46757382_RIGHT_primer_0: TTGTCATGTCATCTGTCCCATT
Product Length	212
Contact	Ksenija Gasic

Figure 7. An example Marker page. JBrowse is hyperlinked from the Alignments section and MapViewer page is hyperlinked from the Map Positions section.



Data Templates

Instructions

Data templates

The table below lists the templates used in BIMS. The template name corresponds to the name of Excel worksheet and it is not case-sensitive. You can download an indivual template in an Excel format form a "download" link in the table. The templates can be combined into one Excel file before loading to the database.

Template name	Details	Download	Description
Property	view	download	Properties
Contact	view	download	Contact information
Dataset	view	download	Trials of your program
Descriptor	view	download	Traits used in phenotyping data of your program
Site	view	download	Locations of the phenotyping data collected
Accession	view	download	Accessions of your program
Cross	view	download	Crosses of your program
Progeny	view	downl	y
Marker	view	downl	
Phenotype	view	downl	
Phenotype (Long Form)	view	downl	
Genotype (Long Form)	view	download	
Image	view	download	

Click **here** to download all template in one file.

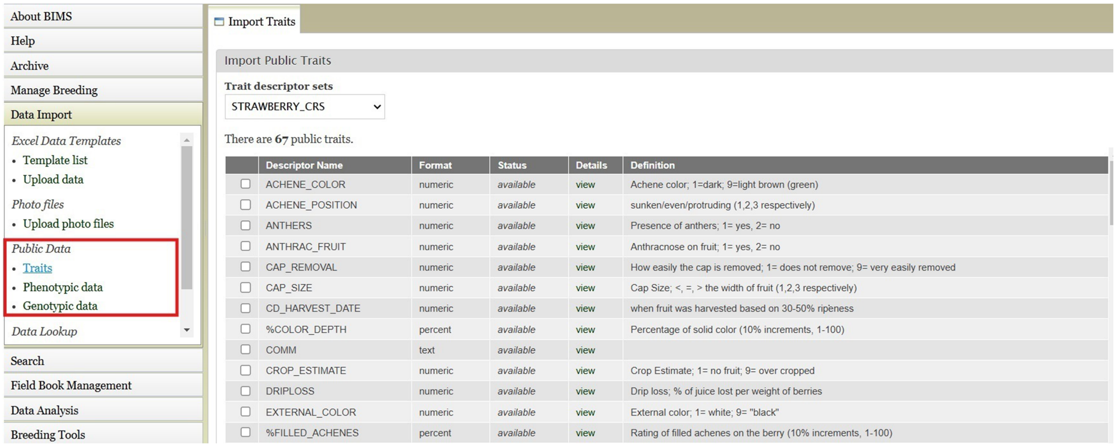
*Click **here** to go Data Analysis → Statistical Analysis to view stats from multiple datasets.*

Generate an input file for Field Book.

□ Please check if you want to save the field file in you account.

Please choose custom fields from the three sections below (Trait, accession and sample properties). The chosen fields become columns in a new input file. Please **click here** to change the custom fields.

Figure 8. BIMS interface. (a) Manage Program page where users can create a program and add members. (b) Data Import page where users can download data templates and upload their data. (c) Trait section of the Manage Program page where users can view the distribution and the statistical values of the trait values for each trait. (d) Search page where users can filter accessions by various categories including trait cutoff values to build, view, save, and download datasets. (e) Analysis page where users can compare trait values of multiple datasets. (f) Field Book Management page where users can send and receive files to Field Book App using files or Breeding Application Programming Interface.



