HortScience 43(5):1365–1366. 2008.

Availability of Genotypic Data for USDA-ARS National Plant Germplasm System Accessions Using the Genetic Resources Information Network (GRIN) Database

Gayle M. Volk[1] and Christopher M. Richards

USDA-ARS National Center for Genetic Resources Preservation, 1111 South Mason Street, Ft. Collins, CO 80521

Additional index words. GRIN database, molecular marker, microsatellite, bioinformatic, genetic resources

***Abstract*. The USDA-ARS National Plant Germplasm System (NPGS) provides critical genetic resources to researchers and breeders worldwide. Users of the NPGS materials need access to data for genetic and descriptive characteristics of the plant materials. New tables and codes have been added to the Germplasm Resources Information Network (GRIN) database to hold raw data relating to molecular markers and alleles. The revised tables accommodate multiple marker types; provide raw data for individuals; accept polyploid data; and provide a record of methods, standards, and control values. A long-term goal is to make the GRIN molecular tables fully interoperable with the National Center for Biotechnology Information database as well as bioinformatic databases (model organism and clade organism databases). The development of this capacity provides critical data infrastructure for future genotype–phenotype association studies and gene discovery.**

USDA's National Plant Germplasm System (NPGS) maintains the world's largest living collection of plant genetic resources. The NPGS is tasked with acquiring, preserving, characterizing, and distributing the over 450,000 accessions. Major components of the system include more than 20 active field evaluation sites, a base collection for long-term storage, quarantine, taxonomy, and plant exploration units. Critical information, including passport data and phenotypic characteristics about these plant materials, is available through the Genetic Resources Information Network (GRIN) database (Mowder and Stoner, 1988; Perry et al., 1988; USDA, ARS, National Genetic Resources Laboratory, 2008).

Stakeholders need access to genotypic as well as phenotypic data for accessions within the collection, and increasingly, there is a need to retrieve associated genotypic data. These data include markers for strain identification, genetic polymorphism data for diversity studies, and mapped markers used for QTL analysis. The ARS Action Plan for National Program 301, "Plant Genetic Resources, Genomics, and Genetics Improvement," calls for an increased database capacity for molecular marker genotypes and genetic profiles that facilitates linkages between phenotypic and genotypic data tables in the GRIN database and allows for broader interoperability (National Program 301 Action Plan, 2008).

The USDA-ARS Database Management Unit (DBMU) is primarily responsible for developing and maintaining the automated data retrieval system (GRIN) for the collection and dissemination of germplasm information for the NPGS. The development and implementation of the molecular tables were collaborative efforts between the DBMU and the USDA-ARS National Center for Genetic Resources Preservation with input and advice from researchers in the NPGS.

The GRIN database previously held only information relating primarily to taxonomy, passport, inventories, distributions, and phenotypic data (USDA, ARS, Database Management Unit, 2007). The new molecular tables in GRIN add the capacity to associate genotypic data with the existing phenotypic data. The revised tables accommodate multiple marker types; provide raw data for individuals; accept polyploid data; and provide a record of methods, standards, and control values. The revised tables also are structured so that interoperability with other databases will be possible.

Database Components

The genomics community continues to generate large data sets comprised of sequence data from expressed sequence tag (EST) studies, mapping projects, and fragment analysis for diverse species. Some of these species have designated bioinformatic model organism databases or clade organism databases that directly support these projects and provide a curated repository for molecular data. However, such database resources are not available for many of the thousands of species maintained in the NPGS. The development of GRIN molecular tables, therefore, serves as a resource for a great number of taxa for which there currently exists no alternative database.

Genetic resources in the NPGS are maintained as "accessions"; these are individual sample units in the collection that may be classified as cultivars, genetic stocks, landraces, and wild relatives of crop species. The accessions themselves may be as genetically uniform as a single genotype (in the case of clonal crops) or as heterogeneous as a bulked sample from a wild population. These numbered accessions are assigned either plant inventory numbers or site-specific local numbers and can be further classified into "inventories" or subgroups of accessions that result from regeneration events, specific individuals within a diverse accession, or other groupings.

