



Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós-Graduação em Biologia Molecular

**DECIFRANDO A DIVERSIDADE E ESTRUTURA GENÉTICA
DE COLEÇÕES DE GERMOPLASMA DE *ZEA MAYS* L. POR
MEIO DA ALELOTIPAGEM DE SNPs**

BIANCA DE SOUSA ALCÂNTARA

Brasília, DF
2025

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós-Graduação em Biologia Molecular

**DECIFRANDO A DIVERSIDADE E ESTRUTURA GENÉTICA
DE COLEÇÕES DE GERMOPLASMA DE *ZEA MAYS* L. POR
MEIO DA ALELOTIPAGEM DE SNPs**

BIANCA DE SOUSA ALCÂNTARA

Orientador: Dr. Dario Grattapaglia

Dissertação de mestrado apresentada ao
Departamento de Biologia Celular do
Instituto de Ciências Biológicas da
Universidade de Brasília como requisito
parcial para a obtenção do título de Mestre
em Biologia Molecular.

Brasília, DF
2025

BIANCA DE SOUSA ALCÂNTARA

**DECIFRANDO A DIVERSIDADE E ESTRUTURA GENÉTICA
DE COLEÇÕES DE GERMOPLASMA DE *ZEA MAYS* L. POR
MEIO DA ALELOTIPAGEM DE SNPs**

Dissertação de mestrado apresentada ao
Departamento de Biologia Celular do
Instituto de Ciências Biológicas da
Universidade de Brasília como requisito
parcial para a obtenção do título de Mestre
em Biologia Molecular.

Banca Examinadora:

Prof. Dr. Dario Grattapaglia (Orientador) – (UnB)

Prof. Dr. Marcelo de Macedo Brígido (Examinador Interno) – (UnB)

Dr. Márcio Elias Ferreira (Examinador Externo) – EMBRAPA

Dr. Márcio de Carvalho Moretzsohn (Suplente) – EMBRAPA

Dedico este trabalho aos meus pais, Hécio e Rosângela Alcântara, que me deram apoio imensurável durante toda a minha jornada científica. Que durante toda a vida se sacrificaram para que eu pudesse alcançar lugares ainda mais distantes e voos ainda mais altos.

AGRADECIMENTOS

A Deus, primeiramente, pois sem Ele essa jornada não seria cumprida.

À Universidade de Brasília (UnB) e especialmente ao Programa de Pós-Graduação em Biologia Molecular, representados por todos os professores e funcionários, pela oportunidade de realização deste curso.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF) pelo apoio financeiro que tornou possível a realização desse trabalho.

À EMBRAPA Recursos Genéticos e Biotecnologia (Cenargen) e, em especial, o Laboratório de Genética Vegetal, representados por todos os pesquisadores e técnicos, pela infraestrutura de trabalho, e pela disposição e genuíno interesse em ajudar os alunos.

À EMBRAPA Milho e Sorgo e especialmente a Dr^a. Flávia Franca Teixeira pela disposição no fornecimento de sementes e informações indispensáveis a esse trabalho.

Ao meu orientador, Dr. Dario Grattapaglia, não apenas por ter aceitado orientar esse trabalho, mas por ser uma inspiração e um grande exemplo a ser seguido. Pelos ensinamentos valiosos durante essa jornada, pelo zelo, pela revisão atenta, pelo cuidado durante esse processo e pelos conselhos que certamente fizeram toda diferença para o meu desenvolvimento profissional e pessoal.

Ao Dr. Guilherme Ferreira Simiqueli por toda ajuda e suporte dados sem nenhuma hesitação.

À Dr^a Marília Pappas por me acolher desde o início da minha jornada científica com muito zelo, apoio, incentivo e carinho.

Ao Dr. Marcelo de Macedo Brígido, Dr. Márcio Elias Ferreira e Dr. Márcio de Carvalho Moretzsohn, por terem aceitado participar da Banca Examinadora e pelas contribuições valiosas a esse trabalho.

Aos meus amigos do laboratório, Lorena, Wellington e Nayara, por todos os momentos compartilhados dentro e fora do laboratório que certamente fizeram toda a diferença para que eu avançasse durante todo o processo.

Aos meus avós Neuton (*in memoriam*) e Maria de Lourdes por sempre terem me feito acreditar que eu era capaz de alçar longos voos.

