Advances in legume research in the genomics era

Ashley N. Egan\textsuperscript{A,D} and Mohammad Vatanparast\textsuperscript{B,C}

\textsuperscript{ADepartment of Bioscience, Aarhus University, Ny Munkegade 116, DK-8000 Aarhus C, Denmark.}
\textsuperscript{BDepartment of Geosciences and Natural Resource Management, University of Copenhagen, DK-1958 Frederiksberg C, Denmark.}
\textsuperscript{CDepartment of Botany, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA.}
\textsuperscript{DPresent address, Department of Biology, Utah Valley University, 800 W University Parkway, Orem, UT 84058, USA. Corresponding author. Email: aegan@uvu.edu}

Abstract. Next-generation sequencing (NGS) technologies and applications have enabled numerous critical advances in legume biology, from marker discovery to whole-genome sequencing, and will provide many new avenues for legume research in the future. The past 6 years in particular have seen revolutionary advances in legume science because of the use of high-throughput sequencing, including the development of numerous types of markers and data useful for evolutionary studies above and below the species level that have enabled resolution of relationships that were previously unattainable. Such resolution, in turn, affords opportunities for hypothesis testing and inference to improve our understanding of legume biodiversity and the patterns and processes that have created one of the most diverse plant families on earth. In addition, the genomics era has seen significant advances in our understanding of the ecology of legumes, including their role as nitrogen fixers in global ecosystems. The accumulation of genetic and genomic data in the form of sequenced genomes and gene-expression profiles made possible through NGS platforms has also vastly affected plant-breeding and conservation efforts. Here, we summarise the knowledge gains enabled by NGS methods in legume biology from the perspectives of evolution, ecology, and development of genetic and genomic resources.

Additional keywords: crop genomes, Fabaceae, genome-wide research, Leguminosae, next-generation sequencing, phylogenomics, RADseq, sequence capture, target enrichment.

Received 4 March 2019, accepted 16 July 2019, published online 30 September 2019

Introduction

In the late 20th century, Sanger sequencing (Sanger et al. 1977) transformed biology and medicine, enabling many genetic advances, the greatest being completion of the human genome project (International Human Genome Sequencing Consortium 2001). Three decades later, second- or next-generation sequencing (NGS) ushered in the genomics era, producing massive amounts of sequence data at a fraction of the cost and time. For comparison, the cost of sequencing a human genome in September of 2001 by Sanger sequencing was \~\$95 million; today sequencing the same genome will cost less than \$1500 (The Human Genome Research Institute; see https://www.genome.gov/sequencingcosts/, accessed 28 August 2019).

Numerous NGS methods have been introduced, each with strengths and weaknesses (for review, see Egan et al. 2012; Soltis et al. 2013; Reuter et al. 2015). NGS technologies provide unprecedented opportunities in fields such as crop genomics, molecular systematics, evolutionary genomics or plant breeding, prompting scientific understanding across the tree of life. In particular, plant biology has blossomed through NGS applications, including transcriptomics (e.g. Matasci et al. 2014; Wen et al. 2015), phylogenomics (Ruhfel et al. 2014; Wickett et al. 2014; Soltis et al. 2018), genome-wide single-nucleotide polymorphism (SNP) sequencing by genome-reduction techniques (Andrews et al. 2016; Jiang et al. 2016) and whole-genome sequencing (Zhang et al. 2011; Koboldt et al. 2013).

Leguminosae (Fabaceae) is the third-largest family of flowering plants after Orchidaceae and Asteraceae (Lewis et al. 2005), comprising \~770 genera and \~19 500 species (Legume Phylogeny Working Group 2013a, 2017). Our understanding of the classification and evolutionary relationships within the family has been transformed in recent years, with the impact of NGS methods as summarised by Doyle (2013). The family was recently reclassified from the classical three subfamilies (Caesalpinioideae, Mimosoideae, Papilionoideae) into six subfamilies, corresponding to the six main clades, on the basis of an international effort involving \~100 scientists, using molecular systematics (Fig. 1; Legume Phylogeny Working Group 2013b, 2017). Fabaceae is
distributed in all of the world’s vegetation types (biomes) except polar ice. Legumes are second only to grasses in economic importance, with uses common to nearly all facets of life, including food, medicine, oils, timber, fibres, industry, fodder, soil stabilisation and soil enrichment (Graham and Vance 2003). Legume research has benefited from NGS innovations, with studies employing genomic sequencing techniques to address questions in evolution, ecology, conservation and plant breeding. The sequencing of legume genomes, both nuclear and plastid, is rapidly transforming our understanding of legume evolution (Doyle 2013), not least through development of a variety of methods, as well as their impacts on legume subdisciplines, from systematics to plant breeding. This review expands on that of Doyle (2013), focusing especially on work published during the intervening 6 years. We summarise recent insights across selected topics and review ongoing research regarding NGS technologies and how they are contributing to our understanding of legumes.

Evolutionary aspects

Given the incredible diversity and vast ecological and economic importance of legumes, understanding the patterns and processes underlying the evolution of the legume family is an important endeavour. Robust inferences of phylogenetic trees are fundamental to any subsequent analyses, above or below the species level. In this section, we summarise knowledge gains in legume evolutionary biology afforded by genomic advances.

Phylogenetics v. Phylogenomics

Investigation of evolutionary relationships among plants by using molecular data began in the late 1980s. Pioneering legume studies included size polymorphism of chloroplast DNA in *Pisum* L. (Palmer et al. 1985) and nuclear rDNA (rDNA) repeat-length and restriction-enzyme site locations in soybean and relatives (Doyle and Beachy 1985). Sanger sequencing revolutionised phylogenetics through ease of use and reproducibility (Sanger et al. 1977). The chloroplast gene *matK* is the most comprehensively sequenced phylogenetic marker in legumes. Wojciechowski et al. (2004) produced one of the earliest generic-level phylogenies of legumes by using 330 *matK* gene sequences, demonstrating monophyly of Papilionoideae and resolving multiple papilionoid subclades. Many subsequent studies have used *matK* (e.g. Bruneau et al. 2008; Simon et al. 2009; Stefanovic et al. 2009; Cardoso et al. 2012; de Queiroz et al. 2015; Egan et al. 2016; Snak et al. 2016), culminating in the Legume Phylogeny Working Group (2017) phylogeny that included 3842 *matK* sequences representing 3696 species (~20% of the family) and 698 of the 765 genera that had been recognised. Whereas *matK* robustly supported each of the newly recognised six subfamilies, basal nodes key to understanding subfamilial relationships and early legume evolution remained unresolved, which was likely because of the lack of resolving power available from a single marker.

The use of nuclear markers for legume phylogenetics has lagged behind chloroplast markers, largely because of the bi-parental inheritance and the high incidence of gene duplications in the nuclear genome, which make orthology assessment and primer design difficult (Zimmer and Wen 2013). The advent of genomics has transformed and facilitated nuclear marker discovery (Zimmer and Wen 2015). For example, Scherson et al. (2005) screened several nuclear loci from the *Medicago truncatula* Gaertn. genome for phylogeny reconstruction in the hyper-diverse legume genus *Astragalus*. Similarly, Choi et al. (2006) tested 274 putative single-copy genes garnered from comparative analysis of 15 legume genomes from six species and identified 129 single-copy loci that were tested across 95 legume species.

As helpful as NGS methods are for improving nuclear-marker discovery, nuclear loci that are putatively single-copy in one lineage may not be so in others, a phenomenon that makes finding the ‘silver bullet’ of nuclear markers difficult in plants (e.g. Manzanilla and Bruneau 2012), and which argues for having many different markers to mitigate issues with a few. This is where the true utility of NGS technologies comes in. The move from single amplicon-based Sanger sequencing to NGS-based, simultaneous sequencing of numerous markers is rapidly transforming our understanding of legume evolution (Doyle 2013), not least through development of a variety of new approaches for NGS marker data for phylogenetic and population-genetic studies. These include microsatellites, RNAseq or transcriptomics (Wang et al. 2009; Wen et al. 2015), restriction site-associated DNA tags (RADseq; Miller et al. 2007), genotyping-by-sequencing (GBS; Elshire et al. 2011), genome skimming (Straub et al. 2012), targeted enrichment (also known as sequence capture or hybrid enrichment; Gnirke et al. 2009), simultaneous amplicon sequencing (Bybee et al. 2011), and whole-genome sequencing (e.g. Stein et al. 2018; Grover et al. 2019).

Studies have capitalised on massively parallel sequencing for discovery and optimisation of microsatellite markers for population studies, for example, using transcriptome sequencing, including in legumes (e.g. Chapman 2015; Vatanparast et al. 2016; Sathyanarayana et al. 2017; Haynsen et al. 2018). Other popular NGS methods for population genetics include RADseq and GBS, which are also employed in
phylogenetics below the genus level. For instance, Wong et al. (2015) used GBS across 60 accessions of seven species of *Lens* and produced a phylogeny in which all seven species were reciprocally monophyletic. Grillo et al. (2016) employed RADseq across 191 accessions of *Medicago truncatula* in Europe to investigate population structure and screen for candidate symbiosis genes. They found evidence that suggests that one gene, *DMII*, is under adaptive selection. Their work detailed five distinct genetic clusters and aided in correct identification of species. That said, work by others has found that different parameter perturbations produced significant differences in phylogenetic networks of species relationships within *Medicago* by using RADseq (Blanco-Pastor et al. 2018), suggesting caution when using population-level markers above species level.

RNAseq is often used as the first line for phylogenetic-marker development, whether for microsatellite development, amplicon sequencing (e.g. Chapman 2015) or targeted enrichment. For example, Vatanparast et al. (2018) used 30 transcriptomes across the legume family to select over 500 nuclear markers for targeted enrichment, which were tested across 25 legume taxa. This same target set proved useful as far out as Rosales (M. Vatanparast and A.N. Egan, unpubl. data), showing that targeted enrichment can be useful for both lower- and higher-level phylogenomics, as has been found by others (Kadlec et al. 2017; Chau et al. 2018).

