CROP BREEDING AND APPLIED BIOTECHNOLOGY

ARTICLE

The numbers game of soybean breeding in the United States

Caio Canella Vieira[1] and Pengyin Chen[1*]

Crop Breeding and Applied Biotechnology
21(S): e387521S10, 2021
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332021v21Sa23

Abstract: *Soybean [Glycine max (L.) Merr.] represents one of the most essential crops to the world's economy and food security due to its unique seed composition. Public soybean breeding programs in the United States played an important role in developing the genetic basis of American soybean and discovering many economically important traits. After the passage of the Plant Variety Protection Act (PVP) in 1970 and the authorization to patent living matter in 1980, private companies have dominated the market share of commercial soybean varieties and public breeding programs shifted the efforts towards basic and applied research and education of the next generation of plant breeders. The short history of soybean breeding combined with a very narrow genetic basis derived from few ancestors can only make us reflect on all the innovations yet to be unveiled and the multiple possibilities to explore the unique traits that the golden miracle bean offers.*

Keywords*: Crop domestication, genetic diversity, breeding pipeline, predictive breeding.*

INTRODUCTION

Soybean [*Glycine max* (L.) Merr.], often referred to as the "*golden miracle bean"* due to its unique seed composition and versatile uses, represents the largest and most concentrated segment of global agricultural trade (Gale et al. 2019). With a constantly growing world population and demand for not only sufficient but high-quality food, soybean rises as a crop that delivers the highest amount of protein per hectare and accounts for over 60% of total global oilseed production (United States Department of Agriculture 2021). In the United States, the soybean value chain has an economic impact of $115.8 billion per year and supports over 357.000 people involved in the sector (LMC International 2019). In Midwestern states where soybean production is predominant, the contribution to the gross domestic product can reach as high as ten percent overall (LMC International 2019). But it was not until World War II that soybean became a major row crop in the United States. It was during this period that early breeding efforts shifted the plant architecture from a viny morphology to an upright plant, and the value of its seed composition was extensively explored (Singh and Hymowitz 1999, Anderson et al. 2019).

The soybean domestication and primary center of origin are tied to the Shang dynasty and date back to 1600 to 1046 B.C. in the eastern half of North China (Singh and Hymowitz 1999, Sleper and Shannon 2003, Anderson et al. 2019). The migration of populations and consolidation of new territories resulted in the expansion of soybean to central and south China and peninsular Korea by

***Corresponding author:**
E-mail: chenpe@missouri.edu
ORCID: 0000-0002-2663-7734

Received: 21 June 2021
Accepted: 27 June 2021
Published: 05 July 2021

[1] University of Missouri - Fisher Delta Research Center, 147 W. State Highway T, Portageville, MO 63873, US

the first century A.D. (Singh and Hymowitz 1999). Soybean was first introduced in the Western hemisphere in the late 16th century in Europe, and in 1765 Samuel Bowen introduced soybeans from China to Savannah, Georgia where it was used to manufacture soy sauce and soybean noodles (Hymowitz and Harlan 1983). In this communication, the authors intent to briefly review the history of soybean becoming an economically important crop in the U.S. and progressive stages of breeding effort towards genetic gain and overall improvements on yield, seed quality, defensive traits, and herbicide technologies. In addition, the authors take a look at the current status of soybean breeding and genetic improvement, accomplishments, bottleneck, and challenges, and then finally a future outlook for new technologies and tools towards yield breakthroughs.

THE PATH TO BECOME A MAJOR CROP IN THE UNITED STATES

Soybeans were initially grown in the United States as a forage crop for livestock feed, pasture, silage, and hay (Probst and Judd 1973). The economic value of the soybean seed composition started to gain attention in the early 20th century when George Washington Carver promoted soybeans for their elevated protein content and soil-health benefits when incorporated in crop rotation. Osborne and Mendel (1917) suggested the use of soybean as a protein source for human consumption under the premise of supporting the demand for a "cheaply produced and easily obtained source of all nutrients, and particularly of suitable proteins and fats". Their observation of normal growth in rats being fed with a soybean-based ration indicated that the crop contained an adequate amount of macro and micronutrients for proper development and could be widely used for human nutrition (Osborne and Mendel 1917).

World War II was a milestone in the soybean path to becoming a major crop in the United States. Tremendous national and international demand for oil and protein-based products impelled extensive efforts in soybean research and production (Shurtleff and Aoyagi 2004). Soybean research, which until then was mostly small-scale based on trial and error, became a multidisciplinary science-based conglomerate including plant breeding and genetics, plant physiology, plant nutrition, plant pathology, and soil-related sciences (Singh and Hymowitz 1999). Soybean cultivars were predominantly selections of superior plants from introduced accessions such as CNS (PI 548445), S-100 (PI 548488), and Mandarim (PI 548378) where plant breeders selected upright, non-shattering plants with local adaptability, high yield potential, and superior tolerance to biotic and abiotic stressors. A cooperative program between the United States Department of Agriculture (USDA) and State Agricultural Experiment Stations (AES) contributed to the development of the first soybean varieties based on hybridization breeding techniques (Shurtleff and Aoyagi 2004). The cross-combination of diverse parental lines resulted in allelic combinations once unattainable and drastically contributed to increasing soybean yield and total production. In addition, this period also marked the widespread implementation of mechanization in crop management operations, which significantly reduced harvest losses and increased realized yields across the country (Strand 1948). Between 1941 and 1942, soybean production in the United States increased almost 80% (106 to 188 million bushels), and the country became the largest producer in the world for the first time (USDA-NASS 2020a). This marked the rise of soybeans as *the golden miracle bean* and a crop with unprecedented impact on the world's economy and food security. It also represents a milestone in soybean research, particularly the consolidation of soybean breeding and genetics.

