ANALYSIS OF YIELD AND GENETIC DIVERSITY WITHIN NORTH DAKOTA STATE UNIVERSITY SOYBEAN [*GLYCINE MAX* (L.) MERR] CULTIVARS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Forrest Jay Hanson

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Plant Sciences

May 2024

Fargo, North Dakota

North Dakota State University
Graduate School

Title

ANALYSIS OF YIELD AND GENETIC DIVERSITY WITHIN NORTH DAKOTA STATE UNIVERSITY SOYBEAN [*GLYCINE MAX* (L.) MERR] CULTIVARS

By

Forrest Jay Hanson

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Carrie Miranda
Chair

Dr. Nonoy Bandillo

Dr. Paulo Flores

Approved:

5/28/2024
Date

Dr. Richard Horsley
Department Chair

ABSTRACT

Soybean [*Glycine max* (L.) Merr.] has historically been subject to extensive breeding efforts aimed at enhancing yield and other agronomic traits. Relatively new to North Dakota, breeding efforts began at North Dakota State University (NDSU) in 1986, where yield gains and genetic diversity were previously not well-characterized. This research investigates yield improvements and genetic diversity within NDSU cultivars. Era trial data analysis of 28 cultivars reveals incremental yield gains amidst considerable variability. Exploration into ancestral pedigree records of 29 released NDSU cultivars identified 49 founders with genetic contributions. Coefficient of parentage estimates revealed only 10 founders collectively contribute over 70% of all NDSU germplasm. Utilizing SNP-based analyses, intricate relationships among cultivars and founders are outlined, offering useful insights for informed breeding strategies. This study underscores the intricate nature of yield advancements and genetic diversity within the NDSU soybean breeding program, accentuating the importance of genetic diversity in plant breeding populations.

ACKNOWLEDGMENTS

My sincere gratitude goes to my major advisor, Dr. Carrie Miranda, who provided me nothing short of a wonderful opportunity of pursuing a Master's Degree of Plant Sciences at North Dakota State University. I will forever be grateful to her for the guidance, support, and immeasurable knowledge she provided me throughout the course of my research. Her knowledge and passion towards plant breeding and research is contagious and supplied me with invaluable experiences and connections within the soybean and plant breeding communities.

I would also like to extend my thanks and appreciation to Dr. Nonoy Bandillo and Dr. Paulo Flores, members of my thesis supervisory committee. I am grateful to them for improving my research approach and overall thesis through the offering of their time, suggestions, encouragement, and wealth of knowledge.

It is my pleasure to extend my sincere gratitude and appreciation towards Dr. Gustavo Kreutz and Ben Harms for their continuous support throughout my research activities. Their assistances throughout plot planning, field seasons, data interpretation, and research was extremely valuable for completing my research.

I must also extend my gratitude to all other current and former lab members: Dr. Wisdom Edzesi, Clara Mvuta, Cole Williams, Pete Gregoire, Bayan Shukr, Jennifer Obirih-Opareh, Gilda MejÃa, Ashley Cooper, Vicki Magnusson, Aaron Froemke, Ross Lockhart, Evan Omerza, Jacob Thiong, and all undergraduate student workers. I also thank former soybean breeder, Dr. Ted Helms, for creation of the cultivars analyzed throughout this research.

Thank you to the funding support from the North Dakota Soybean Council, North Central Soybean Research Program, and the United Soybean Board. Without their funding, this research

would not be possible. With this, I am grateful to North Dakota State University and the Department of Plant Sciences for the opportunity to pursue a higher education.

Furthermore, I would like to express my gratitude to my parents, Jason and Leah, and siblings, Morgan, Cole, and Mason, for their continuous support, guidance, love, and encouragement. Lastly, thank you to my family and friends who have who have been encouraging and supportive of me throughout every step.

DEDICATION

This work is dedicated to Dr. Carrie Miranda. Her persistence, guidance, and passion encouraged me to apply for graduate school. Since then, I have found a deeper appreciation towards plant breeding and research. I could not have asked for a better advisor and mentor to spend over two years working under. I am forever grateful for the guidance and support you provided me, ultimately setting me on a path to pursue a future career I know I will enjoy. It was my greatest pleasure and honor to work for you. I would also like to dedicate this work to my parents, Jason and Leah, for all you have done for me to pursue my dreams with constant love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ac.....Acre.

AMS.....Ammonium Sulfate.

ANOVA. .....Analysis of Variance.

bu.....Bushels.

C.....Celsius.

CO_2.....Carbon Dioxide.

CP.....Coefficient of Parentage.

CV.....Coefficient of Variance.

F .....Fahrenheit.

Gal.....Gallons.

GPA.....Gallons per Acre.

GT .....Glyphosate-tolerant.

IDC.....Iron Deficiency Chlorosis.

K.....Genetic Clusters.

LD-KNNi .....Linkage Disequilibrium K-Nearest Neighbors imputation.

M.....Million.

MAF.....Minor Allele Frequency.

MG .....Maturity Group.

ND.....North Dakota.

NDSU.....North Dakota State University.

oz.....Ounce.

PCA.....Principal Component Analysis.

pH.....Potential Hydrogen.

PI.....Plant Introduction

PM.....Physiological Maturity.

pt .....Pint.

PVP .....Plant Variety Protection.

RCBD.....Randomized Complete Block Design.

REI .....Re-entry Interval.

RM .....Relative Maturity.

SCN.....Soybean Cyst Nematode.

SNP .....Single Nucleotide Polymorphism.

SP .....Subpopulation.

TASSEL.....Trait Analysis by Association, Evolution, and Linkage.

US .....Unites States.

USDA.....United States Department of Agriculture.

USDA-AMS.....United States Department of Agriculture Agricultural Marketing Service.

USDA-ARS.....United States Department of Agriculture Agricultural Research Service.

VCF.....Variant Call Format.

LIST OF SYMBOLS

° .....Degrees.

% .....Percentage.

® .....Registered Trademark.

LIST OF APPENDIX FIGURES

1. LITERATURE REVIEW

1.1. Soybean Origin and Domestication

Soybean [*Glycine max* (L.) Merr] is a globally grown leguminous crop used primarily for protein meal and oil content for both human and animal consumption, as well as recent upsurges in demand for use in industrial products (Hartman et al., 2011; Wilson, 2008). Despite being native to China, Japan, Korea, and Russia, it is commonly believed that domestication occurred in China (Li et al., 2010) approximately 3,000 to 5,000 years ago from wild soybean (*Glycine soja* Seib. et Zucc.) (Carter Jr et al., 2004; Hyten et al., 2006). Centers of origin are geographic regions where any particular species first evolved. Several studies have investigated the levels of genetic diversity within regions native to soybean. Supporting this, a study analyzing accessions from China, Japan, and Korea found the mean genetic distance, a measure used for quantifying the degree of genetic variation within a population, to be highest among Chinese accessions, while Japanese and Korean accessions were more closely linked (Li & Nelson, 2001). Similar results were found by Bandillo et al. (2015), showing that Japanese and Korean accessions were closely related to another but diverged from Chinese accessions. This suggests the possibility of ancient soybean transport from China to Korea (Bandillo et al., 2015).

Domestication occurs when humans perform selection on a wild species. When selection occurs for extended periods of time, wild species transition into cultivated species, genetic bottlenecks are established, and genetic diversity is diminished (Hyten et al., 2006). Domestication of G. soja into G. max was the first of three genetic bottlenecks soybean has endured (Song et al., 2015). Several regions of domestication within China have been suggested, yet none have been confirmed. These regions include the Yellow River Valley of central China (Dong et al., 2004; Li et al., 2008), the Manchurian region of northeastern China (Deasy, 1939),

and southern China (Guo et al., 2010). Of these regions, it is commonly believed that the Yellow River Valley is most likely the center of origin for soybean due to having higher amounts of genetic diversity than the other regions (Li et al., 2008). Wang et al. (2016) observed many Chinese regions in which levels of genetic diversity increased as distance from the respective centers of origin decreased (Wang et al., 2016). Their study found the Wei and Hanjiang rivers, both located in the Yellow River Valley, displayed the most genetic diversity (Wang et al., 2016). Although these findings support the hypotheses of the Yellow River Valley being soybean's center of origin, an exact center of origin for soybean has not yet been distinctly determined.

1.2. Soybean in the United States

Soybean was first introduced to the United States (US) from China in 1765 (Hymowitz & Harlan, 1983). Soybean was originally grown primarily as hay and forage for livestock. Between the original introduction in 1765 and 1898, Europe and Asia both reintroduced soybean to the US (Tavaud-Pirra et al., 2009). Production of soybean in the US was in high demand as World War II began (September, 1939), as fear of other oil and feed sources would be diminished throughout the war (Fornari, 1979).

Since 1940, soybean production in the US has increased from 4,807,000 to 82,356,000 harvested acres (ac) in 2023 (USDA, 2024). During this same time, average US soybean yields have increased from 16.2 to 50.6 bushels per acre (bu/ac) (USDA, 2024). This 83-year span resulted in a 1,613% and 212% increase in US harvested acres and yield, respectively. In 2023, Illinois, Iowa, Minnesota, Indiana, and Ohio led the US in soybean production, producing 648.9 million (M), 573M, 349.4M, 334.3M, and 274.3M bu, respectively (USDA, 2024). North Dakota (ND) was the ninth highest ranked state, totaling 218.7M bu (USDA, 2024). Soybean production

thrives in these states due to advantageous climates, favorable soil fertility, and robust infrastructure. Soybean production increases within the US can also be attributed in part to breeding efforts focused heavily on developing high-yielding cultivars that are well-adapted to localized environmental conditions. However, as global population increases, so does the demand for soybean and other crops. In a study analyzing maize, rice, wheat, and soybean yields to determine if yield increases would suffice estimated global needs by 2050, it was determined that soybeans' 1.3% per year yield increase would not be enough (Ray et al., 2013). Thus, it is essential that soybean research and breeding methods are optimized for the improvement of the crop and the global needs that rely upon it.

