Genomic research on soybean and its impact on molecular breeding

Man-Wah Li[a,b,*,†], Bingjun Jiang[c,†], Tianfu Han[c], Guohong Zhang[d], and Hon-Ming Lam[a,b,*]

[a]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
[b]Centre for Soybean Research of the State Key Laboratory of Agrobiotechnology and School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong, China
[c]MARA Key Laboratory of Soybean Biology (Beijing), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China
[d]Institute of Dryland Agriculture, Gansu Academy of Agricultural Sciences, Key Laboratory of Northwest Drought Crop Cultivation of Chinese Ministry of Agriculture, Lanzhou, China
*Corresponding authors: e-mail address: limanwah@cuhk.edu.hk; honming@cuhk.edu.hk

Contents

† Authors contributed equally.

Advances in Botanical Research
ISSN 0065-2296
https://doi.org/10.1016/bs.abr.2022.02.021

Abstract

Soybean, with its high protein and oil contents as well as nitrogen-fixing ability, is an important component in climate-smart agriculture. Yet, to meet the demands on production and consumption, it is necessary to continue to improve its agronomic traits related to yield, adaptability, and nutritional values. Benefiting from advances in genome sequencing technologies, soybean research has blossomed in the past decade since the release of the first soybean reference genome. Improvements in the quality of soybean reference genomes in recent years have facilitated ever higher levels of soybean research. New features such as open chromatin regions and long non-coding RNAs in the soybean genome have also been revealed. To take advantage of the genetic diversity inherent in the soybean germplasm collections, genotyping-by-sequencing, high-resolution mapping, and subsequent gene functional analyses are now routine processes to precisely identify functional markers and genes for molecular breeding. Molecular breeding techniques allow the efficient stacking or modification of desirable genes, and greatly improve the rate and precision of generating new soybean varieties. In this chapter, a brief summary will be given on the current progress in soybean genomic research and its impact on soybean molecular breeding.

1. Introduction

Archaeological evidence suggested that human consumption of soybean can be dated back to >8,000 years before present (BP) in East Asia (Lee, Crawford, Liu, Sasaki, & Chen, 2011). Since it is not known whether the soybean seeds found in archaeological sites were wild-gathered or field-grown, the actual beginning of soybean domestication is still largely debatable. In general, the undomesticated or wild soybean is classified as *Glycine soja* (*G. soja*) while the domesticated or cultivated soybean is classified as *Glycine max* (L.) Merr. (*G. max*). Despite being classified into different species, wild and cultivated soybeans can cross-breed to produce fertile offspring (Kim et al., 2010; Li et al., 2014), and thus wild soybeans have been regarded as reservoirs of genetic materials for the improvement of cultivated soybeans. Some intermediate accessions between wild and cultivated soybeans could sometimes be regarded as semi-wild soybeans and landraces. The definition is fuzzy and is largely dependent on how they were collected, regardless of their true origins. Genomic analyses suggested that cultivated soybean originated from a *G. soja*/*G. max* lineage that had diverged from other *G. soja* lineages 0.27–0.8 million years ago (MYA). Yet these estimations were based on a small number of soybean genomes and several assumptions. More accurate predictions should be available in upcoming large-scale genomic analyses. While *G. max* was derived from

a limited genetic background, the substantial genetic resources contained in wild soybeans could potentially be important for soybean improvement.

Over 70,000 cultivated and 20,000 wild soybean accessions are maintained by different seed banks worldwide (Li et al., 2020; Liu et al., 2020; Valliyodan et al., 2021). To make good use of these germplasm resources, sequencing, analyzing and publishing their genomes would be an essential step. Since the official release of the first soybean reference genome in 2010 (Schmutz et al., 2010), genomic research has been expanding, contributing greatly to the molecular breeding of soybean.

2. A survey of published soybean reference genomes

2.1 A brief overview of the soybean genome

The current soybean genome resulted from two rounds of recent whole-genome duplication (WGD). The first round of whole-genome duplication occurred around 58 MYA, common to all members of Papilionoideae (Schmutz et al., 2010)Kim et al., 2010; Schmutz et al., 2021. The second round happened about 13 MYA and was specific to *Glycine* spp. The resultant soybean genome comprises of 20 pairs of chromosomes which are morphologically identical under the microscope during metaphase (Findley et al., 2010). The genome size of soybean was estimated using flow cytometry to be 1.1–1.15 Gb (Arumuganathan & Earle, 1991; Shultz et al., 2006), which, in general, agrees with the 0.88–1.1 Gb estimate by K-mer frequency, based on sequencing reads obtained in recent genomic studies (Li et al., 2014; Qi et al., 2014). It is predicted that there are 45,000–65,000 protein-coding genes in the soybean genome, of which ~50,000–55,000 are of high confidence (Li et al., 2014; Liu et al., 2020; Qi et al., 2014; Schmutz et al., 2010; Xie et al., 2019). Around 60–70% of these genes exist as paralogues as a consequence of WGDs (Schmutz et al., 2010). Although gene redundancy could be an obstacle for genetic studies, it was found that just altering one gene out of a paralogous pair was normally sufficient to create observable phenotypic changes in plant traits during domestication (Lu et al., 2017; Wang et al., 2016, 2020). Similar to other crops, 50%–60% of the soybean genome is made up of transposable elements (TE) dominated by long terminal repeat (LTR) retrotransposons (Schmutz et al., 2010; Xie et al., 2019). Moreover, the soybean genome also contains ~40,000 long non-coding RNA (lncRNA) gene loci which produce transcripts that are not traditionally considered to be functional or protein-encoding (Lin et al., 2020), constituting a relatively

unexplored component of the soybean genome. For regulatory elements, >20,000 open chromatin regions potentially containing transcription factor-binding *cis* elements have been identified in a wild soybean genome (Huang et al., 2021), providing an area of focus for future research on soybean transcription regulation.

2.2 Currently available high-quality reference genomes and their characteristics

2.2.1 First generation of soybean reference genomes

The official release of the Williams 82 (Wm82) reference genome in 2010 (Schmutz et al., 2010) marked the beginning of the genomic era of soybean research. A few *de-novo* genome assemblies or reference-guided genome assemblies were subsequently released (Table 1). These first-generation soybean genome assemblies were mostly built on short sequencing reads from next-generation sequencing platforms, except Wm82 which was constructed from large-scale Sanger sequencing. Limited by the short sequencing read length and the complexity of the soybean genome, these genome assemblies were usually fragmented, with short contig N50 ranging from 20 to 190 Kb. Among these soybean genome assemblies, the most commonly used was Wm82, owing to its better contiguity over the others. These published genome assemblies have been widely used as references for resequencing projects, for variation calling and association mapping (see below and other sections), speeding up the progress of soybean research. However, the short reads were not able to resolve the genomic regions with low complexity or duplications, leading to bubbles in the de Brujin graph and resulting in contigs that were usually relatively short and fragmented. The consequence is that it becomes challenging to anchor the short contigs onto scaffolds and chromosomes, leaving large numbers of gaps, which in turn results in ambiguous gene models spanning these gaps. Such gap-filled genomes are not suitable for structural variation comparisons and make a lot of downstream analyses problematic. Fortunately, this issue is now partially resolved by second-generation soybean genome assemblies.

2.2.2 Second generation of soybean reference genomes

The breakthrough in long-read sequencing has brought soybean genome assemblies to another level. Contig assemblies using long sequencing reads and scaffolding using Hi-C sequencing, occasionally supplemented with optical mapping, have now become the standard for the second-generation soybean genome assembly pipeline. The first second-generation soybean

Table 1 First-generation soybean genome assemblies.

Variety	Genome ID and version information	Cultivated/wild	Year of official release	Assembled size[a] (Mb)	Contig N50 (Kb)	Scaffold N50	Reference
Williams 82	Wm82.a1	Cultivated	2010	955	189.4	47.8 Mb	Schmutz et al. (2010)
Menjing1	W05	Wild	2014	808	24.2	401 Kb	Qi et al. (2014)
Lanxi 1	–	Wild	2014	895	21.7	51 Kb	Qiu et al. (2014)
Williams 82	Wm82.a2	Cultivated	2015	955	182.8	48.6 Mb	Schmutz et al. (2010)
Enrei	–	Cultivated	2015	905	NA	NA	Shimomura et al. (2015)
ZYD04569	GsojaA	Wild	2014	813	9	18.3 Kb	Li et al. (2014)
PI507600	GsojaB	Wild	2014	895	22.2	57.2 Kb	Li et al. (2014)
PI407222	GsojaC	Wild	2014	841	8	17 Kb	Li et al. (2014)
ZYD03247	GsojaD	Wild	2014	985	11	48.7 Kb	Li et al. (2014)
ZYD02878	GsojaE	Wild	2014	920	27	65.1 Kb	Li et al. (2014)
ZYD00401	GsojaF	Wild	2014	886	24.3	52.4 Kb	Li et al. (2014)
PI578344B	GsojaG	Wild	2014	878	19.2	44.9 Kb	Li et al. (2014)

[a]Assembled size without gap.

reference genome released was Zhonghuang 13 (Gmax_ZH13), published in 2018 (Shen et al., 2018). It was quickly updated to version 2 in 2019 with a seven-fold improvement in the contig N50 (Shen et al., 2019). Within three years, around 40 more high-quality soybean reference genomes were released (Table 2). In general, most of the second-generation soybean reference genomes have a contig N50 at the megabase level, a hundred times longer than those of the first-generation assemblies. The drastic improvement in the contiguity of the reference genomes allows better gene model predictions, identification and comparison of large structural variations, and association studies with higher precision (Liu et al., 2020; Xie et al., 2019).

2.2.3 Currently available soybean pan-genomes

A single reference genome cannot effectively represent the whole species. Although there are more than 40 high-quality reference genomes of soybean available, using all of them as references in a single study would involve tedious analyses of mostly redundant sequences. Although whole-genome resequencing enables comparisons between large natural populations (see Section 2.2.4), variation calling relies heavily on the information contained within the reference genome. Therefore, features absent in the reference genome are usually overlooked and structural variations are usually missed.

A pan-genome has the potential to solve the above issues. It is built by collapsing the genomic information of multiple genome assemblies into nonredundant sequences or a set of nonredundant genes (Lei et al., 2021). In brief, the set of sequences or genes shared by all genome assemblies are defined as the core genome or core gene set (Lei et al., 2021). Those shared by some genomes are defined as dispensable or variable genome or genes (Lei et al., 2021). Private sequences and genes are those unique to a specific genome assembly. A gene-based pangenome is easy to build as it is only concerned with the mostly conserved genes or protein-coding sequences, and is especially suitable for analyzing genomes with a distant relationship. A sequence-based pangenome, on the other hand, is usually more computationally demanding and is suitable for analyzing closely related genomes with small numbers of sequence variations. It is more information-intensive as it also considers the non-coding elements in the genomes, which can be important in molecular breeding as well.

So far, there are four pangenomic studies on soybean (Table 3). Compared with those built from wild soybeans or a collection of both wild and cultivated soybeans, a pangenome built solely from cultivated soybeans would have a higher proportion of core genes (Table 3). Although more

Table 2 Second-generation soybean genome assemblies.

Variety	Genome ID and version information	Wild/cultivated	Year of official release	Assembled size (Gb)	Contig N50 (Mb)	Scaffold N50 (Mb)	Reference
Zhonghuang 13	Gmax_ZH13	Cultivated	2018	1.025	3.46	51.87	Shen et al. (2018)
Zhonghuang 13	Gmax_ZH13_v2.0	Cultivated	2019	1.011	22.6	NA	Shen et al. (2019)
Mengjin 1	W05	Wild	2019	1.013	3.3	50.7	Xie et al. (2019)
Williams 82	Wm82v4	Cultivated	2019	0.978	0.419	20.4	Valliyodan et al. (2019)
Lee	Lee v1.1	Cultivated	2019	1.016	0.037	15.0	Valliyodan et al. (2019)
PI 483463	Glycine soja v1.1	Wild	2019	0.985	0.011	4.4	Valliyodan et al. (2019)
Wenfeng 7	IGA1001	Cultivated	2021	0.996	1.74	50.7	Chu et al. (2021)
Hefeng 25	IGA1002	Cultivated	2021	0.987	2.92	49.51	Chu et al. (2021)
GsojaF	IGA1003	Wild	2021	0.975	1.55	48.77	Chu et al. (2021)
Zhonghuang 35	IGA1004	Cultivated	2021	1.001	1.41	50.49	Chu et al. (2021)
Zhonghuang 13	IGA1005	Cultivated	2021	0.988	4.65	50.28	Chu et al. (2021)
Jingyuan	IGA1006	Cultivated	2021	0.995	4.27	50.57	Chu et al. (2021)
Huaxia 3	IGA1007	Cultivated	2021	0.986	6.16	50.7	Chu et al. (2021)
Williams 82	IGA1008	Cultivated	2021	0.993	1.88	49.8	Chu et al. (2021)
PI 562565	SoyW01	Wild	2020	1.021	23.9	52.1	Liu et al. (2020)
PI 549046	SoyW02	Wild	2020	1.007	20.5	21.9	Liu et al. (2020)