GENETIC BASE OF SOYBEAN VARIETIES IN THE UNITED STATES

The historical domestication of soybean, the introduction in the United States, and further intensification of selection of superior soybean plants strictly based on yield potential and adaptability to targeted environments have resulted in many genetic bottlenecks that severely limited the availability of genetic diversity and improvement in soybean. It is estimated that 50% of genetic diversity and over 80% of rare alleles were lost during domestication and artificial selection, which may include economically important genes that could enhance seed quality-related traits as well as tolerance to biotic and abiotic stressors (Hyten et al. 2006). An influx of soybean accessions from different countries occurred in the late 19th century after the establishment of the Office of Foreign Seed and Plant Introduction within the USDA and resulted in over 10,000 soybean accessions to further explore the genetics and economic value of the crop (Morse et al. 1949).

However, the soybean ancestors in the United States, which is defined as a founding stock with no known pedigree (Gizlice et al. 1994), are limited to only 17 accessions that contribute to 86% of the parentage of modern cultivars and may explain the significant loss of diversity and rare alleles during domestication and selection (Hyten et al. 2006). More

than half of the genetic diversity encountered in public cultivars can be explained by only six ancestors: Mandarin, CNS, Richland (PI 548406), S-100, and the two unknown parents of Lincoln (Gizlice et al. 1994). Interestingly, the same pattern of diversity is observed in Brazil, where 14 ancestors account for nearly 92% of the genetic diversity and CNS, S-100, Roanoke (PI 548485), and Tokyo (PI 548493) contribute to over 55% of the diversity (Wysmierski and Vello 2013). A large portion of genetic diversity from both countries is derived from the common ancestors CNS, S-100, Tokyo, and PI 54610. The genetic differences between North American and Brazilian soybean cultivars are likely caused by alleles inherited from Mandarin, Richland, Harrow (PI 548298), and Mukden (PI 548391) in North American cultivars, and Roanoke, PI 60406, Arksoy (PI 548438), Haberlandt (PI 548456), and Bilomi (PI 240664) in Brazilian cultivars (Gizlice et al. 1994, Wysmierski and Vello 2013). These ancestors may explain both common and unique phenotypic variance observed between North American and Brazilian soybean cultivars, including seed quality-related traits, biotic and abiotic stress tolerance, and overall agronomic and yield-related traits.

CNS, S-100, and PI 54610 are ancestors from China (Jiangsu, Jilin, and Heilongjiang, respectively), whereas Tokyo represents the major ancestor from Japan. These ancestors have contributed with key economically important alleles conferring biotic and abiotic tolerance, as well as adaptive traits in modern soybean cultivars (Table 1). For instance, CNS has been reported to carry the original alleles conferring tolerance to major soybean diseases including stem canker (*Diaporthe phaseolorum* var. *caulivora*) (Keeling 1982), phytophthora rot (*Phytophthora sojae* M.J. Kaufmann & J.W. Gerdemann) (Kilen et al. 1974), powdery mildew (*Microsphaera diffusa* Cooke & Peck)(Lohnes and Bernard 1992), bacterial leaf pustule (*Xanthomonas phaseoli* var *sojensis* Starr & Burk) (Hartwig and Lehman 1951), and peanut mottle virus (Chang et al. 2017). S-100 has been particularly known for carrying alleles conferring abiotic tolerance, including tolerance to salt (Lee et al. 2004, Lee et al. 2009), synthetic abscisic acid (Sloger and Caldwell 1970), and sulfentrazone (Hulting et al. 2001); it also represents the ancestor that conferred resistance to Asian soybean rust (*Phakopsora pachyrhizi* Syd. & P. Syd) (Monteros et al. 2010). S-100 and Tokyo were critical throughout the domestication process by conferring resistance to seed shattering (Funatsuki et al. 2014). Tokyo and CNS were also reported to carry the alleles conferring tolerance to soybean mosaic virus strains G1 and G5 (Wang et al. 2005), and PI 54610 has been found to carry alleles conferring tolerance to both sudden death syndrome (*Fusarium virguliforme* O'Donell & T. Aoki) and frogeye leaf spot (*Cercospora sojina* Hara) (Baker et al. 1999, Mueller et al. 2003, Mian et al. 2008).

Prior to the establishment of hybridization techniques, the ancestors were the predominant soybeans grown in the United States, including selections of their superior single plants or natural-occurring crosses in the ancestors (Morse and Cartter 1939). After hybridization, these were used as parents to develop soybean varieties that later constituted

Table 1. Major tolerance to biotic and abiotic stressors inherited by ancestors in the United States and Brazil

Name	PI Number	Origin	Stressor tolerance	References
CNS	PI 548445	China	Bacterial Leaf Pustule	Hartwig and Lehman (1951)
			Frogeye Leaf Spot	Baker et al. (1999), Mian et al. (2008)
			Peanut Mottle Virus	Chang et al. (2017)
			Phytophthora Rot	Kilen et al. (1974)
			Powdery Mildew	Lohnes and Bernard (1992)
			Soybean Mosaic Virus	Wang et al. (2005)
			Stem Canker	Keeling (1982)
			Asian Soybean Rust	Monteros et al. (2010)
S-100	PI 548488	China	Salt	Lee (2004), Lee et al. (2009)
			Shattering	Funatsuki et al. (2014)
			Sudden Death Syndrome	Mueller et al. (2003)
			Sulfentrazone	Hulting et al. (2001)
			Synthetic Abscisic Acid	Sloger and Caldwell (1970)
PI 54610	PI 54610	China	Frogeye Leaf Spot	Baker et al. (1999), Mian et al. (2008)
			Sudden Death Syndrome	Mueller et al. (2003)
Tokyo	PI 548493	Japan	Shattering	Funatsuki et al. (2014)
			Soybean Mosaic Virus	Wang et al. (2005)

the genetic basis of soybean cultivars in the United States. Soybean varieties "Lincoln" (PI 548362), which is estimated to provide 24% of the genetic base of cultivars in the northern US, and "Lee" (PI 548656), which contributes to 46% of the genetic base of cultivars in the southern US (Gizlice et al. 1994), were the first relevant progenies obtained through hybridization. Interestingly, northern and southern US varieties have maintained a separate pattern of diversity or gene pool to the present day where programs within a similar latitude location share little to no genetic diversity but rather distinct genetic diversity across latitudes. Studies have identified clear distinct genetic clusters dividing southern and northern varieties, where the minimum overlapping observed is likely due to shared ancestors (Wolfgang and Charles An 2017). Recently public breeding programs from diverse geographical regions have focused on germplasm exchange aiming to generate unique and superior allelic combinations for maximum genetic improvement possible (Chen et al. 2020, Chen et al. 2021).