Genotypic data for NPGS accessions are represented by broad-based diversity analyses of polymorphisms to more narrowly focused QTL studies. In our revised database, evaluations using molecular data are classified into categories to facilitate searching on these criteria in the future. The four new tables added to GRIN accommodate specific molecular data types: amplified fragment length polymorphism, allozyme, sequence, microsatellite, restriction fragment length polymorphism, and their variants, including CAPS and SCAR markers. In addition, sequence-based markers will be accommodated including genomic and genic (EST) derived single nucleotide polymorphisms.

The Marker table, Marker Citation table, Genotypic Assay table, and Genetic Observation tables comprise the new GRIN Genetic observation area (Figs. 1 and 2). The Marker table contains general information about the markers used in the assay. These data are specific to each crop and links to the Crop tables with the crop number. Information within the Marker table could originate from prior studies cited in the literature. The Marker Citation table is formatted to hold the literature citations for the markers and links will eventually enable users to acquire the cited literature from materials available online. The Genotypic Assay table describes the methods used by the laboratory where the genotypic data were collected. Selected fields include: amplification, detection, scoring, and controls. Assays are linked to the evaluation table that documents a phenotypic or molecular data collection event. Finally, the Genetic Observation table holds the raw data matrix (allele calls or sequence alignment) for each

Received for publication 4 Mar. 2008. Accepted for publication 5 Apr. 2008.

This work was funded by institutional sources.

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The authors acknowledge and appreciate their close working relationship with the programmers at the Database Management Unit.

[1] To whom reprint requests should be addressed; e-mail Gayle.Volk@ars.usda.gov

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-11-29 via free access

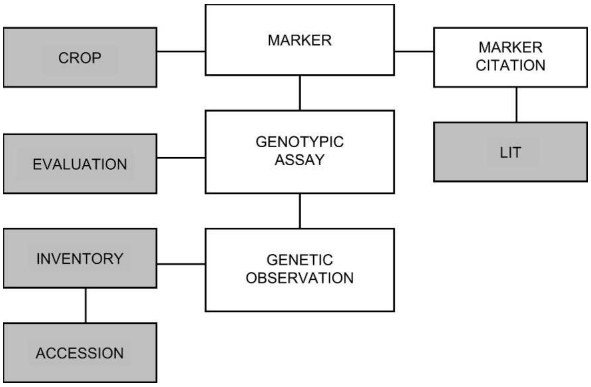

Fig. 1. Relationships among molecular tables (not shaded) and connected preexisting tables within GRIN (shaded).

Marker table (mrk)
Marker nmber (mrkno)
Crop number (cropno)
Site code (site)
Marker (marker)
Synonyms (synonyms)
Repeat motif (repeat_motif)
Primers (primers)
Standard assay conditions (assay_conditions)
Range products (range_products)
Genbank number (genbank_no)
Map location (map_location)
Position (position)
Comment (cmt)
Polymorphic type (poly_type)

Genetic observation table (gob)
Genetic observation number (gobno)
Genotypic assay number (gano)
Inventory identifier (ivid)
Individual (indiv)
Genetic observation (gob)
Genbank link (genbank_link)
Image link (image_link)

Genotypic assay table (ga)
Genotypic assay number (gano)
Marker number (mrkno)
Evaluation number (eno)
Method (method)
Scoring method (scoring_method)
Control values (control_values)
Number of observed alleles (no_obs_alleles)
Maximum observed alleles (max_gob_alleles)
Size alleles (size_alleles)
Unusual alleles (unusual_alleles)
Comment (cmt)

Marker Citation Table (mcit)
Citation number (citno)
Marker number (mrkno)
Reference abbreviation (abbr)
Citation title (cittitle)
Author of publication (author)
Citation year (cityr)
Citation reference (citref)
Comment (cmt)

Fig. 2. Columns within the Marker, Genotypic assay, Genetic observation, and Marker citation tables in the Genetic Observation area of GRIN.

individual within an inventory (linking to the Inventory tables of GRIN) for a genotypic assay. Presentation of data at the individual level is a new feature for GRIN; now specific seeds (or individuals) within an inventory can be genotyped and documented. These data are numerical, binary, or alphabetic (for sequence data) depending on the marker type. When additional data are available, it is provided as a downloadable spreadsheet link. The Genetic Observation table has the optional capacity of holding additional data such as an image of the gel, autoradiogram, or trace file.

