Published online 5 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D1181–D1188***
doi: 10.1093/nar/gkv1159

Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family

Sudhansu Dash[1,†], Jacqueline D. Campbell[2,†], Ethalinda K.S. Cannon[3,†], Alan M. Cleary[1,4], Wei Huang[2], Scott R. Kalberer[5], Vijay Karingula[2], Alex G. Rice[1], Jugpreet Singh[6], Pooja E. Umale[1], Nathan T. Weeks[5], Andrew P. Wilkey[2], Andrew D. Farmer[1,*] and Steven B. Cannon[2,5,*]

[1]National Center for Genome Resources, Santa Fe, NM 87505, USA, [2]Dept. of Agronomy, Iowa State University, Ames, IA 50011, USA, [3]Dept. of Computer Science, Iowa State University, Ames, IA 50011, USA, [4]Dept. of Computer Science, Montana State University, Bozeman, MT 59715, USA, [5]USDA-ARS Corn Insects and Crop Genetics Research Unit, Crop Genome Informatics Lab, Iowa State University, Ames, IA 50011, USA and [6]ORISE Fellow, USDA-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

Received August 14, 2015; Revised October 16, 2015; Accepted October 19, 2015

ABSTRACT

Legume Information System (LIS), at http://legumeinfo.org, is a genomic data portal (GDP) for the legume family. LIS provides access to genetic and genomic information for major crop and model legumes. With more than two-dozen domesticated legume species, there are numerous specialists working on particular species, and also numerous GDPs for these species. LIS has been redesigned in the last three years both to better integrate data sets across the crop and model legumes, and to better accommodate specialized GDPs that serve particular legume species. To integrate data sets, LIS provides genome and map viewers, holds synteny mappings among all sequenced legume species and provides a set of gene families to allow traversal among orthologous and paralogous sequences across the legumes. To better accommodate other specialized GDPs, LIS uses open-source GMOD components where possible, and advocates use of common data templates, formats, schemas and interfaces so that data collected by one legume research community are accessible across all legume GDPs, through similar interfaces and using common APIs. This federated model for the legumes is managed as part of the 'Legume Federation' project (accessible via http://legumefederation.org), which can be thought of as an umbrella project encompassing LIS and other legume GDPs.

INTRODUCTION

With nearly 20 000 species, approximately two dozen of which are domesticated, legumes are enormously important in natural and agro-ecosystems. Legumes play key roles in both animal and human nutrition, providing about a third of human dietary nitrogen (1,2). Seeds from grain legume species are sources of oil, protein and starch, for both animal and human consumption, and legumes play an important role as animal forages. The legumes also claim two important model organisms, *Medicago truncatula* and *Lotus japonicus*, which have served to elucidate the genetic basis of the legume-rhizobial symbiosis that serves to fix atmospheric nitrogen for use by the plant.

Much has changed since the inception of LIS about 15 years ago (3) around the time of the initial publication of the *Arabidopsis thaliana* genome (4). Today, a genome sequence is a routine starting point for a species research community. Decreased costs have enabled the sequencing of the genomes of more than 10 legume species (at the time of writing), along with many transcriptomes and other genomic resources. While this offers exciting opportunities for legume crop improvement and biological studies, the rapid increase of sequence data also poses challenges for data storage, integration and utilization. Legume Information System (LIS: http://legumeinfo.org) is a web portal supporting research in legume genomics, genetics and breeding. The data and tools at LIS help address questions both in

*To whom correspondence should be addressed. Tel: +1 515 294 6971; Fax: +1 515 294 9359; Email: steven.cannon@ars.usda.gov
Correspondence may also be addressed to Andrew D. Farmer. Tel: +1 505 995 4464; Fax: +1 505 995 4461; Email: adf@ncgr.org
†These authors contributed equally to the paper as first authors.

and across legume species and for plants in general. Currently, LIS has whole-genome and related data for nine important legume species (common bean, mungbean, pigeonpea, chickpea, *Lotus japonicus*, *Medicago truncatula*, soybean (in cooperation with SoyBase http://soybase.org) and two wild peanut species (in cooperation with PeanutBase, http://peanutbase.org (5)). The goal is to provide a per-species reference resource for genomic and genetic research in legumes, and an integrated resource for accessing, visualizing and comparing genetic and molecular data across multiple legume species.