The first targeted enrichment-gene set in legumes selected 50 nuclear loci from the *Medicago* genome to resolve relationships among six *Medicago* species and other genera of tribe Trifolieae (de Sousa et al. 2014). Targeted enrichment may be especially useful for resolving relationships among rapidly radiated, species-rich groups by virtue of the large amount of data produced. For example, Nicholls et al. (2015) applied a targeted enrichment method using transcriptomes from three species to isolate 264 nuclear loci for sequencing *Inga* Mill., a genus of ~300 neotropical rainforest trees that diversified rapidly during the late Miocene (2–10 million years ago). Of these loci, 194 were used for phylogeny reconstruction across 22 *Inga* species, resulting in a highly resolved phylogenetic tree. Similarly, Ojeda et al. (2019) used 289 nuclear loci identified from transcriptomes across Detarioideae to construct a phylogenomic hypothesis of relationships within the *Anthonotha* clade, discovering an overall general trend towards petal reduction in this florally diverse group. In addition to the target-gene sets outlined here, several other, as yet unpublished, gene sets have been generated for legumes.

As legume researchers adopt NGS for phylogenomics, it could be worth considering selection and adoption of a core set of nuclear target genes for use across legumes, along lines similar to the 353 nuclear-gene set for targeted enrichment in angiosperms (Johnson et al. 2019). Doing so would also facilitate barcoding efforts (Hollingsworth et al. 2016). With the recent addition of the first genome sequences for mimosoids and cercidoids (Table 1, Fig. 2), plus a significant number of new transcriptomes, it would now be possible to design a legume-wide, legume-specific bait set more efficiently. Such a design might include genes specifically selected for phylogenetics, as well as genes that are related to particular legume functional traits such as nodulation, compound-leaf development or floral symmetry, as has been done for Caryophyllales (Moore et al. 2018). Whereas the number of target gene sets used in legumes will undoubtedly increase, including a core subset of genes in every target-gene set would enable published sequences from different studies to be combined for wider analyses. Doing so would foster wide collaboration among legume systematists, while still enabling project-specific objectives to be met. That said, issues with the conflation of orthology and paralogy need to be dealt with when using universal gene sets because of whole-genome duplications (WGDs) and differential gene birth and death events, issues that are particularly troubling if targets are used for DNA barcoding.

**Species diversity**

Understanding the dynamics of diversification within the third-largest species-rich plant family is complex and only a few attempts have been made to estimate species diversification and speciation–extinction rates within legumes (e.g. Sanderson and Wojciechowski 1996; Richardson et al. 2001; Scherson et al. 2008). Koenen et al. (2013) estimated species diversification rates and tested for rate shifts by using species-level, time-calibrated phylogenies of the following four legume clades: *Calliandra* Benth., *Indigofereae*, *Lupinus* L. and *Mimosa* L., finding evidence for significant among-lineage diversification-rate variation. A major challenge for diversification studies is having a well resolved phylogeny, a non-trivial task when dealing with rapid, recent radiations (Hughes et al. 2015). The most species-rich legume tribe, *Galegeae*, includes *Astragalus*, the largest genus of any biological group, with nearly 3000 species (Kazempour Osaloo et al. 2003; Podlech et al. 2014), as well as *Oxytropis* DC., its sister genus with between 310 and 450 species (Malsheev 2008). Recent molecular studies of these groups exemplify the difficulties of resolving phylogenies for rapidly radiating lineages (e.g. Azani et al. 2017; Bagheri et al. 2017). Where phylogenies based on data from single or a few genes lack resolution, genomic-scale data may provide a solution. Application of anchored enrichment of 527 gene regions (i.e. targeted enrichment) in *Oxytropis* produced a robust phylogeny compared to relying on conventional markers (Shahi Shavvon et al. 2017).

Next-generation sequencing data provide more than just phylogenetic resolution and their application to elucidating adaptive radiations and diversification is just beginning. This is exemplified in ground-breaking studies on *Lupinus*, a genus of ~280 species, with a series of nested and parallel rapid radiations in North and South America (Hughes and Eastwood 2006) where the genus exhibits startling diversity in growth form and habitat in the Andes, a radiation that is likely to have been spurred by Pleistocene glacial cycles (Nevado et al. 2018). Furthermore, lineage-specific diversification rates were detected across the phylogeny, but the ability to ascertain the processes underpinning rapid species radiations remained elusive. Today, investigating adaptive radiations from a genomic perspective is shedding light on how speciation and trait diversification occur. Contreras-Ortiz et al. (2018) used RADseq to investigate species diversification in Andean lupines and found evidence for both adaptive, ecological and non-adaptive, geographical drivers influencing their radiation.
### Table 1. Legume genomes published as of 11 May 2019

Some genomes were sequenced by multiple groups and not all are summarised here, in which case the genome summarised has the first author underlined. Gb, gigabases; Mb, megabases; bp, base pairs; nr, not reported; SNP, single-nucleotide polymorphism.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Relevance</th>
<th>Genome size</th>
<th>Length assembled</th>
<th>Number of genes</th>
<th>Method</th>
<th>Sequence produced</th>
<th>Assembly status</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference quality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Glycine max</em> (soybean) Williams 82&lt;sup&gt;A&lt;/sup&gt;</td>
<td>Protein and oil crop</td>
<td>1115 Mb</td>
<td>955 Mb</td>
<td>56 044</td>
<td>Sanger shotgun &amp; BAC sequencing</td>
<td>Iterative</td>
<td>8.04× coverage; 17191 contigs of L50 182.8 kb; 1190 scaffolds of L50 47.8 Mb; 20 chromosomes produced with BAC-end sequences</td>
<td>Schmutz et al. 2010; Shen et al. 2018</td>
</tr>
<tr>
<td><em>Glycine soja</em> (wild soybean)</td>
<td>Wild relative</td>
<td>1115 Mb</td>
<td>915.4 Mb</td>
<td>nr</td>
<td>Illumina (genome resequencing); 454 (variant validation)</td>
<td>48.8 Gb</td>
<td>52× coverage; 75 195 contigs of N50 ~250 bp; re-sequenced reference: this genome was assembled based on soybean reference</td>
<td>Kim et al. 2010; Xie et al. 2019</td>
</tr>
<tr>
<td><strong>Reference draft quality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Arachis duranensis</em></td>
<td>A genome diploid wild ancestor</td>
<td>1.25 Gb</td>
<td>1211 Mb</td>
<td>36 734</td>
<td>Illumina</td>
<td>325.73 Gb</td>
<td>154× coverage; 76 5406 contigs of N50 22.3 kb; 63 5392 scaffolds of N50 948 kb; assembled to 10 pseudomolecules using BAC sequences and GBS-derived genetic map</td>
<td>Bertioli et al. 2016; Chen et al. 2016</td>
</tr>
<tr>
<td><em>Arachis ipaensis</em></td>
<td>B genome diploid wild ancestor</td>
<td>1.56 Gb</td>
<td>1512 Mb</td>
<td>41 840</td>
<td>Illumina</td>
<td>416.59 Gb</td>
<td>163× coverage; 86 9435 contigs of N50 23.5 kb; 75 9499 scaffolds of N50 5.34 Mb; assembled to 10 pseudomolecules using BAC sequences and GBS-derived genetic map</td>
<td>Bertioli et al. 2016; Lu et al. 2018</td>
</tr>
<tr>
<td><em>Arachis hypogaea</em> cv. Tifrunner (cultivated peanut)</td>
<td>Crop and oil plant</td>
<td>~2.7 Gb</td>
<td>2.54 Gb</td>
<td>90 519</td>
<td>PacBio (sequencing), Illumina (polishing), Hi-C (scaffolding)</td>
<td>173.6 Gb</td>
<td>76.74× coverage; 17.75 M reads of average length 9784 bp; 4,037 contigs of N50 1.5 Mb; 384 scaffolds of N50 134.0 Mb; assembled to 20 chromosomes using Hi-C</td>
<td>Bertioli et al. 2019; Chen et al. 2019</td>
</tr>
<tr>
<td><em>Cajanus cajan</em> var. <em>Asha</em></td>
<td>Crop plant</td>
<td>833.07 Mb</td>
<td>605.78 Mb</td>
<td>48 680</td>
<td>Illumina, Sanger BAC-end libraries</td>
<td>237.2 Gb</td>
<td>163.4× coverage; 173 708 contigs of N50 21.95 kb; 137 542 scaffolds of N50 536 kb; anchored to 11 chromosomes using BAC-ends and genetic map</td>
<td>Varshney et al. 2012</td>
</tr>
<tr>
<td><em>Cajanus cajan</em> var. <em>Asha</em>&lt;sup&gt;A&lt;/sup&gt;</td>
<td>Crop plant</td>
<td>858 Mb</td>
<td>648.2 Mb</td>
<td>56 888</td>
<td>454, Illumina</td>
<td>195.4 Gb</td>
<td>360 028 contigs of N50 5341 bp</td>
<td>Mahato et al. 2018</td>
</tr>
</tbody>
</table>

