The legume information system and associated online genomic resources

Joel Berendzen1 | Anne V. Brown2 | Connor T. Cameron1 | Jacqueline D. Campbell2,3 | Alan M. Cleary1 | Sudhansu Dash1 | Samuel Hokin1 | Wei Huang2 | Scott R. Kalberer2 | Rex T. Nelson2 | Sven Redsun1 | Nathan T. Weeks2 | Andrew Wilkey4 | Andrew D. Farmer1 | Steven B. Cannon2

1National Center for Genome Resources, Santa Fe, NM, USA
2USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA, USA
3Department of Computer Science, Iowa State University, Ames, IA, USA
4ORISE, Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA, USA

Correspondence
Steven Cannon, Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA, USA.
Email: steven.cannon@ars.usda.gov

Present address
Jacqueline D. Campbell, National Center for Genome Resources, Santa Fe, NM, USA and Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA, USA.

Funding information
US. Department of Agriculture, Agricultural Research Service, Grant/Award Number: 5030-21000-069-00D; NSF, Division of Biological Infrastructure: Federated Plant Database Initiative for the Legumes, Grant/Award Number: 1444806

Abstract
The Legume Information System (LIS; https://legumeinfo.org) houses genetic and genomic data, integrated in various online tools to allow comparative genomic analyses. The website and database maintain data for more than two dozen species, particularly focusing on crop and model species and holding data for other diverse species of taxonomic interest. Major analysis features include genome browsers, sequence-search tools, legume-focused gene families and a phylogenetic tree viewer, a gene annotation service (which places a submitted gene into a gene family and phylogenetic tree), an interactive microsynteny and pan-genome viewer, a novel viewer of genetic variant data, genetic maps and viewers, a Data Store for data sets such as assemblies and annotations, InterMine instances for querying genetic and genomic data, and a tool for viewing geographic distributions of germplasm accessions. LIS also integrates with several other legume data resources and tools, including PeanutBase (https://peanutbase.org), SoyBase (https://soybase.org), Medicago Hapmap (https://medicagohapmap2.org), Alfalfa Breeder’s Toolbox (https://alfalfatoolbox.org), and the Legume Federation (https://legumefederation.org).

KEYWORDS
comparative genomics, CoNekT, determinacy, gene families, genome context viewer, InterMine, Legume Information System, pan-genome, ZZBrowse

1 | INTRODUCTION

The mission of the Legume Information System (LIS; https://legumeinfo.org) is to facilitate research and crop improvement for the many legume species that are important in global agriculture. LIS is a collaborative project between the USDA-Agricultural Research Service and the National Center for Genome Resources (Dash et al., 2016; Gonzales et al., 2005; Gonzales, Gajendran, Farmer, Archuleta, & Beavis, 2007). LIS houses genetic and genomic data for 30 legume species, with extensive support for 18 of these, as of late 2020.

Strengths of LIS include a Data Store, to serve as a one-stop shop for static, versioned legume genetic and genomic data sets such as genome assemblies and annotations; genome browser support for many species; a gene family and phylogenetic viewer; a micro- and macro-synteny viewer; a tool for viewing geographic distributions of germplasm accessions; InterMine instances for querying genetic and
genomic data; tools for exploring gene expression information; and viewers for genetic variant data, including variants associated with traits through published genome-wide association study (GWAS) analyses. We will describe each of these features below.

We first give an overview of the holdings and main tools at LIS and then illustrate the use of some of these tools and resources with a case study to investigate growth habit in cowpea and soybean.

2 MATERIALS AND METHODS

LIS holds, at time of writing, substantial genetic and genomic data sets for more than two dozen legume species, as well as some specialized cross-cutting data sets such as legume-focused gene families. Many of these are explorable via various graphical user interfaces (GUIs), such as genome browsers, genetic map viewers, sequence search tools, or query and report interfaces for features such as genes and markers. Genome browsers are provided for sequenced crops and models: crops ranging from *Arachis hypogaea* (peanut) to *Vigna unguiculata* (cowpea) and models including *Medicago truncatula* and *Lotus japonicus*.

2.1 The Data Store

To provide access to the underlying data used in visualization and query tools, LIS also maintains a Data Store (https://legumeinfo.org/data/), which holds data sets for download. These data sets include genome assemblies, gene models and sequences, genetic markers, maps, Quantitative Trait Loci (QTLs), GWAS results, and expression data.

