EWERTON LÉLYS RESENDE

PHENOTYPIC AND GENOTYPIC VARIATION WITHIN SOYBEAN CULTIVARS

LAVRAS - MG
2024

EWERTON LÉLYS RESENDE

PHENOTYPIC AND GENOTYPIC VARIATION WITHIN SOYBEAN CULTIVARS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, área de concentração em Genética e Melhoramento de Plantas, para obtenção do título de Doutor.

Orientador
Prof. Dr. Adriano Teodoro Bruzi

LAVRAS - MG

2024
FICHA CATALOGRÁFICA

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Resende, Ewerton Lelys.
Phenotypic and genotypic variation within soybean cultivars / Ewerton Lelys Resende. - 2024.
97 p. : il.

Orientador(a): Adriano Teodoro Bruzi.

Tese (doutorado) - Universidade Federal de Lavras, 2024.
Bibliografia.

1. *Glycine max* L Merril. 2. Intracultivar Variability. 3. Conservation Breeding. I. Bruzi, Adriano Teodoro. II. Título.

EWERTON LÉLYS RESENDE

PHENOTYPIC AND GENOTYPIC VARIATION WITHIN SOYBEAN CULTIVARS

VARIAÇÃO FENOTÍPICA E GENOTÍPICA EM CULTIVARES DE SOJA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, área de concentração em Genética e Melhoramento de Plantas, para obtenção do título de Doutor.

APROVADA em 26 de março de 2024.

Dr. Arthur Tavares de Oliveira Melo	Inova Genética LTDA
Dr. José Baldin Pinheiro	ESALQ
Dr. Evandro Novaes	UFLA
Dra. Heloisa Oliveira dos Santos	UFLA

gov.br
Documento assinado digitalmente
ADRIANO TEODORO BRUZI
Data: 17/06/2024 19:21:16-0300
Verifique em https://validar.iti.gov.br

Orientador
Prof. Dr. Adriano Teodoro Bruzi

LAVRAS - MG
2024
AGRADECIMENTOS

Primeiramente, agradeço a Deus e a toda minha família, incluindo pais e avós, e em especial à minha esposa Vanessa Pereira, pelo apoio incondicional.

Ao Professor Dr. Adriano Teodoro Bruzi, expresso minha gratidão pela orientação, amizade, disponibilidade e pelos ensinamentos valiosos que levarei comigo por toda a vida.

Aos membros da banca, agradeço a disponibilidade e pelas contribuições valiosas ao trabalho.

A todos os colegas da Pesquisa Soja, pela convivência harmoniosa e pelas amizades construídas.

Um agradecimento especial aos amigos Mateus Piza e Danyllo Oliveira, pois sem a ajuda deles a conclusão deste trabalho não seria possível.

À Empresa GDM, pelo suporte nas análises de genotipagem e pela oportunidade de acompanhar as análises no laboratório.

Expresso minha gratidão a todos os funcionários de limpeza, de campo e secretariado, com um agradecimento especial à Zélia e à Rafaela pelo suporte e disponibilidade constantes.

Agradeço a todos os professores do PPGGM e aos membros do GEN pela atenção e pelos conhecimentos compartilhados.

À Universidade Federal de Lavras - UFLA e ao programa de pós-graduação em Genética e Melhoramento de Plantas - PPGGM, sou grato pela oportunidade que me foi concedida.

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

A soja desempenha um papel fundamental no agronegócio brasileiro, especialmente porque o país é o maior produtor mundial dessa cultura. Há uma busca constante por genótipos homozigotos e puros nos programas de melhoramento de soja comercial. No entanto, é importante reconhecer que a homogeneidade genética e a pureza não se mantêm estáveis a médio e longo prazo devido à notável instabilidade e flexibilidade do genoma, o que leva à variabilidade intracultivar. Assim, o objetivo deste trabalho foi investigar a variabilidade intracultivar em linhagens de soja, tanto fenotipicamente quanto genotipicamente. Foram também comparadas diferentes estratégias de classificação de progênies de soja e selecionadas as progenies mais promissoras para características que exibem variabilidade. Durante a safra de 2017/2018, 47 plantas de seis cultivares de soja foram utilizadas para estabelecer os tratamentos. Essas progênies, juntamente com o controle (checks), foram submetidas a ensaios em duas safras subsequentes, 2018/2019 e 2019/2020. As características avaliadas incluíram produtividade (YIELD), Maturidade Absoluta (FM), Dias para Floração (DF) e Altura de Plantas (PH). Além disso, a genotipagem foi realizada em 288 genótipos usando um chip com 1329 marcadores (SNPs). Os dados fenotípicos foram analisados usando modelos mistos para avaliar variância, herdabilidade, precisão e o coeficiente de variação, bem como para calcular um índice multi caracteristicas. A correlação de postos de Spearman (ρ) e o índice de coincidência (IC), foram calculados usando valores genéticos preditos derivados de diferentes estratégias analíticas, como o BLUP e GBLUP. O estudo revela a existência de variação intracultivar fenotípica e genotípica entre as cultivares avaliadas. O grau de variação observado difere entre as cultivares. O potencial tanto para seleção quanto para descarte está presente em todas as características avaliadas individualmente e também quando se emprega um índice multi caracteristica. No entanto, a correlação entre previsões genotípicas e fenotípicas para fins de seleção é baixa. Em contraste, essa correlação é maior para descarte, sugerindo que os dados genotípicos são mais aplicáveis para identificar e descartar as progênies indesejáveis. Além disso, as características de altura de plantas (PH) e dias para floração (DF) demonstram uma maior consistência entre previsões genotípicas e fenotípicas, indicando sua maior utilidade nos processos de seleção e descarte dentro de um índice multi caracteristicas. Estes resultados enfatizam a importância de considerar a previsibilidade específica da característica ao aplicar informações genômicas à estratégias de melhoramento e validam o índice multi caracteristicas como uma ferramenta útil para melhorar a eficiência dos programas de melhoramento.

Palavras-chave: *Glycine max* L. Merril.; variabilidade intracultivar; melhoramento conservativo

ABSTRACT

Soybean plays a fundamental role in the Brazilian agribusiness, mainly because the country is the world's largest producer of this crop. Consequently, there is a constant pursuit of homozygous and pure genotypes in commercial soybean breeding programs. However, genetic homogeneity and purity do not remains stable over the medium and long terms due to the notable instability and flexibility of the soybean genome, which leads to intracultivar variability. Thus, the objective of this study was to investigate intracultivar variability in soybean lines, using both phenotypic and genotypic data. Efforts were also made to compare different strategies for ranking soybean progenies and to select the most promising lines for traits exhibiting variability. During the 2017/2018 crop season, 47 plants from six soybean cultivars were used to set the genetic treatments. These progenies, along with control, were subjected to trials across two subsequent crop seasons, 2018/2019 and 2019/2020. Phenotypic traits such Grain Yield (YIELD), Full Maturity (FM), Days to Flowering (DF), and Plant Height (PH) were estimated. Additionally, genotyping was performed for 288 genotypes using a chip with 1329 SNP markers. The phenotypic data were analyzed using mixed models to evaluate variance, heritability, accuracy, and coefficient of variation, as well as to estimate a multi-trait index. Spearman's rank correlation (ρ) and the coincidence index (IC) were computed using predicted genetic values from different analytical strategies like BLUP and GBLUP. This study reveals the existence of both phenotypic and genotypic intracultivar variation among the evaluated cultivars. The degree of variation observed differs among cultivars. The potential for both selection and discarding is present across all evaluated traits individually and also when employing a multi-trait index. However, the correlation between genotypic and phenotypic predictions for the purpose of selection is low. However, this correlation is higher for discarding, suggesting genotypic data is more applicable for identifying and discarding undesirable progenies. Furthermore, traits such plant height (PH) and days to flowering (DF) showed high consistency between genotypic and phenotypic predictions, indicating they are efficient in the selection and discarding processes within a multi-trait index. These results emphasize the importance of considering trait-specific predictability when applying genomic information to breeding strategies and validate the multi-trait index as a useful tool for improving the efficiency of breeding programs.

Key Words: *Glycine max* L. Merrill.; intracultivar variability; conservation breeding

LISTA DE FIGURAS

LISTA DE TABELAS

Impactos sociais, tecnológicos, econômicos e culturais

O melhoramento de soja no Brasil desempenha um papel crucial no contexto social, tecnológico, econômico e cultural do país, e de acordo com o Levantamento da CONAB 2024, indica que a cultura da soja foi a que mais cresceu no Brasil nas últimas cinco décadas, tanto que de 1978 até 2024, esse valor representa um aumento impressionante de mais de 1500%, sendo que a área cresceu em pouco mais de 400%, demonstrando o aumento de produtividade. Novas abordagens de melhoramento para aumentar a velocidade do lançamento de novas cultivares no mercado são fundamentais. Assim, essa tese objetivou investigar a variabilidade intracultivar em linhagens de soja, tanto fenotipicamente quanto genotipicamente, comparando diferentes estratégias de classificação de progênies de soja, mostrando plausível a seleção de progênies mais promissoras para características que exibem variabilidade, possibilitando o lançamento de melhores cultivares de soja. Demonstrando impacto e participação da sociedade externa à UFLA como parceiros estratégicos, incluindo empresas de melhoramento e estudantes que participaram diretamente ou indiretamente, tanto da graduação quanto da pós graduação. Os impactos do trabalho também foram de cunho social, das quais as áreas temáticas mais relacionadas ao trabalho foram tecnologia e produção, comunicação e educação. Além disso, os impactos estão alinhados aos 17 Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU), incluindo a erradicação da pobreza, fome zero e agricultura sustentável, trabalho decente e crescimento econômico, indústria, inovação e infraestrutura, e ação contra a mudança global do clima. A soja é uma das principais culturas agrícolas do Brasil, sendo uma fonte significativa de renda para os agricultores e um componente essencial na economia nacional. Através do melhoramento genético, os pesquisadores têm desenvolvido variedades de soja mais produtivas, resistentes a pragas e doenças, e adaptadas a diferentes condições climáticas, contribuindo para aumentar a produtividade e a sustentabilidade da agricultura. Além disso, o melhoramento de soja tem impactos sociais significativos, uma vez que a melhoria da produtividade e da qualidade das safras beneficia diretamente os agricultores, suas famílias e as comunidades rurais, promovendo o desenvolvimento e a melhoria das condições de vida no campo. Do ponto de vista tecnológico, o avanço no melhoramento genético da soja tem impulsionado a inovação no setor agrícola, com a introdução de variedades mais eficientes e adaptadas, contribuindo para a modernização e a competitividade do agronegócio brasileiro. Além disso, o cultivo de soja tem impactos econômicos significativos, uma vez que o Brasil é um dos maiores produtores e exportadores mundiais desse grão, gerando divisas e contribuindo para a balança comercial do país. Em suma, trabalhos de melhoramento de soja não apenas impulsionam a produtividade agrícola, mas também promovem o desenvolvimento social, tecnológico e econômico, consolidando a soja como um pilar fundamental da agricultura e da economia mundial.

Social, technological, economic and cultural impacts

The soybean breeding in Brazil plays a crucial role in the country's social, technological, economic, and cultural context. According to the 2024 CONAB Survey, soybean cultivation has experienced the most significant growth in Brazil over the past five decades. From 1978 to 2024, this value represents an impressive increase of over 1500%, with the area expanding by just over 400%, demonstrating an increase in productivity. Novel breeding approaches to accelerate the release of new cultivars into the market are essential. Consequently, this thesis aimed to investigate the intra-cultivar variability in soybean lines, both phenotypically and genotypically, comparing different strategies for classifying soybean progenies. This study demonstrated the feasibility of selecting more promising progenies for traits exhibiting variability, enabling the release of improved soybean cultivars. It showcased the impact and participation of external society at UFLA as strategic partners, including breeding companies and students who participated directly or indirectly, both at the undergraduate or graduate level. The impacts of this work were also of a social nature, with the thematic areas most related to the work being technology and production, communication, and education. Furthermore, the impacts are aligned with the 17 Sustainable Development Goals (SDGs) of the United Nations (UN), including poverty eradication, zero hunger and sustainable agriculture, decent work and economic growth, industry, innovation and infrastructure, and action against global climate change. Soybean is one of Brazil's primary agricultural crops, serving as a significant source of income for farmers and an essential component of the national economy. Through genetic improvement, researchers have developed more productive soybean varieties resistant to pests and diseases and adapted to different climatic conditions, contributing to increased productivity and agricultural sustainability. Moreover, soybean breeding has significant social impacts, as improving crop productivity and quality directly benefits farmers, their families, and rural communities, promoting development and improving living conditions in rural areas. From a technological perspective, advances in soybean genetic improvement have driven innovation in the agricultural sector, introducing more efficient and adapted varieties, contributing to the modernization and competitiveness of Brazilian agribusiness. Additionally, soybean cultivation has significant economic impacts, as Brazil is one of the world's largest producers and exporters of this grain, generating foreign exchange and contributing to the country's trade balance. In summary, soybean breeding efforts not only drive agricultural productivity but also promote social, technological, and economic development, solidifying soybean as a fundamental pillar of agriculture and the global economy.

SUMÁRIO

PRIMEIRA PARTE

1. INTRODUÇÃO GERAL

No melhoramento de plantas autógamas, assim como nas alogamas, existe uma busca constante por genótipos homozigotos e puros, conhecidos como linhagens. Na cultura da soja, o objetivo é obter linhagens como produto comercial, enquanto no caso do milho, as linhagens homozigotas são empregadas na sintetização de híbridos. No entanto, deve-se ressaltar que a homogeneidade e pureza genética não se mantêm de forma estável a médio e longo prazo, devido à notável instabilidade genômica, que se refere à propensão de um genoma a sofrer alterações, e à flexibilidade, que se refere à capacidade do genoma de se adaptar a novas condições ambientais ou pressões seletivas. O genoma está sujeito a mudanças nas sequências de bases e a remodelações estruturais (TOKATLIDIS, 2015).

Este dinamismo do genoma está associado à complexidade do gene. No passado, um gene era considerado como um segmento de DNA localizado em uma posição específica de um cromossomo determinado, e que participava da expressão fenotípica de um certo caráter (RAMALHO *et al*., 2021). No entanto, com as novas ferramentas de estudos moleculares, um novo conceito de gene foi proposto. Nesse novo entendimento, é definido como uma unidade funcional de hereditariedade que pode incluir regiões codificadoras (éxons), regiões não codificadoras (íntrons) e elementos regulatórios (promotores, realçadores), que juntos contribuem para a produção de uma proteína funcional ou molécula de RNA, sendo responsáveis pela expressão fenotípica de um determinado caráter (PORTIN e WILKINS, 2017).

Além da complexidade inerente aos genes, outros mecanismos podem modificar o genoma, sendo notáveis a heterogeneidade remanescente, mutações, elementos transponíveis, mecanismos epigenéticos, recombinação não-homóloga e mutações cromossômicas. Essas modificações têm o potencial de induzir alterações nos fenótipos das cultivares e gerar variabilidade natural, resultando na instabilidade das linhagens ao longo dos anos de cultivo. Por outro lado, também podem ser encaradas como oportunidades nos programas de melhoramento.

No contexto do melhoramento genético da cultura da soja, significativos recursos financeiros e operacionais são alocados, juntamente com um extenso período, para o lançamento de novas cultivares. Portanto, a avaliação e seleção de plantas superiores dentro das cultivares existentes ganham relevância como uma alternativa viável para o desenvolvimento de novas cultivares. Isso tem sido documentado por diversos autores na

cultura da soja. (ACHARD *et al*., 2020; AMARAL *et al*., 2019; FASOULA e BOERMA, 2005; FASOULA e BOERMA, 2007) assim como em outras culturas, tais como milho, batata, trigo, algodão (LIU *et al*., 2018; MARAND *et al*., 2019; NINOU *et al*., 2022, TOKATLIDIS *et al*., 2008), abrangendo diferentes abordagens de fenotipagem e, principalmente, genotipagem.

Na cultura da cevada, a investigação da presença de variação genética para tolerância/resistência ao vírus do nanismo amarelo da cevada - PAV (BYDV-PAV) foi conduzida em cinco cultivares comerciais e duas variedades locais tunisianas. Os resultados evidenciam considerável variação tanto nas variedades locais adaptadas quanto nas cultivares comerciais melhoradas (GHANEM *et al*., 2018).

A investigação da variabilidade genética pode ser conduzida por meio do emprego de diferentes tipos de marcadores moleculares, sendo os SNPs (do inglês Single Nucleotide Polymorphism - polimorfismo de nucleotídeo único) um dos mais relevantes atualmente. Os SNPs apresentam estabilidade, alta frequência, facilidade de automatização e custo por dado mais acessível (THOMSON, 2014). Embora os custos tenham diminuído com o advento do sequenciamento de nova geração (NGS), eles ainda permanecem elevados quando se adota o sequenciamento de todo o genoma. Uma alternativa para solucionar essas questões é a genotipagem por sequenciamento GBS (do inglês Genotyping-by-Sequencing) (CHUNG *et al*., 2017).

O método GBS emprega uma estratégia de sequenciamento multiplex na construção de bibliotecas reduzidas e representativas para plataformas de NGS, utilizando um sistema de códigos de barras (barcodes) para aumentar a eficiência, a um custo menor se comparado a outros métodos de genotipagem (ELSHIRE *et al*., 2011). O GBS é simples, específico, altamente reprodutível e rápido, devido à detecção simultânea de SNPs e genotipagem. Esta característica é particularmente vantajosa no contexto do melhoramento de plantas, quando um grande número de genótipos necessita ser genotipado (CHUNG *et al*., 2017). Com a necessidade de maior eficiência na utilização das marcas genotipadas, tem-se preferido usar o GBS-targeted SNPs, onde são selecionadas marcas específicas para o pool genético correspondente, o qual apresenta uma maior porcentagem de SNPs chamados (CLARKE *et al*., 2014).

Diferentes metodologias podem ser empregadas para avaliar o valor genético dos indivíduos. O Melhor Preditor Linear Não Viesado (BLUP) é um preditor apropriado para estimar valores genéticos aditivos de indivíduos sob seleção. O BLUP permite a utilização simultânea de praticamente todas as fontes de informação provenientes de diversos

ensaios, avaliados em um ou mais locais, resultando em estimativas mais precisas dos valores genotípicos (RESENDE, *et al*., 2012). Outro método amplamente utilizado para a predição de características complexas na agricultura, baseado no genoma, é a melhor predição genômica não viesada, conhecida pela sigla GBLUP (VANRADEN, 2008). O GBLUP envolve o uso de uma matriz de relacionamento genômico observado, em vez de uma matriz de relacionamento esperada baseada em pedigree (VIANA *et al*., 2022).

Diante do exposto, o objetivo deste estudo foi investigar a variabilidade intracultivar em linhagens de soja, tanto do ponto de vista fenotípico quanto genotípico. Adicionalmente, procurou-se comparar diferentes estratégias de ranqueamento de progênies de soja e selecionar as linhagens mais promissoras com base em características de interesse agronômico.

2. REFERENCIAL TEÓRICO

2.1. Cultura da soja no Brasil e no mundo

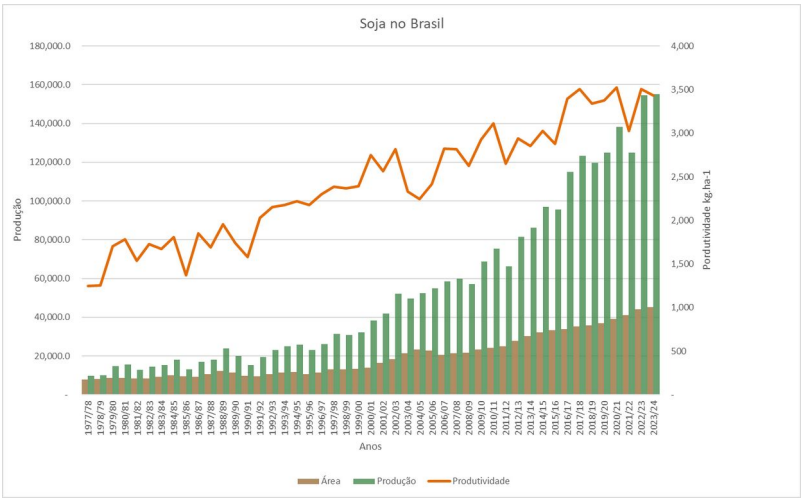
É amplamente reconhecido que a soja cultivada atualmente (*Glycine max* (L.) Merr.) foi domesticada a partir da espécie selvagem (G. soja Sieb. & Zucc.) na China há aproximadamente 5.000 anos. Posteriormente, ela foi introduzida na Coreia e no Japão por volta de 2.000 anos atrás, e sua chegada à América do Norte ocorreu em 1765 (SEDIVY; WU; HANZAWA, 2017). Contudo, a expansão significativa da cultura no território norte-americano, com relevância econômica, teve início em 1930, consolidando-se rapidamente como uma das principais culturas (CÂMARA, 2015).

No Brasil, a soja foi introduzida em 1882 por Gustavo Dutra, no estado da Bahia. No entanto, as cultivares inicialmente trazidas não obtiveram sucesso satisfatório devido às diferentes condições climáticas desse estado em relação à região de origem nos Estados Unidos. Posteriormente, em 1908, imigrantes japoneses trouxeram a soja para São Paulo, em uma latitude próxima de 22° Sul, e as primeiras observações ocorreram no Instituto Agronômico de Campinas - IAC (SEDIYAMA, 2015). Entretanto, o verdadeiro sucesso veio na região sul do país, onde as condições climáticas são mais similares àquelas da região de cultivo estadunidense. Em 1914, o professor E. Craig iniciou pesquisas no estado do Rio Grande do Sul, e em 1949 a primeira exportação de soja desse estado foi registrada (CÂMARA, 2015).

Até a década de 1970, o cultivo da soja estava principalmente concentrado na região sul do Brasil, ou seja, em latitudes mais altas (ALMEIDA *et al*., 1999). A notável

expansão descrita proporcionou ganhos expressivos na produção da cultura no Brasil, que atualmente ostenta o título de maior produtor mundial de soja, atingindo uma produção de 154,6 milhões de toneladas na safra 2022/23. Esse valor representa um aumento impressionante de mais de 1500% em comparação ao total produzido nos últimos 46 anos, conforme ilustrado na Figura 1 (CONAB, 2024).

Figura 1- Gráfico de produção, área e produtividade de soja ao longo dos anos no Brasil.



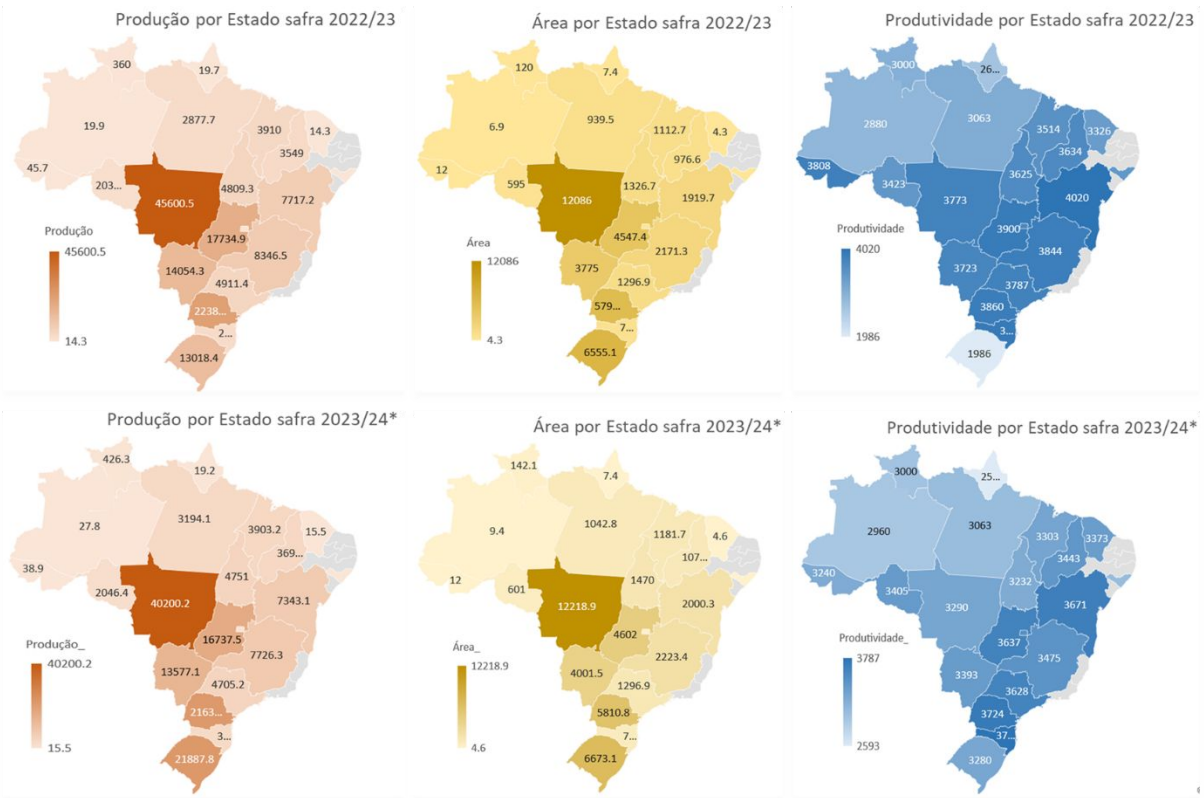
Fonte: Adaptado de CONAB (2024).

Contudo, estimativas de agosto de 2023 (FAS/USDA, 2023) indicavam uma expectativa de expansão para 163 milhões de toneladas na safra 2023/24. Entretanto, devido às condições climáticas desfavoráveis, caracterizadas por baixa pluviosidade e altas temperaturas nas regiões Centro-Oeste, Sudeste e Norte, a produção estimada foi revisada para 155 milhões de toneladas, conforme indicado pela CONAB em sua estimativa de janeiro de 2024. Vale ressaltar que outras entidades de dados apresentam projeções menores, mas, mesmo diante dessas variações, o Brasil mantém sua posição de liderança, seguido pelos Estados Unidos, Argentina e China. (Figura 2).

Figura 2 - Produção de soja por países em 2022/23 de acordo com dados do USDA.

Fonte: Adaptado de USDA (2024)

Figura 3 - Representação da produção, área e produtividade por estado brasileiro nas safras 2022/23 e 2023/24 de acordo com levantamento da Conab 2024



Fonte: Adaptado de CONAB (2024).

Em nível estadual, destaca-se Mato Grosso como o principal produtor nacional de soja. Entretanto, os desafios climáticos enfrentados durante a safra 2023/24 resultaram em perdas significativas, totalizando uma redução de 12% em comparação com a safra anterior, o que equivale a uma diminuição de 5 milhões de toneladas, conforme ilustrado na . Essa queda na produção representa um cenário desafiador para a região. Logo após Mato Grosso, na classificação dos maiores produtores estaduais, estão os estados do Rio Grande do Sul, Paraná, Goiás, Mato Grosso do Sul, Minas Gerais e Bahia. A distribuição geográfica dos polos de produção destes estados evidencia a relevância e a diversidade da contribuição de diferentes regiões brasileiras para a produção nacional de soja.

É crucial destacar que o Rio Grande do Sul enfrentou desafios significativos em termos de produtividade devido às condições de baixa pluviosidade que afetaram a região por duas safras consecutivas, abrangendo os períodos de 2021/22 e 2022/23. Em algumas localidades do estado, essa situação persistiu por três safras, conforme relatado pela Emater-RS (CANAL RURAL, 2023). Contudo, na safra 2023/24, as chuvas retornaram à região Sul, resultando em um aumento de produtividade notável, alcançando um incremento de 68% ().

Por outro lado, a Bahia destacou-se como o estado com as maiores produtividades na safra 2022/23, atingindo uma média de 4020 kg/ha. Esse desempenho foi atribuído a um sólido pacote tecnológico e condições climáticas favoráveis na região. No entanto, na safra de 2023/24, houve uma redução de 8% na produtividade, evidenciando a volatilidade das condições climáticas e a influência direta sobre os resultados agrícolas.

Em virtude da importância global da cultura e, apesar das adversidades climáticas que afetaram a produção, como a estiagem que causou quebras de safra no Rio Grande do Sul e em Santa Catarina em 2019/20, e as dificuldades enfrentadas no ano 2021/22, incluindo atrasos na semeadura devido à escassez e/ou atrasos de chuvas e atrasos e perdas na colheita devido ao excesso de precipitação e dias nublados, a soja continua a desempenhar um papel crucial na agricultura nacional.