<sup>A</sup> Reference quality and draft quality are based on the quality of reference sequence for these genomes.
<table>
<thead>
<tr>
<th>Species</th>
<th>Type</th>
<th>Size (Mb)</th>
<th>Annotation (Mb)</th>
<th>Flowcell</th>
<th>Coverage</th>
<th>Contig N50</th>
<th>Scaffold N50</th>
<th>Pseudomolecules</th>
<th>MAPS</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cicer arietinum</em> (chickpea)</td>
<td>Crop plant</td>
<td>738</td>
<td>544.73</td>
<td>Illumina</td>
<td>28 269</td>
<td>153 Gb</td>
<td>207.32 coverage; 62 619 contigs of N50 23.54 kb; 7 163 scaffolds of N50 40 Mb; eight pseudomolecules produced with help of BAC sequences and genetic map</td>
<td>Varshney et al. 2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Cicer arietinum</em> L. (chickpea)</td>
<td>Crop plant</td>
<td>740</td>
<td>519.8</td>
<td>454, Illumina</td>
<td>27 571</td>
<td>57 Gb</td>
<td>15× 454 coverage; 181 462 scaffolds of N50 77.3 kb; 8 pseudomolecules produced with help of BAC sequences and genetic map</td>
<td>Jain et al. 2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Lotus japonicus</em> (bird’s-foot trefoil)</td>
<td>Model legume</td>
<td>472</td>
<td>315.1</td>
<td>Sanger shotgun and BAC sequencing; Illumina added later</td>
<td>Iterative</td>
<td>39 734</td>
<td>Sanger: 2.4× coverage; 109 986 contigs; 110 940 scaffolds; six chromosomes produced with BAC and TAC sequences. Illumina 40× coverage; 23 572 contigs with N50 (anchored) of 118 kb; further hypothetical chromosome 0 produced</td>
<td>Sato et al. 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Lupinus angustifolius</em> (narrow-leafed lupin)</td>
<td>Health food</td>
<td>924</td>
<td>609</td>
<td>Illumina</td>
<td>33 076</td>
<td>150.4 Gb</td>
<td>162.8× coverage; 1068 669 contigs of N50 4246 bp; 14 379 scaffolds of N50 (?); 20 pseudomolecules derived from genetic map</td>
<td>Hane et al. 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Medicago truncatula</em> (barrel medic)</td>
<td>Model legume</td>
<td>465</td>
<td>384.5</td>
<td>Sanger, BAC end, Optical Mapping, 454, Illumina</td>
<td>Iterative</td>
<td>50 894</td>
<td>All data types (v.4.0): contig N50 102 kb; scaffold N50 4.24 Mb; eight pseudomolecules produced from optical maps and GBS-based linkage map</td>
<td>Young et al. 2011; Tang et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Phaseolus vulgaris</em> (common bean)</td>
<td>Model bean and protein crop</td>
<td>587</td>
<td>537.2</td>
<td>Sanger, 454, and Illumina; PacBio added later</td>
<td>Iterative</td>
<td>36 995</td>
<td>All data types (v.2.1):1044 contigs of N50 1.9 Mb; 478 scaffolds of N50 49.7 Mb; 11 chromosomal pseudomolecules produced with BACs and linkage map</td>
<td>Schmutz et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Trifolium pratense</em> (red clover)</td>
<td>Forage legume</td>
<td>420</td>
<td>309</td>
<td>Illumina, Sanger BAC-end sequencing</td>
<td>nr</td>
<td>40 868</td>
<td>Hybrid assembly: 30× coverage; 39 904 scaffolds of N50 223 kb; seven pseudomolecules produced with BACs, a physical, and two genetic maps</td>
<td>De Vega et al. 2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Trifolium subterraneum</em></td>
<td>Annual wild relative of forage crop</td>
<td>540</td>
<td>512</td>
<td>Illumina, 454, Bionano</td>
<td>157 Gb (Illumina + 454)</td>
<td>290.7× coverage; 27 257 super-scaffolds of N50 410.5 kb; eight pseudomolecules produced by genetic map of &gt;35 000 SNPs and a Bionano physical map</td>
<td>Hirakawa et al. 2016; Kaur et al. 2017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organism</td>
<td>Relevance</td>
<td>Genome size</td>
<td>Length assembled</td>
<td>Number of genes</td>
<td>Method</td>
<td>Sequence produced</td>
<td>Assembly status</td>
<td>References</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-------------</td>
<td>------------------</td>
<td>-----------------</td>
<td>--------</td>
<td>-------------------</td>
<td>----------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Trifolium subterraneum</em>^A^</td>
<td>Annual wild relative of forage crop</td>
<td>540 Mb</td>
<td>403.4 Mb</td>
<td>nr</td>
<td>Illumina, 454, Hi-C</td>
<td>nr</td>
<td>48× coverage (Hi-C); 46453 contigs of N50 22.77; 5285 scaffolds of N50 56.3 Mb; eight chromosome-length scaffolds with lengths ranging from 49.5 to 65.2 Mb produced from Hi-C mapping coverage (Hi-C); 36516 contigs of N50 21.9 kb; 3883 scaffolds of N50 703 kb; 11 pseudomolecules produced with GBS-based genetic map production; 56.3 Mb; eight chromosome-length scaffolds with lengths ranging from 49.5 to 65.2 Mb</td>
<td>Hirakawa et al. 2016; Dudchenko et al. 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Vigna angularis</em> (adzuki bean)</td>
<td>Protein crop</td>
<td>612 Mb</td>
<td>443 Mb</td>
<td>26857</td>
<td>Illumina, 454</td>
<td>172 Gb</td>
<td></td>
<td>Kang et al. 2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Vigna radiata var. radiata</em> (mungbean)</td>
<td>Protein crop</td>
<td>579 Mb</td>
<td>421 Mb</td>
<td>22427</td>
<td>Illumina, 454</td>
<td></td>
<td>25 922 contigs of N50 41.8 kb; 2748 scaffolds of N50 1516 kb; 11 pseudomolecules produced by GBS-based genetic map production</td>
<td>Kang et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Vigna unguiculata</em> (cowpea)</td>
<td>Protein crop</td>
<td>620 Mb</td>
<td>519.4 Mb</td>
<td>29773</td>
<td>PacBio (genome), Illumina (gene prediction)</td>
<td>56.8 Gb</td>
<td>~6 million PacBio reads of N50 14.5 kb for 91× coverage; 11 pseudomolecules produced using two optical maps and 10 genetic maps comprising &gt;44 000 SNPs</td>
<td>Lonardi et al. 2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draft quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ammopiptanthus nanus</em></td>
<td>Endangered desert shrub</td>
<td>889 Mb</td>
<td>823.74 Mb</td>
<td>37188</td>
<td>PacBio (genome), Illumina (gene prediction)</td>
<td>64.72 Gb (PacBio); 55.97 Gb (Illumina)</td>
<td>72.59× coverage; 7.92 million reads of N50 12.79 kb; 1099 contigs of N50 2.76 Mb</td>
<td>Gao et al. 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Cercis canadensis</em> (redbud)</td>
<td>Horticultural tree</td>
<td>301 Mb</td>
<td>330 Mb</td>
<td>34023</td>
<td>Illumina</td>
<td>335.07 Gb</td>
<td>220.74× coverage; gap-filled scaffold N50 of 12.883 bp; gap-filled scaffold N50 of 421 kb</td>
<td>Griesmann et al. (2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Dipteryx oleifera</em></td>
<td>Tropical timber species</td>
<td>1.89 Gb</td>
<td>1.16 Gb</td>
<td>nr</td>
<td>Illumina, PacBio</td>
<td>50.78 Gb</td>
<td>44× Illumina coverage, 5× PacBio coverage. Combined assembly: 381 857 contigs of N50 81.94 bp</td>
<td>Jimenez-Madrigal 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Faidherbia albida</em> (apple- ring acacia)</td>
<td>Agroforest crop</td>
<td>661 Mb</td>
<td>653.7 Mb</td>
<td>28979</td>
<td>Illumina</td>
<td>nr</td>
<td>nr</td>
<td>Chang et al. 2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Glycine latifolia</em></td>
<td>Wild relative</td>
<td>1.13 Gb</td>
<td>939 Mb</td>
<td>54475</td>
<td>10X Genomics, Chromium linked reads on Illumina</td>
<td>78.39 Gb</td>
<td>66 825 contigs of N50 (?); 42 539 scaffolds of N50 53.5 kb; 20 pseudomolecules produced using two genetic maps and the G.max genome</td>
<td>Liu et al. 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>Type</td>
<td>Genome Size</td>
<td>Drafted Size</td>
<td>Assembly Size</td>
<td>Coverage</td>
<td>Special Notes</td>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------------------------</td>
<td>-------------</td>
<td>--------------</td>
<td>---------------</td>
<td>------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycyrrhiza uralensis (Chinese licorice)</td>
<td>Medicinal legume</td>
<td>400 Mb</td>
<td>379 Mb</td>
<td>34 445</td>
<td>PacBio, Illumina</td>
<td>387.29 Gb 817× (Illumina) and 16.86× (PacBio) coverage. Hybrid assembly: 72 148 contigs of N50 7324 bp; 12 528 scaffolds of N50 109.3 kb</td>
<td>Mochida et al. 2017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lablab purpureus (Hyacinth bean)</td>
<td>Ancient crop</td>
<td>423 Mb</td>
<td>395.5 Mb</td>
<td>20 946</td>
<td>Illumina</td>
<td>nr</td>
<td>nr</td>
<td></td>
<td>Chang et al. 2019</td>
<td></td>
</tr>
<tr>
<td>Lupinus angustifolius (Narrow-leafed lupin)</td>
<td>Health food</td>
<td>1152 Mb</td>
<td>598 Mb</td>
<td>57 807</td>
<td>Illumina</td>
<td>31 Gb</td>
<td>26.9× coverage; 457 917 contigs of N50 5.8 kb; 234 534 scaffolds of N50 12.5 kb; RAD-seq produced 20 linkage groups but pseudomolecules not built</td>
<td>Yang et al. 2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mimosa pudica (Sensitive thigmotropic plant)</td>
<td></td>
<td>896 Mb</td>
<td>557 Mb</td>
<td>33 108</td>
<td>Illumina</td>
<td>370.07 Gb</td>
<td>287.65× coverage; gap-filled scaffold N50 of 11 069 bp; gap-filled scaffold N50 of 117.9 kb</td>
<td>Griesmann et al. 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nissolia schotti (Schott’s desert perennial yellowhood)</td>
<td></td>
<td>471 Mb</td>
<td>466 Mb</td>
<td>36 369</td>
<td>Illumina</td>
<td>144.06 Gb</td>
<td>149.48× coverage; gap-filled scaffold N50 of 10 065 bp; gap-filled scaffold N50 of 179.7 kb</td>
<td>Griesmann et al. 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trifolium pratense (Red clover)</td>
<td>Forage legume</td>
<td>420 Mb</td>
<td>314.6 Mb</td>
<td>47 398</td>
<td>Illumina</td>
<td>24.6 Gb</td>
<td>58.8× coverage; 236 989 contigs of N50 2397 bp; 176 760 scaffolds of N50 4750 bp</td>
<td>Ivšánek et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vigna radiata var. sublobata</td>
<td>Wild relative</td>
<td>501.7 Mb</td>
<td>423 Mb</td>
<td>22 834</td>
<td>Illumina</td>
<td>nr</td>
<td>45 606 contigs of N50 16.4 kb; 81 616 scaffolds of N50 214 kb</td>
<td>Kang et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vigna reflexo-pilosa var. glabra</td>
<td>Wild relative</td>
<td>967.8 Mb</td>
<td>791.6 Mb</td>
<td>41 844</td>
<td>Illumina</td>
<td>nr</td>
<td>163 809 contigs of N50 4.6 kb; 29 166 scaffolds of N50 63 kb</td>
<td>Kang et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vigna subterranea (Bambara groundnut)</td>
<td>Food crop</td>
<td>550 Mb</td>
<td>535 Mb</td>
<td>31 707</td>
<td>Illumina</td>
<td>nr</td>
<td>Nr</td>
<td></td>
<td>Chang et al. 2019</td>
<td></td>
</tr>
<tr>
<td>Unassembled Vigna angustifolius var. nipponensis</td>
<td>Wild relative</td>
<td>562 Mb</td>
<td>nr</td>
<td>nr</td>
<td>Illumina</td>
<td>43.9 Gb</td>
<td>72.04× coverage</td>
<td>Kang et al. 2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vigna nakashimae</td>
<td>Wild relative</td>
<td>756 Mb</td>
<td>431 Mb</td>
<td>23 197</td>
<td>Illumina</td>
<td>36.2 Gb</td>
<td>47.91× coverage</td>
<td>Lestari et al. 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vigna nepalensis</td>
<td>Wild relative</td>
<td>610 Mb</td>
<td>nr</td>
<td>nr</td>
<td>Illumina</td>
<td>45 Gb</td>
<td>80× coverage</td>
<td>Kang et al. 2015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Genome was improved by addition of next-generation sequence data after initial publication and statistics reported are of the improved version.*
Using transcriptomes from slowly and rapidly diversifying lupin lineages, Nevado et al. (2016) verified, for the first time in plants, the role of adaptive evolution in rapid radiations. Rapidly diversifying lineages had two to three times more positively selected genes than did slowly diversifying lineages, suggesting a genome-wide response to adaptation. Further, the rapidly diversifying Andean lineage exhibited a higher gene-expression divergence than did the slowly diversifying lineages, suggesting that underlying genomic shifts in expression happened during adaptive radiations. Also, shifts in gene-expression level were non-randomly clustered around significant evolutionary time points, including at the base of the Andean clade when lupins moved into novel, extreme montane environments, and near a branch signifying a shift from annual to perennial life-history. NGS has gone beyond simply detecting shifts in diversification rates, by enabling the determination of the how and why behind diversification in plants.