In order to help organize data, make it Findable, Accessible, Interoperable, and Reusable (FAIR; Wilkinson et al., 2016), and to aid in its curation and use, the LIS team has adopted a set of practices for consistently organizing, naming, formatting, describing, and tracking the data. These standardized patterns are helpful both for outside users and for LIS developers and curators. The standardization has an important additional benefit of making the data findable and directly usable by applications within LIS and by other projects. The main organizing principles are as follows:

The Data Store web interface mirrors a file system of limited depth, with main Genus_species directories, which hold data “collections.” Collections are groups of thematically related files from a major, versioned analysis or study. Examples are genome assemblies, annotations, maps, genotype collections, expression, and QTLs.

Collection names employ a naming scheme: accession_name, type +version, and random-key, for example, Wm82.gnm2.DTC4, Wm82.gnm2.ann1.RVB5, and Wm82.gnm2.div.0SZD. In these examples, the accession_name is Wm82 (abbreviated from Williams 82); the type +version is gnm2, gnm2.ann1, and gnm2.div, indicating genome assembly version 2, genome annotation 1 (on assembly 2), and diversity set (on assembly 2). The random key for each of these is the four-letter alphanumeric string at the end. This unique key is also used in each file name contained within collections and in metadata that describes each collection. The key, then, serves to concisely tie the files and the metadata together.

Within each collection, there are four metadata files: a README, a MANIFEST file that describes each data file, a MANIFEST file that indicates correspondences between filenames in the Data Store and any prior filenames, and a CHECKSUM file. The metadata files all use the “YAML” format, which is both human-readable and is sufficiently structured to be machine-readable. The README has 26 fields (entries), such as identifier (the collection key name), provenance (the data origin), source (the URL address of the data, if appropriate), and subject (a brief description of the data). The README ends with the fields file_transformation (indicating modifications to files, if any) and changes (indicating modifications, if any, to the collection after its initial version). The Data Store interface is implemented using h5ai (https://larsjung.de/h5ai/).

2.2 DSCensor: Tools for summarizing and conducting quality control on Data Store contents

The Data Store contents are indexed and described in the “DSCensor” (https://dscensor.legumeinfo.org), so named for its dual role as a “sensor” of data in the Data Store, and as a “censor” of data sets that may have problems that need evaluation or fixing by project curators—problems such as malformed general feature format (GFF) structure, noncanonical naming patterns, and unexpected assembly statistics or annotation results. The DSCensor uses the MultiQC tools (Ewels, Magnusson, Lundin, & Kaller, 2016), which were developed to describe, analyze, and report on data from multiple bioinformatic analyses—for example, at a sequencing and annotation facility. Plots and tables provide a helpful overview of data characteristics—for example, assembly sizes, scaffold N50s, proportion of gaps in assemblies, and BUSCO results (Seppey, Manni, & Zdobnov, 2019) for gene models, to provide an indication of the completeness of assemblies and annotations.

Source code for DSCensor is available at https://github.com/legumefederation/dscensor.

2.3 Genome browsers, for exploring genes, expression information, synteny, and genomic organization

LIS holds GBrowse (Stein, 2013) and JBrowse (Buels et al., 2016) genome browsers for more than a dozen species—and close to two dozen, if counting partner sites PeanutBase and SoyBase (*Arachis* and *Glycine* genomes are linked from LIS and integrated across these sites via queries and browser track links). It is worth noting that the files maintained in the Data Store are indexed for direct use by modern genome browsers, so that annotation tracks may be served from the Data Store by browsers at alternate sites (e.g., http://medicagohapmap2.org serving *Medicago* annotations whose curation...
has been centralized at the Data Store). Similarly, command line tools such as samtools can be used to access subsets of gene sequences or genomic regions via calls to the https-hosted indexed fasta files.

One strength of the browsers at LIS is the cross-species synteny tracks, which identify regions of correspondence between species represented in the present browser and other related species. These synteny tracks are calculated using comparisons between protein sequences in the various species. This allows for identification across significant evolutionary distances (e.g., Glycine to Medicago); it also allows for estimates of evolutionary distance in each synteny block, in the form of Ks values calculated by comparing corresponding coding sequences within the blocks. These are reported, for each synteny block, as median Ks values between the present species and the target species for a given synteny track. These precalculated synteny tracks are also loaded from the Data Store into the InterMine instances to facilitate cross-mine traversal by syntenic region.

