MASC Phenomics- 2006 Subcommittee Report
Prepared by Eva Huala and Sean May (co-chairs)

The Phenome subcommittee is meant to track the progress of efforts including development and availability of biological materials for phenotyping, efforts to develop and use standardized phenotype descriptions including ontologies, and the availability of materials and data in stock centers and databases.

Biological resources for phenotyping

Insertion lines: A total of 28,607 genes (92%) have been tagged with insertions in the combined set of all insertion collections, and 23,509 (76%) of these genes contain insertions into the coding region, according to the SIGnAL website (http://signal.salk.edu/cgi-bin/tdnaexpress). The insertion sets were generated from the Salk, SAIL, GABI-Kat, FLAG FST, JIC, Wisconsin, RIKEN and CSHL projects. A list of projects with brief descriptions, location of data and links can be found at TAIR: http://arabidopsis.org/portals/mutants/worldwide.jsp. Most sets are available for ordering from both NASC and ABRC, with the exception of the RIKEN lines (available from RIKEN) and the FLAG FST lines (available from INRA).

Homozygous mutant lines: The Salk "phenome-ready" project to produce homozygous mutant T-DNA lines is in the process of generating homozygous mutants for 24,357 Arabidopis protein-coding genes. Currently (as of March 31 2006), seeds for 1989 lines (affecting 1740 genes) are available from the *Arabidopsis* Biological Resource Center (ABRC) and will also be available from NASC. The remainder of the Salk homozygous lines should be completed in 2007, and seeds will be available in quantities sufficient to allow partial or full sets of lines, or pools of lines, to be utilized by researchers for research projects and/or screening. A second large project at RIKEN is now preparing homozygous mutant transposon-tagged lines for approximately 5,000 *Arabidopsis* genes. RIKEN BRC (http://www.brc.riken.jp/lab/epd/Eng/) will start the distribution of homozygous seeds by the end of this year. In addition to these two large collections, 642 Salk and SAIL homozygous insertion lines from the community have been donated to ABRC as of April 2006 and are available for ordering.

RNAi lines: The AGRIKOLA project (http://www.agrikola.org/) has cloned 21,862 gene-specific tags into binary hairpin RNA vectors and made them available to the public through NASC. These lines target a total of 19,202 genes. The final targets to be achieved in mid-2006 will be over 25,000 tags for over 20,000 genes. In addition, 786 sets of transformed lines and 6755 individual transformed lines are currently available for ordering from NASC, out of a final target of around 3,000 sets of transformed lines. The AGRIKOLA lines are also being made available by ABRC (200 lines currently in-house). A basic phenotypic description of the transformed lines is being recorded using an EAV model developed in collaboration with NASC. This phenotypic data should be available in mid-2006.

Artificial miRNAs: Greg Hannon (CSHL), Rob Martienssen (CSHL) and Detlef Weigel (MPI) have submitted an *Arabidopsis* 2010 proposal to create a universal library of 3 artificial miRNAs per gene. In addition, artificial miRNAs for tandemly duplicated and paralogous (recent segmental duplication) genes are planned.

Ontology development

Several efforts are underway to provide the controlled vocabulary needed to describe phenotypes in a standardized way.

POC: The goal of the POC (Plant Ontology Consortium, http://www.plantontology.org/) is to provide a standardized vocabulary to describe anatomy, morphology, and growth and developmental stages in *Arabidopsis*, maize, rice, Fabaceae and Solanaceae, with future plans to extend the vocabularies to other plant families. This standardized vocabulary will serve as a basis for cross-species queries for phenotype

and gene expression information, an essential capability that will permit researchers working with crop species to leverage *Arabidopsis* research results into agricultural advances. POC has released ontologies covering plant structure (release date July 2004, currently 753 terms) and growth and developmental stages (release date July 2005, currently 274 terms). These ontologies are now in use at TAIR and NASC as well as several other databases including Gramene, MaizeGDB, SGN, BRENDA, Genevestigator and ArrayExpress. Requests for additional terms can be made by using the Feedback button on the POC website.