2.2. Melhoramento genético da soja

Desde os primórdios da domesticação, quando o melhoramento era realizado mais como uma forma de arte do que como ciência, até os dias de hoje, a cultura da soja vem passando por melhorias contínuas. A produtividade nos Estados Unidos aumentou significativamente, praticamente quadruplicando, saindo de 740 kg/ha em 1924 para 3400 kg/ha em 2023 (FAS/USDA, 2024). No Brasil, os ganhos em produtividade foram ainda mais expressivos, com uma taxa média de aumento de cerca de 52 kg/ha por ano no período de 1977 a 2023, chegando a uma média nacional de 3530 kg/ha (CONAB, 2024).

Para alcançar esses avanços, foram realizadas numerosas pesquisas científicas visando o desenvolvimento de novas cultivares, bem como o aprimoramento das técnicas agronômicas. No Brasil, em particular, o melhoramento genético das cultivares pode ser dividido em três etapas importantes, que serão descritas a seguir. (SPECHT *et al*., 2014).

A primeira etapa consistiu na introdução de cultivares do sudeste dos Estados Unidos para o cultivo na região sul do Brasil. As cultivares importadas apresentavam um hábito de crescimento determinado, como a Hardee, Bossier e Davis. Outras cultivares, como Santa Rosa e Paraná, eram linhagens experimentais desenvolvidas nos Estados

Unidos, mas receberam nomes brasileiros para fins de comercialização (SPECHT *et al*., 2014). Inicialmente, esse método de melhoramento trouxe resultados significativos, principalmente devido à semelhança entre os ambientes nos Estados Unidos a região sul do Brasil.

A segunda etapa foi de extrema importância para impulsionar a produção de soja em todo o território nacional. Como é sabido, a soja é sensível ao fotoperíodo, ou seja, é considerada uma planta de dias curtos. Devido a essa característica, as cultivares inicialmente introduzidas tinham uma adaptabilidade restrita a regiões com latitudes acima de 23° S. No entanto, com a identificação dos alelos responsáveis pela expressão do fenótipo de período juvenil longo (PJL), tornou-se possível cultivar a soja em faixas mais abrangentes de latitudes e épocas de semeadura (EMBRAPA, 2021). Isso permitiu o cultivo de soja inicialmente em latitudes de 16° S e posteriormente em latitudes de 10° S. Nessa fase, várias cultivares importantes foram desenvolvidas, como a FT Cristalina, BRS 232, CD 202, CD 215, MGBR-46 (Conquista), entre outras.

Após superar a limitação relacionada à resposta ao fotoperíodo, o melhoramento da soja concentrou-se em aumentar a produtividade e melhorar a resistência a doenças, como o cancro da haste (*Diaporthe phaseolorum var. meridionalis F.A. Fern*) e a mancha olho de rã (*Cercospora sojina Hara*). Além disso, houve um foco no combate ao nematoide do cisto (*Heterodera glycines Ichinohe*), que continua sendo um patógeno significativo na região central do país até os dias atuais, bem como em outros nematoides, como o *Meloidogyne spp e o Pratylenchus brachyurus* (SPECHT *et al*., 2014).

A terceira fase do melhoramento da soja no Brasil teve início em 2000, com a mudança do hábito de crescimento das cultivares de determinado para indeterminado. Isso possibilitou a consolidação da segunda safra de milho e também do algodão após a colheita da soja. As cultivares de crescimento indeterminado permitiram o cultivo de plantas mais produtivas e precoces, uma vez que continuam crescendo após o florescimento. Além disso, as cultivares precoces desempenharam um papel importante na redução dos efeitos da ferrugem asiática (*Phakospora pachyrhizi* Sydow), pois essas cultivares têm maior probabilidade de escapar da doença e requerem menos aplicações de fungicida(SPECHT *et al*., 2014).

A partir de 1990, o cenário do melhoramento genético da soja no Brasil passou por uma transformação significativa com a entrada das multinacionais no país, especialmente em resposta à lei de proteção de cultivares, o que motivou as empresas a investirem no

melhoramento de plantas autógamas. Em 1996, a Monsanto iniciou suas atividades com cultivares convencionais.

No entanto, a partir de 2005, a base genética predominante das cultivares de soja tornou-se a transgênica denominada Roundup Ready RR (BAYER, 2021). As primeiras variedades RR cultivadas comercialmente no Brasil foram originárias da Argentina, e a cultivar Anta82RR, em particular, foi amplamente cultivada, principalmente devido à sua precocidade. Posteriormente, outras empresas, como GDM, Pioneer, Syngenta, Basf e Bayer, também implementaram programas de melhoramento de soja transgênica. Atualmente, a adoção de soja transgênica é de 99% das áreas cultivadas (CONAB, 2021).

Em 2013, a Monsanto lançou a tecnologia Intacta RR2 PRO, conferindo resistência não apenas ao glifosato, mas também às lagartas, devido à introdução de um transgene que sintetiza a proteína Cry1Ac (BAYER, 2021). Essa tecnologia rapidamente conquistou a adesão dos produtores, e várias empresas incluíram cultivares com essa tecnologia em seu portfólio, como é o caso das cultivares Intacta 2 (Bayer). A empresa Bayer lançou comercialmente a tecnologia INTACTA 2 XTEND® na soja, representando a terceira geração de transgênicos trazida pela Bayer ao mercado nacional. Além de ser tolerante ao herbicida glifosato, a plataforma também introduz uma nova ferramenta para o controle de plantas daninhas: o Dicamba, que oferecerá um amplo controle de plantas de folhas largas, além de proporcionar proteção adicional contra pragas-alvo, como Spodoptera cosmioides e Helicoverva armigera. Outro exemplo é a Conkesta Enlist (Corteva), uma cultivar de soja que, além de ser tolerante aos herbicidas 2,4-D, glifosato e glufosinato, também conta com a biotecnologia Bt de amplo espectro de controle de lagartas, representando avanços significativos nesse sentido.

No início de 2016, foi lançada a primeira cultivar de soja com a tecnologia Cultivance, fruto de uma parceria entre BASF e Embrapa. Essa tecnologia confere à soja resistência aos herbicidas do grupo químico das imidazolinonas, que são inibidores da enzima ALS (BAYER, 2016). Ainda durante o ano agrícola 2016/2017, a Bayer CropScience disponibilizou a tecnologia Liberty LinkTM em 11 cultivares. Essa tecnologia, já comercializada em países como EUA e Canadá, proporciona resistência ao herbicida Glufosinato de Amônio.

A recente revolução da biotecnologia é a tecnologia CRISPR (Repetições Palindrômicas Curtas Agrupadas e Regularmente Interespaçadas), que tem marcado uma nova era na agricultura com sua capacidade de manipulação genômica. Em um evento crucial ocorrido em 1º de setembro de 2022, a CTNBio emitiu um parecer favorável à soja

editada via CRISPR como material não transgênico, representando um marco significativo na reunião de grande importância (EMBRAPA, 2022). Esse parecer positivo foi concedido à Embrapa para a utilização da tecnologia CRISPR na edição do genoma da soja, com o objetivo principal de desativar fatores anti-nutricionais específicos, culminando na validação da soja resultante como convencional, ou seja, não transgênica. Essa aprovação oficial do projeto representa um importante marco para a tecnologia CRISPR na agricultura.

A técnica amplamente adotada de edição genômica usando CRISPR/Cas9 se destaca como uma estratégia inovadora para induzir mutações direcionadas, visando aprimorar características em culturas agrícolas. Esse contexto estabelece as bases para compreender as implicações do avanço da tecnologia CRISPR na agricultura, como evidenciado pelo parecer positivo emitido pela CTNBio (VERMA *et al*., 2023).

2.3. Fonte de variabilidade genética em cultivares de soja

Cultivares com um pool genético homogêneo, como as linhagens obtidas por autofecundação, são teoricamente consideradas puras, uniformes e estáveis. Entretanto, a plasticidade e flexibilidade do genoma tornam possível que essas cultivares sofram modificações genéticas ao longo do tempo, resultando em alterações fenotípicas. Vários autores têm demonstrado a presença de variabilidade em linhagens de soja para diversas características, possibilitando a seleção de genitores superiores (AMARAL *et al*., 2019; FASOULA; YATES; BOERMA, 2012; TOKATLIDIS, 2015).

Esse fenômeno é conhecido como variação intracultivar, referindo-se à variação genética presente entre as plantas dentro de uma mesma cultivar ou variedade (HAUN *et al*., 2011). Essa variação ocorre devido a diversos mecanismos, como heterogeneidade remanescente, falta de pureza genética das sementes, mutações, elementos transponíveis, mecanismos epigenéticos, recombinação não-homóloga e mutações cromossômicas.

Dentre os mecanismos mencionados, destaca-se a heterozigosidade remanescente (HR). No processo de autofecundação, ocorre uma redução de cinquenta por cento nos loci em heterozigose a cada geração. Portanto, após sete gerações de autofecundação, a proporção de loci em homozigose é de aproximadamente 99,2%. Ainda que os loci em heterozigose se apresentem em proporções reduzidas, eles podem segregar, originando variação fenotípica.

Em um estudo conduzido por (FASOULA; YATES; BOERMA, 2012) utilizando marcadores moleculares microssatélites (SSR - *Single Sequence Repeat*) em três cultivares

de soja, observou-se variação entre 82% e 93% decorrente da HR. Mihelich *et al*., 2020, demonstrou com mais de 20 mil acessos genotipados utilizando SNPs, a presença sugerida de heterozigosidade remanescente em alguns indivíduos. A exploração da HR tem se mostrado vantajosa no melhoramento da soja, permitindo a seleção de indivíduos superiores provenientes de cultivares elite (FASOULA e BOERMA, 2007; SEBASTIAN *et al*., 2010)

Outro mecanismo biológico é a mutação. Ainda que seja um evento raro, quando ocorre, provoca alterações permanentes no DNA. As mutações podem ser prejudiciais em nível individual, podendo resultar na morte do organismo se ocorrerem em posições vitais na sequência do DNA. Contudo, as alterações genéticas adaptativas aumentam as chances de sobrevivência de uma espécie ao longo de sua evolução (ALBERTS *et al*., 2017).

Essas mutações têm origem em duas fontes fundamentais: erros na replicação do DNA e/ou lesões químicas no material genético. Os erros na replicação limitam a precisão do pareamento de bases durante o processo, e apesar de existirem mecanismos de revisão e correção nas células, alguns erros não são identificados. Além disso, as moléculas de DNA podem sofrer danos por substâncias químicas ou radiação, levando a modificações em sua estrutura (WATSON *et al*., 2015).

Durante a mutagênese, observam-se diversos tipos de alterações genéticas, como inserções, deleções, variações no número de cópias, rearranjos cromossômicos e movimento de elementos móveis. Inicialmente, os melhoristas de plantas usavam de mutações naturais como uma fonte de variação genética para melhorar e desenvolver variedades de culturas. O conceito de melhoramento genético por mutação foi introduzido para criar maior diversidade genética entre as espécies de culturas, visando aprimorar características como resistência a doenças e pragas, tolerância a estresses abióticos e melhoria nutricional (SALGOTRA, 2023).

A mutagênese é especialmente relevante em espécies de reprodução assexuada, como alho e batata, onde desempenha um papel significativo na introdução de variabilidade genética. Além disso, na cultura do trigo, observou-se a ocorrência de um SNP a cada 540 pares de base entre 12 cultivares (WEBER et al., 2012), destacando a importância da compreensão das variações genéticas para o melhoramento de culturas.

No caso do milho, há evidências de variação intracultivar relatada em linhagens duplo-haploides (BOGENSCHUTZ e RUSSELL, 1986). Essas linhagens manifestaram acumulação significativa de mudanças para características quantitativas, superando as taxas de mutação presentes na literatura. Na cultura do tabaco, Lemos *et al*. (2023)

mostraram maior variação em linhagens provenientes de DH em comparação com linhagens obtidas por meio de autofecundações, corroborando os relatos anteriores.

Uma alternativa para gerar variabilidade é através da transposição, um mecanismo de recombinação genética que movimenta elementos genéticos de um local do DNA para outro, chamados de elementos de transposição (ET) ou transposons (WATSON *et al*., 2015). Esses elementos têm pouca seletividade em relação às sequências de inserção. Como resultado, podem se inserir em genes e/ou sequências regulatórias, causando alterações na expressão gênica, gerando novos alelos e, em alguns casos, até novos genes, assim gerando variabilidade (HIRSCH e SPRINGER, 2017).

Esse fenômeno foi estudado pela citogeneticista e ganhadora do Prêmio Nobel Barbara McClintock. Ela identificou em plantas de milho (*Zea mays*) os elementos genéticos (Ds e Ac) capazes de transpor suas localizações cromossômicas (KIM, 2017). Os elementos transponíveis dividem-se em duas classes: elementos da classe 1, ou retrotransposons, que se movem por meio do mecanismo "cópia e cola", intermediados pela enzima transcriptase reversa; e elementos da classe 2, ou DNA transposons, que se movem por meio do mecanismo "corte e cola", intermediados pela enzima transposase. (KIM, 2017).

A quantidade de elementos transponíveis em cada espécie está diretamente associada ao tamanho do genoma. Por exemplo, o genoma de *Zea mays* possui cerca de 2300 Mbp, e 84,2% desse total é devido a elementos transponíveis. Já no caso da soja, *Glycine max*, que possui um genoma de 1115 Mbp, aproximadamente 5,3% consistem em elementos transponíveis (OLIVER; MCCOMB; GREENE, 2013).

Na literatura, diversos exemplos de variabilidade intracultivar são encontrados. Stone *et al*., (2019) buscaram identificar a diversidade intracultivar em diferentes fornecedores comerciais de sementes de seis cultivares de melão. Para isso, utilizaram 32 marcadores SSR e identificaram diferenças nos lotes de sementes para uma mesma cultivar.

A variabilidade intracultivar tem sido amplamente explorada na indústria vinícola, e as cultivares "Grenache" e "Tempranillo" são bons exemplos dessa abordagem. Ambas as cultivares têm origem espanhola, mas são extensivamente cultivadas em regiões vinícolas ao redor do mundo (IBÁÑEZ *et al*., 2012). Em ambas as cultivares, foi observada uma considerável diversidade intracultivar, com 49 e 76 clones certificados para "Tempranillo" e "Grenache", respectivamente, incluindo mutações somáticas em ambas as cultivares que resultaram até mesmo em cultivares de uvas brancas. Os

programas de seleção intracultivar são atualmente uma ferramenta interessante para a adaptação de vinhedos às mudanças climáticas. Em um estudo conduzido por BUESA *et al*. (2022), a variabilidade intracultivar na Eficiência de Uso da Água (WUE, em inglês) dentro da cultivar "Grenache" foi confirmada.

Gethi *et al*., (2002) ao trabalharem com milho, identificaram variação entre linhagens de milho usando 44 marcadores SSR. Eles utilizaram seis linhagens endogâmicas de diferentes programas de melhoramento, públicos e privados, e observaram variação na frequência gênica de 7,6% entre os programas. Esse nível de variação pode ser atribuído a diferentes métodos utilizados para manter as linhagens. Mesmo dentro dos programas, as linhagens demonstraram diferenças genéticas significativas na média, representando 4,6% da variação total. O autor também destaca que, mesmo sendo pequena, essa variação nas linhagens indica que a homozigosidade em todos os loci não pode ser presumida.

Em outro estudo realizado na cultura do milho, com uma das linhagens mais conhecidas, a B37, foram utilizados 10 marcadores SSR para estimar a variação entre e dentro de duas populações dessa linhagem. Foi possível identificar, em média, 1,9 variantes alélicas diferentes por locos para as duas populações B37, com uma heterozigosidade média esperada (He) de 0,170 para a população de MRI-Kneja e 0,046 para a população IFC – Pleven (KOSTOVA *et al*., 2014).

2.4. Marcadores Moleculares: Avanços e Aplicações no Melhoramento Genético

Ao longo dos anos, houve uma evolução significativa nos marcadores moleculares. A tecnologia de marcadores de DNA começou com o desenvolvimento da técnica de RFLP (*Restriction Fragment Length Polymorphism*) nos anos 80 (BOTSTEIN *et al*., 1980). Com a ascensão dos métodos que empregam a amplificação de fragmentos de DNA via PCR (*Polymerase Chain Reaction*), novos marcadores surgiram, incluindo RAPD (*Random Amplified Polymorphism DNA*), que amplifica fragmentos polimórficos de DNA, AFLP (*Amplified Fragment Length Polymorphism*), que amplifica seletivamente grupos de fragmentos genômicos com enzimas de restrição, e os microssatélites, também conhecidos como *Single Sequence Repeats* (SSRs), caracterizados pela amplificação de fragmentos de DNA contendo sequências repetidas em tandem de um a sete pares de bases, frequentemente encontradas em abundância no genoma de plantas.

Os microssatélites destacam-se por serem altamente informativos, multi-alélicos, codominantes, altamente repetíveis e transferíveis entre espécies (GWINNER *et al*.,

2017). Vários estudos empregaram esse marcador para analisar a diversidade genética na cultura da soja, como demonstrado por autores como (CLEVER *et al*., 2020; GWINNER *et al*., 2017), que revelaram uma baixa diversidade nas cultivares disponíveis no mercado no Brasil e em Uganda.

Em um estudo conduzido por (FASOULA; YATES; BOERMA, 2012), o objetivo era identificar a variabilidade intra-cultivar genética de progênies de soja selecionadas a partir de plantas únicas em baixa densidade. Utilizando 144 marcadores SSR, o estudo identificou que a variabilidade fenotípica para conteúdo de proteína nos grãos e peso das sementes possui componentes genéticos na variação. Três cultivares foram utilizadas no estudo: Cook e Haskell, linhagens derivadas de F5 lançadas em 1991 e 1993, respectivamente, e Benning, uma linhagem derivada de F4 lançada em 1995. Ficou evidente que grande parte da variabilidade se deve à heterozigosidade remanescente nas plantas selecionadas individualmente para compor a cultivar (Abertura do bulk). Mais especificamente, 82% das variações dos alelos SSR foram observadas na Benning, 93% na Hadkell e 82% na Cook. Outras fontes de variação estão relacionadas a mutações, transposição e modificações epigenéticas.

Hoje, sistemas de marcadores de alta densidade baseados em chips de SNPs e plataformas de sequenciamento de alto rendimento são amplamente utilizados para analisar a diversidade, uma vez que os SNPs estão distribuídos por todo o genoma (YU *et al*., 2021). Além disso, a genômica tem experimentado avanços significativos nos últimos anos, impulsionados pelas capacidades de sequenciamento de DNA em larga escala e pela redução de custos associados a essas tecnologias.

Os SNPs são marcadores sequenciados, e a primeira tecnologia de sequenciamento, a de Sanger, surgiu em 1977 (MARDIS, 2017). Essa tecnologia foi usada para sequenciar genomas como o do bacteriófago phi-x174, com apenas 5375 pares de bases (SANGER *et al*., 1977), e mais tarde, o Projeto Genoma Humano (PGH). O PGH consumiu cerca de 3 bilhões de dólares e levou treze anos para sequenciar 3 bilhões de nucleotídeos.

Posteriormente, no início dos anos 2000, surgiram as plataformas de sequenciamento de segunda geração, reduzindo drasticamente os custos de sequenciamento genômico. Atualmente, o sequenciamento do genoma humano é 50.000 vezes mais barato em comparação com a técnica de Sanger (GOODWIN *et al*., 2016). O sequenciamento de segunda e terceira geração, ou Next Generation Sequencing (NGS), tem a capacidade de gerar centenas de milhões a bilhões de bases de DNA por ciclo. A

metodologia envolve a fragmentação aleatória do DNA, ligação de adaptadores, amplificação por PCR e detecção dos sinais gerados durante o sequenciamento (METZKER, 2010; NADEEM *et al*., 2018)

O sequenciamento de nova geração (NGS) revolucionou a genômica, permitindo a produção em larga escala de sequências nucleotídicas. No caso da soja, um re-sequenciamento de 31 genótipos identificou um conjunto de 205.614 SNPs, fornecendo um recurso genômico valioso (LAM *et al*., 2011) para o melhoramento genética da espécie. Apesar disso, esses esforços ainda não são ideais para genotipagem em larga escala de genótipos, devido ao custo.

Para reduzir ainda mais os custos, foram desenvolvidos métodos que envolvem o sequenciamento de apenas uma pequena fração do genoma. Três abordagens principais são descritas na literatura: Bibliotecas de Representação Reduzida (RRL), sequenciamento de DNA associado a locais de restrição (RAD) e Genotipagem por sequenciamento (GBS) (BOYLE *et al*., 2013).

O GBS surgiu como uma abordagem de sequenciamento de representação reduzida para amostras multiplexadas, possibilitando sequenciamento de alto rendimento em todo o genoma. A flexibilidade e o baixo custo do GBS o tornam uma ferramenta versátil para pesquisas em genética de plantas e melhoramento genético (POLAND e RIFE, 2012).

A metodologia do GBS, proposta por Elshire *et al*. (2011), compreende várias etapas, como extração de DNA, clivagem do DNA com enzimas de restrição, ligação de adaptadores com ou sem códigos de barras às extremidades dos fragmentos clivados, amplificação por PCR e análise bioinformática para detecção de SNPs. Um desafio dessa abordagem é a clivagem aleatória, em comparação com o GBS direcionado, que se concentra em regiões específicas do genoma para genotipagem, oferecendo uma opção de alto rendimento, personalizável e flexível.

O GBS identifica principalmente SNPs, os marcadores mais abundantes no genoma, tornando-os ideais para análises em grande escala. Os SNPs têm sido utilizados para explorar a diversidade intraespecífica, construir mapas de haplótipos, realizar estudos de associação do genoma (GWAS), descoberta de marcadores moleculares, análises de ligação genômica, seleção genômica e sequenciamento do genoma (HE *et al*., 2014).

Dentre os diversos usos dessa ferramenta a verificação da diversidade genética é um tema crucial em estudos fundamentais e aplicados. Na cultura da soja, foi encontrado, em média, um SNP a cada 2000 pares de bases em regiões codificantes e um SNP a cada 191 pares de bases em regiões não codificantes. Em comparação, em Arabidopsis, com

125 milhões de pares de bases, espera-se uma taxa de 1,75 novos SNPs devido a mutações por geração (WEBER *et al*., 2012), demonstrando a constante geração de variabilidade.

Em um estudo comparativo da diversidade genética entre genótipos de soja dos Estados Unidos e da China, LIU *et al*. (2017) propuseram examinar e comparar a variabilidade genética desses genótipos, empregando SNPs como marcadores genéticos. Os resultados obtidos revelaram uma significativa disparidade na diversidade genética entre os acessos de soja originários desses dois países. Notavelmente, os valores do conteúdo de informações de polimorfismo (PIC) para os acessos chineses foram mais elevados, registrando 0.2643, enquanto para os acessos norte-americanos esses valores foram mais baixos, medindo 0.2408. Essa observação indica que a população de soja na China exibe uma gama mais abrangente de variações genéticas em comparação com os genótipos dos Estados Unidos.

No estudo da diversidade genética utilizando SNPs, a partir da matriz de dissimilaridade, é possível calcular as distâncias entre os indivíduos. Uma das métricas mais empregadas para dados de SNP é o coeficiente de Roger, uma adaptação da distância euclidiana. Esse coeficiente busca uma distância euclidiana média entre todos os loci, justamente para superar a disparidade no número de loci avaliados em diferentes estudos, garantindo a comparabilidade das estimativas de distância (CRUZ; FERREIRA; PESSONI, 2011; LIU *et al*., 2017).

Após a construção da matriz de distâncias, segue-se com a análise de agrupamento, cuja finalidade é agrupar indivíduos geneticamente semelhantes em clusters e separar os geneticamente distintos. Diversas abordagens são utilizadas, incluindo o Método de Tocher, o Método de Tocher Modificado, o SAHN (s*equential, agglomerative, hierarchical, nonoverlapping*), o método da ligação simples, o método da ligação completa (complete linkage), o método da ligação média entre grupos, o UPGMA (*Unweighted pair-group method using arithmetic averages*) e o método da variância mínima de Ward. Utilizando uma dessas métricas, é possível representar os genótipos mais próximos e os mais distantes através de um dendrograma (CRUZ; FERREIRA; PESSONI, 2011).

Para avaliar a consistência dos agrupamentos, a matriz de coeficiente de semelhança cofenético é empregada. Nessa abordagem, após a criação do dendrograma, uma nova matriz de dissimilaridade ou similaridade é gerada entre os genótipos avaliados, permitindo a correlação entre elas. Quanto maior a correlação, menor é a distorção causada pelo agrupamento (CRUZ; FERREIRA; PESSONI, 2011). Entre os métodos de

agrupamento descritos anteriormente, o UPGMA demonstrou a maior correlação cofenética.

Uma alternativa adicional para visualizar a diversidade genética é a análise de componentes principais (PCA), que transforma um conjunto original de variáveis em outro conjunto de dimensões equivalentes, mantendo propriedades importantes. A análise agrupa os indivíduos de acordo com sua correlação ou falta de correlação; isto é, os indivíduos são agrupados devido à sua semelhança ou diferença dentro da população, representada pela variação no conjunto de características que define o indivíduo. O Componente Principal 1 tem maior importância do que o PCA2. Essa técnica se mostrou consistente com o método de estruturação e análise *neighbor-joining tree* na cultura da soja (LIU *et al*., 2017).

Dentre as análises para populações, pode-se destacar o STRUCTURE, um programa disponível gratuitamente para análise populacional desenvolvido por Pritchard *et al*. (2000). O STRUCTURE analisa diferenças na distribuição de variantes genéticas entre populações com um algoritmo iterativo Bayesiano, colocando amostras em grupos cujos membros compartilham padrões semelhantes de variação. Esse programa identifica populações a partir dos dadose atribui indivíduos a essas populações que representam o melhor ajuste para os padrões de variação encontrados (JOBLING *et al*., 2004; PRITCHARD *et al*., 2000; WAPLES e GAGGIOTTI, 2006).

STRUCTURE usa uma abordagem sistemática de agrupamento bayesiano aplicando estimativa de Markov Chain Monte Carlo (MCMC). O processo MCMC começa atribuindo aleatoriamente indivíduos a um número pré-determinado de grupos, depois as frequências alélicas são estimadas em cada grupo e os indivíduos são reatribuídos com base nessas estimativas de frequência. Isso é repetido muitas vezes, normalmente compreendendo 100.000 iterações, que resulta em uma convergência progressiva em direção a estimativas confiáveis de frequência alélica em cada população e probabilidades de adesão de indivíduos a uma população (PORRAS-HURTADO et al., 2013).

A distribuição da variabilidade genética, tanto entre quanto dentro das populações, pode ser avaliada e testada pela metodologia AMOVA (análise de variância molecular). Nesse método, a matriz de distâncias entre todos os pares de genótipos é empregada em uma análise de variância hierarquizada, resultando em estimativas de componentes de variância análogas às estatísticas F de Wright. Essa metodologia é suficientemente flexível para incorporar diversas matrizes de entrada alternativas, correspondentes a

diferentes tipos de dados moleculares, bem como diferentes suposições evolutivas, sem alterar a estrutura fundamental da análise. A significância dos componentes de variância e hipóteses estatísticas é testada por meio de abordagem permutacional, evitando a suposição de normalidade, que é comum em análises de variância, mas inadequada para dados moleculares (EXCOFFIER; SMOUSE; QUATTRO, 1992).

Dentre as medidas de diversidade genética, destaca-se a Heterozigosidade Esperada (He), também denominada índice de diversidade genética. Seus valores variam de 0 a 1, indicando baixa a alta variabilidade genética. Onde o valor 0 indica que todos os indivíduos da população são homozigotos para o mesmo alelo, ou seja, não há variação genética nesse loci. Por outro lado, um valor de 1 indica que todos os indivíduos são heterozigotos, representando a máxima variação genética nesse loci. Valores intermediários indicam diferentes níveis de variabilidade genética. Quando a heterozigosidade esperada é alta, isso significa que a população apresenta uma ampla variedade de alelos em seus loci, sugerindo que os indivíduos são geneticamente distintos uns dos outros.

Uma medida popular de variação genética é a heterozigosidade média esperada no equilíbrio de Hardy-Weinberg. Nei 1973, chamou essa medida de índice de diversidade genética e a definiu como a proporção média de heterozigotos por loco em uma população de acasalamento aleatório ou como a probabilidade de que dois alelos selecionados aleatoriamente e independentemente de um pool gênico representem alelos diferentes.

Outra medida de diversidade é a diversidade genética observada dentro de subpopulações, representada pela heterozigosidade observada (Ho). Os valores de diversidade genética observada, em termos de variação, referem-se à quantidade e variedade de diferenças genéticas presentes em uma população ou grupo de indivíduos. Quanto maior a diversidade genética observada, mais diferenças genéticas existem entre os indivíduos dessa população, indicando uma ampla gama de alelos e genótipos diferentes, o que pode ser benéfico de várias maneiras. Em contraste, quando a diversidade genética observada é baixa, significa que há menos diferenças genéticas entre os indivíduos da população. A partir dos valores de heterozigosidade observada e esperada, calculam-se as estatísticas F de Wright (NEI; TAKEO; RANAJIT, 1975), fornecendo informações sobre a estrutura genética da população.