2.4 | Gene families, phylogenetic viewer, and sequence annotation tool, for exploring ortholog and paralog relationships

LIS provides access to a set of legume-focused gene families, both for download in several formats and for query and visual exploration. The current families were calculated using 14 proteomes from sequenced legume species and five from selected outgroup species. Special effort was made to circumscribe the families to include evolutionary events (speciations and whole genome duplications) within the timeframe of the legume family (extending back about 65 Mya) but not older events—avoiding the angiosperm triplication (Koenen et al., 2020). This tends to give families of reasonable size, typically including the whole genome duplication that predated the origin of the papilionoid subfamily (Koenen et al., 2020; Ren, Huang, & Cannon, 2019; Stai et al., 2019).

The gene families are available for download via the Data Store and for searching and browsing via https://legumeinfo.org/search/phylotree. In the latter instance, families can be filtered by gene content and count, key word, or family name. Families can also be searched by gene from the gene search page (https://legumeinfo.org/search/gene). Within the phylootree viewer, genes are colored by species and also identified using a five-letter prefix, which takes three letters from the genus and two from the species epithet, for example, glyma for *Glycine max* and medtr for *M. truncatula*.

The tree visualizations (Figure 4) are interactive. Clicking on a gene name exposes links relevant to that gene (e.g., to the gene page within LIS or a comparable one at Phytozome). The link “Find similar genomic contexts” links to the genome context viewer (GCV), described next. Interior nodes of the trees are also active elements allowing sets of genes that are children of subtrees to be used as inputs to other tools (e.g., for list creation in the InterMine instances to be described later). Large trees can also be filtered down through either collapsing interior nodes or through focusing on subsets of taxa through interaction with the taxon composition display. All such changes to the visible content of the tree will also be reflected in the multiple sequence alignment viewer that is available through the multiple sequence alignment (MSA) visualization button near the top of the tree display. LIS also provides a tool (https://legumeinfo.org/annot) to annotate user-provided sequences and to place them into phylogenetic trees for viewing. For a supplied sequence, this translates the sequence (if a nucleotide sequence was provided), then searches against selected legume proteome datasets, conducts an InterProScan analysis for domain homology (Jones et al., 2014), searches and assigns the sequence to the LIS gene families using HMMER (Finn, Clements, & Eddy, 2011), aligns the sequence, and recalculates and displays the gene tree. Functional descriptions are provided using the Automated Human Readable Descriptions (AHRD) pipeline (Jozashoori, Jozashoori, & Schoof, 2019).

The LIS phylogenetic viewer and annotation tool are implemented as Tripal modules (Sanderson et al., 2013; Spoor et al., 2019), with source code available at https://github.com/legumeinfo/tripal_phylotree and https://github.com/legumeinfo/tripal_funnotate, respectively. Code for placing user sequences in phylogenetic trees (a function associated with the annotation service) is available at https://github.com/LegumeFederation/lorax.

2.5 | The GCV, for interactive microsynteny and pan-genome exploration

The GCV (https://legumeinfo.org/gcv) is a web app for visualizing related genomic regions across a set of species or taxa (Cleary & Farmer, 2018). The GCV operates at the level of genes in genomic regions, typically spanning several hundred Kb, although this is configurable and very flexible. The program first identifies regions from the set of target taxa that have similar gene content as the query region and then aligns those regions, using gene families as the “alphabet” used in the alignment. Using gene families as the primary units in the analysis allows for very fast searches and alignments, making it possible to find and align corresponding regions from dozens of taxa.

GCV can be configured to operate on any collection of related genomes. At LIS, GCV operates on genomes across the papilionoid subfamily. Another instance of GCV is available at SoyBase (https://soybase.org/gcv), operating on a collection of *Glycine* accessions, including both several cultivated lines and the wild soybean, *Glycine soja*.

GCV makes extensive use of microservices internally and can provide this data via application programming interface (API) calls from other software. This is done, for example, by ZZBrowse, to identify syntenic regions between indicated species. This service architecture also helps enable GCV to function in federated contexts, drawing data from various nonlocal sources, as has been done for an instance for Arabidopsis, for example (https://gcv-arabidopsis.ncgr.org) (Pasha et al., 2020).