Continued

Table 2 Second-generation soybean genome assemblies.—cont'd

Variety	Genome ID and version information	Wild/cultivated	Year of official release	Assembled size (Gb)	Contig N50 (Mb)	Scaffold N50 (Mb)	Reference
PI 578357	SoyW03	Wild	2020	1.015	26.8	52.3	Liu et al. (2020)
Zhutwinning2	SoyL01	Landrace	2020	0.999	23.3	50.6	Liu et al. (2020)
Zi Hua No.4	SoyL02	Landrace	2020	1.012	23.1	52.3	Liu et al. (2020)
Tong Shan Tian E Dan	SoyL03	Landrace	2020	1.040	21.3	51.3	Liu et al. (2020)
58-161	SoyL04	Landrace	2020	1.005	23.0	51.4	Liu et al. (2020)
PI 398296	SoyL05	Landrace	2020	1.060	22.1	51.2	Liu et al. (2020)
Zhang Chun Man Cang Jin	SoyL06	Landrace	2020	0.998	21.5	50.7	Liu et al. (2020)
Feng Di Huang	SoyL07	Landrace	2020	1.005	23.7	50.9	Liu et al. (2020)
Tie Jia Si Li Huang	SoyL08	Landrace	2020	1.000	22.6	51.1	Liu et al. (2020)
Shi Sheng Chang Ye	SoyL09	Landrace	2020	1.028	23.1	51.0	Liu et al. (2020)
Xu Dou No.1	SoyC01	Cultivated	2020	1.004	23.5	51.5	Liu et al. (2020)
Tie Feng No.18	SoyC02	Cultivated	2020	1.011	23.8	51.2	Liu et al. (2020)
Ju Xuan No.23	SoyC03	Cultivated	2020	1.008	23.9	51.3	Liu et al. (2020)

Wan Dou No.28	SoyC04	Cultivated	2020	1.002	22.7	50.9	Liu et al. (2020)
Amsoy	SoyC05	Cultivated	2020	0.992	22.7	50.3	Liu et al. (2020)
Yu Dou No.22	SoyC06	Cultivated	2020	1.008	23.0	52.0	Liu et al. (2020)
Jin Dou No.23	SoyC07	Cultivated	2020	1.009	21.8	51.2	Liu et al. (2020)
Qi Huang No. 34	SoyC08	Cultivated	2020	1.002	22.4	50.9	Liu et al. (2020)
Han Dou No.5	SoyC09	Cultivated	2020	1.004	22.6	50.4	Liu et al. (2020)
PI 548362	SoyC10	Cultivated	2020	1.005	19.8	51.0	Liu et al. (2020)
Ji Dou No.17	SoyC11	Cultivated	2020	1.025	18.8	51.4	Liu et al. (2020)
Dong Nong No.50	SoyC12	Cultivated	2020	1.025	20.0	50.8	Liu et al. (2020)
Hei He No.43	SoyC13	Cultivated	2020	1.010	23.8	51.1	Liu et al. (2020)
Ke Shan No.1	SoyC14	Cultivated	2020	1.007	23.0	51.6	Liu et al. (2020)
Jidou 17	JD17	Cultivated	Preprint	0.965	18.0	NA	Yi et al. (2021)
Fiskeby III (PI 438471)	Fiskeby v1.1	Cultivated	Pre-release	0.992	NA	NA	DOE-JGI (2021)

Table 3 Pangenome studies on soybean.

Number of accessions	Assembled pan-genome size (Mb)	Core genome size (Mb)	Total no. of gene families	Core gene families	Reference
7 wild soybeans	986.3	790.1	59,080	28,716	Li et al. (2014)
3 wild soybeans	–	–	57,492	20,623	Liu et al. (2020)
9 landraces					
15 cultivars					
204 cultivated	1,086	–	54,531	49,431	Torkamaneh, Lemay, and Belzile (2021)
157 wild soybeans	1,213	–	51,414	44,654	Bayer et al. (2021)
723 landraces					
228 cultivars					
2 unclassified					

accessions were involved in building PanSoy (Torkamaneh et al., 2021), due to the lower genetic diversity in cultivated soybean, the size of this pangenome is not any bigger and the number of core gene families is higher than that built from just seven wild soybean accessions (Li et al., 2014).

2.2.4 Large-scale soybean resequencing projects

While it is still costly to build high-quality reference genomes for all soybean accessions, whole-genome resequencing is still an affordable way to capture decent genetic variations for genomic analyses. Since the official release of the Wm82 reference genome in 2010, a number of large-scale soybean resequencing projects have been published (Table 4). In the early phase, resequencing studies mainly focused on the analysis of genetic diversity. They all came to the same long-known but previously unproven conclusion that wild soybeans have higher genetic diversity than cultivated soybeans (Chung et al., 2014; Lam et al., 2010; Qiu et al., 2014; Zhou et al., 2015). Furthermore, domestication- and adaptation-related genomic loci were rediscovered in these studies, improving our understanding of the domestication process. With the reduction in sequencing costs, resequencing

Table 4 List of large-scale soybean resequencing studies.

Number of accessions	Average sequencing depth/mapping depth	Reference genome	Total number of SNPs	Genetic diversity	Reference
17 Wild soybeans	-/5×	Williams 82 (Wm82.a1)	6,318,109	Wild: $\theta_\pi = 2.97 \times 10^{-3}$	Lam et al. (2010)
17 Cultivated soybeans				Cultivated: $\theta_\pi = 1.89 \times 10^{-3}$	
1 Wild soybean	52×/43×	Williams 82 (Wm82.a1)	2,504,985		Kim et al. (2010)
9 Semi-wild soybeans	-/3.0×	Williams 82 (Wm82.a1)	7,704,637[a]	Wild: $\pi = 2.173 \times 10^{-3}$	Qiu et al. (2014)
1 Semi-wild soybean	-/41.4×			Semi-wild: $\pi = 1.416 \times 10^{-3}$	
1 Wild soybean	-/55.0×			Cultivated: $\pi = 1.332 \times 10^{-3}$	
6 Wild soybeans	17×/14×	Williams 82 (Wm82.a1)	9,028,250	Wild: $\pi = 1.08 \times 10^{-3}$	Chung et al. (2014)
10 Cultivated soybeans				Cultivated: $\pi = 0.46 \times 10^{-3}$	
62 Wild soybeans	-/11×	Williams 82 (Wm82.a1) + *G. soja* var. IT182932 genome-specific sequences	9,790,744	Wild: $\pi = 2.94 \times 10^{-3}$	Zhou et al. (2015)
130 Landraces				Landraces: $\pi = 1.40 \times 10^{-3}$	
110 Improved cultivars				Improved cultivars: $\pi = 1.05 \times 10^{-3}$	
28 Brazilian cultivars	-/14.8×	Williams 82 (Wm82.a2.v1)	5,835,185	Not available	dos Santos et al. (2016)
102 Canadian short-season soybean	11x[b]/>1x	Williams 82 (Wm82.a2.v1)	4,071,378	Not available	Torkamaneh et al. (2018)

Continued

Table 4 List of large-scale soybean resequencing studies.—cont'd

Number of accessions	Average sequencing depth/mapping depth	Reference genome	Total number of SNPs	Genetic diversity	Reference
195 Cultivars	-/16.3×	Williams 82 (Wm82.a2.v1)	10,116,707	Not available	Kajiya-Kanegae et al. (2021)
3 Wild soybeans					
41 Wild soybeans	12.9×/-	Gmax_ZH13	31,870,983	Not available	Liu et al. (2020)
632 Landraces					
1354 Improved cultivars					
85 wild soybean	-/>10x	Gmax_ZH13	12,506,103	Not available	Lu et al. (2020)
153 Landraces					
186 Cultivars					
279 Landrace	-/6.03x	Williams 82 (Wm82.a2.v1)	6,341,742	Not available	Li, Li, et al. (2020)
134 Chinese cultivars	-/8×	Williams 82 (Wm82.a2.v1)	4,163,977	$\pi = 1.28 \times 10^{-3}$	Qi et al. (2021)
429 Cultivars	15×/-	Williams 82 (Wm82.a2.v1)	7,869,806	Not available	Valliyodan et al. (2021)
52 Wild soybeans					
51 Landraces	11×/-	Williams 82 (Wm82.a2.v1)	6,333,721	Not available	Yang et al. (2021)
199 Cultivars					

[a]43 soybean accessions, including those from Lam et al. (2010) and Kim et al. (2010), were used for SNP-calling.
[b]Median sequencing depth/minimum mapped depth.

of larger populations becomes readily accessible. As a consequence, these studies have generated a large number of molecular markers for various uses, including mapping, DNA chip construction, marker-assisted breeding and genomic selection. At the same time, the increased size of populations being resequenced has enabled association studies to identify single-nucleotide polymorphisms (SNPs) associated with the causal locus or gene for a specific trait (see Section 3.3).

3. Quantitative trait loci/association mapping using whole-genome sequencing methods

3.1 Basics of genetic mapping for soybean breeding

The availability of functional genes and their associated markers are the foundation of molecular breeding. Unlike many other model plants, there is only a limited collection of publicly available systematic mutants for soybean due to the high maintenance and labor costs (Brown et al., 2021). Thus, the mutant-based forward-genetics approach for the identification of causal genes or associated markers for certain traits is not favored in soybean research. At the same time, the efficiency of soybean transformation is also not comparable with other model crops. As a result, the pace of using a reverse-genetics approach to determine the gene function also lags far behind the other major crops. This means the availability of markers and genes associated with specific traits for soybean molecular breeding is also limited.

Association mapping, either through quantitative trait locus (QTL) mapping or genome-wide association studies (GWAS), is the common way to identify the genomic element associated with a specific phenotype. Before the publishing of the soybean reference genome, there were three major limitations in soybean association mapping. First of all, the number of available molecular markers was limited, and genotyping them was normally laborious, directly affecting the resolution of the map. There are more than 1,000 documented simple-sequence repeat (SSR) markers in the soybean genome (Song et al., 2004). Considering the polymorphism of these SSR markers in the target population and the labor cost of genotyping them, only a fraction of these markers would normally be used. What makes this worse is that these markers are usually not evenly distributed in the genome. Thus, the physical distance between markers can be up to several megabases. In such a case, the resolution of mapping and efficiency of breeding would be low. Secondly, without the whole-genome sequence, the positions of

genetic markers were mostly expressed in relative terms to the rate of recombination (i.e. genetic distances expressed in centimorgans). Thus, the location and range of the mapped genomic region could be very fuzzy. On some occasions, the relative positions of markers in different genetic maps vary, which makes comparison between maps problematic. Finally, sequence and gene information within the mapped region could be completely unknown. Hence the identification of the causal element of the phenotype of interest would require tremendous amounts of follow-up work.

Genome sequencing and the construction of reference genomes have revolutionized association mapping in soybean. By mapping the resequencing reads to the reference genome or making comparisons between reference genomes, one can theoretically identify all the variations between genomes that can be used as markers for mapping or breeding.

Since the cost of sequencing may still be unaffordable for some laboratories, instead of carrying out costly whole-genome resequencing and *de novo* genome assembly, genotyping-by-sequencing (GBS)/reduced representation sequencing (RRS) can be an economical option for soybean research. There are quite a number of variants of GBS/RRS, such as restriction site-associated DNA sequencing (RAD-seq), double-digestion RAD sequencing (ddRAD-seq), specific-locus-amplified fragment sequencing (SLAF-seq), Diversity Arrays Technology sequencing (DArTseq), low pass/shallow whole-genome sequencing (LP-WGS) and so on. By sequencing only specific parts of the genome, the sequencing depth of each marker can thus be higher with the same number of total sequencing reads as whole-genome resequencing. These methods may sacrifice the marker density in return for a lower cost and higher precision. Still, the actual marker density is normally at least an order of magnitude higher than the traditional SSR markers, thus the mapping resolution can be greatly improved. This is especially suitable for biparental populations with a limited number of recombination events.

3.2 Important QTLs identified by whole-genome sequencing

Before the release of the soybean reference genome, QTL mapping was mainly done with SSR or restriction fragment length polymorphism (RFLP) markers. After the release of the Wm82 reference genome, reference-based mapping or sequencing-mapping blossomed. Single-nucleotide polymorphism (SNP) and insertion/deletion (INDEL) markers became more readily available. Up to date, there are over 200 published reference-based mapping

studies regarding all sort of soybean traits. To keep it simple, we will use those key studies that include gene function validation as examples to illustrate the effectiveness of reference-based mapping.

3.2.1 Growth and development

Low-depth whole-genome resequencing is prone to sequencing error as well as leaving gaps in the coverage at certain genomic regions. To overcome the low coverage of the genome due to low-depth whole-genome resequencing, "bin" markers (Xie et al., 2010) are normally adopted instead of individual SNPs. The genotype of a bin marker is defined by the genotypes of an array of SNPs in a pre-determined window, such that it can tolerate sequencing error, ambiguous base-calling and missing data due to low sequencing coverage. This strategy is better used in a biparental population with limited recombination events. A recombinant inbred (RI) population between a wild and a cultivated parent was genotyped through low-depth whole-genome resequencing at 1X, for the mapping of QTLs governing plant height and growth habit (Wang et al., 2021). SNPs were converted into bin markers before mapping. With more than 6,000 bin markers, this study identified major QTLs related to plant height and growth habit (Wang et al., 2021). Through optical mapping and a comparison between high-quality second-generation soybean reference genomes, a copy number variation (CNV) in *gibberellin 2-oxidase 8* genes was found within these QTLs. Since each unit of the CNV spans ~50 kb, it would not have been able to be resolved using the first-generation soybean reference genomes. The increase in the copy number of *gibberellin 2 oxidase 8* genes is widespread among soybean cultivars, and it reduces internode elongation and suppresses the trailing growth habit (Wang et al., 2021).