Para a avaliação da presença de endogamia em uma população, a comparação entre a Heterozigosidade Observada (HO) e a Heterozigosidade Esperada (HE) é crucial. Quando a Heterozigosidade Observada é menor que a Heterozigosidade Esperada

(calculada sob Equilíbrio de Hardy-Weinberg), sugere-se a ocorrência de endogamia na população. Por outro lado, se a Heterozigosidade Observada é maior que a Heterozigosidade Esperada, é possível suspeitar de um rompimento no isolamento reprodutivo. Essa análise comparativa entre HO e HE fornece insights valiosos sobre os padrões de acasalamento e a estrutura genética da população em estudo.

Por outro lado, um valor de Ho próximo a 1 sugere que a diversidade genética observada é alta na população. Isso significa que existem muitos alelos diferentes em vários loci e, possivelmente, uma ampla gama de genótipos entre os indivíduos. Populações com valores próximos a 1 são geralmente mais adaptáveis a mudanças ambientais e têm maior potencial evolutivo.

Outro conceito relevante é a Frequência do Alelo Menor (MAF), utilizado na genética para quantificar a frequência de um alelo menos comum em uma população em relação aos outros alelos presentes em um determinado loco (posição no genoma). Em termos simples, a MAF refere-se à proporção do alelo menos frequente em uma população.

Os valores de MAF variam de 0 a 0,5, sendo que um MAF próximo a 0 indica que o alelo menos frequente é raro na população, encontrado apenas em uma pequena proporção dos indivíduos. Por outro lado, um MAF próximo a 0,5 indica que o alelo menos frequente é mais comum na população, estando presente em cerca da metade dos indivíduos (CRUZ; FERREIRA; PESSONI, 2011).

2.5. Modelos Mistos

Na década de 70, Henderson (1973), visando solucionar problemas na área de melhoramento animal, criou um modelo de seleção, denominado Modelo Misto, que contempla o método do — Melhor Preditor Linear Não Viesado – *Best Linear Unbiased Prediction* (BLUP) e estimativas de componentes de variância através das médias do modelo de máxima verossimilhança restrita (REML), que possibilita um processo de seleção mais acurado, análise de dados desbalanceado (RESENDE *et al*., 2016).

A abordagem estatística sob modelos mistos é utilizada para descrever dados de experimentos cuja estrutura de tratamentos envolve fatores fixos e aleatórios. Um modelo linear misto é representado da seguinte forma:

$$y = X\beta + Za + \varepsilon$$

em que: y é o vetor de observações ou fenótipos; X é matriz do modelo referente aos efeitos fixos b; β é o vetor dos efeitos fixos; Z é a matriz do modelo referente aos

efeitos aleatórios a; a é o vetor dos efeitos aleatórios ou valores genéticos aditivos (VGA) das progênies; ε é o vetor de erros, com e ~ N(0;R). Para esse modelo assume-se que os efeitos aleatórios de progênies são normalmente distribuídos, ou seja, a ~ N(0,G), sendo a matriz de covariâncias genéticas aditivas dos VGA das progênies.

Existem inúmeras vantagens práticas acerca da aplicação do procedimento REML/BLUP em análises de estudos genéticos. Além da estimação dos componentes de variância e predição acurada e não viesada dos valores genéticos de forma simultânea, a abordagem permite lidar com estruturas de dados complexas, associados a diferentes anos, locais e delineamentos; e pode ser aplicado a dados desbalanceados e a delineamentos não ortogonais (RESENDE e DUARTE, 2007). Por exemplo, nas fases preliminares do processo seletivo, quando os genótipos são numerosos e a quantidade de material propagativo restringe a avaliação em experimentos com repetições, é corrente a utilização de delineamentos que são naturalmente não ortogonais, como o de blocos aumentados (FEDERER, 1956). O efeito de shrinkage, que se caracteriza pelo encolhimento das médias ajustadas dos tratamentos (genótipos) em relação à média geral, depende do número de repetições e da herdabilidade. Esse efeito é desejável, uma vez que demonstra a propensão de se aproximar do valor genotípico verdadeiro (PIEPHO *et al*., 2008).

No caso do BLUP, os candidatos à seleção são variáveis aleatórias e o mérito de cada candidato é a soma da média da população mais o valor predito da variável aleatória associado ao candidato. Assim Segundo Bernardo (2010), no caso de dados desbalanceados, o procedimento BLUP retorna a predições mais confiáveis do que as obtidas pelo método dos quadrados mínimos (MQM). Entretanto, na ocorrência de balanceamento dos dados, BLUP e MQM muitas vezes não fornecem resultados distintos, o que leva a falsa impressão de que a escolha do estimador não é importante. Panter e Allen (1995) avaliando estratégias de escolha de genitores em soja compararam o método dos quadrados mínimos com o método de modelos mistos via BLUP. Em todos os casos analisados ficou evidenciado a superioridade dos cruzamentos obtidos por meio do BLUP, além de menor erro padrão e maior correlação dos valores preditos com o desempenho observado dos genótipos.

O BLUP foi utilizado primeiramente no melhoramento animal, onde o alto custo de fenotipagem bem como a impossibilidade da replicação de indivíduos, fez com que essa metodologia se tornasse conveniente. O método, assumindo normalmente um modelo genético aditivo, foi utilizado extensivamente em avaliações de gado de leite. Contudo até

a década de 90, as avaliações do BLUP em espécies vegetais foram bem restritas (BERNARDO, 1994).

Assim, a partir da década de 90, Bernardo (1994) propôs a utilização do BLUP para predição da performance de híbridos simples de milho baseado em dados fenotípicos e coeficientes de coancestria estimados por pedigree ou marcadores RFLPs. Desde então, a predição no melhoramento genético vegetal tem sido extensivamente estudada e novos modelos vêm sendo testados. Em seguida a metodologia foi estendida para a inclusão do genótipo das marcas para identificação de QTLs.

Entre os métodos paramétricos está o GBLUP (*Genomic Best Linear Unbiased Prediction*), que vem sendo amplamente utilizado no melhoramento vegetal. O GBLUP utiliza a relação genômica para estimar o mérito genético de um indivíduo. A equação é semelhante à do BLUP, mas inclui uma matriz de relação genômica, que captura as relações genéticas entre os indivíduos com base nos marcadores. Para tal, uma matriz de correlações genéticas é utilizada, estimada a partir de informações de marcadores de DNA. Essa matriz define a covariância entre indivíduos baseada na similaridade ao nível genômico (CROSSA *et al*., 2013).

O G-BLUP assume que todos os efeitos dos marcadores são normalmente distribuídos e têm variância igual (MEUWISSEN *et al*., 2001). Este método utiliza informações de marcadores genéticos para calcular associações entre indivíduos com base na matriz de relação genômica (HABIER *et al*., 2007). A equação de modelos mistos lineares utilizado foi:

$$Y = Xb + Z\mu a + \varepsilon$$

em que: y é o vetor de valores fenotípicos (N x 1, em que N é o número de indivíduos); b é o vetor de efeitos fixos (p x 1, onde p é o número de efeitos fixos considerados); μa é o vetor de efeitos genéticos aditivos dos indivíduos (N x 1); ε é o vetor de resíduos do modelo, onde ε ~ N (0, Iσ^2Ɛ) e σ^2Ɛ é a variância residual. X (N x p) e Z (N x N) são as matrizes de incidência para b e μa, respectivamente. A estrutura de variância dada por μa ~ N (0, Ga σa) em que σa é a variância aditiva e Ga (N x N) é a matriz de parentesco genômica para efeitos aditivos.

Sob essas configurações, a equação de modelos mistos genômicos para a predição de μa através do BLUP genômico (G-BLUP) é equivalente a:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + IG - 1_a / \left(\frac{\sigma_\varepsilon^2}{\sigma_a^2} \right) \end{bmatrix} \begin{bmatrix} b \\ \mu a \end{bmatrix} = \begin{bmatrix} X'y \\ Z\acute{y} \end{bmatrix}$$

em que σ_a^2 se refere à variância genética total da característica e σ_ε^2é a variância residual. Sem padronização a variação genética de cada loco é dada por $\frac{\sigma_a^2}{n}$, onde n está relacionada com o número de marcadores utilizados e é dada por $n = 2 \sum pi\ (1 - pi\ n\ i)$ (GIANOLA *et al*., 2009; RESENDE *et al*., 2008), onde *pi* é a frequência dos alelos dos loci *i,* e *ni* o tamanho da subpopulação i-ésima.

Outra estratégia para a estimação dos BLUPs envolve a integração da matriz de parentesco. Em comparação, o BLUP com informações de parentesco é denominado Valor Genético de Melhoramento ou *Breeding Value*, e a variância obtida é de natureza aditiva. O coeficiente de coancestria (ϕ), também conhecido como coeficiente de kinship, consanguinidade ou parentesco, constitui uma medida clássica de relação genética entre indivíduos. Este parentesco é determinado com base na genealogia ou, preferencialmente, por meio de marcadores moleculares, proporcionando uma avaliação mais precisa. Nesse contexto, é desenvolvida uma matriz aditiva de parentesco genético, denotada por A, na qual cada elemento é duas vezes o coeficiente de coancestria 2ϕi entre os pares de indivíduos. O coeficiente de parentesco é definido como a probabilidade de dois gametas selecionados aleatoriamente terem alelos idênticos por descendência (GAUTASON *et al*., 2023)

Assim, tais informações sobre a existência de similaridade genética podem ser incorporadas ao modelo, possibilitando a obtenção de resultados mais precisos (por meio da consideração da informação de parentesco) e imparciais (por meio da consideração da correlação entre os valores genéticos na estimação dos efeitos fixos) (RESENDE, 2012). Vale ressaltar que, pelo método dos quadrados mínimos, essas informações são desconsideradas.

Em uma análise realizada por Piepho *et al*. (2008), a incorporação da informação de parentesco conduziu a predições mais precisas do valor genético e erro quadrático médio menor em comparação com aquelas obtidas sem essa informação. Adicionalmente, observou-se que essa inclusão não apenas eleva a acurácia das avaliações genéticas, mas também aprimora a eficiência do processo de seleção, resultando em ganhos genéticos

superiores quando contrastados com situações em que a matriz de parentesco não é utilizada (NUNES *et al*., 2008)

Sob essas configurações, a equação de modelos mistos genômicos para a predição de μa através do BLUP parentesco é equivalente a:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right) \end{bmatrix} \begin{bmatrix} b \\ \mu a \end{bmatrix} = \begin{bmatrix} X'y \\ Z\acute{y} \end{bmatrix}$$

Onde A é a matriz de parentesco.

REFERÊNCIAS

ACHARD, F et al. **Single nucleotide polymorphisms facilitate distinctness-uniformity-stability testing of soybean cultivars for plant variety protection.** Crop Science, v. 1, n. June 2019, p. 1–24, 2020.

ALBERTS, B. et al. **Biologia molecular da célula**. 7. ed. [S.l.]: Garland Science, 2017. 1464 p.

ALMEIDA, L. A de et al. **Melhoramento da soja para regiões de baixas latitudes. In: Recursos Genéticos e Melhoramento de Plantas para o Nordeste Brasileiro**. 1. ed. Brasília: Embrapa Recursos Genéticos e Biotecnologia, 1999. p. 100–115.

AMARAL, L. de O et al. **Pure line selection in a heterogeneous soybean cultivar**. Crop Breeding and Applied Biotechnology, v. 19, n. 3, p. 277–284, 2019.

BAYER. **Agricultura.** 2016. Disponível em: https://www.basf.com/br/pt/products-and-industries/agriculture.html. Acesso em: 22 nov. 2023.

BAYER. **INTACTA 2 XTEND**. 2021. Disponível em: https://www.bayer.com.br/midia/sala-de-imprensa/crop-science/releases/intacta-2-xtend-sera-apresentada-em-campo-pela-primeira-vez-durante-o-evento-gigantes-da-soja.php. Acesso em: 22 mar. 2023

BERNARDO, R. **Breeding for quantitative traits in plants.**2. ed. Woodbury, Minnesota: [s.n.], 2010. 400 p.

BERNARDO, R. **Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids**. Crop Science, v. 34, n. 1, p. 20-25, 1994. DOI: 10.2135/cropsci1994.0011183X003400010003x.

BOGENSCHUTZ, T. G.; RUSSELL, W. A. **An evaluation for genetic variation within maize inbred lines maintained by sib-mating and self-pollination**. Euphytica, v. 35, p. 403-412, 1986.

BOTSTEIN, D et al. **Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms.** The American Journal of Human Genetics, v. 32, p. 314–331, 1980.

BOYLE, B.; SONAH, H.; BASTIEN, M.; IQUIRA, E. **An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping.** PLoS ONE, v. 8, n. 1, p. 1–9, 2013.

BUESA, I et al. **Unraveling the Physiological Mechanisms Underlying the Intracultivar Variability of Water Use Efficiency in Vitis vinifera "Grenache"**. Plants, v. 11, 3008, 2022. DOI: 10.3390/plants11213008.

CÂMARA, G. M. de S. **Introdução ao agronegócio soja**. Piracicaba, SP: USP/ESALQ/LPV, 2017. Disponível em: https://edisciplinas.usp.br/pluginfile.php/5746644/mod_resource/content/1/LPV%200584%202017%20-%20REVISAO%20Soja%20Apostila%20Agronegocio%20%282%29.pdf. Acesso em: 06 mar. 2024.

CANAL RURAL. **China e a qualidade da soja brasileira: exigências de proteína e óleo**. 29 ago. 2022. Disponível em: https://www.canalrural.com.br/projeto-soja-brasil/china-qualidade-soja-brasileira-exigencias-proteina-oleo/. Acesso em: 20 abr. 2023.

CANAL RURAL. **Produtores de soja do RS colhem menos de 20 sacas por hectare**. [S.l.]: Canal Rural, 2023. Disponível em: https://www.canalrural.com.br/nacional/rio-grande-do-sul/produtores-soja-rio-grande-do-sul-produtividade-baixa/. Acesso em: 18 dez. 2023.

CHUNG, Y. S.; CHOI, S. C.; JUN, T.; KIM, C. **Genotyping-by-Sequencing: a Promising Tool for Plant Genetics Research and Breeding.** Horticulture, Environment, and Biotechnologyv. 58, n. 5, p. 425–431, 2017.

CLARKE, S et al. **Genotyping-by-Sequencing (GBS) in Sheep: Comparison to SNP chips and Whole Genome Sequencing.** Invermay Agricultural Centre, v. p. 610, 2014.

CLEVER, M et al. **Genetic diversity analysis among soybean genotypes using SSR markers in Uganda**. African Journal of Biotechnology, v. 19, n. 7, p. 439–448, 2020.

CONAB. **Acompanhamento da Safra Brasileira. Boletim da Safra 2023**, v. 9, n.5 Terceiro levantamento, p. 60, 2024.

CORTEVA AGRISCIENCE. **Tecnologias 2022**. 14 ago. 2022. Disponível em: https://www.corteva.com.br/produtos-e-servicos/tecnologias/sistema-enlist.html. Acesso em: 20 abr. 2023.

CROSSA, J et al. **Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing**. G3, Bethesda, v. 3, n. 11, p. 1903-1926, 2013.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. 2. Ed, UFV, 2020.

ELSHIRE, R. J et al. **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.** PLoS ONE, v. 6, n. 5, p. 1–10, 2011.

EMBRAPA SOJA. **Comunicado sobre apodrecimento de vagens de soja na safra 2020/21** A. p. 2, 2021.

EMBRAPA SOYBEANS. **Gene-edited to reduce anti-nutritional factors, soybeans get green light**. In: Embrapa Soybeans, 2022. Disponível em: https://www.embrapa.br/en/busca-de-noticias/-/noticia/73468020/gene-edited-to-reduce-anti-nutritional-factors-soybeans-get-green-light. Acesso em: 17 dez. 2023.

EMBRAPA. **Tecnologias de Produção de Soja**. 1. ed. Londrina: Embrapa Soja, 2020.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. **Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes**: Application to Human Mitochondrial DNA Restriction Data. **Genetics**, v. 131, n. 2, p. 479–491, 1992. DOI: 10.3354/meps198283.

FAS/USDA. **World Agricultural Production**, 2024. Disponível em: https://fas.usda.gov/data/world-agricultural-production . Acesso em: 03 mar. 2024.

FASOULA, V. A.; BOERMA, H. R. **Divergent selection at ultra-low plant density for seed protein and oil content within soybean cultivars**. Field Crops Research, v. 91, p. 217–229, 2005.

FASOULA, V. A.; BOERMA, H.R. **Intra-Cultivar Variation for Seed Weight and Other Agronomic Traits within Three Elite Soybean Cultivars**. Crop Science, v. 47, p. 367–373, 2007.

FASOULA, V. A.; YATES, J. L.; BOERMA, H. R. **SSR-Marker Analysis of the Intracultivar Phenotypic Variation Discovered within 3 Soybean Cultivars**. Heredity, v. 103, n. 4, p. 570–578, 2012.

FEDERER, W. T. Augmented (or hoonuiaku) designs. **Hawaiian Planter's Record**, v. 55, n. 2, p. 191-208, 1956.

GAUTASON, E. et al. **Impact of kinship matrices on genetic gain and inbreeding with optimum contribution selection in a genomic dairy cattle breeding program**. Genetics Selection Evolution, [s.l.], v. 55, n. 48, 2023. DOI: 10.1186/s12711-023-00826-x.

GETHI, J. G et al. **SSR variation in important U.S. maize inbred lines**. Crop Science, v. 42, n. 3, p. 951–957, 2002.

GHANEM, H et al. **Exploiting intra-cultivar variation to select for Barley yellow dwarf virus-PAV (BYDV-PAV) resistance in barley**. Canadian Journal of Plant Science, v. 98, n. 4, p. 930-946, 2018. DOI: 10.1139/cjps-2017-0364.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. **Coming of age: Ten years of next-generation sequencing technologies**. Nature Reviews Genetics, v. 17, n. 6, p. 333–351, 2016.

GWINNER, R et al. **Genetic diversity in Brazilian soybean germplasm**. Crop Breeding and Applied Biotechnology, v. 17, p. 373–381, 2017.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. **The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values**. Genetics, v. 177, n. 4, p. 2389-2397, fev. 2008. DOI: 10.1534/genetics.107.081190.

HAUN, W. J. et al. **The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82**. Plant Physiology, v. 155, n. 2, p. 645–655, 2011.

HE, J et al. **Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding**. Frontiers in Plant Science, v. 5, n. SEP, p. 1–8, 2014.

HIRSCH, C. D.; SPRINGER, N. M. **Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1860, n. 1, p. 157–165, 2017.

IBÁÑEZ, J. et al. **Genetic Origin of the Grapevine Cultivar Tempranillo**. American Journal of Enology and Viticulture, v. 63, p. 549–553, 2012.

JOBLING, M. A.; HURLES, M.; TYLER-SMITH, C. **Human Evolutionary Genetics: Origins, Peoples and Disease. New York: Garland Science** - Taylor and Francis Group, 2004.

KIM, N. S. **The genomes and transposable elements in plants: are they friends or foes?** Genes & Genomics, v. 39, n. 4, p. 359–370, 2017.

KOSTOVA, A et al. **Genetic Variation Within and Among Populations of Maize Inbred B37 Revealed By SSR Markers**. Biotechnology & Biotechnological Equipment, v. 28, n. 18, p. 37–44, 2014.

LAM, H et al. **Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection.** Nature Genetics, v. 42, n. 12, p. 42–52, 2011.

LEMOS, R. C et al. **Comparison between doubled haploid lines and lines obtained via the bulk method in tobacco**. Crop Breeding and Applied Biotechnology, p. 22, 2023.

LIU, N et al. **Intraspecific variation of residual heterozygosity and its utility for quantitative genetic studies in maize**. BMC Plant Biology, v. 1, p. 1–15, 2018.

LIU, Z et al. **Comparison of Genetic Diversity between Chinese and American Soybean (***Glycine max* **(L .)) Accessions Revealed by**. v. 8, n. November, p. 1–13, 2017. DOI: 10.3389/fpls.2017.02014.

MARAND, A. P et al. **Residual heterozygosity and epistatic interactions underlie the complex genetic architecture of yield in diploid potato**. Genetics, v. 212, p. 317–332, 2019.

MARDIS, E. R. **DNA sequencing technologies: 2006-2016**. Nature Protocols, v. 12, n. 2, p. 213–218, 2017.

METZKER, M. L. **Sequencing technologies the next generation**. Nature Reviews Genetics, v. 11, n. 1, p. 31–46, 2010.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. **Prediction of total genetic value using genome-wide dense marker maps**. Genetics, Bethesda, v. 157, n. 4, p. 1819-1829, abr. 2001.

MIHELICH, N T. et al. **Characterization of genetic heterogeneity within accessions in the USDA soybean germplasm collection**. The Plant Genome, v. 13, n. 1, p. 1–15, 2020.

NADEEM, M. A et al. **DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing**. Biotechnology and

Biotechnological Equipment, v. 32, n. 2, p. 261–285, 2018. DOI: 10.1080/13102818.2017.1400401.

NEI, M. Mint: **Analysis of Gene Diversity in Subdivided Populations**. Proceedings of the National Academy of Sciences, v. 70, n. 12, p. 3321-3323, 1973. DOI: 10.1073/pnas.70.12.3321

NEI, M.; TAKEO, M.; RANAJIT, C. **The Bottleneck Effect and Genetic Variability in Populations**. v. 29, n. 1, p. 1–10, 1975.

NINOU, E et al. **Utilization of Intra-Cultivar Variation for Grain Yield and Protein Content within Durum Wheat Cultivars**. Agriculture, v. 12, n. 5, p. 661, 2022. Disponível em: https://www.mdpi.com/2077-0472/12/5/661/htm. Acesso em: 11 maio 2023.

NUNES, A.R.; RAMALHO, M.A.P.; FERREIRA, D.F. **Inclusion of genetic relationship information in the pedigree selection method using mixed models**. Genetics and Molecular Biology, v. 31, n. 1, p. 73-78, 2008. Disponível em: http://www.sbg.org.br. Acesso em: 20 nov. 2023.

OLIVER, K. R.; MCCOMB, J. A.; GREENE, W. K. **Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity**. Genome Biology and Evolution, v. 5, n. 10, p. 1886–1901, 2013.

PANTER, D. M.; ALLEN, F. L. **Using Best Linear Unbiased Predictions to Enhance Breeding for Yield in Soybean: I. Choosing Parents**. Crop Science, v. 35, n. 2, p. 397-405, mar. 1995. DOI: 10.2135/cropsci1995.0011183X003500020020x.

PIEPHO, H.P et al. **BLUP for phenotypic selection in plant breeding and variety testing.** Euphytica, [S.l.], v. 161, p. 209–228, 2008. DOI: 10.1007/s10681-007-9449-8.

POLAND, J. A.; RIFE, T. W. **Genotyping-by-Sequencing for Plant Breeding and Genetics.** The Plant Genome, v. 5, n. 3, p. 92–102, 2012.

PORRAS, H et al. **An overview of STRUCTURE: applications, parameter settings, and supporting software**. Frontiers in Genetics, v. 4, art. 98, 29 may 2013. DOI: 10.3389/fgene.2013.00098.

PORTIN, P.; WILKINS, A. **The evolving definition of a gene**. Genetics, v. 205, n. 11, p. 1353–1364, 2017.

PRITCHARD, J. K et al. **Association Mapping in Structured Populations**. American Journal of Human Genetics, v. 67, n. 1, p. 170-181, jul. 2000. DOI: 10.1086/302959.

RAMALHO, M. A.P et al. **Genética na Agropecuária**. 6. Ed, UFLA, 2021.

RESENDE, M. D. V et al. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa, MG: Universidade Federal de Viçosa, 2012.

RESENDE, M. D. V. de; DUARTE, J. B. **Precisão e controle de qualidade em experimentos de avaliação de cultivares**. Pesquisa Agropecuária Tropical, v. 37, n. 3, p. 182–194, 2007.

RESENDE, M. D. V. E. A et al. **Selection index with parentes, population, progenies and generations effects in autogamous plant breeding**. Crop Science, v. 56, n. 2, p. 530-546, Madison, 2016.

SALGOTRA, R.K.; CHAUHAN, B.S. **Genetic Diversity, Conservation, and Utilization of Plant Genetic Resources**. Genes, [S.l.], v. 14, 174, 2023. DOI: 10.3390/genes14010174.

SANGER, F.; NICKLEN, S.; COULSON, A. R. **DNA sequencing with chain-terminating inhibitors**. Proceedings of the National Academy of Sciences, v. 74, n. 12, p. 5463-5467, 1977.

SEBASTIAN, S. A et al. **Context-specific marker-assisted selection for improved grain yield in elite soybean populations**. Crop Science, v. 50, n. 4, p. 1196–1206, 2010.

SEDIVY, E. J.; WU, F.; HANZAWA, Y. **Soybean domestication: the origin, genetic architecture, and molecular bases**. New Phytologist, v. 214, n. 2, p. 539–553, 2017.

SEDIYAMA, T. **Melhoramento Genético da Soja**. Londrina - PR, p. 352, 2015.

SPECHT, J. E et al. **Yield Gains in Major U.S. Field Crops.** In: ASA, CSSA, AND SSSA (Ed.). Yield Gains in Major U.S. Field Crops, 2014. V. N. p. 311–356.

STONE, S.; BOYHAN, G.; MCGREGOR, C. **Inter- and intracultivar variation of heirloom and open-pollinated watermelon cultivars**. HortScience, v. 54, n. 2, p. 212–220, 2019.

THOMSON, M. J. **High-Throughput SNP Genotyping to Accelerate Crop Improvement.** Plant Breeding and Biotechnology, v. 2014, n. 3, p. 195–212, 2014.

TOKATLIDIS, I. S et al. **Variability within cotton cultivars for yield, fibre quality and physiological traits**. The Journal of Agricultural Science, v. 146, n. 4, p. 483–490, 2008.

TOKATLIDIS, I. S. **Conservation breeding of elite cultivars.** Crop Science, v. 55, n. 6, p. 2417–2434, 2015.

VANRADEN, P. M. **Efficient Methods to Compute Genomic Predictions.** Journal of Dairy Science, v. 91, n. 11, p. 4414-4423, Nov. 2008.

VERMA, V et al. **CRISPR-Cas: A robust technology for enhancing consumer-preferred commercial traits in crops.** Frontiers in Plant Science, v. 14, 1122940, 2023. DOI: 10.3389/fpls.2023.1122940.

WAPLES, R. S.; GAGGIOTTI, O. **What is apopulation? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity.** Molecular Ecology, v. 15, n. 6, p. 1419–1439, 2006. DOI: 10.1111/j.1365-294X.2006.02890.x.

WATSON, J. D et al. **Biologia molecular do gene**. Porto Alegre: Artmed Editora, 2015.

WEBER, N et al. **Editor's Choice: Crop genome plasticity and its Relevance to food and feed safety of genetically engineered breeding stacks**. Plant Physiology, v. 160, n. 4, p. 1842–1853, 2012.

YU, D. et al. **Genetic diversity and population structure of popcorn germplasm resources using genome-wide SNPs through genotyping-by-sequencing**. Genetic Resources and Crop Evolution, v. 68, n. 6, p. 2379–2389, 2021. DOI: 10.1007/s10722-021-01137-0.

SEGUNDA PARTE – ARTIGO

ARTICLE: ARE THERE INTRACULTIVAR GENETIC VARIABILITY IN SOYBEAN LINES?

Artigo redigido conforme a norma para publicação periódica científica NBR 6022 (ABNT, 2018)

ABSTRACT

We aimed to explore the intracultivar variability of soybean lines by conducting both phenotypic and genotypic analyses and to evaluate strategies for ranking the soybean progenies to identify the most promising lines. During the 2017/2018 crop season, plants from six soybean cultivars—SYN1359S IPRO, P98Y11, BMX6160, 97R73 RR, NS7000 IPRO, and NA5909—were used to set the genetic treatments. For each of these six cultivars, 47 plants were selected to generate progenies and, along with the controls, were subjected to trials across two subsequent growing seasons, 2018/2019 and 2019/2020. Phenotypic traits such as Grain Yield (YIELD), Full Maturity (FM), Days to Flowering (DF), and Plant Height (PH) were analyzed. Additionally, 288 samples (progenies and controls) were genotyped by a chip with 1329 SNPs using the Ion S5™ XL System. Phenotypic data were analyzed using mixed models, and the genotypic data underwent quality control and imputation of missing data before population analysis. This analysis included measures such as observed and expected heterozygosity, hierarchical clustering (UPGMA), and principal component analysis (PCA). Best linear unbiased prediction (BLUP) and genomic BLUP (GBLUP) were then estimated, and a comparison between them was made using Spearman's rank correlation (ρ) and the coincidence index (IC), as proposed by Hamblin and Zimmermann (1986). Additionally, multivariate genotypic and phenotypic selection was conducted using the selection index proposed by Mulamba and Mock (1978). The study reveals the existence of both phenotypic and genotypic intracultivar variation among the assessed cultivars. The potential for both selection and discarding are present across all evaluated traits individually and also when employing a multi-trait index. The correlation between genotypic and phenotypic predictions for the selection is low. In contrast, it is higher for discarding. Traits like plant height (PH) and days to flowering (DF) demonstrate a higher consistency between genotypic and phenotypic predictions, indicating their greater utility in the selection and discarding processes within a multi-trait index. These results emphasize the importance of considering trait-specific predictability when applying genomic information to breeding strategies and validate the multi-trait index as a useful tool for improving the efficiency of plant breeding programs.

