Received: 5 August 2019 | Revised: 25 September 2019 | Accepted: 28 September 2019

DOI: 10.1002/leg3.16

COMMENTARY

Legume Science WILEY

The future of legume genetic data resources: Challenges, opportunities, and priorities

Guillaume J. Bauchet[1] | Kirstin E. Bett[2] | Connor T. Cameron[3] | Jacqueline D. Campbell[4] | Ethalinda K.S. Cannon[4] | Steven B. Cannon[5] | Joseph W. Carlson[6] | Agnes Chan[7] | Alan Cleary[3] | Timothy J. Close[8] | Douglas R. Cook[9] | Amanda M. Cooksey[10] | Clarice Coyne[11] | Sudhansu Dash[3] | Rebecca Dickstein[12] | Andrew D. Farmer[3] | David Fernández-Baca[4] | Samuel Hokin[3] | Elizabeth S. Jones[13] | Yun Kang[14] | Maria J. Monteros[14] | María Muñoz-Amatriaín[15] | Kirankumar S. Mysore[14] | Catalina I. Pislariu[16] | Chris Richards[17] | Ainong Shi[18] | Christopher D. Town[7] | Michael Udvardi[14] | Eric Bishop von Wettberg[19] | Nevin D. Young[20] | Patrick X. Zhao[14]

[1]Boyce Thompson Institute, Ithaca, New York, USA

[2]Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

[3]National Center for Genome Resources, Santa Fe, New Mexico

[4]Department of Computer Science, Iowa State University, Ames, Iowa, USA

[5]USDA-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, Iowa, USA

[6]Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA

[7]J. Craig Venter Institute, Rockville, Maryland

[8]Department of Botany and Plant Sciences, University of California, Riverside, California, USA

[9]Department of Plant Pathology, University of California Davis, Davis, California, USA

[10]CyVerse, University of Arizona, Tucson, Arizona

[11]USDA-ARS Plant Germplasm Introduction and Testing, Washington State University, Pullman, Washington, USA

[12]Department of Biological Sciences, University of North Texas, Denton, Texas

[13]Institute of Biotechnology, Cornell University, Ithaca, New York, USA

[14]Noble Research Institute LLC, Ardmore, Oklahoma

[15]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado

[16]Department of Biology, Texas Woman's University, Denton, Texas

[17]USDA-ARS National Center for Genetic Resources Preservation, Fort Collins, Colorado

[18]Department of Horticulture, University of Arkansas, Fayetteville, Arkansas

[19]Department of Plant and Soil Science, University of Vermont, Burlington, Vermont

[20]Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota

Recommended citation: LGDWG (2019)

This paper is the product of the Legume Genomic Data Working Group, who discussed and debated topics presented here and are listed in alphabetical order. A subset of the members of the Working Group, indicated by Guillaume J. Bauchet, Kirstin E. Bett, Jacqueline D. Campbell, Ethalinda K.S. Cannon, Steven B. Cannon, Joseph W. Carlson, Agnes Chan, Timothy J. Close, Amanda M. Cooksey, Clarice Coyne, Rebecca Dickstein, Andrew D. Farmer, Samuel Hokin, Elizabeth S. Jones, Maria J. Monteros, María Muñoz-Amatriaín, Chris Richards, Michael Udvardi, Eric Bishop von Wettberg, Nevin D. Young, assembled and edited the manuscript. All listed authors commented on and approved the final manuscript.

Correspondence
Steven Cannon, USDA-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA.
Email: steven.cannon@ars.usda.gov

Funding information
Agricultural Research Service, Grant/Award Number: 5030-21000-062-00D; Division of Biological Infrastructure, Grant/Award Number: #1444806; U.S. Department of Energy, Grant/Award Number: DE-AC02-05CH112; USDA Agricultural Research Service, Grant/Award Number: 5030–21000-062-00D; Federated Plant Database Initiative for the Legumes, Grant/Award Number: 1444806

Abstract

Legumes, comprising one of the largest, most diverse, and most economically important plant families, are the subject of vibrant research and development worldwide. Continued improvement of legume crops will benefit from the recent proliferation of genetic (including genomic) resources; but the diversity, scale, and complexity of these resources presents challenges to those managing and using them. A workshop held in March of 2019 addressed questions of data resources and priorities for the legumes. The workshop identified various needs and recommendations: (a) Develop strategies to effectively store, integrate, and relate genetic resources collected in different projects. (b) Leverage information collected across many legume species by standardizing data formats and ontologies, improving the state of metadata about datasets, and increasing use of the FAIR data principles. (c) Advocate for the critical role that curators exercise in integrating complex datasets into databases and adding high value metadata that enable downstream analytics and facilitate practical applications. (d) Implement standardized software and database development practices to best leverage limited developer time and expertise gained from the various legume (and other) species. (e) Develop tools and databases that can manage genetic information for the world's plant genetic resources, enabling efficient incorporation of important traits into breeding programs. (f) Centralize information on databases, tools, and training materials and establish funding streams to support training and outreach.

KEYWORDS
breeding, database, genotyping, informatics, international collaboration, legumes, reference genomes

1 | INTRODUCTION

Legumes (Fabaceae) comprise the third largest plant family, with more than 20,000 species (LPWG, 2017). Roughly two dozen crop legumes are critical within food systems worldwide, because of their overall high nutritional composition, nitrogen-fixing capacityand other keyroles inagroecosystems (Meyer, DuVal, & Jensen, 2012; Smýkal et al., 2015; Smýkal, Nelson, Berger, & von Wettberg, 2018). Continued development of legume varieties is essential for addressing new environmental challen-. ges due to climate change and needs for improved production and quality traits

The NSF and USDA funded a planning session in late 2004 to develop objectives and goals for cross-legume genomics research with the participation of approximately 50 legume researchers. The meeting was named "Cross-legume Advances Through Genomics" (CATG) and resulted in a white paper and meeting report (Gepts et al., 2005), which laid out strategic goals over a 3-, 5-, and 10-year timeline.

In March of 2019, 15 years after the initial 2004 meeting, a follow-up workshop was held, with the participants comprising the Legume Genomic Data Working Group, to assess the state of legume genomics research and to develop a strategic plan for the coming decade in this field, with support from the NSF (Legume Federation project, award 1444806), the USDA, and the Noble Research Institute. Below, we report the main recommendations from this meeting.

1.1 | Reviewing the roadmap, 15 years after the CATG meeting

With rapid advances in genomic technologies over the last decade, most objectives laid out in the CATG meeting (Gepts et al., 2005) were exceeded. The CATG plan called for the development of reference genome assemblies and resources for three species to be used as genomic and biological models: *Medicago truncatula*, *Lotus japonicus*, and *Glycine max* (soybean), and to further develop genomic tools for *Arachis hypogaea* (peanut) and *Phaseolus vulgaris* (common

Recommended citation: LGDWG (2019)

This paper is the product of the Legume Genomic Data Working Group, who discussed and debated topics presented here and are listed in alphabetical order. A subset of the members of the Working Group, indicated by Guillaume J. Bauchet, Kirstin E. Bett, Jacqueline D. Campbell, Ethalinda K.S. Cannon, Steven B. Cannon, Joseph W. Carlson, Agnes Chan, Timothy J. Close, Amanda M. Cooksey, Clarice Coyne, Rebecca Dickstein, Andrew D. Farmer, Samuel Hokin, Elizabeth S. Jones13, Maria J. Monteros, María Muñoz-Amatriaín, Chris Richards, Michael Udvardi, Eric Bishop von Wettberg, Nevin D. Young, assembled and edited the manuscript. All listed authors commented on and approved the final manuscript.

