

18 and 19 February 2008 at ICRISAT

Platform Development Meeting to discuss Use Case 3.

SP3 Use Cases 2008 Document 1

Introduction

An overview of the Molecular Breeding Information System was discussed. It was agreed that there were four components:

- Molecular Breeding Design Tool (MBDT)
- Sample and marker tracker (Molecular LIMS)
- Molecular Selection Tool (MoSel)
- Data loading application

The first and third of these components are applications which need to be developed by the ICRISAT/CIMMYT team, the second and fourth are information system components mostly concerned with LIMS functions which will be implemented separately by ICRISAT and CIMMYT. The ICRISAT team has started work on the design tool, and the CIMMYT team has spent time considering the design of the selection tool. Hence it was decided that we should continue with that division of labor for the time being and look for opportunities to share components across tools.

Molecular Breeding Design Tool (MBDT)

The design tool has the following functions:

- Assist breeders in selecting parental germplasm based on phenotypic data, known adaptation, economic importance.
 - What traits are present and which are absent in potential recipients?
 - Find and display known genes and QTL available for supplying the missing traits.
 - List potential donors for missing traits.
 - Develop and display trait maps

This function relies heavily on integration and assimilation of external knowledge and information. Initially the application interface should facilitate appropriate queries, but even then most results are likely to come from external analysis and the interface should enable capture of the information from that external analysis.

- Verify and display genotypic and trait association data
 - Check availability of genotyping data for potential recipients and donors.

- If not available, formulate lists of germplasm and required marker analysis and pass the information to the Molecular LIMS for field and lab requests.
- Display graphical genotypes and known trait associations for potential recipients.
- Display genotype and trait association information on potential donors.
- Analyse compatibility between donors and recipients
 - Show donor and recipient regions
 - Choose markers (exact or flanking) for important regions. These must be polymorphic between donors and recipients, as close as possible to genes of interest, and should mark regions for introgression from donors or specific regions for retention of genotype in recipient germplasm.
 - Identify potential linkage drag
 - Analyse background similarity via genetic distances of coefficients of parentage.
 - Select background markers, polymorphic between recipients and donors and suitably spaced over the genome.
 - Select recipient and donor germplasm, design crosses and pass crossing lists to the Molecular LIMS for making crosses.
- Specify target genotypes

This function will benefit substantially from integration of simulation tools like QULine to pick crossing strategies, population sizes and most compatible parents. However this will not be attempted in year 1 of the use case development.

The MBDT has two main display formats, one for trait maps and one for displaying genotyping and trait association data for specific recipients and donors. The first has a mapping flavor and technology could draw heavily on the CMTV, while the second is a graphical genotyping tool which also lies at the heart of the third component - MOSEL.

Sample and marker tracker (Molecular LIMS)

The Molecular LIMS has the role of driving and coordination field and lab work to facilitate molecular breeding. It must accommodate the following functions:

- Sample tracking for germplasm - crosses and progeny
- Field lay-out and field book preparation
- Capture of field evaluations linked to germplasm samples
- Sample tracking for DNA extraction and analysis
- Marker and protocol identification for the molecular lab
- Specification of Sample by Marker analysis requirements

- Allow partitioned display of foreground and background markers. Problems with the generalized approach include how, or whether, to distinguish foreground and background markers. How to indicate relative importance of different segments (perhaps importance can be coded and displayed on target genotypes by line thickness or color intensity).
- Order, group and filter test genotypes
 - Compute ranking scores of test genotypes using weighted averages of distances to target genotypes. A problem is to select weights for different loci. Critical loci to be retained or introgressed should have high weights while background loci should have lower weights. One possibility is to compute separate distances for background and foreground markers then filter test lines on an acceptable threshold of foreground proximity and rank remaining lines on distance to background markers.
 - One could compute pairwise distances between lines (including test lines, target genotypes and parental lines) based on genetic and phenotypic values and show ordination plots or dendrograms.
 - Filter, sort and scroll genotype displays for test lines in proximity to target and parental lines.
- Choose lines and crossing schemes for further development
 - If the filtered set of test lines is relatively small these can be inspected individually to select a few for backcrossing, intercrossing or selfing. Many factors are used in the final choice including any available phenotypic data, as well as the number and specificity of crossovers required to approach the target genotype. It will be important to capture these decision rules from users and build them into the analytical tools.
 - Users could mark regions requiring crossover to approach the target and select a crossing scheme then the tool could compute the (map based) probability of observing the required crossover events.
 - Best lines are selected and sent to the Molecular LIMS as crossing lists for the next cycle.
 - Markers fixed in recipient populations should also be identified so that they do not need to be processed for subsequent generations.

•

•Data Loading Application

The Data Loading Application is required to allow users to identify data to be published to a central database for future use. At ICRISAT this application is need to feed selected data to ICRIS. At CIMMYT the update procedures from ICIS local to central will be used.

Development Decisions

- Development teams at ICRISAT and CIMMYT have agreed to use the Eclipse RCP environment to develop the tools.
- Existing or new projects will be set up on CropForge using Subversion for code maintenance and exchange.
- The GCP Wiki will be used to exchange design and development information until such time as a more formal bug and feature tracking system is required.
- Development of MBDT and MOSEL can progress under the control of individual teams, but prototypes will be exchanged as soon as possible and each team should act as alpha testers for the other.
- Common code modules should be identified as soon as possible (during the design phase) and decisions on which team should contribute the module should be taken.
- The first common code decision involved a graphical heat map display tool. Teams will collaborate on researching available, existing components which allow the display of large matrices using heat strip technology. Sufficient control is required to allow ordering, spacing and filtering of rows and columns.