Development of LIS is a collaborative effort between the National Center for Genome Resources (NCGR) and the USDA Agricultural Research Service (ARS). The website was originally available at http://comparative-legumes.org (3,6), but was moved in 2013 to http://legumeinfo.org. Along with the change in domain name, the new genomic data portal (GDP) is implemented with Tripal (7,8), which consists of a set of Drupal modules for developing GDPs, and Chado (9) a database schema for biological data. LIS uses additional Generic Model Organism Database (GMOD) and other open-source software where possible, including CMap (10), GBrowse (11) and JBrowse (12).

The decision in 2013 to adapt the system of tools built around the Chado framework such as Tripal was strategic. It enables LIS to take advantage of a broad community of developers, and to more easily exchange information with other Chado-based GDPs. With the amount of genomic data increasing so rapidly, across so many disparate species, it is important to take advantage of the specialized knowledge within various legume research communities, and also the distributed development efforts of various GDP projects.

As of late 2015, there are at least five legume GDPs that use Tripal and Chado: KnowPulse (http://knowpulse.usask.ca), CoolSeasonFoodLegume (https://www.coolseasonfoodlegume.org, 13), PeanutBase (http://peanutbase.org), Alfalfa-Genome (http://alfalfa-genome.org) and MedicagoGenome (http://medicago.jcvi.org/MTGD/, 14). Additionally, there are several other legume GDPs that use various combinations of custom interface code and GMOD components. In order to help reduce the amount of redundant development and to enable more seamless access to data, a NSF project was initiated in 2015. 'The Federated Plant Database Initiative for the Legumes' (or 'Legume Federation'), seeks to 'better integrate legume databases through data standards, distributed development and comparative analysis' (Legume Federation project, http://legumefederation.org).

The Legume Federation project, though distinct from LIS, helps coordinate development on several important framework components for LIS and other legume GDPs, including: (i) standards for data formats, metadata and web service protocols and ontology use; (ii) an open repository for data exchange (in development using iPlant's DataStore (15)); (iii) new modules for Tripal and other GMOD components (see discussion of the BLAST user interface and Phylotree modules below) and (iv) configurations and programmatic interfaces in packages such as JBrowse and InterMine (16) to enable seamless navigation and data access across sites.

LIS development also follows the spirit of the Legume Federation philosophy in that development of some legume content is intentionally left to other GDPs, but is accessible seamlessly from various points within LIS. This is true, for example, with soybean (data are maintained at SoyBase (6,17)), peanut (data are maintained at PeanutBase), Medicago truncatula (data are maintained both within LIS and at MedicagoGenome and MedicagoHapmap).

Below, we describe key data and features of LIS, including: synteny mappings among genomes, phylogenetic trees to enable traversal among species, a genomic context viewer to find and explore regions with local conservation of gene content and order, genetic maps and mapped traits (quantitative trait loci; QTL).

DATA AND TOOLS IN LIS

Species

A summary of the data and tools available for each species can be found on the species overview page, accessible via the 'Species' menu. LIS currently maintains data for 15 species: alfalfa (Medicago sativa), chickpea (Cicer arietinum), common bean (Phaseolus vulgaris), cowpea (Vigna unguiculata), garden pea (Pisum sativum), lentil (Lens culinaris), mungbean (Vigna radiata), narrow-leafed-lupin (Lupinus angustifolius), pigeonpea (Cajanus cajan), white clover (Trifolium repens), the two model species Lotus japonicus and Medicago truncatula. Data for the tetraploid peanut species Arachis hypogaea (cultivated peanut), and two wild diploid progenitors of cultivated peanut, A. ipaensis and A. duranensis, are primarily maintained at and linked to PeanutBase. Similarly, data for soybean (Glycine max) are primarily maintained and linked to at SoyBase (Supplemental Table S1).

The amount and type of data varies for these species. Those with genome sequences (and browsers) include common bean (Pv1.0; (18)), chickpea (Ca1.0; (19)), pigeonpea (Cc1.0; (20)), soybean (Gmax2.0; (21)), *Medicago truncatula* (Mt4.0; (22)), *Lotus japonicus* (Lj2.5; (23)), *Arachis ipaensis* and *A. duranensis* (Araip1.0 & Aradu1.0; (5)). There are genetic maps for most species listed above. QTL data are collected at LIS for common bean, and extensive QTL data are maintained for soybean at SoyBase and for peanut at PeanutBase.

Navigation

The tools and data sets at LIS are available from two starting points: from a species, or from a content/data type. For example, one can start browsing the genomes or maps from the species page or go to the genome browsers page and select a species there. This organizational pattern has been maintained for genomes, traits and maps, keyword search, QTL, genes, BLAT, BLAST, gene family and domains.

