

Integrating Genetic with Genomic Data in Legume InterMines

Sam Hokin, Andrew Farmer

[<shokin@ncgr.org>](mailto:shokin@ncgr.org), [<adf@ncgr.org>](mailto:adf@ncgr.org)



National Center for Genome Resources

Vivek Krishnakumar

[<vkrishna@jcvi.org>](mailto:vkrishna@jcvi.org)



J. Craig Venter
INSTITUTE



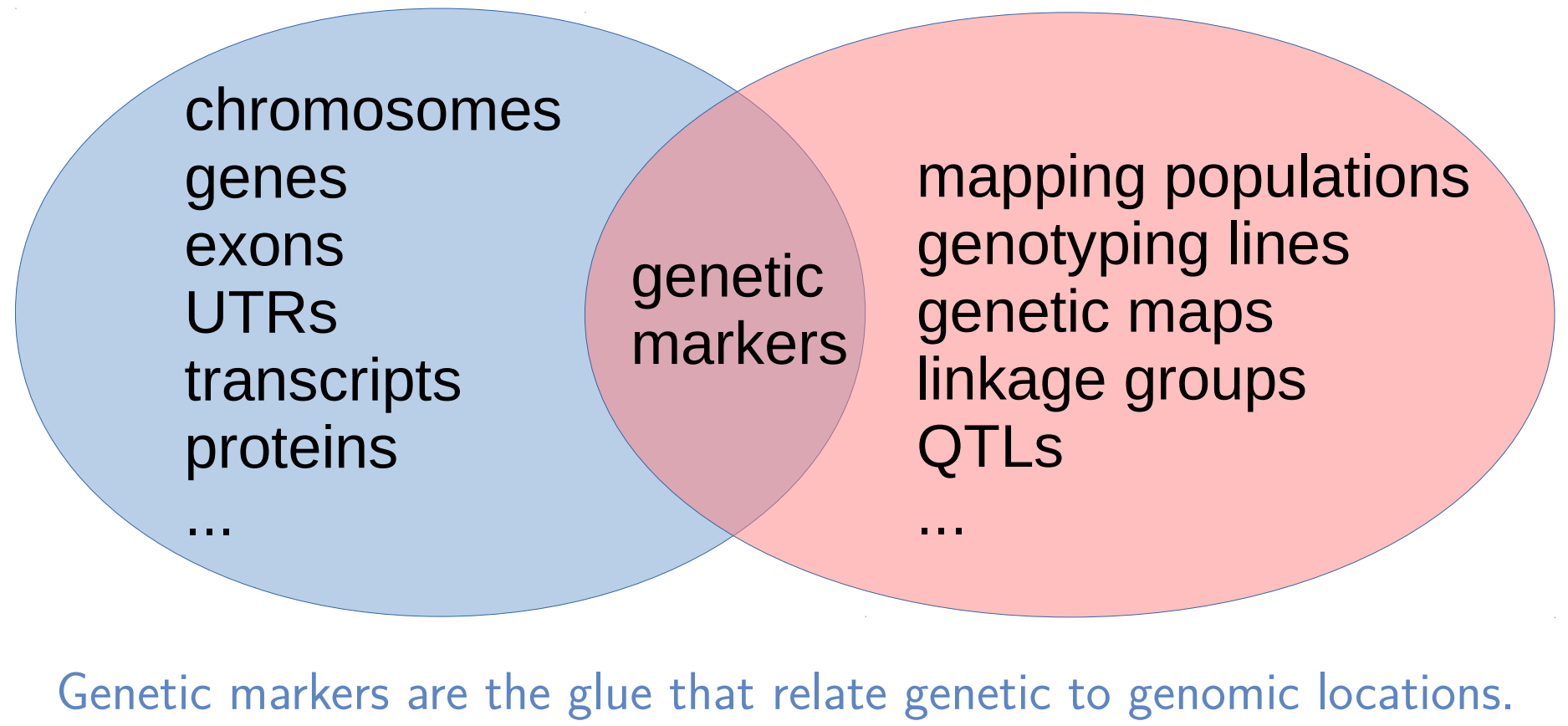


INTRODUCTION

InterMine [<intermine.org>](http://intermine.org) is an open-source, extensible, data warehouse developed by the Micklem Lab at Cambridge with contributions from developers around the world. The core app is written for FlyMine [<flymine.org>](http://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>](http://legumeinfo.org) as well as SoyBase [<soybase.org>](http://soybase.org) and PeanutBase [<peanutbase.org>](http://peanutbase.org). The LIS and PeanutBase sites use the Tripal extension of the Drupal web application, using Chado databases which include genetic data.

