What’s New in the Legume Information System and the Federated Legume Database Initiative

Jacqueline D. Campbell
Rebuilt from the ground up...

http://legumeinfo.org

A collaborative, community resource to facilitate crop improvement by integrating genetic, genomic, and trait data across legume species. (Cite LIS)

NEWS
- Jan 2016 update: geographic map-based displays of germplasm data, genome context viewer enhancements, LIS in NAR 2016 Database issue
- Dec 2015 update: Red clover genome; speedier genome context viewer; new help & FAQ features; improved gene pages
- Nov 2015 update: Bean markers from many publications have been added; many enhancements to displays of gene-based synteny and phylogeny perspectives

More News
Contributors

Jacqueline D. Campbell
Ethy K. S. Cannon
Wei Huang
Vijay Karingula
Andrew P. Wilkey

Sudhansu Dash
Alan M. Cleary\(^1\)
Alex G. Rice
Pooja E. Umale
Andrew D. Farmer
\(^1\)Montana State University

Steve B. Cannon
Scott R. Kalberer
Nathan T. Weeks
Jugpreet Singh\(^2\)
\(^2\)ORISE Fellow
Gene/Genome Centric

- Gene description
- Gene sequence
- Gene families

Trait Centric

- Traits
- Markers
- Germplasm

Species

Gene models

Genomes

Gene families

QTL

Markers

Traits
But how can LIS ...

... help me with my research or breeding project?
Legume sleuthing ~ LIS style

Case 1: The search for the trait

Case 2: The adventure of the synteny

Case 3: A case of germplasm

“The game is afoot.”
-Sherlock Holmes
The Adventure of the Abbey Grange
• Starting from a trait of interest can we traverse to a genomic location?

• Searching for seed weight trait
Interactive consensus genetic map

- QTL data from multiple studies
- Determined by flanking markers & percentile calculation
- Markers common between published & consensus maps calculate linkage distance
QTL Details

Reference: common bean
PvConsensus_GaleanoFernandez2011_a

Seed weight 3-2 (QTL_seed)

CMap feature details...

View QTL details...

QTL Overview

<table>
<thead>
<tr>
<th>QTL Name</th>
<th>Seed weight-3-2</th>
</tr>
</thead>
</table>

Trait Information

<table>
<thead>
<tr>
<th>Trait Name</th>
<th>Seed weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trait Description</td>
<td>Weight of a seed.</td>
</tr>
<tr>
<td>Trait Class</td>
<td>seed</td>
</tr>
</tbody>
</table>

QTL Information

| Trait Unit | g |
| Treatment | |
| Organism | Phaseolus vulgaris (common bean) |
| Map(s) | PvConsensus_GaleanoFernandez2011_a (linkage group: Pv02); Ceratina_x024404_a (linkage group: B02) |
| Comments | The QTL sw2.2 was also detected at Popayan A, but the Darien QTL was chosen for representation on the genetic map. Seed weights were estimated during all three seasons. |

Publication Information

| Publication | Blair, Iriarte et al., 2008a |
| QTL Symbol | sw2.2 |
| Trait Name | Seed weight |
| Trait Description | 100 Seed weight |
**Clue:** Seed weight QTL with marker data

<table>
<thead>
<tr>
<th>QTL Overview</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trait Name</strong></td>
<td>Seed weight 3-2</td>
</tr>
<tr>
<td><strong>Trait Information</strong></td>
<td></td>
</tr>
<tr>
<td>Trait name</td>
<td>Seed weight</td>
</tr>
<tr>
<td>Trait Description</td>
<td>Weight of a seed.</td>
</tr>
<tr>
<td>Trait Class</td>
<td>seed</td>
</tr>
<tr>
<td><strong>QTL Information</strong></td>
<td></td>
</tr>
<tr>
<td>Trait Unit</td>
<td>g</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
</tr>
<tr>
<td>Organism</td>
<td><em>Phaseolus vulgaris</em> (common bean)</td>
</tr>
<tr>
<td>Map(s)</td>
<td><em>PvConsensus_GaJanoFernandez2011</em> (linkage group: Pv02); <em>Carinza_v_G24404</em> (linkage group: B02)</td>
</tr>
<tr>
<td>Comments</td>
<td>The QTL sw2.2 was also detected at Poyayan A, but the Darien QTL was chosen for representation on the genetic map. Seed weights were estimated during all three seasons.</td>
</tr>
<tr>
<td><strong>Publication Information</strong></td>
<td></td>
</tr>
<tr>
<td>Publication</td>
<td>Blair, Iriarte et al., 2006a</td>
</tr>
<tr>
<td>Publication QTL Symbol</td>
<td>sw2.2</td>
</tr>
<tr>
<td>Publication Trait Name</td>
<td>Seed weight</td>
</tr>
<tr>
<td>Publication Trait Description</td>
<td>100 Seed weight</td>
</tr>
</tbody>
</table>