Import Public Traits

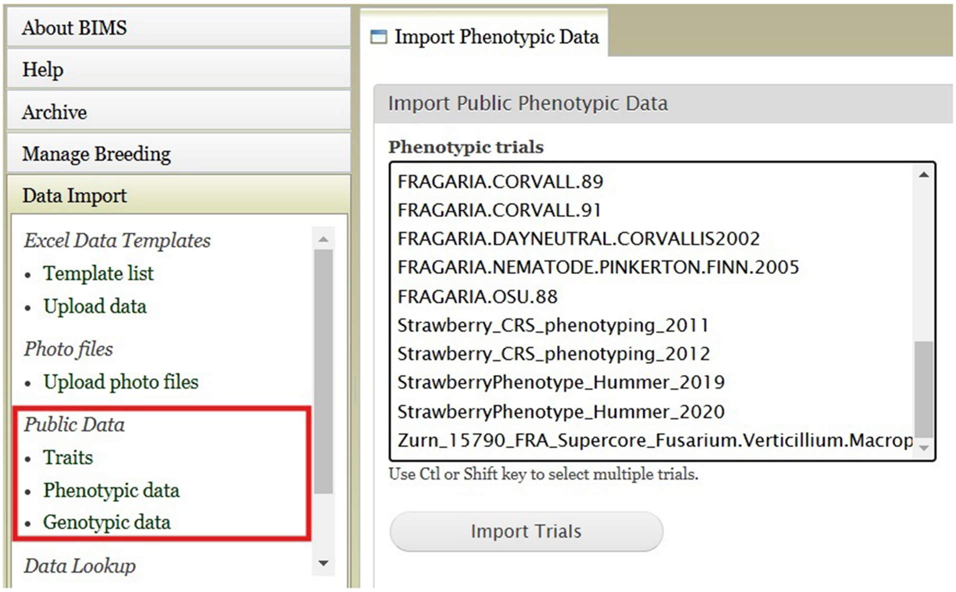
Trait descriptor sets: STRAWBERRY_CRS

There are **67** public traits.

	Descriptor Name	Format	Status	Details	Definition
□	ACHENE_COLOR	numeric	*available*	view	Achene color; 1=dark; 9=light brown (green)
□	ACHENE_POSITION	numeric	*available*	view	sunken/even/protruding (1,2,3 respectively)
□	ANTHERS	numeric	*available*	view	Presence of anthers; 1= yes, 2= no
□	ANTHRAC_FRUIT	numeric	*available*	view	Anthracnose on fruit; 1= yes, 2= no
□	CAP_REMOVAL	numeric	*available*	view	How easily the cap is removed; 1= does not remove; 9= very easily removed
□	CAP_SIZE	numeric	*available*	view	Cap Size; <, =, > the width of fruit (1,2,3 respectively)
□	CD_HARVEST_DATE	numeric	*available*	view	when fruit was harvested based on 30-50% ripeness
□	%COLOR_DEPTH	percent	*available*	view	Percentage of solid color (10% increments, 1-100)
□	COMM	text	*available*	view	
□	CROP_ESTIMATE	numeric	*available*	view	Crop Estimate; 1= no fruit; 9= over cropped
□	DRIPLOSS	numeric	*available*	view	Drip loss; % of juice lost per weight of berries
□	EXTERNAL_COLOR	numeric	*available*	view	External color; 1= white; 9= "black"
□	%FILLED_ACHENES	percent	*available*	view	Rating of filled achenes on the berry (10% increments, 1-100)

Figure 9. BIMS page in GDR where users can download Strawberry Crop Ontology.

nuts (almond), pulses (chickpea, fava bean, lentil, pea, and common bean), and ornamentals (apple, cherry, and rose). Grown commercially throughout the USA, the annual US value of production of these crops averaged $26.445 billion between 2016 and 2022 (Table 2). These crops are important for food, feed, and fiber supply and are the economic backbone of many rural communities. These databases are trusted, well established, widely used, community-driven, and highly