In the Southern United States, 17 ancestors contributed to over 90% of the genes in cultivars adapted to this growing region, of which ancestors CNS and S-100 contributed to nearly 50% of the genetic diversity. CNS and S-100 are also the two most common ancestors in Brazilian soybean cultivars. This is likely related to better adaptation of these two common ancestors due to similar geographical and environmental features between the two regions. First-generation progenies derived from these ancestors account for approximately 60% of the genetic diversity in Southern U.S. cultivars, including Lee (PI 548656), Hill (PI 548654), Volstate (PI 548494), Ogden (PI 548477), and D49-2491 (Gizlice et al. 1994). In the Northern United States, the two unknown parents of Lincoln alone accounted for over 25% of the genetic variability in Northern cultivars. Additionally, Mandarin (PI 548378), Richland (PI 548406), Harrow (PI 548298), and Mukden (PI 548391) contributed to an additional 40% of the genetic variability. These indicate that Northern United States cultivars do not share common ancestors with Brazilian cultivars and are likely genetically distant due to different geographical and environmental features. As for the first-generation progenies, Lincoln (PI 548362), Harosoy (PI 548573), and Clark (PI 548533) accounted for nearly 50% of total genetic diversity in the Northern varieties (Gizlice et al. 1994).

We attempted to summarize the most relevant soybean varieties that dominated the U.S. production acreage from the 1940s to the current days in the Southern (Table 2) and Northern (Table 3) U.S. The predominant varieties were divided into decades starting in the 1940s and classified based on their maturity group. In the Southern varieties, maturity groups 3 and 4 were classified as 'Early', 5 and 6 as 'Mid', and above 6 as 'Late'. In the Northern varieties, maturity groups 00, 0, and 1 were classified as 'Early', 2 and 3 as 'Mid', and above 3 as 'Late'. These varieties constituted the chosen reference checks for each maturity group in the USDA Uniform Trials Southern (https://www.ars.usda.gov/southeast-area/stoneville-ms/crop-genetics-research/docs/uniform-soybean-tests/) and Northern States (https://www.ars.usda.gov/midwest-area/west-lafayette-in/crop-production-and-pest-control-research/docs/uniform-soybean-tests-northern-region/), which is a traditional collaborative trial focused on speeding up the advancement and non-biased evaluation of public soybean varieties. Public soybean breeders yearly submit the most advanced soybean breeding lines to be evaluated for yield performance in a multi-state replicated yield trial and characterize the disease resistance and quality traits and compare with the check varieties in order to make a release decision.

Public soybean breeding programs discovered and identified the majority of key economically important traits and established the genetic basis of modern soybean varieties. Advancements from the public sector have contributed to the approximate 370% increment in yield and 90,000% increment in production from the 1920s (740 kg ha^{-1} and 134 thousand metric tons, respectively) to the present day (3.470 kg ha^{-1} and 123 million metric tons, respectively) (Specht et al. 1999, Koester et al. 2014, Spetch et al. 2014, USDA-NASS 2020a). Before the passage of the Plant Variety Protection Act (PVP) in 1970, the development and release of soybean varieties in the United States were primarily conducted by public breeding programs. In 1980 with the authorization to patent living matter by the US Supreme Court, the dynamics of soybean breeding changed dramatically and private institutions found in soybean breeding a business model with a recurrent attractive return of investment. The rapid ascension and dominance of market share by private seed companies became more prominent with the development and wide adoption of patented biotechnology traits. In addition to profits from exclusive genetics, private seed companies enjoy the revenues from these traits and the agrochemicals often associated with said cropping system.

The first herbicide-tolerant soybean cultivar was developed by Monsanto in the 1980s and commercialized later in 1996. The trait conferred resistance to over-the-top applications of the herbicide glyphosate (*Roundup Ready*®) and intensely modified the cropping system once prevailed by conventional soybean varieties. In 2008, the second generation

of the technology (R2Y) was available in the United States under the name *Roundup Ready 2*® and was marketed as a higher-yielding version of the first generation technology (RR1). In 2009, a new herbicide-resistant soybean cultivar was commercialized by Bayer CropScience (sold in 2017 to BASF) and conferred resistance to over-the-top applications of the herbicide glufosinate (*LibertyLink*®*).* In 2016, Monsanto released the first soybean resistant to over-the-top applications of the herbicide dicamba (Xtend®) and quickly became the predominant soybean system in the United States with the expectation of being effective against herbicide-tolerant invasive weed species. In the same year, Dow AgroSciences released the first soybean with resistance to over-the-top applications of both herbicides 2,4-D and glyphosate (*Enlist duo*®). This technology was further advanced by the addition of resistance to over-the-top applications of the herbicide glufosinate, making the *Enlist E3*® technology first available in soybeans in 2019.

Currently, public breeding programs represent a small share of the soybean acreage in the United States and over 94% of the market share is accounted by varieties containing herbicide-tolerance traits (USDA-NASS 2020b). Public programs have shifted their efforts from supplying the national needs for soybean varieties towards basic and applied research, including trait identification, development of novel breeding technologies, and integration of multidisciplinary approaches to enhance breeding efficacy and genetic gain, as well as training and educating graduate students who will become the next generation plant breeders. The improved germplasm and released varieties from public programs are often provided to private entities as free genetics or licensed to interested parties as products or breeding materials.