**Details**

<table>
<thead>
<tr>
<th>Details</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Marker</td>
<td>BM152</td>
</tr>
<tr>
<td>Flanking Marker Low</td>
<td>GAT381</td>
</tr>
<tr>
<td>Flanking Marker High</td>
<td>BM156</td>
</tr>
<tr>
<td>LOD</td>
<td></td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>45.52</td>
</tr>
<tr>
<td>Marker R2</td>
<td>0.17</td>
</tr>
<tr>
<td>Total R2</td>
<td>0.72</td>
</tr>
<tr>
<td>Additivity</td>
<td>10.24</td>
</tr>
</tbody>
</table>
# QTL Seed Weight: Marker Overview

<table>
<thead>
<tr>
<th>Name</th>
<th>BM152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker Type</td>
<td>microsatellite</td>
</tr>
<tr>
<td>Accession</td>
<td>n/a</td>
</tr>
<tr>
<td>Organism</td>
<td><em>Phaseolus vulgaris</em> (common bean)</td>
</tr>
<tr>
<td>Source Description</td>
<td></td>
</tr>
<tr>
<td>Primer1 Name</td>
<td>BM152.p1</td>
</tr>
<tr>
<td>Primer1 Sequence</td>
<td>AAGAGGAGGTGCAAACCTAAATCG</td>
</tr>
<tr>
<td>Primer2 Name</td>
<td>BM152.p2</td>
</tr>
<tr>
<td>Primer2 Sequence</td>
<td>CCGGACTTGGACAGAAGAC</td>
</tr>
<tr>
<td>Genetic Map(s)</td>
<td></td>
</tr>
<tr>
<td>Cerinza_x_G24404_a</td>
<td>linkage group: Cerinza_x_G24404_a-B02</td>
</tr>
<tr>
<td>DOR364_x_BAT477_a</td>
<td>linkage group: DOR364_x_BAT477_a-B02</td>
</tr>
<tr>
<td>Publication(s)</td>
<td>Blair, Galeano et al., 2012a</td>
</tr>
<tr>
<td></td>
<td>Blair, Iriarte et al., 2006a</td>
</tr>
<tr>
<td></td>
<td>Gaitan-Solis, Duque et al., 2002a</td>
</tr>
<tr>
<td>Sequence length</td>
<td>474bp</td>
</tr>
<tr>
<td>Sequence</td>
<td>&gt;BM152</td>
</tr>
<tr>
<td></td>
<td>ACATTTGATGTCGAGATTCTTTTATTATCCATTTACACCATGGTACTGAAATGAGAGTTTTAAGAGGAAATGGTTAAGGAGGACGCTGAAATACTTTAATCGATGTAAGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGTATGAGCAATCACAAACCCACCA</td>
</tr>
<tr>
<td></td>
<td>TGACAGGATTCATCCGGCATGGCAGCAATAATTCCTCCATGTTGGGCTGAGCTGCCGCTGCCCTTCACCAACACCACACT</td>
</tr>
<tr>
<td></td>
<td>CCTCCACGTACCATCCATTCATGCTACCCACCCAGCTTTCAACACACACACCCAGAGGATTCCTCCATCTTACAAATCAGT</td>
</tr>
<tr>
<td></td>
<td>AAGCTTCCATCATGTCAGTTTCTCCATCAGATGCTTCCTCCACAAACATACCCACCACTCCCTCCCTTAGAGT</td>
</tr>
</tbody>
</table>
QTL Seed Weight: Marker Positions
Clue: Three useful markers

Flanking Marker High
BM156

Flanking Marker Low
GATS91

Nearest Marker
BM152

Clue: Seed Weight QTL within <1.1 Mb region
Clues

- QLT for seed weight found in common bean
  - With associated marker data
- Three useful markers
  - Markers mapped to common bean genome
  - Show genomic location of trait

Conclusion

Our search lead us to a seed weight QTL with three markers. These markers were mapped on the genome and indicated the location of the trait.

