

**Barley Coordinated Agricultural Project Work Plan FY07 (4/1/07 – 3/31/08) Objective 2:
THT Development**

Julie A. Dickerson, Roger Wise, Iowa State University

1) Describe the research, education, and outreach activities you are planning for the next year (4/1/07 – 3/29/08)

THT Development (Dickerson, Wise)

An initial version of the website that supports curation tools for the entry/editing of data in both spreadsheet and single record format for trait and pedigree data will be deployed at the end of the first quarter 2007. The database schema will be adapted from the Germinate schema. The first version website will allow searching by breeder, location, pedigree, or measured traits. The users will be able to refine searches and download their results in a spreadsheet compatible form. The site will be tested and refined over the second quarter.

Storage of SNP data will be added to the site in the second half of 2007. Initially, the data will be only available after lines have been selected. Tools will be added to allow searches by common haplotypes. Visualization tools for SNP data such as CMAP or adaptations from the HapMap project will be added.

File format development. QTLMiner and SNP data formats will be finalized. Output data file formats for download will also be deployed. All traits will be fully defined with value ranges and “reasonable” inputs given for developing a robust data verification process.

Data curation.

Data from the 2006 cropping season needs to be standardized against the list of proposed traits for review and upload to THT. Collect descriptions of measurement processes from all participants. The first set of Illumina SNP data will be available in mid-2007. The data must be curated to be sure that lines are correctly named. The calls will be stored in a web-searchable format and the raw data will be compressed and be made available for download on THT.

Data Integration

The main task is to develop a structure for integrating information from a variety of websites including GrainGenes, Gramene, WebHarvEST, and Germinate. This may involve web services and the development of common viewing tools. Plan is to send out a draft document on proposed interactions then schedule a set of web meeting(s) with the different databases to discuss coordination.

Education and Outreach:

The main outreach component will be to make THT publicly accessible on the web and to publicize it to the Barley community.

2) List specific outcomes and deliverables that will be accomplished in the first 6 months (4/1 – 9/30). These will be used as benchmarks for your bi-annual progress report.

- THT website will be available for search by pedigree and trait by members of the BarleyCAP team under password protection until the datasets from the 2006 season are publicly available. The public website capability will be available when the data is approved for public release.
- The Germinate schema and haplotype viewing tools will be adapted for use by the BarleyCAP project and will be made available for use.
- Data pipelines for all data sources will be developed so that incoming data from Oregon and North Dakota can be downloaded and entered into the database.

3) List specific outcomes and deliverables that will be accomplished in the second 6 months (10/1 – 3/31). These will be used as benchmarks for the bi-annual progress report.

- Users will be able to format input data files for the QTLminer program from the results of searches.
- Users will be able to display and manipulate haplotype data on the THT website.
- Develop a functional specification for linking BarleyCAP data to other sites such as GrainGenes and PLEXdb.

Barley Coordinated Agricultural Project Year 1 Progress Report
Objective 2: THT Development
FY06 (4/1/06 – 03/31/07)
Julie A. Dickerson, Roger Wise, Iowa State University

1. Research, education, and outreach activities in FY06 (4/1/06 – 03/31/07)

THT Development (Dickerson, Wise)

We began the process of developing The *Hordeum* Toolbox (THT), an integrated public resource for discovery and dissemination of large-scale SNP data sets as it applies to genetic, phenotypic, and trait data.

Use Case Development: Use cases for three different users, a data curator, plant breeders and academic researchers were developed and circulated to the project team. Comments from the potential user base were incorporated into these documents. User scenarios have also been developed to guide website development.

Contractor Statement of Work Development: A functional specification has been developed for THT using an outside contractor, BioNeos to develop search tools and user interfaces for THT.

Germinate Schema Review: A draft version of the SCRI Germinate database schema was received in mid-February 2007. We have reviewed the schema and are revising it to focus on the specific data being generated by the CAP project.

Phenotype Variable Definition: Requests for clarification and details on the measured phenotypes used in the CAP project were sent out in early February 2007. These details include parameter ranges, units and measurement descriptions. Partial details were received in early March. This information is important to ensure data integrity in the database and to provide automated data checking on the input of the data.

Data Collection (Kling, Dickerson)

File format development. File formats and standardized spreadsheets developed for pedigree information, phenotypic data, QTLMiner, and SNP data were agreed upon with the respective research groups. BarleyCAP breeders, THT developers, and input from TASSEL and GERMINATE (SCRI) teams were used to develop scheme for pedigree input. Lists of standardized annotations for data sets were developed in consultation with staff at GrainGenes and Gramene for defining traits. Formats for checking means and standard errors from each testing site developed.

Data curation(Kling, Dickerson, Smith, Bradbury).

Pedigree files for all of the pilot OPA2 genotypes were created for uploading into THT. Input of the pedigrees for the 960 CAP lines for 2006 is in progress. Data from the 2006 cropping season has just started to come in, for review and upload to THT.

Data Integration (Carolla-Blake (and GrainGenes Team),Dickerson, Wise, Close)

Integrate SNP and gene expression data. Every SNP marker will be mapped back to the sequences on the Barley1 GeneChip using the assemblies in the HarvEST database and the gene locus names. All SNPs will also be labeled according to the proposal from David Marshall at the SCRI.

The GrainGenes project continues to add genetic maps, barley QTLs and new SNP markers to the GrainGenes database. QTL catalogued previously from the BeerGenes and BarleyWorld projects are being validated in the originating literature and added to GrainGenes. New QTLs from publications since 2003 are being noted and records created. The Barley QTL Community Curation Workbook (http://wheat.pw.usda.gov/ggpages/qtl_worksheet_beta.xls) is available on GrainGenes, and has been beta-tested by several members of the database community. This data will be linked to THT.

A web-based version of HarvEST (<http://www.harvest-web.org/>) developed by the Close Lab debuted this fall. Links to this database will be used to relate the SNP data to expression data at PLEXdb and to physical maps.

Education and Outreach:

- Discussed the Barley CAP project informally at the International Rice Genome Conference in Montpellier, France (October, 2006) and at the Data Warehouse Workshop in Halle, Germany (December, 2006). (Dickerson)
- Information about the Barley CAP project was displayed on one of several posters in an ARS booth at the New York Empire Farm Days in August 2006 (Bradbury).
- Presentation entitled “The Hordeum Toolbox - A public resource for integrating pedigree, SNP, EST, and phenotypic data from breeding germplasm” at the Plant Ontology Consortium Workshop in Ithaca in May. (Kling)
- Photographs of the Barley CAP nurseries in Bozeman Montana were taken, and posted with annotations at <http://hordeum.oscs.montana.edu/capnursery/index.htm> .

2. Specific outcomes and deliverables accomplished in FY06 (4/1 – 3/31).

- Web site design based on use case analysis is underway. Registered the domain names *hordeumtoolbox.org* and *thehordeumtoolbox.org* for THT web site.
- Germinate schema was altered in accordance with the CAP data needs.
- Finalized plans for submission of breeder data from Jennifer Kling using a spreadsheet interface. Developed system for pedigree data entry and compiled data for the pilot OPA2 genotypes
- Developed a list of measured traits and their units and establishing identifiers for all participants using a uniform numbering scheme from Kevin Smith.
- File formats for data exchange with QTL Miner established.
- Developed table design for SNP data results. Reported allelic data for each line will be searchable by users and raw data files will be available in zip files if users wish to apply their own environment to the data.
- Researched off-site data storage options for geographically diverse data storage. All data tables will be backed up monthly and stored at a commercial vendor.