1.3. Genetic Diversity

Ensuring vast ranges of genetic diversity is essential for the success of plant breeding programs, as it directly impacts potential genetic gains through selective breeding efforts. Genetic diversity is a serious current concern for soybean's future, considering diversity losses have been, and are an unavoidable aspect of elite-by-elite breeding. This is due to breeders often selecting elite cultivars as progenitors for most breeding combinations, which expedites the creation of agronomically-improved lines, at the cost of narrowing the gene pool (Viana et al., 2022). As previously stated, domestication was the first event soybean underwent to initiate losses in genetic diversity. Hyten et al. (2006) reported the domestication bottleneck resulted in 50% diversity reductions.

The second diversity reduction soybean underwent is explained through the few landraces introduced to North America. Introduction and reintroduction of soybean to the US resulted in only 80 ancestors accounting for 99% of all parentage of 258 cultivars released between 1947 and 1988 (Gizlice et al., 1994). In their study, ancestors were defined as any

founding stock with no known pedigree. However, the term “ancestor” is commonly used to describe any previous relative in which a cultivar has descended from. To eliminate confusion, the term “founder” is used here to describe any germplasm that was brought over to serve as founding stock in the initiation of US soybean production, as not every ancestor can be considered a founder. These founders serve as the initial stock that is responsible for all North American cultivars. Furthermore, 95% of the current North American genetic base was complete by 1970 (Carter Jr et al., 2004). It has also been estimated that the United States Department of Agriculture Agricultural Research Service (USDA-ARS) germplasm collection (19,625 accessions at the time) had an effective population size represented by only 106 accessions (Xavier et al., 2018). Several large-scale diversity analyses for soybean genetic bases have been completed for several countries, including but not limited to Brazil (Wysmierski & Vello, 2013), China (Cui et al., 2000; Li et al., 2008), Japan (Zhou et al., 2000), and the US (Delannay et al., 1983; Gizlice et al., 1994).

These relatively few introduced founders provided the founding genetic stock for US soybean until cultivars resulting from plant breeding were released in the 1940s (Specht et al., 2014). Since then, the practice of using elite cultivars as parental material derived from the few founders has proven to be productive in North American soybean history (Bruce et al., 2019). However, convenience of constantly utilizing limited quantities of superior germplasm as parental stock for breeding purposes has contributed to the reduction of the present genetic base and is the third genetic bottleneck soybean has undergone in the US.

1.4. Maturity Groups and Relative Maturity

Garner and Allard (1920), Borthwick and Parker (1938), and Kantolic and Slafer (2001) are among many researchers who have studied the effects of photoperiod and temperature on

various growth and development processes within soybean (Borthwick & Parker, 1938; Garner & Allard, 1920; Kantolic & Slafer, 2001). In soybean cultivation, understanding these effects is crucial, as soybeans are known to be photoperiod-sensitive, meaning reproductive development is heavily influenced by day length. Additionally, overnight temperatures play significant roles in shaping reproductive development.

In soybean, maturity groups (MG) refer to areas of suggested adaptation based on the latitude and climatic factors affecting the variation in days requires to reach physiological maturity (PM) for a geographic region (Zhang et al., 2007). Across North America, 13 different MG classifications are present, spanning from 000 to X, where 000 is the northernmost MG, and X is southernmost (Bandillo et al., 2015; Zhang et al., 2007). Soybean production in the US spans across MG 00 to VIII. Within a single MG, the date of reaching PM can vary from 10 to 18 days (Scott & Aldrich, 1970). To increase accuracy of a MG label, MG are further subdivided into categories denoted by decimal values (.0 to .9) known as relative maturity (RM). Relative maturity is a metric referring to the number of days it takes for a particular variety to reach PM under certain growing conditions. Selection of cultivars with a RM suited for planting is crucial to maximize yields, as it ensures they are well-adapted to the specific growing environment. Positioned in the northernmost region of the US, ND is limited to growing early MG 00, 0, and I (Berglund, 2002) and more specifically RM 00.6 – 1.3. This is one of the distinct challenges of producing soybean in ND.

1.5. Soybean in North Dakota

Soybean in ND began in 1942 where 4,000 harvested acres initiated the state's production history, yielding an average of 10 bu/ac that year (USDA, 2024). In ND, 1997 marked the first year that soybean exceeded one million harvested acres (29.5 bu/ac average),

while harvested acres in 2023 exceeded six million in total, with an average yield of 35.5 bu/ac (USDA, 2024). Based on these yield data and duration of production, the average rate of yield gain for ND can be determined as an average of 0.31 bu/ac per year.

These data for ND yield improvements are comparable to that of Rincker et al. (2014) and Specht et al. (2014). Rinker and collaborators evaluated yield and other agronomic traits for 168 cultivars across MG II – IV, and reported an average annualized yield improvement of 0.32 bu/ac per year (Rincker et al., 2014). Specht and collaborators performed a similar study analyzing on-farm soybean yield improvements across an 80-year period, and reported an average annualized yield improvement of 0.35 bu/ac per year (Specht et al., 2014).

However, when comparing statewide yield averages from 2019 to 2023, ND averaged only 32.3 bu/ac, while states like Illinois, Indiana, and Iowa averaged 61, 57.7, and 57.7 bu/ac, respectively (USDA, 2024). Despite ND sharing similar yield gains per year to MG II – IV studies, ND still falls short on soybean yield in comparison to other states. This potentially could be explained by ND's average of 10 bu/ac starting point, which created a low enough base level to see vast improvements over the duration of increased production and improvement. A second hypothesis for lower yields could possibly be due to the lack of crop management length within ND compared to other states, as soybean production initiated in other states before becoming suitable to ND environments. A third possibility is the restriction of available MG 00 and 0 germplasm, as availability of MG 00 and 0 germplasm is more limited (relative to other MG germplasm) due to less land area coverage of production. Along with these three possibilities, state-to-state yield discrepancies are undoubtedly amplified by the wide range of environmental conditions that exists in the midwestern US.

1.6. Environmental Conditions in North Dakota

Environmental stresses are among the main factors that negatively affect the yield of adapted cultivars. This stress occurs when one or more environmental factors are deficient or abundant to an extent that reduces crop quality and yield. Environmental stresses can be either abiotic (air and soil temperature, water availability, salinity, potential hydrogen (pH), ultraviolet radiation, heavy metals) or biotic (weeds, insects, nematodes, bacteria, viruses, fungi).

Environmental conditions across ND vary significantly. There are large temperature and precipitation variances, both seasonally and daily. Over a 30-year period from 1980 to 2010, annual precipitation ranged from 13 to 20 inches per year, increasing from west to east (North Dakota Game and Fish, 2019). During this same time period, annual temperatures averaged around 40°F (4.4°C) statewide, with cooler temperatures in the north compared to the south (North Dakota Game and Fish, 2019). Soybean can withstand cold temperatures that fall as low as 59°F (15°C), but significant yield reductions may occur when temperatures fall to 50°F (10°C) (Board & Kahlon, 2011). While temperatures during a ND growing season do not typically fall below this mark, late frost (after emergence) and early frost (before harvest and/or PM) can occur, resulting in yield losses. Freezing temperatures during the reproductive growing stages R1-R5 may result in yield losses up to 70%, while freezing at R6 may result in 25% losses (Board & Kahlon, 2011).

Certain regions in ND face specific challenges. Acidic soils are prevalent in south-central and southwestern ND (Seelig, 2000). In addition, saline seeps affect an estimated 100,000 acres of western ND farmland (Doering & Sandoval, 1976). Saline seeps pose serious problems to yield and production, as they produce provide high levels of salt concentration, increased soil salinity, and land deterioration (Seelig, 2000). This combination of low rainfall, warmer

temperatures, low pH, and high-salinity soils within western ND contribute to a drought-like environment, dragging down the statewide yield average. The majority of North Dakota's soybean production stems from central and eastern regions closer to and within the Red River Valley, where climatic conditions all favor soybean production.

Approximately 90% of total dry weight of crops is a result of carbon dioxide (CO_2) assimilated through photosynthesis (Zelitch, 1982). Because photosynthesis and canopy light interception both play crucial roles in determining crop growth and yield, environments with larger quantities of sunlight should have more optimized photosynthesis, and yield in return (Fageria et al., 2006). Given the latitude of ND, longer durations of sunlight and optimized photosynthesis should be expected when compared to lower latitude regions. However, longer durations of sunlight do not always correlate to warmer temperatures. Therefore, lower statewide yields could be partly attributed to drought-like conditions and temperature-related stresses, such as early and late frosts.

1.7. North Dakota State University Soybean Breeding Program

Soybean breeding at North Dakota State University (NDSU) first started in 1986 under professor Dr. Theodore Helms (Myrdal, 2022), who was the breeder for 34 years until retiring in 2020 (North Dakota Soybean Council, 2020). Helms released his first cultivar "Council" in 1994, and since then NDSU has released 40 soybean cultivars in total (Myrdal, 2022). Currently, the NDSU breeding program is focused on cultivar development using superior germplasm. Program focuses include breeding for phytophthora, soybean cyst nematode (SCN), and iron-deficiency chlorosis (IDC) resistances, as well as drought tolerance specialized for western ND growers, and improved yields. In addition to producing both conventional and glyphosate-tolerant (GT) commodity soybean, the program also works in developing both high protein (tofu)

and small-seeded (natto) food grade soybeans, and recently implemented high-oleic breeding efforts.

1.8. Research Objectives

The introduction of this research began in 2022 after analyzing field data from the 2021 growing season. Albeit a year where ND experienced drought conditions, the 2021 yield data for certain NDSU cultivars was much lower than expected. The previous analyses and comparisons of ND yield data to the Rinker et al. (2014) and Specht et al. (2014) studies sparked the desire to conduct research in similar practices using NDSU cultivars. Understanding how yields have changed among NDSU cultivars since the program's inception can provide useful information about the program's past and also future development.

The main purpose of this study was to determine how yields for released NDSU cultivars have changed since the first variety was released. However, with the importance of genetic diversity gaining popularity among public breeding programs, collection and analysis of that data would serve as supplemental data. A multitude of factors, both genetic and environmental, influence a crop's yield. Since a breeder has little to no control over most environmental conditions, focusing on improving genetics is important. In this research, the genetic aspect to be evaluated is diversity.