Table 2. Major Reference soybean varieties in the Southern U.S. from the 1940s to 2020

Period	Maturity[1]	Reference soybean varieties[2]
1943-1950	Early	**S-100 (MO)**, Perry (IN), Wabash (IN)
	Mid	**Ogden (TN)**, **D49-2491 (ARS-MS)**
	Late	Roanoke (NC), **Volstate (TN)**, Mamotan (ARS-MS)
1951-1960	Early	Perry (IN)
	Mid	**S-100 (MO),** Dorman (ARS-MS), **Hill (ARS-MS)**
	Late	Roanoke (NC), Jackson (NC), **Lee (ARS-MS)**
1961-1970	Early	Kent (IN)
	Mid	**Hill (ARS-MS)**
	Late	Lee 68 (AR), Bragg (FL), Jackson (NC), **Lee (ARS-MS)**, Hood (ARS-MS)
1971-1980	Early	Columbus (KS), Crawford (KS), Kent (IN)
	Mid	**Hill (ARS-MS)**, Essex (VA)
	Late	Lee 68 (AR), Bragg (FL), Braxton (FL), Tracy (ARS-MS)
1981-1990	Early	Douglas (KS), Delsoy 4500 (MO)
	Mid	Essex (VA)
	Late	Braxton (FL), Thomas (GA), Centennial (ARS-MS), LeFlore (ARS-MS)
1991-2000	Early	KS 4694 (KS), Manokin (MD), Delsoy 4710 (MO)
	Mid	Holladay (NC), Essex (VA), Hutcheson (VA)
	Late	Stonewall (AL), Boggs (GA), Cook (GA), Haskell (GA), Prichard (GA), Bedford (MS), Brim (NC), Dillon (SC), LeFlore (ARS-MS)
2001-2010	Early	*DK 4866 (DK)*, *DK 4868 (DK)*, LD00-3309 (IL), LN97-15076 (IL), KS 4602N (KS), KS 4694 (KS), Manokin (MD), *AG 3904 (AG)*, *AG 4201 (AG)*, *AG 4403 (AG)*, *AG 4603 (AG)*, *AG 4903 (AG)*
	Mid	Osage (AR), DPL 5114 (DP), *AG 5501 (AG)*, *AG 5605 (AG)*, 5002T (TN), 5601T (TN), JTN 5503 (ARS-TN), Hutcheson (VA)
	Late	Benning (GA), Boggs (GA), Boggs RR (GA), Cook (GA), Haskell (GA), Haskell RR (GA), Prichard (GA), Prichard RR (GA), N7002 (GA), NC-Roy (NC), Dillon (SC)
2011-2020	Early	*DK 4866 (DK)*, LD00-3309 (IL), LD06-7620 (IL), *AG 3803 (AG)*, *AG 4103 (AG)*, *AG 4135 (AG)*, *AG 43X7 (AG)*, *AG 4403 (AG)*, *AG 46X7 (AG)*, *AG 4903 (AG)*, *AG 49X6 (AG)*
	Mid	Osage (AR), UA 5612 (AR), S11-20124 (MO), *AG 53X6 (AG)*, *AG 55X7 (AG)*, *AG 5606 (AG)*, *P95Y70 (PNR)*, 5002T (TN), 5601T (TN), Ellis (TN), TN11-5140 (TN), JTN 5203 (ARS-TN), JTN 5503 (ARS-TN)
	Late	G03-1187RR (GA), G04-1618RR (GA), *AG 6534 (AG)*, *AG 74X8 (AG)*, *AG 7934 (AG)*, N05-7432 (NC), N7002 (NC), N8001 (NC), N8002 (NC), NC-Dilday (NC), NC-Dunphy (NC), NC-Roy (NC), NC-Wilder (NC), Dillon (SC), TN08-109 (TN)

[1] Maturity: Classification based on maturity group, where 'Early' includes maturity groups 3 and 4; 'Mid' includes maturity groups 5 to 6; 'Late' includes maturity groups above 6. [2] Reference Soybean Varieties: Highlighted varieties represent the major varieties that constitute the genetic basis in Southern U.S; Italicized varieties are soybean developed by private companies.

Table 3. Major Reference soybean varieties in the Northern U.S. from the 1940s to 2020