Key Words: *Glycine max* L. Merrill.; intracultivar variability; genomic selection; BLUP; GBLUP.

1. INTRODUCTION

Selection of pure and homozygous lines is a fundamental goal of plant breeding programs. Soybean lines are typically developed to be used as a cultivar. Nonetheless, the genetic stability of these lines is prone to intra-cultivar variation, which can manifest as genomic changes and structural alterations over time (TOKATLIDIS, 2015). Several mechanisms, including residual heterozygosity, mutations, transposable elements, epigenetic modifications, process of cross-pollination and chromosomal mutations, can induce genomic changes which contributes to phenotypic variability within cultivars. These genomic alterations may impact the stability of lines over time but can also present opportunities for breeding programs.

The development of new soybean cultivars requires significant investment, and the evaluation and selection of superior plants within existing cultivars is considered a cost-effective breeding strategy. This approach has been documented in soybean research and applied to other crops such as corn, potato, wheat, and cotton, where a combination of phenotyping and genotyping methods is employed (ACHARD *et al*., 2020; AMARAL *et al*., 2019; FASOULA and BOERMA, 2005; FASOULA and BOERMA, 2007; TOKATLIDIS, 2015; LIU *et al*., 2018; MARAND *et al*., 2019; NINOU *et al*., 2022; TOKATLIDIS *et al*., 2008).

For instance, genetic variation in barley for resistance to barley yellow dwarf virus-PAV (BYDV-PAV) among local Tunisian varieties and commercial cultivars has been identified, suggesting breeding potential for disease resistance (GHANEM *et al*., 2018). However, selection within a soybean cultivar raises important considerations in the context of plant variety protection, which is essential for intellectual property rights and the encouragement of breeding programs (ACHARD *et al*. 2020).

In the assessment of genetic variability, the Single Nucleotide Polymorphisms (SNPs) markers are highly valued due to their stability and universality. Even with the cost reductions associated with next-generation sequencing (NGS), whole-genome sequencing remains expensive, and Genotyping-by-Sequencing (GBS) provides an economical alternative for genotyping. Also, GBS employe multiplex sequencing and barcodes to process numerous samples efficiently, which is beneficial for plant breeding due specificity and high throughput (CHUNG et al., 2017).

To assess the genetic value of an individual, various methodologies are available. The Best Linear Unbiased Prediction (BLUP) is a reliable method for estimating the additive genetic values of individuals under selection, utilizing data from multiple trials and locations. Genomic Best Linear Unbiased Prediction (GBLUP) is another method employed for predicting complex traits in agriculture, which uses an observed genomic relationship matrix rather than an expected pedigree-based relationship matrix (VANRADEN, 2008; VIANA *et al*., 2022).

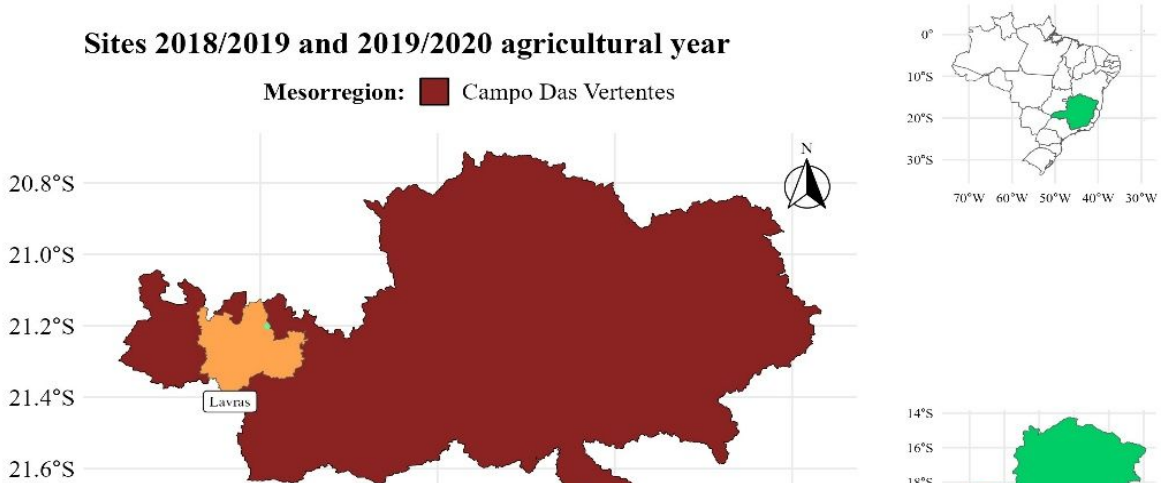
In this way, we aimed to investigate the intracultivar variability of soybean lines through phenotypic and genotypic analyses and to compare different strategies for ranking soybean progenies to select the most promising lines based on interest agronomic traits.

2. MATERIALS AND METHODS

2.1. Site and weather conditions

The phenotyping experiments were carried out at the Muquém Farm, which is part of the Center for Scientific and Technological Development in Agriculture from the Federal University of Lavras (UFLA). The farm's geographic coordinates are 21°14' South latitude and 45°00' West longitude, and it is positioned at an altitude of 918 meters. Located in the city of Lavras in the state of Minas Gerais, Brazil, the farm's location and layout are illustrated in figure 1.

Figure 1 - Representation of the experiment implementation sites in Campo das Vertentes, Minas Gerais – Brazil.

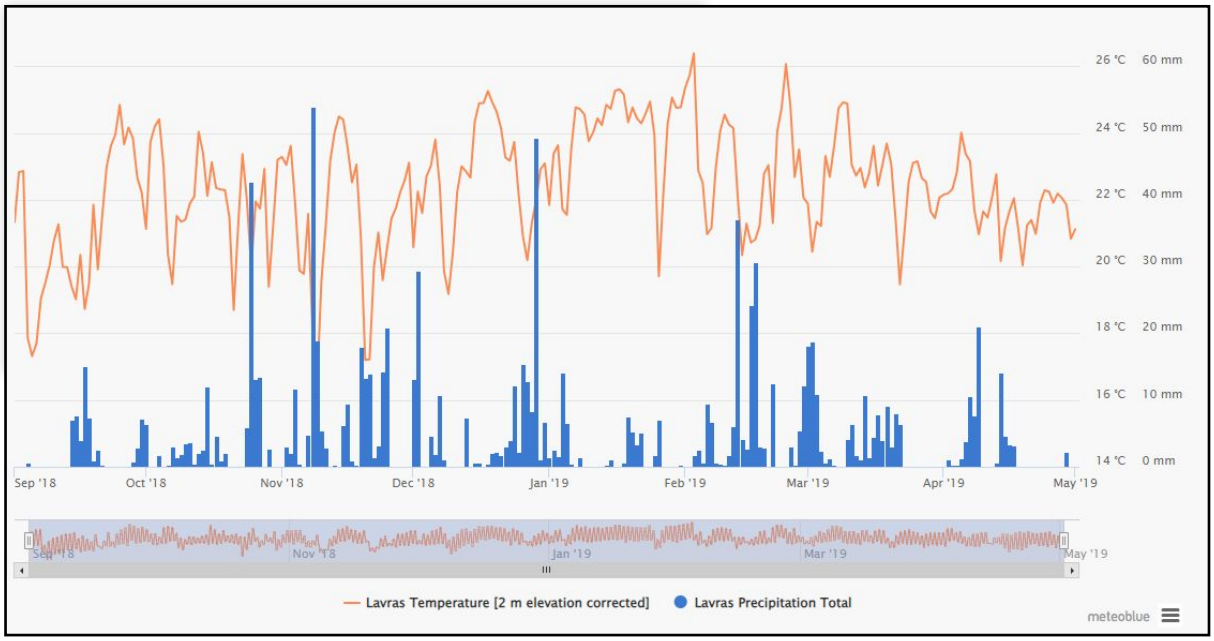


Source: From the author (2024).

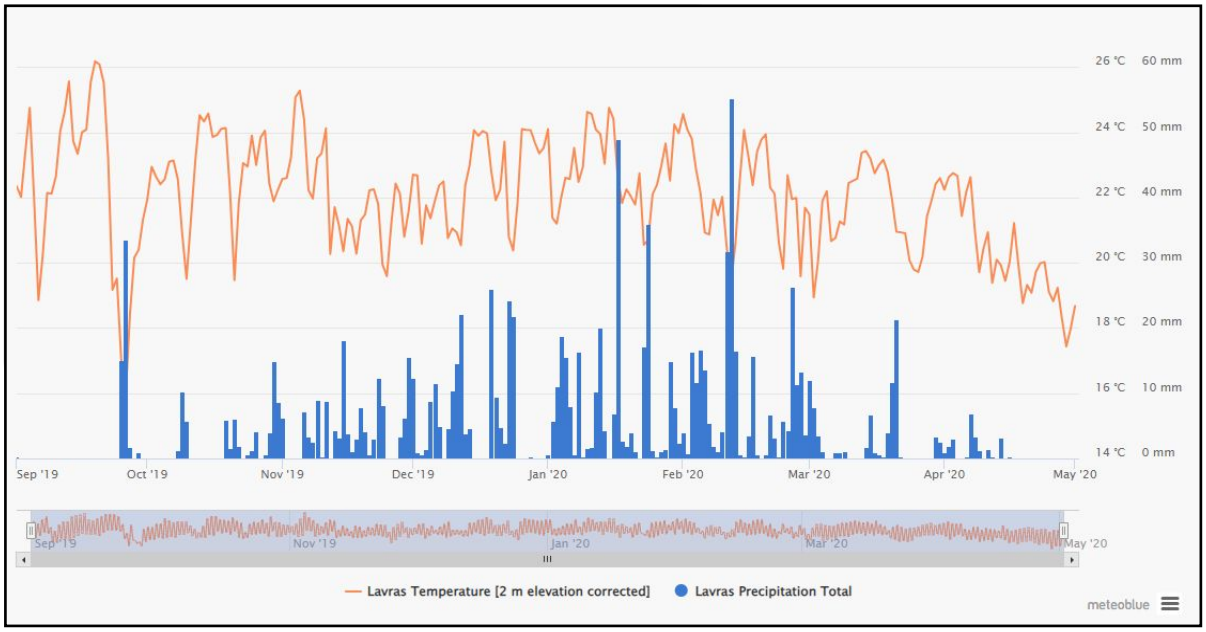
Additionally, the weather conditions during the period of the experiments are presented in Error: Reference source not found. These graphics provide information on the precipitation levels as well as the maximum, minimum, and mean temperatures recorded.

Figure 2 - Precipitation and mean temperature in the city of Lavras, MG. A the 2018/2019 growing season and B the 2019/2020 growing season.

A



B



Source: Adapted from METEOBLUE (2024).

The studies started in the 2017/2018 cropping season, with the selection of plants from the SYN1359S IPRO, P98Y11, BMX6160, 97R73 RR, NS7000 IPRO, and NA5909 cultivars to establish the genetic treatments. The specific characteristics of these cultivars are detailed in Table 1. Each cultivar was growed in a population scheme consisting of 500 plants per cultivar. From each population, 80 plants were pulled. Out of these 80 plants, a total of 47 were selected to progenies trials. These progenies were then evaluated alongside control samples across two cropping seasons, in the years 2018/2019 and 2019/2020.

Table 1 - Soybean cultivars used for obtaining the evaluated progenies, MG (maturity group)

Cultivars	MG	Launched Year	Resistances	Origin
SYN1359S IPRO	5.9	2014	Resistances: Glyphosate, Stem Canker; Cyst Nematode Race 3; some caterpillars	Syngenta
BMX6160	6.	2012	Moderately resistant: Phytophthora; Frog-eye Leaf Spot. Resistant: Stem	Brasmax

	0		Canker;	
NA5909	6.4	2008	Resistant: Glyphosate; Stem Canker; Frog-eye Leaf Spot; Bacterial Blight;	Nidera
NS7000 IPRO	6.7	2012	Resistant: Glyphosate; some caterpillars	Nidera
97R73 RR	7.3	2013	Resistant: Glyphosate;	Pioneer
P98Y11	8.1	2006	Resistant: Glyphosate; Cyst Nematode Races 1 and 3	Pioneer

Source: From the author (2024).

2.2. Experimental Design

The experiment conducted during the 2018/2019 cropping season utilized an incomplete block design (IBD), specifically a simple lattice arrangement of 17x17, resulting in a total of 288 treatments. These treatments included 282 progenies, in addition to 6 control varieties, which were SYN1359S IPRO, P98Y11, BMX6160, 97R73 RR, NS7000 IPRO and NA5909. The experimental plots were set up as single rows, with two meters in length, with a spacing of 50 centimeters between rows.

In the 2019/20 crop season, the progenies were split into two separate experiments based on their maturity group: early and late. The early experiment included 46 progenies from each of the early-maturing cultivars (SYN1359S IPRO, BMX6160, and NA5909), as well as the six control cultivars mentioned in Table 1. The late experiment comprised 46 progenies from each of the late-maturing cultivars (NS7000, P98Y11, and 97R73), along with the same six control cultivars listed in Table 1. Both experiments were conducted with three replications in a 12x12 lattice design. Each experimental plot consisted of two rows, with two meters in length.

2.3. Experiment Execution

The sowing was done manually using the no-tillage system (NTS). Fertilization included inoculation with *Bradyrhizobium japonicum*, which was applied via directed

injection into the planting furrows. For pest and disease control, insecticides and fungicides were utilized, with applications made as necessary to manage pest populations.

In the context of weed control after emergence, the herbicide glyphosate was sprayed at a rate of 2 L ha^{-1}. In cases of necessity, other herbicides were also applied. Mineral supplementation was applied according to the recommendations for the crop. Other cultural practices were carried out based on the crop's requirements and prevailing weather conditions (SOARES *et al*., 2015).

The following traits were measured:

• Grain Yield (YIELD): The yield is reported as kilograms per hectare, adjusted to a moisture content of 13%.

• Days to Flowering (DF): This represents the number of days from sowing to the R2 stage, at which point 50% of the plants exhibit full flowering.

• Full Maturity (FM): This is the number of days from sowing until the R8 stage (full maturity) is reached, defined as the point when 90% of the plants in the plot have attained this stage, according to the criteria set by Fehr and Caviness (1977).

• Plant Height (PH): The height of the plants was recorded at harvest. Three plants were randomly selected from each plot for measurement, which was taken from the base of the plant to the insertion point of the uppermost leaf (in centimeters).

2.4. DNA Extraction, Library Preparation, and Sequencing

Initially, the seeds of the 288 genotypes were sown in trays and placed in a growth chamber maintained at 25°C. The seeds were grown until the emergence of the first trifoliate and six leaf punches were collected from each genotype. The samples were immediately frozen and stored at -80°C to halt any metabolic activity and preserve their integrity. Subsequently, the samples were lyophilized (freeze-dried) to remove moisture, which further ensures their preservation for future analysis.

The extraction and genotyping steps were conducted in collaboration with GDM Seeds Inc., located in Cambé-PR. DNA extraction was performed using the KLEARGENE commercial kit, which involved a multi-step process to isolate and purify DNA. Initially, cell lysis was achieved using mechanical disruption with beads in a device known as the PaintShaker, which facilitates the breaking of cell walls. This step was followed by a chemical lysis phase, which included the addition of detergents and chaotropic salts to further disrupt cellular structures and release DNA.

Subsequent to lysis, DNA washing and purification steps were carried out to remove contaminants such as proteins, salts, and polysaccharides. This was done using specific buffers provided in the kit and 100% ethanol to ensure the DNA was clean and free of impurities. Finally, the purified DNA was eluted from the purification matrix with an elution buffer to obtain it in a usable form. The elution step was optimized to achieve an uniform DNA concentration of 10 ng/µl for all DNA samples, which is suitable for downstream genotyping applications. (KLEARGENE, 2022). A Chip for sequencing was used to target 1329 single nucleotide polymorphism (SNP) markers for genotyping. Sequencing was performed on the ION S5 sequencer from Thermo Fisher Scientific.

The amplification reaction included the AgriSeq Amplification Mix, Ion AmpliSeq™ Primer Pool specific to the SNPs of interest, and nuclease-free water to avoid any contamination from DNases or RNases. The thermocycling process for amplification involved several steps: enzyme activation, DNA strand denaturation, primer annealing, and DNA synthesis. Next, unique adapters or barcodes were ligated to the amplified DNA fragments to identify each sample uniquely. This step involved the addition of a Barcode Reaction Mix and another round of thermocycling.

After adapters bind, the library was normalized to ensure sequencing coverage across samples and a purification step was done to remove any non-amplified fragments. The purified and normalized libraries was prepared for loading onto the ION CHEF system, which automates template preparation and chip loading for the ION S5 sequencer.

The final step was sequencing those libraries on the Ion S5™ XL System, which can deliver up to 80 million reads per run. Bioinformatic analysis was then performed using Torrent Suite™ Software, which processed the sequencing data on a computer connected to the Ion Torrent™ server. The outcome of this analysis was a matrix in xlsx format, listing the genotypes alongside their corresponding SNP markers.

2.5. Phenotypic Data analysis

Data were analyzed adopting a mixed-model approach. The experiments from each year, categorized by maturity groups, were individually analyzed using model one, which incorporate the recovery of interblock information. Residual normality was assessed using the Shapiro-Wilk test (SHAPIRO and WILK, 1965), and homogeneity of variances across the experiments was evaluated using Hartley's maximum *F* test (HARTLEY, 1950).

(1)

$$\bar{y} = \mu + X_r \tau_r + X_t \tau_t + X_g u_g + X_b u_b + X_p u_p + \varepsilon$$

where:

$\bar{y}$: Observed value for the analyzed trait.

μ: constant associated with all observations.

$X_r \tau_r$: vector of replicate fixed effect.

$X_t \tau_t$: vector of checks or test fixed effect.

$X_g u_g$: vector of progenies effect (random), $g \sim N(0, I\sigma_g^2)$.

$X_b u_b$: vector of block effect aligned with replications (random), $b \sim N(0, I\sigma_b^2)$.

$X_p u_p$: vector of population effects the six cultivars (random), $p \sim N(0, I\sigma_p^2)$.

ε: vector of associated error effect (random), $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$.

Following the assessment for homogeneity of residual variances, a combined analysis across environments was done in accordance with the second model. The matrix of residual variances and covariances was configured with a diagonal structure as a response to identified heterogeneity within the dataset. The next phase involved modeling the variance and covariance matrix for the genotype-by-environment interaction. For this purpose, an extended form of the factor analytic (FA) structure was applied, guided by the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) to determine the most suitable model.

(2)

$$\bar{y} = \mu + X_r \tau_r + X_c \tau_c + X_t \tau_t + X_p u_p + X_a u_a + X_g u_g + X_{ga} u_{ga} + X_b u_b + \varepsilon$$

where:

$\bar{y}$: Observed value for the analyzed trait.

μ: constant associated with all observations.

$X_r \tau_r$: vector of replicate fixed effect.

$X_c \tau_c$: vector of maturity group fixed effect.

$X_t \tau_t$: vector of checks or test fixed effect.

$X_p u_p$: vector of population effects the six cultivars (random), $b \sim N(0, I\sigma_p^2)$.

$X_a u_a$: vector of environment effects the years (random), $a \sim N(0, I\sigma_a^2)$.

$X_g u_g$: vector of progenies effect (random), $g \sim N(0, I\sigma_g^2)$.

$X_{ga} u_{ga}$: vector of interaction effect progenies × environment (random), $ga \sim N(0, FA_1 \otimes I\sigma_{ga}^2)$.

$X_b u_b$: vector of block effect aligned with replications (random) $b \sim N(0, I\sigma_b^2)$.

ε: vector of effect of associated errors (random), $\varepsilon \sim N\left(0, \bigoplus_{j=1}^{j} I\sigma_{\varepsilon}^{2}\right)$¿¿.

Progeny, population, and total heritability were estimated using equations 3, 3.1, and 3.2, respectively.

(3)

$$h^2_{prog} = \frac{\hat{\sigma}^2_{prog}}{\hat{\sigma}^2_{prog} + \frac{\hat{\sigma}^2_{GxE}}{j} + \frac{\hat{\sigma}^2_{\varepsilon}}{jn}}$$

(3.1)

$$h^2_{pop} = \frac{\hat{\sigma}^2_{pop}}{\hat{\sigma}^2_{pop} + \hat{\sigma}^2_{prog} + \frac{\hat{\sigma}^2_{GxE}}{j} + \frac{\hat{\sigma}^2_{\varepsilon}}{jn}}$$

(3.2)

$$h^2_{total} = \frac{\hat{\sigma}^2_{pop} + \hat{\sigma}^2_{prog}}{\hat{\sigma}^2_{pop} + \hat{\sigma}^2_{prog} + \frac{\hat{\sigma}^2_{GxE}}{j} + \frac{\hat{\sigma}^2_{\varepsilon}}{jn}}$$

where:

h^2_{Prog}: heritability within populations using estimator.

h^2_{Pop}:heritability among populations using estimator.

h^2_{total}:heritability among populations using estimator.

n: number of replications;

j: number of environments;

$\hat{\sigma}^2_{Prog}$:mean additive genetic variance among populations.

$\hat{\sigma}^2_{GxE}$:mean additive genetic variance of the genotype by environment interaction.

$\hat{\sigma}^2_{\varepsilon}$: residual variance.

The accuracy of populations ($r_{g\hat{g}_{pop}}$), progenies ($r_{g\hat{g}_{prog}}$) and the overall accuracy ($r_{g\hat{g}_{total}}$¿ were calculated using estimators 4, 4.1, and 4.2, respectively.

(4)

$$r_{g\hat{g}_{pop}} = \sqrt{h^2_{pop}}$$

(4.1)

$$r_{g\hat{g}_{prog}} = \sqrt{h^2_{prog}}$$

(4.2)

$$r_{g\hat{g}_{total}} = \sqrt{h^2_{total}}$$

The coefficient of variation for experimental variance (CV_e) for each variable was calculated according to equation five.

(5)

$$CV_e = \frac{\sqrt{\sigma^2_\varepsilon}}{\overline{X}_i}$$

where:

CV_e: Experimental Coefficient of Variation.

σ^2_ε: residual variance.

$\overline{X}$: overall average of predicted values for Variable *i*.

The magnitude of genetic variation for the evaluated traits was measured at the progeny level (CV_{prog}¿ and at the population level (CV_{pop}¿, according to the estimators from equations 6 and 6.1, respectively.

(6)

$$CV_{prog} = \frac{\sqrt{\sigma^2_{prog}}}{\overline{X}}$$

(6.1)

$$CV_{pop} = \frac{\sqrt{\sigma^2_{pop}}}{\overline{X}}$$

where:

CV_{prog}: Progeny Genetic Coefficient of Variation.

CV_{pop}: Population Genetic Coefficient of Variation.

$\hat{\sigma}^2_{Prog}$: mean additive genetic variance among populations.

$\hat{\sigma}^2_{pop}$: mean additive genetic variance among populations.

$\overline{X}$: overall average of predicted values for Variable *i*.

2.6. Selection, Imputation, and Coverage of SNPs

The filtering and selection of SNPs is a critical step in genomic analysis. Monomorphic SNPs and those with a Minor Allele Frequency (MAF) below 5% were removed. The call rate threshold for individuals was set at 75%, where genotypes with a call rate lower than this threshold were removed. In the end, a total of 605 selected SNPs and 267 genotypes composed the genotypic data set.

For the imputation of missing data, the LD-kNNi algorithm was used, as per model 7 (MONEY *et al*., 2015).

(7)

$$d_l(s_1, s_2) = c + \frac{1}{n} \sum_{p \in L(p_i)} ¿\, g(s_1, p) - g(s_2, p) \vee ¿¿$$

where:

c: constant associated with the method (adopted as one)

$\frac{1}{n}$ normalizing term, where n is the number of SNPs in the summation $\sum_{p \in L(p_i)}$.

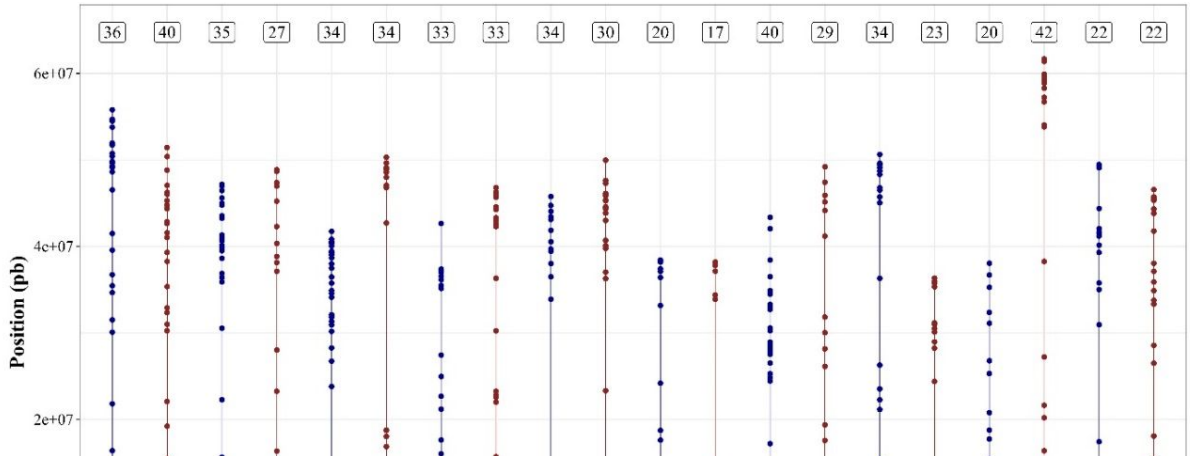
p is the position of SNP i.

$L(p_i)$: set of l SNPs in strong linkage disequilibrium with the SNP to be imputed.

$g(s_1, p)$ is the genotype of sample s at position p.

At the end of the filtering and imputation process, the distribution of SNPs across the 20 soybean chromosomes is presented in Figure 3 and this representation is based on SNPs positions and the actual size of the chromosome is not depicted in the image, only the equidistant positioning of the selected SNPs.

Figure 3 - Number and equidistance of SNPs across the 20 soybean chromosomes.



Source: From the author (2024).

2.7. Population Analyses

The expected heterozygosity (*He*), also known as genetic diversity, is a measure of genetic variation within a population and for multiple loci, the average heterozygosity of a population (*He*) can be calculated using estimator 8:

(8)

$$\overline{He}_i = \frac{1}{L}\sum_{j=1}^{L} H_{i(j)}$$

where:

$\overline{He}_i$: the estimate of heterozygosity at the (*j*)-th locus;

$H_{i(j)}$: the estimate of heterozygosity at the j-th locus;

L: the number of loci sampled.

The average observed heterozygosity, often denoted as (Ho), is calculated by directly observing the proportion of heterozygous individuals at each locus within a population and then averaging these proportions across all loci. The formula to estimate the average observed heterozygosity across multiple loci is presented in equation 9:

(9)

$$\overline{H_o} = 1 - \sum_{k}\sum_{i} P_{kii}/np$$

where:

P_{kii}: represents the proportion of homozygote *i* in sample *k*.

n: the number of samples;

p: frequency of allele k.

The genetic distance between populations was calculated using the Prevosti (1975) method and the resulting distance matrix was employe in a hierarchical clustering analysis using the unweighted pair group method with arithmetic mean (UPGMA). The starting point for the clustering was the smallest distance and this method was chosen due its cophenetic correlation, estimated by the Mantel test with 10,000 permutations. Model 10 was used to construct a dendrogram (FERREIRA, 2018).

The Mantel test measures the extent to which hierarchical clustering explains the original distance matrix and based on cophenetic correlation the UPGMA was superior to other methods (Figure 1.1).

(10)

$$\overline{d}_{(i,j)t} = \frac{d_{(i)t} + d_{(j)t}}{2}$$

where:

$\overline{d}_{(i,j)t}$: average dissimilarity between populations *i* and *j* considering population *t* (fusion point in the dendrogram);

$d_{(i)t}$: dissimilarity of population *i* considering population *t*;

$d_{(j)t}$: dissimilarity of population *j* considering population *t*.