The code for GCV is available at https://github.com/legumeinfo/gcv.
2.6 | InterMine interface and data warehouse for selected legume species

As part of the National Science Foundation (NSF) Legume Federation project (Bauchet et al., 2019), a set of InterMine instances has been populated for seven legume genera (as of 2020), as well as an integrating LegumeMine that provides access to the data in the genus-specific mines, linking among species via gene family relationships. The legume mines are accessible at https://mines.legumeinfo.org. Intermine is a “data warehouse,” built originally to hold genomic data for fly, yeast, and nematode, but extended and made into a generic system for holding and providing access to genetic and genomic data (Kalderimis et al., 2014; Lyne et al., 2015). The InterMine interfaces provide a large collection of query templates to help users drill into genomic data sets; and it also provides methods for constructing custom queries. A particular strength of the mines is their ability to work with lists and genomic regions in nontrivial ways—for example, intersecting a list of genes with particular expression profiles, with genes falling within selected GWAS regions.

2.7 | Tools for exploring expression data at LIS

LIS currently houses tissue and developmental expression atlas data for five species. These data can be accessed in several ways.

First, the data are available, in flat-file formats, in the Data Store, in collections indicated with “exp” (this and similar terms are searchable in the Data Store user interface).

Second, gene expression heat maps are available in genome browsers, for those species with published tissue expression atlases.

Third, a custom Drupal module displays expression data in the corresponding gene pages (https://legumeinfo.org/lis_expression/demo; source code: https://github.com/legumeinfo/lis_expression). In addition to providing expression profiles for each gene, the module can also identify genes that have similar expression profiles. The facility also extends to displaying expression of other paralogous sequences (if present) within a gene family. The module also has the provision for supplying the expression data along with its metadata in a json format.

Fourth, LIS has adopted the CoNekT analysis and visualization suite (Proost & Mutwil, 2018), to provide a comparative view of the expression levels of orthologous genes and to filter genes with a desired profile or level in a tissue (https://conekt.legumeinfo.org/). The tool groups samples into equivalent/corresponding tissues common among the represented species for the purpose of comparison and using gene family data for ortholog determination and applies normalizations to permit comparisons between species.

Fifth, the mines have gene expression bar charts on gene report pages and expression heat maps on gene list analysis pages. One can write queries to select genes based on expression level, tissue, or other attributes such as GO terms.

2.8 | Methods for exploring geographic distributions of legume species, accessions, and traits

LIS has a web app (https://legumeinfo.org/germplasm/map) for visualizing the geographic origins of accessions from legume species held in the USDA Germplasm Resources Information Network (GRIN). This tool allows for searches by genus or species, as well as application of various filters and viewing options. Users can also upload their own data, for example, to display quantitative or qualitative trait data for accessions in GRIN, or to display a researcher’s accessions, they have supplied with latitude and longitude coordinates and species, trait, or population structure data.

Source code for the LIS germplasm map is available at https://github.com/legumeinfo/lis_gis

2.9 | The GCViT, for exploring and comparing genetic variants (SNPs)

The Genotype Comparative Visualization Tool (GCViT; Wilkey, Brown, Cannon, & Cannon, 2020) at LIS enables researchers to explore variant data, as is typically provided by large-scale genotyping projects that generate Variant Call Format (VCF) files (Danecek et al., 2011). In a typical GCViT session, a user will select a genotyping project (which points to a VCF file held in the LIS Data Store) and then select one or more accessions within the project for display and comparison. Displays show counts of variants (single nucleotide polymorphisms [SNPs]) in bins across chromosomes, with counts either being absolute (i.e., number of SNPs in a bin at some location for an accession) or relative differences (i.e., the number of “same” or “different” SNPs between two selected accessions, in a bin at some location). Counts (absolute or of similarities or differences) can be rendered in several ways: as histograms, intensity heat maps, or haplotypes. Haplotypes are rendered as colored lines or blank extents (i.e., binary indications of presence or absence), indicating that counts of SNP differences (or similarities) between two accessions are above a selected threshold.

2.10 | Comparative GWAS exploration with ZZBrowse

ZZBrowse (https://zzbrowse.legumeinfo.org) is an interactive web application for GWAS and QTL visualization. Originally developed by the Baxter Laboratory (Ziegler, Hartsock, & Baxter, 2015) for single-species GWAS, we have adapted it for comparative GWAS and QTL between two species. It is described with a use case below. Species currently available include common bean, cowpea, pigeonpea, soybean, M. truncatula, and Arabidopsis thaliana; most data are provided directly through the LIS datastore. ZZBrowse depends on data from GWAS and QTL studies, which requires extraction and curation of the data from the associated research studies. The LIS team has therefore...
been working with research groups to curate and add data from their own studies, depositing the data in the Data Store in formats ready to be loaded into ZZBrowse, InterMine, and related LIS databases.