Period	Maturity	Reference soybean varieties[2]
1941-1950	Early	Hawkey (IA), Illini (IL), **Mandarin (PI)**, Ottawa (PI)
	Mid	Dunfield (IN), **Richland (IN)**, **Lincoln (IL)**
	Late	Gibson (IN), Wabash (IN)
1951-1960	Early	Chippewa (IL), Acme (ON), Ottawa (PI), Grant (WI)
	Mid	Hawkey (IA), **Lincoln (IL)**, Shelby (IL)
	Late	**Clark (IL)**, Wabash (IN)
1961-1970	Early	Chippewa (IL), Portage (MB), Acme (ON), Merit (ON), Grant (WI)
	Mid	Corsoy (IA), Hawkey (IA), Harosoy 63 (IL), Shelby (IL), Wayne (IL)
	Late	**Clark (IL)**, Clark 63 (IL), Cutler (IN)
1971-1980	Early	Chippewa (IL), Portage (MB), Evans (MN), Hodgson (MN), Steele (MN), Swift (MN), Merit (ON)
	Mid	Corsoy (IA), Cumberland (IA), Woodworth (IA), Wayne (IL)
	Late	Union (IL), Cutler (IN)
1981-1990	Early	Portage (MB), Evans (MN), Dawson (MN), Glenwood (MN), Hodgson (MN), McCall (MN), Sibley (MN)
	Mid	Corsoy (IA), Cumberland (IA), Elgin (IA), Harper (IA), Kenwood (IA), Resnik (OH)
	Late	Union (IL), Spencer (IN), Sparks (KS), Morgan (MD)
1991-2000	Early	Glenwood (MN), Lambert (MN), McCall (MN), M84-916 (MN), Parker (MN)
	Mid	A94-77421 (IA), IA 2021 (IA), IA 3010 (IA), Kenwood (IA), Iroquois (IL), Resnik (OH)
	Late	Spencer (IN), HS93-4118 (OH), Stressland (OH)
2001-2010	Early	*AG 1501 (AG), AG 1602 (AG), AG 2302 (AG), AG 2403 (AG)*, Lambert (MN), M94-135066 (MN), M98-227065 (MN), McCall (MN), MN 0071 (MN), MN 1410 (MN), Parker (MN), RG200R (ND), Sheyenne (ND)
	Mid	*AG 2403 (AG)*, A99-315026 (IA), IA 2021 (IA), IA 2068 (IA), IA 2094 (IA), IA 3010 (IA), IA 3023 (IA), U03-827101 (NE), HF9667-2 (OH)
	Late	*AG 3401 (AG), AG 3505 (AG), AG 4103 (AG)*, AG 4201 (IL), LD00-3309 (IL), LN97-15076 (IL), HS93-4118 (OH)
2011-2020	Early	*AG 005X8 (AG), AG 00632 (AG), AG 11X8 (AG), AG 17X8 (AG)*, MN 0071 (MN), MN 0083 (MN), MN 1410 (MN), Dickey (ND), Sheyenne (ND), Stutsman (ND), U06-814223 (NE)
	Mid	*AG 25X8 (AG)*, IA 2094 (IA), IA 2102 (IA), IA 3023 (IA), LD11-2170 (IL), U03-827101 (NE)
	Late	*AG 4033 (AG)*, LD00-3309 (IL), LD06-7620 (IL), LD07-3395bf (IL)

[1] Maturity: Classification based on maturity group, where 'Early' includes maturity groups 3 and 4; 'Mid' includes maturity groups 5 to 6; 'Late' includes maturity groups above 6. [2] Reference Soybean Varieties: Highlighted varieties represent the major varieties that constitute the genetic basis in Northern U.S; Italicized varieties are soybean developed by private companies.

THE NUMBERS GAME OF SOYBEAN BREEDING

As the development of a new soybean variety relies heavily on superior allelic combinations randomly obtained through recombination, plant breeding is often considered a numbers game. But the numbers game of soybean breeding goes beyond the probability of obtaining the desired allelic combination. Plant breeders consistently need to play the numbers game in multiple aspects and stages of a breeding pipeline, including early decisions regarding how many parental lines and which lines to be entered in a crossing block, how many crossing combinations and hybridizations attempts to be executed, how many generations to be advanced prior to lines selections and field trials, and how many progeny rows to be derived and selected from a single cross. Then molecular characterization-related decisions such as how many lines and in which generation to be genotyped and how many molecular markers to cover enough genetic variability enter the numbers game. Later they will have to decide how many locations, replications, and years for yield trials, how many breeding lines to be advanced through the pipeline, and ultimately how many breeding lines to be released as varieties - and often how many units of seeds to be curated by foundation seed organizations. Certainly, each technical aspect of a breeding pipeline could be a numbers game of its own. But there are multiple ends in a breeding program, including managing human, natural and financial resources. Finding the right balance between workload, workforce, and resource availability in a public breeding program can be as much of a numbers game as developing the superior soybean variety with desired traits.

A traditional public soybean breeding pipeline takes four to six years to release a new cultivar or germplasm (Figure 1). The following pipeline and numbers are based on the soybean breeding program at the University of Missouri – Fisher Delta Research Center (MU-FDRC) and may differ from different programs in the United States.

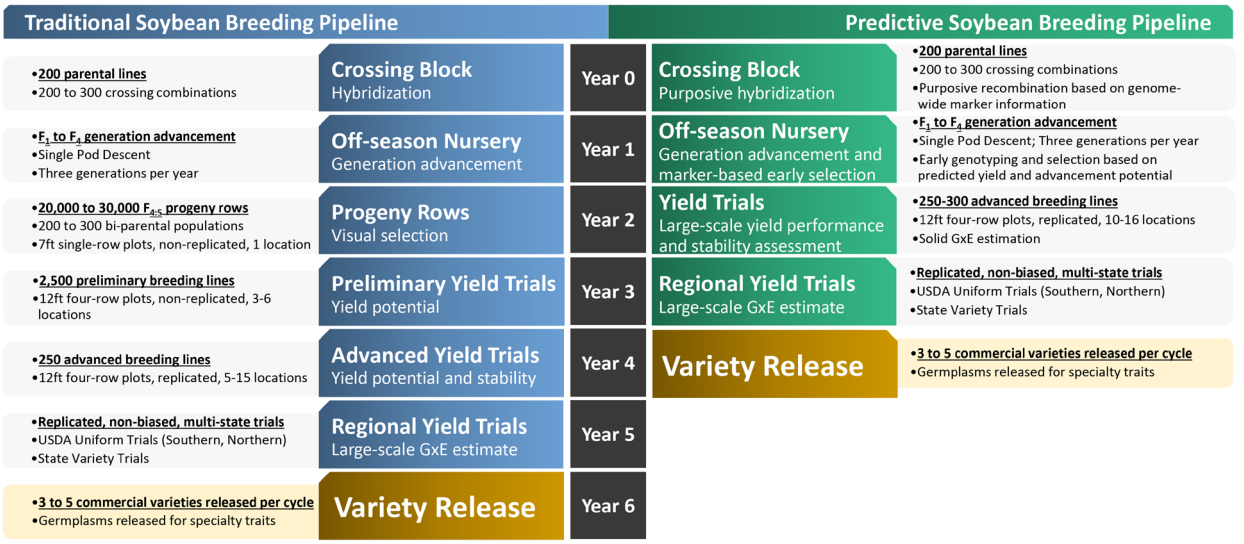

Figure 1. The traditional breeding cycle of a public soybean breeding pipeline in the United States and the possibility to enhance genetic gain based on a predictive breeding pipeline.