The genetic regulation of growth period is a major focus of soybean genomic research. A number of genes and loci regulating growth period have been identified. By genotyping 308 RI lines originating from a very late-flowering and an extremely early-flowering parent using 2b-RAD (Wang et al., 2020), a variant of ddRAD-seq (Wang, Meyer, McKay, & Matz, 2012), a genetic map consisting of 3,454 markers was constructed. In total, 15 QTL regions were identified, with some of them containing reported genes controlling flowering time (Wang et al., 2020). Specifically, in the QTL *qFT12-2*, a circadian clock-related gene, *GmPRR37*, was identified to control the photoperiodic flowering time that helped soybean adapt to the regional conditions during domestication (Wang et al., 2020). The same genes were also discovered in GWAS (Li et al., 2020; Lu et al., 2020).

The cultivation of soybean at low latitudes usually suffers from low yields due to early flowering under inductive short-day conditions. The adoption of soybean varieties with the long-juvenile (LJ) trait, which have a prolonged vegetative phase under short-day conditions, can significantly improve the yield. PI 159925 and BR121 are two soybean varieties showing the LJ trait while Harosoy is a normal variety sensitive to day-length. QTL mapping using two genetic populations originating from PI 159925 x Harosoy and BR121 x Harosoy identified the same QTL region on chromosome 4 governing the LJ phenotype (Lu, Zhao, et al., 2017). An orthologue of the circadian clock gene, *EARLY FLOWERING 3* (*ELF3*), encoding a component of the evening complex, was found to carry the variations that could explain the LJ phenotype (Lu, Zhao, et al., 2017). Complementation of the loss-of-function variants with a functional copy of *ELF3* can significantly reverse the LJ phenotype (Lu, Zhao, et al., 2017). Further analyses also supported that variants of other evening complex genes may also contribute to soybean adaptation to low latitudes (Lu, Zhao, et al., 2017).

3.2.2 Yield potential

A 100-seed weight study was carried out using 198 RI lines generated from a wild (ZYD7) and a cultivated (HN44) soybean (Lu et al., 2017). Interestingly, an elite RI line, R245, was found having seed weights exceeding those of the parental lines, suggesting that the alleles of some loci from the wild parent ZYD7, which has a lower seed weight than the cultivated one, may contribute to the increased seed weight of the RI line. The 198 RI lines were resequenced at 2X depth for SNP-calling and bin-map construction. Fourteen QTLs associated with the 100-seed weight were mapped. It was found that R245 had 13 alleles from the cultivated parent and one from the wild parent in these 14 seed weight QTLs. Of this particular locus from the wild parent that contributed to a large seed size, analyses identified 2 out of 22 genes bearing polymorphisms between the two parental lines. These two genes encode protein phosphatase 2C (PP2C) and EamA-like (EAL), respectively (Lu, Xiong, et al., 2017). Functional analyses in Arabidopsis showed that the ectopic expression of *PP2C* from ZYD7 could increase the plant and seed sizes (Lu, Xiong, et al., 2017). The analysis of 72 wild soybeans and 94 cultivars showed that only 60% of the cultivars carried the ZYD7 *PP2C* allele (Lu, Xiong, et al., 2017), suggesting that the yield of some of the cultivars could be further improved by introducing the ZYD7 *PP2C* alleles.

3.2.3 Abiotic stress tolerance

It has been known that soybean salt tolerance was controlled by a single dominant gene since 1969 (Abel, 1969). The salt tolerance QTL was later mapped to linkage group N (chromosome 3) of the soybean genome using SSR and RFLP markers (Lee et al., 2004). A decade later, the determinant gene was finally uncovered by one of the pioneer sequencing-based QTL mapping studies in soybean (Qi et al., 2014). The salt tolerance QTL was mapped to a 978-kb region in chromosome 3 of the soybean genome using a bin map constructed from 96 RI lines with a 1X sequencing depth (Qi et al., 2014). It was further narrowed down to a 388-kb region using SNP markers developed from the resequencing of 31 soybean germplasms (Lam et al., 2010; Qi et al., 2021). After analyzing a basket of genomic data, it was found that *GmCHX1* was the determinant gene for salt tolerance (Qi et al., 2014). The protein sequence of GmCHX1 is conserved in all salt-tolerant soybean accessions. Any mutation affecting the protein sequence or reducing its expression could be detrimental (Qi et al., 2014). The identity of GmCHX1 (also known as GmSALT3 or Ncl) was later confirmed by another mapping study using 367 F_5 RI lines and 5,769 $F_{5:6}$ individuals using a non-sequencing-based mapping method (Guan et al., 2014). This information has also led to the breeding of a few new multi-stress-tolerant soybean varieties (Li, Wang, et al., 2020).

Apart from *GmCHX1/GmSALT3/Ncl*, another genomic study has also identified a key gene corresponding to soybean salt tolerance during germination (Zhang et al., 2019). By combining the mapping results from an RI population consisting of 184 lines and a GWAS using 211 soybean accessions, a 560-kb region on chromosome 8 was pinpointed to determine the salt tolerance of soybean at germination (Zhang, Liao, et al., 2019). To identify the causal variations within the QTL region, the parental lines were resequenced for variant-calling against the Wm82 reference genome. After identifying 29 genes with variations between the parents, qRT-PCR was used to detect the genes that were distinctively responsive to salt treatment and differentially expressed between parents. A cation diffusion facilitator-encoding gene (*GmCDF1*) was finally isolated as the potential candidate. Functional analyses suggested that GmCDF1 probably mediates Na^+ and K^+ homeostasis by regulating the expression of ion transporter genes such as *GmSOS1* and *GmNHX1* (Zhang, Liao, et al., 2019).

3.2.4 Biotic stress resistance

Sclerotinia stem rot is a disease that can devastate a soybean crop. Using 149 $F_{5:20}$ RI lines for QTL mapping and 261 soybean accessions for GWAS, a

region on chromosome 13 was found to correspond to the sclerotinia stem rot resistance in both populations (Zou et al., 2021). A glutathione S-transferase (GST)-encoding gene (*GmGST*) was differentially expressed between susceptible and resistant lines upon *Sclerotinia sclerotiorum* challenge (Zou et al., 2021). The overexpression of *GmGST* from the resistant lines can improve the resistance toward *S. sclerotiorum* in both resistant and susceptible lines (Zou et al., 2021). At the same time, the CRISPR/cas9-mediated mutation of *GmGST* in the resistant line also reduced its resistance toward the pathogen (Zou et al., 2021).

3.2.5 Seed coat color

The transition from a pigmented to a non-pigmented seed coat was an important trait of domestication of soybean. It has long been known that this transition was due to the tissue-specific RNA silencing of *chalcone synthase* genes in the presence of interfering RNAs generated by the *I* locus on chromosome 8 of the soybean genome (Todd & Vodkin, 1996; Tuteja, Zabala, Varala, Hudson, & Vodkin, 2009). Nevertheless, the causal variation was not uncovered till 2019, when the seed coat color variation of a soybean RI population from a wild and a cultivated parent was also mapped to the *I* locus through QTL mapping (Xie et al., 2019). By comparing the wild soybean reference genome with bacterial artificial chromosome clones corresponding to the *I* locus of cultivated soybean, it was found that there was an inversion within the *I* locus that caused a head-to-head fusion between a *subtilisin* gene and a *chalcone synthase* gene (Xie et al., 2019). The fused gene produced chimeric transcripts that led to the RNA silencing of the *chalcone synthase* gene family, thus resulting in an opaque seed coat. The segmental inversion was also confirmed by optical contigs as well as a high-quality reference genome published later (Valliyodan et al., 2019; Xie et al., 2019). Using PacBio long read sequencing, the variation in the pigmentation of the hilum was revealed to be also controlled by another chimeric fusion in the *I* locus (Kajiya-Kanegae et al., 2021).

3.3 Current findings and limitations of genome-wide association-mapping

The first comprehensive sequencing-based GWAS can be dated back to 2015 (Zhou et al., 2015) where 302 soybean accessions consisting of wild soybeans, landraces, and improved cultivars were sequenced at an average depth of 11X (Zhou et al., 2015). Both known and novel genomic regions

associated with domestication-related phenotypes were identified using GWAS (Zhou et al., 2015), demonstrating that sequencing-based GWAS on soybean is feasible. Since then, a number of GWAS have been carried out on yield-related traits such as yield stability (Quero et al., 2021), seed shape (Zhao, Li, et al., 2019), and seed weight (Zhang et al., 2021; Zhao, Dong, et al., 2019). Other GWAS were also performed on plant architecture (Seck, Torkamaneh, & Belzile, 2020), stress performance (Jing et al., 2021; Ravelombola et al., 2020; Steketee, Schapaugh, Carter, & Li, 2020), nitrogen fixation-related traits (Ray et al., 2015) and other traits (Dhanapal et al., 2016). Although these studies may not provide sufficient evidence to pinpoint the causal genes, they have provided valuable information for marker-assisted selections and further studies. A few examples are discussed below to illustrate how GWAS were applied to soybean.

Apart from the previously mentioned *J* locus (Lu, Zhao, et al., 2017; Yue et al., 2017), other genes/loci were also found to contribute to the later-flowering trait in soybean cultivars developed in these regions. A GWAS using 329 soybean accessions, including 165 from low latitudes, identified a locus named *Tof16* on chromosome 16 that controls the flowering time under short-day conditions (Dong et al., 2021). After fine mapping and sequence comparisons, *Tof16* was found to encode LHY1a, a circadian clock gene (Dong et al., 2021). A loss-of-function mutation of *LHY1a* induced by CRISPR/cas9 caused delayed flowering in soybean under short-day conditions, through the derepression of *E1* and the repression of *GmFT2a* and *GmFT5a* (Dong et al., 2021). Loss-of-function haplotypes of *LHY1a* are widely distributed in tropical soybean accessions, indicating its importance in soybean adaptation to low-latitude regions (Dong et al., 2021). In the same study, the 4 *LHY* homologues in the soybean genome were demonstrated to function redundantly in regulating flowering time and yield under short-day conditions (Dong et al., 2021).

Using RAD-seq, 182 wild soybean accessions were genotyped to obtain 72,574 SNPs for GWAS on salt tolerance (Jin et al., 2021). This is one of the few GWAS using only wild soybean accessions. Among the 209 genes associated with the 11 SNPs significantly linked to salt tolerance, nine were directly identified as salt tolerance-related, including a gene corresponding to *GmCHX1/GmSALT3/Ncl* (Jin et al., 2021). For the uncharacterized gene, *GsERD15B* (*Early Responsive to Dehydration 15B*), although there was no polymorphism identified in the protein sequence between salt-tolerant and salt-sensitive accessions, a 7-bp insertion was spotted in its promoter in most of the salt-sensitive ones (Jin et al., 2021), leading to reduced expression.

Overexpression of *GsERD15B* was shown to enhance the salt tolerance of soybean, probably through the activation of other stress-responsive genes (Jin et al., 2021).

Interactions between genetic loci could sometimes make these loci hard to identify. A GWAS involving 809 soybean accessions initially identified 150 loci significantly associated with 57 traits (Fang et al., 2017). However, some of the well-characterized loci, such as *dt2* which corresponds to semi-determinate growth, were not detected in this initial study due to the epistatic effect of the loci (Fang et al., 2017). To solve this problem, the 809 accessions were divided into subgroups according to their genotypes at the trait-associated loci identified in the first round, for a second round of GWAS. By employing this strategy, 95 additional secondary loci were identified, including a locus on chromosome 19 that interacts with *Ln* on chromosome 20 to control leaf area (Fang et al., 2017).

Another study genotyped 200 soybean germplasms by SLAF sequencing to obtain 28,926 high-quality SNPs for GWAS on isoflavone contents (Wu et al., 2020). Of the significantly associated SNPs, a peak was consistently identified on chromosome 8, which was associated with total isoflavone content and daidzein content across all the tested environments (Wu, Li, et al., 2020). Together with the results from biparental mapping and fine mapping, a mitogen-activated protein kinase (MPK)-encoding gene, *GmMPK1*, was identified. The overexpression of *GmMPK1* in low-isoflavone soybean accessions, either by transient expression in hairy root or by stable transformation, could increase the isoflavone contents in hairy root and seeds, respectively (Wu, Li, et al., 2020). The direct link between *GmMPK1* and isoflavone accumulation remains unknown. However, it is suggested that GmMPK1 participates in a MAPK-mediated signal transduction pathway to regulate isoflavone production upon pathogen challenge (Wu, Li, et al., 2020).

On some occasions, SNPs alone cannot provide sufficient resolution for association mapping. Other sequence variations such as INDELs could then be used to improve the marker density. While soybean is one of the most important oil crop, efforts have also been spent on discovering oil content related genes or loci. A seed protein and oil QTL was mapped to a 4-Mb region on chromosome 15 using a biparental RI population (Zhang et al., 2020). To further delimit this QTL, an association study was done using 631 soybean accessions. After analyzing the association between the protein/oil content and the 79,725 SNPs and 17,088 INDELs within the 4-Mb region, the QTL was finally narrowed down to a 34-kb region (Zhang et al., 2020).