TABLE 1 Legume genomes sequenced to date

Common name	Latin name	Clade	Num	Quality	References
Soybean	*Glycine max*	Phaseoloid	4	High	(Schmutz et al., 2010; Shen et al., 2018; Shimomura et al., 2015; Valliyodan et al., 2019)
Wild soybean	*Glycine latifolia*	Phaseoloid	1	Medium	(Liu, Chang, Hartman, & Domier, 2018)
Wild soybean	*Glycine soja*	Phaseoloid	3	High	(Kim et al., 2010; Valliyodan et al., 2019; Xie et al., 2019)
Pigeon pea	*Cajanus cajan*	Phaseoloid	1	Medium	(Varshney et al., 2012)
Common bean	*Phaseolus vulgaris*	Phaseoloid	2	High	(Schmutz et al., 2014; Vlasova et al., 2016)
Mung bean	*Vigna radiata*	Phaseoloid	1	Medium	(Kang et al., 2014)
Adzuki bean	*Vigna angularis*	Phaseoloid	3	Medium	(Kang et al., 2015; Sakai et al., 2015; Yang et al., 2015)
Cowpea	*Vigna unguiculata*	Phaseoloid	1	High	(Lonardi et al., 2019)
Bambara groundnut	*Vigna subterranea*	Phaseoloid	1	Scaffs	(Chang et al., 2019)
Hyacinth bean	*Lablab purpureus*	Phaseoloid	1	Scaffs	(Chang et al., 2019)
Ji xue teng	*Spatholobus suberectus*	Phaseoloid	1	Medium	(Qin et al., 2019)
Jequirity bean	*Abrus precatorius*	Phaseoloid	1	Scaffs	https://www.ncbi.nlm.nih.gov/genome/74709
Velvet bean	*Mucuna pruriens*	Phaseoloid	1	Scaffs	https://www.ncbi.nlm.nih.gov/genome/71552
Birdsfoot trefoil	*Lotus japonicus*	Robinioid	1	Medium	(Sato et al., 2008)
Barrel medic	*Medicago truncatula*	Galegoid	17	High	(Pecrix et al., 2018; Tang et al., 2014; Young et al., 2011; Zhou et al., 2017)
Alfalfa	*Medicago sativa*	Galegoid	1	Medium	(Monteros et al., 2018)
Pea	*Pisum sativum*	Galegoid	1	High	(Kreplak et al., 2019)
Chickpea	*Cicer arietinum*	Galegoid	2	Medium	(Parween et al., 2015; Varshney et al., 2013)
Wild chickpea	*Cicer reticulatum*	Galegoid	1	Medium	(Gupta et al., 2017)
Wild chickpea	*Cicer echinospermum*	Galegoid	1	Scaffs	(von Wettberg et al., 2018)
Red clover	*Trifolium pratense*	Galegoid	1	Medium	(De Vega et al., 2015)
White clover	*Trifolium repens*	Galegoid	1	Medium	(Griffiths et al., 2019)
Subterranean clover	*Trifolium subterraneum*	Galegoid	1	Scaffs	(Hirakawa et al., 2016)
Narrow-leaved lupin	*Lupinus angustifolius*	Genistoid	1	Medium	(Hane et al., 2017)
Cultivated peanut	*Arachis hypogaea*	Dalbergioid	3	High	(Bertioli et al., 2019; Chen et al., 2019; Zhuang et al., 2019)
Wild peanut (A)	*Arachis duranensis*	Dalbergioid	1	Medium	(Bertioli et al., 2016)
Wild peanut (B)	*Arachis ipaensis*	Dalbergioid	1	Medium	(Bertioli et al., 2016)
Schott's yellowhood	*Nissolia schottii*	Dalbergioid	1	Scaffs	(Griesmann et al., 2018)
Partridge pea	*Chamaecrista fasciculata*	Mimosoid	1	Scaffs	(Griesmann et al., 2018)
Bull horn acacia	*Vachellia collinsii*	Mimosoid	1	Scaffs	https://www.ncbi.nlm.nih.gov/genome/82426
White carrob tree	*Prosopis alba*	Mimosoid	1	Scaffs	https://www.ncbi.nlm.nih.gov/genome/79095
Sensitive plant	*Mimosa pudica*	Mimosoid	1	Scaffs	(Griesmann et al., 2018)
Apple-ring acacia	*Faidherbia albida*	Mimosoid	1	Scaffs	(Chang et al., 2019)
Redbud tree	*Cercis canadensis*	Cercidoid	1	Scaffs	(Griesmann et al., 2018; Stai et al., 2019)

Note. The "Num" column gives the number of available genome assemblies known to the authors as of mid-2019. The "Quality" column gives an approximate indication of assembly completeness and contiguity: "Scaffs" indicating a scaffold-only assembly, not placed into chromosomal pseudomolecules; "High" generally indicating a highly contiguous, reference-quality assembly; and "Medium" indicating a more fragmented assembly.

bean). Genome assemblies have been generated for each of these species and for numerous other legumes (Table 1). Several species, such as cowpea and soybean, also have high-throughput genotyping platforms, which are being used to develop genotypic catalogs of germplasm collections and to support genomics-based trait development and breeding projects.

It is clear that some predictions from the CATG workshop were on target while others did not anticipate the rapid technological advances that would propel the data generation and analytical capabilities of our field. The recommendations were on target regarding completion of model genome assemblies and selected resources for comparative genomics. There has also been impressive progress in the availability of genetic and genomic resources through both generalist websites (GenBank, EBI) and specialist sites (Table 2). Further, there has been considerable tool development for exploration of legume genomic resources; a

TABLE 2 Web resources for legume genetics and genomics

Web resource	URL	Species
Phytozome	https://phytozome.jgi.doe.gov	Various
Legume Federation	https://www.legumefederation.org	Tools and links to other legume genomic resources
Legume Information System	https://legumeinfo.org	Adzuki bean, chickpea, common bean, Lotus, lupin, *Medicago truncatula*, mungbean, pigeonpea, red clover, cowpea, soybean, peanut, others
LIS InterMines	https://mines.legumeinfo.org	InterMine interface to most data at legumeinfo.org
Cool Season Food Legume database	https://www.coolseasonfoodlegume.org	Pea, lentil, chickpea, faba bean
Know Pulse	https://knowpulse.usask.ca	Chickpea, common bean, faba bean, lentil, pea
Kazusa Genome Database	http://www.kazusa.or.jp/genome	*Lotus japonicus*, clover (Trifolium)
LegumeIP	https://plantgrn.noble.org/LegumeIP	*Medicago truncatula*, soybean, Lotus, pigeonpea, chickpea, common bean
SoyBase	https://www.soybase.org	Soybean
PeanutBase	https://peanutbase.org	*Arachis hypogaea*, *Arachis ipaensis*, *Arachis duranensis*
Medicago truncatula genome database	http://www.medicagogenome.org	*Medicago truncatula*
INRA Medicago Bioinf. Resource	https://medicago.toulouse.inra.fr	*Medicago truncatula*
Medicago truncatula hapmap project	http://www.medicagohapmap.org http://www.medicagohapmap2.org	*Medicago truncatula*
Alfalfa Breeder's Toolbox	https://www.alfalfatoolbox.org	Alfalfa (*Medicago sativa*)
Vigna Genome Server	https://viggs.dna.affrc.go.jp	Vigna species: cowpea, mungbean, adzuki bean

Note. The general or clade-oriented resources are given first, followed by species- or genus-specific resources. See additional resources at https://www.legumefederation.org/en/species/.

current list is maintained at https://www.legumefederation.org/en/tools/.

In retrospect, the CATG recommendations were too conservative in terms of the advances possible in 5 to 10 years. A range of genome sequencing and assembly technologies has led to the sequencing of most crop and model legumes—currently (as of mid-2019) comprising nearly three dozen legume species and more than 60 genome assemblies (Table 1).

Another dynamic that was not predicted in the 2004 meeting was the effect of the highly energetic, largely independent international efforts focused on many legume model and crop species. While it might have been optimal to focus research and resource-development efforts on a few models, the institutional funding patterns and interests of national research projects have resulted in parallel genomics projects. For example, high-quality soybean assemblies have been generated in the U.S., Japan, and China (Table 1). There are similar stories for chickpea (the U.S./India and Canada), peanut (the U.S. and China), *M. truncatula* (the U.S. and France), and common bean (the U.S. and Mexico). In general, these parallel efforts have produced complementary resources, since genome assemblies are being developed from different accessions, and often using different technologies that generate assemblies and annotations with different characteristics.

Below, we summarize the main conclusions and recommendations of the 2019 meeting, following the organization of four thematic topics from the meeting.

2 | CROSS-CUTTING THEMES

2.1 | Data management, curation, and utilization

Genomic science has become increasingly data-centric—for example, through high-density genotyping, genome sequencing, and high-throughput phenotyping—but the value of data cannot be fully realized without careful attention to how it is processed and analyzed following the initial phases of generation. Increased sequencing capacity has led to development of intraspecific resources, adding additional complexities. This fast-paced data generation presents significant challenges for data curation, whose costs and complexity are often underestimated in grant-based research. Large, complex datasets require expert evaluation and often modification in order to make them widely useful. For example, genomic features such as genes, markers, and gene expression data are typically presented in a genome browser and added to systems for BLAST searches and other tools. These are relatively complex tasks requiring expert attention.

Data curation seeks to enable effective use of the developed genetic and genomic datasets, so that the information generated can be used more broadly to answer biological questions and facilitate practical breeding applications. In this regard, the use of FAIR data principles is crucial. FAIR is an acronym for: Findable (e.g., data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier); Accessible (e.g., data are deposited in a permanent, trusted repository); Interoperable (e.g., data are in a standard format; metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation); and Reusable (e.g., data and collections have clear usage licenses and provide accurate information on provenance; Wilkinson et al., 2016).