A breeding cycle usually starts with roughly 200 parental lines, of which anywhere from 300 to 400 crossing combinations will be attempted in summer in the United States. Through agreements and partnerships with private institutions, breeding programs can have access to herbicide-resistant technologies such as *Roundup Ready®*, *LibertyLink®*, *Enlist®*, and *Xtend®*. The F_1 hybrid seeds are advanced to F_4 generation using the modified single-pod descent method (Fehr 1987) in an off-season nursery where they can conduct three growing seasons in a year. Roughly 100-150 F_4 single plants per bi-parental population are harvested and threshed separately, and the $F_{4:5}$ seeds are grown and evaluated as progeny two years later. Conducting field trials in a soybean breeding program is labor and cost-intensive. The mechanization of field experiments with precision planters and plot combines equipped with GPS systems becomes essential to efficiently conduct soybean breeding on a commercial scale. It is common for public soybean breeding programs to grow 20,000.00 to 30,000.00 progeny rows in a year. The approximate cost for planting one progeny row is $10, resulting in a total cost of $200,000 to $300,000. The purpose of the progeny trial in a conventional breeding pipeline is to visually select the best breeding lines (top 10-15%) based on agronomic traits (such as maturity, plant height, lodging, and disease reactions), uniformity, and overall yield potential for further advancement through yield trials in the subsequent years. In year 2, approximately 2,500 selected breeding lines are tested in preliminary yield trials (PYT) across four environments using non-replicated 12-ft four-row plots. Given the cost of $25 to plant, manage and harvest a 12-ft four-row yield plot, the PYT stage can cost as much as $250,000. In year 3, the selected lines (approx. top 10%) from PYT are entered in advanced yield trials (AYTs). These lines are characterized both phenotypically and genotypically for their seed-quality related traits (protein, oil, and carbohydrate and fatty acid profiles) and response to various biotic stressors including plant-parasitic nematodes (soybean cyst [*Heterodera glycines* Ichinohe], southern-root-knot [*Meloidogyne incognita* (Kofold & White) Chitwood], and reniform nematodes [*Rotylenchulus reniformis* Linford & Oliveira]), stem canker, frogeye leafspot, sudden death syndrome, and phytophthora rot, and abiotic stressors including drought, flooding, and salt tolerance. The AYTs are grown in 12-ft four-row plots often replicated (2-3 replications) across multiple environments (5-15), which results in an estimated cost of $180,000. Although not common, public breeding programs may grow AYT lines in-house and also collaborate with public and private institutions to increase the number of yield plots and locations to better assess genotype by environment interactions. The best breeding lines (top 5-7%) from AYT are selected to be entered in regional trials (e.g., USDA Southern and Northern Uniform Trials and State Variety Trials), and upon satisfactory performance across multiple environments, promising lines are proposed for release to the state seed committee. In summary, the traditional public breeding pipeline takes four to six years to release a new soybean variety and the breeding cycle can cost from $600,000 up to $800,000.

INTEGRATION OF LARGE-SCALE LAYERS OF DATA TO ENHANCE GENETIC GAIN

As the fields of genomics and phenomics as well as the complexity of predictive analytics and AI-based models keep evolving at a fast pace, plant breeders now have access to powerful technologies to predict unobserved phenotypic values and support their advancement decisions. For instance, genome-wide selection can predict unobserved phenotypic values by estimating the effects of multiple loci across the genome (Crossa et al. 2017). Predicted phenotypic values, as well as classification-based metrics, can be implemented earlier in the breeding pipeline to discard inferior genotypes and potentially reduce yield trials in numbers and duration (Jarquin et al. 2014, Stewart-Brown et al. 2019). High-density molecular data can also be used to early characterize genotypes for multiple agronomic traits and responses to biotic and abiotic stressors, which in turn can assist in the selection of parental lines as well as designing effective crossing combinations (Neyhart et al. 2019). For instance, plant breeders may use molecular data to reduce segregation among traits (keep or increase desired fixed alleles) of interest among the crossing parents while simultaneously promoting recombination to achieve novel allelic combinations. In addition, aerial-based robotic platforms equipped with multiple sensors can rapidly and non-destructively collect an extensive amount of data which, when integrated with advanced computer vision, artificial intelligence, and big data analytics, can estimate phenotypes that are otherwise labor and cost-intensive including plant maturity (Zhou et al. 2019, Trevisan et al. 2020), plant height (Zhou et al. 2021), canopy coverage (Moreira et al. 2019), yield performance (Herrero-Huerta et al. 2020, Maimaitijiang et al. 2020, Zhou et al. 2021), and biotic and abiotic tolerance (Zhou et al. 2021). In the early stages (such as the progeny row stage) of a breeding pipeline where representative yield observations are limited and/or not possible to be obtained, the adoption of phenomics can assist breeders to make decisions with more confidence and quickly advance genotypes throughout the pipeline (Moreira et al. 2020). Integration of large-scale layers of data can quickly and precisely narrow down promising lines and reduce the need to evaluate inferior materials in the field, which in turn can drastically improve breeding efficiency and enhance genetic gain. Comparing to the traditional breeding pipeline, the use of predictive breeding may help plant breeders to not only reduce costs, time, and space but enhance genetic gain by reducing the length of the breeding cycle, as well as allowing the breeders to have a clear knowledge of the genetics of the materials early in the pipeline (Figure 1).