À minha família, Hércio, Rosângela, Natália e Almir, pelo apoio incondicional, por não medirem esforços para o meu crescimento pessoal e profissional e pelo incentivo sem medidas. Por tudo que eu sou, por tudo que consegui conquistar, pela felicidade que tenho e principalmente por todo amor a mim dado.

Aos meus grandes amigos, Kellen e Raphael, que são irmãos de coração, por todo esforço para que eu pudesse superar cada obstáculo em meu caminho e chegar até aqui. Por serem meu suporte em todos os momentos, pelo apoio imensurável e pela torcida entusiasmada.

RESUMO GERAL

O Banco Ativo de Germoplasma de Milho (BAG Milho) conta com mais de 4.000 acessos preservados em sua coleção de base. A partir de informações fenotípicas e de origem geográfica foi criada a Coleção Nuclear de Milho, uma amostra representativa da coleção de base, com máxima representatividade de variabilidade genética e mínima redundância. Adicionalmente, foi montada uma coleção de acessos classificados como raças indígenas e/ou originários de terras indígenas. Embora mais de 80% dos acessos do BAG tenham sido caracterizados de acordo com descritores morfológicos da cultura do milho, até o momento, pouco se sabe a respeito da diversidade molecular ampla no genoma, dentro e entre os acessos destas coleções. O conhecimento da diversidade, ancestralidade comum e estrutura genética em nível molecular de uma coleção de germoplasma é uma informação importante para validar e complementar as informações fenotípicas e de origem no desenho de estratégias de conservação e uso de recursos genéticos e melhoramento. Um dos desafios da análise genômica de coleções de germoplasma compostas por grandes números de acessos geneticamente heterogêneos (populações ou variedades não endogâmicas), é lidar com a variação genética existente dentro dos acessos. A genotipagem de vários indivíduos por acesso torna o esforço demorado e caro. A alelotipagem do acesso, ao contrário, estimando as frequências alélicas de uma amostra agrupada de indivíduos de um acesso, representa uma alternativa elegante para análise genética de alta acurácia e baixo custo. Um conjunto de 3.526 SNPs distribuídos no genoma do milho, interrogados no chip *EMBRAPA 65kMultiespécies* foi usado neste estudo genético. Para avaliar a acurácia do método, três métodos de montagem de pool foram avaliados para quatro acessos diferentes, comparando as estimativas de frequências alélicas obtidas pela razão de intensidade alélica normalizada dos dois alelos ao SNP em pools de 16 plantas por acesso, em três réplicas, com a estimativa das mesmas plantas genotipadas individualmente. Três métodos foram testados: (1). Pool equimolar de DNA extraído individualmente e concentração mensurada em espectrofotometria em equipamento “Nanodrop”; (2). O mesmo que (1), mas o DNA mensurado por fluorimetria com “picogreen”; (3). Extração de DNA a partir de uma amostra consolidada (pool ou bulk) de tecidos foliares das várias plantas, pesados com precisão de 0,1 grama. As frequências alélicas estimadas a partir da genotipagem do DNA preparado com o método (3) de extração foliar em bulk, apresentaram o menor desvio quadrático médio (6,3%) e a maior correlação (0,984) com as frequências alélicas estimadas a partir de plantas individualmente. Usando este método, pools de DNA de 24 plantas de cada um dos 293 acessos da Coleção Nuclear Brasileira de milho e 188 acessos da coleção de milho indígena foram alelotipados. Um conjunto de 64 réplicas biológicas também foi alelotipado para avaliação de reprodutibilidade avaliada pela estimativa de diferenciação genética de Wright. Estimativas precisas das frequências alélicas foram obtidas com erro de amostragem desprezível estimado por um F_{ST} médio entre réplicas de 0,0038. Análises de distância genética e estrutura foram realizadas para os 481 acessos juntamente com dados dos mesmos SNPs alelotipados obtidos da literatura para acessos de milho e Teosinto conservados pelo centro internacional CIMMYT (*Centro Internacional de Mejoramiento de Maíz y Trigo*) representativos de germoplasma tropical, subtropical e temperado. As coleções nuclear e indígena apresentaram diferenciação genética moderada, com um F_{ST} médio de 0,136 revelando um padrão de estrutura populacional no qual acessos da coleção nuclear se sobrepuseram a acessos de origem indígena e germoplasma CIMMYT. Entretanto, a coleção indígena apresentou maior diferenciação entre os acessos ($F_{ST} = 0,140$) em comparação à coleção nuclear ($F_{ST} = 0,109$), o que sugere a necessidade de uma atualização da composição da coleção nuclear incluindo mais acessos indígenas e assim ampliar a diversidade genética representada no germoplasma disponível da espécie. O agrupamento por k-means detectou sete grupos, que continham proporções variáveis