The improved capacity for holding genotypic data in GRIN has made apple, pear, blueberry, hops, hazelnut, and cacao allelic data publicly available. Additional data will be uploaded as data sets are published. The availability of allelic data enables users to analyze data sets for sets of individuals that include the desired phenotypic and genotypic data. For example, unified records of morphological and molecular characters are critical basic research in association mapping studies and gene discovery as well as applied research in core subset development and collection management. By providing detailed methods, citations, and control values, users can evaluate the quality of the data provided.

Future Directions

It is imperative that databases evolve to become interoperable with other genomic, genebank, and environmental (or geographical information system) databases. Databases must have compatible fields for such interoperability to be possible. Interoperability allows for the combination and synthesis of data from disparate sources using middleware services (Casstevens and Buckler, 2004). An example of interoperability in the biodiversity discipline is the Global Biodiversity Information Facility (www.gbif.org) of which the NPGS is a data provider. Additionally, direct links have already been established between sequence data in GRIN and NCBI. These direct links can also connect accession numbers in GRIN to sequence data for those accessions maintained at NCBI. In the future, the connectivity between these databases will be fully established (currently it is only functional for a small number of accessions).

It is also recognized that model organism and clade organism databases provide users with valuable map, marker, and genomic data. Complementary features of these databases and the passport and phenotypic data available in GRIN enhance the prospects of future interoperability. Potential fields for interoperability between GRIN and other databases include, but are not limited to, accession and inventory identifiers, markers, control values, and primers. As interoperability standards are established, current column formats may have to be slightly modified to enable smooth interconversions and reach-through capacity, which would allow users to initiate a search query in one database and have it return data from another.

Curated databases are continuously adapting and improving to meet the needs of their users. The original design of the GRIN database was robust and it has succeeded in its mission to digitize germplasm requests and to provide data for NPGS accessions. However, user needs continue to evolve and scientists, breeders, and cooperators now require raw genetic data sets to be made available and downloadable directly from the public web interface. It is currently possible to download tables of allelic data from GRIN by electing to "Search GRIN" from the homepage and then clicking on "Data Queries" and selecting the "List of Genetic Markers" (USDA, ARS, National Genetic Resources Laboratory, 2008). Data are downloaded by clicking on buttons at the bottom of selected marker or evaluation pages. A primary goal is to develop an improved GRIN public interface to simplify queries and enable users to download genetic data in alternative formats. This endeavor will likely be a collaboration among GRIN users and developers. These future alterations could involve adopting established ontologies and formats that will facilitate interoperability with other databases in the future.

Literature Cited

Casstevens, T.M. and E.S. Buckler. 2004. GDPC: Connecting researchers with multiple integrated data sources. Bioinformatics 20:2839–2840.

Mowder, J.D. and A.K. Stoner. 1988. Information systems. Plant Breed. Rev. 7:57–65.

National Program 301 Action Plan. 15 Feb. 2008. <http://www.ars.usda.gov/research/programs/programs.htm?np_code=301&docid=13280>.

Perry, M., A.K. Stoner, and J.D. Mowder. 1988. Plant germplasm information management system: Germplasm Resources Information Network. HortScience 23:57–60.

USDA, ARS, Database Management Unit. 2007. Plant data dictionary, National Plant Germplasm System.

USDA, ARS, National Genetic Resources Laboratory. 2008. Germplasm Resources Information Network (GRIN). [online database] National Germplasm Resources Laboratory, Beltsville, MD. 15 Feb. 2008. <http://www.ars-grin.gov/npgs/>.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-11-29 via free access