# Integrating Genetic with Genomic Data in Legume InterMines

Sam Hokin, Andrew Farmer
<shokin@ncgr.org>, <adf@ncgr.org>

National Center for Genome Resources

Vivek Krishnakumar
<vkrishna@jcvi.org>

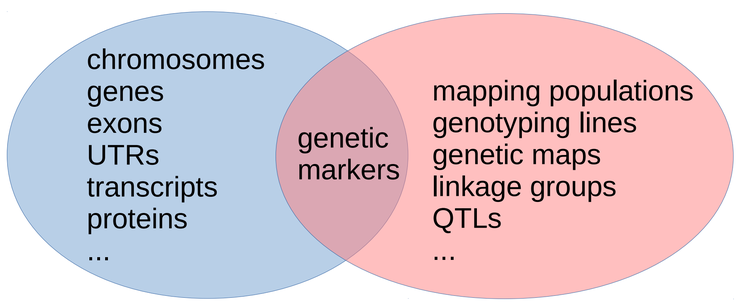
J. Craig Venter Institute

## INTRODUCTION

**InterMine** <intermine.org> is an open-source data warehouse developed by the Miklem Lab at Cambridge with contributions from developers around the world. The core app is written for FlyMine <flymine.org>, an integrated database for *Drosophila* and *Anopheles* genomics. **The core InterMine data model does not include genetic data**.

**The USDA-funded Legume Information System (LIS)** collects and provides both genomic and genetic data for legumes on the main LIS site <legumeinfo.org> as well as SoyBase <soybase.org> and PeanutBase <peanutbase.org>. The LIS and PeanutBase sites use the Tripal extension of the Drupal web application, using chado databases which have been modified to include genetic data.

**The NSF-funded Legume Federation**, in turn, was chartered to to provide a "one-stop shop" for legume breeders, geneticists and biologists, federating data sources such as LIS, JCVI's MedicMine, and others. For that purpose, **we have extended InterMine to include genetic data** and we have added new visualization and analysis tools.




Genetic markers are the glue that relates genetic to genomic locations.

Genetic data is largely independent of genomic data: genetic features are positioned on linkage groups in centi-Morgans, while genomic features are positioned on chromosomes in sequence coordinates. **Genetic markers have both genetic and genomic locations.** We make use of this to place genetic features on the genome.

## OBJECTIVES

1. Create a convenient environment for plant breeders and biologists to study legume genomic and genetic data.
2. Provide query and visualization tools to analyse genetic traits along with genomic features.
3. Enable cross-species analysis amongst legumes and other plants.

## DATA MODEL

### Genetic data

The core InterMine data model does not contain genetic data. We have added the following classes:

**Germplasm** describes the parents of a cross used in a mapping experiment.

**MappingPopulation** is a container for a genotyping experiment: **parents**, **geneticMaps**, **geneticMarkers** and **genotypingLines**.

**GenotypingLine** describes a particular plant line, e.g. an RIL, in a **mappingPopulation**.

**GeneticMarker** is a **SequenceFeature** representing a SNP or other marker, belonging to one or more **mappingPopulations**, with a **chromosomeLocation** as well as **linkageGroupPositions**. It is related to **QTLs** in the original data sources. It has **associatedGenes** which overlap its genomic location.

**GenotypeValue** is a single value on the markers × lines genotyping matrix with the value of the parental allele (usually denoted by "A" or "B").

**GeneticMap** contains the **linkageGroups** derived from one or more **mappingPopulations** (more than one in the case of a consensus map).

**LinkageGroup** contains the **geneticMarkers** and **QTLs** with their positions and ranges. Each belongs to one **geneticMap**.

**QTL** is a quantitative trait locus with one or more **linkageGroupRanges**. It has **spannedGenes** which overlap its **associatedMarkers** on the genome. (QTL–marker associations are given in the primary data sources, not by comparing the QTL to markers on linkage groups. There are often many more markers within a QTL's genetic range than are explicitly associated with that QTL by the original scientists.)