de acessos da coleção nuclear e indígena. Padrões de estrutura genética baseados na distância genética, conforme definido pelos dados do SNP, foram encontrados com sobreposição parcial com classificações dos acessos de germoplasma usados no estabelecimento das coleções, tais como tipos de grãos de milho, nível de domesticação e origem do material coletado. No entanto, algumas discrepâncias importantes também foram detectadas onde tais classificações não correspondiam ao que os dados de DNA mostraram. Esses resultados indicam que informações genômicas não enviesadas derivadas dos dados do SNP fornecem informações adicionais importantes que refletem de forma mais precisa a origem ancestral dos acessos de germoplasma. Propomos que os dados genômicos gerados neste trabalho sejam prontamente integrados aos dados de passaporte disponíveis dos acessos estudados para chegar a uma representação mais precisa do germoplasma de milho conservado nas duas coleções.

Palavras-chave: *Zea mays* L.; Germoplasma; Alelotipagem; Raças Indígenas, Bulk de DNA; Diversidade Genética; SNPs.

GENERAL ABSTRACT

The Brazilian Maize Germplasm Bank (Maize BAG) has over 4.000 accessions preserved in its base collection. This BAG also has the Brazilian Maize Core Collection, a representative sample of the base collection, and a collection of accessions classified as indigenous landraces and/or originating from indigenous lands. Although over 80% of the BAG accessions have been characterized according to morphological descriptors of the maize plant, to date, little is known about the genome-wide molecular diversity effectively preserved in these collections. Knowledge of the diversity, common ancestry and genetic structure of a germplasm collection at the molecular level is important information to validate and complement phenotypic and origin information in the design of conservation and use strategies for genetic resources and targeted improvement. One of the challenges of genomic analysis of germplasm collections composed of large numbers of genetically heterogeneous accessions (populations or non-inbred varieties) is how to deal with the genetic variation existing within the accessions. Genotyping several individuals per accession makes the effort time-consuming and expensive. Accession allelotyping, on the other hand, estimating the allele frequencies of a pooled sample of individuals of an accession, represents an elegant alternative for high-precision and low-cost genetic analysis. A set of 3.526 SNPs distributed in the maize genome, interrogated on the *EMBRAPA 65kMultispecies* chip, was used in this genetic study. To evaluate the accuracy of the method, three pool assembly treatments were evaluated for four different accessions, comparing the allele frequency estimates obtained by the normalized allele intensity ratio of the two alleles at the SNP in pools of 16 plants, in three replicates, with the estimate of the same plants genotyped individually. Three methods were tested: (1). Equimolar pool of individually extracted DNA and concentration measured by spectrophotometry in “Nanodrop” equipment; (2). Same as (1), but DNA measured by fluorimetry with “picogreen”; 3. DNA extraction from a consolidated sample (pool or bulk) of leaf tissues of the various plants, weighed to the nearest 0.1 gram. The allele frequencies estimated from the allelotyping of DNA prepared with method (3) of bulk leaf extraction, showed the lowest root mean square deviation (6.3%) and the highest correlation (0.984) with the allele frequencies estimated from individual plants. Using this method, DNA pools from 24 plants of each of the 293 accessions of the Brazilian Maize Core Collection and 188 accessions of the indigenous maize collection were allelotyped. A set of 64 biological replicates was also allelotyped for evaluation of reproducibility assessed by Wright's genetic differentiation estimate. Accurate estimates of allele frequencies were obtained with negligible sampling error estimated by a mean F_{ST} among replicates of 0.0038. Genetic distance and structure analyses were performed for the 481 accessions together with data from the same allelotyped SNPs obtained from the literature for maize and Teosinte accessions conserved by the international center CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) representative of tropical, subtropical and temperate germplasm. The core and indigenous collections showed moderate genetic differentiation, with an average F_{ST} of 0.136 revealing a pattern of population structure in which accessions from the core collection overlapped with accessions of indigenous origin and CIMMYT germplasm. However, the indigenous collection showed greater genetic differentiation between accessions ($F_{ST} = 0.140$) compared to the core collection ($F_{ST} = 0.109$), which suggests the need to update the composition of the latter including more indigenous accessions and thus expand the genetic diversity represented in the available germplasm of the species. Clustering by k-means detected seven groups, that contained variable proportions of core and indigenous accessions. Patterns of genetic structure based on genetic distance, as defined by SNP data, were found to partially overlap with classifications of the germplasm accessions used in the establishment of the