The secondary analysis of this study was to determine the amount of genetic diversity among released NDSU cultivars. The first objective was to utilize pedigree records to determine the parentage of NDSU cultivars dating back to the founders to allow visualization of a complete pedigree for the breeding program. The second objective was to quantify the genetic base of the released NDSU cultivars by using the pedigree relationship to determine coefficient of parentage (CP). The third objective was to utilize genotype data to display genetic relatedness and assess

genetic diversity among released NDSU cultivars and between released NDSU cultivars and founders.

Both the yield gains and genetic diversity of the released cultivars were previously not well characterized. It will be essential for the NDSU soybean breeding program to assess the trends of yield data and genetic diversity among previously released cultivars in its ambition to improve yields during future cultivar development.

2. MATERIALS AND METHODS

2.1. Germplasm Selection

Cultivars from the NDSU soybean breeding program were chosen based on available germplasm on campus. Of the 40 released NDSU cultivars, 11 were excluded due to lack of available germplasm. The cultivars analyzed in this study were released between 1994 and 2021 (Table 1). Within the 29 cultivars, 17 are conventional commodity, six are glyphosate-tolerant (GT) commodity, three are specialty food grade tofu, and three are specialty food grade natto. In this analysis of genetic diversity, no distinction between specialty and commodity soybean was made. Although tofu and natto cultivars tend to yield less than commodity cultivars, all yield were taken at face value.

Table 1. Cultivar Entries, Assigned PI number, Year of Release, and Relative Maturity Rating.

Studied NDSU Cultivars			
Cultivar	PI Number	Release Year	RM
Council	PI 587091	1994	0.6
Traill	PI 596541	1997	0.0
Norpro	PI 603900	1998	0.4
Jim	PI 602897	1998	00.7
Barnes	PI 614831	2000	0.3
Walsh	PI 615586	2001	0.3
Sargent	PI 615585	2001	0
Nornatto	PI 631437	2002	0.4
RG200RR	PI 632259	2002	0.0
Nannonatto	PI 631438	2002	0.4
LaMoure	PI 634813	2003	0.7
ProSoy	PI 638511	2005	0.8
Pembina	PI 638510	2005	00.6
RG607RR	PI 645465	2006	0.7
Sheyenne	PI 647867	2007	0.8
Cavalier	PI 654358	2008	00.7
Ashtabula	PI 655938	2009	0.4
ND1100S	PI 664265	2011	00.9
ND1406HP	PI 673929	2014	0.6
ND Henson	PI 675334	2015	0.0
ND17009GT	PI 686350	2017	00.9
ND Bison	PI 680568	2017	0.7
ND Benson	PI 686348	2018	0.4
ND Stutsman	PI 686349	2018	0.7
ND18008GT	PI 689514	2018	00.8
ND Rolette	PI 699922	2019	00.9
ND Dickey	PI 701371	2020	0.7
ND2108GT73	PI 699334	2021	0.8
ND21008GT20	PI 699333	2021	00.8

2.2. Locations of Yield Analysis

The field research consists of data from several locations throughout eastern ND over two growing seasons. The first year (2022) consisted of two locations: Casselton and Grandin, ND.

Due to planting and harvesting errors, data collected for Grandin's first growing season was unreliable, so it was excluded from the analysis. The second year (2023) consisted of six locations: Casselton, Grandin, Hatton, LaMoure, Lisbon, and Milnor, ND. Planting dates for each location were as follows: Casselton (5/22/22 & 5/26/23), Grandin (5/19/22 & 5/22/23), Hatton (5/25/23), LaMoure (5/23/23), Lisbon (5/31/23), and Milnor (5/29/23). Final locations of data included Casselton (2022), Grandin (2023), Hatton (2023), LaMoure (2023), and Lisbon (2023). Large amounts of plant injury from SCN and IDC were present within Casselton's experiment. Herbicide carryover was present within Milnor's, causing over half the experiment to become non-salvageable. Due to these issues, data from both Casselton and Milnor for the second growing season were removed from analyses. The remaining five locations were analyzed as environments.

2.3. Field Experiment

In each location, plots were planted in two 14.93-foot rows (measure given with appropriate driver and driven gears) with 30-inch row spacing and three replications of each cultivar in a randomized complete block design (RCBD). A total of 32 cultivars were used in the experiment, 29 of which were NDSU cultivars, while two were Minnesota releases and one South Dakota release. All plots for both years were planted with an Almaco 4-row SeedPro planter with an integrated Skytrip system (Almaco, Nevada, IA, USA).

Across all locations, a pre-emergent herbicide tank mix was sprayed following planting, targeted towards both grass and broadleaf weeds. For five acres of spraying, the tank-mix included: Cornerstone® 5 Plus (30 oz ac^{-1}), Valor® EZ (3 oz ac^{-1}), Dual Magnum (1.5 pt ac^{-1}), Destiny® HC adjuvant (1.5 pt ac^{-1}), InterLock® adjuvant (6 oz/15 gal ac^{-1}), and Ammonium Sulfate (AMS; 12lb/100 gal $water^{-1}$). A rate of 15 gallons-per-acre (GPA) was used, and a re-

entry interval (REI) of 24 hours was followed. All spraying was done with a 3-point John Deere® sprayer with a 30-foot boom. Throughout the growing season, the experiments did not receive any herbicide or fungicide applications. In Casselton, Sefina® (3 oz ac^{-1}) was applied via crop duster to control aphids. Manual weed control was used within all experiments. Flowering and maturity notes were also taken throughout the season. For all locations in both years, the number of days to flowering and flower color were recorded for every plot when 50% or more of the cultivars' flowers were present. In all experiments, maturity was recorded as the number of days after August 31st that 95% of the pods had reached mature pod color.

Plots in the first year were harvested using an Almaco SP20 2-row combine (Almaco, Nevada, IA, USA). In the second year, plots were harvested using a Zürn 150 2-row plot combine (Zürn Harvesting GmbH & Co. KG, Kapellenstraβe, Schöntal-Westernhausen, Germany). Weight and moisture from each replication were recorded by the combine. Before calculating yield, weight measurements from the combine were converted into bushels and moisture was standardized at 13%. With this data, yield (bu/ac) was calculated.

Yield data was graphed in Excel, using yield as the y-axis component and year of release as the x-axis component. For each yield graph, a linear trendline was added along with respective R^2 values. Analysis of yield data excludes all three non-NDSU cultivars, as well as Norpro, in which seed contamination throughout environments provided inaccurate yields. This reduced the number of NDSU cultivars with yield data from 29 to 28.

2.4. Pedigrees of Studied Germplasm

The initial source of pedigree information for NDSU cultivar parental data was Plant Variety Protection (PVP) records gathered from the United States Department of Agriculture Agricultural Marketing Service (USDA-AMS) (https://www.ams.usda.gov/). With this

information, the majority of the remaining parental pedigree data collected was done through SoyBase's Pedigrees Database (Brown et al., 2021). Parental data was recorded for each ancestral accession in each generation until an introductory founder was reached. Any additional pedigree data missing from SoyBase (https://soybase.org) was collected again through PVP records (USDA-AMS) and the Uniform Soybean Tests for the Northern Region, serviced by the USDA-ARS (https://www.ars.usda.gov/). Pedigree information was then formatted and visualized using GraphvizOnline (https://dreampuf.github.io/GraphvizOnline/) allowing historical connections to be made for ancestral data.

2.5. Coefficient of Parentage

First, compiled pedigree information traced NDSU experimental and released lines back to the founders. Coefficient of parentage, or additive relationship, was calculated using the pedigree information and the R package "AGHmatrix" (Amadeu et al., 2023). This matrix was subset into four matrices: founders and NDSU released cultivars, founders and NDSU experimental lines, first progeny and NDSU released cultivars, and first progeny and NDSU experimental lines. First progeny describes any cultivar with a founder as an immediate parent. NDSU experimental lines are any lines created at NDSU, but were not released as a cultivar. Using the additive relationship matrix, average contribution was calculated by averaging the additive relationship of each founder to each released cultivar and experimental line, and the additive relationship of all first progeny to each released cultivar and experimental line.

2.6. Sequencing Data

Whole genome sequencing was completed for NDSU cultivars in 2020. At the time of sequencing, ND2108GT73 and ND21008GT20 had not yet been released and sequencing for these two cultivars had not been done. This reduced the number of NDSU cultivars with

sequencing data from 29 to 27. Plant tissue was grown and DNA extracted by the Bilyeu lab at the University of Missouri. Samples were sent to GENEWIZ (Azenta Life Sciences, South Plainfield, NJ, USA) for short-read whole genome sequencing at 15x coverage. Read mapping to Wm82.a2.v1 and variant calls were also completed by GENEWIZ.

The whole genome re-sequence dataset of 27 ND soybean cultivars was subsetted to only contain Single Nucleotide Polymorphism (SNP) positions overlapping with Illumina Infinium SoySNP50K iSelect Beadchip (SoySNP50K) in Wm82.a2.v1 coordinate system (Song et al., 2015). The resulting Variant Call Format (VCF) file was annotated with SNP names using the "annotate" function in bcftools (Danecek et al., 2021). The SoySNP50K genotypic data for ND founder accessions was downloaded from SoyBase (Grant et al., 2010) and merged with the subsetted ND soybean cultivars dataset using the merge function in bcftools (Danecek et al., 2021). Only 42 of the 49 program founders had 50K data available on SoyBase. The resulting dataset was converted to HapMap format using Trait Analysis by Association, Evolution, and Linkage (TASSEL) (Bradbury et al., 2007) and contained 42,195 SNP positions (Bradbury et al., 2007). Three subsets of the merged 50K data were used for analysis. The first included all NDSU cultivars and available founders 50K data with the 42,195 positions.

Three subsets of the merged 50K data were used for the analysis. The first included the 27 NDSU cultivars and available founders 50K data with the 42,195 positions. The 50K data from the founders had an average of less than 1% missing data, while the NDSU cultivars averaged 24.5% missing data. Due to this disparity, Linkage Disequilibrium K-Nearest Neighbors imputation (LD-KNNi) was conducted in TASSEL (Money et al., 2015). This set was filtered in TASSEL for minor allele frequency (MAF) < 0.1 and heterozygosity > 0.2, which resulted in 29,378 SNP positions. The second subset included all NDSU cultivars and nine of the

top 10 contributing founders by coefficient of parentage. Using the same filtering parameters, this dataset was filtered to 23,357 SNPs. The final subset included only NDSU cultivars, and was filtered to 19,320 SNPs. All three subsets were exported from TASSEL in VCF format for further analyses.