The **NSF-funded Legume Federation** [<legumefederation.org>](http://legumefederation.org), in turn, was chartered to provide a “one-stop shop” for legume breeders, geneticists and biologists, federating data sources such as LIS [<legumeinfo.org>](http://legumeinfo.org), MTGD [<medicagogenome.org>](http://medicagogenome.org), and others. For that purpose, **we have extended InterMine to include genetic data** and we have added new visualization and analysis tools.



Genetic data is largely independent of genomic data: genetic features are positioned on linkage groups and measured in centiMorgans (cM), while genomic features are positioned on chromosomes and measured in base pairs (bp). **Genetic markers can have both genetic and genomic locations.** We make use of this to relate genetic features to 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 analyze genetic traits along with genomic features.
3. Enable cross-species analysis amongst legumes and other plants in order to help crop improvement efforts.

DATA MODEL

Genetic data

In InterMine, data entities are represented by a Java **Class**, with attributes and references to other **classes**. We have added the following genetic data classes to the core data model:

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

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

GenotypingLine describes a particular plant line, e.g. a recombinant inbred line, 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.

GenotypeValue is a single value on the markers \times lines genotyping matrix with the value of the parental allele

(often denoted by “A” or “B” or “H”).

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 linkage group 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 researchers.)

Genomic data modifications

Several additions and extensions to the core InterMine data model are needed to accommodate 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**.

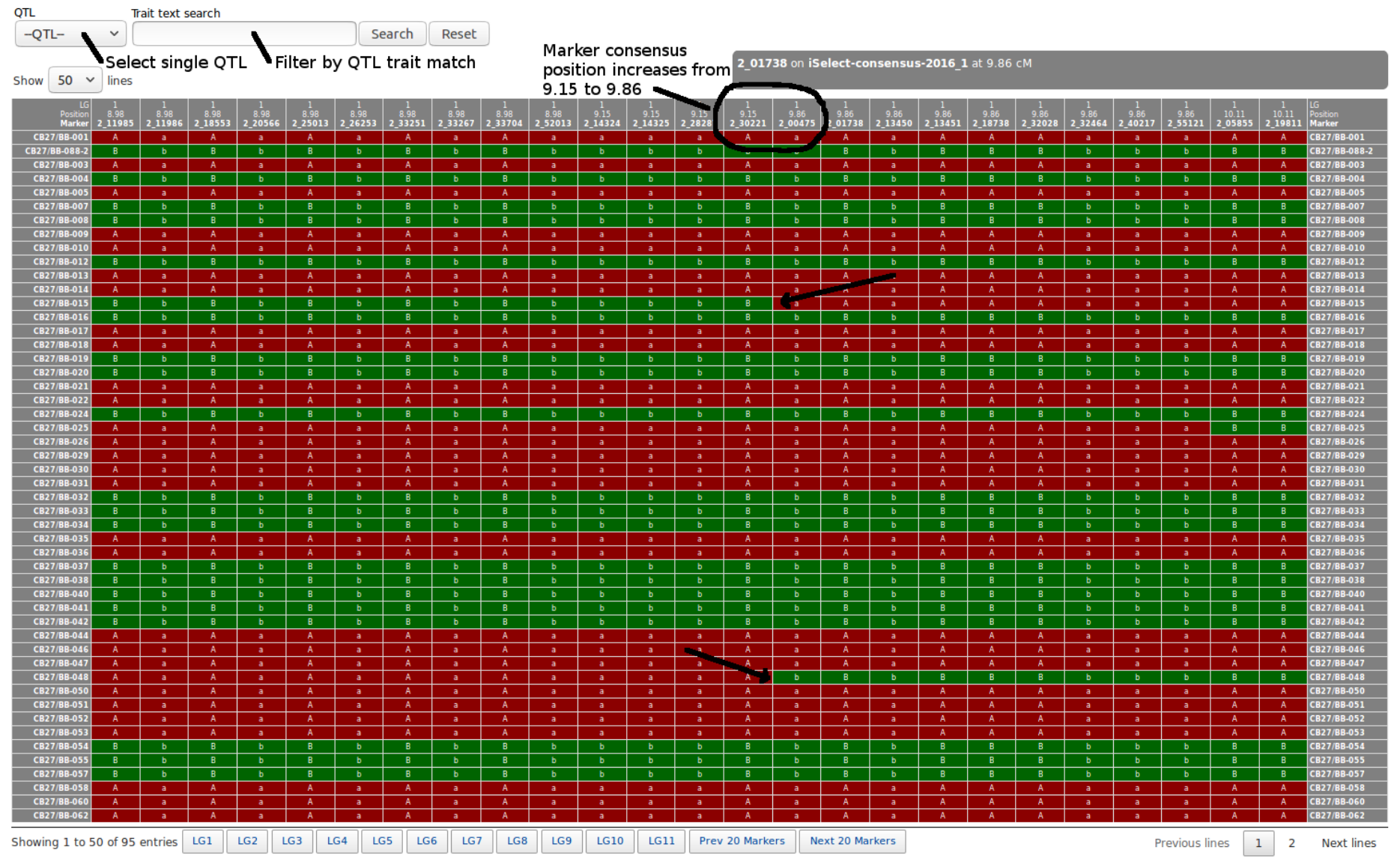
GeneFamily describes a Phytozome gene family, referencing a **consensusRegion**.