All features and data are available via the top menu tabs and sub-navigation fields. The most significant features and data sets are also available via a set of buttons on the front page. These include access to genes, gene families, several search functions and the genome browsers. In addition,

many of the data types are interlinked, so that a search for protein domains can easily lead the user to the set of genes annotated with that domain or to the set of gene families in which the domain is significantly conserved among the members.

Genome browsers

For legumes with sequenced genomes, LIS provides both GBrowse and JBrowse genome browser (see Species section above). The soybean genome is hosted at SoyBase. The genomes of the two wild progenitors of peanut, *A. duranensis* and *A. ipaensis*, are hosted at PeanutBase. Direct links to the genome browsers at the LIS sister sites are available at multiple locations within LIS. There are several particular strengths of the browsers at LIS, PeanutBase and SoyBase, including: (i) extensive synteny mappings among these species (see discussion below); (ii) informative gene annotations, with links from gene models to other resources such as Phytozome (24); (iii) genetic markers where available; (iv) identification of genomic landmarks such as pericentromeric and centromeric boundaries, based on transposon and gene densities and (v) mappings of features (such as gene models and transcriptome sets) from other selected relevant species. Additionally, the LIS and affiliated browsers are configured for speed at all viewing scales. This has been accomplished by serving the data from a 48-core server, and (in GBrowse) by managing all track data as separate SQLite data sets, so each can be rendered separately.

Traits, maps and markers

LIS provides genetic linkage maps from 12 legume species, using the comparative map viewer, CMap (Figure 1). Each of the maps can also be downloaded as a file from the map overview page and the download page.

QTL data are accessible through the 'QTL' quick link on the LIS home page, via the Traits and Maps page and the search menu. QTL can be searched by species, trait class (e.g. development, organic, root, yield), and either by its published QTL symbol or unique identifier, assigned by LIS curators. Each record in the table of search results can be examined in detail, including information about the trait, experimental conditions, QTL statistics and measurements, and associated markers (Figure 1C). QTL are placed on two maps, the map published by the paper, and projected onto a consensus map, which permits comparing QTL from all papers for a species. The consensus map can be viewed in CMap. The projection is performed by LIS curators, using flanking markers and a percentile calculation (Figure 1A). If flanking markers are not given in the publication, markers that flank the associated marker are used. Markers that are common between the published and consensus maps are used to calculate the linkage distance and the QTL is projected onto the consensus map accordingly. Currently LegumeInfo contains QTL data for common bean and crosslinks to peanut and soybean QTL in PeanutBase and SoyBase, respectively.

Detailed information about each map, marker, trait and QTL, along with information about the experimental conditions and procedures, mapping populations and parental germplasm is curated from publications. Information is entered into a standardized spreadsheet, then verified and loaded into the Chado tables via Perl scripts. Because the curation process is slow and resource intensive, the LIS project has focused on developing and sharing data-collection templates, in order to make use of community participation by authors of new QTL and mapping studies. A simplified version of this spreadsheet is available for researchers who wish to submit their own data sets. These templates are available via the 'Submit Data' tab from the main navigation page.

As terminology differs among publications and research communities, making cross-comparisons difficult, data curation from publications requires human selection and interpretation. In addition to collecting the information indicated by the spreadsheet template, LIS curators establish standard names, terms and definitions for traits and maps. These are associated with the names used in the publication and with standard ontologies like the Gene Ontology (GO), Plant Ontology (PO), Trait Ontology (TO) and species-specific ontologies like the soybean ontology (SOY), where possible. The ontology terms in use at LIS and allied sites are periodically updated and are available for download from the QTL download page (http://legumeinfo.org/QTL_download).

This genetic data collection spreadsheet template is currently being modified by representatives from multiple species and GDPs with the objective of providing a standard for collecting and sharing trait and genetic data. Current versions of the QTL and marker template spreadsheet are available for download at LIS at http://legumeinfo.org/QTL_download. Updated versions should be available as well at all Legume Federation member sites by mid-2016.

Synteny

Synteny is one of two primary methods that LIS offers for traversing between corresponding genomic features in different legume species (the other being phylogenetic trees). The genome browser has tracks for showing syntenic regions with other legumes. Tracks are turned on via the Select Tracks at the top of the GBbrowse utility. Clicking on any synteny feature gives information about that region (median Ks value and coordinates), and enables navigation to the corresponding syntenic feature in the other genome. Coordinates for all synteny features can be download using the GBrowse download icon for any trackset. Other synteny data (paralog pairs, alignments, Ks values) are available by request.