2.7. SNP-based Dendrogram

The first subset of merged data (27 NDSU cultivars and founders) containing 29,378 positions was used to construct a dendrogram in R software version 4.3.2 (R Core Team, 2023). The "snpgdsVCF2GDS" function in the "SNPRelate" package (Zheng et al., 2012) was used to convert the filtered VCF file to a gds object. Using the same R package, dissimilarities for each pair of individuals were calculated using the "snpgdsDiss" function and a hierarchical cluster analysis was performed on the dissimilarity matrix using the "snpgdsHCluster" function. The SNP-based dendrogram was created using the R package "ggtree" (Yu, 2022) to allow the visual assessment of relatedness among individuals.

2.8. Heatmap

The heatmap was created using R software version 4.3.2 (R Core Team, 2023), with the second dataset (27 NDSU cultivars and nine of the top 10 program founders) that contained 23,357 SNPs. The R package "AGHmatrix" (Amadeu et al., 2023) was used to calculate the genomic relationship matrix between all pairs of individuals, and the package "heatmap3" (Zhao et al., 2021) was used to create the heatmap that visualized the matrix values.

2.9. Population Structure

Population structure figures were created to visualize only the NDSU cultivars (third subset), as well as the NDSU cultivars and nine of the top 10 founders (second subset) using R software version 4.3.2 (R Core Team, 2023). For each set, a principal component analysis (PCA)

was performed and the percentage of variance explained by each PCA component was computed using the "pca" function in the R package "LEA" (Frichot & François, 2015). The appropriate number of genetic clusters (*K*) to be used in the population structure analysis was visually determined by generating a screeplot of the scores of each PCA component. Then, ancestral admixture coefficients were estimated using "LEA" and visualized using the "barplot" function in base R. This was done for both sets of individuals at three ($K = 3$) and five ($K = 5$) genetic clusters.

3. RESULTS AND DISCUSSION

3.1. Era Trial Yield Data

For 29 NDSU cultivars (Table 1), yield data were collected as environments in 2022 and 2023.These were grown in five locations: Casselton, Grandin, Hatton, LaMoure, and Lisbon, ND. Data were analyzed by ANOVA for coefficient of variance (CV) to determine usefulness (data not shown). Data at a single location was considered useful if CV was below 30%, no locations were excluded. Means from multilocation ANOVA were plotted based on cultivar year of release (Figure 1). A linear best-fit trend line was added to the data to visualize yield gains across cultivar release years. The best-fit trend line statistic value (R^2) was calculated as 0.1161.

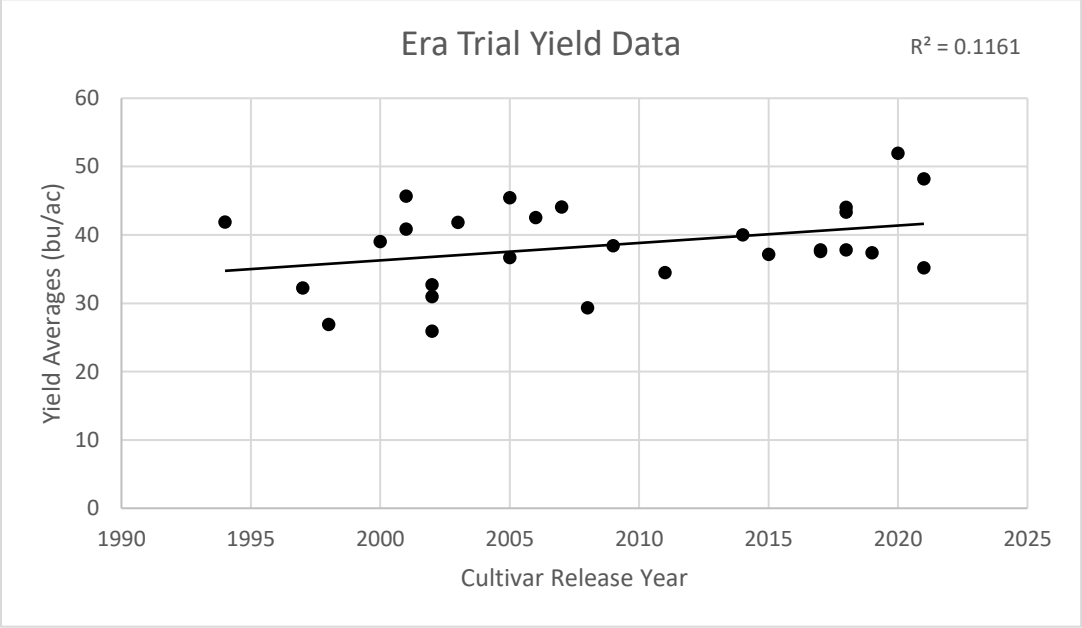

Figure 1. Era Trial Yield Data.

Scatter plot of cultivar release year vs cultivar yields. Yield data was averaged for each cultivar replicate per environment across both 2022 and 2023 growing season. Each cultivar data point is the average yield of three replicates across five environments. A linear best-fit trend line was added to the data, and the best-fit trend line statistic (R^2) value is displayed in the top right.

Overall, the best-fit trend line shows that yield gains among released cultivars are increasing gradually. Variation among released cultivar yields is responsible for these gradual

gains. However, the low R^2 value suggests the trend line does not have high accuracy. As years progressed, cultivar yield often averaged less than the previously released cultivars. An example of this is observed among the first and second released cultivars in 1994 and 1997, respectively, where yield averages dropped over 10 bu/ac. Trends similar to this example occur throughout the dataset. However, one reason minor contribution to this variation is due to the analysis including six specialty food grade cultivars, where food grade quality is prioritized, and lower yields are expected. When food grade cultivars were removed, the R^2 value increases slightly (data not shown). Moving forward, the NDSU soybean breeding program will continue this experiment to collect more yield data, incorporating any newly released future cultivars. Data for Norpro was not removed from cultivar entry list (Table 1), but was removed from yield analysis (Figure 1).

3.2. Pedigree

Pedigrees for the 29 released NDSU cultivars dating back to the founders were determined by utilizing ancestral pedigree records (Figure 2). In constructing the parental pedigree for NDSU released cultivars, lineage traced back to the founders encompassed a total of 443 pedigree-related accessions. Of the 443 total accessions, 40 are NDSU cultivar releases (29 highlighted) and 49 are founders. For seven of the 443 accessions, either partial or no parental data was found. In these cases, parental data was either not recorded, lost, or unknown. A total of 19 generations are present within the pedigree. Overall, the shallow distance between founders and NDSU cultivars likely stems from increased levels of inbreeding, a commonality within all North American soybean, and limited time to increase generations since NDSU soybean breeding efforts initiated in 1986. Repetitive use of released cultivars as parental material for new stock is observed throughout NDSU cultivars, suggesting lack of access to MG 00 and 0 breeding materials. Large rates of yield gains among the studied NDSU cultivars should not be

anticipated given they are highly genetically similar. To improve genetic diversity within the NDSU program, new germplasm has been and will continue to be introduced to target specific agronomic traits. New germplasm has been introduced through collaboration with other breeding programs and use of germplasm collections. The USDA Soybean Germplasm Collection serves as a vast repository of alleles for soybean breeding programs worldwide. It contains a wide range of *G. soja* accessions from native countries (China, Korea, Japan, and Russia) and *G. max* accessions from 87 countries (Song et al., 2015). Yield, herbicide tolerance, drought tolerance, and varying disease resistances are all targeted areas of current breeding efforts to expand diversity while improving productivity.

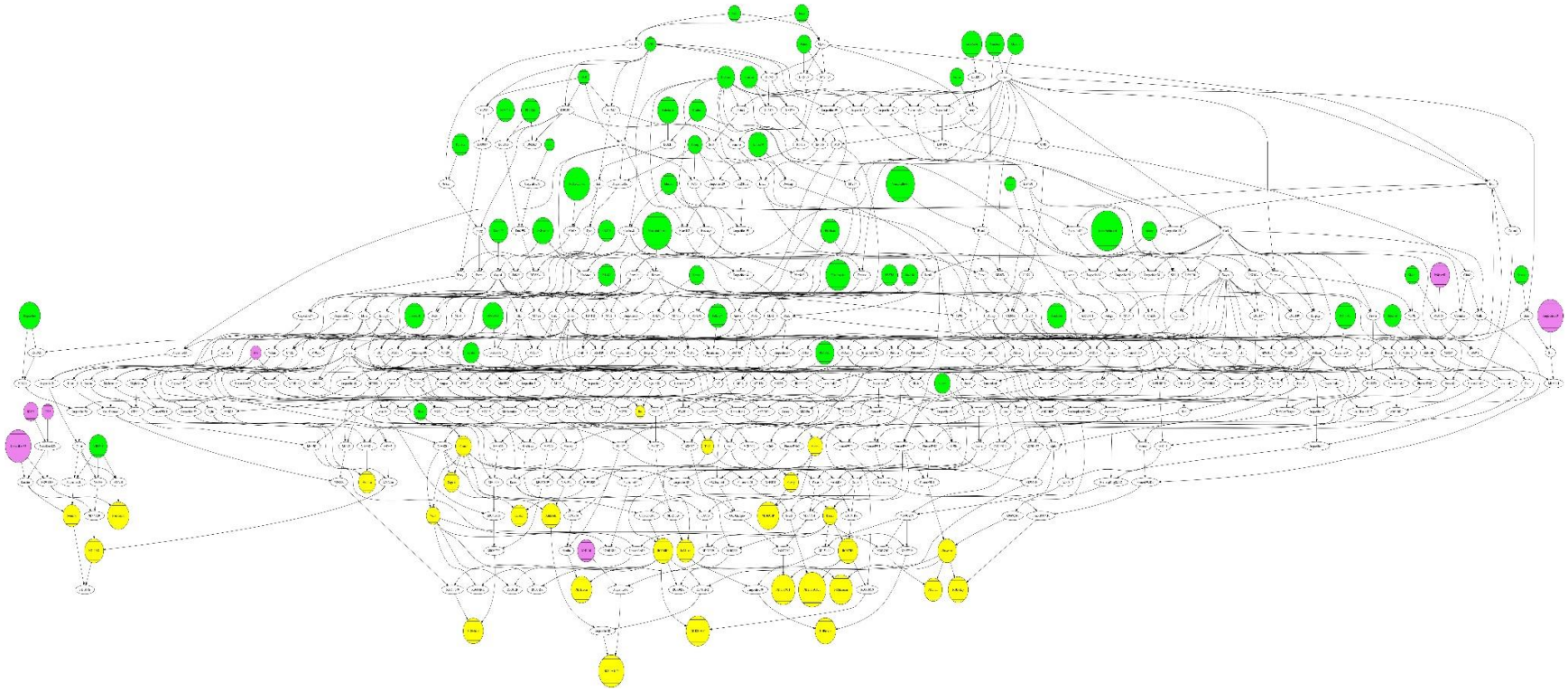

Figure 2. Pedigree Tree.