Within this 34-kb region, a sugar transporter-encoding gene (*GmSWEET39*) was suggested to be the causative gene. A CC deletion in *GmSWEET39* has led to a truncation of the C-terminus of the resulting protein, which was associated with the low protein-high oil trait (Zhang et al., 2020). Apparently, high oil content was a more favorable trait in cultivated soybean, and the allele with the CC deletion was selected during soybean domestication, elevated from <3% prevalence in wild soybeans to over 85% in improved cultivars (Zhang et al., 2020).

In a GWAS using 219 soybean accessions with seed oil content ranging from 9.64% to 27.69% in six environmental conditions, 110 SNPs were identified to associated with the oil content (Zhang et al., 2019). Collapsing the 110 SNPs yield 3 QTL regions. Within the QTL on chromosome 20, a domestication related oleosin-encoding gene, *GmOLE1*, was identified based on the *Fst* value. Variations in the promoter of GmOLE1 was highly associated with the expression and oil content (Zhang, Zhang, et al., 2019). Ectopic overexpression of *GmOLE1* in soybean not only improved the seed oil content, it also increased overall seed yield per plant, and germination rate of the transgenic seeds (Zhang, Zhang, et al., 2019). The tradeoff was the reduction of protein content and 100-seed weight (Zhang, Zhang, et al., 2019).

The polyunsaturated fatty acid content determines the stability of soybean oil and also nutritional consequence. The major fatty acid in soybean oil in linoleic acid, which is a polyunsaturated fatty acid. A GWAS using 510 Chinese soybean accessions, with a linoleic acid range from 36.22% to 72.18%, identified 612 significant SNPs distributed on 9 chromosomes (Di et al., 2022). Among these region, a APETALA2/ethylene responsive element-binding protein (AP2/EREBP) transcription factor encoding gene on chromosome 4 was stably identified in 3 different conditions (Di et al., 2022). This gene was named *GmWRI4* after its homologue *WRINKLED1* (*AtWRI1*) from Arabidopsis. The expression of GmWRI4 was significantly negatively correlated with the linoleic acid (Di et al., 2022). Low linoleic acid was resulted in transgenic soybean seed overexpressing GmWRI4 due to the repression of fatty acid desaturase encoding gene *GmFAD2-1A* and *GmFAD2-1B* (Di et al., 2022).

Symbiotic nitrogen fixation is a unique feature of legumes that is agriculturally and environmentally important. A total of 5,294,054 markers, including both SNPs and INDELs, from 496 soybean accessions were used for GWAS on nodule numbers (Zhang et al., 2021). Only one locus spanning ~105 kb on chromosome 2 was found significantly passing the

Bonferroni threshold. Although there are 20 protein-coding gene models within the region, the only TIR-NBS-LRR-encoding gene, namely *GmNNL1*, was assumed to be the most probable candidate as TIR-NBS-LRR proteins are usually involved in plant-microbe interactions. Sequence analyses categorized the haplotypes of *GmNNL1* into two classes, either with or without a SINE-like transposon insertion. Genetics and functional analyses suggested that *GmNNL1* without the SINE-like transposon is the dominant functional allele that inhibits nodulation (Zhang, Wang, et al., 2021). The wide distribution of *GmNNL1* with a SINE insertion in both wild and cultivated soybeans suggests that the reduced symbiotic specificity was evolutionarily beneficial. It might give soybean a better nitrogen fixation ability and therefore might have improved its fitness in both natural and field environments.

4. Impacts of genomic research on soybean molecular breeding

4.1 Marker-assisted selection

Discovering the function-linked molecular markers and even identifying the underlying gene are imperative for introducing and merging elite alleles into germplasms to enhance soybean cultivar performance. Many types of molecular marker systems have been successfully developed and widely applied in various researches. For example, before high-throughput sequencing technologies were widely available, researchers and breeders made use of low-density restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD), sequence characterized amplified region (SCAR), simple sequence repeat (SSR), and cleaved amplified polymorphic sequence (CAPS). However, in the genomic era, with the release of soybean reference genomes and an increasing amount of genome-wide association studies on diverse traits, high-density molecular markers such as SNPs and INDELs are more frequently employed.

With the help of soybean reference genomes and high-throughput sequencing, many molecular markers have been developed with higher-than-ever linkages to the traits of interest, among which some are the functional markers located within the causal genes. This progress has significantly promote the efficiency of marker-assisted selection. Furthermore, marker-assisted selection together with recurrent backcrossing is an excellent strategy to introduce elite alleles into elite germplasms. A few examples of such successes are highlighted here.

Resistance to pod shattering is a domesticated trait that retains the seeds on the plant until harvest. A major QTL, *qPdh1*, corresponding to pod shattering resistance, were first located in the region between two SSR markers, Sat_366 and Sat_093, and then finally identified as a gene encoding a Dirigent-like protein (Funatsuki et al., 2006, 2014). These linked markers have been used to successfully assist the breeding of several pod-shattering-resistant cultivars, such as "Sachiyutaka A1 gou" (Hajika et al., 2016), "Fukuyutaka A1 gou" (Yamada et al., 2017), "Enreinosora" (Yamada et al., 2017), and "Kotoyutaka A1 gou" (MAFF, Japan).

Similarly, several QTLs associated with soybean cyst nematode (SCN) resistance, such as *Rhg4*, *rhg1s*, *rhg1g*, *rhg2g* and *rhg2s*, were discovered and their mapping information has been used in different breeding projects. For example, "Yukihomare R" and "Suzumaru R" are respectively derived from "Yukihomare" and "Suzumaru" through recurrent backcrossing. Apart from acquiring *rhg2g* from "Gendenshirazu," "Suzumaru R" also acquired *rhg1s* and *Rhg4* from PI 84751, resulting in a cultivar that is resistant to SCN races 1 and 3 (Kurosaki et al., 2017; Suzuki et al., 2017). As for soybean mosaic virus (SMV) resistance, Kato et al. (2016) successfully conferred the resistance against SMV strains C and D on a leading variety, Ohsuzu, by transferred the *Rsv3* marker from a U.S. variety, Harosoy (Kato et al., 2016). Oki et al. (2011) introduced two QTLs, CCW-1 and CCW-2, into "Fukuyutaka" in developing the variety "Fukuminori" with antibiosis resistance (Oki et al., 2011). A null allele of kunitz trypsin inhibitor (KTI) from PI542044 was introgressed into JS97-52, DS9712 and DS9814 to get KTI-free lines (Kumar, Rani, Rawal, & Mourya, 2015; Maranna et al., 2016). To speed up the soybean improvement process, a fast-breeding strategy, integrating off-site generation advancement and marker-assisted selection to shorten the generation time, was established to allow at least four generations per year under natural growing conditions (Fang et al., 2021). In these successful cases, marker-assisted selection significantly promoted the efficiency of breeding. However, it appears that this method works better with the qualitative traits with high heritability, but the benefits to quantitative traits such as yield are still limited.

4.2 Genomic selection

Instead of directly selecting the markers associated with major-effect QTLs or functional genes using marker-assisted selection (MAS), genomic selection (GS) estimates the genomic-estimated-breeding-values (GEBVs) by considering the effects of all markers as a whole. It has been demonstrated

that GS using 500 SNP markers could outperform MAS using 22 loci identified via GWAS for seed weight (Zhang, Song, Cregan, & Jiang, 2016). GS requires a training population to supply the genotype and phenotype information to establish a statistical model for the estimation of the marker effects. In turn, the GEBVs of the individuals in the test population can be calculated and cross-validated with the actual phenotype. To obtain more reliable GEBVs, the training cycle can be repeated until satisfactory.

With the emerging SNP and INDEL markers linked to all kinds of traits, GS can be used for both background and foreground selections. GS in soybean is in the early stage and apparently still requires a lot of optimization. For example, for soybean maturity and yield components, the training population size apparently was the most important factor in improving the prediction accuracy, more than the choice of statistical models and marker density. On average, ~9.1% improvement in accuracy can be observed by doubling the population size. However, once the population size exceeded 2,000, the effect on prediction accuracy improvement plateaued (Xavier, Muir, & Rainey, 2016). Similarly, to carry out GS for quantitative resistance toward *Phytophthora sojae*, when the population size was limited, the predictive power was improved with an increase in marker density of up to 1,000 SNPs, and the improvement in prediction accuracy diminished with any further increase in marker number (Rolling, Dorrance, & McHale, 2020). In this case, a simplified BARCSoySNP6K developed from the SoySNP50K assay would be sufficient for GS (Song et al., 2020).

One obstacle for GS is the missing genotypes, which were normally imputed, but they may not reflect the genotype effects. With the emergence of deep-learning technology, a framework based on convolutional neural networks (CNNs) was developed to predict the GEBV of soybean without the imputation of missing genotypes (Liu et al., 2019). The deep-learning model showed an overall improvement over conventional statistic models in predicting the yield, seed moisture and oil contents (Liu et al., 2019). Instead of using the entire set of SNPs, a GS study on SCN tolerance used only the SNPs from GWAS with logarithm of the odds (LOD) scores higher than 2 (Ravelombola et al., 2020). This strategy improved the GS accuracy by two times when compared to another strategy that employed the whole set of SNPs (Ravelombola et al., 2020). Furthermore, the same study also demonstrated that GS accuracy was dependent on the SNP set, training model, and training population size (Ravelombola et al., 2020). Similarly, using just 231 SNPs associated with amino acid contents for GS gave better GEBVs than those generated using 23,279 random SNPs (Qin et al., 2019).

A QTL-allele matrix of seed oil, oleic acid and linolenic acid contents was constructed for the Chinese soybean landrace population (CSLRP) and was used for the GS of optimal crosses with high transgressive potentials (Zhang et al., 2018). Since phenotype is normally determined by the haplotype instead of individual SNPs alone, haplotype-based analyses are expected to be able to better infer the phenotype. This has been confirmed by a GS using preselected markers based on haplotype blocks, which effectively improved the prediction accuracy for grain yield (Ma et al., 2016).

4.3 Random mutagenesis

Random mutagenesis is a traditional way of introducing variations into the soybean genome, usually achieved through physical methods such as irradiating soybean seeds with high-energy radiations or chemical methods such as treating soybean seeds with chemical mutagens to induce changes in their DNA. Other methods could involve the use of random T-DNA or transposon insertions into the genome to alter endogenous gene functions. These treated seeds would then be germinated and screened for the desired traits. These random processes may occasionally create beneficial mutations that improve certain traits of soybean. However, the outcomes of these processes are highly unpredictable, and the causal mutations were largely unknown before the genomic era. Multiple mutations could occur in the same treated seed and result in simultaneous changes in other traits besides the targeted one. Nevertheless, randomly induced mutagenesis could create new alleles that do not exist naturally, and thus, in its own right, is still an important tool in genetic studies and crop improvement especially when the causal gene of a certain trait is still unknown.

A soybean mutant collection was created by fast neutron bombardment of the seeds of soybean cultivar M92-220 (Bolon et al., 2011). Characterization of the genomes of these mutants, by either comparative genomic hybridization or resequencing, identified all sorts of mutations including deletion, tandem duplication, and translocation, etc. (Bolon et al., 2011, 2014). This mutant collection showed a wide range of variations in seed composition such as changes in protein and oil contents, and visible phenotype such as alterations in plant architecture (Bolon et al., 2011). Another soybean mutant collection was produced by ethyl methane sulfonate (EMS)-induced mutagenesis (Espina et al., 2018). The M2 plants also displayed a wide range of phenotypic variations. Such mutants are valuable for genetic studies and crop improvement.

It is relatively straightforward to identify the causal mutation using forward genetic approaches. In a nutshell, a mutant of interest can be crossed with an unrelated soybean variety. The causal mutation can then be identified through a bulked segregant analysis (BSA) of their offsprings. For example, the *chlorophyll-deficient gold yellow trifoliolate leaves* (*gyl*) mutation was mapped by BSA to a region on chromosome 13 containing a chlorophyll synthesis-associated gene (Li et al., 2017). The causal mutation resulting in brown-seededness in a fast-neutron mutant in the yellow-seeded Williams 82 background was identified using comparative genome hybridization (CGH) (Stacey et al., 2016). CGH analyses using five brown-seeded BC_1F_2 plants identified three deletions in common (Stacey et al., 2016). In another study, a fast neutron-induced chromosomal translocation between chromosomes 8 and 13 corresponding to the high-sucrose and low-oil phenotype of soybean seed was identified by BSA, CGH and genome resequencing. Using a genomic approach, this mutation may also be mapped by QTL mapping (Dobbels et al., 2017). This inter-chromosomal translocation may have disrupted a β-ketoacyl-[acyl carrier protein] synthase 1 (KAS1), which is essential for fatty acid biosynthesis, and led to the inefficient synthesis of oil from sucrose in seeds (Dobbels et al., 2017). PE2166 was a mutant of Pungsannamul with elevated α-linolenic acid (Johnson et al., 2021). An RI population was produced by crossing PE2166 with Daepung. A QTL corresponding to elevated α-linolenic acid was mapped to chromosome 5 containing a homeodomain-like transcriptional regulator-encoding gene that bears the causal mutation (Johnson et al., 2021). In cases where the major causal gene of the unmutated phenotype is well characterized, checking the sequence of the causal gene directly may provide the answer to the type and location of the mutation. For example, six mutants were identified to have high seed stearic acid contents compared to the wild type (Lakhssassi et al., 2017). As it is known that the metabolic enzyme stearoyl-acyl carrier protein desaturase (SACPD-C) is responsible for the conversion of stearic acid to oleic acids, all six mutants were verified to carry mutations in the *SACPD-C* gene (Lakhssassi et al., 2017).