collections, such as maize grain types, domestication level and origin of the collected material. However, some important discrepancies were also detected where such classifications did not match what the DNA data showed. These results indicate that the unbiased genomic information derived from SNP data provides important additional and alternative information that more precisely reflects the ancestral origin of the germplasm accessions. We propose that the genomic data generated in this work be promptly integrated with the available passport data of the studied accessions to arrive at a more accurate representation of the maize germplasm conserved in the two collections.

Keywords: *Zea mays* L.; Germplasm; Allelotyping; Indigenous Landraces, DNA pooling; Genetic Diversity; SNPs.

LISTA DE FIGURAS

CAPÍTULO 1

Figura 1 – Fluxograma esquemático da montagem de bulks	65
Figura 2 – Gráficos de correlação (<i>correlation plot</i>) entre frequências alélicas estimadas por meio de bulks de DNA (eixo Y) e calculadas a partir de dados genotípicos (eixo x) para os mais de 3430 SNPs (pontos no gráfico).....	75

CAPÍTULO 2

Figura 1 – Composição e classificação dos acessos da coleção nuclear	104
Figura 2 – Composição e classificação dos acessos da coleção indígena	105
Figura 3 – Estruturas populacionais dos acessos detectadas pela análise de agrupamento k-means com base em diferentes matrizes de distância genética (Euclidiana, F_{ST} e distância de Nei)	107
Figura 4 – Diagramas de agrupamento k-means envolvendo todos os acessos (BRA_IND_XU) com base nas diferentes matrizes de distância genética em função dos dois primeiros componentes principais estimados pela respectiva matriz de distância	108
Figura 5 – Dendrograma UPGMA construído com base na matriz de base em F_{ST} (índice de fixação de Wright) envolvendo acessos da coleção nuclear, indígena e de bancos de germoplasma internacionais publicados por Xu et al. 2022	110
Figura 6 – Distribuição dos 109 acessos alocados no grupo 1 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	112
Figura 7 – Distribuição dos 15 acessos alocados no grupo 2 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	113
Figura 8 – Distribuição dos 69 acessos alocados no grupo 3 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	115
Figura 9 – Distribuição dos 91 acessos alocados no grupo 4 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	116
Figura 10 – Distribuição dos 98 acessos alocados no grupo 5 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	117
Figura 11 – Distribuição dos nove acessos alocados no grupo 6 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	119
Figura 12 – Distribuição dos 103 acessos alocados no grupo 7 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem	120
Figura 13 – Definição do número de grupos genéticos com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos	121
Figura 14 – Definição do número de subgrupos genéticos dentro de cada um dos dois grupos definidos em análise anterior com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos	122

Figura 15 – Proporções médias da ancestralidade genômica dos indivíduos de cada um dos grupos genéticos detectados pelo STRUCTURE para os 494 acessos.....	123
Figura 16 – – Comparação entre (A) a dispersão de acessos nos sete grupos genéticos detectados pelo agrupamento k-means (A) e (B) nos cinco grupos genéticos detectados pela análise de estrutura com base em indivíduos simulados com os software STRUCTURE.....	125
Figura 17 – Sobreposição da classificação racial dos 494 acessos estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright)	127
Figura 18 – Sobreposição da classificação racial dos 188 acessos indígenas estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright)	129
Figura 19 – Sobreposição da origem geográfica de coleta dos 494 acessos estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright)	131
Figura 20 – Sobreposição da classificação quanto ao tipo de grão dos 494 acessos estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright)	132

LISTA DE TABELAS

CAPÍTULO 1

Tabela 1 – Esquema experimental dos métodos de preparação das amostras de DNA agrupadas em bulks com as respectivas repetições 66

Tabela 2 – Estimativas da raiz do erro quadrático médio (RMSE), correlação (r) e coeficiente de determinação (R^2) entre as frequências alélicas estimadas a partir do bulk e as frequências alélicas calculadas a partir de genótipos de amostras individuais 73