### Polyploidy

Whole-genome duplication (WGD), or polyploidy, is a major evolutionary process underlying speciation in plants (Van de Peer et al. 2017), and legumes are no exception (Doyle 2012). Polyploid crop legumes, such as peanut, alfalfa and soybean are well known and, in soybean (Glycine max [L.] Merr.), studies have confirmed genomic evidence for both ancient and recent polyploidisation events just before the radiation of the genus c. 10–12 million years ago, and between 40 and 66 million years ago (Egan and Doyle 2010). Legume WGDs were further characterised by Cannon et al. (2015), by using transcriptomic and genomic data from 20 diverse legumes and 17 outgroups to determine that WGDs coincide with the origins of major legume lineages. A follow-up study incorporating genome data for the genus Cercis L. (Cercidoideae), suggested that Cercis may represent the only extant legume lineage lacking a polyploid...
history, providing a plausible hypothesis of what the ancestral legume genome looked like (Stai et al. 2019). In contrast, Koenen et al. (2019) used thousands of nuclear genes and 72 protein-coding chloroplast genes to find evidence for WGDs at the stem of all legumes, as well as nested WGDs subtending radiation of subfamilies Papilionoideae and Detarioideae. Koenen et al. (2019) also described difficulties in resolving the initial divergence of the legume family tree and suggested that polyploidy may play a key role in the lack of support for deep-branching relationships in the family. Further improvements in long-read NGS platforms may solve such recalcitrant nodes and facilitate research into polyploidy and its ramifications.

Genome sequencing also enables us to detect and characterise more recent WGDs. Allopolyploidy within Glycine has been shown to be rampant and complex, but genomic data have helped unravel relationships within the genus by using transcriptomics (Bombarely et al. 2014) and GBS (Sherman-Broyles et al. 2017). Similarly, comparison of rDNA, genomic and fluorescence in situ hybridisation data and whole plastome sequences generated by Illumina-based sequencing of Stylosanthes scabra Vogel, an important forage legume, and its hypothesised genome donors S. hamata (L.) Taub. or S. seabrana B.L.Maass & 't Mannetje (A genome) and S. viscosa (L.) Sw. (B genome), have provided evidence for an allopolyploid origin of S. scabra and showed the genomic impacts of subsequent homogenisation following ‘genomic shock’ (Marques et al. 2018). Capitalising on a target-gene set derived from the Medicago truncatula genome (de Sousa et al. 2014), Eriksson et al. (2017) tested the hypothesis that M. prostrata is a homoploid hybrid, while accounting for the impact and signature of introgression from M. sativa as a contributor of genetic variation therein. Eriksson et al. (2018) characterised the allopolyploid origin of two Medicago species and showed the importance of allele phasing. Using NGS methods to investigate the mechanisms of polyploidisation can resolve duplicated regions by phasing and comparative analyses.

Ecological aspects
Quantification and integration of information on species, genetic, population and ecosystem diversity from NGS-based metagenomic methods can contribute to biodiversity and conservation assessment and understanding of evolutionary history, population processes, community assembly, and ecosystem equilibrium and services (Papadopoulou et al. 2015; de la Harpe et al. 2017). As primary mediators of nitrogen fixation, legumes are an integral source of fixed nitrogen within terrestrial communities from grasslands to forests. Here, we review some of the recent advances obtained through genomic studies within ecological arenas.

Forestry and range management
From providing essential ecosystem and natural resources to sustaining natural and human infrastructures, legumes are fundamental and abundant components of biomes across the globe, from desert sands to towering trees of the Amazon rainforests and temperate grasslands where legume biodiversity is directly related to the overall biological diversity and community health. Metagenomic comparison of above- and below-ground plant-species richness in a grassland employed Sanger sequencing and Roche 454 pyrosequencing to illustrate that below-ground diversity was higher, demonstrating the power of NGS methods to detect dormant plant diversity (Hiesalu et al. 2012). Assessments of the portion of biodiversity that legumes represent in grasslands, and perhaps other ecosystems, through metagenomic barcoding may help in maintaining and managing biodiversity in ecosystems, be they native or range managed, particularly as loss of legume biodiversity is directly linked to a decline in the nitrogen budget and the overall health and biodiversity of plant communities (Spehn et al. 2002).

Legumes are particularly prevalent in Neotropical forests and savannas, where plot data suggest that they make up more than 11% of species (Oliveira-Filho et al. 2013; Yahara et al. 2013). With increasing pressures on forest-ecosystem dynamics, including fragmentation, over-logging of particular species, and wholesale deforestation, understanding how to mitigate the effects of such events and their toll on ecosystem health, as well as the roles that legumes play therein, are urgent needs. Genomic methods are poised to help (Neale and Kremer 2011). Creation of genomic-marker sets is an important starting point. For example, RADseq has been used to generate 330 SNPs for population genetic analyses in Robinia pseudacacia L., an economically important eastern North American tree widely cultivated and invasive in Europe and elsewhere (Verdu et al. 2016); employment of such markers may help forest managers capitalise on its use and limit its dynamic spread. Similarly, several studies have generated SNP marker sets for Dipteryx Schreb., an economically important genus of Neotropical canopy trees, various species of which are threatened. These include development of microsatellite markers (Soares et al. 2012), a nuclear and plastid SNP MassArray panel (Honorio Coronado et al. 2019), and a draft genome (Jimenez-Madrigal 2018), providing useful genetic data resources for management and conservation through detailed GD assessment. Gailing et al. (2017) used RADseq to produce a framework genetic-linkage map for Gleditsia triacanthos L., a common North American hardwood forest tree, providing an important genetic resource for future quantitative trait-locus mapping.

Landscape genomic data coupled with an assessment of GD can promote sound forest and range management, quantify the impacts of deforestation, fragmentation, climate change and habitat restoration. For example, Acacia koa A.Gray, a Hawaiian endemic and one of two dominant canopy hardwood tree species in Hawaiian forests, is under increasing pressures from logging and changing climate. Gugger et al. (2018) used GBS to assess GD across 311 Acacia koa trees sampled over its geographical, elevational and climatic range, and found evidence for genetic differentiation among islands and a strong association between genetic structure and the mean annual rainfall. These results suggest that changing rainfall patterns could cause a genetic offset between adaptation and extant populations, placing future survival of this species at risk. Such knowledge can be used for future management planning. Similarly, Grando (2015) assessed GD across native and restored populations of Piptadenia gonoacantha (Mart.) J.F.Macbr., a
species that is often used in reforestation across the Brazilian Atlantic Forest because of its rapid growth and regeneration. These data provided insights into which native populations should be used as seed sources for restoration efforts and determined that GD was similar across native and restored sites, which is evidence of successful capture and maintenance of biodiversity during restoration efforts. As NGS becomes more affordable for high sample numbers, landscape genomics and GD assessment using such data can ensure sound scientific foundations for natural-resource management and biodiversity conservation in the face of global change.

Conservation biology and genetics

As one of the largest plant families and key components of tropical and temperate forest ecosystems, legumes are the focus of global conservation biology and genetics global change. Resource management and biodiversity conservation in the face of such data can ensure sound scientific foundations for natural-resource management and biodiversity conservation in the face of global change.

Invasion biology

The numerous markers generated by reduced-representation genomic methods can also resolve recent population divergences such as those arising from anthropogenic plant invasions (Chown et al. 2015) and help understand whether pre-adaption to the novel environment or rapid adaptive changes account for invasions. For example, Helliwell et al. (2018) used Illumina shotgun sequencing of rosid species in a Nebraska grassland to isolate ~80 plastid genes for 45 species, 22 of which were legumes. This enabled the reconstruction of a plastid phylogeny for a complete community for divergence dating and estimation of conservation metrics, illustrating the potential of genomics methods for conservation-biology research.