Synteny mappings have been made among the cool-season species (*Medicago*, *Lotus*, chickpea), among the warm-season species (pigeonpea, common bean, soybean) and between each species and the major models (*Medicago* and *Lotus*). Synteny calculations involves identification of orthologous genes by selecting reciprocal top hits between chromosome pairs from an all-against-all blastp search of genes between two species (evalue threshold of 1e-10), followed by DAGchainer (25) to predict chains of syntenic genes in complete genomes. Lastly, proportions of synonymous-site changes (Ks) between orthologs are calculated, and synteny blocks are filtered for block-wise median

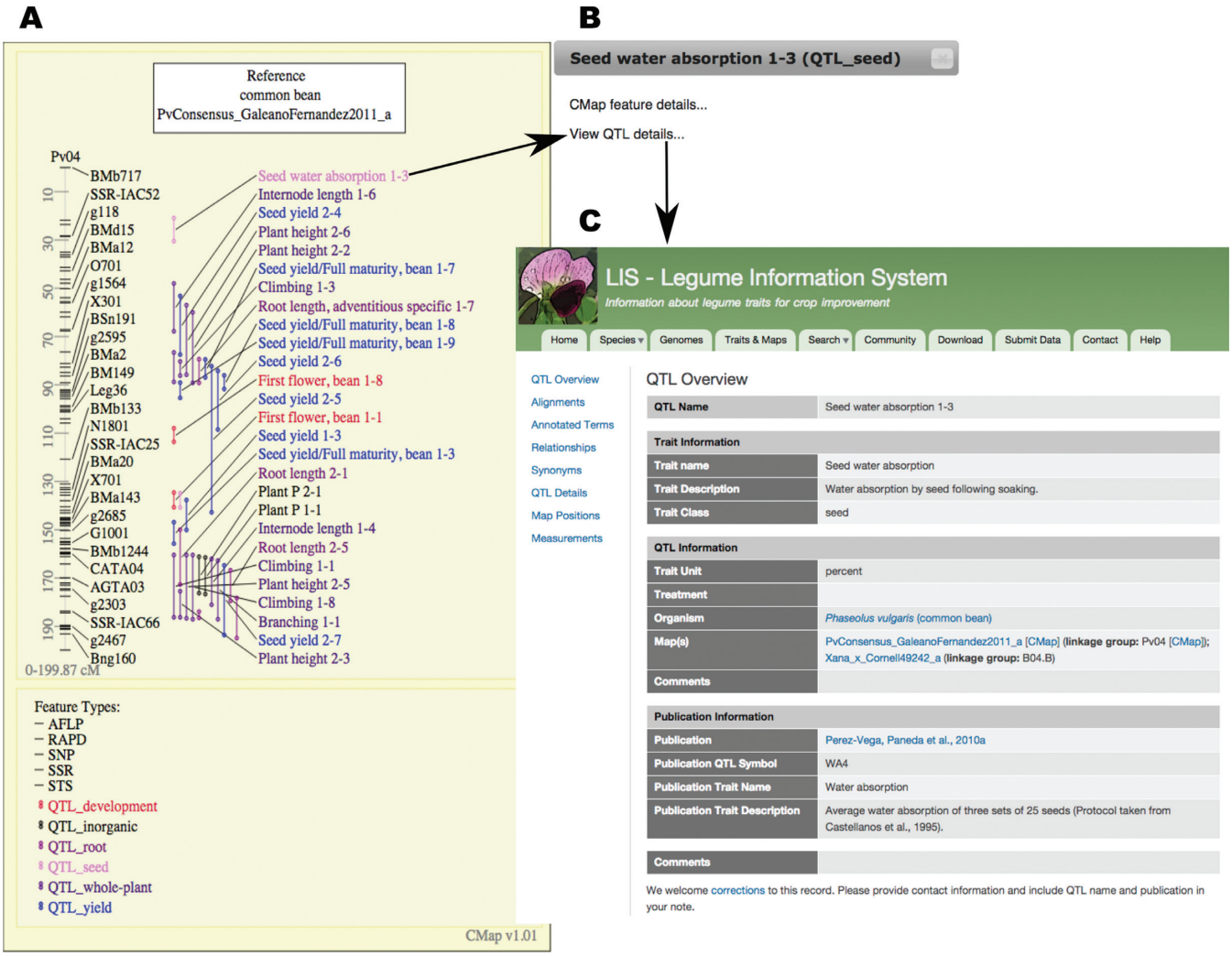

Figure 1. Screenshot of an interactive CMap. **(A)** The QTL from different experiments are consolidated into a single consensus view. **(B)** Clicking on the QTL of interest on the CMap opens a menu with links to more information. **(C)** After selecting 'View QTL details…' the user is redirected to the Tripal page about that particular QTL.