Case 1: The search for the trait
Legume sleuthing ~ LIS style

Case 1: The search for the trait

Case 2: The adventure of the synteny

Case 3: A case of germplasm

“There is nothing like first-hand evidence.”
-Sherlock Holmes
A Study in Scarlet
Case 2: The adventure of the synteny

Clue: Seed Weight QTL within <1.1 Mb region

Gene Model | Seed weight symbol | Chrom | Start | End
--- | --- | --- | --- | ---
Phvul.001G000500 | CDLB1 | Chr01 | 144,309 | 146,169
Phvul.001G003700 | EXPO10 | Chr01 | 341,806 | 344,139
Phvul.002G029500 | DDM1 | Chr02 | 3,063,228 | 3,069,240
Phvul.002G029700 | DWF4 | Chr02 | 3,098,498 | 3,103,000

Supplementary Table 19. Candidate common bean seed weight genes.
Genome Browser to Gene page

**Description**

- **Gene Model Name**: phavu.Phvul.002G029700
- **Organism**: Phaseolus vulgaris (common bean)
- **Description**: Cytochrome P450 superfamily protein; IPR001128 (Cytochrome P450); GO:0005606 (iron ion binding), GO:0020037 (heme binding), GO:0055114 (oxidation-reduction process)
- **Gene Family**: phytozome_10_2.59227799
- **mRNA and protein identifiers**
  - phavu.Phvul.002G029700.1
  - phavu.Phvul.002G029700.2

**Actions**

- More at LGS gene search page
- View assembly sequence
- View assembly sequence with 1kb flanking
Gene page to Gene Family

<table>
<thead>
<tr>
<th>Description</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene Model Name</strong></td>
<td>phavu.Phvul.002G029700</td>
</tr>
<tr>
<td><strong>Organism</strong></td>
<td>Phaseolus vulgaris (common bean)</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Cytochrome P450 superfamily protein; IPR001128 (Cytochrome P450); GO:0005506 (iron ion binding), GO:0020037 (iron ion binding), GO:0016593 (oxidation-reduction process)</td>
</tr>
<tr>
<td><strong>Gene Family</strong></td>
<td>phytozome_10_2.59227799</td>
</tr>
<tr>
<td>mRNA and protein identifiers (also see Sequences tab)</td>
<td>phavu.Phvul.002G029700.1, phavu.Phvul.002G029700.2</td>
</tr>
</tbody>
</table>
- Gene families:
  - Legume focused
  - Based on (and linked to) Phytozome families
Clue: 2\textsuperscript{nd} seed weight candidate gene in same gene family

Supplementary Table 19. Candidate common bean seed weight genes.
Links to more clues
Genome Context Viewer (GCV)

Displays genomic contexts of orthologs, in terms of gene families

- Hover gene to see name, highlight in dotplots
- Click gene for links to more info
- Hover track for all gene names
- Click track to list all genes and track options
Links to more clues: Phytozome

Phylogram

phytozome_10_2.59227799: Cytochrome P450 superfamily protein IPR001128 (Cytochrome P450) GO:0005506 (iron ion binding), GO:0020037 (heme binding), GO:0055114 (oxidation-reduction process) ***. AT3G50660.1

[Diagram of Phytozome phylogram]

- view organism: Phaseolus vulgaris (common bean)
- view feature: phavu.Phvul.002G029700.1
- Find similar genomic contexts for: Phvul.002G029700
- view at Phytozome: Phvul.002G029700
- view at LIS P1.0 GBrowse: Phvul.002G029700
Gene: Phvul.002G029700

**Gene Info**

- **Organism**: Phaseolus vulgaris
- **Locus Name**: Phvul.002G029700
- **Transcript Name**: Phvul.002G029700.1 (primary)
- **Other transcripts**: Phvul.002G029700.2
- **Location**: Chr02:3098881..3103000 forward
- **Description**: (M=3) K09587 - cytochrome P450, family 90, subfamily B, polypeptide 1 (steroid 22-alpha-hydroxylase)

**Protein domain view**

Clue: More information about the gene
Links to more clues: SoyBase
Clue: Highly expressed in pod and seed tissues
Links to more clues: PeanutBase
Clue: Highly expressed in seed tissues
Case 2: The adventure of the synteny