Conservation often involves efforts to characterise genetic diversity within vulnerable or endangered species by using microsatellite markers. Genomic sequencing has revolutionised discovery of such markers (Zalapa et al. 2012), whether species-specific (Abdelkrim et al. 2009) or for use across a broader taxonomic group (Hodel et al. 2016), including for several legumes (e.g. Borges et al. 2015; Morris et al. 2016). For example, microsatellite markers developed from the Lotus japonicus (Regel) K. Larsen genome (Sato et al. 2008) were used through cross-species amplification in Lotus sessilifolius DC., a species endemic to Macaronesia (Yang et al. 2018). To determine which populations should be prioritised for conservation management, eight populations across four islands were assessed for their GD across 11 microsatellite markers, highlighting a population from Tejina-Milán as strongly distinct and of low genetic diversity relative to the others, suggesting this as one population to target. Next-generation sequencing platforms coupled with amplicon sequencing can now be used to obtain microsatellite data en masse (Zhan et al. 2017).

Although microsatellites are a time-tested and effective tool for assessing GD, issues of small sample sizes and low numbers of markers limit their power for assessing population structure and dynamics. Reduced representation techniques such as GBS or RADseq, or targeted enrichment and amplicon-sequencing methods linked to RADseq and GBS (e.g. GTseq and Rapture), offer cost-effective methods for generating orders of magnitude more marker sites (thousands to millions of SNPs) than do microsatellites (Meek and Larson 2019) for assessing GD and establishing population and species relationships. For example, Harrison et al. (2019) garnered thousands of SNPs by GBS to compare genetic variation among infraspecific taxa of the Astragalus lentiginosus Hook. species complex including var. piscinensis Barneby, that inhabits just 8 km². They showed that, in spite of rarity, significant genetic diversity and population structure exists within and among varieties. Harrison et al. (2019) exemplified the power of NGS to produce large numbers of SNPs for population-genomic and conservation studies.

Nitrogen fixation

Legumes provide numerous inputs to agricultural and natural ecosystems, with perhaps the most important being soil enrichment by nitrogen fixation. The majority of legume
species have the ability to form symbioses with rhizobial bacteria that transform atmospheric nitrogen to ammonia, making atmospheric nitrogen bioavailable in the soil as amino acids and other cellular constituents. Metagenomic approaches are commonly used to characterise soil microbiomes (Andújar et al. 2015), advancing understanding of the interactions of legumes and the environment through nodulation and nitrogen fixation (e.g. Afkhami and Stinchcombe 2016). Birnbaum et al. (2018) investigated symbionts of Acacia rostellifera Benth. across the natural-soil fertility gradient of the Jurien Bay Dune chronosequence. Using Illumina-based nif metabarcoding, they delineated which species of Rhizobiaceae inhabited nodules and how the composition of nodule symbionts changed with soil fertility. They noted that the older soils with the lowest soil phosphorus had more unclassified operational taxonomic units, suggesting a shift to a unique set of nitrogen-fixing bacteria more adapted to limited soil fertility. Associations between legume species and their root-nodule symbionts can be both generalist and specific, but our understanding of how specificity is controlled remains fragmentary. Keller et al. (2018) attempted to delineate the hows and whys of host specificity by meta-sequencing three Lupinus nodulomes, discovering differential compatibility between lupine and Bradyrhizobium and that different hormone, secondary metabolite and plant-defence mechanisms activated depended on host compatibility. Similarly, GBS was used to show lack of local adaptation to different prevailing soil species of Ensifer between northern and southern populations of Medicago lupulina L. in North America (Harrison et al. 2017). Knowledge of which nitrogen-fixing bacteria are optimal is of key importance to maximise crops yields. High-throughput sequencing of nodules or rhizospheres has determined the symbionts of legume crops, forage legumes and invasive species, including cowpea (Chidebe et al. 2018), lucerne (alfalfa; Wigley et al. 2017), rooibos tea (Le Roux et al. 2017) and silver wattle (Kamutando et al. 2017). Furthermore, characterisation and sequencing of bacterial genomes, including those that are key to nitrogen fixation, is now routine. For example, a new species, Rhizobium hidalgonense, was recently isolated and characterised from a Phaseolus vulgaris L. nodule growing in acidic soil (Yan et al. 2017), and draft genome sequences of Bradyrhizobium (Tian et al. 2015) and another new species, Ensifer aridi (Le Quéré et al. 2017), have been completed and characterised.

The evolutionary origin of nodulation has baffled researchers for many years, and particularly whether it evolved once or multiple times, whether some sort of cryptic precursor could have predisposed lineages in the nitrogen-fixing clade of angiosperms to evolve nodulation, and whether polyploidy may have been involved in its origination (Werner et al. 2014; Doyle, 2016). Griesmann et al. (2018) used a comparative genomics approach to address this question by sequencing genomes from across the nitrogen-fixing clade including several non-nodulating species, to look for genes known to be vital to nitrogen-fixing nodulation (NFN). They discovered that all NFN genes were conserved in all but one nodulating species and found evidence for multiple independent losses of the nodule-inception (NIN) gene in 10 of 13 non-nodulating species, attesting to the key role of this gene within the nodulation pathway, and suggesting multiple evolutionary losses of nodulation (van Velzen et al. 2019). Comparative analysis of the legume Medicago truncatula and the non-legume Parasponia andersonii Planch. also supported the idea of a single origin and multiple losses of nodulation across the nitrogen-fixing angiosperm clade (van Velzen et al. 2018, 2019). However, others disagree, citing differences in transcriptome profiles as evidence for a two-step process in origination of nodulation (Battenberg et al. 2018). In addition to working out its origin, understanding the subsequent evolution of nodulation is also important. One unique example is the ability of some members of photosynthesising Bradyrhizobium strains to prompt nodulation and fix nitrogen in the absence of nodABC genes that are key to modulating nodule formation. Strains of this bacterium form symbioses with Aeschynomene evenia C. Wright, and transcriptomics has enabled the creation of a gene-map to delineate genes involved in this unique type of symbiosis (Chaintreuil et al. 2016).

Genetic and genomics resources

The genomics era and all genetic knowledge owes homage to the pea (Pisum sativum L.), the legume crop that captured Gregor Mendel’s attention and led to his monumental discovery of genetic inheritance. The importance of nitrogen-fixing legumes as the most important protein and rotation crops has driven high-throughput sequencing and genomics to establish genetic and genomic data to underpin plant-breeding research. Here, we discuss some of these resources and subsequent discoveries.

Sequenced genomes

With soybean providing nearly 70% of the world’s edible protein, the completion of the soybean genome (Schmutz et al. 2010) marked an important milestone in legume research. Even though the soybean genome was not sequenced using NGS methods, it remains a gold standard in plant-genome sequencing, providing a vital reference for other work, including the assembly and annotation of other legume genomes. Draft genome sequences of the model legumes Lotus japonicus (Sato et al. 2008) and Medicago truncatula (Young et al. 2011), and common bean, Phaseolus vulgaris (Schmutz et al. 2014), based on Sanger sequencing, have been improved to reference quality by using NGS data (e.g. Tang et al. 2014).

Plant genomes are complex relative to other eukaryotic genomes, owing to large genome sizes, higher repetitive fraction, prevalence of polyploidy and difficulty of obtaining high-molecular weight DNA caused by presence of the cell wall, polysaccharides and secondary metabolites that impair enzymes or damage DNA (Jiao and Schneeberger 2017). Nevertheless, the advent of NGS has enabled whole-genome sequencing of a rapidly growing number of species and multiple accessions within species (Fig. 2, Table 1), as costs have fallen (from the ~US$20 million cost of the soybean genome; Marris 2008) and techniques have advanced such that a single laboratory can now rapidly produce multiple genomes. For example, Griesmann et al. (2018) sequenced draft genomes of four legume genera for comparative genomic analysis of nodulation genes, and Chang et al. (2019) sequenced three African orphan legume crop species to support crop breeding. Liu et al. (2019) set out to...
bar code 761 vascular plants from the Riuli Botanical Garden by using NGS whole-genome sequencing, 71 of which are legumes (because identification and genome assembly remain incomplete, these genomes are not listed in Table 1).

Owing to plant genome size and complexity, NGS-based plant genomes often have poorer assembly statistics than do those of vertebrates and are commonly assembled only to draft status. However, a draft genome is still worthwhile! For example, genomic sequencing of multiple accessions of the recent domesticate, *Lupinus angustifolius* L., has allowed mapping of key disease resistance and domestication traits (Yang et al. 2013; Hane et al. 2017). Even without assembly to draft status, sequencing the full genomic content of an organism can yield important information. For example, Kang et al. (2015) sequenced the genomes of *Vigna angularis* var. nipponensis (Ohwi) Ohwi & H.Ohashi and *V. nepalensis* Tateishi & Maxted, wild relatives of cultivated adzuki bean, *V. angularis* var. *angularis* (Willd.) Ohwi & H.Ohashi. Even though the reads were not fully assembled, nor genes called *de novo*, variation in the two wild relatives called against the draft genome of the cultivated bean provided insights into the timing of domestication and variation across the genus.

Early NGS platforms provided vast amounts of data; however, limitations, such as short read length and limited read output, made genome assembly from such platforms alone challenging. As technologies have improved and new mapping methods have been devised, draft genome assemblies can now be improved on after the fact. For example, *Trifolium subterraneum* (Trifolium subterraneum L.), an annual relative of the forage legumes *T. repens* L. (white clover) and *T. pratense* L. (red clover), was sequenced using Illumina and 454 pyrosequencing as a reference in the genus *Trifolium* L. (Hirakawa et al. 2016), producing 27228 scaffolds representing the draft genome, TSUd_r1.1. Subsequently, Kaur et al. (2017) applied a Bionano Genomics (San Diego, CA, USA) optical map and a transcriptome atlas to improve the TSUd_r1.1 assembly by anchoring unplaced contigs, correcting mis-assemblies, and improving gene annotation, resulting in the coalescence of 264 contigs into 97 super-scaffolds representing 43% of the genome and a 1.4-fold increase in the scaffold N50 to create the Tsub_Refv2.0 assembly. Similarly, also building on TSUd_r1.1, application of Hi-C contact mapping, a chromosome conformation technology, corrected misjoins, anchored and oriented scaffolds into eight chromosome-length pseudomolecules that included 95% of the sequenced bases in the input assembly, with the resulting TrSub3 assembly improving the initial scaffold N50 from 287 kb to 56 Mb (Dudchenko et al. 2018).