PATO: The Phenotype, Attribute and Trait Ontology (PATO) is currently under development as part of the NIH-funded cBiO project (http://www.bioontology.org/). The purpose of this ontology is to facilitate description of phenotypes using controlled vocabulary terms arranged in a syntax line referred to as the Entity, Attribute and Value (EAV) model. Entity terms describe physical entities such as body part, cell type, subcellular structure, developmental stage or biological process affected, and can be drawn from other ontologies such as PO (Plant Ontology) and GO (Gene Ontology). Attribute terms describe the trait that is altered, and Value terms describe the nature of the alteration in the mutant or natural variant (For example Entity:leaf, Attribute:color, Value:yellow could be used to describe a mutant with yellow leaves). The core participants in development of PATO include Monte Westerfield at ZFIN (the zebrafish database), Michael Ashburner at Flybase (the Drosophila database) and Ida Sim of UCSF, leader of a group that will use the ontology to describe the results of human clinical trials, with Georgios Gkoutos in the UK serving as the main PATO curator. The POC organizer (Katica Ilic) has been in contact with the project organizers and there is a shared opinion that the plant community should participate more actively in PATO development to ensure that this ontology becomes more suitable for description of plant mutants and natural variants. To facilitate the development of PATO for use in plant biology, curators from plant databases are encouraged to request new PATO terms through the SourceForge website (http://sourceforge.net/tracker/?group_id=76834&atid=595654). TAIR is planning to participate in a PATO content development workshop in mid-May 2006.

Phenotyping efforts

No public domain effort to date has released data systematically describing the overall phenotype of large numbers of *Arabidopsis* gene knockouts. Rather, the emphasis has been on investigating specific traits and on describing natural variants and QTLs. Large scale phenotyping of knockouts has been carried out in the past by the private sector (for example at Paradigm genetics – see Boyes et al. 2001, Plant Cell 13:1499–1510). However, the phenotype data from these projects have not been made publicly available. A group at INRA (Granier et al., New Phytol. 2006;169:623-35) has described a robotic setup for large scale phenotyping of general morphological traits but no large data releases to the community have resulted to date.

Examples of recent efforts aimed at phenotyping for specific traits include ion profiling of large numbers of mutants (Salt, Plant Physiol. 2004;136:2451-6) and searching for cell wall mutants using MALDI-TOF MS to detect structural changes in cell wall polymers (Lerouxel et al., Plant Physiol. 2002;130:1754-63).

Efforts to identify large numbers of natural variants include a project in Detlef Weigel's group to phenotype about 15 F2 populations from crosses among the "Nordborg 96" ecotype sets, examining 500 plants of each population for a variety of morphological traits including flowering time. Individual plants will be phenotyped and also genotyped at about 80 loci each, to produce a large number of QTL maps. It is planned to store the phenotypes in a genotype-phenotype database, as a prototype for the future integration of phenotyping data sets from many different sources.

Phenotyping at the level of individual genes or small sets is ongoing, and this information is primarily captured in the literature. Phenotype descriptions have been published for approximately 3000 loci,

estimated from the number of loci at TAIR that are associated to published papers containing the word "mutant" in the abstract. TAIR is working to extract this information from the literature.

Storage and display of phenotype data

NASC currently displays PO and PATO controlled vocabulary phenotype annotations for 1,897 mutant lines and natural variants, and will continue to add additional annotations in the future.

POC currently displays NASC's PO annotations for 1,897 mutant and natural variant lines. The consortium is working on displaying plant phenotype descriptions on the POC website as a combination of controlled vocabulary terms and free text (using PO terms to describing the affected part (entity), and using free text for attribute and value).

TAIR contains phenotype descriptions for mutants in approximately 1100 AGI loci, collected from the literature and from ABRC stock data. TAIR is currently working to separate free text phenotype descriptions from other descriptive information relating to germplasms and add search and display capabilities specific to phenotype. In the next 12 months TAIR expects to begin adding PO and PATO controlled vocabulary phenotype annotations through literature curation and community submissions.

(Additional report contributors: Ian Small, Randy Scholl, Detlef Weigel, Minami Matsui and Katica Ilic)