Pedigree relationships of 29 NDSU cultivars tracing back to historical program founders. Yellow coloring depicts NDSU cultivars, green depicts founders, and purple depicts any cultivar with at least one unknown parent. Aside from the 11 non-colored released NDSU cultivars, all remaining non-colored accessions represent any other ancestor.

3.3. Coefficient of Parentage

Coefficient of parentage (CP) was determined using pedigree information and relationship estimates in the "AGHmatrix" package in R. One kinship value was assigned for each combination of individuals, and coefficient of parentage was calculated as the average of all kinship values for each founder. The top 10 contributors are listed in Table 2. Mandarin (Ottawa) is the top contributing founder at 24.19%, followed by A.K. (Harrow) at 9.88%, and Strain 171 at 6.43%. The cultivar Mandarin (Ottawa) was originally grown primarily in Canada and Minnesota, where it was used successfully in cultivar development (Hymowitz & Bernard, 1991). Mandarin (Ottawa) was one of the first commercial cultivars grown in ND that yielded similarly to others, but provided height without lodging and also produced pods higher off the ground, making harvest easier (Stoa, 1950). This combination provides strong evidence as to why Mandarin (Ottawa) is responsible for nearly a quarter of the NDSU germplasm. Over 70% of all germplasm within the NDSU pedigree traced back to just 10 founders. The remaining 39 founders present within the NDSU pedigree account for remaining germplasm, indicating the dependability of certain cultivars throughout historical breeding efforts.

Table 2. Top 10 Contributing Founders to Released NDSU Cultivars based on Coefficient of Parentage.

Top 10 Contributing Founders	
Founder	Contribution (%)
Mandarin (Ottawa)	24.19
A.K. (Harrow)	9.88
Strain 171	6.43
Mandarin	5.97
Richland	5.86
Manchu	5.30
Mukden	4.56
Strain 18	3.16
Fiskeby III	2.92
201-14-20	2.35
Total	70.62

3.4. SNP-based Dendrogram

The SNP-based dendrogram displays comparisons between 27 NDSU cultivars and 42 program founders (Figure 3). Two NDSU cultivars were released after genotyping was completed (ND2108GT73 and ND21008GT73) and genetic information for seven founders present within the program were not available (Strain 171, Kogane Jiro, Palmetta, 680+993+994, PI 191110-1, PI 95560, and PI 171862). All germplasm included in the SNP-based dendrogram were clustered into two groups, comprising the founders labeled in black and NDSU cultivars labeled in red (Figure 3).

The branch that centrally splits the founders into two groups represents a significant divergence in genetic relatedness. The founders located on the branch excluding NDSU cultivars (Komata No. 79 through PI 88788) suggest a lower genetic similarity to founders on the opposite branch containing founders (Dunfield through Mandarin) and NDSU cultivars (Figure 3). Within the NDSU cultivar branch, the three specialty food grade natto cultivars (Nannonatto, ND1100S,

and Nornatto) formed a cluster. Similarly, the three specialty food grade tofu cultivars (ND1406HP, ProSoy, and Norpro) also formed a cluster. High-yielding commodity cultivars (ND Bison, ND Dickey, Sheyenne, and ND Stutsman) formed another distinct cluster within the NDSU cultivars. A fourth cluster including cultivars Traill, RG200RR, ND18008GT, Jim, and Pembina was also observed. The last distinct cluster contains cultivars Barnes, RG607RR, ND17009GT, ND Henson, and ND Benson. Council, Walsh, Sargent, Cavalier, and Ashtabula group together, but relatedness among these cultivars appears to be relatively low. Low relatedness is also observed among ND Rolette and LaMoure (Figure 3).

Clustering and branching are two forms of assessing relatedness among founders and cultivars within the SNP-based dendrogram. However, vertical distance of each cluster or connection provides information on relatedness as well. Increased distance between two connected founders or cultivars represents lower relatedness than a short vertical distance. Within NDSU cultivars, Traill and RG200RR display the least vertical distance connecting them, implying they share the highest degree of genetic relatedness among any two NDSU cultivars. Sheyenne and ND Stutsman share the second least vertical distance, followed by ND1406HP and ProSoy. Within the founders, minimal vertical distance is observed among multiple founding pairs: Mandarin and Mandarin (Ottawa), A.K. (Harrow) and Illini, Arksoy and Ralsoy, and 201-14-20 and Fiskeby III (Figure 3). This is because Mandarin (Ottawa) is a selection from Mandarin, A.K. (Harrow) and Illini are both selections from A.K., and Ralsoy is a selection from Arksoy (Bernard et al., 1988). In the case of 201-14-20 and Fiskeby III, they are full-sibs (Gizlice et al., 1994).

The branch including founders located closest to the NDSU cultivar branch contains four of the top 10 contributing founders: Mandarin, Mandarin (Ottawa), Strain 18, and Richland. Horizontal analysis of the SNP-based dendrogram provides insight to founders more closely related to NDSU cultivars. However, given their stronger relations to other founders, certain founders with higher contribution percentages are located further away from NDSU cultivars. Examples of this can be seen with A.K. (Harrow), Fiskeby III, and 201-14-20. This is likely due to these founders sharing a higher degree of genetic relatedness to other founders located within their clusters than to NDSU cultivars.

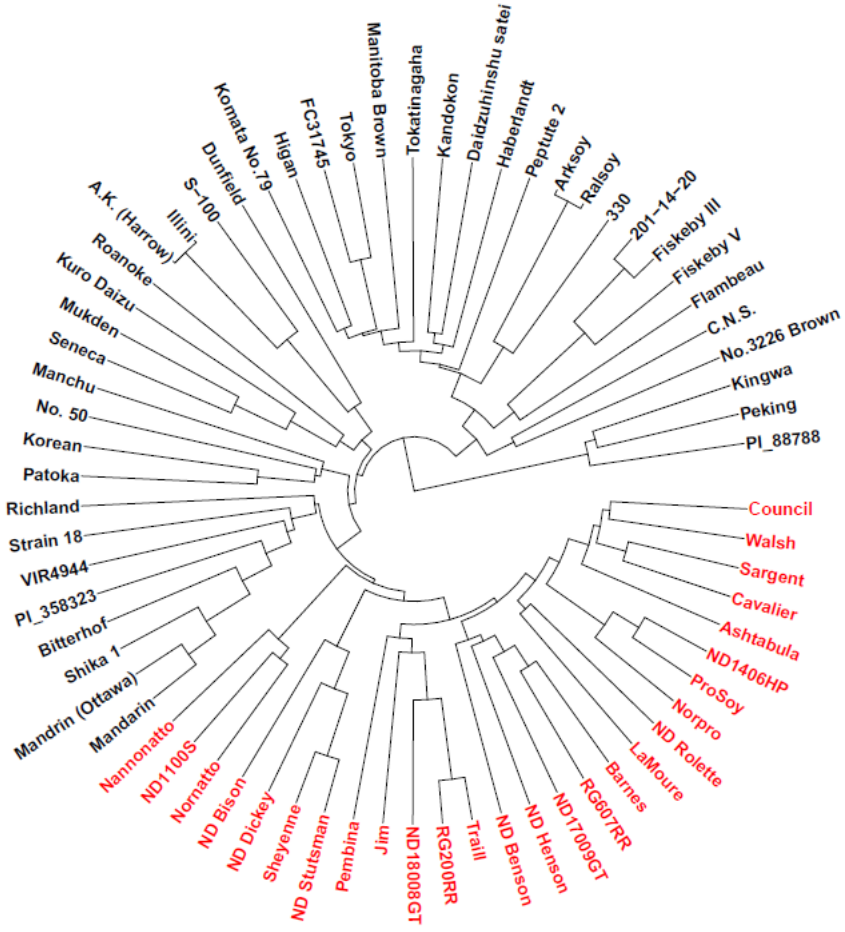

Figure 3. SNP-based Dendrogram.

SNP-based dendrogram displaying 42 of 49 program founders (black), and 27 NDSU cultivars (red).

3.5. Heatmap

Genetic relationships among released NDSU cultivars and top contributing program founders are displayed as a heatmap (Figure 4). The first cluster displays a strong genetic relationship between full-sib founders 201-14-20 and Fiskeby III, exemplified by deep red coloring. The second cluster formed displays increased relatedness among specialty food grade natto (Nannonatto, Nornatto, and ND1100S) cultivars. Hierarchical clustering within the heatmap is congruent with the SNP-based dendrogram analysis. The third cluster includes founders Mandarin and Mandarin (Ottawa), which was originally a selection of Mandarin itself.

Within the hierarchical clustering on the heatmap, the first three clusters discussed all fall on a separate branch than the rest of the founders and NDSU cultivars studied (Figure 4). This difference in clustering location is the main difference compared to the SNP-based dendrogram clustering, where all NDSU cultivars were grouped together. A reason for this difference is that the SNP-based dendrogram analyzes 42 founders while the heatmap only analyzes nine of the top 10 contributing founders. When analyzing the SNP-based dendrogram horizontally, the closest cluster of NDSU cultivars to founders was the specialty food grade natto lines. When analyzing NDSU cultivars and 42 founders, the natto cultivars are more closely related to other NDSU cultivars than the majority of the founders. However, when analyzing NDSU cultivars and only nine founders, the natto cultivars group with select founders due to the decreased sample size.