Data curation is critical for the quality and longevity of databases and repositories to ensure the maintenance of the research infrastructure for legume and plant biology research. Skilled data curation is critical for uploading datasets that meet thresholds for experimental design, adequate replication for statistical inference, and suitable protocols prior to broader use by the community.

Recommendations:

- Recognize data management and curation as a complex, high-level role, requiring specific and significant budgeting in proposals.
- Give specialized training to data curators, including in genomic technologies, data handling and analysis, and database management.
- Promote the usage of FAIR data standards. The data management component of grant proposals should include funding allocated to collecting and storing data according to FAIR standards, as well as the training of students and researchers in proper data handling and preparation methods.

2.2 | Training, outreach, and documentation to better integrate bioinformatics with plant biology and breeding

Bioinformatics, breeding tools and methods are often complex, relying on understanding of various applications and analysis workflows in Unix command-line environments, with understanding of statistical methods and algorithm limitations. Bioinformatics is often taught separately from, and with little integration with, plant biology and breeding. Rapid advances in the field require continued training for both students and practicing researchers. There is a need for further collaboration across multiple disciplines, and tools developed to utilize genomics-based approaches should streamline interoperability across multiple disciplines, species, and scales.

Recommendations:

- Foster increased communication between developers, researchers, collection curators, biologists, breeders, and other end-users, to ensure that resources and visualizations can address practical applications.
- Encourage annual meetings of PIs, postdoctoral fellows, and students from all plant disciplines and groups involved in breeding to discuss ideas, review available resources and identify research gaps and opportunities.
- Encourage video and static tutorials from bioinformatics developers when new tools are developed and released.

2.3 | General development needs and practices: Formats, standards, and computational resources

Several open-source frameworks for constructing and maintaining genomic websites and databases have matured, including: Tripal, https://tripal.info (Sanderson et al., 2013); InterMine, http://intermine.org (Kalderimis et al., 2014; Smith et al., 2012); and Solgenomics.org (Fernandez-Pozo et al., 2015). These frameworks are focused on the delivery of biological data and tools, integration and reuse of software components, implementation of standard web service APIs (Application Programming Interfaces), and flexible search queries.

Challenges include the difficulty of making user interfaces both easy to use and sufficiently powerful in terms of the scale and speed of analysis. Further, web technologies and underlying frameworks continue to change, and require ongoing software and security updates for any online database or website. Facilitating interoperability among websites adds an additional level of complexity to the software engineering challenge. Some of these objectives and challenges have been further described by the AgBioData consortium (Harper et al., 2018).

Recommendations:

- Promote greater use of standard APIs by genomic data portals to facilitate cross-site access and data sharing.
- Utilize common user interfaces and tools to ease the burden on users. Encourage standardization of visualization methods, data storage methods, and frameworks for making genetic and genomic data.
- Increase access to computational resources with sufficient storage, memory, and processor capacities for tasks such as genome assembly and annotation, with the capacity to support large-scale queries or exploration of whole genome comparisons among multiple species.
- Make analysis tools that can be used in stand-alone installations, e.g. through containerization, to enable use in regions with limited or unreliable internet access, such as in remote field locations and developing countries.

3 | GENOMIC DATA AND BIOLOGY

3.1 | Overview

It is now possible, with modest funding, to generate a high-quality draft genome assembly for most species (the exceptions being

unusually large genomes like faba bean, *Vicia faba*, or genomes with particular complexities, such as alfalfa, *Medicago sativa*, whose autotetraploid genome has made assembly extra challenging). Similarly, it has become relatively straightforward to generate gene predictions, transcriptome assemblies, RNA-seq atlases, large marker sets, and other next-generation sequencing-enabled datasets. After such a resource is generated, the best practice is to make it available in an accepted (and validated) format, with sufficient description, using established "minimum information standards" for the data, e.g., MIAME (Brazma et al., 2001) or MIAPA (Leebens-Mack et al., 2006), and then to deposit it in a permanent and public repository such as GenBank or a generalist repository with DOI-issuing capability such as Data Dryad or Zenodo. For most of these steps, format validators are useful but human curation is required to assess, describe, and often to correct the computational products.

The proliferation of datasets raises additional questions and challenges: How should multiple, related resources (e.g., multiple genome assemblies for a species) be handled? In what ways can they be usefully compared and integrated (e.g., into a pan-genome)? What kinds of metrics are most useful for describing the characteristics of a genomic resource? What standards of quality should be met before an analysis combining disparate resources is likely to yield insights into underlying biology rather than differences in technical approach? What types of evidence and time points should be considered to evaluate gene-specific differences in genotypes (e.g., susceptible vs. tolerant)?

An important factor in legume genomic biology is that of polyploidy (or whole-genome duplication; WGD)—both for genome assembly and for utilization. It appears that nearly all legume species have undergone at least one WGD since the time of their most recent common ancestor ~55–60 million years ago (Cannon et al., 2015; Stai et al., 2019). Legume crops that have only the papilionoid WGD, such as chickpea, common bean, and pea, behave as diploids genetically. Species with more recent WGD include peanut (~10 thousand years; Bertioli et al., 2019), white clover (~15–28 thousand years; Griffiths et al., 2019), soybean (~13 million years ago; Schmutz et al., 2010), and lupin (triploidy ~27 million years ago; Hane et al., 2017). For the more recent polyploids, the higher ploidy impacts genomic studies, genetic studies (making it more difficult to identify mutants) and breeding strategies (e.g., requiring dosage-sensitive genotyping assays in autotetraploid alfalfa). An intriguing finding from the tetraploid peanut genome is that an important source of genetic diversity is from dosage changes in alleles, as genes or regions between subgenomes convert or invade the other subgenome (Bertioli et al., 2019).

3.2 | Assembly and annotation quality and consistency

Quality metrics for reference genome assemblies and their annotations are often not provided by the sites housing them, and current assembly statistics for quality are generally insufficiently informative. In particular, "reference" and "draft" are poorly defined, and measures like N50 and L50 are inconsistently or interchangeably used. Efforts to define lineage-appropriate core gene families for assessment of annotation completeness are useful, e.g. BUSCO (Waterhouse et al., 2017) and coreGF (Bel et al., 2012; Veeckman, Ruttink, & Vandepoele, 2016), but interpretation of results using these standards is complicated by the history of polyploidy with legumes. Alignment of genome assemblies with genetic linkage maps or optical maps can provide an additional resource for comparisons and can be used to identify contigs or scaffolds that correspond to the same chromosome or help identify chimeric assemblies.

Not every assembly or annotation needs to be of the highest quality; a fragmentary draft assembly may be sufficient to align sequencing reads to identify sequence variants. Nevertheless, low-quality assemblies and annotations can cause problems if inappropriately integrated into other analyses—for example, including rough annotations in which chimeric, fragmentary, haplotypes, or transposon gene-calls may "pollute" downstream analyses such as gene families. Similarly, the characterization of genes associated with lineage-specificity or their status as a "core" or "dispensable" gene within the context of a pan-genomic analysis will depend heavily on the completeness of the individual datasets used in the determination.

(Salzberg, 2019). There is also recognition that standard genome assembly and annotation methods don't capture all the features of interest—for example, methylation status, chromatin features, or recombination hotspots (Mei, Stetter, Gates, Stitzer, & Ross-Ibarra, 2018).

Recommendations:

- Improve metrics for assembly and annotation quality and standardize methods for applying the metrics across multiple legume species.
- Increase consistency and comparability for assembly and annotation tools. This will need to be an ongoing effort, as technologies continue to evolve.
- Increase support and standardization for "unusual features" such as small-RNAs, structural-genomic features such as chromatin accessibility, methylation, and epigenetic gene regulation. Further, consider distinct features of particular genomes such as heterozygosity, ploidy levels, disease resistance hotspots, and transposable elements in large genomes.

3.3 | Pan-genomes: Supporting multiple reference genomes for a single species

For species that have had multiple genomes sequenced and assembled, characterization of the pan-genome will be important for representing the gene complement and sequence diversity present in a species and to avoid biases that might be introduced in comparisons against a single accession. Similar to the problems inherent in constructing gene families or other types of cross-species comparative analysis from inputs of variable quality, pan-genome-based inferences will be dependent on the quality of the data, the diversity of

genotypes sequenced, and the nature of the inputs used to construct them. For example, many aspects of diversity can be addressed by resequencing data and pan-genomes can be constructed from genotypes inferred from such data. However, low coverage resequencing would be of limited value to assess regions of complex structural variation, including those found in rapidly evolving resistance-gene clusters in plant genomes.

