New Challenges for Data Management in Genebanks

C.M. Richards and G.M. Volk
USDA, ARS National Center for Genetic Resources Preservation
1111 S. Mason St., Fort Collins, CO, 80521
USA

Keywords: ex situ conservation, genebanks, population genetics, ecological data

Abstract
The use of genetic resources for crop improvement has undergone a fundamental shift. Continued progress will be dependent upon the natural variation contained within the world’s genebanks. Genebanks must manage their collections in ways that promote their utilization by increasing access to not only the living materials but also to the vast amounts of data that describe them. Documentation of ecological, phenotypic, pedigree, and genetic data associated with genebank accessions is critical for the efficient use of genetic resources. We must develop a capacity within the genetic resources community to integrate genomic data from gene discovery projects to improve core genebank management and services.

INTRODUCTION
Genebanks serve a key role in providing genetic diversity for a host of interrelated research. Like most biological repositories, genebanks are responsible for not only the integrity of the samples they contain but also a set of associated data that identifies and describes accessions in the collection. Genebanks are unique in that samples are living—a major part of collection management is the maintenance of viability over time. Plant genebanks maintain viable propagules in the form of field collections, seeds, vegetative tissues preserved cryogenically and tissue culture samples. The USDA National Plant Germplasm System (NPGS) contains over 500,000 accessions representing roughly 13,000 species. The continued acquisition and maintenance of genetic diversity has been motivated more by the certainty of loss of local landraces and the vulnerability of wild populations in the face of increased habitat loss than from any specific evidence for agricultural worth. Moreover, the storage and maintenance of large collections is an investment that seems particularly cost effective compared to the risk of losing potentially valuable genetic resources (Pardey et al., 2001). Yet large collections often contain redundancies and are logistically cumbersome to work with if characterization data are incomplete.

Genebanks have an obligation to make their genetic diversity useful and accessible for breeding and research. Breeders, once the primary users of accessions, are being joined by molecular geneticists focused on identifying genes of agronomic or adaptive significance. As large biological databases containing genetic linkage maps and EST sequences have become more accessible, the technical process of locating quantitative trait loci in model systems has become routine. These advances have proven to be a powerful new tool for characterizing plant genotype-phenotype relationships. Empirical evidence suggests that much of the genetic variation useful to agricultural improvement is not recognizable in the plant phenotype ( Tanksley and McCouch, 1997). Therefore, instead of screening for promising phenotypes, continuing agricultural improvement may rely on allele mining at loci of agronomic importance within wild germplasm, guided by some information about phylogeny, population structure and genetic diversity (Charlesworth et al., 2001; Alonso and Ecker, 2006).

Plant collection expeditions actively seek novel as well as representative wild germplasm for inclusion within genebanks (Fig. 1). This inclusion of wild relatives is critical for agricultural improvement. The idea that natural variation contains untapped...
genetic diversity useful for agricultural traits has greatly influenced the scope of worldwide collecting efforts; however, only a few modern varieties have been developed directly using wild materials. The reason may be partly because of the need for pre-breeding lines that serve as a first step in a backcrossing program and partly because the characterization of the germplasm is inadequate or inaccessible.

INFORMATION MANAGEMENT WITHIN GENE BANKS

In the past, genebank databases focused primarily on tracking the source history and distribution of each plant inventory. In addition to inventory control (availability, location), databases increasingly play a role in the identification of desirable accessions for targeted programs. Germplasm collections that are well-characterized using standardized, defined descriptors are more frequently used than those that are poorly described (Rubenstein et al., 2006). Key associated data include documentation of original habitats, sampling methods, collection maintenance, pedigree, phenotypic information (biochemical, biotic and abiotic resistance, phenological, morphology, production), and molecular variation. This annotation is critical for sample authentication, calibration, and characterization.

Ecological Data

Traditionally, ecological data has included latitude, longitude, elevation, country, state/province, and locality information. Other fields such as the accuracy of georeferencing data, soil conditions, slope, aspect, land element types, land use, landscape narratives, International Union for Conservation of Nature (IUCN) classification, plot area, sampling method, and quality of sample are also valuable. Climatic information (air temperatures, rainfall, precipitation, hardiness zone) may also be available through local sources or through federal databases by cross-referencing latitude-longitude data.

In genebank collections, ecological data is most frequently associated with the original source location, but it also may be useful to document the environmental conditions of seed production during grow-outs and phenotypic characterizations. In addition to enhancing databases for the public by adding fields for ecological data, genebank managers can improve their own operations by making use of such data. Consideration of associated ecological data increases the efficiency in planning and performing collection trips. For example, the field of landscape genetics combines spatial and genetic data to estimate population history and the structure of local adaptation (Manel et al., 2003). Such analyses could be used to identify key geographical regions for further collections. Combining genetic and ecological information can also help identify hot spots of genetic diversity that may provide valuable sources for novel germplasm. Thus, ecological information can also inform managers on how collections can be improved.

Genetic Data

Genetic marker data can be particularly useful for identifying genetically similar accessions (fingerprinting) and can assist in the prioritization of collection materials (Benson et al., 2001; Gokirmak et al., 2009). Genetic data may also be used to assemble small diversity reference sets for use in associative mapping or marker development (Volk et al., 2005; Richards et al., 2009). However, genetic data contained in genebank databases will be most valuable if key reference samples and standards are also included. Associated pedigree information could allow researchers to determine related germplasm with similar traits in different genetic backgrounds. Pedigree data may also be valuable in identifying key genes underlying traits of agronomic importance (Bink et al., 2008).

Until recently, genetic marker data were not available in the Germplasm Resources Information Network (GRIN), the database for the NPGS (Volk and Richards, 2008). Although GRIN can maintain limited amounts of marker data, it is not currently capable of holding large EST or sequence datasets. We are hopeful that improvements in interoperability among molecular biological databases including the GRIN-Global
initiative and the Global Portal for Genetic Resources, clade oriented databases and genebank databases will provide access to large-scale molecular data. Although making genetic variation accessible to researchers may require adjustments in the way data are distributed among databases, a community-based approach will increase the capacity for biodiversity informatics.

CONCLUSIONS

Genebanks have core functions that will benefit from greater integration of ecological and genetic datasets. The demand on genebanks to provide collection information has increased as technologies enable the exchange of electronic biodiversity information. Enhanced ecological documentation can serve to improve collection procedures to better capture the genetic diversity of wild populations. In addition, better documentation of field sites in which phenotypic data have been collected will assist in our understanding of genotype by environment interactions. Molecular documentation of inventories will aid in identifying genetic shifts that may have occurred during storage and regeneration. Combined with phenotypic data, molecular data will also provide information on collection diversity, aid in the assembly of core subsets, and improve regeneration protocols. Interoperable databases that provide key inventory, phenotype, ecological, and genetic data will ensure that genebank materials can be efficiently used for the development of new varieties and new crops.

ACKNOWLEDGEMENTS

We would like to thank Christina Walters, Pat Reeves, Rich Hannan, Angela Baldo, Clare Coyne, and Dale Lockwood for helpful discussions.

Literature Cited

Figures

Fig. 1. Diagram demonstrating the flow of both plant genetic resources and data within genebanks and associated databases. Plant icons represent germplasm within genebanks and bar codes represent data maintained for genebank materials.