### Genomic data modifications

Several additions to the core InterMine data model are needed to accomodate genomic data pulled from the LIS databases, in particular, Phytozome gene families, as well as the new connection to genetic markers.

**Gene** has additional references to **geneFamily**, **homologues** and **spanningQTLs**.

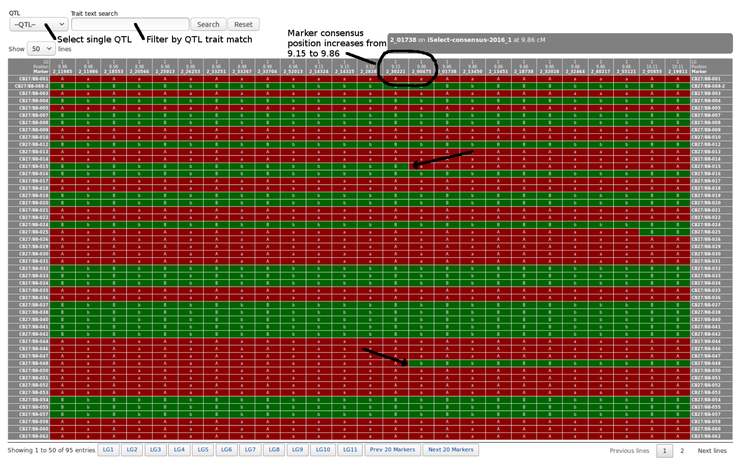
**GeneFamily** describes a Phytozome gene family, referencing a **consensusRegion**.

**ConsensusRegion** is a **BioEntity** which holds the sequence associated with a **geneFamily**.

## VISUALIZATIONS

### Mapping Populations

A genotyping experiment conducted on a mapping population is displayed on the Mapping Population report, with a color-coded markers × lines matrix inspired by Flapjack.

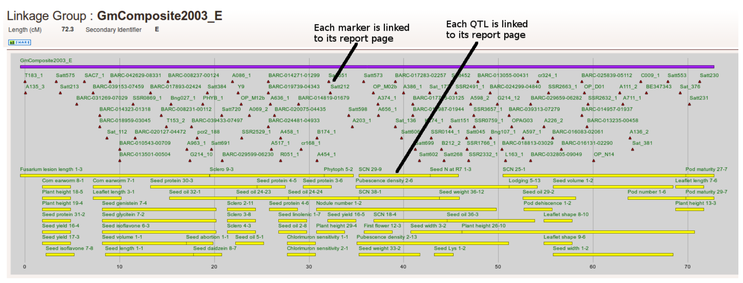



Two allele exchanges in this mapping population indicate recombination contributing to a 0.71 cM increase in linkage group distance on the consensus map.

One can filter the markers by a single QTL or QTLs with traits that match a search term.

### Linkage Groups

Linkage groups are displayed in a horizontal diagram with markers and QTLs shown at their genetic locations. The markers and QTLs link to their report pages.

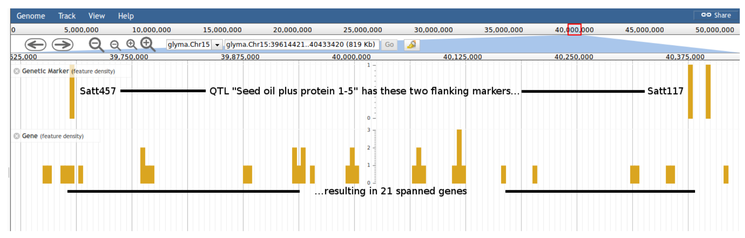



Soybean consensus linkage group GMComposite2003_E has 143 QTLs and 189 markers, 94 of which have been located on chromosome 15.