Ks values. This filtering step is needed because of whole-genome duplications (WGD) in the legume family. A WGD occurred just prior to the radiation of the papilionoid subfamily in the legumes, affecting all species included at LIS; and another WGD occurred in the Glycine lineage at ≈10 Mya. For the genome browsers for soybean (at SoyBase) and *Medicago truncatula*, synteny blocks have also been calculated to show these paralogous WGD-derived duplications.

Gene families and the Phylotree module

Gene family and phylogenetic tree viewers at LIS provide evolutionary context and (where available) functional information about groups of genes. An interactive tree viewer shows genes from eight sequenced legume species. At the time of writing, gene families at LIS are based on the Phytozome v10.2 angiosperm-level gene families, but are enriched for legume species. Legume species included in the LIS gene families are soybean, common bean, Medicago truncatula, pigeonpea, chickpea and two wild diploid Arachis species. Non-legume species (included as phylogenetically- or functionally-informative outgroups) are peach (in the fabid clade, along with legumes), Arabidopsis, grape, tomato (in the asterid clade), rice and maize (monocots) and Amborella (the most distant outgroup).

Basing the LIS gene families on the Phytozome angiosperm families has the advantage of enabling linkages and integration with tools and resources at Phytozome. At the same time, there would be advantages in more tightly-circumscribed families, containing only legume sequences and near outgroup species. These will be available in 2016, along with the current Phytozome-based angiosperm superfamilies, and mappings between the legume- and angiosperm-level families.

Gene families can be searched with the Phytozome gene family ID, one or more words found in the family description, GO terms and InterPro (26) gene descriptions. Results can be restricted by family size or species composition. In the results table, the family name links to a phylogram display of the family. The membership counts under

the combined and separate species headings link to a list of the member genes. The families are also available through a REST Web service that returns a family ID when provided with a legume gene ID.

In the interactive phylogram display, clicking a legume gene node in the tree brings up a menu leading to detailed information about the gene, provided by LIS and other GDPs (Figure 2C). The internal (white) nodes link to the Genomic Context viewer (described below) for the genes in the node's subtree (Figure 2B). Clicking the root node of the tree provides the option of either viewing the entire tree in the Genomic Context Viewer (Figure 2), or exploring multiple sequence alignments in various forms, including in the interactive Jalview alignment viewer (27). Further details on analysis methodologies are available using the Analysis link provided for each tree, using a standard Chado/Tripal approach to providing provenance information.

Gene family descriptions are derived from homology-based functional analysis of consensus sequences from each family (AHRD: https://github.com/groupschoof/AHRD), and include InterPro and GO identifiers which may be included in the search. The comparative genomics data have been further enriched by protein domain annotations that are derived from InterProScan (28) analysis of protein sequences against InterPro protein signature databases. The code for the phylogenetic tree viewer and associated interface components are available as a Tripal module at github.com/ncgr/tripal_phylogeny.

Genomic context viewer

The Genomic Context Viewer displays corresponding regions around a selected gene or set of genes in a subtree. By default, the regions extend out by 10 genes upstream and down from the selected genes. Each sequence of genes is called a 'track' and there is one track for each species or chromosome in the set of genes in the subtree. Each gene in a Context View is colored by the gene family it belongs to as indicated in the legend. In the case where a single gene is used as the selection, that gene's context track will be used as a query to find other tracks annotated with similar gene family content in the database (Figure 2E). Synteny between these tracks is determined via a modified Smith-Waterman or Repeat alignment algorithm (29). The modification makes it possible to correctly align inversions and segmental tandem duplications, events which occur frequently at the scale of multi-genic segments in plant genomes, but are outside the scope of traditional sequence alignment algorithms

The Context Viewer can be used for discovering functional gene groups, gene-level insertion/deletion/duplication events, and other structural disruptions to core syntenic regions. By interlinking between the single gene-family phylogenetic perspective and the multi-genic genomic context, hypotheses formed regarding evolutionary relationships (and the implied conservation of functional relationships) from one view can be re-considered from the alternative perspective afforded by the other view.