One obstacle with generating randomly-induced mutants is that the locations of the mutations are unknown and could occur at multiple locations on the genome. Thus, using a reverse genetic approach to study the function of a given gene using randomly-induced mutants could be confusing. However, with the availability of soybean reference genomes and next-generation sequencing platforms, the high-throughput detection of

mutations in targeted genes becomes more feasible (Rigola et al., 2009). In brief, DNAs of the mutants are mixed by multidimensional pooling. Targeted regions are amplified and indexed from different pools. The amplicons are then sequenced on next-generation sequencing platforms to identify the desired mutations. The loss-of-function mutant of *GmDNJ1* was isolated by such a strategy. A study on the *dnj1* mutant demonstrated that *GmDNJ1* played crucial roles in plant growth and heat stress tolerance (Li et al., 2021). A similar technology called TILLING-by-Sequencing (+) (TbyS[+]), which uses magnetic breads to enrich the target amplicon before sequencing, has successfully identified mutants from three desaturase gene families involved in the soybean fatty acid biosynthesis pathway (Lakhssassi et al., 2021). Furthermore, next-generation sequencing can also detect copy number variations in the mutant lines. Using genotyping-by-sequencing technology, copy number variations of up to 50 Mb were detected in a fast-neutron mutant population (Lemay et al., 2019).

The application of the activation tagging technique could induce the expressions of tagged genes while at the same time providing an anchor with a known sequence for later identification of the tagged genes. A variant of the rice *mPing*-based activation tag called *mmPing20* was adopted in soybean research (Johnson et al., 2021). Expression analyses showed that *mmPing20* could actually activate the expressions of nearby genes in the soybean genome (Johnson et al., 2021). The drawback of this method is that the activation of genes requires exogenous DNA. As a result, the resulting mutant is regarded as a genetically-modified organism, which is in general not suitable for commercialization as a food crop.

4.4 Genome editing in soybean

Compared to traditional transgenesis, new crop varieties generated by genome editing are in general more acceptable to the public. The main reason is that although the process still requires the introduction of exogenous DNA into the plant, the foreign DNA can be completely removed afterward, leaving only the edited site in the genome. Genomic research has played a fundamental role in genome editing, especially for those highly duplicated genomes such as that of soybean. Targets for editing are usually determined by homology search, genetic studies, mapping, or association studies. Building good reference genomes with precise annotations followed by in-depth studies is the key for identifying the right targets for editing.

4.4.1 TALEN

Transcription activator-like effector nucleases (TALENs) are sequence-specific DNA endonucleases engineered from TAL effectors, where the transcription activation domain is replaced by an endonuclease domain, specifically the *Fok*I domain, and the DNA-binding domain is a customizable stretch of 34-amino-acid repeats. These repeats vary only at amino acid positions 12 and 13, so that each repeat binds one specific nucleotide on the DNA (Sprink, Metje, & Hartung, 2015). Thus, by designing a unique combination of repeats, TALEN can act like a pair of scissors to make a double-stranded break (DSB) in a specific location on the DNA determined by the corresponding combination of repeats. The DSB is usually repaired via two main pathways including the error-prone nonhomologous end-joining (NHEJ) and the homology-directed repair (HDR) to introduce insertions, deletions, substitutions, or even larger chromosomal rearrangements. So far, TALEN-based genome editing has successfully been applied in diverse plants and crops, such as Arabidopsis, tomato, rice, barley and wheat (Sprink et al., 2015). As for soybean, there were also several successful cases. It has been demonstrated that engineered TALENs can disrupt both copies of *fatty acid desaturase 2* genes (*FAD2-1A* and *FAD2-1B*) to decrease the levels of polyunsaturated fats to increase the stability of soybean oil against oxidation and make it healthier for human consumption (Haun et al., 2014). Another mutation of *FAD3A* was stacked into the *fad2-1a fad2-1b* mutant to further decrease the linoleic acid level from 5.1% in the double mutant to 2.7% in the triple mutant and increase the oleic acid from 77.5% to 82.2% (Demorest et al., 2016). The first gene-edited crop commercialized in the U.S. was also a high-oleic low-linolenic acid soybean engineered using TALEN (Li, Wang, et al., 2020).

4.4.2 ZFN

Zinc-finger nuclease (ZFN) is another synthetic engineered nuclease with a *Fok*I nonspecific endonuclease domain fused to a customized array of engineered zinc fingers, where each finger typically recognizes three nucleotides (Urnov, Rebar, Holmes, Zhang, & Gregory, 2010). Thus, with a well-designed array of zinc fingers, a ZFN can also help edit the genome at a specific location. In soybean, Curtin et al. (2011) first demonstrated that custom-designed ZFNs could carry out transmissible mutations of soybean genes, specifically *DCL4a* and *DCL4b* in their work (Curtin et al., 2011). Not only can they create INDELs, ZFNs have also been demonstrated to integrate a large DNA fragment containing four genes into the soybean

genome through NHEJ. However, the rate of successful incorporation in the regenerated plant was only 0.03%. Therefore further improvement in this technique is required (Bonawitz et al., 2019).

4.4.3 CRISPR/Cas9

The CRISPR (clustered regularly interspaced short palindromic repeat)/Cas9 (CRISPR-associated endonuclease 9) system is an emerging DSB technology with great promise. Different from the previous TALEN and ZFN platforms which rely on protein-nucleotide interactions to recognize a specific DNA sequence, the CRISPR/Cas9 system generally uses a simple short guide RNA (gRNA) to recognize a specific DNA sequence followed by a protospacer adjacent motif (PAM) to generate a DSB in the targeted sequence. More specifically, for the *Streptococcus pyogenes* CRISPR/Cas9 system, the PAM consensus sequence is 5'-NGG-3', and the cleavage normally occurs between the third and fourth nucleotides upstream of the PAM. The cleaved DNA will then be repaired by NHEJ or HDR in the presence of a sequence with homology from either an exogenous or an endogenous source.

The history of CRISPR/cas9 genome editing in soybean is short. Due to the low efficiency of stable transformations in whole soybean plants, transient expression methods, such as soybean hairy root transformation, were first utilized to demonstrate that CRISPR/Cas9 system can effectively edit both endogenous and exogenous genes, including the *bar* gene, *GFP*, *GmFEI2*, *GmSHR*, *Glyma06g14180* (a homolog of *TTG1*), *Glyma08g02290* (a homolog of *KUP4*), and *Glyma12g37050* (a homolog of *ETR1*) (Cai et al., 2015; Jacobs, LaFayette, Schmitz, & Parrott, 2015; Sun et al., 2015). Furthermore, HDR gene integration and replacement were also demonstrated to be feasible through cotransforming the target site-specific Cas9-sgRNA and donor DNA constructs by particle bombardment (Li et al., 2015). Since then, CRISPR/Cas9 has been increasingly utilized in various aspects of soybean research and breeding.

The CRISPR/Cas9 system has significantly facilitated the precise identification of the causal genes for important traits, especially when combining its use in the construction of a knock-out mutant by targeted mutagenesis with a conventional overexpression-complementation experiment. For example, *Rj4*, a gene that restricts certain *Bradyrhizobium* strains, was originally reported to be encoded by *Glyma.01g165800*. However, it was later found to be encoded by *Glyma.0165800-D*, the neighboring gene

duplicate of *Glyma.01g165800*. Knocking out *Glyma.01G165800-D* by CRISPR/cas9 abolished the function of *Rj4*, i.e. allowing the nodulation by *B. elkanii*, but knocking out *Glyma.01G165800* did not (Tang, Yang, Liu, & Zhu, 2015). This provided concrete evidence to support that *Glyma.01G165800-D* is in fact *Rj4*. A fast neutron-induced six-gene deletion was found to be associated with the short-trichome phenotype (Campbell et al., 2019). Among the six genes, *Glyma.06g145800* encoding a homologue of *Arabidopsis* Constitutive Expression of PR Genes 5 (CPR5) was the strongest candidate (Campbell et al., 2019). A panel of mutants with different numbers of base deletions in *CPR5* were generated by CRISPR/cas9 genome editing. All these mutants showed a significant reduction in the trichome length, with was highly consistent with the fast-neutron mutant phenotype (Campbell et al., 2019). Similarly, the causal gene *GmMS1* (*male-sterility 1*), located in a 39-kb region where five genes were deleted, was demonstrated by CRISPR/cas9 to be *Glyma13G114200* (Jiang et al., 2021). *GmKIX8-1* encodes a KIX domain-containing protein. A CRISPR/cas9 mutant of *GmKIX8-1* showed increased cell proliferation and enlarged organ size. It was shown to be the causative gene for the major seed weight QTL qSw17-1 (Nguyen, Paddock, Zhang, & Stacey, 2021). As mentioned previously, although there are limited mutant collections for soybean, the CRISPR/Cas9 system has greatly facilitated the forward genetics studies on soybean.

In addition to basic research, CRISPR/Cas9 also enables the direct change in elite traits to promote soybean breeding and immediate applications. It is generally agreed that flowering time and maturity are the most important traits for soybean breeding. Research is focused on producing soybeans with times to maturity that are suited to different latitudes. For example, the flowering integrator *GmFT2a* and *GmFT5a* were knocked out to delay flowering time (Cai et al., 2018, 2020). The resulting double mutant *ft2aft5a*, which displays a late-flowering phenotype, is suitable for low-latitude high-temperature environments. On the other hand, to promote the adaptation to high-latitude environments where early flowering is required, CRISPR/Cas9-edited *E1* shortened the time to flowering as desired (Han et al., 2019). While there is a basket of maturity-related genes found in the soybean genome, strategic editing of them could provide more refined controls on soybean ecological adaptations. Furthermore, seven cryptochrome genes (*GmCRYs*) in soybean were edited. Among them, GmCRY1s was found to regulate soybean shade avoidance in

response to reduced blue light, through modulating gibberellin metabolism (Lyu et al., 2021), thus providing a new target for improving soybean performance under high-density cultivation in the field.

CRISPR/cas9 has also been applied to improve qualitative traits. The Δ12-fatty acid desaturase II (FAD2) family of genes were targeted by CRISPR/Cas9 to improve the oil profile, similar to the process using TALEN technology (Do et al., 2019; Wu et al., 2020). Gly m Bd 28K and Gly m Bd 30K are two major allergens present in soybean. Their nullification by CRISPR/Cas9 has been demonstrated to be a feasible way of producing hypoallergenic soybeans (Sugano et al., 2020).

To potentially improve disease resistance in soybean by introducing variations to existing genes, two regions of tandem-duplicated arrays of NBS-LRR (nucleotide-binding-site leucine-rich-repeat) gene families, Rpp1 and Rps1, were targeted by CRISPR/Cas9 to induce rearrangements and produce novel chimeric paralogues that could lead to new disease resistance specificities (Nagy et al., 2021). CRISPR/cas9 can not only target protein-coding genes in the soybean genome, but it can also be used to edit other regulatory elements. For example, it has been demonstrated that the knockout of microRNA miR169c by CRISPR/cas9 could promote nodulation through the repression of soybean *Nuclear Factor-Y Subunit A* (*GmNFYA*) (Xu et al., 2021).

4.4.4 Potential applications of base-editing using CRISPR/Cas9

The commonly used CRISPR/Cas9 system usually disrupts the gene function by introducing a frame-shift variation. However, the complete knockout of a gene may lead to undesirable pleiotropic effects. In fact, variations at the base level could lead to enhanced or diminished protein functions through the alteration of three-dimensional structures, enzymatic activities or interactions with other molecules. Recent genomic researches have identified rich sequence variations in the soybean genome (Kajiya-Kanegae et al., 2021; Lam et al., 2010; Valliyodan et al., 2021; Yang et al., 2021; Zhou et al., 2015). Instead of incorporating a random combination of these variations through traditional breeding, the precise introduction of a desired sequence variation by base editing would be a much better option. In soybean, a CRISPR/Cas9-mediated base-editing tool was developed through combining the $Cas9^{D10A}$ nickase, the rat cytosine deaminase (APOBEC1) or the uracil glycosylase inhibitor (UGI). This system successfully introduced

an edited base at the target position in *GmFT2a* and *GmFT4a* (Cai et al., 2020). The mutation from C to G in *GmFT2a* has led to a significant delay in flowering time when compared to the wild type, but a slightly earlier flowering time when compared to the *ft2a* frameshift-knockout line (Cai, Chen, et al., 2020). This example has sufficiently demonstrated the potential of base editing in soybean. However, the current base-editing tool is still limited and need to be further improved.

5. Perspectives

Although soybean is the major dietary source of plant-based protein, its yield is still much lower than those of other major crops including rice, wheat, and maize (FAOstat [retrieved on Oct 5, 2021, from http://www.fao.org/faostat/en/#data/QCL]). Furthermore, due to population explosion and climate change, food security has become the major challenge in the coming century. Thus, increasing yield, improving yield stability, and enhancing added values are the major goals for soybean breeding. To achieve these goals, genome sequencing has fueled soybean research in the past decade, accumulating vast amounts of information to serve as the basis for soybean breeding.