Whereas Illumina sequencing produces the most data for the lowest cost, long-read sequencing is increasingly affordable and can sequence through repetitive regions that often make plant genomes difficult to assemble. For example, Lonardi et al. (2019) used PacBio (Pacific Biosciences of California, Inc., Menlo Park, CA, USA) to sequence the genome of cowpea (*Vigna unguiculata* [L.] Walp) by using 56.8 Gb of sequence data with a read N50 of 14595 bp, i.e. longer than many *de novo* assembled contigs and scaffolds. In addition, they employed two Bionano optical maps and a novel ‘stitching’ assembly method that combined eight different assemblies created using three different programs. Long-read lengths coupled with iterative cross-checking methods to remove chimeric joins enabled a highly accurate and complete assembly that was able to push through the long repetitive regions that break other plant assemblies into numerous contigs. They also detected a 4.2-Mb chromosomal inversion that may confer resistance to a parasitic weed, *Striga gesnerioides* (Willd.) Vatke. Cross-platform genome-sequencing strategies that use both short- and long-read platforms coupled with physical mapping are, thus, good contemporary approaches for plant-genome sequencing, including enhancement of existing assemblies from draft to reference quality.

As the number of sequenced genomes has grown, genetic and genome-scope databases have been developed, including Phytozome (phytozome.jgi.doe.gov, accessed 30 May 2019) and the Legume Information System (legumeinfo.org, accessed 30 May 2019; Dash et al. 2016), to facilitate access to genome content and information. These databases can be integrated with programs and platforms such as CyVerse (cyverse.org, accessed 30 May 2019) to enable public access to datasets, management and integration of personal data and access to high-performance computing platforms, thus enabling collaborative and critical breakthroughs through the combined power of genetic and computational resource platforms. However, none of these databases integrates other types of information, such as morphological-trait variation or geographical-distribution data, prompting legume systematists to devise a legume portal that can integrate across platforms and data types (Bruneau et al. 2019). Development of integrated database systems across genomic, taxonomic, geographical, morphological, population and expression-profile data will enhance our ability to parse and use information from NGS data and plant genomes.

**Organelle genomes**

With the abundance of chloroplast DNA, compared to nuclear, within a cell, 10–20% of NGS reads are chloroplast, such that entire chloroplast genomes can be assembled from even low-coverage genome skimming. Furthermore, the shorter length and less-repetitive nature of the chloroplast makes assembly easier, even for degraded herbarium material (Bakker et al. 2016). Thus, the use and sequencing of chloroplast genomes has rocketed forward with the advent of NGS methods. Some have even suggested the use of whole chloroplast genomes as DNA barcodes for plants (Li et al. 2015).

Comparative phylogenetic analyses have shown that structural changes in the chloroplast genome are often synapomorphies for large legume clades (Wang et al. 2018). For example, the loss of one copy of the inverted repeat marks the inverted repeat lacking clade (IRLC), a large papilionoid clade including tribes Ciceriaceae, Hedysareae, Trifolieae and Fabeeae (Vicieae), among others (Wojciechowski et al. 2000). Recent sequencing of chloroplasts from eight Cercidoideae genera showed structural diversification characteristic of that subfamily (Wang et al. 2018). In contrast, comparative analyses of chloroplast genomes determined that Papilionoideae have reduced genome sizes and are more divergent from the ancestral angiosperm chloroplast-genome.
organisation than other subfamilies (Schwarz et al. 2015). Structural changes have also been identified at a generic level. *Trifolium subterraneum* L. was shown to have a highly unusual chloroplast genome greatly expanded by repetitive regions (Cai et al. 2008). To determine the evolutionary origin of this unique plastome type, Sveinsson and Cronk (2014) sequenced eight other *Trifolium* chloroplast genomes, showing that the expanded *T. subterraneum*-type plastome is shared by members of ‘core *Trifolium*’, providing a synapomorphy for what they called the ‘refractory clade’. The expansion of the inverted repeat region across the large *Ingea* + *Acacia s.s.* clade of mimosoid legumes (Wang et al. 2017) has provided another synapomorphic structural plastid mutation apparently characterising that clade, and is associated with shifts in evolutionary rate or selection pressures on proximate chloroplast gene regions (Mensous et al. 2017).

Numerous chloroplast genomes have been sequenced in legumes; the NCBI archive of full-length chloroplast genomes includes 284 accessions (97.2% circularised, 2.8% linear; 66.4% legumes; the NCBI archive of full-length chloroplast genomes 12 million years ago (Tosso et al. 2018). Of the nearly 300 legume accessions with sequenced plastochrons, 95 are from *Acacia*, a lineage of >1000 species. Williams et al. (2016) used whole chloroplast sequences of 65 *Acacia* species to build a robust backbone constraint tree, adding amplicon sequences of four chloroplast and two nuclear ribosomal loci for 508 other *Acacia* species, to produce a more robust phylogeny than from amplicon sequences alone. This suggests that combining legacy Sanger sequencing datasets with large, phylogenomic datasets will likely yield positive results.

Another byproduct of NGS is sequencing of the mitochondrion, which has significantly fewer genomes sequenced than does the chloroplast (Table 2). The first legume mitochondrion sequenced was that of *Vigna radiata* (L.) Wilczek (Alverson et al. 2011) and the Sanger-based sequence of the *Vicia faba* L. mitochondrion was not far behind (Negrak 2013). The first legume mitochondria sequenced by NGS methods were *Lotus japonicus* and *Milletia pinnata* (L.) Panigrahi, on the Illumina platform (Kazakoff et al. 2012). Soon after, the mitochondrial genomes of soybean (Chang et al. 2013) and *Medicago truncatula* (Bi et al. 2016) were sequenced on the 454 platform, and *Vigna angularis* was compiled using both 454 and Illumina sequence reads (Naito et al. 2013). The paucity of sequenced plant mitochondrial genomes is largely due to the highly recombinant nature of mitochondria, which makes assembly difficult. Shi et al. (2018) combined PacBio with Illumina reads to investigate the repetitive complement of the *Styphnolobium japonicum* (L.) Schott mitochondrial genome, and discovered that small repeats (<100 bp) had a disproportionate impact on the evolution of *Styphnolobium* mitochondria through mediation of recombination along intronic regions. The first non-papilionoid legume mitochondrial to be sequenced was *Leucaena trichandra* (Zucc.) Urb., completed using PacBio (Kovar et al. 2018) to obtain long-reads, enabling assessment of variable assemblies and investigation of mitochondrial genome-size variation in legumes. Further, overlaying transcriptomic data enabled comparative study of RNA editing among *Leucaena* species, providing knowledge that can yield useful information regarding close species relationships and hybrid origins (Kovar et al. 2018).

Despite these advances, the repetitive nature of the mitochondrial and its evolutionary implications are only

Table 2. List of mitochondrial genomes sequenced in Fabaceae from NCBI

<table>
<thead>
<tr>
<th>Species</th>
<th>Length (bp)</th>
<th>Platform</th>
<th>Accession number</th>
<th>GI number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acacia ligulata</em></td>
<td>698 138</td>
<td>Illumina</td>
<td>MH933866.1</td>
<td>1552 055 398</td>
<td>Sanchez-Puerta et al. 2019</td>
</tr>
<tr>
<td><em>Ammopiptanthus mongolicus</em></td>
<td>475 396</td>
<td>Illumina</td>
<td>NC_039660.1</td>
<td>1511 253 925</td>
<td>Yu et al. 2018</td>
</tr>
<tr>
<td><em>Castanospermum australe</em></td>
<td>542 079</td>
<td>Illumina</td>
<td>MK426679.1</td>
<td></td>
<td>Zhang et al. 2019</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>402 558</td>
<td>454</td>
<td>NC_020455.1</td>
<td>476 507 670</td>
<td>Chang et al. 2013</td>
</tr>
<tr>
<td><em>Glycine soja</em></td>
<td>402 545</td>
<td>Illumina</td>
<td>NC_039768.1</td>
<td>1511 246 382</td>
<td>Asaf et al. 2018</td>
</tr>
<tr>
<td><em>Leucaena trichandra</em></td>
<td>722 009</td>
<td>PacBio</td>
<td>NC_039738.1</td>
<td>1511 244 784</td>
<td>Kovar et al. 2018</td>
</tr>
<tr>
<td><em>Lotus japonicus</em></td>
<td>380 861</td>
<td>Illumina</td>
<td>NC_016743.2</td>
<td>387 866 040</td>
<td>Kazakoff et al. 2012</td>
</tr>
<tr>
<td><em>Medicago truncatula</em></td>
<td>271 618</td>
<td>454</td>
<td>NC_029641.1</td>
<td>1003 725 997</td>
<td>Bi et al. 2016</td>
</tr>
<tr>
<td><em>Milletia pinnata</em></td>
<td>425 718</td>
<td>Illumina</td>
<td>NC_016742.1</td>
<td>372 450 249</td>
<td>Kazakoff et al. 2012</td>
</tr>
<tr>
<td><em>Senna occidentalis</em></td>
<td>447 106</td>
<td>Illumina</td>
<td>NC_038221.1</td>
<td>1442 330 107</td>
<td>Kang et al. 2019a</td>
</tr>
<tr>
<td><em>Senna tora</em></td>
<td>566 589</td>
<td>Illumina</td>
<td>NC_038053.1</td>
<td>1436 049 411</td>
<td>Kang et al. 2019b</td>
</tr>
<tr>
<td><em>Styphnolobium japonicum</em></td>
<td>484 916</td>
<td>Illumina, PacBio</td>
<td>NC_039596.1</td>
<td>1509 839 239</td>
<td>Shi et al. 2018</td>
</tr>
<tr>
<td><em>Vicia faba</em></td>
<td>588 000</td>
<td>Sanger</td>
<td>KC189947.1</td>
<td>442 803 095</td>
<td>Negrak 2013</td>
</tr>
<tr>
<td><em>Vigna angularis</em></td>
<td>404 466</td>
<td>Illumina, 454</td>
<td>NC_021092.1</td>
<td>501 594 995</td>
<td>Naito et al. 2013</td>
</tr>
<tr>
<td><em>Vigna radiata</em> var. <em>radiata</em></td>
<td>401 262</td>
<td>Sanger</td>
<td>NC_015121.1</td>
<td>323 149 028</td>
<td>Alverson et al. 2011</td>
</tr>
</tbody>
</table>
beginning to be understood, particularly with regard to cross-species interactions. That said, Sanchez-Puerta et al. (2019) presented an elegant study of host–parasite horizontal gene transfer between the holoparasite Lophophytum mirabile Schott & Endl. and its host Acacia ligulata A.Cunn., showing that ~60% of the L. mirabile mitochondrial genome is derived from its host, including 34 of 43 protein-coding genes (also see Sanchez-Puerta et al. 2017; Kovař et al. 2018). In addition, ~26 of its native genes were replaced by host genes through homologous recombination, and with a large portion of intergenic regions also host-derived. This work provided incontrovertible support for mitochondrial-to-mitochondrial horizontal gene transfer, a phenomenon well documented in Amborella trichopoda Baill. (Rice et al. 2013). Furthermore, Kovař et al. (2018) also showed that non-coding mitochondrial DNA was horizontally transferred, suggesting capture of an entire mimosoid mitochondrial genome during the evolutionary history of the Lophophytum parasite. Knowledge such as this may shed light on the incredible length diversity of plant mitochondrial genomes, which range from 271,618 bp in Medicago truncatula to 698,138 bp in Acacia ligulata. At the time of writing, 15 legume mitochondria had been sequenced (Table 2). As this number expands, broader comparative analyses become possible; sequencing of the wild soybean, Glycine soja (Asaf et al. 2018), provided insights into a likely progenitor to the soybean mitochondrion, whereas that of Ammopiptanthus mongolicus (Kom.) S.H.Cheng provided a glimpse into a Tertiary relic (Yu et al. 2018). Comparative analysis of the most recently sequenced legume mitochondrion, namely that of Castanospermum australe A.Cunn & C.Fraser, enabled phylogenetic analysis of 33 mitochondrial genomes across legumes, producing a fully supported phylogeny (Zhang et al. 2019), suggesting that despite the fewer phylogenetically informative sites than in chloroplast genomes (Palmer and Herbon 1988), the mitochondrion has potential utility for molecular systematics of legumes.