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 \times lines* matrix inspired by Flapjack.

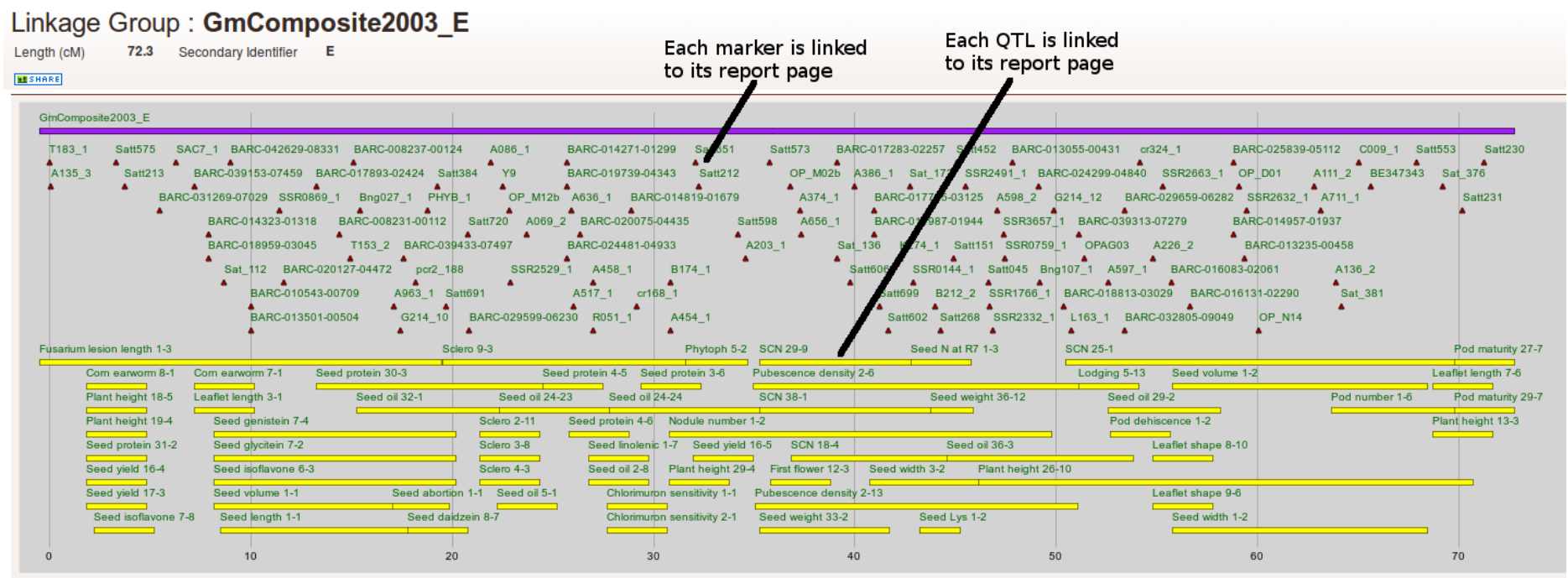


Two allele exchanges in this mapping population indicate recombination, contributing to a 0.71 cM increase in linkage group distance on the consensus map.