Pan-genomes may be represented by a graph data structure, in which each accession is represented by a separate path (Eggertsson et al., 2017; Computational Pan-Genomics Consortium, 2018). In contrast, a single reference assembly has a simple linear coordinate system. The representation of gene complements in pan-genomes remains an area of active research and discussion. Other open questions involve approaches for naming features in a pan-genome, for updating and versioning a pan-genome as additional assemblies become available, and for accessing and extracting information. Opportunities exist to optimize the construction, representation and analysis of pan-genomes—for example, to provide useful information to help breeders identify plants for crossing to generate optimal gene combinations, alleles, and complementary genes. A recent use-case is tomato, in which a pan-genome analysis helped identify traits in tomato domestication (Gao et al., 2019).

Recommendations:

- Adopt and adapt emerging methods and data standards to handle pan-genomes leveraging examples from other communities (such as those developed for humans, *Drosophila*, *Arabidopsis*, and *Brachypodium*).
- Foster increased communication with other communities who have or are developing pan-genomes. Coordinate with developers, breeders and other end-users, to ensure that pan-genome visualizations can address practical applications.

3.4 | Functional, expression, and comparative resources

Transcriptome datasets can provide a reference for expressed genes in different genotypes, organs, tissues, and cell types throughout plant development and under "ideal" or environmental or biotic stress conditions. Gene expression datasets compared against a common framework can facilitate comparisons across experiments, genotypes, and species. Gene families, whether derived from transcriptome assemblies or genome assemblies, enable identification of ortholog sets, creating a bridge between species. Such ortholog sets are available for most crop and model legumes at https://legumeinfo.org.

Microfluidic and droplet-based technologies for assaying expression and other functional genomics data at the level of individual cells have recently been applied to *Arabidopsis* roots in several studies (Denyer et al., 2019; Jean-Baptiste et al., 2019; Ryu, Huang, Kang, & Schiefelbein, 2019), suggesting that the plant community as a whole will soon be in a position to benefit from these exciting technologies previously inaccessible to them due to limitations on ready dissociability imposed by the plant cell wall.

There is also a need for continued inter-specific comparative analysis. Several legume genera have multiple domestications (e.g., *Vigna*, *Phaseolus*, and *Lupinus*), but in general, it is not known whether the molecular causes underlying these domestication syndromes are shared or species-specific. A number of traits have evolved multiple times across the family, such as geocarpic fruits (e.g., in *Arachis*/peanut and *Vigna*/bambara groundnut), tubers (e.g., in *Phosphocarpus*/winged bean, *Tylosema*/marama bean, and *Apios*/potato bean), or transitions between growth forms (e.g., tree and herb forms in many lineages). As more genomic resources become available across the legume family, comparative analysis of the repeated evolution of such traits will be increasingly tractable and powerful. For species with limited funding opportunities, the ability to leverage genetic information from better characterized plant relatives is invaluable.

Recommendations:

- Create comprehensive gene expression atlases for all crop legumes and make them available through websites with long-term support. These atlases are important for defining lists of conserved candidate genes across species or those unique to a certain species, and for generating gene-based markers associated with specific traits.
- Develop standards for replications, growing conditions, and tissue types to establish reference gene atlas databases for the various legume species.
- Improve documentation of annotation methods and evidence to support gene annotations and strategies to facilitate cross-species comparisons.
- Develop methods and standards for naming annotated genes that take into account gene families, haplotypes, and pan-genomes.
- Use established "minimum information" (Brazma et al., 2001) standards for describing gene expression datasets.
- Develop an encyclopedia of genes that underlie domestication and quality traits for important legume species.

4 | GENOTYPE AND PHENOTYPE MEASUREMENTS AND ASSOCIATIONS: ONTOLOGIES

4.1 | Overview

The ability to efficiently and densely genotype many accessions enables powerful analyses. The data can be used to determine genetic relationships among and within accessions, to determine population structure in plant genetic resource collections, and to make informed decisions about managing a collection based on genetic diversity. In breeding applications, markers with known associations can be used for marker-assisted selection, to identify novel genetic variants within genes, and for genomic selection (GS).

For many legume crops, gene banks contain thousands of accessions which can be narrowed down to a subset of core accessions that represent the genetic diversity of the entire collection, to facilitate phenotyping for traits that are time-consuming or labor intensive to evaluate. Genetic information can be used to help identify redundancies or gaps in the germplasm to inform plant exploration or new plant genetic resource collection initiatives. Combined with genotypic data, phenotypic information about a core set of accessions or in some cases, the entire collection of a species, can be used in genome-wide association studies (GWAS) or if a key gene for a specific trait (e.g., disease resistance) exists, to identify novel alleles for the gene(s) of interest in the accessions evaluated.

4.2 | Data comparison between genotyping platforms

Data collected from different genotype array platforms can be difficult to compare or for combined analyses. Cross-platform comparisons require shared variants between any two platforms, which may be limited to a small portion of the markers. Even for common variants, the design for different genotyping platforms may not be on the same DNA strand. In that case, even when using the same genotyping platform, merging and comparing datasets can be difficult. Data formats may differ, genotype identifiers may not be comparable, the reporting DNA strand may not be uniform, and markers may have been designed on different versions of genome assemblies. Open-source genotyping data management systems such as GOBii, http://gobiiproject.org and Gigwa, https://southgreen.fr/content/gigwa (Sempéré et al., 2016) can be used to load and organize data with varying formats, so that data can then be extracted with a common data format and markers across platforms can be aligned to common variants. Imputation methodologies can be applied to align different genotyping methodologies to standard sets of markers, and variants and markers generated from differing genotyping methodologies can be aligned to common reference genome variants, although this requires significant bioinformatics resources imputation platforms (Wang et al., 2018). Haplotype graphs can reduce this complexity and consolidate the information from different genotyping technologies into common haplotypes. The process of imputing SNPs may be easier to implement in some species vs. others considering mode of reproduction (self vs. outcross) and ploidy level.

Recommendations:

- Increase comparability across genotyping platforms. Use common SNP/feature sets where possible, and make genotyping results available in well-established formats, with validation. Improve imputation methods for inferring missing data points when compared to common SNP databases.
- Increase the use of genotyping data management systems to help align data formats and outputs to facilitate comparison between experiments or genotyping runs. These systems should associate markers with their underlying variants wherever possible, and document design and reporting strands to enable transformation to a common allele strand output.
- Explore tools and procedures that facilitate consolidation of different marker technologies, such as methods that enable imputation to a common variant or haplotype set. Provide a central SNP and genetic variant repository for legumes, whether provided by a member of the INSDC (International Nucleotide Sequence Database Collaboration) or by the legume community itself.

4.3 | Expansion beyond single nucleotide to structural variations

Sequence variant analysis needs to expand from studying SNPs and small insertions and deletions (INDELs) to include larger structural variations (SVs) such as inversions, translocations, and copy-number variations (CNVs) including presence–absence variations (PAVs), which can significantly affect phenotypes. For example, in soybean a CNV at *Rhg1* increases resistance to the cyst nematode *Heterodera glycines* (Cook et al., 2012). SVs can also be linked to domestication events (Lye & Purugganan, 2019). A comprehensive characterization of structural variation in legume genomes would require the use of long-read sequencing technologies and the development of de novo assemblies to resolve highly repetitive regions and to eliminate reference bias.

Recommendations:

- Expand variant analysis to include complex variations: presence–absence variations (PAVs), copy-number variations (CNVs), and larger structural variations (SVs).
- For particularly large or repeat-dense species, such as faba, pea, and lentil, utilize exome targeted sequencing methods such as exome capture for identifying PAVs and CNVs.

4.4 | GWAS: Meta-analysis, resources, and repositories

To better facilitate the combination of GWAS studies in plants, phenotypic characterizations need to be comparable (Zhao et al., 2019). This is difficult if the terminology used to describe traits and trait variation differs. Use of consistent, common descriptors, or ontologies is therefore necessary (Shrestha et al., 2010). Although the use of ontologies by research communities over the last decade has expanded, further developments are required (Ćwiek-Kupczyńska et al., 2016; Krajewski et al., 2015; Walls et al., 2019). Ontologies for several legume crops (e.g., chickpea, faba, lentil, and soybean) are relatively well-developed and can be accessed at http://www.cropontology.org.

Also, improved data management and sharing of crop genotype, phenotype, and GWAS datasets are needed to facilitate their integration and use. A need exists to develop open source databases and repositories for genotype and phenotype datasets coupled with easy-to-use GWAS pipelines with embedded standardized genotype data to identify SNP-trait association in legume crops. A similar concept

and approach was developed for the model plant *A. thaliana* and other plants, e.g., AraGWAS (Togninalli et al., 2018), AraPheno (Seren et al., 2017), easyGWAS (Grimm et al., 2017). Furthermore, as GWAS data is often scattered across publications, there is a need for a centralized repository of GWAS results (e.g., human GWAS catalog) which follows data standards for studies, traits, variants, accessions, and base pair locations.