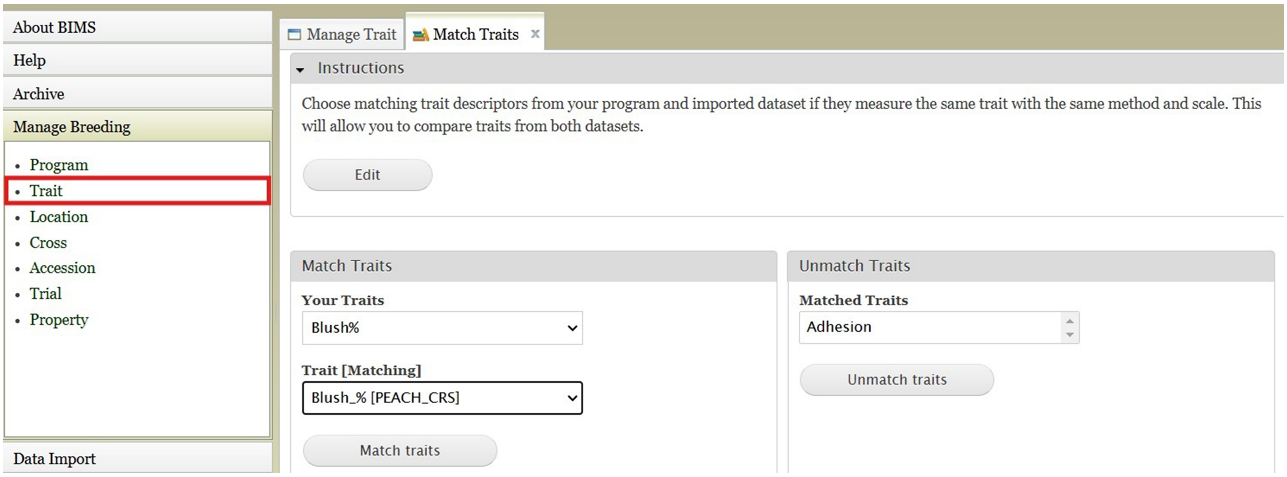


Phenotypic trials

FRAGARIA.CORVALL.89
FRAGARIA.CORVALL.91
FRAGARIA.DAYNEUTRAL.CORVALLIS2002
FRAGARIA.NEMATODE.PINKERTON.FINN.2005
FRAGARIA.OSU.88
Strawberry_CRS_phenotyping_2011
Strawberry_CRS_phenotyping_2012
StrawberryPhenotype_Hummer_2019
StrawberryPhenotype_Hummer_2020
Zurn_15790_FRA_Supercore_Fusarium.Verticillium.Macrop

Use Ctl or Shift key to select multiple trials.

Figure 10. An example BIMS page where users can download public phenotype data.



Choose matching trait descriptors from your program and imported dataset if they measure the same trait with the same method and scale. This will allow you to compare traits from both datasets.

Your Traits: Blush%

Trait [Matching]: Blush_% [PEACH_CRS]

Matched Traits: Adhesion

Figure 11. An example BIMS page where users can match trait descriptors from their datasets and imported datasets.

impactful. They are used in every US state and territory and are the databases of choice of many researchers around the world. Since their establishment (2003 for GDR, 2011 for CGD, GDV, and PCD, and 2012 for CottonGen), the usage of databases has been increasing significantly. Over the past 5 years, they have been collectively cited in 2192 publications, visited by 418 254 users from 185 countries with 7.5 million pages accessed (Table 2).

Data types, data integration, and user interface

Our databases contain whole genome assembly, genes/mRNAs, genetic markers, genetic maps, QTLs, MTLs, phenotype, genotype, haplotype, and publication data. Figure 4 summarizes the data types, data analysis, data integration, and user interface in our databases. Data integration is achieved through various analyses and manual curation. Data with sequences, such as transcripts, markers, and individually cloned genes, are integrated with whole genome data by sequence alignment. Genetic map and whole genome data are integrated when SNPs in whole genome data are used in genetic maps. QTLs are also integrated with whole genome data when the colocalizing markers are aligned to the whole genome. Phenotypic descriptors and trait names of QTLs and haplotype blocks are integrated by associated Trait Ontology. Genotype data are integrated with whole genome data by the markers. Genotype data and phenotype data are also integrated by germplasm. Additionally, genomic regions from related species are integrated by synteny analysis. All the data that are associated with the genomes of one species can be transferred to another species through synteny data. Orthologs identified by the synteny analysis between multiple genome assemblies of the same species represent the same genes, serving as important data integration for users.

Over the years, new types of data and metadata have been added to our databases without changing the underlying database schema due to the ontology-driven and modular nature of Chado. For example, haplotype data and SNP array data, as well as the association of Trait Ontology terms with various data types, such as markers, germplasm, and QTLs, were added without drastic schema changes.

As shown in Fig. 4, users can access the data using various search pages and graphic interfaces. MegaSearch (Fig. 5) allows users to search and download various types of data using comprehensive categories and customize the output. Search results for haplotype and genotype data provide a comprehensive view (Fig. 5) and downloading functionality. Individual data pages, such as gene/mRNA page (Fig. 6) and marker page (Fig. 7), allow users to access extensively integrated data. When data are available, graphic interfaces, such as MapViewer, Expression Heatmap, Synteny Viewer, and JBrowse, can be accessed through the link from the gene/mRNA page (Fig. 6) and marker page (Fig. 7). The graphic interfaces can also be accessed directly from the tool menu.

BIMS is a tool that provides individual breeders with a secure and comprehensive online breeding management system that allows breeders to store, manage, archive, and analyze their private breeding data. Once breeders have their accounts in one of our databases, they can create their own program, add members, upload their data, and edit their data to manage their accession, phenotype, and genotype data (Fig. 8). Breeders can create and maintain multiple programs if they have different crops and/or different projects. For example, breeders can create an additional BIMS program in addition to their own program to manage data from a collaborative project that involves multiple breeders. Since breeders can operate their BIMS program on their own, it does not require any work from the database team except helping them with questions.