Upon completion of the hierarchical clustering, the cutoff point for group formation was estimated using the Mojena method (MOJENA, 1977) as outlined in model 11.

(11)

$$PCM = \frac{\sum_{m} \overline{d}_{(i,j)t}}{m} + \sigma_{\overline{d}_{(i,j)t}} * k$$

where:

PCM: Mojena cutoff point;

$\sum_{m} \overline{d}_{(i,j)t}$: sum of the distances at the *m* fusion points in the dendrogram;

$\sigma_{\overline{d}_{(i,j)t}}$: standard deviation of the distances at the *m* fusion points in the dendrogram;

k: a constant, assumed to be 1.25 as suggested by Milligan and Cooper (1985).

To evaluate the genetic relationships among soybean lineages and the variation within populations, a pairwise distance matrix was calculated using the Pairwise method (PARADIS,

2011). The resulting genetic distance matrix was then used to conduct a principal coordinate analysis (PCA) to visualize the genetic variation. The PCA was plotted on a Cartesian plane, considering the first two principal components, where the specific explanatory capacity was determined by the eigenvalues. Confidence ellipses were added to the PCA plot, assuming a multivariate t-distribution at a 0.05 probability level.

2.8. Genomic Selection

The phenotypic data utilized were from two years, and the genotypic data were obtained from sequencing each line. To predict GEBV (Genomic Estimated Breeding Values), the GBLUP (Genomic Best Linear Unbiased Prediction) model outlined in model 12 (HABIER et al., 2007; VAN RADEN, 2008) was employed. Additionally, the values of BLUEs (Best Linear Unbiased Estimation) were used based on model 12 to avoid including the shrinkage effect twice.

(12)

$$\overline{y} = X_f \tau_f + X_g u_g + \varepsilon$$

where:

$\overline{y}$: Phenotypic BLUEs Vector.

$X_f \tau_f$: vector of fixed effects (intercept).

$X_g u_g$: matrix for individual additive genetic effects (random), $g \sim NM(0, G\sigma_g^2)$;

ε: Residual Variance.

For this model, the G matrix of individuals genetic effects is comprised by the variance-covariance matrix among individuals, also known as the kinship matrix. This matrix is estimated according to equation 13 (ENDELMAN and JANNINK, 2012).

(13)

$$G = \frac{WW'}{2\sum pq}$$

where:

G: Additive genomic relationship matrix.

W: Centered genotypic matrix for the markers.

pq: Allelic frequency of the markers.

The efficiency of the predictors model was assessed using predictive ability (equation 14), bias (equation 14.1), and Root Mean Square Error (RMSE) (equation 14.2).

These measures were used to evaluate the performance of the GBLUP model in predicting values and to compare it against other estimators. The GBLUP model was found to be the best based on these assessments.

(14)

$$\rho_{\hat{y}\,y} = \frac{\sigma_{\hat{y}\,y}}{\sqrt{\left(\sigma^2_{\hat{y}} * \sigma^2_{y}\right)}}$$

where:

$\rho_{\hat{y}\,y}$: Pearson correlation between BLUP and *GEBV*.

$\sigma_{\hat{y}\,y}$: Covariance between BLUP and *GEBV*.

$\sigma^2_{\hat{y}}$: Variance of BLUP.

σ^2_{y}: Variance of *GEBV*.

(14.1)

$$V_{\hat{y}\,y} = 1 - \beta_{BLUP;GEBV}$$

where:

$V_{\hat{y}\,y}$: bias between phenotypic BLUP and GEBV.

$\beta_{BLUP;GEBV}$: coefficient of the regression model 14.1.1.

¿.1.1)

$$\overline{y} = \beta_0 + \beta_i x_i + \varepsilon$$

where:

$\overline{y}$: predicted response variable (phenotypic BLUP).

β_0, β_i: coefficients of the regression model;

x_i: linear effect in the model for the predictor variable (GEBV).

ε: Residual variance.

(14.2)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (BLUP - GEBV)^2}{n}}$$

where:

RMSE: Root Mean Square Error between BLUP and GEBV

n: number of treatments.

2.9. Univariate and Multivariate Genotypic and Phenotypic Selection

To verify the efficiency in selecting the best individuals among and within populations, a selection intensity (SI) of 15% was applied to the 267 common progenies to both genomic and phenotypic selection, which resulted in the selection of 41 progenies among populations and seven progenies within the six studied soybean cultivars.

The expected selection gain was estimated by estimator 15, and the percentage of gain was estimated by estimator 15.1.

(15)

$$SG = \overline{X}_s - \overline{X}$$

(15.1)

$$PSG = \frac{SG}{\overline{X}_s} * 100$$

where:

SG: Selection Gain for trait *i*;

PSS: Percentage Selection Gain for trait *i*;

$\overline{X}_s$: average of the predicted values of the selected progenies.

Assuming that markers are spread throughout the soybean genome and that the application of genomic selection methods takes into account the selection of all markers used as input for the model, selection using a multivariate strategy for the studied traits was used to evaluate the efficiency of the selection. The selection index proposed by Mulamba and Mock (1978) was adopted, calculated by summing the points (MMI) derived from the adjusted phenotypic averages for each trait under consideration, according to equation 16.

(16)

$$MMI_{(i)} = \sum_{k-1}^{n} u_k r_{ik}$$

where:

$MMI_{(i)}$: value of the Mulamba and Mock index associated with genotype *i*;

u_k: the weight of the evaluated trait *k*;

r_{ik}: the rank associated with the average of genotype *i* relative to trait *k*

Furthermore, a selection index was utilized to rank progenies, favoring genotypes with reduced stature, shorter flowering time and maturity, and increased yield. These characteristics were considered equally important, and each trait was equally weighted in the selection process. The Mulamba and Mock Index (MMI) was employed to summarize the phenotypic data for each trait based on the Best Linear Unbiased Predictions (BLUPs). Subsequently, the resulting MMI values were used in model 12 to estimate the Genomic Estimated Breeding Values (GEBVs), and genotypes were selected based on these GEBVs.

2.10. Coincidence Analysis

To assess the consistency of the genotype rankings between treatments, Spearman's rank correlation coefficient (ρ) was calculated. The significance of the rank correlation was tested using a t-test at a 0.05 probability level. Additionally, the Coincidence Index (CI), as proposed by Hamblin and Zimmermann (1986), was computed to evaluate the agreement among the genotypes selected by estimator 17.

$$CI = \frac{(A - C)}{(M - C)} \tag{17}$$

where:

A: number of coincident progenies between the different analysis strategies.

M: number of selected progenies as a function of selection intensity (15%).

C: number of coincident progenies due to chance, the product being C x SI.

2.11. Computational Aspects

The R software (R Core Team, 2020) version 4.1.3 was employed for data handling and for the fitting of the utilized models. Data manipulation and the generation of the graphical representations presented in this study were facilitated by the tools available in the Tidyverse package (WICKHAM *et al*., 2019).

The AFEchidna package (ZHANG *et al*., 2021) was utilized for fitting mixed linear models to compute the variance-covariance matrix with an analytic factor structure. The imputation algorithm was executed using the LinkImpute software (MONEY *et al*., 2015). Population analyses were conducted with the aid of the packages adegenet v2.1.3 (JOMBART, 2008) and ape v5.6-2 (PARADIS and SCHLIEP, 2018).

3. RESULT AND DISCUSSION

For individual analysis, the results were split by cycle: early and late-maturing. Results of the individual analysis for early maturing in the crop season 2018/19 and 2019/20, across four traits, are presented in Appendix, Table 1.1. In the 18/19 season, a significant difference was observed for YIELD within the test, progenies, and population, as well as among the three unfolding populations. A similar pattern was observed for FM (full maturaty), although to a lesser extent in the test. Across the population, there was a significant difference for all traits in both growing years and within the early maturing group. During the 19/20 season, significant differences were observed for YIELD within the test, population, and notably within SYN1359. Additionally, for the plant height in progenies originating from NA5909, a significant difference was recorded (TABLE 1.1). The difference for DF (days to flowering) was identified exclusively within the NA5909 population.

For late maturing cultivars, the results of the four traits evaluated in both 2018/19 and 2019/20 years are presented in Appendix, Table 1.2. In the 18/19 season, a significant difference was observed for YIELD within the test, progenies, and population, as well as between the three late maturing cultivars. In the subsequent 19/20 season, significant differences were detected for all traits within the variety 97R73. Furthermore, at the progeny level, there was a significant difference for all traits. Notably, the variety 98Y11 exhibited a significant difference in yield and in plant height (Table 1.2).

The high total heritability (h^2_{total}) observed for most traits in late-maturing crops suggests a greater contribution of genetic variation relative to environmental variation, thereby indicating that selection could be effective at both the population and progeny levels. The progeny selection accuracy ($r_{g\hat{g}_{prog}}$), reflects the correlation between the true breeding values and the predicted breeding values for progenies, showed medium to high magnitudes on average. The range of accuracy for YIELD was from 0.35 in the EARLY 19/20 season to 0.82 in the LATE 18/19 season. This indicates that the accuracy of selecting progenies based on their yield performance varied between the early and late maturing groups and between the two crop seasons, with higher values observed in the late maturing group for the 18/19 season.

Population selection accuracy ($r_{g\hat{g}_{pop}}$), which measures the accuracy of selection at the population level, exhibited high magnitudes, with values ranging from 0.71 to 0.51. These estimates for the variables, in general were higher for the late-maturing group in both years, as

shown in Appendix, Tables 1.1 and 1.2. This suggests that the selection of populations based on their genetic potential was more precise in the late maturing group.

It is important to note that the accuracy of selection takes into account the variation between progenies, which includes more than just the experimental error quantified by the coefficient of variation (CV). Therefore, high values of F-statistic for the source of variation progenies correspond to increased accuracy and experimental precision associated with scientific observations. High F-statistic values indicate that a significant portion of the observed variation can be attributed to genetic differences between progenies, which is crucial for effective selection in breeding programs (LORENZO *et al*,. 2015).

The experimental precision varied due to different environments and evaluated traits. The estimates of CVe ranged from 1.13 for Full Maturity (FM), LATE 19/20 to 17.91 for YIELD, EARLY 18/19. The CVgprog is relatively low compared to CVgpop, suggesting more variation among populations than within for most of the traits. The experimental coefficient of variation (CVe) is lower than CVgpop, indicating that environmental variation is less than the genetic variation among populations. The experimental coefficient of variation (CVe) was generally lower suggesting that environmental conditions were relatively uniform and well-controlled.

In the multi-environment analysis, a significant variation was detected for progenies for the traits YIELD, PH, and FM (Ninou et al. (2022) found significant variation within improved commercial cultivars of durum wheat for grain yield and protein content. Similarly, significant intracultivar differences over two years and among three locations were observed for cotton yield. The same study also identified intracultivar variation for fiber quality traits such as length and micronaire, while fiber strength and uniformity showed no significant variation. Physiological traits like leaf carbon isotope discrimination, ash content, and potassium concentration also exhibited intracultivar variation (TOKATLIDIS et al., 2008).). Variation was detected among populations for all traits, as well as between early and late maturity cultivars. The variation among progenies within each population was significant for all traits for the late maturity cultivars 97R73 and NS7000. For the cultivars 98Y11, NA5909 and BMX6160, the variation was detected for PH, FM and DF. In addition, for SYN1359, the

variation was detected for all traits except for PH (Ninou et al. (2022) found significant variation within improved commercial cultivars of durum wheat for grain yield and protein content. Similarly, significant intracultivar differences over two years and among three locations were observed for cotton yield. The same study also identified intracultivar variation for fiber quality traits such as length and micronaire, while fiber strength and uniformity showed no significant variation. Physiological traits like leaf carbon isotope discrimination, ash content, and potassium concentration also exhibited intracultivar variation (TOKATLIDIS et al., 2008).).

Phenotypic variation within cultivars has been extensively reported in soybean research (AMARAL et al., 2019; ACHARD et al., 2020; FASOULA and BOERMA, 2005; FASOULA and BOERMA, 2007; TOKATLIDIS, 2015), corroborating the results of the present study. The presence of intracultivar phenotypic variation was also showed for different crops such corn, potato, wheat, cotton and barley (LIU *et al*., 2018; MARAND *et al*., 2019; NINOU *et al*., 2022; TOKATLIDIS *et al*., 2008).

Ninou *et al*. (2022) found significant variation within improved commercial cultivars of durum wheat for grain yield and protein content. Similarly, significant intracultivar differences over two years and among three locations were observed for cotton yield. The same study also identified intracultivar variation for fiber quality traits such as length and micronaire, while fiber strength and uniformity showed no significant variation. Physiological traits like leaf carbon isotope discrimination, ash content, and potassium concentration also exhibited intracultivar variation (TOKATLIDIS *et al.,* 2008).

"Grenache" and "Tempranillo" grape cultivars are good examples of how intracultivar variability has been widely explored in the wine industry. Even though both cultivars have Spanish origins, they are extensively cultivated in wine regions around the world (IBÁÑEZ *et al.*, 2012). "Tempranillo" and "Grenache" have 49 and 76 certified clones, respectively, including somatic mutations in both that have even resulted in white grape cultivars. Intravarietal selection programs are currently a useful tool for adapting vineyards to climate change, and Buesa *et al.* (2022) confirmed intracultivar variability in Water Use Efficiency (WUE) within the "Grenache" cultivar.

Heritability is a measure of the proportion of variation in a trait that can be attributed to genetic differences among individuals, as opposed to environmental factors. In the multi-environment analysis, the highest values of heritability and accuracy were observed for PH, suggesting that this trait is likely controlled by a small number of genes and the selection for this trait could be effective at the progeny level. In contrast, the coefficient of

variation due to environmental effects (CVe) was very low for FM, showing the greatest precision for this trait (TABLE 2). Low heritability and accuracy values were observed for YIELD, as expected, due to its high sensitivity to genotype-by-environment (G × E) interactions, which is consistent with the findings of Mendonça *et al*. (2020). The lower heritability for YIELD and FM indicates that improving these traits may require more complex strategies, such as the selection of specific genotypes that perform well under specific environmental conditions.

Besides the trait itself, the genetic structure of a population is also crucial in heritability estimation. In self-pollinated plants, we often see high heritability within a population due to homozygosity at many loci, reducing genetic variation for certain traits and making the individuals more genetically similar to each other (Ramalho *et al.,* 2012). The higher heritability within a population is present in Table 2.

Based on the genotypic dataset, the coefficient of parentage for the 267 progenies analyzed using the UPGMA clustering method highlights that the cultivars are closely related due to shared ancestry (FIGURE 4). For example, cultivars 97R73 and P98Y11 were grouped together because they both originated from the same breeding company (Corteva - Pioneer). Even though they have different maturity groups, the genetic background could be similar. In contrast, the cultivars NA5909 and NS7000 IPRO exhibit differences, and despite both originating from the same breeding company (Syngenta - Nidera), this variation may be due to the cultivars being derived from different relative maturity groups (RMGs). These findings agree with those reported by Mendonça *et al.* (2022), indicating that varieties from the same breeding company, particularly those from identical RMGs, have substantial genetic similarity.

Table 2 - Results for the multi-environment analysis for Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD) and Plant Height (PH).

SV	Effect	YIELD	PH	FM	DF
MS_{Test}	F	ns	ns	**	**
σ^2_{Prog}	R	**	**	**	ns
σ^2_{Pop}	R	**	**	**	**
σ^2_{Early}	R	**	**	**	**
$\sigma^2_{SYN\,1359}$	R	**	ns	**	**
$\sigma^2_{NA\,5909}$	R	ns	**	**	**
$\sigma^2_{BMX\,6160}$	R	ns	**	**	**
σ^2_{Late}	R	**	**	**	**
$\sigma^2_{97\,R73}$	R	**	**	**	**
$\sigma^2_{NS\,7000}$	R	**	**	**	**

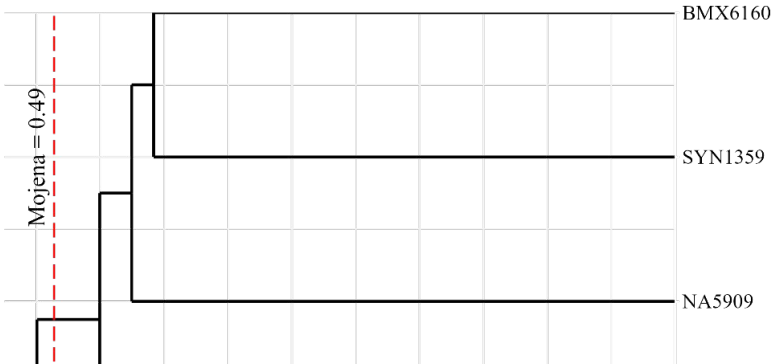
$\sigma^2_{98Y\,11}$	R	ns	**	**	**
σ^2_e	R	ns	ns	**	**
σ^2_{GxE}	R	**	ns	**	**
Mean	-	4655.08	76.44	115.13	45.19
h^2_{prog}	-	0.17	0.86	0.12	0.48
h^2_{pop}	-	0.18	0.96	0.86	0.96
h^2_{total}	-	0.30	0.99	0.88	0.98
$r_{g\hat{g}_{prog}}$	-	0.41	0.93	0.35	0.69
$r_{g\hat{g}_{pop}}$	-	0.39	0.98	0.93	0.98
$r_{g\hat{g}_{Total}}$	-	0.55	0.99	0.94	0.99
CV_{prog} (%)	-	3.49	3.42	0.36	0.92
CV_{pop} (%)	-	3.58	18.43	2.53	6.67
CV_e (%)	-	16.61	8.67	2.11	3.74

SV: Source of variation; MS_{Test}—¿ mean square of checks, σ^2_{Prog}—¿ progenies variance, σ^2_{Pop}—¿ population variance, Variance among ¿ , $\sigma^2_{NS\,7000}$, $\sigma^2_{98Y\,11}$, $\sigma^2_{SYN\,1359}$, $\sigma^2_{NA\,5909}$, $\sigma^2_{BMX\,6160}$ ¿ σ^2_e—¿ residual variance, DF and FM (days); PH (cm); YIELD (Kg ha-1); Wald test for fixed effects (F) and LRT (Likelihood Ratio Test) for random effects (R); "ns" indicates 'not significant', whereas asterisks (*) indicate levels of significance, with one asterisk for $p < 0.05$ and two asterisks for $p < 0.01$. h^2_{pop}: Heritability among populations; h^2_{prog}: heritability within populations modified; h²total: total heritability = heritability among + within populations modified; $r_{g\hat{g}_{prog}}$: accuracy on the progeny-mean basis; $r_{g\hat{g}_{pop}}$: accuracy on the population; $r_{g\hat{g}_{Total}}$: accuracy total; CV_{pop}: population coefficient of variation in percentage; CV_{prog}: coefficient of variation of progeny or progeny within population, in percentage terms; CV_e: experimental coefficient of variation in percentage terms.

Source: From the author (2024).

The cophenetic correlation coefficient is 0.92 and graphically presented in the appendix, Figure 1.1. It was found to be significant according to the Mantel test. A high cophenetic correlation coefficient suggests that the dendrogram accurately reflects the genetic distances among the progenies, with minimal distortion from the clustering process. In other words, the dendrogram reliably represents the actual genetic relationships within the population (SARAÇLI and DDOĞAN, 2013).

Figure 4 - Hierarchical Cluster Analysis Dendrogram of 267 Soybean Genotypes Based on Provesti's Absolute Genetic Distance, derived from the analysis of 605 SNP markers, with the Mojena test applied to determine the optimal number of clusters.

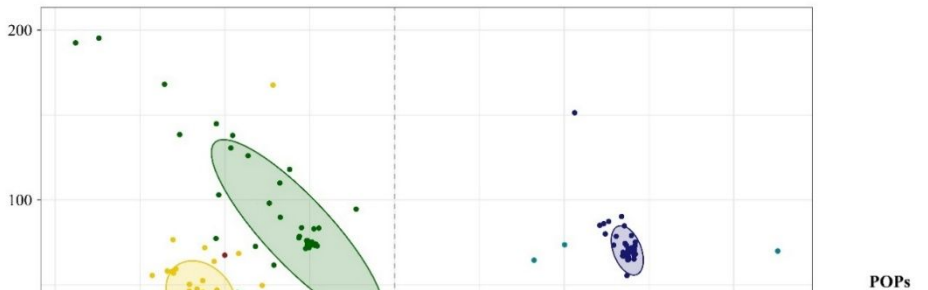


Source: From the author (2024).

Principal component analysis (PCA) based on SNP markers was employed to evaluate the structure of the populations. The first two principal components accounted for 47.11% of the total genetic variability (Figure 5). The analysis revealed six major clusters corresponding to the origins of the progenies from previous cultivars. A distinct separation of cultivars 97R73 and P98Y11 from the others corroborates with the dendrogram findings. Notably, the PCA provided a visualization of the high level of variation within the P98Y11 and NA5909 populations, as evidenced by the larger ellipses representing these groups. Conversely, the progenies derived from NS7000 exhibited the smallest variation. This intra-cultivar variation for P98Y11 and NA5909 can be explained by various mechanisms such as residual heterozygosity, mutations, transposable elements, epigenetic modifications, non-homologous recombination, and chromosomal mutations (TOKATLIDIS, 2015).

The use of PCA analyses for expressing variability has been widely used in different studies (ANDRIJANIĆ *et al*. 2023; NAFLATH T; RAJENDRAPRAS; RAVIKUMAR, 2023). Specifically, Andrijanić *et al*. (2023) found that the country of origin and maturity group were the most significant factors to explain the clusterization among European soybean cultivars and some countries also exhibited more variability than others.

Figure 5 - Principal Component Analysis (PCA) Scatter Plot of Pairwise Genetic Distances Among 267 Progenies Using 605 SNP Markers.



Source: From the author (2024).

The genetic variation among soybean progenies derived from six cultivars was assessed by calculating expected heterozygosity (He) values for each population. These He values ranged from 0.047725 in NS7000 to 0.124485 in P98Y11. Additionally, observed heterozygosity (Ho) values were determined for each population to assess the genetic variation within the soybean progenies. The Ho values varied from 0.032058 in NS7000 to 0.112448 in P98Y11, as shown in (Table 3). These findings are consistent with the PCA results, which indicated significant variation in cultivars P98Y11 and NA5909, as reflected by high He and Ho values, indicating a broad genetic base, which suggests potential for intracultivar selection. Conversely, the NS7000 population exhibited the lowest genetic diversity with an He value of 0.047725 and an Ho value of 0.032058 also evidenced by the low dispersion of data points in the PCA. This suggests a more uniform genetic structure within this population (LU *et al*., 2022).

Table 3 - Heterozygosity Metrics for Genetic Diversity Assessment within Populations using 605 SNP Markers. It includes information on Expected Heterozygosity (He), and Observed Heterozygosity (Ho).

Popu-lation	H_e	H_o
SYN1359	0.0929	0.0700
P98Y11	0.1245	0.1125
BMX6160	0.0483	0.0360
97R73	0.0702	0.0477
NS7000	0.0478	0.0321
NA5909	0.0952	0.0897

Source: From author (2024).

Several studies have demonstrated intracultivar variation using molecular tools such as RFLP (Restriction Fragment Length Polymorphism) and SSR (Simple Sequence Repeats). For instance, in sunflower, variability was assessed in four inbred lines using RFLP markers with 30 probe-enzyme combinations (ZHANG *et al.,* 1995). Heterogeneity at various levels was observed within three of the four lines studied. Yates *et al.* (2012) utilized SSR markers to analyze three soybean cultivars and confirmed heterogeneity in protein and oil content, as well as fatty acid composition, as previously reported (FASOULA and BOERMA, 2005). The majority of intracultivar SSR variation was attributed to residual heterozygosity, resulting in allele polymorphism. Additionally, six maize inbred lines obtained from eight different sources were examined for the level of genetic diversity among and within sources using SSR markers (GETHI *et al.,* 2002).

Although few reports have utilized SNP markers to assess intracultivar variation, a study of 36 cultivars using 5,346 SNPs revealed that the levels of intracultivar heterogeneity ranged from 0 to 10%. The highest levels of heterogeneity were observed in two cultivars: 'Essex' exhibited 6% heterogeneity, and 'Evans' exhibited 10% (ACHARD *et al*., 2020).

Mihelich *et al.* (2020) conducted a heterogeneity analysis of 20,087 *Glycine max* and *Glycine soja* accessions from the USDA Soybean Germplasm Collection (SGC). The

study identified high probability intervals of heterogeneity in 4% of the collection, corresponding to 870 accessions. However, the 'Williams 82' soybean accession showed no evidence of heterogeneity, in contrast to the within 'Williams 82' variation reported by Haun *et al.* (2011). The researchers proposed three explanations for the absence of intra-accession variation in 'Williams 82': 1. a genetic bottleneck causing a specific population homogeneity distinct from other varieties; 2. the genotyping sampling was based on three bulks of individuals and might have inadvertently included individuals with identical genotypes; and 3. the 'Williams 82' sample was derived from a single individual, which may not be representative of the accession's potential diversity. This lack of heterogeneity in 'Williams 82' suggests that similar processes could have also obscured the true genetic diversity in other accessions within the SGC.

The observed genotypic variation is consistent with phenotypic variation. Progeny derived from P98Y11 and NA5909 exhibited significant variations for PH, FM, and DF in multi-environment analysis. So, a considerable genetic diversity was found for a within P98Y11 and NA5909 population analysis. indicating substantial genetic diversity within the P98Y11 and NA5909 populations. Furthermore, at an individual level, they displayed considerable variation for yield and plant height.

In a breeding program, multiple F1 populations are generated through parental crossing, involving several rounds of self-mating and subsequent seed increase generations until distinct lines are established. Breeders typically begin the selection process in the F4 generation, although this may vary based on the breeding company's strategy (SILVA *et al*., 2017). It is important to note the frequency of heterozygotes is reduced by half with each generation. Therefore, in each (t) generation, the frequency of the locus with the genotype B^1B^2 will be $(1/2)^{t-1}$. Consequently, if the breeder pulls plants in the F_4 generation, the frequency of heterozygotes remaining is still 12.5%.

This percentage of segregating loci in F4, which subsequently become fixed for alternative alleles after successive cycles of self-pollination results in the presence of slightly different genetic lines in a commercial cultivar. These lines manifest as heterozygous SNP loci when profiled using bulk samples of plants from an individual cultivar and morphological variation inside a cultivar (MIHELICH *et al*., 2020).

One highlighted mechanism of intracultivar variation is remaining heterozygosity (HR). The process of self-fertilization leads to homozygosity and the reduced number of remained heterozygous loci still segregating and cause phenotypic variation. A study carried out by Fasoula, Yates, and Boerma (2012) using SSR markers in soybean cultivars

found 82% to 93% variation attributable to HR. However, even in 100% homogeneous lines, variation can be found. In corn, there is evidence of intracultivar variation reported in double haploid (DH) lines (BOGENSHUTZ and RUSSELL, 1986). These lines exhibited a significant accumulation of changes for quantitative traits, exceeding the mutation rates found in the literature. In tobacco crop, Lemos *et al*. (2023) showed greater variation in lines derived from DH compared to lines obtained through self-fertilization.

Besides that, intracultivar variation can results from mutation, intragenic recombination, unequal crossing over, DNA methylation, excision or insertion of transposable elements, and gene duplication (KIM, 2017; MORGANTE *et al*., 2005; SALGOTRA *et al*., 2023; SANDHU *et al*., 2017).

Given that phenotypic and genotypic variations have been observed across all populations, selection can be employed for soybean breeding. The desired outcome is to obtain cultivars with high grain yield and an early growth cycle, as well as a compact plant height to prevent lodging. Selection can be strategically applied to identify and advance the most promising individuals or to eliminate the least desirable ones. Consequently, 15 % of top-performing genotypes (UP +15%) and 15 % of bottom-performing genotypes (DOWN -15%) were implemented for both the estimation of genomic best linear unbiased predictors (G-BLUP) and phenotypic best linear unbiased predictors (BLUP). Finally, the comparisons of the phenotypic and genotypic predictors were carried out.

Furthermore, four distinct strategies were implemented: 1. Broad-based selection based on progenies without considering the population. 2. Selection conducted within each population. 3. Selection and discard using an index proposed by Mulamba and Mock that overlooks general aspects. 4. Selection and discard using an index proposed by Mulamba and Mock that is specific to each population.

For the general selection of top-performing genotypes, varied outcomes across different traits and predictors were observed. Under the BLUP method, a decrease in the selected mean individuals (SM) compared to the progenies mean (GM) was observed for plant height (PH), with a selection gain (GS) of -2.38 cm and a percentage selection gain (PSG) of -3.22%. The G-BLUP method showed a more pronounced decrease with a GS of -9.93 cm and a PSG of -13.52%. For days to flowering (DF), a slight reduction was observed in both BLUP and G-BLUP methods, with G-BLUP showing a more substantial decrease in SM and a larger negative GS and PSG of -5.28 cm and -14.04%, respectively. Minimal changes were noted for fruit maturity (FM) in the BLUP method, while the G-

BLUP method resulted in a GS of -5.48 days and a PSG of -5.09%. An increase in yield was observed, with the BLUP method showing a GS of 76.42 kg/ha and a PSG of 1.62%, while the G-BLUP method resulted in a more significant GS of 377.74 kg/ha and a PSG of 8.17% (Table 4).