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.

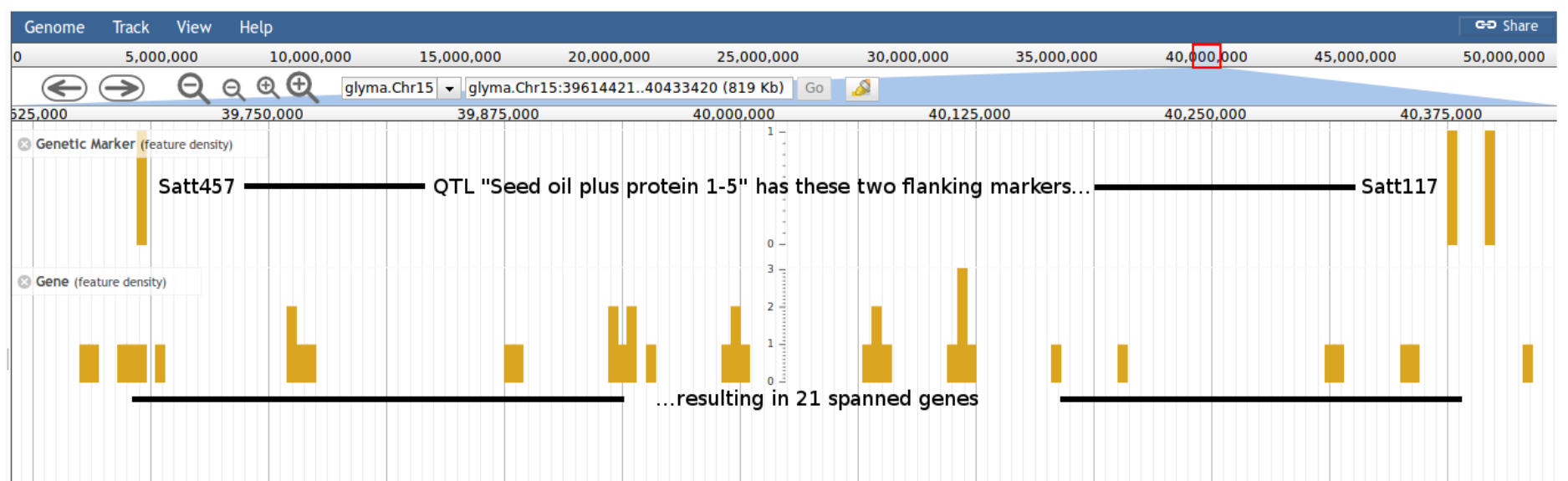


Soybean consensus linkage group GMComposite2003_E has 143 QTLs and 189 markers, 94 of which have been located on chromosome 15.