Recommendations:

- Facilitate cross-study comparisons by using consistent, common descriptors, from established ontologies.
- Extend ontologies from one legume to others, to facilitate the transferability of information from one legume to another.
- Encourage utilization of ontology descriptors described in other crops, and the use of standardized ontologies maintained at http://www.cropontology.org.
- Develop a centralized repository of GWAS results, with rigorous adherence to data standards.
- Develop tools to facilitate cross-species and within-species comparisons and meta-analyses across multiple studies, which can in turn enable comparison of GWAS features across studies and enable identification of causal genetic elements.

5 | BREEDING APPLICATIONS

5.1 | Overview

Breeding progress can be measured through genetic gains per cycle of selection. Management of data for screening materials is critical to catalog the genetic diversity and to make selection decisions. Easy access to a wide range of data, from sources across many disciplines, would facilitate decision-making at every step of a breeding program. Additionally, there is a potential to expand breeding from single crop selections to include breeding for the most favorable interactions with rhizobia and other microbes, for improved symbiotic performance and survival in different soils (Busby et al., 2017; Greenlon et al., 2019).

5.2 | Data collection and integration for breeding needs

For genetic resource collections, users need access to information such as pedigree information, trait phenotyping, researcher attribution, links to genomic data, and a variety of other germplasm data. The Germplasm Resources Information Network (GRIN), maintained by the USDA-ARS, is the online germplasm database for the U.S. National Plant Germplasm System. GRIN maintains a catalog of public germplasm resources at https://npgsweb.ars-grin.gov/gringlobal/search.aspx and contains phenotypic and agronomic data for germplasm held in gene banks in the US and worldwide. A need exists for increased investment in germplasm databases to maximize their use. For example, the trait data for the accessions held in GRIN are not always complete, having been collected at different time-points and under different growing conditions. Descriptors are not easily integrated with other datasets due to the lack of automated download and bulk access.

For discovery research, plant databases hold information on reference genomes, as well as gene and protein annotations (structural and functional), along with quantitative trait loci (QTL), marker-trait associations, genes, and phenotypes. However, many datasets remain buried in research papers, stored in lab computers, or are otherwise unavailable for the broader research community. Varying data formats, different methods of access (web services, bulk downloads, data displayed on web page only, or requiring a log-in), and inconsistent use of metadata and ontologies, make it difficult if not impossible to integrate multiple sets of related data.

For variety development, data management systems need to focus on the rapid collection and aggregation of phenotypic and genotypic data to deliver decision support to breeders within the tight time frames of a breeding cycle (Lulsdorf & Banniza, 2018). Breeding systems need to support pedigree management, seed source inventory, cross tracking, screening trial data, and be connected with decision support tools for activities such as marker-assisted backcrossing and genomic selection. The ability to update prediction models on allele substitution, estimated breeding values and multitrait selection indices as new data from performance trials is produced can increase breeding efficiencies. The Bill and Melinda Gates Foundation has supported the development of systems to improve crop yields in developing countries through funding for the open-source Breeding Management System (https://www.integratedbreeding.net/15/breeding-management-system) and the Cassavabase community of databases for clonally propagated crops. These systems have been expanded to incorporate open-source field-based data collection tools such as Field Book (Rife & Poland, 2014), large-scale genomics data management (GOBii; gobiiproject.org), and decision support tools such as Flapjack (Milne et al., 2010), OptiMAS (Valente et al., 2013), and genomic selection pipelines (Tecle et al., 2014), including those available at http://galaxy-demo.excellenceinbreeding.org.

Recommendations:

- Use standard ontologies, such as the Crop Ontologies (http://www.cropontology.org/) and Plant Trait Ontology (TO; http://www.obofoundry.org/ontology/to.html) to enable comparisons across datasets. Contribute terms as needed to facilitate their use for practical breeding applications.
- Expose and retrieve data through standard web service APIs. The Breeding API, BrAPI (https://brapi.org/) should be implemented and contributions made to keep standards current (Selby et al., 2019).
- Establish standard protocols for phenotypic traits and data collection standards to populate GRIN-Global as the main repository for germplasm and phenotypic descriptors. Use web services for integration with related data.

- Develop tools and standard interfaces that meet breeder use cases (haplotype mining, identification of potentially useful parents/alleles), and the ability to access data from multiple databases (germplasm maintenance and discovery research).
- Researchers should be encouraged to utilize appropriate long-term repositories with commitments to FAIR data principles to facilitate data integration.

5.3 | Breeding data search, acquisition, browsing, visualization, and analysis tools

To enable better access to data, improved tools to find, view, analyze, and acquire data are required. Having to learn different tools and navigation at different websites, along with the need to combine multiple online and stand-alone applications is inefficient and burdensome. To avoid duplicated development efforts and to provide more consistent website navigation, a number of open source frameworks are available.

Recommendations:

- Improve tools to easily find germplasm using a variety of search filters, including geographic origin, traits, marker alleles, and heterotic groups.
- Implement common frameworks to the extent possible and collaborate with others doing similar work; develop initiatives to foster collaborations and leveraging of resources to avoid operating in silos.
- Adapt tools developed by other research communities where possible to address new projects and needs in other legumes.
- Promote comparative analysis tools such as the Genome Context Viewer (Cleary & Farmer, 2018) available at https://legumeinfo.org to evaluate legume gene families and phylogenies.

5.4 | Breeding management tools

Although some breeding tools and management systems exist and continue to be actively developed, some still lack the full functionality required by breeders. This includes support for flexible plot designs and different modes of reproduction (selfing vs. outcrossing), visualization of SNP haplotypes for comparisons of closely related entries, and the ability to define variable thresholds for multiple traits simultaneously as part of a breeding selection index. Breeding databases to date do not have the ability to manage large scale SNP data that can be rapidly extracted into downstream tools. More recently, the GOBii (gobiiproject.org) Genomic Data Management system has been designed for this purpose and can be integrated with any Breeding management systems.

Recommendations:

- Link and augment existing breeding management tools to include additional breeder-centric functionalities (support flexible plot designs and modes of reproduction).
- Integrate breeding management tools with data-rich informatics resources focused on climate, soil, and weather to better address GxE interactions.
- Improve breeding decision support tools that consider multiple traits simultaneously

5.5 | Gene bank collections

Diverse collections of plant genetic resources are critical for breeders because they archive the diversity on which trait improvement depends. Maintenance and characterization of germplasm collections is essential to maximize their utility for addressing current and emerging threats to agricultural productivity.

Characterizing germplasm and gaining easy access to the data continues to be a challenge. As previously mentioned, some of these data are available through GRIN-Global, but are sparse, not easily accessible or downloadable in bulk and not integrated with data maintained in other databases. The trait terms used in GRIN for legume crops have generally not been linked to widely used trait ontologies such as the Plant Trait Ontology (TO; http://www.obofoundry.org/ontology/to.html) and the crop-specific Crop Ontologies (http://www.cropontology.org). Additional trait data are scattered across data resources, lab servers, and in publications in multiple journals.

There is enormous potential for genetic and genomic data to improve the quality and integrity of plant genetic resource management. The prospect of genotyping core or complete germplasm collections is very appealing, as this information can be a powerful tool both for discovering unique alleles and for improving access to collections by assessing collection genetic diversity. In addition, core collections can serve as a diversity panel in evaluating incoming materials for redundancy and for identifying gaps in collections. Documenting an explicit linkage between a physical accession and the derived genotypic sequence information is challenging for a number of reasons. For example, when genotyping germplasm collections, some collections (accessions) are highly heterogeneous (e.g., through mixed seed lots) or heterozygous (e.g., in hybrids, outcrossing, or allozygous crops). Attempts at surveying an entire collection for genotypic variation have often derived purified, single seed descent samples that serve as the basis for sequencing. In addition, the identity of duplicate accessions among global collections can be difficult to detect because gene banks have different synonyms for the same accession and cross referencing metrics are insufficient. There have been discussions within the U.S. National Plant Germplasm System to apply globally adopted unique digital object identifiers (DOIs) to germplasm accessions but various technical challenges remain.

Recommendations:

- Archive collections in central gene banks and maintain them in sufficient quantity to fulfill seed requests.
- Prioritize genotyping of core germplasm collections for multiple legumes using standard sampling protocols, while also considering

the challenge of maintaining and increasing seed lots. Successful examples of this are projects to genotype the entire barley collection (Milner et al., 2019) and the entire U.S. soybean collection (Song et al., 2015). Genotyping entire collections will likely result in the development of new core collections that better represent the genetic diversity of the collection.