For the -15% selection strategy, there was an increase in SM for both methods, with G-BLUP showing a substantial GS of 24.46 cm and a high PSG of 22.68% for the PH trait. For DF, both methods resulted in an increase in SM, with G-BLUP showing a GS of 2.92 days and a PSG of 6.37%. For FM, a slight increase in SM was observed for both methods, with G-BLUP showing a GS of 2.34 days and a PSG of 2.02%. A decrease in yield was noted for both methods, with G-BLUP showing a significant GS of -352.16 kg/ha and a PSG of -9.04% (Table 4).

Table *4* - General Selection for top-performing genotypes UP (+15%), a the bottom-performing genotypes DOWN (-15%), for the traits Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), Plant Height (PH), GM: Mean of Progenies, SM: Selected Mean Individuals, SG: Selection gain , PSG Percentage Selection Gain.

			BLUP		
Direction	**Trait**	**GM**	**SM**	**GS**	**PSG (%)**
UP (+15%)	PH	76.45	74.06	-2.38	-3.22
	DF	45.19	44.93	-0.26	-0.57
	FM	115.14	114.91	-0.23	-0.20
	YIELD	4654.59	4731.01	76.42	1.62
DOWN (-15%)	PH	76.45	78.83	2.39	3.03
	DF	45.19	45.50	0.30	0.67
	FM	115.14	115.39	0.25	0.22
	YIELD	4654.59	4568.30	-86.30	-1.89
			G-BLUP		
Direction	**Trait**	**GM**	**SM**	**GS**	**PSG (%)**
UP (+15 %)	PH	83.37	73.44	-9.93	-13.52
	DF	42.85	37.57	-5.28	-14.04
	FM	113.30	107.81	-5.48	-5.09
	YIELD	4247.52	4625.26	377.74	8.17
DOWN (-15%)	PH	83.37	107.84	24.46	22.68
	DF	42.85	45.76	2.92	6.37
	FM	113.30	115.64	2.34	2.02
	YIELD	4247.52	3895.37	-352.16	-9.04

Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), and Plant Height (PH) The measurement unit for each trait (days for DF and FM, cm for PH, Kg ha^{-1} for YIELD)

Source: From the author (2024).

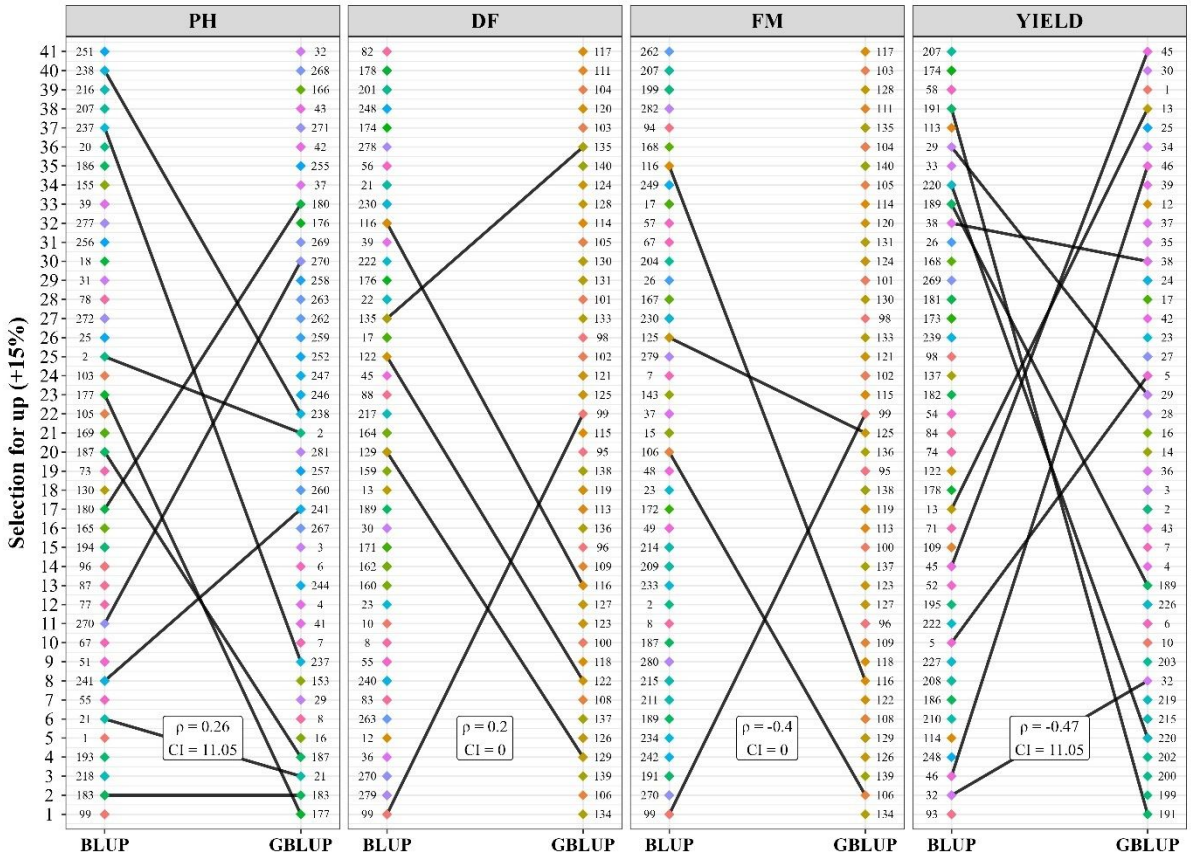
The highlight of 41 progenies selected by both BLUP and G-BLUP for overall selection, for UP +15% and DOWN -15% is presented in a Figure 6. The Coincidence Index proposed by Hamblin and Zimmermann ranged from 11.05% to 0 for UP +15%.

For DF and YIELD a negative Spearman correlation was observed, while for FM and PH, the correlation was positive. For DOWN -15%, the index indicated more coincidence, though in a negative sense, ranging from 13.92% to 0. For PH and DF, a negative Spearman correlation was noted, and for YIELD and FM, the correlation was positive. A positive correlation between BLUP and G-BLUP rankings means both methods agree on which progenies are top and bottom performers. A negative correlation would mean they disagree, ranking progenies differently in terms of performance. It has been reported by different authors, comparing multitrait with single trait genome prediction where multitrait had better results (GAIRE *et al*., 2022; SAPKOTA *et al*., 2020).

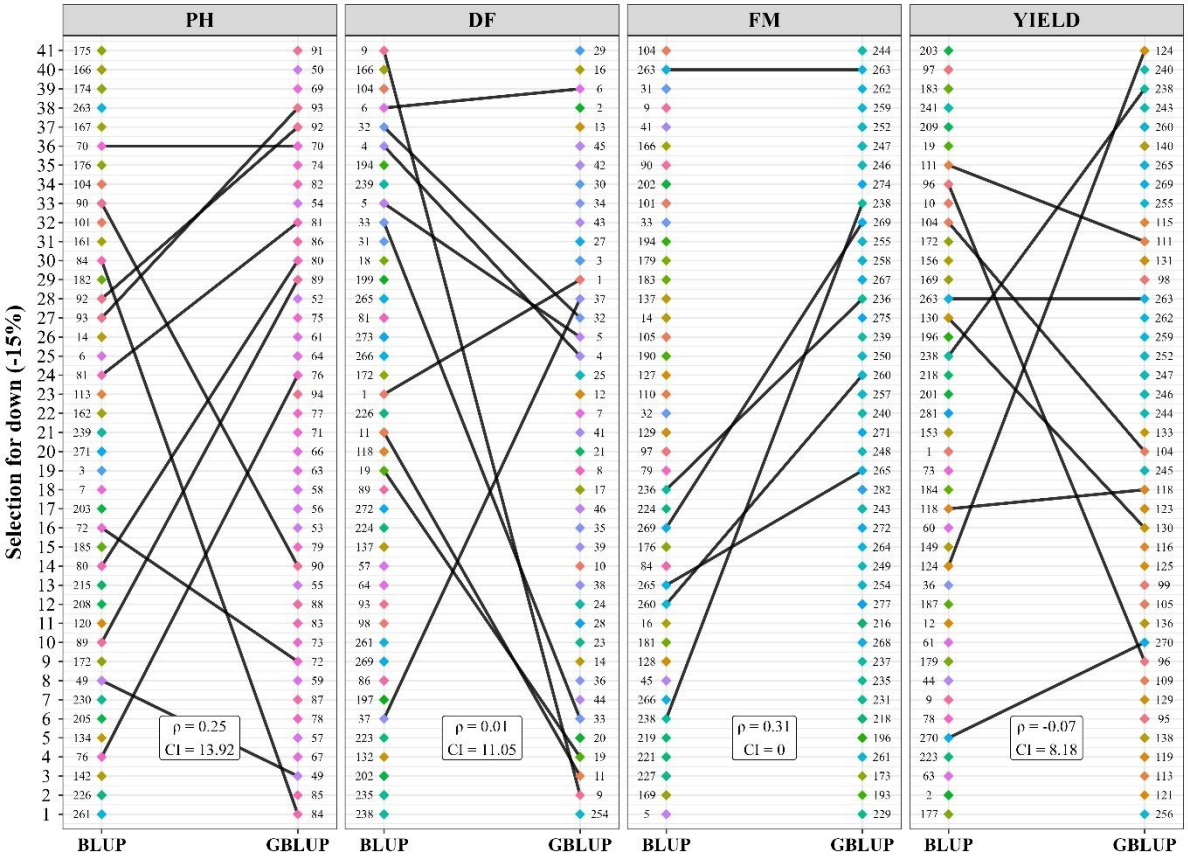
In the general selection, there was a difference in the number of progenies selected from each population when using the BLUP and G-BLUP methods. This can be observed in the appendix, Figura 1.2. Generally, BLUP selected almost the same amount of progenies across each population, while the number selected by G-BLUP varied, and some populations had no individuals selected.

Figure 6 - Coincidence of BLUP and G-BLUP for the generally selected progenies population, for top-performing genotypes UP (+15%), in A, the bottom-performing genotypes DOWN (-15%), in B.

A



B



Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), Plant Height (PH), CI Coincidence Index proposed by Hamblin and Zimmermann, ρSpearman's Correlation
Source: From the author (2024).

When considering the selection within all populations, for the trait PH both BLUP and G-BLUP methods resulted in a decrease in the selected mean individuals (SM) compared to the mean of progenies (GM). The percentage selection gain (PSG) ranged from -1.82% to -3.88% for BLUP and from -2.04% to -9.63% for G-BLUP. Notably, the population P98Y11 exhibited the largest decrease in SM with G-BLUP, with a PSG of -9.63% (The selection for FM showed minimal changes in SM across all populations with both methods. The PSG for BLUP ranged from -0.15% to -0.22%, and for G-BLUP, it ranged from -0.28% to -1.51%. The population 97R73 had the highest PSG under G-BLUP with -1.51%. In contrast to other traits, an increase in SM was observed for YIELD in all populations. The PSG for BLUP ranged from 1.28% to 1.85%, while for G-BLUP, it ranged from 0.92% to 5.48%. The population NA5909 showed the most substantial increase under G-BLUP with a PSG of 5.48% (Table 5).).

For the trait DF, all populations showed a reduction in SM with both selection methods. The PSG for BLUP ranged from -0.39% to -0.76%, while for G-BLUP, it ranged from -0.63% to -3.51%. The population 97R73 had the most significant reduction under G-BLUP with a PSG of -3.51%.

The selection for FM showed minimal changes in SM across all populations with both methods. The PSG for BLUP ranged from -0.15% to -0.22%, and for G-BLUP, it ranged from -0.28% to -1.51%. The population 97R73 had the highest PSG under G-BLUP with -1.51%. In contrast to other traits, an increase in SM was observed for YIELD in all populations. The PSG for BLUP ranged from 1.28% to 1.85%, while for G-BLUP, it ranged from 0.92% to 5.48%. The population NA5909 showed the most substantial increase under G-BLUP with a PSG of 5.48% (The selection for FM showed minimal changes in SM across all populations with both methods. The PSG for BLUP ranged from -0.15% to -0.22%, and for G-BLUP, it ranged from -0.28% to -1.51%. The population 97R73 had the highest PSG under G-BLUP with -1.51%. In contrast to other traits, an increase in SM was observed for YIELD in all populations. The PSG for BLUP ranged from 1.28% to 1.85%, while for G-BLUP, it ranged from 0.92% to 5.48%. The population NA5909 showed the most substantial increase under G-BLUP with a PSG of 5.48% (Table 5).).

Table 5 - Selection Within Population for top-performing genotypes UP (+15%). For the traits Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), Plant Height (PH), GM: Mean of Progenies, SM: Selected Mean Individuals, SG: Selection gain , PSG Percentage Selection Gain.

		BLUP				G-BLUP			
Trait	**POP**	**GM**	**SM**	SG	**PSG (%)**	**GM**	**SM**	SG	**PSG (%)**
PH	97R73	76.70	73.99	-2.71	-3.67	76.36	72.82	-3.55	-4.87
PH	BMX6160	76.36	74.00	-2.35	-3.18	89.15	86.32	-2.84	-3.29
PH	NA5909	76.41	74.40	-2.01	-2.70	74.95	73.45	-1.50	-2.04
PH	NS7000	76.20	74.13	-2.07	-2.79	78.45	77.05	-1.40	-1.82
PH	P98Y11	76.72	73.86	-2.87	-3.88	106.97	97.57	-9.40	-9.63
PH	SYN1359	76.31	74.20	-2.11	-2.85	74.93	72.90	-2.03	-2.78
DF	97R73	45.17	44.97	-0.21	-0.46	42.85	41.40	-1.45	-3.51
DF	BMX6160	45.18	44.95	-0.23	-0.51	37.74	37.05	-0.69	-1.85
DF	NA5909	45.18	44.89	-0.29	-0.66	43.79	42.73	-1.06	-2.49
DF	NS7000	45.19	45.01	-0.18	-0.39	44.46	44.18	-0.28	-0.63
DF	P98Y11	45.20	44.98	-0.23	-0.50	42.52	42.06	-0.46	-1.10
DF	SYN1359	45.22	44.88	-0.34	-0.76	45.63	44.63	-1.00	-2.25
FM	97R73	115.13	114.96	-0.17	-0.15	113.63	111.94	-1.69	-1.51
FM	BMX6160	115.15	114.90	-0.25	-0.22	107.97	107.34	-0.63	-0.58
FM	NA5909	115.13	114.90	-	-0.20	115.32	113.80	-1.52	-1.34

FM	NS7000	115.10	114.85	0.23 - 0.25	-0.22	114.86	114.53	-0.32	-0.28
FM	P98Y11	115.15	114.97	0.18 - 0.22	-0.16	113.35	112.75	-0.60	-0.53
FM	SYN1359	115.15	114.93		-0.19	114.54	113.66	-0.88	-0.78
YIELD	97R73	4648.67	4736.09	87.42	1.85	4221.55	4282.68	61.13	1.43
YIELD	BMX6160	4657.13	4724.17	67.04	1.42	3923.10	4021.00	97.90	2.43
YIELD	NA5909	4650.32	4710.57	60.26	1.28	3988.61	4219.94	231.33	5.48
YIELD	NS7000	4662.01	4742.65	80.64	1.70	4544.14	4586.22	42.08	0.92
YIELD	P98Y11	4656.84	4726.63	69.79	1.48	4217.32	4281.40	64.08	1.50
YIELD	SYN1359	4652.81	4735.77	82.96	1.75	4594.32	4680.15	85.84	1.83

Source: From the author (2024).

The selection gains observed for grain yield across different populations indicates the selection efficiency regards the employed methods. Under the BLUP method, all populations exhibited positive selection gains, with 97R73 achieving 87.42 kg/ha of yield increase. For G-BLUP method, the results were variable but remained positive and the 97R73 population achieved a selection gain of 61.13 kg/ha, lower than the BLUP method. Conversely, BMX6160 exhibited a more pronounced gain of 97.90 kg/ha under G-BLUP, surpassing the BLUP method performance. Notably, NA5909 displayed a remarkable selection gain of 231.33 kg/ha under the G-BLUP method; the highest among all populations and methods, which can be due to a big genetic variation (The selection for FM showed minimal changes in SM across all populations with both methods. The PSG for BLUP ranged from -0.15% to -0.22%, and for G-BLUP, it ranged from -0.28% to -1.51%. The population 97R73 had the highest PSG under G-BLUP with -1.51%. In contrast to other traits, an increase in SM was observed for YIELD in all populations. The PSG for BLUP ranged from 1.28% to 1.85%, while for G-BLUP, it ranged from 0.92% to 5.48%. The population NA5909 showed the most substantial increase under G-BLUP with a PSG of 5.48% (Table 5).).

It is important noticed the populations with more variation, such as P98Y11, generally exhibit higher gains, especially for the trait PH. This is even more evident when using genomic prediction methods. On the other hand, NS7000 presented lower PSG% for most of the traits. These results demonstrate that genetic variability is crucial for the success of breeding programs, as it provides the potential for plant breeding strategies (GOVINDARAJ *et al*., 2015).

When discarding of genotypes (DOWN – 15%) strategy of selection was employed for PH, the PSG values ranged from 2.37% to 4.30% for BLUP and from 1.83% to 4.58% for G-BLUP, indicating a general increase in PH across populations. The highest gain using BLUP was observed in population 97R73 (4.30%), while G-BLUP achieved its highest value in the same population (4.58%). In addition, for DF, the PSG values were positive for both methods, with BLUP ranging from 0.41% to 1.12% and G-BLUP

ranging from 0.33% to 2.96%. (Tokatlidis et al. (2008) employed similar analysis for cotton and the results of the study provided evidence of significant variation within elite cotton cultivars, which is likely due to latent and/or newly developed variation. In general selection, it is more likely to select progenies originated from one population than another due to some populations already having a higher potential as presented in the appendix, Figure 1.2. The same was observed in Durum wheat cultivars, where in general selection, 75% of progenies were from the cultivar Maestrale, and 25% from Svevo (NINOU et al., 2022).).

The highest PSG for BLUP was found for SYN1359 population (1.12%), while G-BLUP had its highest value for BMX6160 population (2.96%). For FM, the PSG values were also positive but smaller, with BLUP values ranging from 0.14% to 0.27%, while the G-BLUP values ranged from 0.25% to 1.01%. The highest PSG for BLUP was observed in BMX6160 population (0.27%), and for G-BLUP, it was in population 97R73 (1.01%). However, the YIELD trait showed a negative PSG across all populations for both BLUP and G-BLUP, indicating a reduction in yield for the bottom-performing genotypes. The PSG values for BLUP ranged from -1.47% to -2.26%, with the largest decrease found for population SYN1359 (-2.26%). The PSG values for G-BLUP method ranged from -1.10% to -3.44%, with the most significant decrease also in population SYN1359 (-3.44%) (Tokatlidis et al. (2008) employed similar analysis for cotton and the results of the study provided evidence of significant variation within elite cotton cultivars, which is likely due to latent and/or newly developed variation. In general selection, it is more likely to select progenies originated from one population than another due to some populations already having a higher potential as presented in the appendix, Figure 1.2. The same was observed in Durum wheat cultivars, where in general selection, 75% of progenies were from the cultivar Maestrale, and 25% from Svevo (NINOU et al., 2022).).

Comparing general selection with within-population selection is important, as the strategy to identify variation within a population can be exploited. General selection provides a good strategy for selection. However, by examining each population individually, the intracultivar variation becomes evident. Tokatlidis *et al*. (2008) employed similar analysis for cotton and the results of the study provided evidence of significant variation within elite cotton cultivars, which is likely due to latent and/or newly developed variation. In general selection, it is more likely to select progenies originated from one population than another due to some populations already having a higher potential as presented in the appendix, Figure 1.2. The same was observed in

Durum wheat cultivars, where in general selection, 75% of progenies were from the cultivar Maestrale, and 25% from Svevo (NINOU *et al*., 2022).

When evaluating selection within populations, the concordance between BLUP and G-BLUP varied across different populations. For top-performing genotypes, the range of concordance ranged from 0 to 49.58%. Notably, the Spearman correlation coefficient was significant and positive (Figure 7) for both P98Y11 population (PH trait; 0.86*) and 93R73 population (DF trait; 0.86*). In contrast, for the bottom-performing genotypes, the concordance was generally low among the populations, with values ranging from 0 to 32.77%. The Spearman's correlation was also significant and positive for P98Y11 population (FM trait; 0.82*) (Error: Reference source not found).

The enhanced coincidence for PH and DF traits can be attributed to increased heritability estimation within populations (h^2_{prog}) and improved accuracy on the progeny-mean basis ($r_{g\hat{g}_{prog}}$). High coincidence index and strong Spearman correlation coefficients between the selection strategies BLUP and G-BLUP were noticed (FIGURES 6 and 7). Volpato *et al*. (2018) reinforced the reliability of the methodologies employed and with a selection proportion set at 20%, the coincidence index estimated for grain yield and absolute maturity were approximately 70% and 90%, respectively.

Table *6*- Selection Within Population the bottom-performing genotypes DOWN (-15%). for the traits Days to Flowering (DF), Full Maturity (FM), Grain Yield in kg/ha(YIELD), Plant Height (PH), GM: Mean of Progenies, SM: Selected Mean Individuals, SG: Selection gain , PSG Percentage Selection Gain.

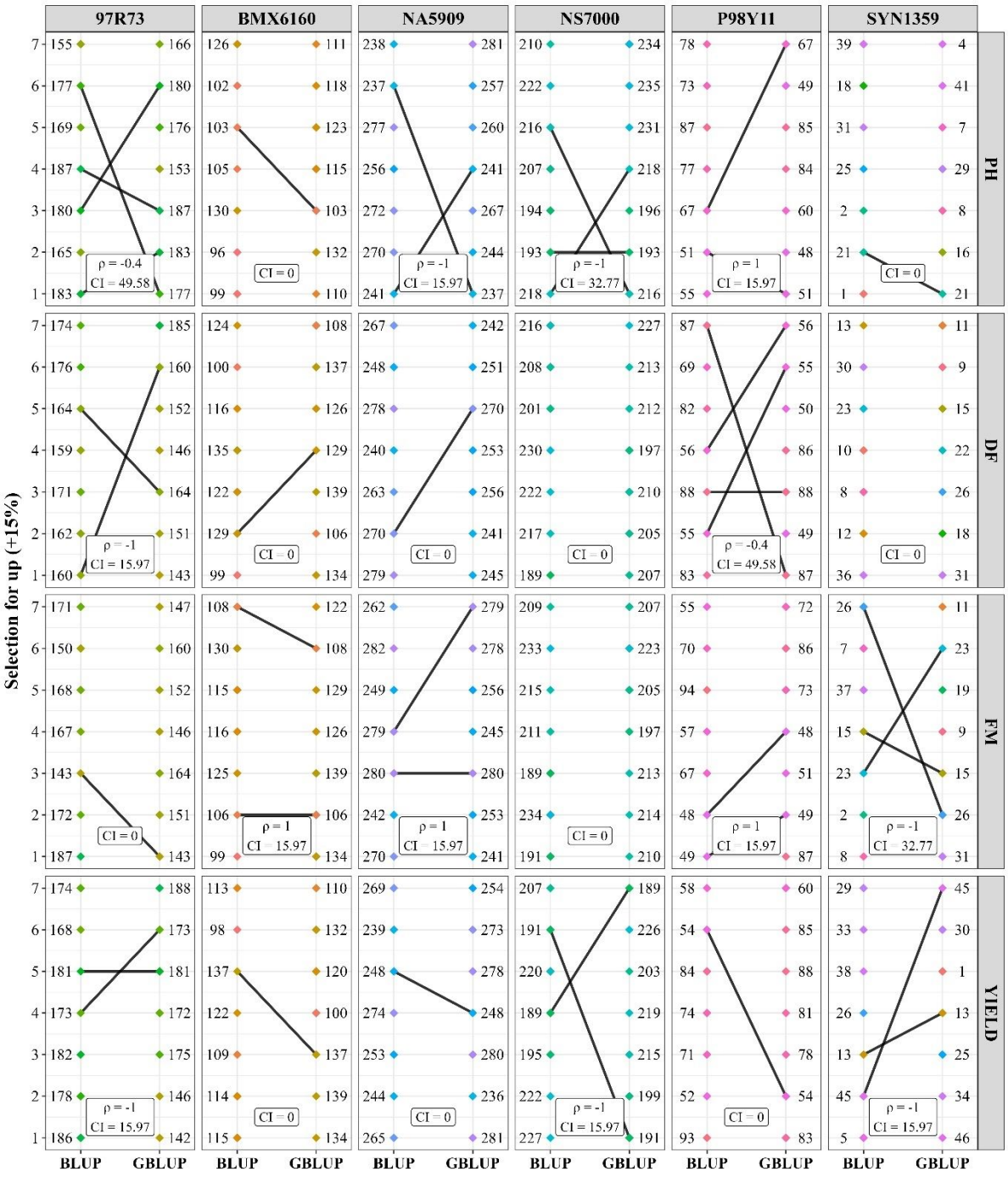
		BLUP				G-BLUP			
Trait	**POP**	**GM**	**SM**	SG	**PSG (%)**	**GM**	**SM**	SG	**PSG (%)**
PH	97R73	76.70	80.15	3.45	4.30	76.36	80.03	3.67	4.58
PH	BMX6160	76.36	78.39	2.04	2.60	89.15	90.81	1.66	1.83
PH	NA5909	76.41	78.26	1.85	2.37	74.95	77.62	2.67	3.44
PH	NS7000	76.20	78.11	1.91	2.44	78.45	80.13	1.68	2.10
PH	P98Y11	76.72	79.08	2.35	2.98	106.97	111.97	5.00	4.47
PH	SYN1359	76.31	78.21	1.90	2.43	74.93	77.21	2.28	2.95
DF	97R73	45.17	45.44	0.27	0.59	42.85	43.94	1.09	2.47
DF	BMX6160	45.18	45.44	0.26	0.58	37.74	38.89	1.15	2.96
DF	NA5909	45.18	45.44	0.26	0.57	43.79	44.48	0.69	1.55
DF	NS7000	45.19	45.43	0.24	0.52	44.46	44.61	0.15	0.33
DF	P98Y11	45.20	45.39	0.19	0.41	42.52	43.04	0.52	1.20
DF	SYN1359	45.22	45.73	0.51	1.12	45.63	46.16	0.53	1.15
FM	97R73	115.13	115.35	0.22	0.19	113.63	114.79	1.16	1.01
FM	BMX6160	115.15	115.46	0.31	0.27	107.97	109.03	1.07	0.98
FM	NA5909	115.13	115.37	0.23	0.20	115.32	116.17	0.84	0.73
FM	NS7000	115.10	115.34	0.24	0.21	114.86	115.14	0.29	0.25
FM	P98Y11	115.15	115.31	0.16	0.14	113.35	113.75	0.39	0.35
FM	SYN1359	115.15	115.45	0.30	0.26	114.54	114.98	0.44	0.38
YIELD	97R73	4648.67	4558.40	-90.27	-1.98	4221.55	4169.80	-51.75	-1.24
YIELD	BMX6160	4657.13	4585.92	-71.21	-1.55	3923.10	3880.29	-42.81	-1.10
YIELD	NA5909	4650.32	4583.12	-67.19	-1.47	3988.61	3892.77	-95.83	-2.46
YIELD	NS7000	4662.01	4582.11	-79.91	-1.74	4544.14	4487.60	-56.54	-1.26
YIELD	P98Y11	4656.84	4565.17	-91.67	-2.01	4217.32	4164.54	-52.78	-1.27
YIELD	SYN1359	4652.81	4550.21	-102.61	-2.26	4594.32	4441.39	-152.93	-3.44

Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), and Plant Height (PH) The measurement unit for each trait (days for DF and FM, cm for PH, Kg ha^{-1} for YIELD)

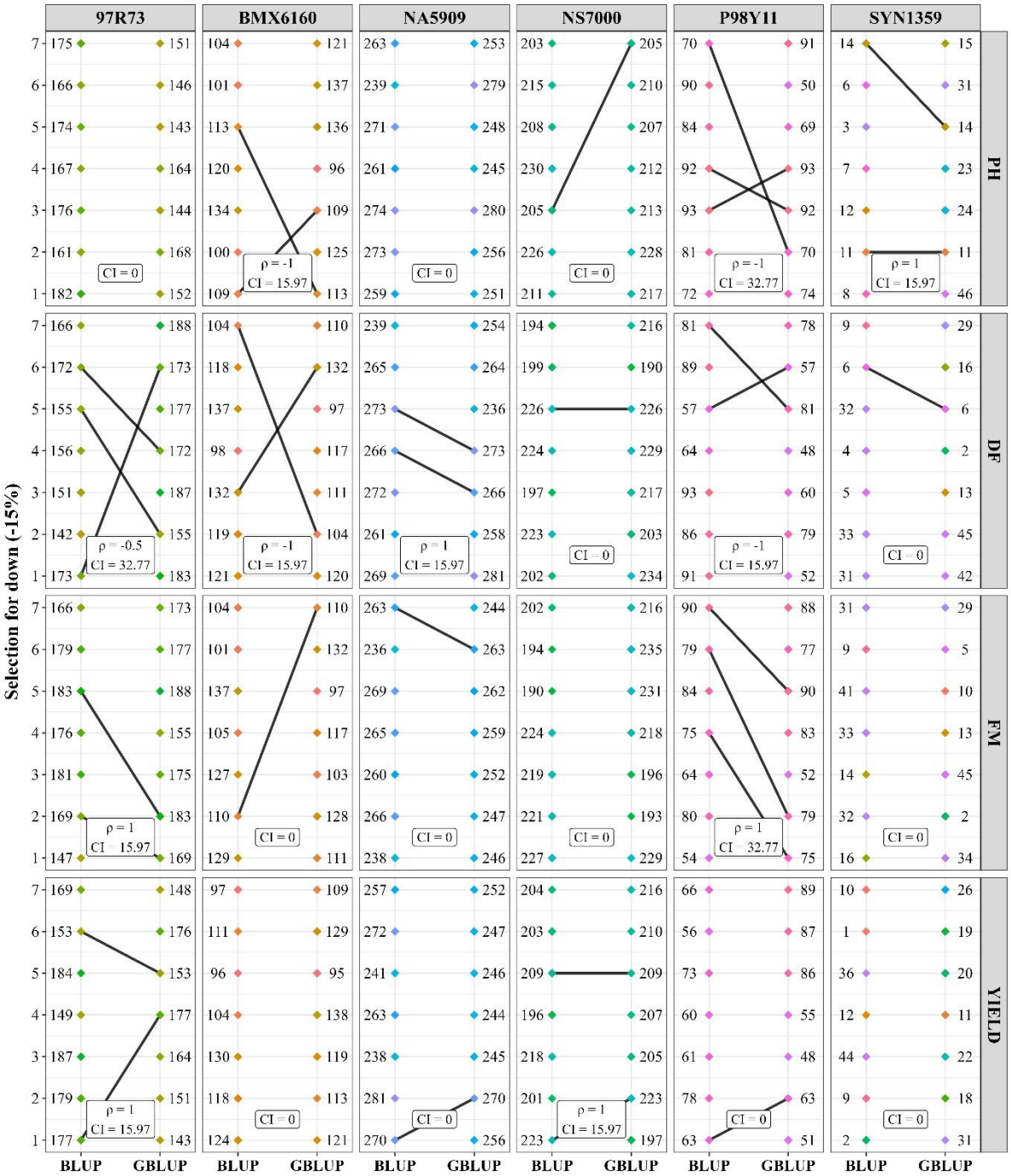
Source: From the author (2024).