The fourth cluster observed within the heatmap (Figure 4) is congruent with the second cluster from the SNP-based dendrogram (Figure 3). This cluster includes high-yielding commodity cultivars: Sheyenne, ND Stutsman, ND Dickey, and ND Bison as well, where relatedness is strongest between Sheyenne and ND Stutsman. Similarly, Traill, RG200RR,

ND18008GT, Jim, and Pembina form a fifth cluster, where Traill and RG200RR share the strongest relationship. The sixth cluster includes three founders: Strain 18, Mandarin, and Mandarin (Ottawa). The last major cluster created entails ND Henson, Barnes, ND Benson, RG607RR, and ND17009GT. Among these cultivars, varying degrees of genetic relatedness can be observed, ranging from light red to orange, suggesting common genetic backgrounds. These findings underscore consistent genetic patterns across the dataset, reflecting the breeding goals and previous selection pressures that shaped the program's background.

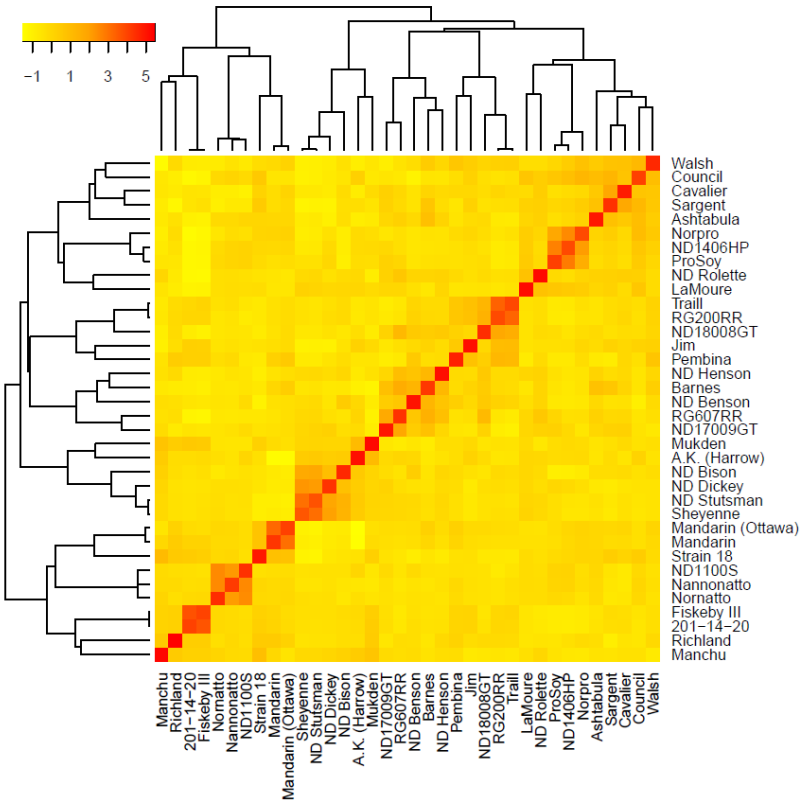

Figure 4. Heatmap of NDSU Cultivars and Nine Founders.

Heatmap for 27 NDSU soybean cultivars and nine of the program's top 10 contributing founders (Strain 171 does not have available genotype data). Each individual square represents the relationship between that cultivar and the respective other. The color gradient depicts levels of relatedness, with darker colors (red) indicating stronger correlations and lighter colors (yellow) suggesting weaker or no correlation. Subpopulations form when multiple cultivars share elevated levels of correlation. Along the top and left sides, dendrograms display the heatmap in a form of hierarchical clustering.

3.6. Population Structure

Population structure figures were created to visualize relationships of genetic background between NDSU cultivars (Figure 5) and between NDSU cultivars and top contributing founders (Figure 6). The appropriate number of genetic clusters ($K = 5$) was determined by generating a screeplot (Figures A1 & A2) of the scores of each PCA component, and analyzing the elbow point of each plot.

Within the subpopulation (SP) results for NDSU cultivars (Figure 5), SP1 best describes cultivars most genetically similar to the cultivar Traill, as Traill is a common parental ancestor. Pembina, RG200RR, Traill, ND18008GT, and Jim all consist of large SP1 backgrounds. This cluster coincides with analyses from both the heatmap and SNP-based dendrogram. Again, SP2 background belongs nearly exclusively to specialty food grade natto cultivars while tofu cultivars contain the largest background percentages within SP3. Council is not a tofu cultivar, but expressed a majority SP3 background. Within Ashtabula, Walsh, Cavalier, and Sargent, SP3 background is derived from Council. Similar to SP1, cultivars most genetically similar to the cultivar Barnes are best described by SP4, as Barnes is a common parental ancestor. This includes RG607RR, Barnes, ND17009GT, ND Benson, and ND Henson. Congruent with other analyses, the high-yielding commodity cultivars are best described by SP5. This includes ND Bison, ND Dickey, Sheyenne, and ND Stutsman, where Sheyenne was used as a parent for each of the other three cultivars. Remaining cultivars expressed combinations of varying SP background percentages, and are not easily categorized. This analysis is most comparable to the hierarchical clustering of NDSU cultivars within the SNP-based dendrogram.

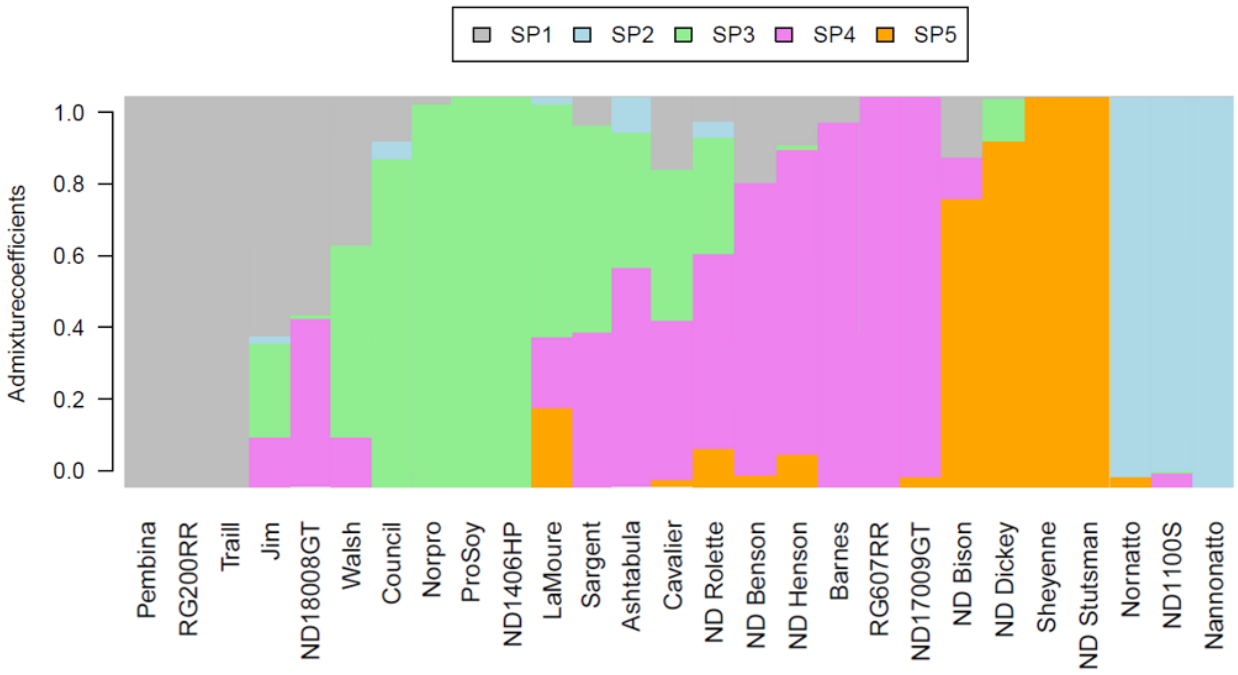

Figure 5. Population Bar Plot for 27 NDSU Soybean Cultivars.

Each colored bar represents the genetic background of a NDSU cultivar proportionally assigned to the *K* clusters (*K* = 5) with the proportions represented by the relative length of each *K* color. Cultivars with similar or identical colored bars are more genetically related to one another, while cultivars differing in bar color are more genetically distinct.

Within the SP results for NDSU cultivars and top contributing founders (Figure 6), cultivars most similar to the cultivar Traill are again best described by SP1. Specialty food grade natto cultivars belong to SP2. High-yielding commodity cultivars are best described by SP3, which also makes up about half of the background of founder A.K. (Harrow). Background of SP4 is best characterized by founders 201-14-20 and Fiskeby III, with large background percentages in Manchu, Mukden, Richland, Strain 18, Mandarin, and Mandarin (Ottawa). Specialty food grade tofu cultivars are grouped in SP5 along with other NDSU cultivars and select founders. Combining NDSU cultivars and top nine founders shifts SP5 to account for a majority of the studied background. Within this analysis, Strain 18, Mandarin, and Mandarin

(Ottawa) are the only founders with over half their background characterized by SP5, although all other founders contain SP5 background apart from Fiskeby III and 201-14-20. This analysis provides supporting evidence to which of the top contributing founders are responsible for donating certain germplasm. For example, Mandarin, Mandarin (Ottawa), and Richland are the only founders with SP1 background. Mandarin, Mandarin (Ottawa), Richland, and Manchu are the only founders with SP2 background. Multiple founders contain varying proportions of SP3, SP4, and SP5 backgrounds. This analysis is most comparable to the heatmap.