- Perform detailed phenotyping of accessions in core collections using standardized protocols for plant growth, replications, locations, and trait ontologies.
- Develop tools to visualize and mine genomics datasets in gene bank collections to increase their application for trait development, prebreeding and breeding purposes.
- Link GRIN descriptors for all legume crops to reference ontologies, such as the Plant Trait Ontology (TO).
- Apply stable data identifiers (DOIs) to germplasm collections and develop recommendations for applying them to heterozygous and heterogeneous seed lots.

5.6 | Rhizobia and other symbionts

The capacity to actively select rhizobial strains using genomic information has improved substantially since the early 2000s. Although legume inoculation as an agronomic practice is nearly a century old, challenges remain in developing rhizobia that are both effective nitrogen-fixing symbionts for particular legume crops and that have high survival in agronomic soils. Rhizobia developed for inoculation in controlled environments have very low survival and performance in real field soils when in competition with the local strains. Researchers would benefit from systematic sequence characterization of full rhizobia collections. Improved technologies to transfer rhizobia or other microbes in seed coatings prior to planting could greatly increase gains by enhancing plant establishment or access to water or nutrients through symbiotic interactions (Ghimire, Charlton, & Craven, 2009). Effective interactions between legumes and symbionts can enhance establishment, drought and salinity tolerance, performance in poor soils, and disease resistance that ultimately could result in higher yields.

Recommendations:

- Systematically improve rhizobial and other symbionts and determine legume breeding lines responsive to these improved symbionts.
- Expand research and development of tools to discover, modify and utilize knowledge about genome-genome interactions between rhizobia/mycorrhizae and legume hosts in agricultural settings.

6 | LEGUME-SPECIFIC DATA RESOURCE OPPORTUNITIES

Rhizobial data resources highlight the fact that many aspects of legume biology are distinctive, calling for either novel or taxon-specific approaches to genomic data management. In the area of annotation, legume genomes including *Medicago truncatula* have played a key role in the elucidation of small secreted protein genes (de Bang et al., 2017). These proteins play key roles in nodule and rhizobial development and hundreds have been discovered across multiple legume genomes. Nevertheless, small proteins are routinely overlooked in genome sequencing and annotation projects. Work in legume genomics has encouraged the development of small secreted protein gene discovery software (Zhou et al., 2013) as well as the reexamination of sequenced plant genomes, including the genomes of nonlegumes (Silverstein et al., 2007). Still, developing standardized conventions for description and naming of short or unusual open reading frame (ORFs) remains a work in progress. Likewise, in mining the most recent version of the *M. truncatula* genome, Pecrix et al. (2018) uncovered hundreds of long, noncoding (lnc) RNAs with apparent roles in nodulation. Exploring these across legume species will require establishing better annotation, nomenclature, orthology relationships, and functional characterization for such unusual genomic elements.

Beyond nodulation, legumes share important taxon-specific data opportunities that must ultimately elucidated from the distinctive lens of legume species. First generation pan-genomes for soybean and *Medicago* have already been developed (Li et al., 2014; Zhou et al., 2017). Soon, much deeper and more extensive pan-genomes will be publicly available in these and other legume species. Likewise, work to define the "ancestral" legume genome and its evolution into present day species has pulled together genomic data across multiple legume sequencing projects (Kreplak et al., 2019; Ren, Huang, & Cannon, 2019; Wang et al., 2017). The data management and sharing standards, especially annotation, orthology, genomic elements, complex variation, haplotypes, and more, are all incomplete and urgently needed to exploit these pan-species and pan-family genome resources. At the individual species level, unique or novel data challenges remain for legume species. The recent publication of the peanut genome (Bertioli et al., 2019; Zhuang et al., 2019) highlights unusual features of subgenome evolution/domestication, while the pea genome (Kreplak et al., 2019) illustrates the impact of massive transposon expansion. For both, data descriptions and standards are in their infancy for their use in legume genomics. Even among long-studied legume traits such as yield, quality, and stress-tolerance (including protein and oil in soybean and pulses, forage quality and winter-hardiness in alfalfa and clover, and pathogens targeting multiple legume species), strategies to enable actionable decision-making by breeding and germplasm researchers still require integration with sequence- and genome-level datasets to be completed.

7 | CONCLUSIONS

The state of legume genomics is characterized by an abundance of data, which offers many opportunities for comparison and combination of datasets. Productive integration and comparison requires data

management practices and methods that have yet to keep up with the pace at which data are being generated. Such standards should include use of consistent metadata, ontologies, and accepted and validated data formats, and deposition of data in well-supported and maintained repositories to facilitate their use. A critical need exists for data curators, improved computational tools and interfaces targeting human end-users, and computer access via APIs. Data generators, curators, software developers, and users of the data should approach the generation of these resources while being mindful of the long-term goals of these efforts: to improve our understanding of legume biology, including interactions with rhizobial symbionts and other biotic and abiotic factors, to promote the efficient stewardship and utilization of legume genetic resources, and to optimize legume improvement for the benefit of farmers and consumers.

ACKNOWLEDGEMENTS

The authors thank legume researchers and data scientists who have shaped the opinions and recommendations of the LGDWG.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Writing – original draft and review and editing: JC, EC, SC, AF, SH. Writing – original draft: GB, KB, JC, AC, TC, AC, CC, RD, EJ, MJM, MM-A, CR, MU, Ev-W, NY. Conceptualization (participated in LGDWG discussions and on initial writing teams): CC, AC, DC, SD, DF, YK, KM, CP, AS, CT, PZ. Project administration: DF-B, CT, AC, AF. Funding acquisition: DF-B, MU. All authors read and approved the final manuscript.

FUNDING INFORMATION

This research was funded by the NSF project "Federated Plant Database Initiative for the Legumes" (1444806), and the USDA Agricultural Research Service project 5030–21000-062-00D. The USDA is an equal opportunity provider and employer. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH112.

DATA AVAILABILITY STATEMENT

None.

ORCID

Steven B. Cannon https://orcid.org/0000-0003-2777-8034

REFERENCES

Boschiero, C., Dai, X., de Bang, T. C., Lundquist, P. K., Pant, P., … Scheible, W. R. Zhuang, Z.,(2017). Genome-wide identification of *Medicago* peptides involved in macronutrient responses and nodulation. Plant Physiology, 175(4), 1669–1689. https://doi.org/10.1104/pp.17.01096

Bel, M. V., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., & Vandepoele, K. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiology, 158(2), 590–600. https://doi.org/10.1104/pp.111.189514

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., … Ozias-Akins, P. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nature Genetics, 48(4), 438–446. https://doi.org/10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., … Schmutz, J. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. Nature Genetics, 51(5), 877–884. https://doi.org/10.1038/s41588-019-0405-z

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., … Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data. Nature Genetics, 29(4), 365–371. https://doi.org/10.1038/ng1201-365

Busby, P. E., Soman, C., Wagner, M. R., Friesen, M. L., Kremer, J., Bennett, A., … Dangl, J. L. (2017). Research priorities for harnessing plant microbiomes in sustainable agriculture. PLoS Biology, 15(3), e2001793. https://doi.org/10.1371/journal.pbio.2001793

Cannon, S. B., McKain, M. R., Harkess, A., Nelson, M. N., Dash, S., Deyholos, M. K., … Leebens-Mack, J. (2015). Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Molecular Biology and Evolution, 32(1), 193–210. https://doi.org/10.1093/molbev/msu296

Chang, Y., Liu, H., Liu, M., Liao, X., Sahu, S. K., Fu, Y., … Liu, X. (2019). The draft genomes of five agriculturally important African orphan crops. GigaScience, 8(3). https://doi.org/10.1093/gigascience/giy152

Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., … Liang, X. (2019). Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. Molecular Plant, 12(7), 920–934. https://doi.org/10.1016/j.molp.2019.03.005

Cleary, A., & Farmer, A. (2018). Genome context viewer: Visual exploration of multiple annotated genomes using microsynteny. Bioinformatics (Oxford, England), 34(9), 1562–1564. https://doi.org/10.1093/bioinformatics/btx757

Computational Pan-Genomics Consortium (2018). Computational pan-genomics: Status, promises and challenges. Briefings in Bioinformatics, 19(1), 118–135. https://doi.org/10.1093/bib/bbw089

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., … Bent, A. F. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science, 338(6111), 1206–1209. https://doi.org/10.1126/science.1228746

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., … Krajewski, P. (2016). Measures for interoperability of phenotypic data: Minimum information requirements and formatting. Plant Methods, 12(1), 44. https://doi.org/10.1186/s13007-016-0144-4

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., … Lang, C. (2015). Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Scientific Reports, 5, 17394. https://doi.org/10.1038/srep17394

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., & Timmermans, M. C. P. (2019). Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. Developmental Cell, 48(6), 840–852.e5. https://doi.org/10.1016/j.devcel.2019.02.022

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., … Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. Nature Genetics, 49(11), 1654–1660. https://doi.org/10.1038/ng.3964