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

PROCESSING

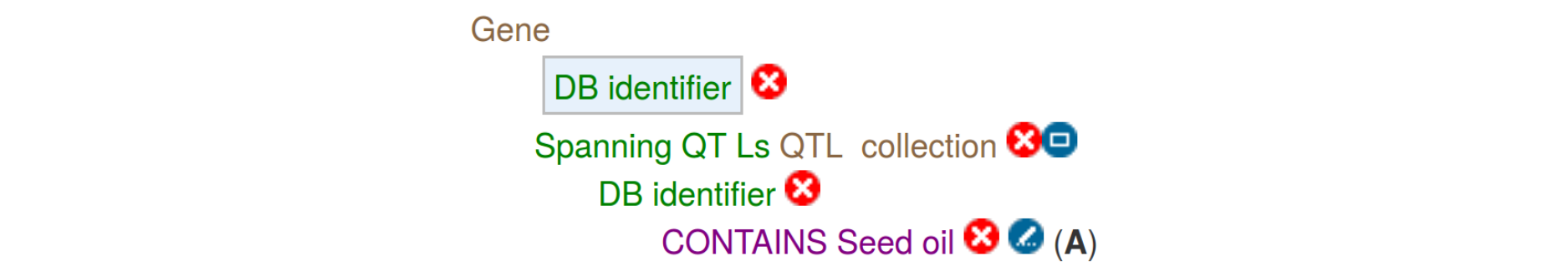
A post-processor runs through the QTLs and finds the genes that are located within the range spanned by the QTL’s associated markers, if the QTL has more than one. The QTL–marker associations are drawn from the original data source (as opposed to using all the markers lying within a QTL’s range on a linkage group). Therefore, 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.

QTL: Seed oil plus protein 1-5 G. max					
Secondary identifier: GmComposite2003_E_Seed oil plus protein 1-5					
2 Associated Genetic Markers					
DB identifier	Type	Length	Chromosome	Start	End
Satt151	microsatellite	289 [105Kb]	glyma.Chr15	39485551	39685559
Satt117	microsatellite	143 [105Kb]	glyma.Chr15	40376691	40376739
1 Linkage Group Ranges					
Begin (cM)	End (cM)	Length (cM)	Linkage Group		
45.09	45.77	0.68	GmComposite2003_E		
21 Spanned Genes					
Organization	DB identifier	Length			
Glycine max	Glycine15G022000	2396 [105Kb]			
Glycine max	Glycine15G022000	8919 [105Kb]			
Glycine max	Glycine15G022000	1546 [105Kb]			
Glycine max	Glycine15G022000	4653 [105Kb]			
Glycine max	Glycine15G022000	2139 [105Kb]			
Glycine max	Glycine15G022000	1152 [105Kb]			
Glycine max	Glycine15G022100	1482 [105Kb]			
Glycine max	Glycine15G022100	1742 [105Kb]			
Glycine max	Glycine15G022100	2329 [105Kb]			
Glycine max	Glycine15G022100	9828 [105Kb]			
Glycine max	Glycine15G022100	371 [105Kb]			
Glycine max	Glycine15G022100	2515 [105Kb]			
Glycine max	Glycine15G022100	1579 [105Kb]			
Glycine max	Glycine15G022200	679 [105Kb]			
Glycine max	Glycine15G022200	5285 [105Kb]			
Glycine max	Glycine15G022100	1111 [105Kb]			
Glycine max	Glycine15G022000	337 [105Kb]			
Glycine max	Glycine15G022100	3844 [105Kb]			
Glycine max	Glycine15G022000	794 [105Kb]			
Glycine max	Glycine15G022200	4941 [105Kb]			
Glycine max	Glycine15G022100	5418 [105Kb]			

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



This SoyMine query returns 7,731 genes spanned by the 29 QTLs containing “Seed oil” that have at least two flanking markers.

NEXT STEPS

The Legume Federation hopes to translate discoveries of the genetic basis underlying key traits made in one species to other species, aiding crop improvement efforts across the family. We will use InterMine to aggregate trait mapping experiments being performed by plant breeders, augmenting the comparative genomics well underway.

What data and tools would YOU like to see in a genetic+genomic mine? Send us your suggestions at [<legumefederation.org/contact>](http://legumefederation.org/contact) or Twitter [@LegumeFed](https://twitter.com/LegumeFed).

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. Westfield, 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. Bozdog, 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.