Clues

• Legume focused Phytozome gene family
  – Has two seed weight candidate genes

• Data in LIS family tree linked to:
  – Phytozome, SoyBase & PeanutBase

• Similar genes in *G. max* and *A. duranensis*
  – Differentially expressed in seed tissue

Conclusion

Our adventure took us through the synteny between legume species. From the LIS gene family tree we visited three other genomic databases.
Case 1: The search for the trait

Case 2: The adventure of the synteny

Case 3: A case of germplasm

“What one man can invent another can discover.”
-Sherlock Holmes
The Adventure of the Dancing Man
Legume Germplasm Resources at LIS

GRIN Connect:

Search for germplasm at USDA GRIN (NPGS) database via LIS:
Use this utility if you want to search the entire USDA GRIN germplasm database from within LIS. You will be able to see and search the GRIN website within a LIS page with a GRIN accession number. You can also check, if it exists, the evaluation data for the germplasm accession from the 'Detailed Accession Observation Page' link in the GRIN webpage.

Trait Evaluation Data:

Search for germplasm evaluation data at LIS:
Germplasm evaluation data (for common crop traits or descriptors) for the following legumes are available via LIS from GRIN. The evaluation data sets have been downloaded from the GRIN website and are served here by LIS).

Geographic Distribution of Legume Germplasm:

Locate the the germplasm accessions in a map.
The legume germplasm accessions can be located on a map using a Geographical Information System (GIS) utility. You can search the distribution of germplasm accessions in a given geographic area, can map multiple species or restrict it to a single legume taxon, look at the accession details and evaluation data if available, and overlay trait evaluation data on the map.
Geographical Information System

[Map showing legume species distribution across different locations with a table listing accession numbers, taxon, crop name, orig country, collection site, acq dates, and elev (m)].
GIS Search: Seed Weight
Clue: location collected and seed weight

<table>
<thead>
<tr>
<th>accession</th>
<th>SEEDWGT</th>
<th>taxon</th>
<th>crop name</th>
<th>orig country</th>
<th>collection site</th>
<th>acq date</th>
<th>elev (m.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI 313307</td>
<td>32</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Puerto Nuevo</td>
<td>1966-04-19</td>
<td></td>
</tr>
<tr>
<td>PI 255309</td>
<td>25</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Purchased in Bacanora.</td>
<td>1959-02-02</td>
<td>300</td>
</tr>
<tr>
<td>PI 309811</td>
<td>22</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Sierra de la Ventora</td>
<td>1965-12-09</td>
<td>1,290</td>
</tr>
<tr>
<td>PI 201494</td>
<td>68</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Market in Villahermosa</td>
<td>1962-05-01</td>
<td>60</td>
</tr>
<tr>
<td>PI 201495</td>
<td>20.2</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Market in Villahermosa</td>
<td>1952-05-01</td>
<td>60</td>
</tr>
<tr>
<td>PI 201496</td>
<td>22.6</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Market in Villahermosa</td>
<td>1952-05-01</td>
<td>60</td>
</tr>
<tr>
<td>PI 440797</td>
<td>40</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>MEX</td>
<td>Papago Indian Village, north</td>
<td>1978-03-01</td>
<td></td>
</tr>
<tr>
<td>PI 529598</td>
<td>15.24, 18</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>USA</td>
<td>Palo Alto, Santa Clara Coun...</td>
<td>1988-03-20</td>
<td></td>
</tr>
<tr>
<td>PI 535242</td>
<td>58</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>USA</td>
<td>Palo Alto, Santa Clara Coun...</td>
<td>1989-02-13</td>
<td></td>
</tr>
<tr>
<td>PI 533382</td>
<td>14</td>
<td>Phaseolus vulgaris</td>
<td>bean</td>
<td>USA</td>
<td>Palo Alto, Santa Clara Coun...</td>
<td>1989-02-13</td>
<td></td>
</tr>
</tbody>
</table>
Clue: location collected and photo response
Case 3: The case of germplasm

Clues

• Multiple ways to search germplasm data
  – GRIN Connect, Trait Evaluation & GIS

• GIS: Graphical Information System
  – Visual tool to not only locate germplasm collection points but also label associated traits

Conclusion

This case lead us to find interesting germplasm trait data.
The case continues ...

UNSOLVED
1. Data Management Requirements

“[i]nvestigators are expected to share with other researchers, at no more than incremental cost & within a reasonable time, the primary data, samples, physical collection & supporting materials created or gathered in the course of work under NSF grants.”