3 | RESULTS AND DISCUSSION

We use this section to present a case study, using tools at LIS to investigate whether there is a common genetic basis for growth habit in cowpea and soybean.

Growth habit has been an important target during domestication of most legume species. Components of growth habit differ in various legumes but can include factors such as vining/indeterminate versus dwarf/determinate growth, prostrate versus upright growth, various branching patterns, pod placement and orientation, stem robustness, and lodging propensity. In this section, we use tools at LIS to investigate whether the viny versus dwarf/determinate trait is common across two legume crops: cowpea and soybean (cowpea data coming from LIS and soybean from SoyBase).

3.1 | Are genetic/trait associations conserved across different legume crops?

The ZZBrowse tool (https://zzbrowse.legumeinfo.org) is designed to explore genetic association studies in two selected species, using genomic synteny as the basis for making the comparison between species. Beginning with ZZBrowse, we first select two species to examine: cowpea and soybean. In Figure 1, cowpea is framed in blue and soybean in orange. Users may import their own GWAS results as well.

In the example shown in Figure 1, we first filter for growth-habit-related traits: “Growth habit” in cowpea (data from LIS) and “Lodging,” “Plant height,” and “Stem termination” in soybean (data from SoyBase). Starting with a region with several growth-habit QTLs on cowpea chromosome 1, clicking on a QTL results in a zoomed-in view of those QTLs; and clicking on a gene within that window triggers a search for syntenic matches in soybean. ZZBrowse aligns itself with help of the GCV. This identifies two soybean regions with synteny: one on soybean chromosome Gm03 and one on Gm19 (these can be toggled under the “Related Regions” selector). The syntenic region on Gm19 shows strong GWAS features for Lodging, Plant height, and Stem termination.

When the Genomic Linkages feature is enabled, selecting a gene in Species 1 calls the GCV web services to identify similar regions in Species 2, using the number of matches and non-matches within a range of neighboring genes (Figure 2). Matching genes are color coded by gene family, in rainbow order relative to their Species 1 sequence, illustrating the degree of synteny between the two species. If more than one similar region exists in Species 2, the Related Regions menu lets you choose which one to view.

The URL automatically reflects the current state of ZZBrowse. To share an interesting result, just copy the current URL and send it to your colleague(s), who can then launch ZZBrowse and recreate the same view.

**FIGURE 1** ZZBrowse, showing “growth habit” associations for cowpea and soybean. The top panel shows cowpea Quantitative Trait Loci (QTLs), across all 11 chromosomes (there are trait associations on chromosomes 1 and 9). The bottom panel shows soybean genome-wide association study (GWAS) features, in a Manhattan plot across all 20 chromosomes.
3.2 What are the evolutionary histories of the candidate genes?

Looking at the gene descriptors (either with mouse rollover or in the “Annotations Table” view), the most likely candidate gene is the well-known determinacy (Dt1) gene (Tian et al., 2010; Zhong et al., 2017), Glyma.19G194300, annotated as “flowering locus protein T.” Clicking on this gene brings up a menu of other links related to this gene. One of these, “Find similar genomic contexts at LIS for: Glyma.19G194300,” links to the GCV at LIS (Figure 3).

This shows (with some filtering for species) synteny for four soybean chromosomal regions (Gm19, Gm03, Gm10, and Gm11) and two cowpea regions (Vu01 and Vu07). In this view, colored triangles represent genes from different gene families (e.g., left-facing
pink representing genes in the flowering locus T protein family). The homoeologous regions in soybean and cowpea represent regions duplicated through the early papilionoid whole genome duplication (WGD) and again through the Glycine WGD for soybean (Ren, Huang, & Cannon, 2019; Schmutz et al., 2010; Stai et al., 2019).

Clicking on a gene in the GCV again brings up a menu of links related to this gene. One of these links is to the gene family view at LIS: https://legumeinfo.org/chado_phylotree/legfed_v1_0.L_18BH5B (Figure 4).