BIMS, when implemented in community databases, can facilitate data standardization and reuse. BIMS users can download community-standard crop ontologies to use as trait descriptors for their private programs In GDR-BIMS, the Strawberry Crop Ontology—developed by the GDR team in collaboration with wider communities—is available (Fig. 9). Similarly, the Blueberry Crop Ontology, developed by community [40], is also available in GDV-BIMS In addition, when public genotype and phenotype data are available in the database, breeders can import them into their private BIMS program. Figure 10 illustrates how BIMS users can view and download publicly available phenotype data in GDR-BIMS. Additionally, BIMS provides functionality for merging trait descriptors when identical traits are measured using different trait descriptors across public and private datasets (Fig. 11). This feature enables users to effectively compare trait data across diverse datasets.

Limitations

There are some limitations to using the Tripal system for building databases. The underlying schema, Chado, is generic and ontology-driven, which can make it nonintuitive for first-time users. However, various data loaders and publications on how to store data are available, as described above. While creating a genome database with Tripal does not require programming, it is advisable for site developers to have strong programming skills in case new tools or customizations are needed for the site.

Conclusion and future direction

Building and maintaining our crop databases using the open-source Tripal genome database toolkit have saved significant time and effort, allowing more time to be spent on data analysis and curation. Data curation, analysis, and integration that keep up to date with new publications are key to the usefulness of a crop database. This makes using an efficient database system crucial, especially for 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 of restructuring the database schema and interfaces. We plan to add more types of data to our databases, such as gene annotation data, in collaboration with community researchers and other crop databases.

Acknowledgements

The authors acknowledge with thanks their funding sources, the crop research community for providing data, support, and feedback, the Tripal and GMOD community of developers for developing and sharing Tripal modules and code, and the AgBioData Consortium and US Land Grant Universities for support.

Conflict of interest: None declared.

Funding

This work was funded by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Specialty Crop Research Initiative projects (2014-51181-2237, 2014-51181-22378, and 2022-51181-38449); USDA NIFA National Research Support Project 10 (NRSP10); Washington State Department of Agriculture Block Grant; USA Dry Pea and Lentil Council; Northern Pulse Growers Association; USDA Agricultural Research Service; and Washington State University. Funding for open access charge: Federal Grant; USDA NIFA NRSP10.

Data Availability

The five databases we described are publicly available: the GDR (www.rosaceae.org), the GDV (www.vaccinium.org), the CGD (www.citrusgenomedb.org), the PCD (www.pulsedb.org), and CottonGen (www.cottongen.org).

References

1. Woodhouse MR, Cannon EK, Portwood JL *et al*. A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol* 2021;**21**:385. https://doi.org/10.1186/s12870-021-03173-5
2. Yao E, Blake VC, Cooper L *et al*. GrainGenes: a data-rich repository for small grains genetics and genomics. *Database* 2022;**2022**:baac034. https://doi.org/10.1093/database/baac034
3. Fernandez-Pozo N, Menda N, Edwards JD *et al*. The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Res* 2015;**43**:D1036–41. https://doi.org/10.1093/nar/gku1165