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., … Mueller, L. A. (2015). The sol genomics network (SGN)—From genotype to phenotype to breeding. Nucleic Acids

Research, 43(Database issue), D1036–D1041. https://doi.org/10.1093/nar/gku1195

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., … Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nature Genetics, 1, 1044–1051. https://doi.org/10.1038/s41588-019-0410-2

Gepts, P., Beavis, W. D., Brummer, E. C., Shoemaker, R. C., Stalker, H. T., Weeden, N. F., & Young, N. D. (2005). Legumes as a model plant family. Genomics for food and feed report of the Cross-Legume Advances Through Genomics Conference. Plant Physiology, 137(4), 1228–1235. https://doi.org/10.1104/pp.105.060871

Ghimire, S. R., Charlton, N. D., & Craven, K. D. (2009). The mycorrhizal fungus, Sebacina vermifera, enhances seed germination and biomass production in Switchgrass (*Panicum virgatum* L). Bioenergy Research, 2(1–2), 51–58. https://doi.org/10.1007/s12155-009-9033-2

Greenlon, A., Chang, P. L., Damtew, Z. M., Muleta, A., Carrasquilla-Garcia, N., Kim, D., … Cook, D. R. (2019). Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. Proceedings of the National Academy of Sciences, 116, 15200–15209. https://doi.org/10.1073/pnas.1900056116

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M. B., … Cheng, S. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science, 361(6398), eaat1743. https://doi.org/10.1126/science.aat1743

Griffiths, A. G., Moraga, R., Tausen, M., Gupta, V., Bilton, T. P., Campbell, M. A., … Anderson, C. (2019). Breaking free: The genomics of allopolyploidy-facilitated niche expansion in white clover. Plant Cell, 31(7), 1466–1487. https://doi.org/10.1105/tpc.18.00606

Grimm, D. G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., … Borgwardt, K. M. (2017). easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies. Plant Cell, 29(1), 5–19. https://doi.org/10.1105/tpc.16.00551

Gupta, S., Nawaz, K., Parween, S., Roy, R., Sahu, K., Kumar Pole, A., … Chattopadhyay, D. (2017). Draft genome sequence of Cicer reticulatum L., the wild progenitor of chickpea provides a resource for agronomic trait improvement. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes, 24(1), 1–10. https://doi.org/10.1093/dnares/dsw042

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

Harper, L., Campbell, J., Cannon, E. K. S., Jung, S., Poelchau, M., Walls, R., … Main, D. (2018). AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database, 2018. https://doi.org/10.1093/database/bay088

Hirakawa, H., Kaur, P., Shirasawa, K., Nichols, P., Nagano, S., Appels, R., … Isobe, S. N. (2016). Draft genome sequence of subterranean clover, a reference for genus Trifolium. Scientific Reports, 6, 30358. https://doi.org/10.1038/srep30358

Jean-Baptiste, K., McFaline-Figueroa, J. L., Alexandre, C. M., Dorrity, M. W., Saunders, L., Bubb, K. L., … Cuperus, J. T. (2019). Dynamics of gene expression in single root cells of *Arabidopsis thaliana*. Plant Cell, 31(5), 993–1011. https://doi.org/10.1105/tpc.18.00785

Kalderimis, 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(W1), W468–W472. https://doi.org/10.1093/nar/gku301

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B. K., … Lee, S. H. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. Nature Communications, 5, 5443. https://doi.org/10.1038/ncomms6443

Kang, Y. J., Satyawan, D., Shim, S., Lee, T., Lee, J., Hwang, W. J., … Lee, S. H. (2015). Draft genome sequence of adzuki bean, *Vigna angularis*. Scientific Reports, 5, 8069. https://doi.org/10.1038/srep08069

Kim, M. Y., Lee, S., Van, K., Kim, T.-H., Jeong, S.-C., Choi, I. Y., … Kim, W. Y. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proceedings of the National Academy of Sciences of the United States of America, 107(51), 22032–22037. https://doi.org/10.1073/pnas.1009526107

Krajewski, P., Chen, D., Ćwiek, H., van Dijk, A. D. J., Fiorani, F., Kersey, P., … Weise, S. (2015). Towards recommendations for metadata and data handling in plant phenotyping. Journal of Experimental Botany, 66(18), 5417–5427. https://doi.org/10.1093/jxb/erv271

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G., … Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. Nature Genetics, 51(9), 1411–1422. https://doi.org/10.1038/s41588-019-0480-1

Leebens-Mack, J., Vision, T., Brenner, E., Bowers, J. E., Cannon, S., Clement, M. J., … Zmasek, C. (2006). Taking the first steps towards a standard for reporting on phylogenies: Minimal information about a phylogenetic analysis (MIAPA). Omics: A Journal of Integrative Biology, 10(2), 231–237. https://doi.org/10.1089/omi.2006.10.231

Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L., Zhang, Z., … Qiu, L. J. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology, 32(10), 1045–1052. https://doi.org/10.1038/nbt.2979

Liu, Q., Chang, S., Hartman, G. L., & Domier, L. L. (2018). Assembly and annotation of a draft genome sequence for Glycine latifolia, a perennial wild relative of soybean. The Plant Journal, 95(1), 71–85. https://doi.org/10.1111/tpj.13931

Lonardi, S., Muñoz-Amatriaín, M., Liang, Q., Shu, S., Wanamaker, S. I., Lo, S., … Alhakami, H. (2019). The genome of cowpea (*Vigna unguiculata* [L.] Walp.). The Plant Journal, 98, 767–782. https://doi.org/10.1111/tpj.14349

LPWG (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). Taxon, 66(1), 44–77. https://doi.org/10.12705/661.3

Lulsdorf, M. M., & Banniza, S. (2018). Rapid generation cycling of an F2 population derived from a cross between Lens culinaris Medik. and Lens ervoides (Brign.) Grande after aphanomyces root rot selection. Plant Breeding, 137(4), 486–491. https://doi.org/10.1111/pbr.12612

Lye, Z. N., & Purugganan, M. D. (2019). Copy number variation in domestication. Trends in Plant Science, 24(4), 352–365. https://doi.org/10.1016/j.tplants.2019.01.003

Mei, W., Stetter, M. G., Gates, D. J., Stitzer, M. C., & Ross-Ibarra, J. (2018). Adaptation in plant genomes: Bigger is different. American Journal of Botany, 105(1), 16–19. https://doi.org/10.1002/ajb2.1002

Meyer, R. S., DuVal, A. E., & Jensen, H. R. (2012). Patterns and processes in crop domestication: A historical review and quantitative analysis of 203 global food crops. The New Phytologist, 196(1), 29–48. https://doi.org/10.1111/j.1469-8137.2012.04253.x

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W. T. B., … Marshall, D. (2010). Flapjack—Graphical genotype visualization. Bioinforma. Oxf. Engl., 26(24), 3133–3134. https://doi.org/10.1093/bioinformatics/btq580

Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., … Stein, N. (2019). Genebank genomics highlights the diversity of a global barley collection. Nature Genetics, 51(2), 319–326. https://doi.org/10.1038/s41588-018-0266-x

Monteros, M. J., He, C., Choi, J., Zhao, P. X., Tayeh, N., Dai, X. … Udvardi M. K. (2018). Development of the alfalfa breeder's toolbox: Integration of genomic, genetic and germplasm resources for alfalfa improvement. https://plan.core-apps.com/pag_2018/abstract/fff6e3855de940b33de03bc4efab46d

Parween, S., Nawaz, K., Roy, R., Pole, A. K., Venkata Suresh, B., Misra, G., … Chattopadhyay, D. (2015). An advanced draft genome assembly of

a *desi* type chickpea (*Cicer arietinum* L). Scientific Reports, 5, 12806. https://doi.org/10.1038/srep12806

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., … Gamas, P. (2018). Whole-genome landscape of *Medicago truncatula* symbiotic genes. Nature Plants, 4(12), 1017–1025. https://doi.org/10.1038/s41477-018-0286-7

Qin, S., Wu, L., Wei, K., Liang, Y., Song, Z., Zhou, X., … Zhang, Z. (2019). A draft genome for Spatholobus suberectus. Scientific Data, 6(1), 1–9. https://doi.org/10.1038/s41597-019-0110-x

Ren, L., Huang, W., & Cannon, S. B. (2019). Reconstruction of ancestral genome reveals chromosome evolution history for selected legume species. The New Phytologist, 223(4), 2090–2103. https://doi.org/10.1111/nph.15770

Rife, T. W., & Poland, J. A. (2014). Field book: An open-source application for field data collection on Android. Crop Science, 54(4), 1624–1627. https://doi.org/10.2135/cropsci2013.08.0579