Figure 7 - Coincidence of BLUP and G-BLUP for the selected progenies within each population, for top-performing genotypes UP (+15%), and the bottom-performing genotypes DOWN (-15%), for the traits Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), and Plant Height (PH), CI Coincidence Index proposed by Hamblin and Zimmermann, ρSpearman's Correlation.

A



B



Source: From the author (2024).

For the top-performing genotypes, the multitrait selection index proposed by Mulamba and Mock (1978) revealed substantial gains. Under BLUP, the Progenies Mean (GM) was 536.00, and the Selected Mean Individuals (SM) was 677.43, resulting in a Selection Gain (SG) of 141.43, which translates to a Percentage Selection Gain (PSG) of 20.88%. In contrast, under G-BLUP, the SG was lower at 40.82, with a corresponding PSG of 7.08%. This indicates that while both methods resulted in positive gains, BLUP was markedly more effective for the top-performing genotypes (Table 7).

For the bottom-performing genotypes, negative selection gains were observed, as expected. With BLUP, the GM was 536.00, and the SM was 390.62, leading to an SG of -145.38 and a PSG of -37.22%. G-BLUP also resulted in a negative SG of -69.90 and a PSG of -15.00%. Although both methods showed a decrease in performance for the bottom-performing genotypes, the decline was more pronounced with BLUP. The application of the General Selection Index for top-performing genotypes showed that BLUP had a significantly higher selection gain compared to G-BLUP (Table 7).

Table 7 - A multi-trait analysis of the General Selection index proposed by Mulamba and Mock (1978), for top-performing genotypes UP (+15%), the bottom-performing genotypes DOWN (-15%), GM: Mean of Progenies, SM: Selected Mean Individuals, SG: Selection gain , PSG Percentage Selection Gain.

	BLUP				G-BLUP			
Direction	GM	SM	SG	PSG (%)	GM	SM	SG	PSG (%)
UP (+15%)	536.0	677.4	141.4	20.8	536.0	576.8	40.8	7.0
DOWN (-15%)	536.0	390.6	-145.3	-37.2	536.0	466.1	-69.9	-15.0

Source: From the author (2024).

The coincidence of BLUP and G-BLUP for the General Selection Index proposed by Mulamba and Mock (1978), for top-performing genotypes (UP +15%) and bottom-performing genotypes (DOWN -15%), is presented in Figure 8. For the top performers, a modest Coincidence Index (CI) of 5.31% was noted, accompanied by a low Spearman's correlation, which was inclined towards negative values. For the bottom-performing genotypes, when genotypes are discarded, G-BLUP and BLUP presented a high value of CI of 79.91%.

Based on these results, it can be inferred that genomics may be particularly useful in discarding poor-performing genotypes, as there is a high coincidence between BLUP and G-BLUP. So, negative selection (the removal of undesirable genotypes), can be confidently carried out using both traditional phenotypic evaluations (BLUP) and genomic evaluations

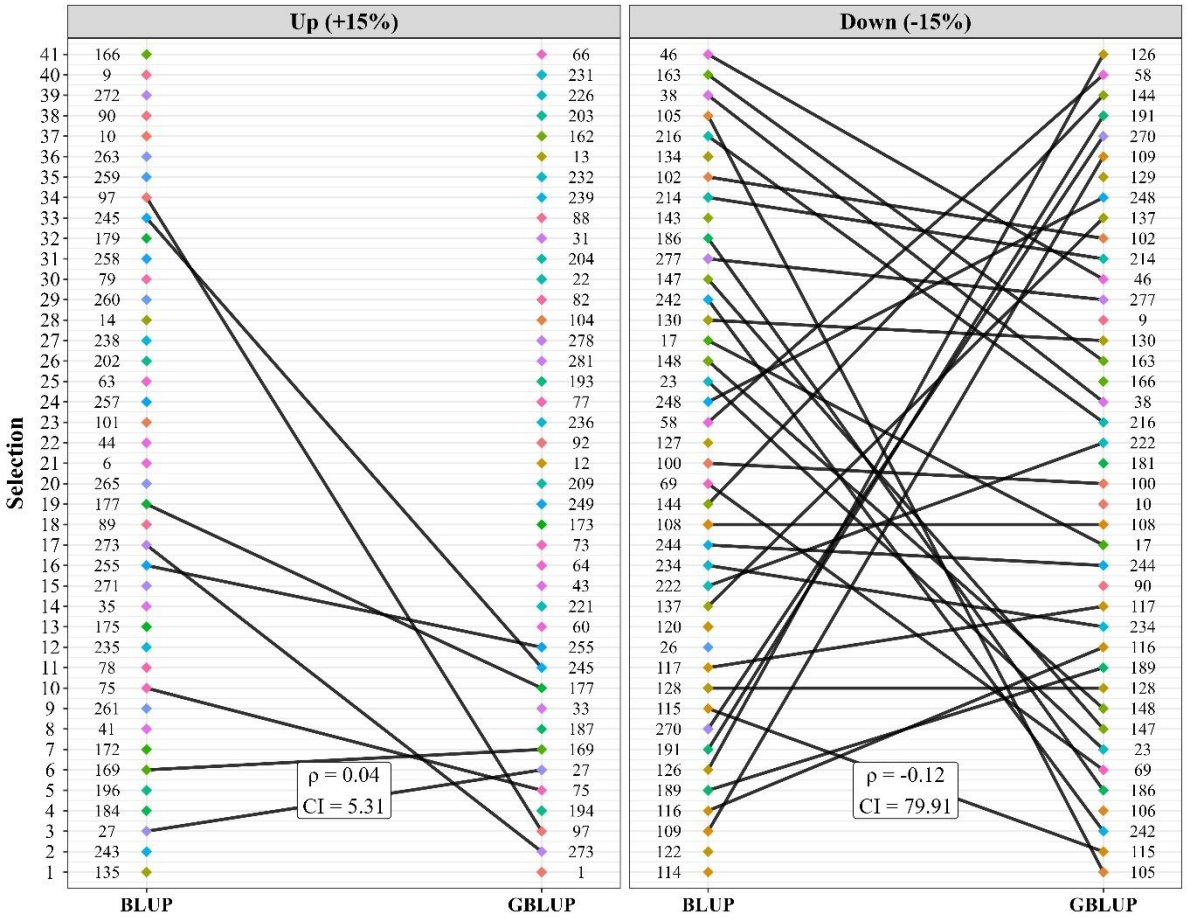
(G-BLUP). This approach is consistent with the strategies of seed companies that utilize genomic selection (GS) early in their breeding programs to weed out the least desirable genotypes (CROSSA *et al*., 2017; JEON *et al*., 2023).

The discrepancy in coincidence between BLUP and GBLUP in the work could be influenced by several factors, as highlighted by Zhang *et al*. (2020) for chickens' immune traits. Different heritability estimations were observed between BLUP and GBLUP. Whether genomic data does not capture all the genetic variance for the traits of interest or whether the markers used are not in strong linkage disequilibrium with the causal genes, this could result in lower coincidence between the methods. Finally, the genetic architecture of the traits under study may be more complex and traits controlled by many genes with small effects or those significantly influenced by the environment might show more variation in predictions between BLUP and GBLUP. (BUDHLAKOTI *et al*. 2022).

Another reason for the lack of coincidence between BLUP and GBLUP is the variation in genotype performance observed across different crop season, as evidenced by the shifting genotype rankings, underscores the significant impact of the crop year on grain yield outcomes. Notably, the genotype x environment (GxE) interaction was significant for nearly all traits, suggesting this interaction is a major factor influencing the YIELD presented on Table 4. Bianchi *et al*. (2020), also found that complex GxE interactions were primarily responsible for the variation seen in soybean progenies. These interactions result in progenies performing inconsistently across various environments, with some excelling only under specific conditions. This inconsistency is a challenge for breeders who aim to select genotypes that consistently perform well across a range of environments, as highlighted by Ramalho *et al*. (2021) and Gesteira *et al*. (2018).

Genotype-by-environment interaction (GxE) has more pronounced impact on plants than in animal breeding, complicating the prediction of plant performance for the following year. The integration of genomic selection (GS) with GxE interaction sparks a range of opinions and debates (ROBERTSEN *et al*., 2019). According to Mendonça *et al*. (2020), in general, G×E models did not demonstrate superior performance compared to mean principal models. Moreover, they found that a training population consisting solely of closely related progenies yielded the highest prediction accuracy for soybean. Conversely, some studies have shown that models incorporating correlations between environments, along with genomic information, attained greater predictive accuracy. This was particularly true in scenarios with a significant number of missing environment (KRAUSE *et al*., 2020).

Figure 8 - Coincidence of BLUP and G-BLUP for General Selection index proposed by Mulamba and Mock (1978), for top-performing genotypes UP (+15%), and bottom-performing genotypes DOWN (-15%), CI Coincidence Index proposed by Hamblin and Zimmermann, ρSpearman's Correlation.



Source: From the author (2024).

The analysis of the selection index for top-performing genotypes (UP +15%) demonstrated that the BLUP methodology yielded higher selection gains (SG) across all populations when compared to G-BLUP. The percentage selection gain (PSG) was also consistently higher under BLUP, indicating a more pronounced improvement in the selected traits. For instance, population 97R73 exhibited a notable SG of 136.92 and a PSG of 20.34% under BLUP, whereas G-BLUP resulted in a lower SG of 39.67 and a PSG of 6.94%.

Conversely, the selection index for bottom-performing genotypes (DOWN -15%) showed negative SG values for both methodologies, as expected due to the selection against undesirable traits. However, the magnitude of reduction was greater under BLUP, with more substantial negative PSG values, suggesting a more aggressive decrease in the frequency of

these genotypes. For example, population BMX6160 showed an SG of -123.81 and a PSG of -37.01% under BLUP, compared to an SG of -53.79 and a PSG of -11.62% under G-BLUP.

The results indicate that BLUP is more effective in amplifying positive traits among top-performing genotypes, while also more aggressively reducing undesirable traits among bottom-performing genotypes. G-BLUP, on the other hand, appears to moderate the selection gains, resulting in less extreme changes in both directions (Table 8). A similar result was observed by Sánchez-Mayor *et al*. (2022) when using the random model BLUP yielded double the selection gain compared to genomic prediction.

		BLUP				G-BLUP			
Direction	**POP**	GM	SM	SG	PSG (%)	GM	SM	SG	PSG (%)
UP (+15%)	97R73	536.1	673.0	136.9	20.3	532.1	571.7	39.6	6.9
	BMX6160	458.3	635.5	177.1	27.8	516.7	564.4	47.7	8.4
	NA5909	575.3	719.9	144.5	20.0	540.5	577.6	37.0	6.4
	NS7000	529.2	641.4	112.2	17.4	540.6	582.1	41.5	7.1
	P98Y11	562.5	669.5	106.9	15.9	546.7	579.9	33.2	5.7
	SYN1359	552.1	683.7	131.5	19.2	538.9	576.7	37.8	6.5
DOWN (-15%)	97R73	536.1	416.2	-119.8	-28.7	532.1	465.6	-66.4	-14.2
	BMX6160	458.3	334.5	-123.8	-37.0	516.7	462.9	-53.7	-11.6
	NA5909	575.3	416.0	-159.3	-38.3	540.5	476.6	-63.9	-13.4
	NS7000	529.2	396.3	-132.8	-33.5	540.6	474.6	-66.0	-13.9
	P98Y11	562.5	447.1	-115.4	-25.8	546.7	499.7	-47.0	-9.4
	SYN1359	552.1	421.5	-130.6	-31.0	538.9	472.4	-66.4	-14.0

Table 8 - A multi-trait analysis of the within the population selection index proposed by Mulamba and Mock(1978), for top-performing genotypes UP (+15%), and bottom-performing genotypes DOWN (-15%), GM: Mean of Progenies, SM: Selected Mean Individuals, SG: Selection gain , PSG Percentage Selection Gain.

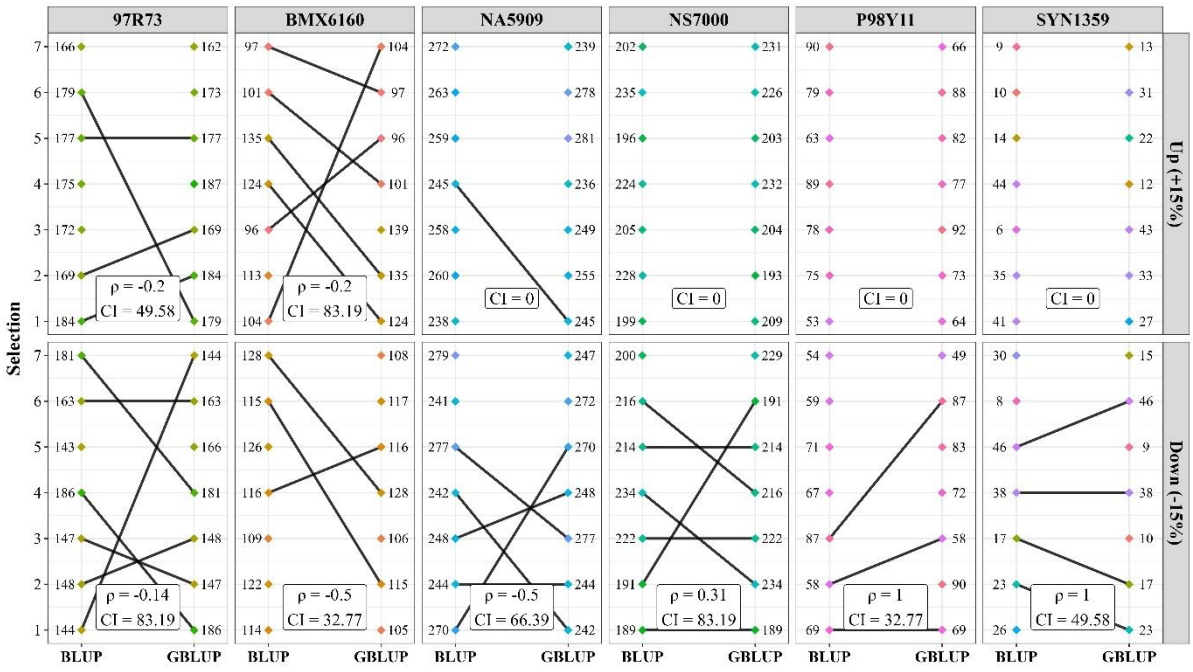
Source: From the author (2024).

When considering selection within populations, the coincidence of BLUP and G-BLUP depends on the population. For top-performing genotypes, the coincidence was high

for 97R73 population with a positive Spearman correlation, and for BMX6160 with a negative Spearman correlation. However, no coincidence was detected in other populations (Error: Reference source not found9). For the bottom-performing genotypes, a high coincidence was observed across all populations, ranging from 32.77% to 83.19%. Significant and positive Spearman correlation was noted for the population SYN1359 (0.86*), while for the other populations, it was not significant.

Residual variation within soybean cultivars was uniform when Benning, Cook, and Haskell cultivars grew up according to common agronomic practices and was sufficient to allow up to seven new morphologically and agronomically distinct cultivars to be selected from individual plants when planting densities were significantly reduced (BOERMA, 2007; FASOULA and BOERMA, 2005; FASOULA and BOERMA, 2005).

Figure 9 - Coincidence of BLUP and G-BLUP for within-population selection index proposed by Mulamba and Mock (1978), for top-performing genotypes UP (+15%), and bottom-performing genotypes DOWN (-15%), CI Coincidence Index proposed by Hamblin and Zimmermann, ρSpearman's Correlation.



Source: From the author (2024).

All the selection strategies employed in the study generated significant results for selecting top-performing genotypes and discarding bottom-performing ones. However, The General Selection Index (Mulamba and Mock, 1978) when applied with BLUP demonstrated

the most considerable gains for the top performers, indicating its potential as the most effective method for enhancing desirable traits.

The selection index proposed by Mulamba and Mock in 1978 has been used in the field of plant breeding, providing a robust statistical framework for the concurrent enhancement of multiple traits within a breeding population. Its significance is rooted in its capacity to amalgamate diverse phenotypic traits assigned a weight based on economic importance and genetic parameters-into a singular composite index. This streamlines the process of identifying elite individuals who contribute to the overall genetic advancement of a crop. In a similar way by Gesteira *et al*., (2018), on the selection of early soybean inbred lines using multiple indices, underscores the efficacy of the Standardized Multiple Selection Index (SMSI) in pinpointing lines that exhibit both phenotypic stability and exceptional performance. The SMSI, which evaluates essential attributes such as grain yield, absolute maturity, oil content, and protein content, has demonstrated its value as a selection tool.

The investigation of intracultivar variation in soybeans has disclose several key insights that have implications for soybean breeding and cultivation. The potential for selection within a cultivar stands out as a promising avenue for enhancing desirable traits while maintaining the overall genetic elite background of the cultivar. This potential for selection within a soybean cultivar also raises important considerations in the context of plant variety protection. To qualify for protection under plant variety protection laws, a new cultivar must typically meet the criteria of Distinctness, Uniformity, and Stability (DUS). These criteria are essential for a cultivar to be legally recognized as a new variety distinguishable from existing ones, which is crucial for intellectual property rights and the motivation of breeding programs. The requirements for DUS testing can vary by country, reflecting the nuances of national agricultural policies and plant variety protection systems. For example, the Brazilian Service for Cultivars Protection (SNPC) requests information of 38 characteristics to assess whether a new soybean cultivar meets the DUS criteria (ACHARD *et al*. 2020). On the other hand, the Brazilian laws about this issue, describe an "essentially derived cultivar" as a plant variety that is primarily developed from a parental cultivar, retaining its essential genetic characteristics while being distinguishable from it. Such a cultivar maintains the core traits of its progenitor but also exhibits unique modifications from the selection process (BRASIL, 1997).

Moreover, the issue of line maintenance over extended periods highlights the importance of stringent seed production and monitoring practices. In essence, for seed production, only different plants are removed, thus rendering the technique more of rouging

than a selection process (PARLEVLIET, 2007). As genetic integrity is paramount for cultivar performance, the ongoing efforts to maintain purity must be as robust as the initial breeding programs that developed these lines. The subtle changes that may occur over generations emphasize the need for vigilance and the adoption of advanced genetic tools to ensure the consistency and reliability of soybean seeds.

Another perspective that emerges when considering intracultivar variation is the multiline cultivars. This strategy aims to enhance yield stability and bolster resilience against both biotic and abiotic stresses. For several soybean breeding programs, the bulk method is commonly employed to manage segregating populations through to the F_3 or F_4 generations. This approach is also used during the evaluation of progenies within families. As a result, the cultivars that are developed under these conditions are, in essence, a composite of multiple lines rather than a single, pure line (TOKATLIDIS, 2015). Furthermore, a study conducted by Carneiro *et al*. (2019) on soybeans using a multiline approach has demonstrated that such cultivars are markedly stable. Their performance exhibits reduced sensitivity to genotype-environment interactions (GxE), which is a significant advantage. According to Vilela *et al*., (2024), the employment of multiline cultivars is highlighted as an effective strategy for managing Asian soybean rust (ASR). The study demonstrated that the use of multiline in soybean provided agronomic performance comparable to the best pure line tested and yielded stable results. Also, the use of resistant cultivars and multiline is effective in reducing the severity of ASR. This stability, coupled with the potential for enhanced disease resistance, makes the multiline approach highly relevant in today's agricultural landscape, which is increasingly influenced by climate change and the imperative for sustainable practices.

4. CONCLUSION

The study reveals the existence of both phenotypic and genotypic intracultivar variation among the assessed cultivars. The degree of variation observed differs, with cultivars P98Y11 and NA5909 exhibiting higher levels of diversity, while NS7000 presents a lower level of variation.

The potential for both selection and discarding is present across all evaluated traits individually and also when employing a multi-trait index. However, the correlation between genotypic and phenotypic predictions for the purpose of selection is low. In contrast, this correlation is high for discarding, suggesting genotypic data is more applicable for identify and remove undesirable progenies.

Furthermore, both plant height (PH) and days to flowering (DF) traits showed a high consistency between genotypic and phenotypic predictions, indicating they are useful in the selection and discarding processes within a multi-trait index. These results emphasize the importance of considering trait-specific predictability when applying genomic information to breeding strategies and validate the multi-trait index as a useful tool for improving the efficiency of breeding programs.

REFERENCES

ACHARD, F. *et al*. **Single nucleotide polymorphisms facilitate distinctness-uniformity-stability testing of soybean cultivars for plant variety protection**. Crop Science, v. 60, p. 2280-2303, 2020. DOI: 10.1002/csc2.20201.

AMARAL, L. de O. *et al*. **Pure line selection in a heterogeneous soybean cultivar**. Crop Breeding and Applied Biotechnology, v. 19, n. 3, p. 277-284, 2019.

ANDRIJANIĆ, Z. *et al*. **Genetic diversity and population structure of European soybean germplasm revealed by single nucleotide polymorphism.** Plants (Basel), v. 12, n. 9, p. 1837, 29 Apr. 2023. DOI: 10.3390/plants12091837.

BERNARDO, R. **Breeding for quantitative traits in plants**. 2. ed. Woodbury: [s.n.], 2010. 400 p.

BIANCHI, M. C. *et al*. Heritability and the genotype×environment interaction in soybean. **Geosciences; Environment,** v. 3, n. 1, p. e20020, 2020. ISSN 2639-6696.

BRASIL. **Lei nº 9.456, de 25 de abril de 1997**. Dispõe sobre a proteção de cultivares e dá outras providências. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 28 abr. 1997. Seção 1, p. 8354.

BUDHLAKOTI, N. *et al*. **Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development of Climate-Resilient Crops**. Frontiers in Genetics, v. 13, Feb 9, 2022. Article 832153. DOI: 10.3389/fgene.2022.832153.

CARNEIRO, A. K. *et al*. **Stability analysis of pure lines and a multiline of soybean in different locations**. Crop Breeding and Applied Biotechnology, v. 19, n. 4, p. 395-401, 2019. DOI: 10.1590/1984-70332019v19n4a56.

CHUNG, Y. S. *et al*. Genotyping-by-Sequencing: a Promising Tool for Plant Genetics Research and Breeding. **Horticulture, Environment, and Biotechnology**, v. 58, n. 5, p. 425-431, 2017.

CROSSA, J. *et al*. **Genomic Selection in Plant Breeding: Methods, Models, and Perspectives.** Trends in Plant Science, v. 22, n. 11, p. 961-975, Nov. 2017. DOI: 10.1016/j.tplants.2017.08.011.

FASOULA, V. A.; BOERMA, H. R. **Divergent selection at ultra-low plant density for seed protein and oil content within soybean cultivars**. Field Crops Research, v. 91, p. 217-229, 2005. DOI:10.1016/j.fcr.2004.07.018.

FERREIRA, D. F. **Estatística multivariada**. 3ª ed. Lavras: Editora UFLA, 2018.

GAIRE, R. *et al*. **Multi-trait genomic selection can increase selection accuracy for deoxynivalenol accumulation because of fusarium head blight in wheat**. Plant Genome, v. 15, 2022. p. e20188. DOI: 10.1002/tpg2.20188.

GESTEIRA, G. D. S. *et al*. **Selection of early soybean inbred lines using multiple indices**. Crop Science, p. 2494-2502, 2018.

GETHI, J. G. *et al*. **SSR variation in important U.S. maize inbred lines**. Crop Science, v. 42, p. 951-957, 2002. DOI:10.2135/cropsci2002.0951.

GHANEM, H. Ben *et al*. **Exploiting intra-cultivar variation to select for Barley yellow dwarf virus-PAV (BYDV-PAV) resistance in barley**. Canadian Journal of Plant Science, v. 98, n. 4, p. 930-946, 2018. DOI: 10.1139/cjps-2017-0364.

GOVINDARAJ, M.; VETRIVENTHAN, M.; SRINIVASAN, M. **Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives**. Genetic Research International, v. 2015, p. 431487, 2015. DOI: 10.1155/2015/431487.

HABIER, D. *et al*. **The impact of genetic relationship information on genome-assisted breeding values**. Genetics, v. 177, n. 4, p. 2389-2397, 2007.

HAMBLIN, J.; ZIMMERMANN, M. J. D. O. J. P. B. R. **Breeding common bean for yield in mixtures.** v. 4, p. 245-272, 1986.

HARTLEY, H. O. J. B. **The maximum F-ratio as a short-cut test for heterogeneity of variance**. v. 37, n. 3/4, p. 308-312, 1950. ISSN 0006-3444.

HAUN, W. J. *et al*. **The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82**. Plant Physiology, v. 155, n. 2, p. 645–655, 2011.

JEON, D. *et al*. **Digitalizing breeding in plants: A new trend of next-generation breeding based on genomic prediction**. Frontiers in Plant Science, v. 14, 1092584, 2023. DOI: 10.3389/fpls.2023.1092584.

JOMBART, T**. Adegenet: a R package for the multivariate analysis of genetic markers.** Bioinformatics, v. 24, n. 11, p. 1403-1405, 2008. DOI:10.1093/bioinformatics/btn129.

KLEARGENE. **KlearGene Plant Plate 384 DNA extraction kits**. Kleargene, p. 3000, 2022.

KRAUSE, M. D. *et al*. **Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models**. Crop Science, [s.l.], 19 Aug. 2020. DOI: 10.1002/csc2.20253.

LIU, N. *et al*. **Intraspecific variation of residual heterozygosity and its utility for quantitative genetic studies in maize**. BMC Plant Biology, v. 1, p. 1–15, 2018.

LORENZO, J. C. *et al*. **Coefficient of Variation Can Identify the Most Important Effects of Experimental Treatments**. Notulae Botanicae Horti Agrobotanici Cluj-Napoca, v. 43, n.3, 2015. DOI: 10.15835/nbha431988.