The gene tree shown in Figure 4 shows the relationships of genes from the gene family that contains the soybean and cowpea determinacy (Dt1) genes (and orthologs from other species). The tree shows generally expected phylogenetic relationships—for example, with the warm-season legume species (soybean, pigeonpea, bean, cowpea, and mung bean—indicated by prefixes glyma, cajca, phavu, vigun, and vigra, respectively) falling in one clade and cool-season legume species falling in another (clover, Medicago, chickpea—indicated by prefixes trip, medtr, and cicar, respectively). At greater evolutionary distances, we see Lupinus angustifolius (prefix lupan) and peanut and its diploid relatives (A. hypogaea and A. duranensis and A. ipaensis; prefixes arahy, aradu, and araip). Also evident in the gene tree are several WGD: one affecting glycine (Schmutz et al., 2010), a triplication affecting Lupinus (Hane et al., 2017; Kroc, Koczyk, Świecicki, Kilian, & Nelson, 2014), and one affecting all of the papilionoid legume species shown here (Ren, Huang, & Cannon, 2019).

3.3 | Is there a signature of selection around the determinacy locus?

Because plant habit is a critical trait for soybean, we might expect to see evidence of selection around this locus—as, indeed, has been reported (Zhong et al., 2017). The expectation would be to see a wildtype (Glycine soja-like) haplotype in indeterminate accessions and alternate haplotype in determinate accessions. We can check this using the GCViT tool at SoyBase (GCViT is available for several species, at the appropriate genomic database sites: peanut at peanutbase.org, bean and chickpea at legumeinfo.org, and soybean at soybase.org).

GCViT produces visualizations of variants from genetic variation collections. These are generally VCF files associated with genotyping projects, derived either from an SNP chip or from genomic resequencing or genotyping-by-sequencing methods. The first step is to select a genotype collection. For soybean, we select the “USB 481” set, from Valliyodan et al. (2020). This has good genome-wide SNP density, and contains a good mix of G. soja and G. max accessions and elite and landrace lines.

We then select genotypes to compare. Although this can be done randomly, to be most informative, we need to know something about the accessions—in this case, about the determinacy status and whether the accessions are G. soja or G. max. This can be done by checking the metadata from the publication (https://doi.org/10.15482/USDA.ADC/1518301) and the phenotype information...
available for soybean at U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search).

We select one *G. soja* accession at random to use as a reference (PI597458); three short (<80 cm) and determinate accessions at random (PI549021, PI361087, and PI594012); and three tall (≥120 cm) and indeterminate accessions at random (PI407729, PI71465, and PI417500). These are plotted in Figure 5.

In Figure 5, although several apparent *G. soja* introgressions are evident (e.g., pericentromeric regions for several accessions on Gm02), the Glyma.19G194300 gene underlying the determinacy locus on Gm19 is of particular interest. This appears as small clear regions, at about 45 Mb, in the three indeterminate accessions, and as colored regions in the three determinate accessions. Although far from definitive, this small association is consistent with selection around this region, with cultivated (determinate) haplotype present in the determinate lines.

**FIGURE 5** Genotype Comparative Visualization Tool (GCViT) view of selected soybean accessions from Valliyodan et al. (2020). On the left of each chromosome, single nucleotide polymorphism (SNP) density is shown (as a gray histogram), for reference accession PI597458, from *Glycine soja*. To the right of each chromosome, the density of SNP differences is shown, for each selected accession compared with the reference accession. Darker colors indicate more SNP within a given bin. Thus, white areas indicate near-identity with the *G. soja* reference. The first three comparison accessions are determinate *G. max* lines. The last three comparison accessions are indeterminate *G. max* lines. Only selected chromosomes are shown, due to space limitations.

4 | SUMMARY

LIS provides numerous tools for accessing and exploring legume genetic and genomic data. A notable strength of LIS is the collection of comparative methods, linking together orthologs through their evolutionary histories, using gene trees, synteny, and homology search tools. The Data Store attempts to provide a one-stop shop for genomic data, organized in a systematic way, with standardized formats and file names, to enable efficient computation on many data sets at a time. The case study presented above illustrates how some of these tools and methods can be used to do “on-line biology” by comparing analyses carried out in different species. Investigating determinacy, we find similar genetic associations occurring in syntenic locations in cowpea and soybean and find possible signature of selection for this trait around the determinacy locus in soybean.