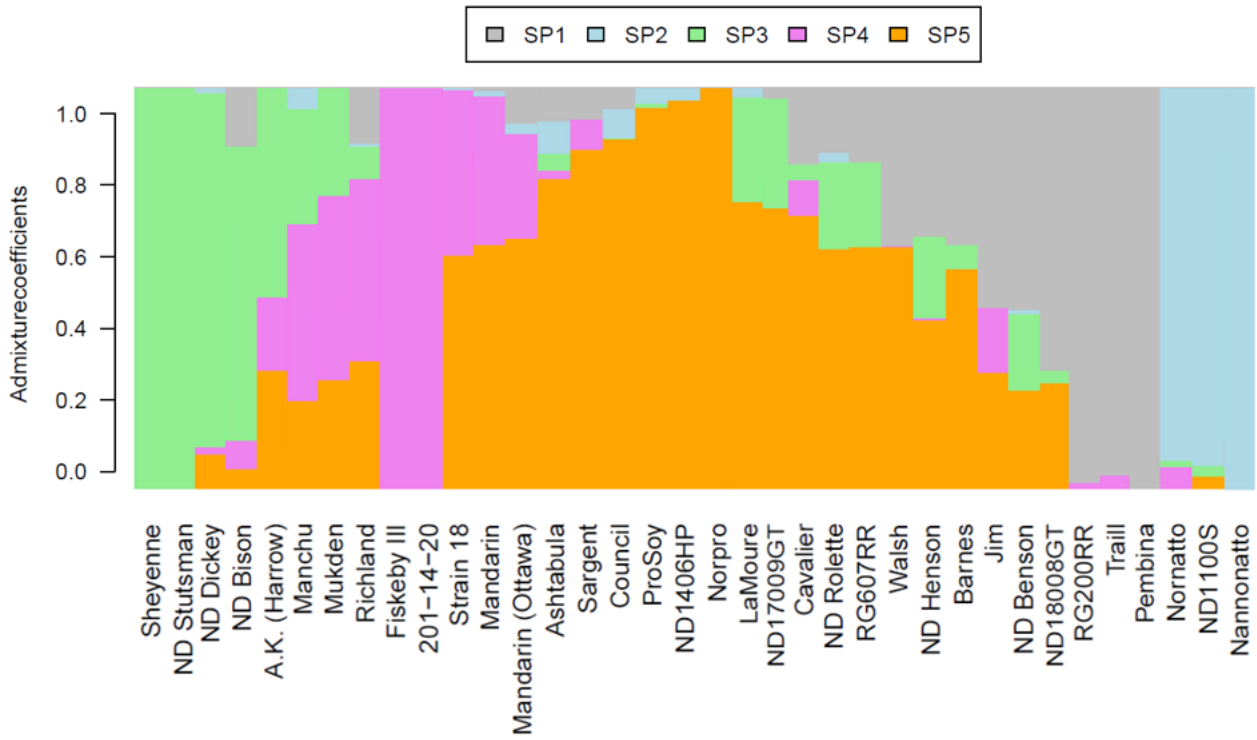

Figure 6. Population Bar Plot for 27 NDSU Soybean Cultivars and Nine Founders.

Each colored bar represents the genetic background of a cultivar proportionally assigned to the *K* clusters (*K* = 5) with the proportions represented by the relative length of each *K* color. Cultivars with similar or identical colored bars are more genetically related to one another, while cultivars differing in bar color are more genetically distinct.

4. SUMMARY

In examining two growing seasons of era trial yield data, NDSU cultivars show that consistent yield progress has been maintained from the first released cultivar in 1994 to the latest in 2021. However, the rate at which yield progression has been achieved needs to be accelerated as future cultivars are released. This is likely due to cultivar releases that were not always focused solely on yield. Highly productive cultivars at the time often lacked a certain disease package necessary for whatever challenges that became present. To combat this, these highly productive cultivars would be used as parental material with another germplasm source that carried solutions to present challenges. Given the complexity of quantitative traits there is often a tradeoff which results in the progeny inheriting the desired trait while losing yield potential. Another possible reason for the fluctuating yields can be attributed to the purpose of the cultivar. Specialty food grade natto and tofu cultivars were developed with the purpose of producing beans that met food grade requirements. Within these cultivars, it is not uncommon to observe lower yield averages than commodity cultivars. The achievement of improving yields can be expedited through improved diversification within the program.

Through pedigree data collection, a complete ancestral pedigree for the NDSU soybean breeding program was visualized. The pedigree visualization illustrates a narrative of intense inbreeding over time, highlighting a heavy reliance of elite germplasm throughout both North American and NDSU soybean breeding efforts. Certain cultivars with desired traits were recycled constantly throughout the production of new cultivars, while others were used very few times. Using the collected pedigree data to calculate coefficient of parentage then helped quantify the extent of inbreeding, shedding light to the pivotal founding contributors to released NDSU cultivars. From this, it was concluded that over 70% of all germplasm within the NDSU

soybean breeding program can be accounted for by only 10 founders. Pedigree data and coefficient of parentage calculations both helped provide information on the potential impact of diversification. At the time of this research, new germplasm has been acquired and used in cross-fertilization with NDSU cultivars to introduce new desired traits.

Deeper insights to the genetic landscape was achieved through SNP-based analyses, including dendrogram, heatmap, and population structure analysis. These analyses revealed distinct patterns of relatedness among the NDSU cultivars and founders.

The SNP-based dendrogram analysis revealed distinct genetic patterns among NDSU cultivars and founders. Founders were grouped into two main clusters (20 and 22), with eight founders displaying increased relatedness to NDSU cultivars. Heatmap analysis corroborated these results, demonstrating congruence with the five NDSU cultivar clusters from the SNP-based dendrogram. Hierarchical clustering differences were observed within the SNP-based dendrogram and founders, likely a result of analyzing 42 founders in the SNP-based dendrogram versus only nine in the heatmap.

Furthermore, population structure bar plot analyses provide insight into the genetic makeup of NDSU cultivars alone, as well as NDSU cultivars and nine of the top 10 contributing founders. Through these analyses, conclusions can be made regarding genetic similarities among NDSU cultivars as well as the founding base responsible for the genetic makeup of all NDSU cultivars. Analysis of NDSU cultivars alone provides the extent of genetic diversity within the studied cultivars. Cultivar clustering from both previous SNP-based analyses resulted in the same subpopulation groupings within the bar plot. However, the population structure bar plot provides additional information to which subpopulation groupings account for remaining genetic

backgrounds of each cultivar. Adding nine of the top 10 contributing founders to this analysis provides a shift in subpopulation makeup most similar to the heatmap analysis.

Together, these approaches have significantly enhanced the understanding and management of genetic diversity within the NDSU soybean breeding program. The comprehensive insights gained from these analyses hold the potential to transcend current yield plateaus observed within the era trial yield data. By understanding the genetic contributions of various founders and elucidating cultivar relatedness through SNP-based analyses, strategic selection of parental lines can thereby introduce novel genetic combinations that have the capacity to break existing yield barriers. This informed approach optimizes utilization of desired genetic resources and ultimately, helps the development of future superior soybean cultivars.

REFERENCES

Amadeu, R. R., Garcia, A. A. F., Munoz, P. R., & Ferrão, L. F. V. (2023). AGHmatrix: genetic relationship matrices in R. *Bioinformatics*, *39*(7), btad445.

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., & Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *The Plant Genome*, *8*(3), plantgenome2015.2004.0024.

Berglund, D. R. (2002). Soybean production field guide for North Dakota and Northwestern Minnesota.

Bernard, R., Juvik, G., Hartwig, E., & Edwards, C., Jr. (1988). Origins and pedigrees of public soybean varieties in the United States and Canada. *Technical Bulletin, US Department of Agriculture*, 1746, 68.

Board, J. E., & Kahlon, C. S. (2011). Soybean yield formation: what controls it and how it can be improved. *Soybean physiology and biochemistry*, 1-36.

Borthwick, H., & Parker, M. (1938). Influence of photoperiods upon the differentiation of meristems and the blossoming of Biloxi soy beans. *Botanical Gazette*, *99*(4), 825-839.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633-2635.

Brown, A. V., Conners, S. I., Huang, W., Wilkey, A. P., Grant, D., Weeks, N. T., . . . Nelson, R. T. (2021). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research*, *49*(D1), D1496-D1501.

Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F., Eskandari, M., & Rajcan, I. (2019). Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. *Theoretical and Applied Genetics*, *132*, 3089-3100.

Carter Jr, T. E., Nelson, R. L., Sneller, C. H., & Cui, Z. (2004). Genetic diversity in soybean. *Soybeans: Improvement, production, and uses*, *16*, 303-416.

Cui, Z., Carter, T. E., & Burton, J. W. (2000). Genetic base of 651 Chinese soybean cultivars released during 1923 to 1995. *Crop Science*, *40*(5), 1470-1481.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008.

Deasy, G. (1939). The Soya Bean in Manchuria. *Economic Geography*, *15*(3), 303-310.

Delannay, X., Rodgers, D., & Palmer, R. (1983). Relative genetic contributions among ancestral lines to North American soybean cultivars 1. *Crop Science*, *23*(5), 944-949.

Doering, E., & Sandoval, F. (1976). Hydrology of saline seeps in the Northern Great Plains. *Transactions of the ASAE*, *19*(5), 856-0861.

Dong, Y., Zhao, L., Liu, B., Wang, Z., Jin, Z., & Sun, H. (2004). The genetic diversity of cultivated soybean grown in China. *Theoretical and Applied Genetics*, *108*, 931-936.

Fageria, N. K., Baligar, V. C., & Clark, R. (2006). *Physiology of crop production*. crc Press.

Fornari, H. D. (1979). The big change: Cotton to soybeans. *Agricultural History*, *53*(1), 245-253.

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925-929.

Garner, W. W., & Allard, H. A. (1920). Effect of the relative length of day and night and other factors of the environment on growth and reproduction in plants. *Monthly Weather Review*, *48*(7), 415-415.

Gizlice, Z., Carter Jr, T., & Burton, J. (1994). Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Science*, *34*(5), 1143-1151.

Grant, D., Nelson, R. T., Cannon, S. B., & Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic acids research*, *38*(suppl_1), D843-D846.

Guo, J., Wang, Y., Song, C., Zhou, J., Qiu, L., Huang, H., & Wang, Y. (2010). A single origin and moderate bottleneck during domestication of soybean (Glycine max): implications from microsatellites and nucleotide sequences. *Annals of Botany*, *106*(3), 505-514.

Hartman, G. L., West, E. D., & Herman, T. K. (2011). Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. *Food Security*, *3*, 5-17.