So far, thousands of soybean accessions have been sequenced in various projects. In most cases, sequencing data are required by publishers of scientific journals to be deposited in public domains, but genetic resources are not. Thus, even though genomic information is available for sophisticated bioinformatic analyses, one cannot test a hypothesis or collect new phenotypic data for analyses without access to the seeds. To get the most out of the published data in order to advance soybean research and breeding, the sharing of soybean germplasm resources between research groups and across national boarders would benefit the whole soybean research community. Taking it a step further, raw phenotypic data should also be systematically shared. In contrast to the volume of available genomic data, phenotypic data are the biggest limiting factor for association studies. The phenotyping of hundreds of soybean accessions is already tedious enough, let alone phenotyping thousands of accessions at once. At the same time, datasets collected across different studies are normally considered not directly comparable among one another. To solve this problem, high-throughput phenotyping systems, either ground-based or unmanned aerial system (UAS)-based, have drawn the attention of researchers (Baek et al., 2020;

Bai, Ge, Hussain, Baenziger, & Graef, 2016; Kaler et al., 2020; Trevisan, Perez, Schmitz, Diers, & Martin, 2020). On the other hand, although directly combining the data from multiple studies for genome-wide association studies (GWAS) is improper, meta-analyses of the results of multiple GWAS have been demonstrated to increase the statistical power and lead to the discovery of new genetic loci (Evangelou & Ioannidis, 2013; Shook et al., 2021).

Simply breeding high-yielding soybean cultivars is apparently not sufficient to fulfill the demand for food in the second half of the 21st century. Serving as a carbon sink and a nitrogen source in a climate-smart agricultural system, new soybean varieties would also need to be highly adaptive to environmental changes and require low agricultural inputs. Although these goals are difficult to attain, new technologies such as machine learning techniques/artificial intelligence have emerged to contribute to both soybean genomic analyses and breeding (Herrero-Huerta, Rodriguez-Gonzalvez, & Rainey, 2020; Yoosefzadeh-Najafabadi, Earl, Tulpan, Sulik, & Eskandari, 2021), giving rise to new possibilities.

Acknowledgments

This work was supported by Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16), Lo Kwee-Seong Biomedical Research Fund and Guangdong Provincial Department of Science and Technology 2020 Key Areas Research and Development Programs: Breeding for High Yield and High Quality New Soybean Cultivars for Tropics and Subtropics (2020B020220008), the China Agriculture Research System (CARS-04) and the CAAS Agricultural Science and Technology Innovation Project. JY Chu copy-edited the text. Any opinions, findings, conclusions or recommendations expressed in this publication do not reflect the views of the Government of Hong Kong Special Administrative Region or the Innovation and Technology Commission.

References

Abel, G. H. (1969). Inheritance of the capacity for chloride inclusion and chloride exclusion by soybeans. *Crop Science*, *9*, 697–698. https://doi.org/10.2135/cropsci1969.0011183X000900060006x.

Arumuganathan, K., & Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter*, *9*, 208–218. https://doi.org/10.1007/BF02672069.

Baek, J., Lee, E., Kim, N., Kim, S. L., Choi, I., Ji, H., et al. (2020). High throughput phenotyping for various traits on soybean seeds using image analysis. *Sensors*, *20*, 248. https://doi.org/10.3390/s20010248.

Bai, G., Ge, Y. F., Hussain, W., Baenziger, P. S., & Graef, G. (2016). A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding. *Computers and Electronics in Agriculture*, *128*, 181–192. https://doi.org/10.1016/j.compag.2016.08.021.

Bayer, P. E., Valliyodan, B., Hu, H. F., Marsh, J. I., Yuan, Y. X., Vuong, T. D., et al. (2021). Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome*, e20109. https://doi.org/10.1002/tpg2.20109.

Bolon, Y. T., Haun, W. J., Xu, W. W., Grant, D., Stacey, M. G., Nelson, R. T., et al. (2011). Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. *Plant Physiology*, *156*, 240–253. https://doi.org/10.1104/pp.110.170811.

Bolon, Y. T., Stec, A. O., Michno, J. M., Roessler, J., Bhaskar, P. B., Ries, L., et al. (2014). Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. *Genetics*, *198*, 967–981. https://doi.org/10.1534/genetics.114.170340.

Bonawitz, N. D., Ainley, W. M., Itaya, A., Chennareddy, S. R., Cicak, T., Effinger, K., et al. (2019). Zinc finger nuclease-mediated targeting of multiple transgenes to an endogenous soybean genomic locus via non-homologous end joining. *Plant Biotechnology Journal*, *17*, 750–761. https://doi.org/10.1111/pbi.13012.

Brown, A. V., Conners, S. I., Huang, W., Wilkey, A. P., Grant, D., Weeks, N. T., et al. (2021). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research*, *49*, D1496–D1501. https://doi.org/10.1093/nar/gkaa1107.

Cai, Y., Chen, L., Liu, X., Guo, C., Sun, S., Wu, C., et al. (2018). CRISPR/Cas9-mediated targeted mutagenesis of *GmFT2a* delays flowering time in soya bean. *Plant Biotechnology Journal*, *16*, 176–185. https://doi.org/10.1111/pbi.12758.

Cai, Y. P., Chen, L., Liu, X. J., Sun, S., Wu, C. X., Jiang, B. J., et al. (2015). CRISPR/Cas9-mediated genome editing in soybean hairy roots. *PLoS One*, *10*, e0136064. https://doi.org/10.1371/journal.pone.0136064.

Cai, Y. P., Chen, L., Zhang, Y., Yuan, S., Su, Q., Sun, S., et al. (2020). Target base editing in soybean using a modified CRISPR/Cas9 system. *Plant Biotechnology Journal*, *18*, 1996–1998. https://doi.org/10.1111/pbi.13386.

Cai, Y. P., Wang, L. W., Chen, L., Wu, T. T., Liu, L. P., Sun, S., et al. (2020). Mutagenesis of *GmFT2a* and *GmFT5a* mediated by CRISPR/Cas9 contributes for expanding the regional adaptability of soybean. *Plant Biotechnology Journal*, *18*, 298–309. https://doi.org/10.1111/pbi.13199.

Campbell, B. W., Hoyle, J. W., Bucciarelli, B., Stec, A. O., Samac, D. A., Parrott, W. A., et al. (2019). Functional analysis and development of a CRISPR/Cas9 allelic series for a *CPR5* ortholog necessary for proper growth of soybean trichomes. *Scientific Reports*, *9*, 14757. https://doi.org/10.1038/s41598-019-51240-7.

Chu, J. S., Peng, B., Tang, K., Yi, X., Zhou, H., Wang, H., et al. (2021). Eight soybean reference genome resources from varying latitudes and agronomic traits. *Scientific Data*, *8*, 164. https://doi.org/10.1038/s41597-021-00947-2.

Chung, W. H., Jeong, N., Kim, J., Lee, W. K., Lee, Y. G., Lee, S. H., et al. (2014). Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Research*, *21*, 153–167. https://doi.org/10.1093/dnares/dst047.

Curtin, S. J., Zhang, F., Sander, J. D., Haun, W. J., Starker, C., Baltes, N. J., et al. (2011). Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiology*, *156*, 466–473. https://doi.org/10.1104/pp.111.172981.

Demorest, Z. L., Coffman, A., Baltes, N. J., Stoddard, T. J., Clasen, B. M., Luo, S., et al. (2016). Direct stacking of sequence-specific nuclease-induced mutations to produce high oleic and low linolenic soybean oil. *BMC Plant Biology*, *16*, 225. https://doi.org/10.1186/s12870-016-0906-1.

Dhanapal, A. P., Ray, J. D., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2016). Genome-wide association mapping of soybean chlorophyll traits based on canopy spectral reflectance and leaf extracts. *Bmc Plant Biology*, *16*, 174. https://doi.org/10.1186/s12870-016-0861-x.

Di, Q., Piersanti, A., Zhang, Q., Miceli, C., Li, H., & Liu, X. Y. (2022). Genome-wide association study identifies candidate genes related to the linoleic acid content in soybean seeds. *International Journal of Molecular Sciences*, *23*, 454. https://doi.org/10.3390/ijms23010454.

Do, P. T., Nguyen, C. X., Bui, H. T., Tran, L. T. N., Stacey, G., Gillman, J. D., et al. (2019). Demonstration of highly efficient dual gRNA CRISPR/Cas9 editing of the homeologous *GmFAD2–1A* and *GmFAD2–1B* genes to yield a high oleic, low linoleic and α-linolenic acid phenotype in soybean. *BMC Plant Biology*, *19*, 311. https://doi.org/10.1186/s12870-019-1906-8.

Dobbels, A. A., Michno, J. M., Campbell, B. W., Virdi, K. S., Stec, A. O., Muehlbauer, G. J., et al. (2017). An induced chromosomal translocation in soybean disrupts a KASI ortholog and is associated with a high-sucrose and low-oil seed phenotype. *G3-Genes Genomes Genetics*, 7, 1215–1223. https://doi.org/10.1534/g3.116.038596.

DOE-JGI. (2021). Glycine max Fiskeby v1.1. Retrieved from https://phytozome-next.jgi.doe.gov/info/GmaxFiskeby_v1_1.

Dong, L., Fang, C., Cheng, Q., Su, T., Kou, K., Kong, L., et al. (2021). Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. *Nature Communications*, *12*, 5445. https://doi.org/10.1038/s41467-021-25800-3.

dos Santos, J. V. M., Valliyodan, B., Joshi, T., Khan, S. M., Liu, Y., Wang, J. X., et al. (2016). Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. *BMC Genomics*, *17*, 110. https://doi.org/10.1186/s12864-016-2431-x.

Espina, M. J., Ahmed, C. M. S., Bernardini, A., Adeleke, E., Yadegari, Z., Arelli, P., et al. (2018). Development and phenotypic screening of an ethyl methane sulfonate mutant population in soybean. *Frontiers in Plant Science*, *9*, 394. https://doi.org/10.3389/fpls.2018.00394.

Evangelou, E., & Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, *14*, 379–389. https://doi.org/10.1038/nrg3472.

Fang, C., Ma, Y. M., Wu, S. W., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology*, *18*, 161. https://doi.org/10.1186/s13059-017-1289-9.

Fang, Y. D., Wang, L. W., Sapey, E., Fu, S., Wu, T. T., Zeng, H. Y., et al. (2021). Speed-breeding system in soybean: Integrating off-site generation advancement, fresh seeding, and marker-assisted selection. *Frontiers in Plant Science*, *12*, 717077. https://doi.org/10.3389/fpls.2021.717077.

Findley, S. D., Cannon, S., Varala, K., Du, J. C., Ma, J. X., Hudson, M. E., et al. (2010). A fluorescence *in situ* hybridization system for karyotyping soybean. *Genetics*, *185*, 727–744. https://doi.org/10.1534/genetics.109.113753.

Funatsuki, H., Ishimoto, M., Tsuji, H., Kawaguchi, K., Hajika, M., & Fujino, K. (2006). Simple sequence repeat markers linked to a major QTL controlling pod shattering in soybean. *Plant Breeding*, *125*, 195–197. https://doi.org/10.1111/j.1439-0523.2006.01199.x.

Funatsuki, H., Suzuki, M., Hirose, A., Inaba, H., Yamada, T., Hajika, M., et al. (2014). Molecular basis of a shattering resistance boosting global dissemination of soybean. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 17797–17802. https://doi.org/10.1073/pnas.1417282111.

Guan, R. X., Qu, Y., Guo, Y., Yu, L. L., Liu, Y., Jiang, J. H., et al. (2014). Salinity tolerance in soybean is modulated by natural variation in *GmSALT3*. *The Plant Journal*, *80*, 937–950. https://doi.org/10.1111/tpj.12695.

Hajika, M., Funatsuki, H., Yamada, T., Takahashi, K., Hishinuma, A., Hirata, K., et al. (2016). Development of a new pod dehiscence-resistant soybean cultivar 'Sachiyutaka A1 gou'. *Bulletin of NARO Institute of Crop Science*, *16*, 1–34.

Han, J. N., Guo, B. F., Guo, Y., Zhang, B., Wang, X. B., & Qiu, L. J. (2019). Creation of early flowering germplasm of soybean by CRISPR/Cas9 technology. *Frontiers in Plant Science*, *10*, 1446. https://doi.org/10.3389/fpls.2019.01446.

Haun, W., Coffman, A., Clasen, B. M., Demorest, Z. L., Lowy, A., Ray, E., et al. (2014). Improved soybean oil quality by targeted mutagenesis of the *fatty acid desaturase 2* gene family. *Plant Biotechnology Journal*, *12*, 934–940. https://doi.org/10.1111/pbi.12201.

Herrero-Huerta, M., Rodriguez-Gonzalvez, P., & Rainey, K. M. (2020). Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. *Plant Methods*, *16*, 78. https://doi.org/10.1186/s13007-020-00620-6.

Huang, M. K., Zhang, L., Zhou, L. M., Yung, W. S., Li, M. W., & Lam, H. M. (2021). Genomic features of open chromatin regions (OCRs) in wild soybean and their effects on gene expressions. *Genes*, *12*, 640. https://doi.org/10.3390/genes12050640.

Jacobs, T. B., LaFayette, P. R., Schmitz, R. J., & Parrott, W. A. (2015). Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnology*, *15*, 16. https://doi.org/10.1186/s12896-015-0131-2.