Tabela 3 – Estimativas do índice de fixação (F_{ST}) entre réplicas biológicas na avaliação do efeito de amostragem de plantas na montagem de bulks por meio do método de extração foliar em massa 76

CAPÍTULO 2

Tabela 1 – Percentual de acessos provenientes das diferentes coleções de germoplasma em cada um dos sete grupos detectados pelo k-means com base na matriz F_{ST} 109

Tabela 2 – Percentual de acessos provenientes das diferentes coleções de germoplasma em cada um dos cinco grupos detectados pela análise de estrutura com o software STRUCTURE realizada com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos 124

Tabela 3 – Mapa de sobreposição da alocação de acessos a grupos genéticos detectados nas análises de estrutura com o software STRUCTURE e agrupamento pelo k-means 125

ANEXO 2

Tabela S1 – Comparação da diferenciação genética (F_{ST}) entre acessos de origem indígena não classificados em raças com acessos classificados em raças indígenas e acessos da coleção nuclear 144

LISTA DE ABREVIATURAS E SÍMBOLOS

A	Adenina
AFLP	Polimorfismo de Comprimento de Fragmento Amplificado
AP	Antes do Presente
BAF	<i>B Allele Frequency</i>
BAG	Banco Ativo de Germoplasma
BPGV	Banco Português de Germoplasma Vegetal
BRA	Acessos da Coleção Nuclear de Germoplasma de Milho
C	Citosina
CENARGEN	Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia
CNV	Variantes de número de cópias
CGIAR	<i>Consulative Group on International Agricultural Research</i>
CIA	Clorofórmio: Álcool Isoamílico, 24:1
CIMMYT	<i>Centro Internacional de Mejoramiento de Maíz y Trigo</i>
CNPMS	Centro Nacional de Pesquisa de Milho e Sorgo
CR	<i>Call rate</i>
CSIR	<i>Savanna Agricultural Research Institute</i>
CTAB	Brometo de Cetiltrimetilamônio
DL	Desequilíbrio de Ligação
DNA	Ácido desoxirribonucleico
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
Esalq	Escola Superior de Agricultura Luiz de Queiroz
EUA	Estados Unidos da América
F	Coeficiente de Endogamia
F_{ST}	Coeficiente de Diferenciação Genética
g	Gramas
G	Guanina
GBS	Genotipagem por Sequenciamento
H_E	Heterozigosidade Esperada
HIV	Vírus da Imunodeficiência Humana
IC	Intervalo de Confiança
IND	Acessos da Coleção Índigena de Milho
Kb	Kilobase
LGV	Laboratório de Genética Vegetal
m	Metros
MAF	Frequências do Alelo Menor
Mb	Megabase
ml	Mililitro
mm	Milímetro
MRIZP	<i>Maize Research Institute Zemun Polje</i>
p	<i>p-value</i>
pb	Par de bases

PCR	Reação de Polimerização em Cadeia
PIC	Conteúdo de Informação Polimórfica
QLTs	<i>quantitative trait locos</i>
r	Coeficiente de Correlação de Pearson
R ²	Coeficiente de Determinação
RAPD	Polimorfismo de DNA Amplificado ao Acaso
RFLP	Polimorfismo de Comprimento de Fragmentos de Restrição
RG	Recursos Genéticos
RMSE	Raiz do Erro Quadrático Médio
SNP	Polimorfismo de Nucleotídeo Único
spp	Espécies
SSR	<i>Simple Sequence Repeats</i> (microsatélites)
T	Timina
UPGMA	<i>Unweighted Pair-Group Method using Arithmetic Averages</i>
USDA	<i>United States Department of Agriculture</i>
VBT	Variação entre Grupos Dividida pelo Total
µl	Microlitro