**ACKNOWLEDGMENTS**

The authors thank legume researchers who have provided data and motivation for this work. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

**CONFLICT OF INTEREST**

The authors declare no conflict of interest.
AUTHOR CONTRIBUTIONS
SC, AF, SK, NW, and SR did the writing—original draft and review and editing. SK, SH, AF, AB, CC, JC, WH, RN, and SC helped with data curation. AW, AC, and AB did the visualization. JB, SD, SH, RN, SR, NW, and AW helped with the software. SC and AF handled the project administration.

FUNDING INFORMATION
This research was funded in part by the NSF project “Federated Plant Database Initiative for the Legumes” (#1444806) and supported in part by the US. Department of Agriculture, Agricultural Research Service, project 5030-21000-069-00D.

DATA AVAILABILITY STATEMENT
All data sets described in this publication are freely available, as described in the manuscript.

ETHICAL STATEMENT
This study did not involve any human or animal testing.

ORCID
Nathan T. Weeks https://orcid.org/0000-0002-4085-6028
Steven B. Cannon https://orcid.org/0000-0003-2777-8034

REFERENCES
Bauchet, G. J., Bett, K. E., Cameron, C. T., Campbell, J. D., Cannon, E. K. S., Cannon, S. B., … Zhao, P. X. (2019). The future of legume genetic data resources: Challenges, opportunities, and priorities. Legume Science, 1, e16. https://doi.org/10.1002/leg3.16
Buels, R., Yao, E., Diesch, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., … Holmes, I. H. (2016). JBrowse: A dynamic web platform for genome visualization and analysis. Genome Biology, 17, 66. https://doi.org/10.1186/s13059-016-0924-1
Cleary, A., & Farmer, A. (2018). Genome context viewer: Visual exploration of multiple annotated genomes using microsynteny. Bioinformatics, 34(9), 1562–1564. https://doi.org/10.1093/bioinformatics/btx757
Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Genome Project Analysis G. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btp330
Dash, S., Campbell, J. D., Cannon, E. K., Cleary, A. M., Huang, W., Kalberer, S. R., … Cannon, S. B. (2016). Legume information system (LegumeInfo.org): A key component of a set of federated data resources for the legume family. Nucleic Acids Research, 44(D1), D1181–D1188. https://doi.org/10.1093/nar/gkv1159
Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354
Finn, R. D., Clements, J., & Eddy, S. R. (2011). HHBlits web server: Interactive sequence similarity searching. Nucleic Acids Research, 39(Web Server issue), W29–W37. https://doi.org/10.1093/nar/gkr367
Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., … Wauth, M. E. (2005). The legume information system (LIS): An integrated information resource for comparative legume biology. Nucleic Acids Research, 33(Database issue), D660–D665. https://doi.org/10.1093/nar/gki128
Gonzales, M. D., Gajendran, K., Farmer, A. D., Archuleta, E., & Beavis, W. D. (2007). Leveraging model legume information to find candidate genes for soybean sudden death syndrome using the legume information system. Methods in Molecular Biology, 406, 245–259. https://doi.org/10.1007/978-1-59745-535-0_11
Hane, J. K., Ming, Y., Kamphuis, L. G., Nelson, M. N., Garg, G., Atkins, C. A., … Singh, K. B. (2017). A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: Insights into plant-microbe interactions and legume evolution. Plant Biotechnology Journal, 15(3), 318–330. https://doi.org/10.1111/pbi.12615
Jones, P., Birns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. Bioinformatics, 30(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
Jozzahoori, S., Jozzahoori, A., & Schoof, H. (2019). AFDP: An automated function description prediction approach to improve accuracy of protein function predictions. arXiv:191006965v1
Kalderinis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., … Micklem, G. (2014). InterMine: Extensive web services for modern biology. Nucleic Acids Research, 42(Web Server issue), W468–W472. https://doi.org/10.1093/nar/gku301
Koenen, E. J. M., Oijeda, D. I., Bakker, F. T., Wieringa, J. J., Kidner, C., Hardy, O. J., … Hughes, C. E. (2020). The origin of the legumes is a complex Paleopolyploid Phylogenomic tangle closely associated with the cretaceous-Paleogene (K-Pg) mass extinction event. Systematic Biology, syaa041. https://doi.org/10.1093/sysbio/syaa041
Kroc, M., Koczyk, G., Swiecicki, W., Kiliann, A., & Nelson, M. N. (2014). New evidence of ancestral polyploidy in the Genistoid legume Lupinus angustifolius L. (narrow-leaved lupin). Theoretical and Applied Genetics, 127(5), 1237–1249. https://doi.org/10.1007/s00122-014-2294-y
Lyne, R., Sullivan, J., Butano, D., Contrino, S., Heimbach, J., Hu, F., … Micklem, G. (2015). Cross-organism analysis using InterMine. Genesis, 53(8), 547–560. https://doi.org/10.1002/dvg.22869
Pasha, A., Subramaniam, S., Cleary, A., Chen, X., Berardini, T., Farmer, A., … Provart, N. (2020). Araport lives: An updated framework for Arabidopsis Bioinformatics. Plant Cell, 32(9), 2683–2686. https://doi.org/10.1105/tpc.20.00358
Proost, S., & Mutwil, M. (2018). CoNeKt: An open-source framework for comparative genomic and transcriptomic network analyses. Nucleic Acids Research, 46(W1), W133–W140. https://doi.org/10.1093/nar/gkx336
Ren, L., Huang, W., & Cannon, S. B. (2019). Reconstruction of ancestral genome reveals chromosome evolution history for selected legume species. The New Phytologist, 229(4), 2190–2193. https://doi.org/10.1111/nph.15770
Sanderson, L. A., Ficklin, S. P., Cheng, C. H., Jung, S., Feltus, F. A., Bett, K. E., & Main, D. (2013). Tripl v1.1: A standards-based toolkit for construction of online genetic and genomic databases. Database: The Journal of Biological Databases and Curation, 2013, bat075. https://doi.org/10.1093/database/bat075
Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. A. (2010). Genome sequence of the paleopolyploid soybean. Nature, 463, 178–183. https://doi.org/10.1038/nature08670
Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome annotation and annotation completeness. Methods in Molecular Biology, 1662, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14
Spoor, S., Cheng, C. H., Sanderson, L. A., Condon, B., Almsaeeed, A., Chen, M., … Ficklin, S. P. (2019). Tripal v3: An ontology-based toolkit for construction of FAIR biological community databases. Database: The Journal of Biological Databases and Curation, 2019, 1–11. https://doi.org/10.1093/database/baz077
Stai, J. S., Yadav, A., Sinou, C., Bruneau, A., Doyle, J. J., Fernandez-Baca, D., & Cannon, S. B. (2019). Cercis: A non-polyploid genomic relic within the generally Polyploid legume family. Frontiers in Plant Science, 10, 345. https://doi.org/10.3389/fpls.2019.00345
Stein, L. D. (2013). Using GBrowse 2.0 to visualize and share next-generation sequence data. *Briefings in Bioinformatics*, 14(2), 162–171. https://doi.org/10.1093/bib/bbt001