Ryu, K. H., Huang, L., Kang, H. M., & Schiefelbein, J. (2019). Single-cell RNA sequencing resolves molecular relationships among individual plant cells. Plant Physiology, 179(4), 1444–1456. https://doi.org/10.1104/pp.18.01482

Sakai, H., Naito, K., Ogiso-Tanaka, E., Takahashi, Y., Iseki, K., Muto, C., … Tomooka, N. (2015). The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. Scientific Reports, 5, 16780. https://doi.org/10.1038/srep16780

Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. Genome Biology, 20(1), 92. https://doi.org/10.1186/s13059-019-1715-2

Sanderson, L.-A., Ficklin, S. P., Cheng, C.-H., Jung, S., Feltus, F. A., Bett, K. E., & Main, D. (2013). Tripal v1.1: A standards-based toolkit for construction of online genetic and genomic databases. Database J. Biol. Databases Curation, 2013, 1–18. https://doi.org/10.1093/database/bat075

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., … Tabata, S. (2008). Genome Structure of the Legume, Lotus japonicus. DNA Research, 15(4), 227–239. https://doi.org/10.1093/dnares/dsn008

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. Nature, 463(7278), 178–183. https://doi.org/10.1038/nature08670

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., … Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics, 46(7), 707–713. https://doi.org/10.1038/ng.3008

Selby, P., Abbeloos, R., Backlund, J. E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O. E., … The BrAPI consortium (2019). BrAPI—An application programming interface for plant breeding applications. Bioinformatics, 35, 4147–4155. https://doi.org/10.1093/bioinformatics/btz190

Sempéré, G., Philippe, F., Dereeper, A., Ruiz, M., Sarah, G., & Larmande, P. (2016). Gigwa-Genotype investigator for genome-wide analyses. GigaScience, 5, 25. https://doi.org/10.1186/s13742-016-0131-8

Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., & Korte, A. (2017). AraPheno: A public database for *Arabidopsis thaliana* phenotypes. Nucleic Acids Research, 45(D1), D1054–D1059. https://doi.org/10.1093/nar/gkw986

Shen, Y., Liu, J., Geng, H., Zhang, J., Liu, Y., Zhang, H., … Tian, Z. (2018). De novo assembly of a Chinese soybean genome. Science China. Life Sciences, 61(8), 871–884. https://doi.org/10.1007/s11427-018-9360-0

Shimomura, M., Kanamori, H., Komatsu, S., Namiki, N., Mukai, Y., Kurita, K., … Katayose, Y. (2015). The Glycine max cv. Enrei Genome for improvement of Japanese soybean cultivars. International journal of Genomics, 2015, 1–8. https://doi.org/10.1155/2015/358127

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., … McLaren, G. (2010). Multifunctional crop trait ontology for breeders' data: Field book, annotation, data discovery and semantic enrichment of the literature. AoB PLANTS, 2010, 11. https://doi.org/10.1093/aobpla/plq008

Silverstein, K. A. T., Moskal, W. A., Wu, H. C., Underwood, B. A., Graham, M. A., Town, C. D., & Vanden Bosch, K. A. (2007). Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. The Plant Journal, 51(2), 262–280. https://doi.org/10.1111/j.1365-313X.2007.03136.x

Smith, R. N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., … Micklem, G. (2012). InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics, 28(23), 3163–3165. https://doi.org/10.1093/bioinformatics/bts577

Smýkal, P., Coyne, C. J., Ambrose, M. J., Maxted, N., Schaefer, H., Blair, M. W., … Varshney, R. K. (2015). Legume Crops Phylogeny and Genetic Diversity for Science and Breeding. Critical Reviews in Plant Sciences, 34(1–3), 43–104. https://doi.org/10.1080/07352689.2014.897904

Smýkal, P., Nelson, M., Berger, J., & von Wettberg, E. (2018). The impact of genetic changes during crop domestication on healthy food development. Agronomy, 8(3), 26. https://doi.org/10.3390/agronomy8030026

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2015). Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes, Genomes, Genetics, 5(10), 1999–2006. https://doi.org/10.1534/g3.115.019000

Stai, J. S., Yadav, A., Sinou, C., Bruneau, A., Doyle, J. J., Fernández-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

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., … Town, C. D. (2014). An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics, 15(1), 312. https://doi.org/10.1186/1471-2164-15-312

Tecle, I. Y., Edwards, J. D., Menda, N., Egesi, C., Rabbi, I. Y., Kulakow, P., … Mueller, L. A. (2014). solGS: A web-based tool for genomic selection. BMC Bioinformatics, 15(1), 398. https://doi.org/10.1186/s12859-014-0398-7

Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., … Grimm, D. G. (2018). The AraGWAS catalog: A curated and standardized *Arabidopsis thaliana* GWAS catalog. Nucleic Acids Research, 46(D1), D1150–D1156. https://doi.org/10.1093/nar/gkx954

Valente, F., Gauthier, F., Bardol, N., Blanc, G., Joets, J., Charcosset, A., & Moreau, L. (2013). OptiMAS: A decision support tool for marker-assisted assembly of diverse alleles. The Journal of Heredity, 104(4), 586–590. https://doi.org/10.1093/jhered/est020

Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S., Brown, A. V., Ren, L., … Plott, C. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. The Plant Journal. https://doi.org/10.1111/tpj.14500

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., … Jackson, S. A. (2012). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nature Biotechnology, 30(1), 83–89. https://doi.org/10.1038/nbt.2022

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., … Cook, D. R. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nature Biotechnology, 31(3), 240–246. https://doi.org/10.1038/nbt.2491

Veeckman, E., Ruttink, T., & Vandepoele, K. (2016). Are we there yet? Reliably estimating the completeness of plant genome sequences. Plant Cell, 28(8), 1759–1768. https://doi.org/10.1105/tpc.16.00349

Vlasova, A., Capella-Gutiérrez, S., Rendón-Anaya, M., Hernández-Oñate, M., Minoche, A. E., Erb, I., … Guigó, R. (2016). Genome and transcriptome analysis of the Mesoamerican common bean and the role of

gene duplications in establishing tissue and temporal specialization of genes. Genome Biology, 17, 32. https://doi.org/10.1186/s13059-016-0883-6

von Wettberg, E. J. B., Chang, P. L., Başdemir, F., Carrasquila-Garcia, N., Korbu, L. B., Moenga, S. M., … Cook, D. R. (2018). Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. Nature Communications, 9(1), 1–13. https://doi.org/10.1038/s41467-018-02867-z

Walls, R. L., Cooper, L., Elser, J., Gandolfo, M. A., Mungall, C. J., Smith, B., … Jaiswal, P. (2019). The plant ontology facilitates comparisons of plant development stages across species. Frontiers in Plant Science, 10, 631. https://doi.org/10.3389/fpls.2019.00631

Wang, D. R., Agosto-Pérez, F. J., Chebotarov, D., Shi, Y., Marchini, J., Fitzgerald, M., … McCouch, S. R. (2018). An imputation platform to enhance integration of rice genetic resources. Nature Communications, 9(1), 3519. https://doi.org/10.1038/s41467-018-05538-1

Wang, J., Sun, P., Li, Y., Liu, Y., Yu, J., Ma, X., … Wang, X. (2017). Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. Plant Physiology, 174(1), 284–300. https://doi.org/10.1104/pp.16.01981

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., … Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution, 35, 543–548. https://doi.org/10.1093/molbev/msx319

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. 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

Xie, M., Chung, C. Y.-L., Li, M.-W., Wong, F.-L., Wang, X., Liu, A., … Lam, H. M. (2019). A reference-grade wild soybean genome. Nature Communications, 10(1), 1216. https://doi.org/10.1038/s41467-019-09142-9

Yang, K., Tian, Z., Chen, C., Luo, L., Zhao, B., Wang, Z., … Wan, P. (2015). Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. Proceedings of the National Academy of Sciences, 112(43), 13213–13218. https://doi.org/10.1073/pnas.1420949112

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., … Roe, B. A. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature, 480(7378), 520–524. https://doi.org/10.1038/nature10625

Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., … Causse, M. (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. Nature Communications, 10(1), 1534. https://doi.org/10.1038/s41467-019-09462-w

Zhou, P., Silverstein, K. A., Gao, L., Walton, J. D., Nallu, S., Guhlin, J., & Young, N. D. (2013). Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). BMC Bioinformatics, 14(1), 335. https://doi.org/10.1186/1471-2105-14-335

Zhou, P., Silverstein, K. A. T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., … Young, N. D. (2017). Exploring structural variation and gene family architecture with de novo assemblies of 15 Medicago genomes. BMC Genomics, 18(1), 261. https://doi.org/10.1186/s12864-017-3654-1

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., … Varshney, R. K. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nature Genetics, 51(5), 865–876. https://doi.org/10.1038/s41588-019-0402-2