4. Brown AV, Conover JL, Woodhouse MR *et al*. SoyBase: a portal for soybean research. *Nucleic Acids Res* 2021;**49**:D1441–50. https://doi.org/10.1093/nar/gkaa1025
5. Jung S, Lee T, Cheng CH *et al*. 15 years of GDR: new data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res* 2019;**47**:D1137–45. https://doi.org/10.1093/nar/gky1030
6. Yu J, Jung S, Cheng CH *et al*. CottonGen: the community database for cotton genomics, genetics, and breeding research. *Plants* 2021;**10**:2805. https://doi.org/10.3390/plants10122805
7. Staton M, Cannon E, Sanderson LA *et al*. Tripal, a community update after 10 years of supporting open source, standards-based genetic, genomic and breeding databases. *Brief Bioinform* 2021;**22**:bbab238. https://doi.org/10.1093/bib/bbab238
8. Spoor S, Cheng CH, Sanderson LA *et al*. Tripal v3: an ontology-based toolkit for construction of FAIR biological community databases. *Database* 2019;**2019**:baz077. https://doi.org/10.1093/database/baz077
9. Ficklin SP, Sanderson LA, Cheng CH *et al*. Tripal: a construction toolkit for online genome databases. *Database* 2011;**2011**:bar044. https://doi.org/10.1093/database/bar044
10. Mungall CJ, Emmert DB. FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 2007;**23**:i337–46. https://doi.org/10.1093/bioinformatics/btm189
11. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
12. Jung S, Cheng CH, Buble K *et al*. Tripal MegaSearch: a tool for interactive and customizable query and download of big data. *Database* 2021;**2021**:baab023. https://doi.org/10.1093/database/baab023
13. Buble K, Jung S, Humann JL *et al*. Tripal MapViewer: a tool for interactive visualization and comparison of genetic maps. *Database* 2019;**2019**:baz100. https://doi.org/10.1093/database/baz100
14. Chen M, Henry N, Almsaeed A *et al*. New extension software modules to enhance searching and display of transcriptome data in Tripal databases. *Database* 2017;**2017**:1–12. https://doi.org/10.1093/database/bax052
15. Jung S, Lee T, Gasic K *et al*. The Breeding Information Management System (BIMS): an online resource for crop breeding. *Database* 2021;**2021**:baab054. https://doi.org/10.1093/database/baab054
16. Jung S, Lee T, Cheng C *et al*, Extension modules for storage, visualization and querying of genomic, genetic and breeding data in Tripal databases. *Database* 2017;**2017**. 10.1093/database/bax092
17. Yu J, Jung S, Cheng CH *et al*. CottonGen: a database of cotton genetics and genomics. *Plants* 2019;**8**:333. https://doi.org/10.3390/plants8090333
18. Buels R, Yao E, Diesh CM *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 2016;**17**:66. https://doi.org/10.1186/s13059-016-0924-1
19. Camacho C, Coulouris G, Avagyan V *et al*. BLAST+: architecture and applications. *BMC Bioinf* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421
20. Karp PD, Midford PE, Billington R *et al*. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* 2021;**22**:109–26. https://doi.org/10.1093/bib/bbz104
21. Jung S, Menda N, Redmond S *et al*. The Chado Natural Diversity module: a new generic database schema for large-scale phenotyping and genotyping data. *Database* 2011;**2011**:bar051. https://doi.org/10.1093/database/bar051
22. Jung S, Lee T, Ficklin SP *et al*. Chado use case: storing genomic, genetic and breeding data of Rosaceae and Gossypium crops in Chado. *Database* 2016;**2016**:baw010. https://doi.org/10.1093/database/baw010
23. Sanderson LA, Almsaeed A, West J *et al*. Tripal/Tripal_jBrowse: Tripal JBrowse 3.0. *Zenodo* 2019. https://doi.org/10.5281/zenodo.3564724
24. Paysan-Lafosse T, Blum M, Chuguransky S *et al*. InterPro in 2022. *Nucleic Acids Res* 2023;**51**:D418–27. https://doi.org/10.1093/nar/gkac993
25. The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics* 2023;**224**:iyad031. https://doi.org/10.1093/genetics/iyad031
26. Wang Y, Tang H, Debarry JD *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**:e49. https://doi.org/10.1093/nar/gkr1293
27. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc Natl Acad Sci U S A* 2019;**116**:13996–4001. https://doi.org/10.1073/pnas.1821905116
28. Sayers EW, Cavanaugh M, Clark K *et al*. Using GenBank and SRA. *Methods Mol Biol* 2022;**2443**:1–25. https://doi.org/10.1007/978-1-0716-2067-0_1
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20. https://doi.org/10.1093/bioinformatics/btu170
30. Grabherr MG, Haas BJ, Yassour M *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52. https://doi.org/10.1038/nbt.1883
31. Gordon D, Green P. Consed: a graphical editor for next-generation sequencing. *Bioinformatics* 2013;**29**:2936–37. https://doi.org/10.1093/bioinformatics/btt515
32. Altschul SF, Gish W, Miller W *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
33. Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res* 1999;**9**:868–77. https://doi.org/10.1101/gr.9.9.868
34. Langmead B, Trapnell C, Pop M *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25. https://doi.org/10.1186/gb-2009-10-3-r25
35. Fu L, Niu B, Zhu Z *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2. https://doi.org/10.1093/bioinformatics/bts565
36. Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol* 2014;**15**:410. https://doi.org/10.1186/s13059-014-0410-6
37. Jung S, Menda N, Redmond S *et al*. The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res* 2014;**42**:D1237–44. https://doi.org/10.1093/nar/gkt1096
38. Cooper L, Elser J, Laporte MA *et al*. Planteome 2024 update: reference ontologies and knowledgebase for plant biology. *Nucleic Acids Res* 2024;**52**:D1548–55. https://doi.org/10.1093/nar/gkad1028
39. García-Recio A, Santos-Gómez A, Soto D *et al*. GRIN database: a unified and manually curated repertoire of GRIN variants. *Hum Mutat* 2021;**42**:8–18. https://doi.org/10.1002/humu.24141
40. Hislop L M *et al*. A Blueberry (Vaccinium L.) Crop Ontology to Enable Standardized Phenotyping for Blueberry Breeding and Research. *horts* 2024;**59**:1433–1442. 10.21273/HORTSCI17676-23

Database, 2025, 00, baaf005, DOI: https://doi.org/10.1093/database/baaf005, Original Article