The context viewer is implemented in Javascript. The D3 Javascript library is used for graphical elements, and all data are fetched via AJAX. The web service providing data to the Context Viewer is currently implemented in the Django web framework, but the user interface could be used against any service providing appropriately structured JSON data. We plan to build a Tripal module for the context viewer so we can better integrate it into LIS and other Drupal/Tripal based sites.

Genes, the 'basket' functionality and protein domains

The list of genes resulting from gene or gene family searches can be saved in a 'basket' for further set-based analyses. This enables, for example, downloading a FASTA file containing genomic or transcript sequence. The basket will persist throughout a user's session and genes can be added and deleted individually or in bulk.

Gene models are annotated with InterPro protein domains to reveal conservation of functional elements across evolutionary time, as well as to provide a functional context in which to interpret the likely impact of intraspecific variation. The protein domain search tool provides information derived from InterProScan analysis of protein sequences against InterPro protein signature databases. The records show the corresponding domain name along with the InterPro term and description, count of genes and gene family consensus sequences that are associated with the domain. The counts are linked to gene and gene family pages for more information.

Sequence search

BLAST. The Tripal BLAST extension module (http://tripal.info/extensions/modules/tripal-blast-ui) is instantiated at LIS and enables researchers to run BLAST (30) against all eight available genomes and their gene models (both CDS and proteins), against the combined genomes, the combined gene models and a set of consensus sequences derived from each gene family. The results show a table of hits, details about each hit including a summary visualization and the option to download the results in XML, text or table form.

BLAT. The BLAT (31) search built into GBrowse is deployed to permit sequence searching via BLAT against each of the eight genomes. A summary of the results are displayed on an ideogram of the genome's chromosomes and clicking a result on the ideogram or name in the result table will link to that position in the genome browser.

Keyword search. The keyword search built into GBrowse is provided to search for genomic features displayed within GBrowse by searching its name and description for search terms. As with the BLAT search, a summary of the results is displayed on an ideogram of the genome's chromosomes and clicking a result on the ideogram or name in the result table will link to that position in the genome browser.

Update and release cycles, versioning and metadata. The LIS project has a monthly update cycle. Update history is available on the right side of the home page and at http://legumeinfo.org/news_page_all. In general, as new major