CONCLUSIONS

The significance of soybean in the world's economy and food security is undebatable. As a versatile crop with unprecedented seed composition, the "*golden miracle bean"* is widely used in the food, feed, and many other industries exploring oil and protein-based products. However, the history of soybean in the United States is recent with the first hybridization occurring not longer than 80 years ago. Public soybean breeding programs have developed the genetic basis and discovered the majority of economically important traits in the United States, and, although private companies have dominated the market share of commercial soybean varieties, public breeding programs are essential to train and educate the next generation of plant breeders and continue to unveil breakthrough traits and advancements in the field. Many soybean breeding programs in the United States are currently going through a transition from a primarily-based conventional breeding approach to the integration of both genomic and phenomic approaches using big data analytics to predict unobserved phenotypes, better understand the interaction between loci, and accelerate the discovery of superior allelic combinations. The numbers game of soybean breeding, which historically has relied on randomly obtained allelic combinations, is steadily becoming a purposive numbers game supported by elegant predictive analytics and large-scale multiple layers of data. The short history of soybean breeding combined with a very narrow genetic basis derived from a handful of ancestors can only make us reflect on all the innovations yet to be discovered and the multiple possibilities to explore the unique traits that soybean offers. The key to future success in improving genetic gain depends on genetic diversity, integration of various technologies, and next-generation breeders and geneticists.

ACKNOWLEDGMENTS

The authors would like to acknowledge the MU – Fisher Delta Research Center soybean breeding team for their efforts and tremendous contribution to the continuous development and release of public soybean varieties adapted to the Mid-south U.S, as well as the curiosity and support in planning, conducting and promoting our research.

REFERENCES

Anderson EJ, Ali LMd, Beavis WD, Chen P, Clemente TE, Diers BW, Graef GL, Grassini P, Hyten DL, McHale LK, Nelson RL, Parrott WA, Patil GB, Stupar RM and Tilmon KJ (2019) Soybean [*Glycine max* (L.) Merr.] Breeding: History, improvement, production and future opportunities. In Al-Khayri J, Jain S and Johnson D (Eds). **Advances in plant breeding strategies: Legumes**. Springer, Cham., p. 431-516.

Baker WA, Weaver DB, Qiu J and Pace PF (1999) Genetic analysis of frogeye leaf spot resistance in PI54610 and Peking soybean. **Crop Science 39**: 1021-1025.

Chang HX, Lipka AE, Domier LL and Hartman GL (2017) Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. **Frontiers in Plant Science 106**: 1139-1151.

Chen P, Shannon G, Ali ML, Scaboo A, Crisel M, Smothers S, Clubb M, Selves S, Vieira CC, Mitchum MG, Nguyen HT, Li Z, Bond J, Meinhardt C, Usovsky M, Li S, Mengistu A and Robbins RT (2020) Registration of 'S14-9017GT' soybean cultivar with high yield, resistance to multiple diseases, and high seed oil content. **Journal of Plant Registrations 14**: 347-356.

Chen P, Shannon G, Scaboo A, Crisel M, Smothers S, Clubb M, Selves S, Vieira CC, Ali ML, Lee D, Nguyen HT, Li Z, Mitchum MG, Bond J, Meinhardt C, Usovsky M, Li S, Mengistu A and Robbins RT (2021) 'S13-1955C': A high-yielding conventional soybean with high oil content, multiple disease resistance, and broad adaptation **Journal of Plant Registrations 15**: 318-325.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, Campos G de los, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J and Varshney RK (2017) Genomic selection in plant breeding: Methods, models, and perspectives. **Trends in Plant Science 22**: 961-975.

Fehr WR (1987) **Principles of cultivar development: Crop species**. Volume 2, Macmillan Publishing Company, New York, 773p.

Funatsuki H, Suzuki M, Hirose A, Inaba H, Yamada T, Hajika M, Komatsu K, Katayama T, Sayama T, Ishimoto M and Fujino K (2014) Molecular basis of a shattering resistance boosting global dissemination of soybean. **PNAS 111**: 1779-17802.

Gale F, Valdes C and Ash M (2019) **Interdependence of China, United States, and Brazil in soybean trade**. U.S. Department of Agriculture, Economic Research Service, 48p.

Gizlice Z, Carter Jnr TE and Burton JW (1994) Genetic base for North American public soybean cultivars released between 1947 and 1988. **Crop Science 34**: 1143-1151.

Hartwig EE and Lehman SG (1951) Inheritance of resistance to the bacterial pustule disease in soybeans. **Agronomy Journal 43**: 226-229.

Herrero-Huerta M, Rodriguez-Gonzalvez P and Rainey KM (2020) Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. **Plant Methods 16**: 1-16.

Hulting AG, Wax LM, Nelson RL and Simmons FW (2001) Soybean (*Glycine max* (L.) Merr.) cultivar tolerance to sulfentrazone. **Crop Protection 20**: 679-683.

Hymowitz T and Harlan JR (1983) Introduction of soybean to North America by Samuel Bowen in 1765. **Economic Botany 37**: 371-379.

Hyten DL, Song Q, Zhu Y, Choi I, Nelson RL, Costa JM, Specht JE, Shoemaker RC and Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. **PNAS 103**: 16666-16671.

Jarquin D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G and Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. **BMC Genomics 15**: 1-10.

Keeling BL (1982) A seedling test for resistance to soybean stem canker caused by *Diaporthe phaseolorum* var. *caulivora*. **Phytopathology 72**: 807-809.

Kilen TC, Hartwig EE and Keeling BL (1974) Inheritance of a second major gene for resistance to phytophthora rot in soybeans. **Crop Science 14**: 260-262.

Koester RP, Skoneczka JA, Cary TR, Diers BW and Ainsworth EA (2014) Historical gains in soybean (*Glycine max* Merr.) seed yield are driven by linear increases in light interception, energy conversion, and partitioning efficiencies. **Journal of Experimental Botany 65**: 3311-3321.

Lee GJ, Boerma HR, Villagarcia MR, Zhou X, Carter TE, Li Z and Gibbs MO (2004) A major QTL conditioning salt tolerance in S-100 soybean and descendent cultivars. **Theoretical and Applied Genetics 109:** 1610-1619.

Lee JD, Shannon JG, Vuong TD and Nguyen HT (2009) Inheritance of salt tolerance in wild soybean (*Glycine soja* Sieb. and Zucc.) accession PI483463. **Journal of Heredity 100**: 798-801.

