

## **GDPC: The Genomic Diversity and Phenotype Connection: accessing data sources via XML web services**

*Terry M. Casstevens<sup>1</sup> and Edward S. Buckler<sup>2</sup>*

<sup>1</sup>*Departments of Statistics and Genetics, North Carolina State University, Raleigh, NC 27695-7566,* <sup>2</sup>*USDA-ARS, Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703*

Research projects on genomic diversity and phenotypes have generated valuable data collections that tend to be abandoned after results are published. Ideally, this data would be made widely available by migrating the data to larger, public databases. The purpose of the Genomic Diversity and Phenotype Connection (GDPC) is to simplify access to data and thereby increase its effective reuse. GDPC is software written in JAVA that works with taxa, loci, environment experiments, genotype experiments, localities, genotypes, and phenotypes. GDPC provides: 1) data sources that are XML web services; 2) programmatic access to data sources; and 3) a front-end application that allows users to retrieve data.

GDPC provides the infrastructure to create connections to data sources that mask the complexities of the data's underlying format/schema. These "GDPC connections" typically support databases and are designed as web services that transfer XML via the SOAP protocol. "GDPC connections" can also be designed to access databases using the JAVA JDBC API. Future plans include developing a connection capable of accessing local flat files. Researchers can create connections to their data that allow them to integrate it with public data. Once in the GDPC format, it is possible to integrate, analyze, and view data from multiple sources. This is a significant advantage over tools that access only one source. A "GDPC connection" already exists to the maize diversity database, Panzea (<http://www.panzea.org>). Other "GDPC connections" are being developed, including one to the comparative cereal database, Gramene (<http://www.gramene.org>). JAVA classes compatible with GDPC are provided to help other organizations make their data "GDPC enabled." Future plans include registering "GDPC connections" with MOBY Central (<http://www.biomoby.org>) to allow users to locate them via that directory service.

Several research projects make their data available via web sites, but programmatic ways to retrieve data generally do not exist. In contrast, GDPC provides a JAVA API that standardizes access to data sources. Programmers can use this API to develop "GDPC aware" front-end applications that perform algorithms relevant to their project. The GDPC Browser and TASSEL (<http://www.maizegenetics.net/bioinformatics/tasselindex.htm>) are applications that currently use this API. The planned development of a Mesquite module (<http://mesquiteproject.org>) will make any "GDPC connection" accessible from the Mesquite toolset.

The GDPC Browser is a front-end application that allows users to retrieve, view, and group data based on property values. With this application, users can connect to a data source(s) and retrieve data based on user defined search criteria. The data properties can be viewed by selecting the individual data elements. Working lists of these elements can be created and sorted based on user needs. These working lists can be saved/opened as XML files. Data can be exported to other formats (i.e. PHYLIP) chosen by the user. This tool and other analysis tools will facilitate QTL linkage mapping and association mapping on a grass and genome wide level.

The source, binaries, documentation, etc. are freely available at:  
[www.maizegenetics.net/gdpc/index.html](http://www.maizegenetics.net/gdpc/index.html).