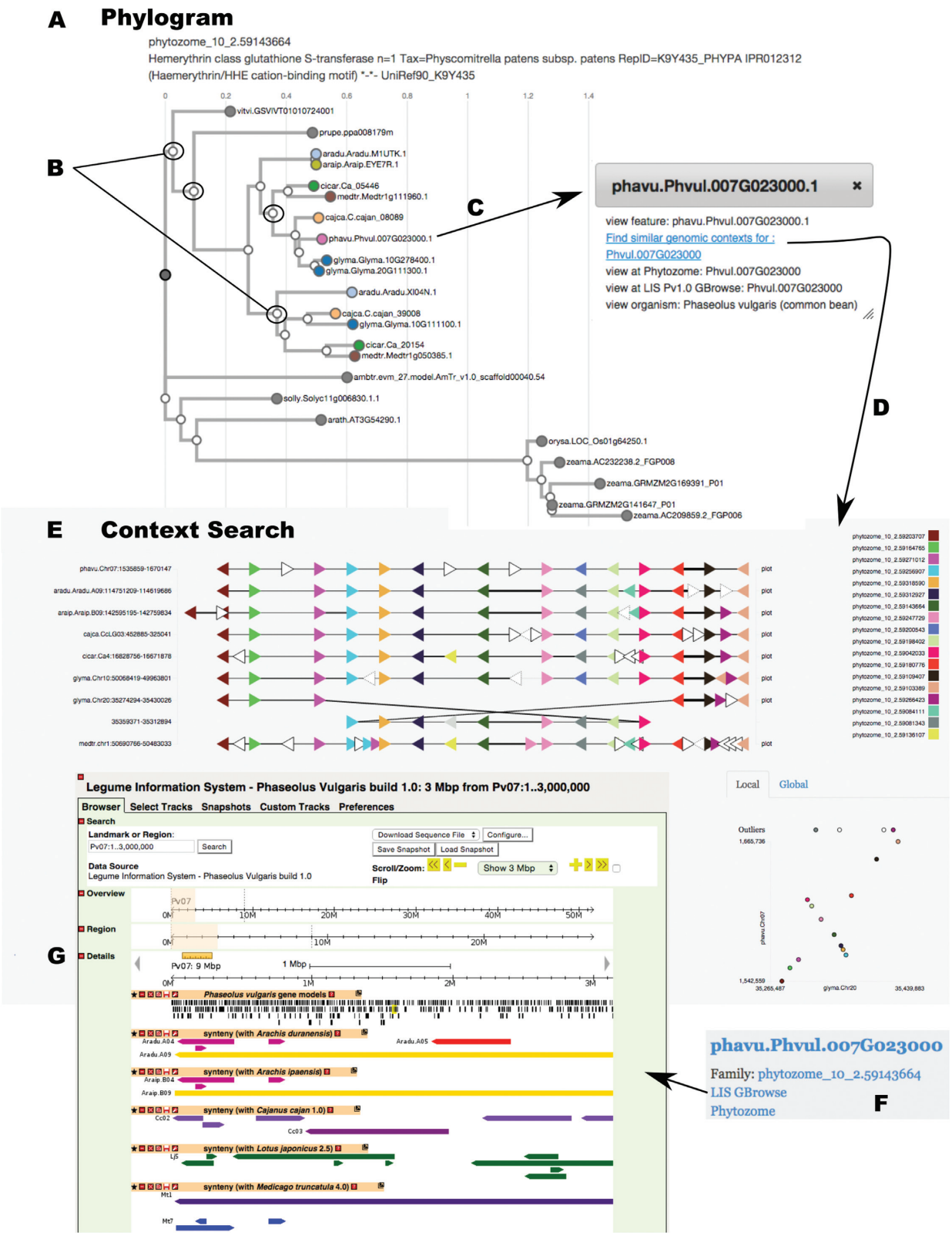

Figure 2. Screenshots of a phylogram, Genomic Context Viewer and genome browser. **(A)** An interactive phylogram display: Clicking on the dark gray root node gives the user the option of viewing all legume member genes in the Genomic Context Viewer or in a multiple sequence alignment format. **(B)** Clicking an internal white node gives the user the option to view the Genomic Context Viewer for the legume genes from that subtree. **(C)** Clicking a colored legume gene node opens a menu with with links to information about that gene and locus, including **(D)** a link to the Genomic Context Viewer, which invokes a search for genomic contexts similar to the one containing the gene selected. **(E)** In the Genomic Context Viewer, genes are colored by gene families, as shown in the legend. Synteny between two species can be seen in a dot plot, with the dots also colored by gene family. **(F)** Clicking on an individual gene within the Genomic Context Viewer presents an option to view the gene in LIS GBrowse. **(G)** Within that GBrowse instance, the selected gene is highlighted and the user can select to view a wide range of synteny, gene models and marker tracks.

genome resources become available, we try to incorporate those as soon as possible, given developer time constraints. Genome and annotation versioning follows the versioning conventions of the genome projects. Where there are stable repositories for genome and annotation resources, we maintain links to those repositories. We generally maintain older versions of genomes and browsers—for example, with genome browsers being maintained for Lotus 2.5 and 3.0, and for soybean 1.0 and 2.0 (Wm82 accession) at partner site SoyBase.

CONCLUSION

LIS will continue to expand its data collection, both in terms of species representation and data richness. Future development plans include a multispecies instance of InterMine (15), additional and continued improvement of gene families, and addition of more genetic map and QTL and GWAS data sets. For these types of complex data, participation of subject-area experts will be important. The number of whole genome sequences and transcriptome data sets are also expected to increase rapidly, so new approaches to housing, displaying and incorporating these data sets will need to be worked out. Expanding interoperability with other legume GDPs will also be critical. This will be accomplished in part through work on data collection and sharing standards and by contributing to the development of reusable software modules.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the many developers of GMOD components used at LIS, particularly including Stephen Ficklin, Lacey Sanderson and Dorrie Main for Tripal, and developers of the Legume Federation sites for cooperation and help during LIS development.

FUNDING

USDA-ARS project funding for the Legume Information System. National Science Foundation [IOS-1444806 to S.B.C., A.D.F., E.K.C.]. Funding for open access charge: USDA-ARS core funds.
Conflict of interest statement. None declared.