**Plant breeding**

Simply put, the genomics era has revolutionised plant breeding in legumes. We cannot possibly summarise it all here; however, some studies are essential for discussion. Major goals of plant breeding are improvement of crop traits useful for humans and adaptation to environment such as increasing yield, freeze and drought tolerance, disease resistance, nutritional quality and seed size. A genome sequence provides the basis for genome-wide association studies (GWAS), functional genomics, quantitative trait-loci (QTL) analysis and linkage studies, SNP variant detection, and genetic modification by CRISPR, among many others. For example, sequencing of the licorice genome (Glycyrrhiza uralensis Fisch.) enabled assessment of genes involved in flavonoid and saponin synthesis, producing candidate genes for improving yield of glycyrrhizin, an active chemical component used in traditional Chinese medicine (Mochida et al. 2017). Of wider importance are several crop genomes, in addition to those already discussed. Chickpea (Cicer arietinum L.) is surpassed only by soybean, peanut and Phaseolus ssp. as a widely grown legume crop (Fig. 3, Table S2, available as Supplementary material to this paper).

Its sequenced genomes (Jain et al. 2013; Varshney et al. 2013), coupled with resequencing of numerous cultivars and wild accessions from 10 countries, enabled discovery of gene regions associated with key domestication, agronomic and disease-resistance traits (Varshney et al. 2013). Likewise, sequencing of Cajanus cajan [L.] Millsp., the pigeon pea (Singh et al. 2012; Varshney et al. 2012; Mahato et al. 2018), one of the most widely cultivated and consumed orphan legume crops in India, highlighted drought-tolerance genes important during its domestication (Varshney et al. 2012). These genomic resources have aided successful translational improvements of both chickpea and pigeon pea (for review, see Varshney 2016).

To illustrate the types of NGS applications and advances made in legume-crop breeding, we focus on Vigna Savi, a pantropical genus of over 100 species of which 10 species have been domesticated and are cultivated in both warm, humid and dry, seasonal climates (Table 3; Harouna et al. 2018). As such, Vigna species provide food for nearly half of the world’s population. Ten species or varieties of Vigna have had their genomes sequenced (Tables 1, 3), including six crop species (Kang et al. 2014, 2015; Chang et al. 2019; Lonardi et al. 2019) and four wild relatives (Kang et al. 2014, 2015; Lestari et al. 2014). The impacts of these genome sequences are just beginning to be realised. For example, comparative searches of the V. angularis and V. radiata genomes for homologues of OSE1-like sequences, a heat-activated retrotransposon isolated from Arabidopsis thaliana (L.) Heynh., produced several key hits in V. angularis, sequences that proved to be polymorphic in different accessions (Masuta et al. 2018). These retrotransposons were associated with accumulation of extrachromosomal DNA and were employed to successfully induce retrotransposition of V. angularis OSE1-like elements in regenerated V. angularis callus tissue, providing a new tool for molecular breeding in Vigna.

Many plant-breeding efforts require a genetic map. RADseq was used to create a high-density genetic-linkage map by using 170 individuals to enable QTL detection of loci related to yield in cowpea (Pan et al. 2017), a species that includes two broad cultivar types, namely, bushy, short-podded grain grown predominantly in Africa, and long-podded climbing vegetables grown mostly in Asia. Genomic scans confirmed that pod length was selected for during domestication of the vegetable variety (Xu et al. 2017). Furthermore, GWAS for genomic regions controlling pod length between cultivars of these two types discovered 72 SNPs whose pod-length association was verified across 299 cowpea accessions. This knowledge, coupled with transcriptomic analysis, suggested the involvement of sugar, gibberellin and nutrition as key factors in pod-length regulation and that cell proliferation rather than cell elongation was key to pod length.

Applications of transcriptomics are often used to study stress responses such as low temperature-stress resistance in V. subterranea, providing gene modules for plant-breeding improvement (Bonthala et al. 2016), among many others. These studies often identify candidate genes useful for functional genomic characterisation. For instance, acidic soils are often characterised by accumulation of aluminium, with aluminium toxicity causing reduced yields in crop plants.
Rice bean (*V. umbellata* (Thunb.) Ohwi & H.Ohashi) is tolerant of soils with high aluminium accumulation. Studies of such resistance in other plants have suggested a role for the abscisic acid (ABA) pathway in dealing with aluminium toxicity. Transcriptomic analysis and functional genomic studies in rice bean support this hypothesis (Fan *et al.* 2019).

Table 3. **Summary of genetic and genomic resources available for species of Vigna**

Including all cultivated species and those wild relatives with next-generation sequencing data. References for genomes can be had from Tables 1, 2 and S1. Nu, nuclear; Mito, mitochondrial; Cp, chloroplast; SRA, number of BIO-projects available in short-read archive (SRA) database of the NCBI as of 28 February 2019; ?, genome sequenced with NGS data in SRA, but not yet published.

<table>
<thead>
<tr>
<th>Common name</th>
<th>Vigna species</th>
<th>Cultivated area or distribution</th>
<th>Genome</th>
<th>Nu</th>
<th>Mito</th>
<th>Cp</th>
<th>SRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moth bean</td>
<td><em>V. aconitifolia</em> (Jacq.) Marechal</td>
<td>Semi-arid India, South-East Asia</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adzuki bean</td>
<td><em>V. angularis</em> (Willd.) Ohwi &amp; H.Ohashi</td>
<td>Eastern Asia, Australia, New Zealand</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>wild relative</td>
<td><em>V. angularis var. nipponensis</em> (Ohwi) Ohwi &amp; H.Ohashi</td>
<td>Japan, Korea, China, Bhutan, Nepal, India (Himalaya)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black gram (urd bean)</td>
<td><em>V. mungo</em> (L.) Hepper</td>
<td>Southern and South-East Asia</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wild relative</td>
<td><em>V. nakashimae</em> (Ohwi) Ohwi &amp; H.Ohashi</td>
<td>Northern China, Korea, northern Japan</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wild relative</td>
<td><em>V. nepalensis</em> Tateishi &amp; Maxted</td>
<td>Bhutan, eastern Nepal, north-eastern India</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mung bean</td>
<td><em>V. radiata var. radiata</em> (L.) R.Wilczek</td>
<td>Southern, eastern and South-East Asia</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>wild relative</td>
<td><em>V. radiata var. sublobata</em> (Roxb.) Verdc.</td>
<td>Asia, Africa, Australia</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creole bean</td>
<td><em>V. reflexo-pilosa</em> Hayata</td>
<td>India, Mauritius, Philippines, Vietnam</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minni payaru</td>
<td><em>V. stipulacea</em> Kunte</td>
<td>India, Sri Lanka, Indonesia, New Guinea</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bambara groundnut</td>
<td><em>V. subterranca</em> (L.) Verdc.</td>
<td>Semi-arid Africa</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rice bean</td>
<td><em>V. umbellata</em> (Thunb.) Ohwi &amp; Ohashi</td>
<td>Asian tropics</td>
<td>X?</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cowpea</td>
<td><em>V. unguiculata</em> (L.) Walp</td>
<td>Semi-arid regions of Africa, Asia, USA, Europe, Central and South America</td>
<td>X</td>
<td>X</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuber cowpea</td>
<td><em>V. vexillata</em> (L.) A.Rich.</td>
<td>Indonesia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Fig. 3. Total production of cultivated legumes by year from 2008 to 2017. Data from FAOSTAT (www.fao.org/faostat/, accessed 25 April 2019). Numbers were summed across all countries, with China counted only once. nes, not elsewhere specified. Data are listed in Table S2.
but its involvement was shown to be dependent on AB15, a transcription factor that mediated changes in cell-wall modification and osmoregulation.