SUMÁRIO

1. INTRODUÇÃO GERAL	18
2. OBJETIVOS	22
2.1 OBJETIVO GERAL	22
2.2 OBJETIVOS ESPECÍFICOS	22
3. REVISÃO BIBLIOGRÁFICA	23
3.1. BOTÂNICA.....	23
3.2 PRODUÇÃO	24
3.3 MERCADO GLOBAL	25
3.4 ORGANIZAÇÃO DO GENOMA.....	26
3.5 RAÇAS DE MILHO NO BRASIL	27
3.5.1 Raças Indígenas	28
3.5.2 Raças Comerciais Antigas	28
3.5.3 Raças Comerciais Recentes	29
3.5.4 Raças Comerciais Exóticas.....	29
3.6 RECURSOS GENÉTICOS	30
3.7 MARCADORES MOLECULARES E ESTUDOS DE DIVERSIDADE GENÉTICA	31
3.7 MARCADORES SNPs.....	37
3.8 ALELOTIPAGEM.....	41
3.9 ESTUDOS DE ORIGEM E EVOLUÇÃO DO MILHO COM MARCADORES MOLECULARES.....	44
4. REFERÊNCIAS BIBLIOGRÁFICAS	49
CAPÍTULO 1 – COMPARAÇÃO DO MÉTODOS DE PREPARAÇÃO DE “BULKS” DE DNA PARA A ESTIMATIVA DE FREQUÊNCIAS ALÉLICAS POR MEIO DA ALELOTIPAGEM DE SNPs EM <i>ZEA MAYS</i>	59
RESUMO.....	59
ABSTRACT	60
1. INTRODUÇÃO.....	61
2. MATERIAIS E MÉTODOS	64
2.1 MATERIAL GENÉTICO	64
2.2 EXTRAÇÃO DE DNA	64
2.3 EXPERIMENTO DE OTIMIZAÇÃO DE MONTAGEM DE BULKS	64
2.4 GENOTIPAGEM DE SNPs	66

2.5 ESTIMATIVA DAS FREQUÊNCIAS ALÉLICAS	67
2.6 FILTRAGEM DE SNP E CONTROLE DE QUALIDADE	67
2.7 ANÁLISE DE DADOS	68
2.7.1 Cálculo de Root-mean-square Error	68
2.7.2 Coeficiente de Correlação de Pearson	69
2.7.3 Coeficiente de Determinação.....	70
2.8 AVALIAÇÃO DO EFEITO DE AMOSTRAGEM DE PLANTAS NA ESTIMATIVA DE FREQUÊNCIAS ALÉLICAS EM RÉPLICAS.....	71
3. RESULTADOS.....	72
3.1 COMPARAÇÃO ENTRE OS MÉTODOS DE MONTAGEM DE BULKS	72
3.2 AVALIAÇÃO DO EFEITO DE AMOSTRAGEM DE PLANTAS NA ESTIMATIVA DE FREQUÊNCIAS ALÉLICAS EM RÉPLICAS.....	76
4. DISCUSSÃO	79
5. REFERÊNCIAS BIBLIOGRÁFICAS	86
CAPÍTULO 2 – DIVERSIDADE E ESTRUTURA GENÉTICA ENTRE ACESSOS CLASSIFICADOS EM RAÇAS COMERCIAIS E INDÍGENAS DO BANCO DE GERMOPLASMA DE MILHO	93
RESUMO.....	93
ABSTRACT	94
1. INTRODUÇÃO.....	95
2. MATERIAIS E MÉTODOS	97
2.1 LEVANTAMENTO DE DADOS DE REPRESENTATIVIDADE DE COLEÇÕES DE GERMOPLASMA BRASILEIRAS	97
2.2 MATERIAL VEGETAL E PREPARAÇÃO DE AMOSTRAS DE DNA EM BULK ...	98
2.3 ALELOTIPAGEM DE SNPs.....	99
2.4 ESTIMATIVA DAS FREQUÊNCIAS ALÉLICAS	99
2.5 OBTENÇÃO DE DADOS GENOTÍPICOS DE MILHO E TEOSINTO DA LITERATURA.....	100
2.6 ANÁLISE DE DADOS	101
2.6.1 Cálculo de matrizes de distância genética entre acessos	101
2.6.2 Agrupamento de acessos com base em clusterização k-means	101
2.6.3 Análise de estrutura genética: reconstituição de indivíduos a partir das frequências alélicas	101
2.6.4 Compartimentalização das análises de agrupamento e estrutura de acordo com a origem dos acessos	102