REFERENCES

1. Graham,P.H. and Vance,C.P. (2003) Legumes: importance and constraints to greater use. *Plant Physiol.*, **131**, 872–877.
2. O'Rourke,J.A., Bolon,Y.T., Bucciarelli,B. and Vance,C.P. (2014) Legume genomics: understanding biology through DNA and RNA sequencing. *Ann. Biology*, **113**, 1107–1120.
3. Gonzales,M.D., Archuleta,E., Farmer,A., Gajendran,K., Grant,D., Shoemaker,R., Beavis,W.D. and Waugh,M.E. (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.*, **33**, D660–D665.
4. The *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.
5. Dash,S., Cannon,E.K.S., Kalberer,S.R., Farmer,A.D. and Cannon,S.B. (2016). PeanutBase and Other Bioinformatic Resources for Peanut. In Wilson,R. and Stalker,T.H., (eds.), *Peanuts: Genetics, Processing, & Utilization*. AOCS Press, Urbana, IL, pp. 241–253.
6. Cannon,S.B., Crow,J.A. and Grant,D.M. (2012) SoyBase and the legume information system: accessing information about the soybean and other legume genomes. In Wilson,R. (ed.), *Designing Soybeans for 21st Century Markets*. AOCS Press, Urbana, IL, pp. 53–66.
7. Ficklin,S.P., Sanderson,L.A., Cheng,C.H., Staton,M.E., Lee,T., Cho,I.H., Jung,S., Bett,K.E. and Main,D. (2011) Tripal: a construction toolkit for online genome databases. *Database*, bar044.
8. Sanderson,L.A., Ficklin,S.P., Cheng,C.H., Jung,S., Feltus,F.A., Bett,K.E and Main,D. (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database*, bat075.
9. Zhou,P., Emmert,D. and Zhang,P. (2006) Using Chado to store genome annotation data. *Curr. Protoc. Bioinform.*, Chapter 9, Unit 9.6, doi:10.1002/0471250953.bi0906s12.
10. Youens-Clark,K., Faga,B., Yap,I.V., Stein,L. and Ware,D. (2009) CMap 1.01: a comparative mapping application for the Internet. *Bioinformatics*, **25**, 3040–3042.
11. Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinform.*, Chapter 9, Unit 9.9. doi: 10.1002/0471250953.bi0909s28.
12. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: A next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
13. Main,D., Cheng,C.-H., Ficklin,S.P., Sanad,M., Jung,S., Lee,T., Zheng,P., Coyne,C., McGee,R. and Mockaitis,K. (2014) The Cool Season Food Legume Database: An Integrated Resource for Basic, Translational and Applied Research. *Proceedings of the International Plant and Animal Genome Conference*, San Diego, CA, USA.
14. Krishnakumar,V., Kim,M., Rosen,B.D., Karamycheva,S., Bidwell,S.L., Tang,H. and Town,C.D. (2015) MTGD: The *Medicago truncatula* Genome Database. *Plant Cell Physiol.*, **56**, e1–e9.
15. Goff,S.A., Vaughn,M., McKay,S., Lyons,E., Stapleton,A.E., Gessler,D., Matasci,N., Wang,L., Hanlon,M., Lenards,A. *et al.* (2011) The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.*, **2**, 34.
16. Smith,R.N., Aleksic,J., Butano,D., Carr,A., Contrino,S., Hu,F., Lyne,M., Lyne,R., Kalderimis,A., Rutherford,K. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, **28**, 3163–3165.
17. Grant,D., Nelson,R.T., Cannon,S.B. and Shoemaker,R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **38**, D843–D846.
18. Schmutz,J., McClean,P.E., Mamidi,S., Wu,G.A., Cannon,S.B., Grimwood,J., Jenkins,J., Shu,S., Song,Q., Chavarro,C. *et al.* (2014) A reference genome for common bean and genome-wide analysis for dual domestications. *Nat. Genet.*, **46**, 707.
19. Varshney,R.K., Song,C., Saxena,R.K., Azam,S., Yu,S., Sharpe,A.G., Cannon,S., Baek,J., Rosen,B.D., Tar'an,B. *et al.* (2012) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.*, **30**, 240–246.
20. Varshney,R.K., Chen,W., Li,Y., Bharti,A.K., Saxena,R.K., Schlueter,J.A., Donoghue,M.T.A., Azam,S., Fan,G., Whaley,A.M. *et al.* (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.*, **30**, 83–89.
21. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
22. Tang,T., Krishnakumar,V., Bidwell,S., Rosen,B., Chan,A., Zhou,S., Gentzbittel,L., Childs,K.L., Yandell,M., Gundlach,H. *et al.* (2014) An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics*, **15**, 312.
23. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.*, **15**, 227–239.
24. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012)

Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
25. Haas,B.J., Delcher,A.L., Wortman,J.R. and Salzberg,S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.
26. Mitchell,A., Chang,H., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
27. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25** 1189–1191.
28. Jones,P., Binns,D., Chang,H., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
29. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
31. Kent,W.J. (2002) BLAT—The BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.