Another focus of plant-breeding efforts involve resistance to pests and pathogens. Micro-RNAs (miRNAs) are short (20–24 nucleotides) non-coding RNAs that act to regulate gene expression, particularly during response to stress, such as that inflicted on a plant by viral infection. The mungbean yellow mosaic India virus (MYMIV) significantly decreases yield across many South-East Asian countries where Vigna species are staple crops. Understanding the mechanisms of infection and host response are key to crop survival. Kundu et al. (2017) identified miRNAs involved in the stress response of V. mungo (L.) Hepper to MYMIV infection, by comparing gene-expression patterns and miRNAs across resistant and non-resistant cultivars, with putative target genes known to be involved in pathogen-stress response, such as NB-LRR, ARF, SOD, SPB, and Basic blue copper protein, linked to and validated as being regulated by miRNAs in stressed and non-stressed plants. Another proverbial pest problem plaguing legume crops is bruchid beetles (Callosobruchus Pic., spp.), which infest seeds in the field, then multiply and destroy seed during storage. Bruchid resistance has been found in wild mungbean, V. radiata var. sublobata (Roxb.) Verdc., and in one mungbean cultivar. Schafleitner et al. (2016) used GBS to map inbred recombinant lines for each of these resistant populations and discovered one QTL associated with bruchid resistance shared between both resistant entities. The markers associated with this QTL were validated as 100% predictive of bruchid resistance, providing an excellent screening tool for developing resistant cultivars. Similar genome-based research across the full spectrum of important legume crops is underway. Rapidly expanding knowledge of gene functional pathways from comparative genomics and the development of genome-based selection, is accelerating and revolutionising legume-crop breeding; these efforts will aid in future crop and food security.

This short summary would be incomplete without some discussion of the genomics of crop wild relatives, which can contribute key genes and diversity to crops to increase pest and disease resistance and extend environmental tolerances of important crop legumes. One of the first crop wild relatives to have its genome sequenced was Glycine soja Sieb. and Zucc. (Kim et al. 2010), one of the first genome re-sequencing projects in plants. In the study of Kim et al. (2010), G. soja was sequenced and its genome assembled against that of soybean, with a comparison between the two suggesting that G. max diverged from G. soja c. 0.27 million years ago, i.e. hundreds of thousands of years before domestication of soybean. The draft genome of Glycine latifolia (Benth.) C.A.Newell & Hymowitz, a perennial wild relative of soybean, was recently sequenced using only linked-reads from a single 10X Genomics (Pleasanton, CA, USA) Chromium library (Liu et al. 2018), presenting a valuable resource of alleles and genes for soybean improvement. Like soybean, the cultivated peanut (or groundnut; Arachis hypogaea L.) is of polyploid origin. The large allotetraploid peanut genome sequence (http://peanutbase.org/peanut_genome, accessed 30 May 2019; Chen et al. 2019; ~2.7 Gb) comprises the two recently diverged subgenomes of its diploid ancestors, Arachis duranensis Krapov. & W.C.Greg, and A. ipaensis Krapov. & W.C.Greg, which were used to assist assembly of the domesticated-peanut genome and detect genetic recombination among peanut subgenomes, providing key information regarding the origin of the cultivated peanut (Bertioli et al. 2016). As climate changes and demands for better yield increase, wild relatives offer plant breeders sources of diverse and adapted traits to incorporate into cultivars.

**Securing the future of legume germplasm diversity**

Given the central roles that legumes play in agriculture, ecosystems and the global nitrogen cycle, securing the future of legume genetic resources is both urgent and of paramount importance. Application of genomics methods, from capturing the genome of a fading species to guiding restoration and management efforts and assisting and speeding up plant breeding, can enable advances in legume research that were previously unattainable that will help secure their future.

A good example is Ammopiptanthus S.H.Cheng, a genus of two evergreen broadleaf desert shrub species endemic to central Asia. This taxonomically isolated genus is hypothesised to be a relic from the Tertiary, having adapted to aridification from a moist and humid climate of the evergreen broadleaf forest characteristic of the Tethyan flora (Zhang et al. 2015). As a relic, Ammopiptanthus has evolved drought-, cold- and wind-resistance, among other stress-tolerant characteristics. Both species are considered threatened because of low seed set and increasing anthropogenic disturbance, with A. nanus (Popov) S. H.Cheng listed as Critically Endangered (www.iucnredlist.org, accessed 19 February 2019). Ammopiptanthus has been the focus of abiotic stress studies, using transcriptomics to understand its drought (Zhou et al. 2012) and cold (Pang et al. 2013) adaptations. These studies have enabled the cloning, characterisation and validation of candidate genes shown as beneficial in other model study systems, conferring salinity and heat tolerance to Escherichia coli by the A. nanus betaine aldehyde dehydrogenase gene (Yu et al. 2014), or cold tolerance by the A. nanus antifreeze gene AnAFP to both E. coli and tobacco (Deng et al. 2014), among many others. Recently, the A. nanus genome was sequenced on the PacBio platform (Table 1; Gao et al. 2018), providing an essential resource for functional genomics and improvement studies, with Ammopiptanthus fast becoming a model for understanding drought and cold tolerance.

Along with alfalfa (Medicago sativa) and several related species, clovers (Trifolium spp.) are important forage crops and key components of natural grazing systems. Red clover (Trifolium pratense) is an excellent short-term hay-rotation crop and pasture or field restoration plant. Its dual use as protein-rich fodder and soil-fertility enhancer led to its early adoption in crop-rotation schemes and is one reason why some tout it as a perfect plant for conservation agriculture, a movement aimed at sustainable intensification of food and forage for increasing world needs (McKenna et al. 2018). The recent sequencing of its genome (Istváné et al. 2014; De Vega et al. 2015) has provided another forage genome resource for comparative analyses to assist breeding (Annicchiarico et al. 2015). Furthermore, transcriptomic studies have enabled advances in seed set (Kovi et al. 2017), drought (Yates et al. 2014) and leaf
senescence (Chao et al. 2018a). For example, Chao et al. (2018b) used PacBio to sequence and analyse full-length transcripts, enabling the detection of over 30,000 novel isoforms and 5,492 alternative splicing events, the majority of which involved intron retention. Using two-dimensional difference gel electrophoresis, Bertrand et al. (2016) characterised the proteome of red clover, particularly with reference to freezing tolerance; selection detailed the involvement of a small number of cold-regulated proteins, suggesting them as potential targets for breeding programs. These studies illustrated the incredible protein diversity arising from the complement of known genes and provided a third layer of information on top of genomic and expression-level knowledge.

The impacts of climate change will be particularly felt on food security, with changes in temperature and rainfall patterns affecting what, where and how crops will flourish. Genomics can help improve food security by providing fundamental molecular-data resources through genome sequencing, comparative genomics and rapid evaluation of genetic variation, to enable assessment of adaptation priority, determine import of specific traits for breeding programs, and discover adaptive genes and traits (Mousavi-Derazmahalleh et al. 2019), as previously discussed. Many have suggested that the potential of legumes to enhance food security and reduce meat consumption has not been realised nor explored adequately (Foyer et al. 2016). Phaseolus, with a sequenced genome and well-developed breeding programs, provides a good example. Key areas of inquiry remain unexplored. For instance, phenotyping of the world’s germplasm collections is important to understand the diversity currently held and how that compares to wild populations (McClellan et al. 2011), and with the advent of genomics techniques, the time is now ripe for doing so en masse. Genotyping-by-sequencing and genome–environment associations within wild populations of common bean have discovered several genomic regions associated with drought adaptation, pinpointing genomic signatures potentially useful for marker-assisted selection (Cortés and Blair 2018). Yet, much still needs to be done. As wild relatives and related species often harbour greater GD than do cultivars, efforts to assess such genetic and phenotypic diversity and unlock the adaptive potential of these entities should be a priority for future food security (Porch et al. 2013). Efforts to identify gaps in germplasm collections for Phaseolus have begun (Ramírez-Villegas et al. 2010), yet, much is still needed (Dohle et al. 2019).

Conclusions

From evolution and ecology to classical and applied genetics, the genomics era has and will continue to contribute to our understanding of legume biology in many important ways, and this is set to expand and accelerate in coming years as sequencing costs continue to fall. Partial or whole genomes of hundreds to thousands of legume species are expected to be available in the near future, with the 10,000 genomes project already targeting over 300 legume genera (Cheng et al. 2018). While such an unprecedented accumulation of genomic data presents significant computational and analytical challenges, there will soon be unparalleled opportunities to address large-scale comparative genomic questions. Analyses of numerous legume genomes will provide massive insights into synteny, micro- and macro-level gene and genome duplication, and chromosome structure coupled with gene expression analyses and recombination perspectives. At the same time, variant detection-panel screening across population-level sampling schemes will enable exceptional advances in understanding patterns and processes of evolution at micro- and macro-scales. Such knowledge will help secure the future of legumes as vital components of ecological and economic security.

Conflicts of interest

A. N. Egan is also an Associate Editor of the ‘Advances in Legume Systematics’ special issue. Despite this relationship, she did not at any stage have Associate Editor-level access to this manuscript while in peer review, as is the standard practice when handling manuscripts submitted by an editor to this journal. 

Declaration of Funding

This research did not receive any specific funding. However, A. N. Egan was supported in part by a subaward from the Carlsberg Foundation to Henrik Balslev (grant #CF14-0245) and by the USA National Science Foundation (DEB-1352217).

Acknowledgements

A. N. Egan thanks Dr Henrik Balslev and the Department of Biosciences at Aarhus University for providing a sabbatical academic home from which to pursue this work and Dr Jeffrey J. Doyle and Dr Colin Hughes, as well as an anonymous reviewer, for helpful comments on the manuscript.

References


A. N. Egan is also an Associate Editor of the ‘Advances in Legume Systematics’ special issue. Despite this relationship, she did not at any stage have Associate Editor-level access to this manuscript while in peer review, as is the standard practice when handling manuscripts submitted by an editor to this journal. Australian Systematic Botany encourages its editors to publish in the journal and they are kept totally separate from the decision-making process for their manuscripts. The authors have no further conflicts of interest to declare.

Declaration of Funding

This research did not receive any specific funding. However, A. N. Egan was supported in part by a subaward from the Carlsberg Foundation to Henrik Balslev (grant #CF14-0245) and by the USA National Science Foundation (DEB-1352217).

Acknowledgements

A. N. Egan thanks Dr Henrik Balslev and the Department of Biosciences at Aarhus University for providing a sabbatical academic home from which to pursue this work and Dr Jeffrey J. Doyle and Dr Colin Hughes, as well as an anonymous reviewer, for helpful comments on the manuscript.

References