LU, Y. *et al*. **High genetic diversity and low population differentiation of a medical plant Ficus hirta Vahl., uncovered by microsatellite loci: implications for conservation and breeding**. BMC Plant Biology, v. 22, n. 334, 2022. DOI: 10.1186/s12870-022-03734-2>. Accessed on: Apr. 1, 2023.

MARAND, A. P. *et al*. **Residual heterozygosity and epistatic interactions underlie the complex genetic architecture of yield in diploid potato**. Genetics, v. 212, p. 317–332, 2019.

MENDONÇA, H. C. *et al*. **Genetic diversity and selection footprints in the genome of Brazilian soybean cultivars.** Frontiers in Plant Science, v. 13, 842571, 2022. DOI: 10.3389/fpls.2022.842571.

MENDONÇA, L. F. *et al*. **Genomic prediction enables early but low-intensity selection in soybean segregating progenies**. Crop Science, v. 60, p. 1-16, 2020. DOI: 10.1002/csc2.20072. Access on: 11 Feb. 2024.

MILLIGAN, G. W.; COOPER, M. C. **An examination of procedures for determining the number of clusters in a data set**. Psychometrika, v. 50, p. 159-179, 1985.

MONEY, D. *et al*. **LinkImpute: fast and accurate genotype imputation for nonmodel organisms**. G3 Genes|Genomes|Genetics, v. 5, n. 11, p. 2383-2390. DOI:10.1534/g3.115.021667

MORGANTE, M. *et al*. **Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize**. Nature Genetics, v. 37, p. 997-1002, 2005. DOI: /10.1038/ng1615.

NAFLATH, T. V. *et al*. **Evaluation of diverse soybean genotypes for seed longevity and its association with seed coat color**. Scientific Reports, v. 13, n. 4313, 2023. DOI: 10.1038/s41598-023-31071-3. Access on: 11 Feb. 2023.

NINOU, E. *et al*. **Utilization of Intra-Cultivar Variation for Grain Yield and Protein Content within Durum Wheat Cultivars**. Agriculture, v. 12, n. 5, p. 661, 2022. Available at: https://www.mdpi.com/2077-0472/12/5/661/htm. Access on: 11 May 2023.

PARADIS, E. **Analysis of Phylogenetics and Evolution with R**. 2. ed. New York: Springer, 2011. XIV, 386 p. DOI: 10.1007/978-1-4614-1743-9.

PARADIS, E.; SCHLIEP, K. **Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R**. Bioinformatics, v. 35, n. 3, p. 526-528, 2018. DOI: 10.1093/bioinformatics/bty633.

PEREIRA, F. C. **Estratégias para seleção de progênies em soja.** Lavras: Universidade Federal de Lavras, 2016.

PIEPHO, H. P. *et al*. **BLUP for phenotypic selection in plant breeding and variety testing**. Euphytica, v. 161, p. 209-228, 2008.

PORTIN, P.; WILKINS, A. **The evolving definition of a gene**. Genetics, v. 205, n. 11, p. 1353–1364, 2017.

PREVOSTI, A.; OCAÑA, J.; ALONSO, G. **Distances between populations of Drosophila subobscura, based on chromosome arrangements frequencies**. Theoretical and Applied Genetics, v. 45, p. 231-241, 1975.

RAMALHO, M. A. P. *et al*. **Aplicações da genética quantitativa no melhoramento de plantas autógamas**. Lavras: UFLA, 2012. 522 p.

RAMALHO, M. A. P. *et al*. **Genética na Agropecuária**. 6. ed. rev. Lavras: UFLA, 2021.

RESENDE, M. D. V. de *et al*. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 2012.

RIBEIRO, F. O. *et al*. **Genotypic selection for soybean earliness. Journal of Agricultural** Science, v. 12, n. 11, p. 169, 2020. ISSN 1916-9752.

ROBERTSEN, C. D.; HJORTSHOJ, R. L.; JANS, L. L. **Genomic Selection in Cereal Breeding**. Agronomy, v. 9, p. 95, 2019. DOI: 10.3390/agronomy9020095. Accessed on: Feb. 23, 2024.

SÁNCHEZ-MAYOR, M. *et al*. **Effect of genotyping strategies on the sustained benefit of single-step genomic BLUP over multiple generations.** Genetics Selection Evolution, v. 54, n. 23, 2022. DOI: 10.1186/s12711-022-00712-y. Available from: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-022-00712-y. Accessed on: Feb 20, 2024.

SANDHU, D. *et al*. **The endogenous transposable element Tgm9 is suitable for generating knockout mutants for functional analyses of soybean genes and genetic improvement in soybean.** PLoS ONE, v. 12, e0180732, 2017 DOI: 10.1371/journal.pone.0180732.

SAPKOTA, S. *et al*. **Multi-Trait Regressor Stacking Increased Genomic Prediction Accuracy of Sorghum Grain Composition**. Agronomy, v. 10, n. 9, p. 1221, 2020. DOI: 10.3390/agronomy10091221.

SARAÇLI, S. *et al*. **Comparison of hierarchical cluster analysis methods by cophenetic correlation**. Journal of Inequalities and Applications, v. 2013, n. 203, 2013. DOI: 10.1186/1029-242X-2013-203.

SHAPIRO, S. S.; WILK, M. B**. An analysis of variance test for normality (complete samples).** Biometrika, v. 52, n. 3/4, p. 591-611, 1965. ISSN 0006-3444.

SILVA, F. L. da *et al*. **Soybean Breeding**. Springer International Publishing, 2017.

SOARES, I. O. *et al*. **Genetic and phenotypic parameters associated with soybean progenies in a recurrent selection program**. Biotechnology, v. 20, 2020. ISSN 1518-7853.

SOARES, I. O. *et al*. **Interaction between soybean cultivars and sowing density**. American Journal of Plant Sciences, v. 6, n. 9, p. 1425, 2015.

THOMSON, M. J. **High-Throughput SNP Genotyping to Accelerate Crop Improvement.** Plant Breeding and Biotechnology, v. 2, n. 3, p. 195–212, 2014.

TOKATLIDIS, I. S. **Conservation breeding of elite cultivars**. Crop Science, v. 55, n. 6, p. 2417–2434, 2015.

TOKATLIDIS, I. S. *et al*. **Variability within cotton cultivars for yield, fiber quality and physiological traits**. The Journal of Agricultural Science, v. 146, n. 4, p. 483–490, 2008.

VANRADEN, P. M. **Efficient Methods to Compute Genomic Predictions**. Journal Of Dairy Science, v. 91, n. 11, p. 4414-4423, 2008. DOI:10.3168/jds.2007-0980.

VIANA, J. M. S. *et al*. **Genomic selection in plant breeding: methods, models, and perspectives**. Trends in Plant Science, v. 27, n. 11, p. 1112-1126, 2022.

VILELA, N. J. D. *et al*. **Multiline is a strategy for homeostasis and Asian Soybean Rust Management in Agriculture**. Genetics and Molecular Research, 2024.

VOLPATO, L. *et al*. **Selection of inbred soybean progeny (Glycine max): an approach with population effect**. Plant breeding, v. 137, n. 6, p. 865-872, 2018. ISSN 0179-9541.

WICKHAM, H. *et al*. **Welcome to the Tidyverse**. Journal Of Open Source Software, v. 4, n. 43, p. 1686, 21 Nov. 2019.

YATES, J. L. *et al*. **SSR-marker analysis of the intracultivar phenotypic variation discovered within 3 soybean cultivars**. Journal of Heredity, v. 103, n. 5, p. 570-578, 2012. DOI:10.1093/jhered/ess015.

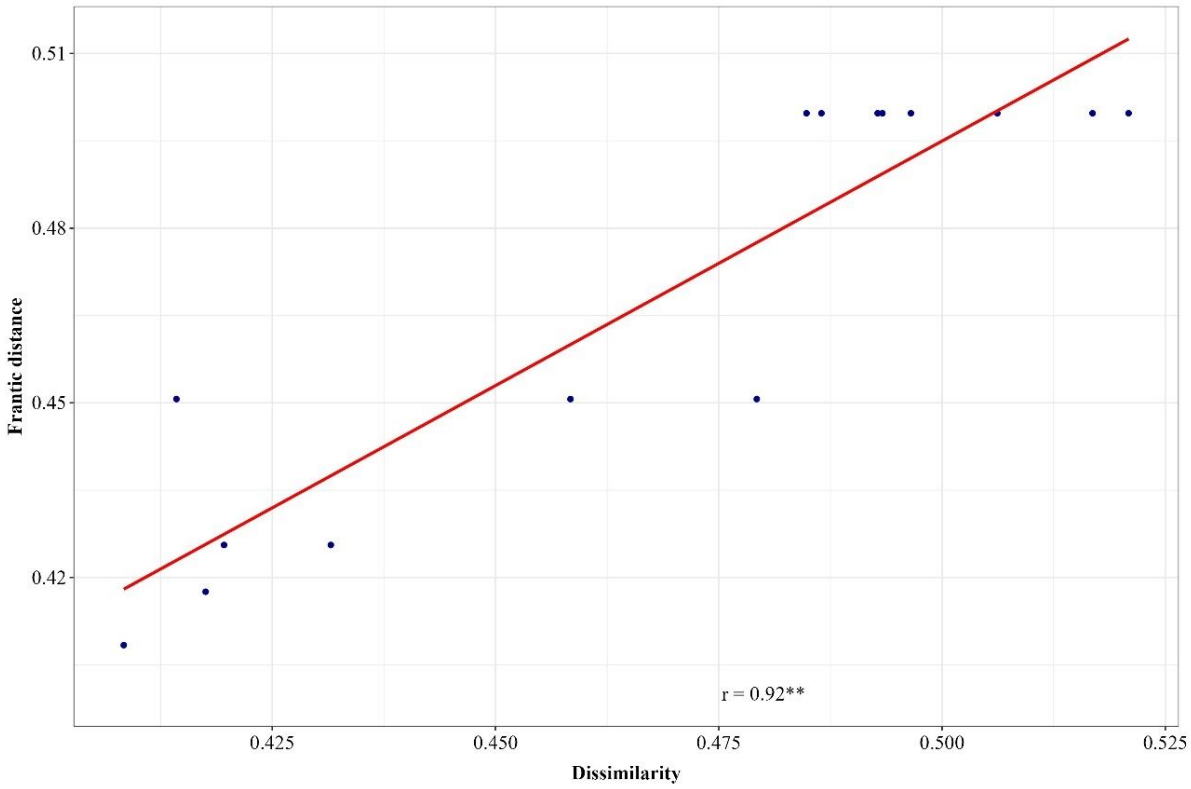
ZHANG, J. *et al*. **Comparison of the Efficiency of BLUP and GBLUP in Genomic Prediction of Immune Traits in Chickens**. Animals (Basel), v. 10, n. 3, p. 419, Mar. 2020. DOI: 10.3390/ani10030419.

ZHANG, W. *et al*. **AFEchidna is an R package for genetic evaluation of plant and animal breeding datasets.** Biorxiv, p. 1-12, 25 Jun. 2021. DOI:10.1101/2021.06.24.449740.

ZHANG, Y. X. *et al*. **Assessment of inter- and intra-inbred line variability in sunflower (*Helianthus annuus*) by RFLPs**. Genome, v. 38, p. 1040-1048, 1995. DOI:10.1139/g95-138.

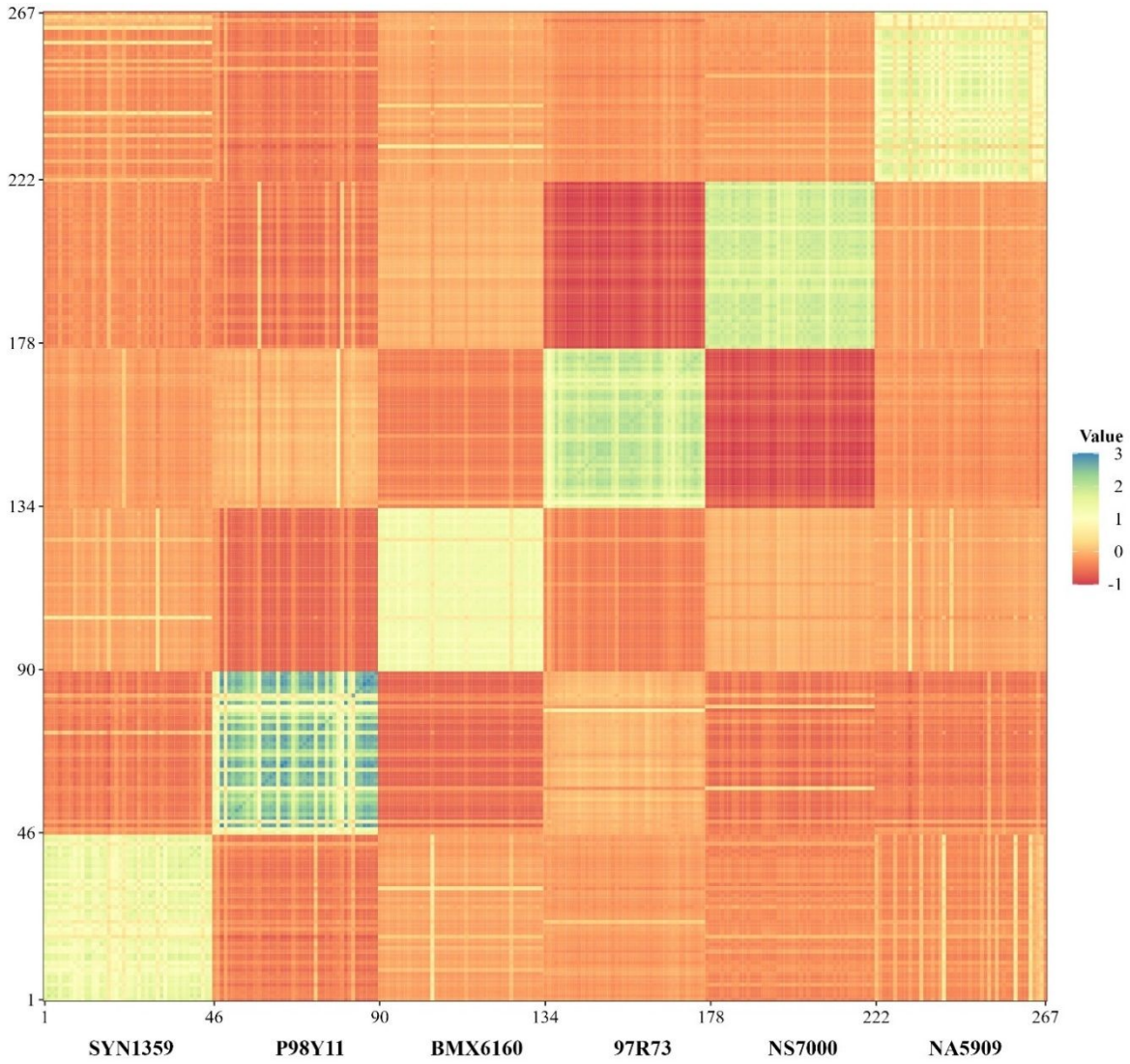
APPENDIX

Figure 1.1 -The cophenetic correlation between the progenies is 0,92. This value was significant at 1% probability by the Mantel test, based on 1,000 resampling.



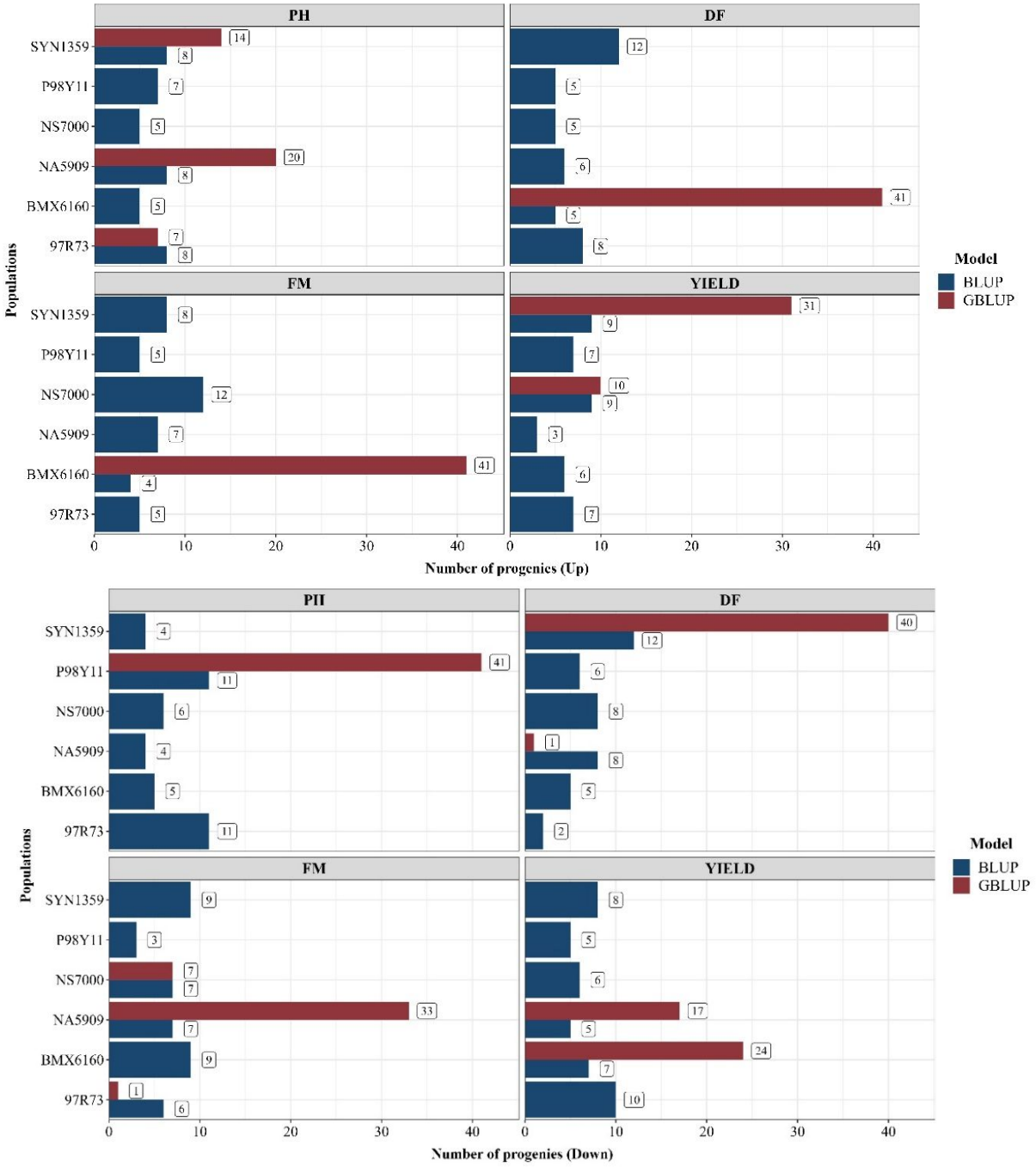
Source: From the author (2024).

Figure 1.2 – Genomic Relationship Matrix for 267 Progenies Based on 605 SNPs



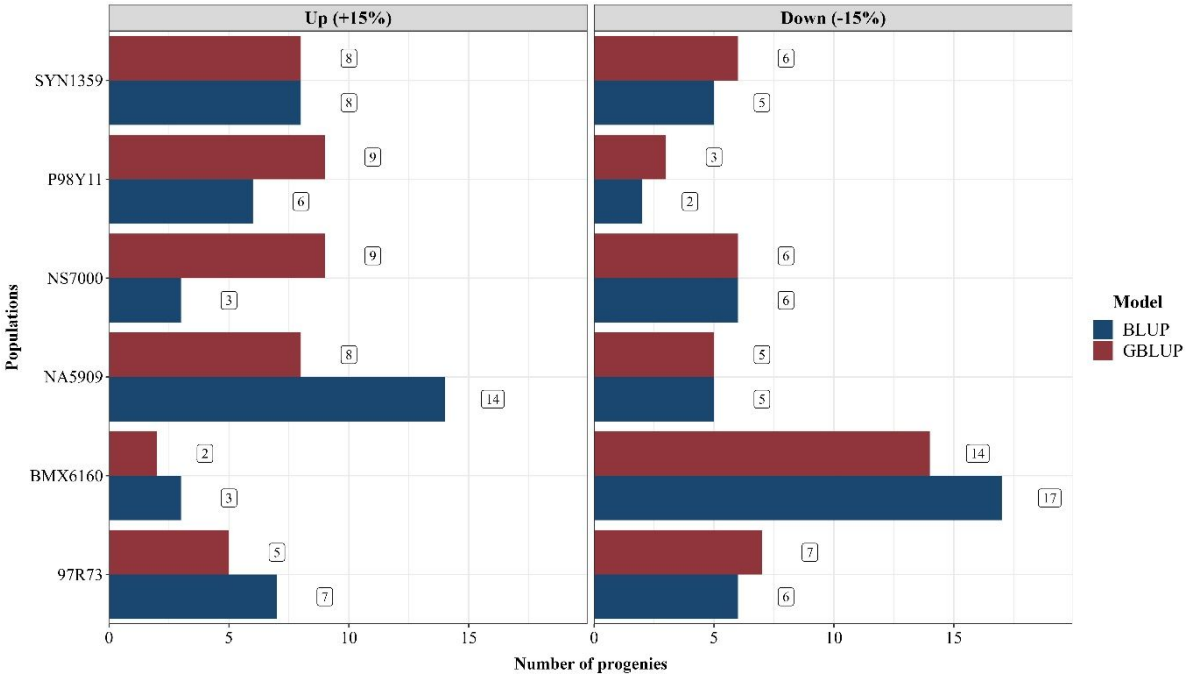
Source: From the author (2024).

Figure 1.3 – Quantity of progenies selected in each population considering BLUP and GBLUP General Selection with 15% of selection for all traits, for top-performing genotypes UP (+15%), the bottom-performing genotypes DOWN (-15%).



Source: From the author (2024).

Figure 1.4 – The number of progenies selected using the General Selection Index proposed by Mulamba and Mock (1978) in each population, considering BLUP and GBLUP for the top-performing genotypes UP (+15%) and the bottom-performing genotypes DOWN (-15%).



Source: From the author (2024).

			EARLY 18/19			EARLY 19/20			
SV	Effect	DF	FM	YIELD	PH	DF	FM	YIELD	PH
MS_{Test}	F	59,37**	7682,37	16171145,16**	7121,55	1544,78	19008,22	45031193,86**	9610,08**
σ^2_{Prog}	R	0,19	3,84**	494146,06**	7,27	0,66**	0,00	27296,65	7,79**
σ^2_{Pop}	R	0,27**	1,49**	114574,38**	60,45**	0,77**	1,30**	38464,72**	82,25**
$\sigma^2_{SYN\,1359}$	R	0,02	2,28*	470287,68*	6,49	0,49	0,43*	147547,80**	5,30
$\sigma^2_{NA\,5909}$	R	0,12	7,39**	506147,06**	14,02*	1,08**	0,09	0,00	18,01**
$\sigma^2_{BMX\,6160}$	R	0,55	2,02	492164,39**	3,34	0,56	0,55*	0,00	6,71
σ^2_e	R	1,09	5,64	658204,49	35,26	2,84	3,15	543704,05	44,15
Mean		45,04	115,35	4530,04	75,16	47,01	125,57	5190,79	79,22
h^2_{prog}		0,15	0,41	0,43	0,17	0,19	0,00	0,05	0,15
h^2_{pop}		0,18	0,14	0,09	0,59	0,18	0,29	0,06	0,61
h^2_{total}		0,30	0,49	0,48	0,66	0,34	0,29	0,11	0,67
$r_{g\hat{g}_{prog}}$		0,50	0,74	0,75	0,52	0,63	-	0,35	0,57
$r_{g\hat{g}_{pop}}$		0,69	0,68	0,66	0,70	0,69	0,70	0,67	0,70
CV_{prog} (%)		0,96	1,70	15,52	3,59	1,74	0,00	3,18	3,52
CV_{pop} (%)		1,16	1,06	7,47	10,34	1,88	0,91	3,78	11,45
CV_e (%)		2,31	2,06	17,91	7,90	3,59	1,41	14,21	8,39

Table 1.1 – Results of Individual Variance Analysis for Late Maturing in the Years 2018/19 and 2019/20 for traits Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), Plant Height (PH).

SV: Source of variation; Effect: Nature of the effect (F - fixed and R- random); MS_{Test} — ¿ mean square of checks, σ^2_{Prog} — ¿ progenies variance, σ^2_{Pop} — ¿ population variance, $\sigma^2_{SYN\,1359}$ — ¿ variance among SYN1359, $\sigma^2_{NA\,5909}$ — ¿variance among NA5909, $\sigma^2_{BMX\,6160}$ — ¿variance among BMX6160 σ^2_e — ¿ residual variance, DF and FM (days); PH (cm); YIELD (Kg ha-1); F-test for fixed effects (F) and LRT (Likelihood Ratio Test) for random effects (A); "ns" indicates 'not significant', whereas asterisks (*) indicate levels of significance, with one asterisk for $p < 0.05$ and two asterisks for $p < 0.01$.h^2pop: Heritability among populations; h^2prog: heritability within populations modified; h^2total: total heritability = heritability among + within populations modified; rggprog: accuracy on the progeny-mean basis; rggpop: accuracy on the population; CVgPop: population coefficient of variation in percentage; CVgProg: coefficient of variation of progeny or progeny within population, in percentage terms; CVe: experimental coefficient of variation in percentage terms.

Source: From the author (2024).

Table 1.2 – Results of Individual Variance Analysis for Late Maturing in the Years 2018/19 and 2019/20 for traits Days to Flowering (DF), Full Maturity (FM), Grain Yield (YIELD), Plant Height (PH).

		LATE 18/19				LATE 19/20			
SV	Effect	DF	FM	YIELD	PH	DF	FM	YIELD	PH
MS_{Test}	F	2415,36	18531,47**	87484805,95**	8530,26	3001,62**	15602,35**	6549,11	175,29*
σ^2_{Prog}	R	0,68	4,43**	992403,98**	12,33**	0,89**	0,40**	15178571,35**	13,50**
σ^2_{Pop}	R	40,98**	58,76**	659543,97**	391,53**	12,38**	14,57**	232974,24**	296,30**
σ^2_{97R73}	R	0,00	6,33**	1471985,69**	18,64	1,18**	0,78**	78391,68**	26,74**
σ^2_{NS7000}	R	1,61*	7,08**	476040,94*	0,00	1,08**	0,84**	418980,50**	10,32
σ^2_{98Y11}	R	1,13	0,97	1206205,80**	11,01	0,43	0,37	167254,56**	12,32*
σ^2_e	R	5,14	11,05	702700,40	41,88	1,84	2,44	112784,49	52,74
Mean		55,64	136,54	5935,22	90,68	59,84	137,81	4437,84	81,44
h^2_{prog}		0,12	0,29	0,59	0,23	0,33	0,14	0,35	0,20
h^2_{pop}		0,88	0,79	0,28	0,88	0,82	0,84	0,11	0,82
h^2_{total}		0,89	0,85	0,70	0,91	0,88	0,86	0,42	0,85
$r_{g\hat{g}_{prog}}$		0,44	0,64	0,82	0,58	0,75	0,56	0,77	0,64
$r_{g\hat{g}_{pop}}$		0,71	0,71	0,69	0,71	0,71	0,71	0,67	0,71
CV_{prog} (%)		1,48	1,54	16,78	3,87	1,58	0,46	10,88	4,51

CV_{pop} (%)	11,51	5,61	13,68	21,82	5,88	2,77	6,31	21,14
CV_{e} (%)	4,07	2,43	14,12	7,14	2,27	1,13	14,77	8,92

SV: Source of variation; Effect: Nature of the effect (F - fixed and R- random); MS_{Test} – ¿ mean square of checks, σ^2_{Prog} – ¿ progenies variance, σ^2_{Pop} – ¿ population variance, σ^2_{97R73} – ¿ variance among 97R73, σ^2_{NS7000} – ¿variance among NS7000, σ^2_{98Y11} – ¿variance among 98Y11σ^2_e – ¿ residual variance DF and AM (days); PH (cm); YIELD (Kg ha-1); F-test for fixed effects (F) and LRT (Likelihood Ratio Test) for random effects (A); "ns" indicates 'not significant', whereas asterisks (*) indicate levels of significance, with one asterisk for $p < 0.05$ and two asterisks for $p < 0.01$. h^2pop: Heritability among populations; h^2prog: heritability within populations modified; h^2total: total heritability = heritability among + within populations modified; rggprog: accuracy on the progeny-mean basis; rggpop: accuracy on the population; CVgPop: population coefficient of variation in percentage; CVgProg: coefficient of variation of progeny or progeny within population, in percentage terms; CV_e: experimental coefficient of variation in percentage terms.

Source: From the author (2024).