Hymowitz, T., & Bernard, R. (1991). Origin of the soybean and germplasm introduction and development in North America. *Use of Plant Introductions in Cultivar Development Part 1*, *17*, 147-164.

Hymowitz, T., & Harlan, J. R. (1983). Introduction of soybean to North America by Samuel Bowen in 1765. *Economic Botany*, *37*, 371-379.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., . . . Cregan, P. B. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences*, *103*(45), 16666-16671.

Kantolic, A. G., & Slafer, G. A. (2001). Photoperiod sensitivity after flowering and seed number determination in indeterminate soybean cultivars. *Field Crops Research*, *72*(2), 109-118.

Li, Y., Guan, R., Liu, Z., Ma, Y., Wang, L., Li, L., . . . Yan, Z. (2008). Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. *Theoretical and applied genetics*, *117*, 857-871.

Li, Y. H., Li, W., Zhang, C., Yang, L., Chang, R. Z., Gaut, B. S., & Qiu, L. J. (2010). Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. *New phytologist*, *188*(1), 242-253.

Li, Z., & Nelson, R. L. (2001). Genetic diversity among soybean accessions from three countries measured by RAPDs. *Crop Science*, *41*(4), 1337-1347.

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., & Myles, S. (2015). LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics*, *5*(11), 2383-2390.

Myrdal, M. (2022). NDSU announces Ted Helms endowed professorship. https://www.ag.ndsu.edu/news/newsreleases/2022/february/ndsu-announces-ted-helms-endowed-professorship.

North Dakota Game and Fish. (2019). Climate. https://gf.nd.gov/wildlife/habitats/climate#:~:text=North%20Dakota's%20average%20annual%20temperature,fifty%20days%20below%200%C2%B0.

North Dakota Soybean Council (2020). Research: The groundwork to innovation. https://ndsoybean.org/wp-content/uploads/2021/01/2020-Research-Update-WEB.pdf.

R Core Team (2023). R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

Ray, D. K., Mueller, N. D., West, P. C., & Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. PloS one, 8(6), e66428.

Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S. R., . . . Davis, V. (2014). Genetic improvement of US soybean in maturity groups II, III, and IV. *Crop science*, *54*(4), 1419-1432.

Scott, W. O., & Aldrich, S. R. (1970). Modern soybean production. *Modern soybean production.*

Seelig, B. (2000). Salinity and sodicity in North Dakota soils.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2015). Fingerprinting soybean germplasm and its utility in genomic research. *G3: Genes, genomes, genetics*, *5*(10), 1999-2006.

Specht, J. E., Diers, B. W., Nelson, R. L., de Toledo, J. F. F., Torrion, J. A., & Grassini, P. (2014). Soybean. *Yield gains in major US field crops*, *33*, 311-355.

Stoa, T. (1950). Growing Soybeans in North Dakota. *Bimonthly Bulletin; 12*, 4.

Tavaud-Pirra, M., Sartre, P., Nelson, R., Santoni, S., Texier, N., & Roumet, P. (2009). Genetic diversity in a soybean collection. *Crop Science*, *49*(3), 895-902.

USDA. (2024). USDA/NASS QuickStats AD-hoc Query Tool. https://quickstats.nass.usda.gov/.

Viana, J. P. G., Fang, Y., Avalos, A., Song, Q., Nelson, R., & Hudson, M. E. (2022). Impact of multiple selective breeding programs on genetic diversity in soybean germplasm. *Theoretical and Applied Genetics*, *135*(5), 1591-1602.

Wang, L.-x., Lin, F.-y., Li, L.-h., Wei, L., Zhe, Y., Luan, W.-j., . . . Li, Z. (2016). Genetic diversity center of cultivated soybean (Glycine max) in China–New insight and evidence for the diversity center of Chinese cultivated soybean. *Journal of Integrative agriculture*, *15*(11), 2481-2487.

Wilson, R. F. (2008). Soybean: market driven research needs. In *Genetics and genomics of soybean*, 3-15.

Wysmierski, P. T., & Vello, N. A. (2013). The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. *Genetics and molecular Biology*, *36*, 547-555.

Xavier, A., Thapa, R., Muir, W. M., & Rainey, K. M. (2018). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genetic Resources*, *16*(6), 513-523.

Yu, G. (2022). *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. CRC Press.

Zelitch, I. (1982). The close relationship between net photosynthesis and crop yield. *Bioscience*, *32*(10), 796-802.

Zhang, L., Kyei-Boahen, S., Zhang, J., Zhang, M., Freeland, T., Watson Jr, C., & Liu, X. (2007). Modifications of optimum adaptation zones for soybean maturity groups in the USA. *Crop Management*, *6*(1), 1-11.

Zhao, S., Yin, L., Guo, Y., Sheng, Q., & Shyr, Y. (2021). heatmap3: An Improved Heatmap Package. https://CRAN.R-project.org/package=heatmap3.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326-3328.

Zhou, X., Carter, T. E., Cui, Z., Miyazaki, S., & Burton, J. W. (2000). Genetic base of Japanese soybean cultivars released during 1950 to 1988. *Crop Science*, *40*(6), 1794-1802.

APPENDIX. SCREEPLOTS EXPLAINING TOTAL GENETIC VARIANCE IN EACH PRINCIPAL COMPONENT

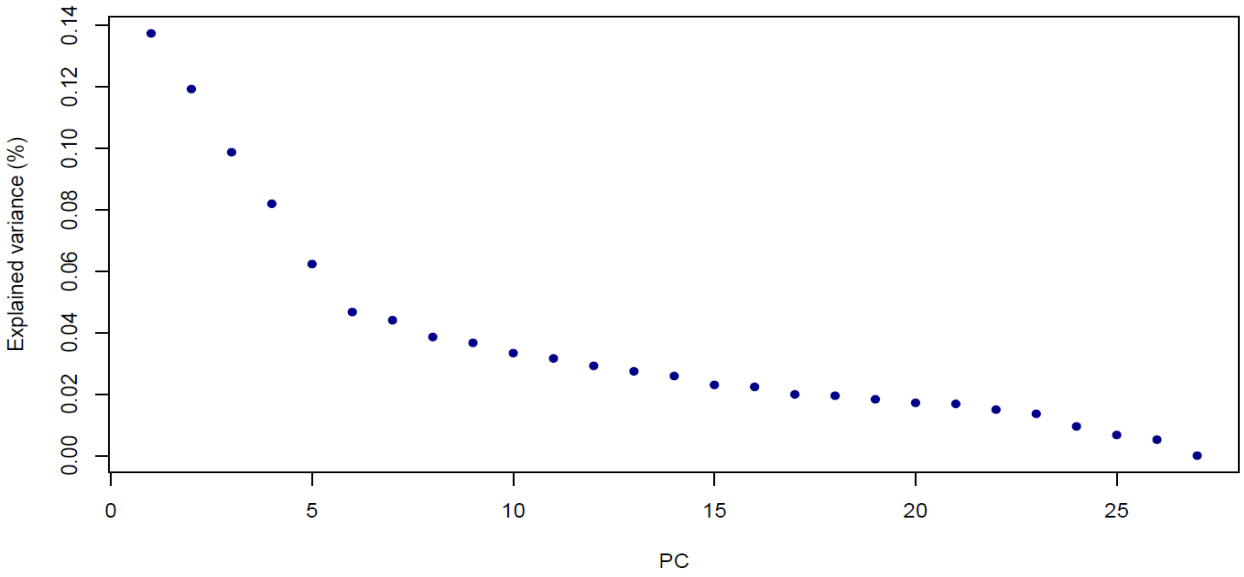

Figure A1. Screeplot Explaining Total Genetic Variance Among NDSU Cultivars.

Explained variance (%) begins to level at PC6, determining the appropriate number of genetic clusters (K = 5). Total explained variance of K = 5 is approximately 50%.

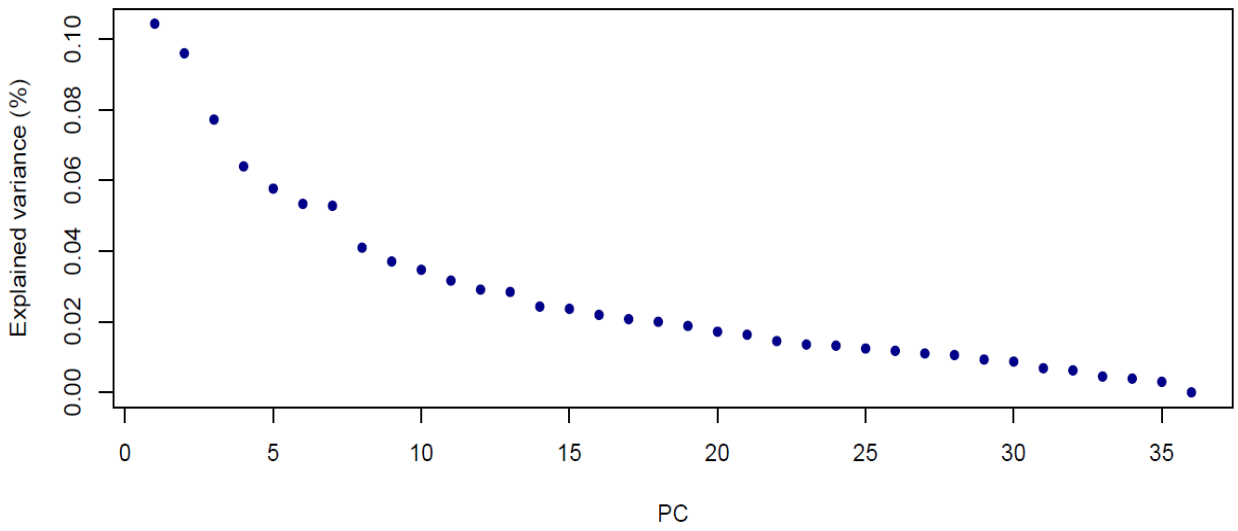

Figure A2. Screeplot Explaining Total Genetic Variance Among NDSU Cultivars and Nine Founders.

Explained variance (%) begins to level at PC6, determining the appropriate number of genetic clusters (K = 5). Total explained variance of K = 5 is approximately 40%.