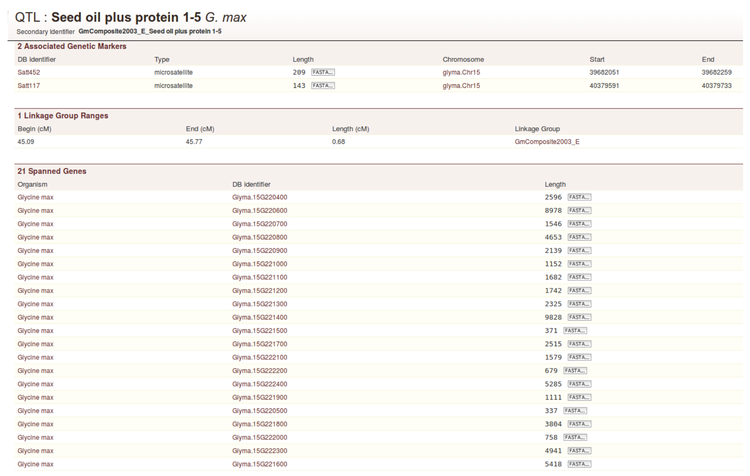
The genetic map report page displays the diagrams for all of its linkage groups.

## PROCESSING

A post-processor runs through the QTLs and finds the genes that are located within the range spanned by the QTL's associated markers, if the QTL has more than one. The QTL–marker associations are drawn from the original data source (as opposed to using all the markers lying within a QTL's range on a linkage group). Therefor, QTLs which have zero or only one associated marker do not have spanned genes.




Soybean QTL Seed oil plus protein 1-5 is flanked by Satt151 and Satt263 leading to 21 spanned genes.




Genes spanned by a QTL via its associated markers are determined in a post-processor and linked on the QTL report page.

## NEXT STEPS

What data and tools would YOU like to see in a genetic+genomic mine?

## References

S. Dash, E. Cannon, S.R. Kalberer, A.D. Farmer, and S.B. Cannon. *Peanuts: Genetics, Processing, and Utilization (AOCS Monograph Series on Oilseeds)*, pages 241–252. Academic Press and AOCS Press, Waltham, MA, 2016. [PeanutBase].

D. Grant, R. T. Nelson, S. B. Cannon, and R. C. Shoemaker. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, 38(Database issue):D843–846, Jan 2010.

A. Kalderimis, R. Lyne, D. Butano, S. Contrino, M. Lyne, J. Heimbach, F. Hu, R. Smith, R. Štěpán, J. Sullivan, and G. Micklem. InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, 42(Web Server issue):W468–472, Jul 2014.

R. Lyne, J. Sullivan, D. Butano, S. Contrino, J. Heimbach, F. Hu, A. Kalderimis, M. Lyne, R. N. Smith, R. Štěpán, R. Balakrishnan, G. Binkley, T. Harris, K. Karra, S. A. Moxon, H. Motenko, S. Neuhauser, L. Ruzicka, M. Cherry, J. Richardson, L. Stein, M. Westerfield, E. Worthey, and G. Micklem. Cross-organism analysis using InterMine. *Genesis*, 53(8):547–560, Aug 2015.

I. Milne, P. Shaw, G. Stephen, M. Bayer, L. Cardle, W. T. Thomas, A. J. Flavell, and D. Marshall. Flapjack–graphical genotype visualization. *Bioinformatics*, 26(24):3133–3134, Dec 2010.

M. Munoz-Amatriain, H. Mirebrahim, P. Xu, S. I. Wanamaker, M. Luo, H. Alhakami, M. Alpert, I. Atokple, B. J. Batieno, O. Boukar, S. Bozdag, N. Cisse, I. Drabo, J. D. Ehlers, A. Farmer, C. Fatokun, Y. Q. Gu, Y. N. Guo, B. L. Huynh, S. A. Jackson, F. Kusi, C. T. Lawley, M. R. Lucas, Y. Ma, M. P. Timko, J. Wu, F. You, P. A. Roberts, S. Lonardi, and T. J. Close. Genome resources for climate-resilient cowpea, an essential crop for food security. *Plant J.*, Oct 2016.

## Acknowledgements

This research was funded by National Science Foundation grant #1444806.