

# **Gramene Genetic Diversity Module:**

## A Plant Genotype-Phenotype Association Data Repository

Jon Zhang<sup>4</sup>, Genevieve DeClerck<sup>1</sup>, Athikkattuvalasu Karthikeyan<sup>1</sup>, Shuly Avraham<sup>5</sup>, Terry Casstevens<sup>4</sup>, Charles Chen<sup>1</sup>, Jer-Ming Chia<sup>5</sup>, Ken Youens-Clark<sup>5</sup>, Aaron Chuah<sup>5</sup>, Paul Derwent<sup>6</sup>, Shiran Pasternak<sup>5</sup>, Liya Ren<sup>5</sup>, William Spooner<sup>7</sup>, Joshua Stein<sup>5</sup>, Jim Thomason<sup>5</sup>, Qi Sun<sup>2</sup>, Sharon Wei<sup>5</sup>, Pankaj Jaiswal<sup>6</sup>, Doreen Ware<sup>3,5</sup>, Edward Buckler<sup>3,4</sup>, Susan McCouch<sup>1</sup>

- 1 Dept. of Plant Breeding and Genetics, 240 Emerson Hall, Cornell University, Ithaca, NY 14853, USA
- 1 Dept. of Plant Breeding and Genetics, 240 Emerson Hall, Cornell University, Ithaca, NY 14853, USA
  2 Computational Biology Service Unit, Life Sciences Core Laboratories Center, Cornell University, Ithaca, NY 14853, USA
  3 USDA-ARS, Robert W. Holley Center, Cornell University, Ithaca, NY 14853, USA
  4 Institute of Genomic Diversity, Cornell University, Ithaca, NY 14853, USA
  5 Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY 11724, USA
  6 Dept. of Botany and Plant Pathology, 3082 Cordley Hall, Oregon State University, Cornallis, OR, 97331, USA
  7 Emposa Delioferantic Leditions, Medicar Text Genome Comput. Hinters Compiler Control Control

- 7 European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Abstract: Identifying genotypic variation has become a routine procedure given the increasing number of high-throughput sequencing and array-based genotyping technologies. However, it is still a challenge for the plant research community to associate this data with phenotypic variation. More specifically, groups attempting to illuminate genotypic-phenotypic associations are encountering bottlenecks in areas of data post-processing, data storage, computationally intense analysis of results. The Gramene Genetic Diversity Module (<a href="www.gramene.org/db/diversity/diversity/view">www.gramene.org/db/diversity/diversity/view</a>) is designed to help overcome these challenges. Using the Genomic Diversity and Phenotype Data Model (GDPDM; maizegenetics.net/gdpdm), this module of Gramene is able to compress genotype data by several orders of magnitude facilitating integration and analysis of genotypic and phenotypic data. Gramene Diversity currently hosts a number of important variation datasets for rice, maize, *Arabidopsis*, sorghum, and wheat. Recent curational activity has been on large-scale SNP genotype datasets in rice, maize and Arabidopsis, with associated phenotype data assigned to ontology term. New terms are created when necessary. In addition to the data, Gramene Diversity offers several options for querying, analyzing, and bulk downloading of the data. Recently we have added a GWAS results visualization tool (www.gramene.org/db/jshc2010), our newest addition to an array of analytical options. Gramene Diversity is updated twice a year as part of the biannual Gramene release, the latest of which (Build 34) was in the Fall of 2011.

## Summary of Datasets Hosted on Diversity

As data storage is a bottleneck, Gramene Diversity hosts many large scale genotype, phenotype, environment, germplasm and association data Currently we host such data for 5 species, Arabidopsis, Maize, Rice, Sorghum, and Wheat using the GDPDM schema

Species	Diversity Data Sources
Arabidopsis	Large scale SNP data from 1001 Genomes and NSF 2010 <i>Arabidopsis</i> Project
Maize	Molecular and Functional Diversity of the Maize Genome Project, <u>Panzea.org</u>
Rice	QTL and SSR based diversity studies from over 20 publications; Large scale SNP data from ricediversity.org and oryzasnp.org, NSF Rice Diversity, OryzaSNP, and Rice Haplotype Map projects
Sorghum	Sorghum Diversity Project
Wheat	Haplotype Polymorphism in Polyploid Wheats and Their Diploid Ancestors project

#### GDPDM: Genomic Diversity and Phenotype Data Model The Gramene Genetic Uses BLOBs for storage Diversity database module Supports full range of ploidy implements the GDPDM schema. GDPDM is designed to capture both Molecular and Environment Associations Phenotypic Diversity data, Germplasms •Geno-pheno link Observations whether this data is a product •Pedigree description •Seed stock Largely of QTL mapping, association •Planting expanded since studies, large scale SNP •Treatment last version genotype assay studies, field Locality data from plant breeding programs, or germplasm evaluation studies Phenotypes Quantitative or qualitative traits GDPDM: www.maizegenetics.net/gdpdm ·Supports ontology intergration

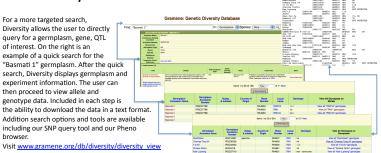
Diversity

Q

The data is available fo hulk download in a variety of formats, hapmap, plink, flapjack directly from the Diversity website. In addition if the user would like to go directly into data analysis, launch links are available, allowing the user to start TASSEL and load the desired data in one click



### **Data Availability**



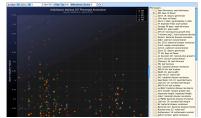
# TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) et/tassel) is a Java software package for evaluating trait associations, evolutionary patterns, and calculating and graphically visualizing data. TASSEL runs locally on your computer with launch links directly from Diversity.

## **Diversity Tools**

Gramene Diversity offers advanced tools for searching, filtering, and analyzing data. TASSEL, left, runs locally on your computer with connectivity with the Diversity database. SNP Query, below, is a web based search tool used to interact with the large scale SNP datasets. GWAS Association Viewer, right, is our new web based GWAS visualization tool.



search tool. It allows the user to search top down, going from a species and publication down to the gene and SNP level or bottom up, from the gene and QTL level up to the publication. Either way the tool will return subsets of data appropriate to the given query and provide download links for the data in various formats.



GWAS Association Viewer (www.gramene.org/db/jshc2010) is the latest entry in our line up of available tools. This tool allows the user to plot GWAS results on an interactive graph. Users can choose specific traits to be plotted, zoom into specific regions on the graph and set specific p-value cutoffs.