Jiang, B. J., Chen, L., Yang, C. Y., Wu, T. T., Yuan, S., Wu, C. X., et al. (2021). The cloning and CRISPR/Cas9-mediated mutagenesis of a male sterility gene *MS1* of soybean. *Plant Biotechnology Journal*, *19*, 1098–1100. https://doi.org/10.1111/pbi.13601.

Jin, T., Sun, Y. Y., Shan, Z., He, J. B., Wang, N., Gai, J. Y., et al. (2021). Natural variation in the promoter of *GsERD15B* affects salt tolerance in soybean. *Plant Biotechnology Journal*, *19*, 1155–1169. https://doi.org/10.1111/pbi.13536.

Jing, Y., Teng, W. L., Qiu, L. J., Zheng, H. K., Li, W. B., Han, Y. P., et al. (2021). Genetic dissection of soybean partial resistance to sclerotinia stem rot through genome wide association study and high throughout single nucleotide polymorphisms. *Genomics*, *113*, 1262–1271. https://doi.org/10.1016/j.ygeno.2020.10.042.

Johnson, A., Mcassey, E., Diaz, S., Reagin, J., Redd, P. S., Parrilla, D. R., et al. (2021). Development of mPing-based activation tags for crop insertional mutagenesis. *Plant Direct*, *5*, e00300. https://doi.org/10.1002/pld3.300.

Kajiya-Kanegae, H., Nagasaki, H., Kaga, A., Hirano, K., Ogiso-Tanaka, E., Matsuoka, M., et al. (2021). Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. *DNA Research*, *28*, dsaa032. https://doi.org/10.1093/dnares/dsaa032.

Kaler, A. S., Abdel-Haleem, H., Fritschi, F. B., Gillman, J. D., Ray, J. D., Smith, J. R., et al. (2020). Genome-wide association mapping of dark green color index using a diverse panel of soybean accessions. *Scientific Reports*, *10*, 5166. https://doi.org/10.1038/s41598-020-62034-7.

Kato, S., Takada, Y., Shimamura, S., Hirata, K., Sayama, T., Taguchi-Shiobara, F., et al. (2016). Transfer of the *Rsv3* locus from 'Harosoy' for resistance to soybean mosaic virus strains C and D in Japan. *Breeding Science*, *66*, 319–327. https://doi.org/10.1270/jsbbs.66.319.

Kim, M. Y., Lee, S., Van, K., Kim, T. H., Jeong, S. C., Choi, I. Y., et al. (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*, 22032–22037. https://doi.org/10.1073/pnas.1009526107.

Kumar, V., Rani, A., Rawal, R., & Mourya, V. (2015). Marker assisted accelerated introgression of null allele of kunitz trypsin inhibitor in soybean. *Breeding Science*, *65*, 447–452. https://doi.org/10.1270/jsbbs.65.447.

Kurosaki, H., Fujita, S., Ohnishi, S., Kosaka, F., Tanaka, Y., Takeuchi, T., et al. (2017). A new soybean variety 'Suzumaru R'. *Bulletin of Hokkaido Research Organization Agricultural Experiment Stations*, *101*, 1–13.

Lakhssassi, N., Colantonio, V., Flowers, N. D., Zhou, Z., Henry, J., Liu, S. M., et al. (2017). Stearoyl-acyl carrier protein desaturase mutations uncover an impact of stearic acid in leaf and nodule structure. *Plant Physiology*, *174*, 1531–1543. https://doi.org/10.1104/pp.16.01929.

Lakhssassi, N., Zhou, Z., Cullen, M. A., Badad, O., El Baze, A., Chetto, O., et al. (2021). TILLING-by-sequencing+ to decipher oil biosynthesis pathway in soybeans: A new and effective platform for high-throughput gene functional analysis. *International Journal of Molecular Sciences*, *22*, 4219.

Lam, H. M., Xu, X., Liu, X., Chen, W. B., Yang, G. H., Wong, F. L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, *42*, 1053–U1041. https://doi.org/10.1038/ng.715.

Lee, G. J., Boerma, H. R., Villagarcia, M. R., Zhou, X., Carter, T. E., Li, Z., et al. (2004). A major QTL conditioning salt tolerance in S-100 soybean and descendent cultivars. *Theoretical and Applied Genetics*, *109*, 1610–1619. https://doi.org/10.1007/s00122-004-1783-9.

Lee, G. A., Crawford, G. W., Liu, L., Sasaki, Y., & Chen, X. X. (2011). Archaeological doybean (*Glycine max*) in East Asia: Does size matter? *PLoS One*, *6*, e26720. https://doi.org/10.1371/journal.pone.0026720.

Lei, L., Goltsman, E., Goodstein, D., AlbertWu, G., Rokhsar, D. S., & Vogel, J. P. (2021). Plant pan-genomics comes of age. *Annual Review of Plant Biology*, *72*, 411–435. https://doi.org/10.1146/annurev-arplant-080720-105454.

Lemay, M. A., Torkamaneh, D., Rigaill, G., Boyle, B., Stec, A. O., Stupar, R. M., et al. (2019). Screening populations for copy number variation using genotyping-by-sequencing: A proof of concept using soybean fast neutron mutants. *BMC Genomics*, *20*, 634. https://doi.org/10.1186/s12864-019-5998-1.

Li, Z. F., Jiang, L. X., Ma, Y. S., Wei, Z. Y., Hong, H. L., Liu, Z. X., et al. (2017). Development and utilization of a new chemically-induced soybean library with a high mutation density. *Journal of Integrative Plant Biology*, *59*, 60–74. https://doi.org/10.1111/jipb.12505.

Li, C., Li, Y. H., Li, Y., Lu, H., Hong, H., Tian, Y., et al. (2020). A domestication-associated gene *GmPRR3b* regulates the circadian clock and flowering time in soybean. *Molecular Plant*, *13*, 745–759. https://doi.org/10.1016/j.molp.2020.01.014.

Li, Z. S., Liu, Z. B., Xing, A. Q., Moon, B. P., Koellhoffer, J. P., Huang, L. X., et al. (2015). Cas9-Guide RNA directed genome editing in soybean. *Plant Physiology*, *169*, 960–970. https://doi.org/10.1104/pp.15.00783.

Li, M. W., Wang, Z. L., Jiang, B. J., Kaga, A., Wong, F. L., Zhang, G. H., et al. (2020). Impacts of genomic research on soybean improvement in East Asia. *Theoretical and Applied Genetics*, *133*, 1655–1678. https://doi.org/10.1007/s00122-019-03462-6.

Li, K. P., Wong, C. H., Cheng, C. C., Cheng, S. S., Li, M. W., Mansveld, S., et al. (2021). GmDNJ1, a type-I heat shock protein 40 (HSP40), is responsible for both Growth and heat tolerance in soybean. *Plant Direct*, *5*, e00298. https://doi.org/10.1002/pld3.298.

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

Lin, X., Lin, W. G., Ku, Y. S., Wong, F. L., Li, M. W., Lam, H. M., et al. (2020). Analysis of soybean long non-coding RNAs reveals a subset of small peptide-coding transcripts. *Plant Physiology*, *182*, 1359–1374. https://doi.org/10.1104/pp.19.01324.

Liu, Y. C., Du, H. L., Li, P. C., Shen, Y. T., Peng, H., Liu, S. L., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, *182*, 162–176.e113. https://doi.org/10.1016/j.cell.2020.05.023.

Liu, Y., Wang, D. L., He, F., Wang, J. X., Joshi, T., & Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in Genetics*, *10*, 1091. https://doi.org/10.3389/fgene.2019.01091.

Lu, S., Dong, L., Fang, C., Liu, S., Kong, L., Cheng, Q., et al. (2020). Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. *Nature Genetics*, *52*, 428–436. https://doi.org/10.1038/s41588-020-0604-7.

Lu, X., Xiong, Q., Cheng, T., Li, Q. T., Liu, X. L., Bi, Y. D., et al. (2017). A *PP2C-1* allele underlying a quantitative trait locus enhances soybean 100-seed weight. *Molecular Plant*, *10*, 670–684. https://doi.org/10.1016/j.molp.2017.03.006.

Lu, S. J., Zhao, X. H., Hu, Y. L., Liu, S. L., Nan, H. Y., Li, X. M., et al. (2017). Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. *Nature Genetics*, *49*, 773–779. https://doi.org/10.1038/ng.3819.

Lyu, X. G., Cheng, Q. C., Qin, C., Li, Y. H., Xu, X. Y., Ji, R. H., et al. (2021). GmCRY1s modulate gibberellin metabolism to regulate soybean shade avoidance in response to reduced blue light. *Molecular Plant*, *14*, 298–314. https://doi.org/10.1016/j.molp.2020.11.016.

Ma, Y. S., Reif, J. C., Jiang, Y., Wen, Z. X., Wang, D. C., Liu, Z. X., et al. (2016). Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). *Molecular Breeding*, *36*, 113. https://doi.org/10.1007/s11032-016-0504-9.

Maranna, S., Verma, K., Talukdar, A., Lal, S. K., Kumar, A., & Mukherjee, K. (2016). Introgression of null allele of Kunitz trypsin inhibitor through marker-assisted backcross breeding in soybean (*Glycine max* L. Merr.). *BMC Genetics*, *17*, 106. https://doi.org/10.1186/s12863-016-0413-2.

Nagy, E. D., Stevens, J. L., Yu, N., Hubmeier, C. S., LaFaver, N., Gillespie, M., et al. (2021). Novel disease resistance gene paralogs created by CRISPR/Cas9 in soy. *Plant Cell Reports*, *40*, 1047–1058. https://doi.org/10.1007/s00299-021-02678-5.

Nguyen, C. X., Paddock, K. J., Zhang, Z., & Stacey, M. G. (2021). *GmKIX8-1* regulates organ size in soybean and is the causative gene for the major seed weight QTL *qSw17-1*. *New Phytologist*, *229*, 920–934. https://doi.org/10.1111/nph.16928.

Oki, N., Komatsu, K., Sayama, T., Ishimoto, M., Takahashi, M., & Takahashi, M. (2011). Genetic analysis of antixenosis resistance to the common cutworm (*Spodoptera litura* Fabricius) and its relationship with pubescence characteristics in soybean (*Glycine max* (L.) Merr.). *Breeding Science*, *61*, 608–617. https://doi.org/10.1270/jsbbs.61.608.

Qi, X., Jiang, B., Wu, T., Sun, S., Wang, C., Song, W., et al. (2021). Genomic dissection of widely planted soybean cultivars leads to a new breeding strategy of crops in the post-genomic era. *The Crop Journal*, *9*, 1079–1087. https://doi.org/10.1016/j.cj.2021.01.001.

Qi, X. P., Li, M. W., Xie, M., Liu, X., Ni, M., Shao, G. H., et al. (2014). Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nature Communications*, *5*, 4340. https://doi.org/10.1038/ncomms5340.

Qin, J., Shi, A. N., Song, Q. J., Li, S., Wang, F. M., Cao, Y. H., et al. (2019). Genome wide association study and genomic selection of amino acid concentrations in soybean seeds. *Frontiers in Plant Science*, *10*, 1445. https://doi.org/10.3389/fpls.2019.01445.

Qiu, J., Wang, Y., Wu, S. L., Wang, Y. Y., Ye, C. Y., Bai, X. F., et al. (2014). Genome re-sequencing of semi-wild soybean reveals a complex *soja* population structure and deep introgression. *PLoS One*, *9*, e108479. https://doi.org/10.1371/journal.pone.0108479.

Quero, G., Simondi, S., Ceretta, S., Otero, A., Garaycochea, S., Fernandez, S., et al. (2021). An integrative analysis of yield stability for a GWAS in a small soybean breeding population. *Crop Science*, *61*, 1903–1914. https://doi.org/10.1002/csc2.20490.

Ravelombola, W. S., Qin, J., Shi, A. N., Nice, L., Bao, Y., Lorenz, A., et al. (2020). Genome-wide association study and genomic selection for tolerance of soybean biomass to soybean cyst nematode infestation. *PLoS One*, *15*, e0235089. https://doi.org/10.1371/journal.pone.0235089.

Ray, J. D., Dhanapal, A. P., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2015). Genome-wide association study of ureide concentration in diverse maturity group IV soybean [*Glycine max*(L.) Merr.] accessions. *G3-Genes Genomes Genetics*, *5*, 2391–2403. https://doi.org/10.1534/g3.115.021774.

Rigola, D., van Oeveren, J., Janssen, A., Bonne, A., Schneiders, H., van der Poel, H. J. A., et al. (2009). High-throughput detection of induced mutations and natural variation using KeyPoint(TM) technology. *PLoS One*, *4*, e4761. https://doi.org/10.1371/journal.pone.0004761.

Rolling, W. R., Dorrance, A. E., & McHale, L. K. (2020). Testing methods and statistical models of genomic prediction for quantitative disease resistance to *Phytophthora sojae* in soybean [*Glycine max* (L.) Merr] germplasm collections. *Theoretical and Applied Genetics*, *133*, 3441–3454. https://doi.org/10.1007/s00122-020-03679-w.

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

Seck, W., Torkamaneh, D., & Belzile, F. (2020). Comprehensive genome-wide association analysis reveals the genetic basis of root system architecture in soybean. *Frontiers in Plant Science*, *11*, 590740. https://doi.org/10.3389/fpls.2020.590740.