LMC International (2019) **The economic impact of U.S. soybeans and end products on the U.S. economy**. Oxford, United Kingdom, 31p.

Lohnes DG and Bernard RL (1992) Inheritance of resistance to powdery mildew in soybeans. **Plant Disease 76**: 964-965.

Maimaitijiang M, Sagan V, Sidike P, Hartling S, Esposito F and Fritschi FB (2020) Soybean yield prediction from UAV using multimodal data fusion and deep learning. **Remote Sensing of Environment 237**: 1-20.

Mian MAR, Missaoui AM, Walker DR, Phillips DV and Boerma HR (2008) Frogeye leaf spot of soybean: A review and proposed race designations for isolates of *Cercospora sojina* Hara. **Crop Science 48**: 14-24.

Monteros MJ, Ha BK, Phillips DV and Boerma HR (2010) SNP assay to detect the "Hyuuga" red-brown lesion resistance gene for Asian soybean rust. **Theoretical and Applied Genetics 121**: 1023-1032.

Moreira FF, Hearst AA, Cherkauer KA and Rainey KM (2019) Improving the efficiency of soybean breeding with high-throughput canopy phenotyping. **Plant Methods 15**: 1-9.

Moreira FF, Oliveira HR, Volenec JJ, Rainey KM and Brito LF (2020)

Integrating high-throughput phenotyping and statistical genomic methods to genetically improve longitudinal traits in crops. **Frontiers in Plant Science 11**: 1-18.

Morse WJ and Cartter JL (1939) **Soybeans: culture and varieties. Farmers' bulletin**. United States Department of Agriculture, Washington D.C., 40p.

Morse WJ, Cartter JL and Williams LF (1949) **Soybeans: culture and varieties**. Farmers' bulletin (U. S. Department of Agriculture), no. 1520.

Mueller DS, Nelson RL, Hartman GL and Pedersen WL (2003) Response of commercially developed soybean cultivars and the ancestral soybean lines to *Fusarium solani* f. sp. *glycines*. **Plant Disease 87**: 827-831.

Neyhart JL, Lorenz AJ and Smith KP (2019) Multi-trait improvement by predicting genetic correlations in breeding crosses. **G3 Genes, Genomes, Genetic 9**: 3153-3165.

Osborne TB and Mendel LB (1917) The use of soybean as food**. Journal of Biological Chemistry 32**: 369-387.

Probst AH and Judd RW (1973) Origin, U.S. history and development, and world distribution. In Caldwell BE (Ed) **Soybeans: improvement, production, and uses**. American Society of Agronomy, Madison, p. 1-15.

Shurtleff W and Aoyagi A (2004) History of world soybean production and trade. In **History of soybeans and soyfoods, 1100 B.C. to the 1980s**. Soyfoods Center, Lafayette (https://www.soyinfocenter.com/HSS/production_and_trade1.php).

Singh RJ and Hymowitz T (1999) Soybean genetic resources and crop improvement. **Genome 42**: 605-616.

Sleper DA and Shannon JG (2003) Role of public and private soybean breeding programs in the development of soybean varieties using biotechnology. **AgBioForum 6**: 27-32.

Sloger C and Caldwell BE (1970) Response of cultivars of soybean to synthetic abscisic acid. **Plant Physiology 46**: 634-635.

Specht JE, Hume DJ and Kumudini SV (1999) Soybean yield potential-a genetic and physiological perspective. **Crop Science 39**: 1560-1570.

Spetch JE, Brian W, Diers D, Nelson L, Toledo JF, Torrion JA and Grassini P (2014) In Smith S, Specht J, Diers B and Carver B (eds) **Yield gains in major U.S. field crops**. CSSA Special Publications, Madison, p. 311-356.

Stewart-Brown BB, Song Q, Vaughn JN and Li Z (2019) Genomic selection for yield and seed composition traits within an applied soybean breeding program. **G3 Genes, Genomes, Genetic 9**: 2253-2265.

Strand EG (1948) **Soybeans in American farming**. USDA, Washington, 67p. (Technical Bulletin 966).

Trevisan R, Pérez O, Schmitz N, Diers B and Martin N (2020) High-throughput phenotyping of soybean maturity using time series UAV imagery and convolutional neural networks**. Remote Sensing 12**: 1-19.

United States Department of Agriculture (2021) Oilseeds: world markets and trade. Available at <https://www.fas.usda.gov/data/oilseeds-world-markets-and-trade>. Accessed in June 2021.

USDA-NASS (2020a) Crop production historical track records; April 2019. Available at <https://www.nass.usda.gov/Publications/Todays_Reports/reports/croptr19.pdf>. Accessed on June 10, 2021.

USDA-NASS (2020b) Corn and soybean acreage. Available at <https://www.nass.usda.gov/Publications/Todays_Reports/reports/acrg0620.pdf>. Accessed on May 10, 2021.

Wang Y, Hobbs HA, Hill CB, Domier LL, Hartman GL and Nelson RL (2005) Evaluation of ancestral lines of U.S. soybean cultivars for resistance to four soybean viruses. **Crop Science 45**: 639-644.

Wolfgang G and Charles An Y-q (2017) Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci. **Molecular Breeding 37**: 1-9.

Wysmierski PT and Vello NA (2013) The genetic base of Brazilian soybean cultivars: Evolution over time and breeding implications. **Genetic Molecular Biology 36**: 547-555.

Zhou J, Yungbluth C, Vong CN, Scaboo A and Zhou J (2019) Estimation of the maturity date of soybean breeding lines using UAV-based multispectral imagery. **Remote Sensing 11**: 1-17.

Zhou J, Zhou J, Ye H, Ali ML, Chen P and Nguyen HT (2021) Yield estimation of soybean breeding lines under drought stress using unmanned aerial vehicle-based imagery and convolutional neural network. **Biosystems Engineering 204**: 90-103.