Tian, Z., Wang, X., Lee, R., Li, Y., Specht, J. E., Nelson, R. L., ... Ma, J. (2010). Artificial selection for determinate growth habit in soybean. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19), 8563–8568. https://doi.org/10.1073/pnas.100088107

Valliyodan, B., Brown, A. V., Wang, J., Gunvant, P., Liu, Y., Otyama, P. I., ... Nguyen, H. T. (2020). Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing [data set]. Figshare. https://figshare.com

Wilkey, A., Brown, A. V., Cannon, S. B., & Cannon, E. K. S. (2020). GCViT: A method for interactive, genome-wide visualization of resequencing and SNP array data. *BMC Genomics* In press, 21, 1–9. https://doi.org/10.1186/s12864-020-07217-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Zhong, L., Yang, Q., Yan, X., Yu, C., Su, L., Zhang, X., & Zhu, Y. (2017). Signatures of soft sweeps across the Dt1 locus underlying determinate growth habit in soybean [Glycine max (L.) Merr.]. *Molecular Ecology*, 26(18), 4686–4699. https://doi.org/10.1111/mec.14209

Ziegler, G. R., Hartsock, R. H., & Baxter, I. (2015). Zbrowse: An interactive GWAS results browser. *PeerJ Computer Science*, 1, e3. https://doi.org/10.7717/peerj-cs.3

**How to cite this article:** Berendzen, J., Brown, A. V., Cameron, C. T., Campbell, J. D., Cleary, A. M., Dash, S., Hokin, S., Huang, W., Kalberer, S. R., Nelson, R. T., Redsun, S., Weeks, N. T., Wilkey, A., Farmer, A. D., & Cannon, S. B. (2021). The legume information system and associated online genomic resources. *Legume Science*, 3(3), e74. https://doi.org/10.1002/leg3.74