“… each Federal agency with over $100M in annual conduct of research & development expenditures to develop a plan to support increased public access to results of research funded by the Federal Government.”

https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
Legume Federation

The Federated Plant Database Initiative for the Legumes.

David Fernandez-Baca (PI)
Ethalinda K. S. Cannon

Andrew D. Farmer

USDA Steven B. Cannon

CYBERSE Jeremy DeBarry

J. Craig Venter Institute Chris Town

Sam Hokin, Vivek Krishnakumar, Nathan Weeks, Jacqueline D. Campbell and Akshay Yadov

Others in ‘Federation projects’: Tripal, Phytozome, SoyBase, KnowPulse, CSFL, MedicagoGenome, USAID chickpea project and Noble Foundation alfalfa project.
Challenges
1. It is important to have specialists for each crop.
   – Genetic characteristics differ
   – Each crop has its own breeding objectives
   – Population and germplasm characteristics
Challenges

1. It is important to have specialists for each crop.
2. Utilize knowledge gained about many legumes.
   - Nodulation, flowering, determinacy, disease resistance, seed patterning, seed development
Challenges

1. It is important to have specialists for each crop.
2. Utilize knowledge gained about many legumes.
3. Genomic data is growing rapidly.

- Common bean (*Phaseolus vulgaris*)
- Pigeon pea (*Cajanus cajan*)
- Chickpea (*Cicer arietinum*)
- Barrel medic / Medicago (*Medicago truncatula*)
- Lotus (*Lotus japonicus*)
- Soybean (*Glycine max*)
- Mungbean (*Vigna radiata*)
- Wild peanut (*Arachis duranensis* and *Arachis ipaensis*)

Red clover (*Trifolium pratense*)
Genome published Nov 2015
1. It is important to have specialists for each crop.
2. Utilize knowledge gained about many legumes.
3. Genomic data is growing rapidly.
4. Many overlapping genomic online databases.
Challenges

1. It is important to have specialists for each crop.
2. Utilize knowledge gained about many legumes.
3. Genomic data is growing rapidly.
4. Many overlapping genomic online databases.
5. Software/tool development is hard.
Legume Federation response
Legume Federation response

- Get easier to use genomic visualization and analysis tools to the researchers.
Legume Federation response

• Get easier to use genomic visualization and analysis tools to the researchers.

• Encourage/help the development of the software.
Legume Federation response

• Get easier to use genomic visualization and analysis tools to the researchers.

• Encourage/help the development of the software.

• Build methods that enable sharing among genomic online databases.
Legume Federation response

• Get easier to use genomic visualization and analysis tools to the researchers.
• Encourage/help the development of the software.
• Build methods that enable sharing among genomic online databases.

• Encourage use of common formats and databases.
Legume Federation tools

CMap
A browser-based tool for the visual comparison of various genetic maps. **Legume Fed will rebuild for greater interactivity.**

CHADO
A standardized database schema for storing biological data.

GBrowse
A set of customizable Drupal modules for constructing biological websites.

Tripal

InterMine
A web-accessible data warehouse for the integration and analysis of complex biological data.
CyVerse (iPlant) tools, methods & storage

- Data Store for major data sets
- Metadata methods (iRODS)
- Integration w/ CoGe & other tools
Shareable & Discoverable

CyVerse (iPlant) tools, methods & storage
Data templates for:
• Maps, Markers, Publications, QTLs, Traits

• Data templates
  ◦ The LegumeInfo and PeanutBase projects are using the following data templates for collecting genetic and QTL data:
  ◦ Download the Excel spreadsheet template
  ◦ Download an example spreadsheet
  ◦ This is currently a topic of very active discussion, as the members of the federation hash out details to accommodate complex genetic data. Please contact us to get involved, or if you would like to submit data for any legume species using the forms above.

“[i]nvestigators are expected to share with other researchers, at no more than incremental cost & within a reasonable time, the primary data, samples, physical collection & supporting materials created or gathered in the course of work under NSF grants.”

Make your data sharable, searchable & discoverable

“Nothing clears up a case so much as stating it to another person.”
-Sherlock Holmes
Silver Blaze

LIS  CYVERSE  SoyBase
PeanutBase
Thank you

Contact us:
Steven Cannon  steven.cannon@ars.usda.gov
http://legumeinfo.org/contact
http://legumefederation.org/contact
Google Group: LegFedCommunity
Twitter: @LegumeFed