Shen, Y. T., Du, H. L., Liu, Y. C., Ni, L. B., Wang, Z., Liang, C. Z., et al. (2019). Update soybean Zhonghuang 13 genome to a golden reference. *Science China-Life Sciences*, *62*, 1257–1260. https://doi.org/10.1007/s11427-019-9822-2.

Shen, Y. T., Liu, J., Geng, H. Y., Zhang, J. X., Liu, Y. C., Zhang, H. K., et al. (2018). *De novo* assembly of a Chinese soybean genome. *Science China-Life Sciences*, *61*, 871–884. https://doi.org/10.1007/s11427-018-9360-0.

Shimomura, M., Kanamori, H., Komatsu, S., Namiki, N., Mukai, Y., Kurita, K., et al. (2015). The *Glycine max* cv. Enrei genome for improvement of Japanese soybean cultivars. *International Journal of Genomics*, *2015*, 358127. https://doi.org/10.1155/2015/358127.

Shook, J. M., Zhang, J., Jones, S. E., Singh, A., Diers, B. W., & Singh, A. K. (2021). Meta-GWAS for quantitative trait loci identification in soybean. *G3-Genes Genomes Genetics*, *11*, jkab117. https://doi.org/10.1093/g3journal/jkab117.

Shultz, J. L., Kurunam, D., Shopinski, K., Iqbal, M. J., Kazi, S., Zobrist, K., et al. (2006). The soybean genome database (SoyGD): A browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of *Glycine max*. *Nucleic Acids Research*, *34*, D758–D765. https://doi.org/10.1093/nar/gkj050.

Song, Q. J., Marek, L. F., Shoemaker, R. C., Lark, K. G., Concibido, V. C., Delannay, X., et al. (2004). A new integrated genetic linkage map of the soybean. *Theoretical and Applied Genetics*, *109*, 122–128. https://doi.org/10.1007/s00122-004-1602-3.

Song, Q. J., Yan, L., Quigley, C., Fickus, E., Wei, H., Chen, L. F., et al. (2020). Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. *Plant Journal*, *104*, 800–811. https://doi.org/10.1111/tpj.14960.

Sprink, T., Metje, J., & Hartung, F. (2015). Plant genome editing by novel tools: TALEN and other sequence specific nucleases. *Current Opinion in Biotechnology*, *32*, 47–53. https://doi.org/10.1016/j.copbio.2014.11.010.

Stacey, M. G., Cahoon, R. E., Nguyen, H. T., Cui, Y. Y., Sato, S., Nguyen, C. T., et al. (2016). Identification of homogentisate dioxygenase as a target for vitamin E biofortification in oilseeds. *Plant Physiology*, *172*, 1506–1518. https://doi.org/10.1104/pp.16.00941.

Steketee, C. J., Schapaugh, W. T., Carter, T. E., & Li, Z. L. (2020). Genome-wide association analyses reveal genomic regions controlling canopy wilting in soybean. *G3-Genes Genomes Genetics*, *10*, 1413–1425. https://doi.org/10.1534/g3.119.401016.

Sugano, S., Hirose, A., Kanazashi, Y., Adachi, K., Hibara, M., Itoh, T., et al. (2020). Simultaneous induction of mutant alleles of two allergenic genes in soybean by using site-directed mutagenesis. *BMC Plant Biology*, *20*, 513. https://doi.org/10.1186/s12870-020-02708-6.

Sun, X. J., Hu, Z., Chen, R., Jiang, Q. Y., Song, G. H., Zhang, H., et al. (2015). Targeted mutagenesis in soybean using the CRISPR-Cas9 system. *Scientific Reports*, *5*, 10342. https://doi.org/10.1038/srep10342.

Suzuki, C., Miyoshi, T., Shirai, S., Yumoto, S., Tanaka, Y., Hagihara, S., et al. (2017). A new soybean variety 'Yukihomare R' introduced resistance for soybean cyst nematode race1 into 'Yukihomare' by marker assisted selection. *Bulletin of Hokkaido Research Organization Agricultural Experiment Stations*, *101*, 33–47.

Tang, F., Yang, S., Liu, J., & Zhu, H. (2015). *Rj4*, a gene vontrolling nodulation specificity in soybeans, encodes a thaumatin-like protein but not the one previously reported. *Plant Physiology*, *170*, 26–32. https://doi.org/10.1104/pp.15.01661.

Todd, J. J., & Vodkin, L. O. (1996). Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. *Plant Cell*, *8*, 687–699. https://doi.org/10.1105/tpc.8.4.687.

Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoughue, L., Cober, E., Rajcan, I., et al. (2018). Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant Biotechnology Journal*, *16*, 749–759. https://doi.org/10.1111/pbi.12825.

Torkamaneh, D., Lemay, M. A., & Belzile, F. (2021). The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. *Plant Biotechnology Journal*, *19*, 1852–1862. https://doi.org/10.1111/pbi.13600.

Trevisan, R., Perez, O., Schmitz, N., Diers, B., & Martin, N. (2020). High-throughput phenotyping of soybean maturity using time series UAV imagery and convolutional neural networks. *Remote Sensing*, *12*, 3617. https://doi.org/10.3390/rs12213617.

Tuteja, J. H., Zabala, G., Varala, K., Hudson, M., & Vodkin, L. O. (2009). Endogenous, tissue-specific short interfering RNAs silence the *chalcone synthase* gene family in *Glycine max* seed coats. *Plant Cell*, *21*, 3063–3077. https://doi.org/10.1105/tpc.109.069856.

Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S., & Gregory, P. D. (2010). Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*, *11*, 636–646. https://doi.org/10.1038/nrg2842.

Valliyodan, B., Brown, A. V., Wang, J. X., Patil, G., Liu, Y., Otyama, P. I., et al. (2021). Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. *Scientific Data*, *8*, 50. https://doi.org/10.1038/s41597-021-00834-w.

Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S. Q., Brown, A. V., Ren, L. H., et al. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. *Plant Journal*, *100*, 1066–1082. https://doi.org/10.1111/tpj.14500.

Wang, Y., Gu, Y. Z., Gao, H. H., Qiu, L. J., Chang, R. Z., Chen, S. Y., et al. (2016). Molecular and geographic evolutionary support for the essential role of GIGANTEAa in soybean domestication of flowering time. *BMC Evolutionary Biology*, *16*, 79. https://doi.org/10.1186/s12862-016-0653-9.

Wang, X., Li, M. W., Wong, F. L., Luk, C. Y., Chung, C. Y. L., Yung, W. S., et al. (2021). Increased copy number of *gibberellin 2-oxidase 8* genes reduced trailing growth and shoot length during soybean domestication. *Plant Journal*, *107*, 1739–1755. https://doi.org/10.1111/tpj.15414.

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*, 808–810. https://doi.org/10.1038/Nmeth.2023.

Wang, L. W., Sun, S., Wu, T. T., Liu, L. P., Sun, X. G., Cai, Y. P., et al. (2020). Natural variation and CRISPR/Cas9-mediated mutation in *GmPRR37* affect photoperiodic flowering and contribute to regional adaptation of soybean. *Plant Biotechnology Journal*, *18*, 1869–1881. https://doi.org/10.1111/pbi.13346.

Wu, D. P., Li, D. M., Zhao, X., Zhan, Y. H., Teng, W. L., Qiu, L. J., et al. (2020). Identification of a candidate gene associated with isoflavone content in soybean seeds using genome-wide association and linkage mapping. *Plant Journal*, *104*, 950–963. https://doi.org/10.1111/tpj.14972.

Wu, N., Lu, Q., Wang, P. W., Zhang, Q., Zhang, J., Qu, J., et al. (2020). Construction and analysis of *GmFAD2-1A* and *GmFAD2-2A* soybean fatty acid desaturase mutants based on CRISPR/Cas9 technology. *International Journal of Molecular Sciences*, *21*, 1104. https://doi.org/10.3390/ijms21031104.

Xavier, A., Muir, W. M., & Rainey, K. M. (2016). Assessing predictive properties of genome-wide selection in soybeans. *G3-Genes Genomes Genetics*, *6*, 2611–2616. https://doi.org/10.1534/g3.116.032268.

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

Xie, W. B., Feng, Q., Yu, H. H., Huang, X. H., Zhao, Q. A., Xing, Y. Z., et al. (2010). Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 10578–10583. https://doi.org/10.1073/pnas.1005931107.

Xu, H. Y., Li, Y. J., Zhang, K. F., Li, M. J., Fu, S. Y., Tian, Y. Z., et al. (2021). miR169c-NFYA-C-ENOD40 modulates nitrogen inhibitory effects in soybean nodulation. *New Phytologist*, *229*, 3377–3392. https://doi.org/10.1111/nph.17115.

Yamada, T., Makita, H., Funatsuki, H., Takahashi, K., Hirata, K., Hishinuma, A., et al. (2017). Causal analysis of yield-increase by introgression of shattering resistance gene *pdh1* in soybean. *Japanese Journal of Crop Science*, *86*, 251–257. https://doi.org/10.1626/jcs.86.251.

Yang, C., Yan, J., Jiang, S., Li, X., Min, H., Wang, X., et al. (2021). Resequencing 250 soybean accessions: New insights into genes associated with agronomic traits and genetic networks. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2021.02.009.

Yi, X., Liu, J., Chen, S., Wu, H., Liu, M., Xu, Q., et al. (2021). Platinum-grade soybean reference genome facilitates characterization of genetic underpinnings of traits. *Research Square*. https://doi.org/10.21203/rs.3.rs-57915/v1.

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: Predicting yield from hyperspectral reflectance in soybean. *Frontiers in Plant Science*, *11*, 624273. https://doi.org/10.3389/fpls.2020.624273.

Yue, Y. L., Liu, N. X., Jiang, B. J., Li, M., Wang, H. J., Jiang, Z., et al. (2017). A single nucleotide deletion in *J* encoding GmELF3 confers long juvenility and is associated with adaption of tropic soybean. *Molecular Plant*, *10*, 656–658. https://doi.org/10.1016/j.molp.2016.12.004.

Zhang, H. Y., Goettel, W., Song, Q. J., Jiang, H., Hu, Z. B., Wang, M. L., et al. (2020). Dual use and selection of GmSWEET39 for oil and protein improvement in soybean. *PLoS Genetics*, *16*, e1009114. https://doi.org/10.1371/journal.pgen.1009114.

Zhang, Y. H., He, J. B., Wang, H. W., Meng, S., Xing, G. N., Li, Y., et al. (2018). Detecting the QTL-allele system of seed oil traits using multi-locus genome-wide association analysis for population characterization and optimal cross prediction in soybean. *Frontiers in Plant Science*, *9*, 1793. https://doi.org/10.3389/fpls.2018.01793.

Zhang, W., Liao, X. L., Cui, Y. M., Ma, W. Y., Zhang, X. N., Du, H. Y., et al. (2019). A cation diffusion facilitator, GmCDF1, negatively regulates salt tolerance in soybean. *PLoS Genetics*, *15*, e1007798. https://doi.org/10.1371/journal.pgen.1007798.

Zhang, J. P., Song, Q. J., Cregan, P. B., & Jiang, G. L. (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theoretical and Applied Genetics*, *129*, 117–130. https://doi.org/10.1007/s00122-015-2614-x.

Zhang, B., Wang, M. D., Sun, Y. F., Zhao, P., Liu, C., Qing, K., et al. (2021). Glycine max NNL1 restricts symbiotic compatibility with widely distributed bradyrhizobia via root hair infection. *Nature Plants*, 7, 239. https://doi.org/10.1038/s41477-021-00872-7.

Zhang, W., Xu, W. J., Zhang, H. M., Liu, X. Q., Cui, X. Y., Li, S. S., et al. (2021). Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean. *Theoretical and Applied Genetics*, *134*, 1329–1341. https://doi.org/10.1007/s00122-021-03774-6.

Zhang, D., Zhang, H. Y., Hu, Z. B., Chu, S. S., Yu, K. Y., Lv, L. L., et al. (2019). Artificial selection on GmOLEO1 contributes to the increase in seed oil during soybean domestication. *PLoS Genetics*, *15*, e1008267. https://doi.org/10.1371/journal.pgen.1008267.

Zhao, X., Dong, H. R., Chang, H., Zhao, J. Y., Teng, W. L., Qiu, L. J., et al. (2019). Genome wide association mapping and candidate gene analysis for hundred seed weight in soybean [*Glycine max* (L.) Merrill]. *BMC Genomics*, *20*, 648. https://doi.org/10.1186/s12864-019-6009-2.

Zhao, X., Li, W. J., Zhao, X. Y., Wang, J. Y., Liu, Z. Y., Han, Y. P., et al. (2019). Genome-wide association mapping and candidate gene analysis for seed shape in soybean (*Glycine max*). *Crop & Pasture Science*, *70*, 684–693. https://doi.org/10.1071/Cp19028.

Zhou, Z. K., Jiang, Y., Wang, Z., Gou, Z. H., Lyu, J., Li, W. Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, *33*, 408–U125. https://doi.org/10.1038/nbt.3096.

Zou, J. A., Li, W. J., Zhang, Y. T., Song, W., Jiang, H. P., Zhao, J. Y., et al. (2021). Identification of *glutathione transferase* gene associated with partial resistance to *Sclerotinia* stem rot of soybean using genome-wide association and linkage mapping. *Theoretical and Applied Genetics*, *134*, 2699–2709. https://doi.org/10.1007/s00122-021-03855-6.