2.7 COMPARAÇÃO DAS CLASSIFICAÇÕES MORFOLÓGICAS E GEOGRÁFICAS COM AGRUPAMENTOS DETECTADOS PELOS DADOS GENÉTICOS.....	103
3. RESULTADOS.....	104
3.1 REPRESENTATIVIDADE DA COLEÇÃO NUCLEAR QUANTO ÀS DIFERENTES CLASSIFICAÇÕES	104
3.2 ANÁLISE DA COLEÇÃO DE ACESSOS DE ORIGEM INDÍGENA.....	106
3.3 ESTRUTURA GENÉTICA DETECTADA PELO AGRUPAMENTO K-MEANS	107
3.4 ANÁLISE DE ESTRUTURA GENÉTICA COM BASE EM GENÓTIPOS INDIVIDUAIS RECONSTITUIDOS	122
3.5 COMPARAÇÃO DAS CLASSIFICAÇÕES MORFOLÓGICAS E GEOGRÁFICAS COM AGRUPAMENTOS DETECTADOS PELOS DADOS GENÉTICOS.....	127
4. DISCUSSÃO	134
5. CONCLUSÃO E PERSPECTIVAS	141
6. REFERÊNCIAS BIBLIOGRÁFICAS	142
ANEXO 1 - PROTOCOLO DE EXTRAÇÃO DE DNA DE PLANTAS COM 3X CTAB...	145
ANEXO 2 - COMPARAÇÃO DA DIFERENCIAÇÃO GENÉTICA ENTRE ACESSOS DE ORIGEM INDÍGENA NÃO CLASSIFICADOS RAÇAS COM ACESSOS CLASSIFICADOS EM RAÇAS INDÍGENAS E ACESSOS DA COLEÇÃO NUCLEAR.	146

1. INTRODUÇÃO GERAL

A preservação da diversidade genética representada pelas inúmeras variedades locais e regionais de milho tem sido reconhecida como um esforço de grande importância para a sustentabilidade do melhoramento genético da cultura e mitigação de sua erosão genética (KHOURY *et al.*, 2022). O trabalho de coleta, preservação, plantio e documentação das inúmeras variedades de milho em bancos ativos de germoplasma (BAG) é o produto dos esforços cooperativos de muitos indivíduos ao longo das últimas sete décadas (ORTIZ *et al.*, 2010). Os recursos genéticos (RG) de milho constituem um tesouro imensurável para a humanidade. As estratégias de conservação do germoplasma do milho e o conhecimento sobre a sua variação e utilização são um esforço de grande importância não apenas para a segurança alimentar, redução da fome e da pobreza, mas também para o fornecimento de uma base sólida de conhecimentos para as futuras gerações de investigadores em diferentes áreas da genética, melhoramento e biotecnologia (WALTERS, 2004; CGIAR, 2007; PRIYANKA *et al.*, 2021).

Estima-se que haja mais de 135.000 acessos de milho conservados em bancos de germoplasma em todo o mundo. Entretanto, esse recurso, em grande parte ainda inexplorado, enfrenta problemas duradouros ao longo dos anos: documentações inconsistentes e, por vezes, incompletas, com difícil acesso das bases de dados, caracterizações fragmentadas, difícil integração entre os recursos de germoplasma e o melhoramento, além da falta de orçamentos adequados para a regeneração, intercâmbio ativo, e caracterização genética do germoplasma (CGIAR, 2007) Praticamente todos os bancos de germoplasma de plantas no Brasil têm enfrentado desafios estruturais para caracterizar RG. É necessário que os RGs existentes deixem de ser apenas um repositório estático de diversidade e passem a contribuir com atividades cientificamente importantes e enriquecedoras (MCCOUCH *et al.*, 2012; MCCOUCH *et al.*, 2020).

A caracterização de recursos genéticos para características fenotípicas complexas tais como produtividade, composição bioquímica, tolerância a estresse, resistência a patógenos etc. é um processo fundamental, porém caro, tecnicamente complexo e demorado, que geralmente contempla apenas uma fração dos acessos de um BAG. A análise genômica com marcadores moleculares, por outro lado, embora não forneça diretamente informações quanto à variação fenotípica, pode ser aplicada a um grande número de acessos provendo rapidamente e de forma econômica dados genéticos, isentos de efeito ambiental, sobre a identidade, diversidade e ancestralidade genética, permitindo inferências sobre redundâncias, estrutura genética, eventos

fundadores, introgressões e assinaturas de seleção natural ou direcional resultantes do processo de domesticação (TANKSLEY; MCCOUCH, 1997). Adicionalmente, marcadores moleculares, quando associados a dados fenotípicos, permitem estudos de genética de associação, mapeamento de QTLs (locos de características quantitativas - *quantitative trait locos*), construção de modelos de seleção genômica (LANGRIDGE; WAUGH, 2019; HÖLKER *et al.*, 2022) e descoberta de variantes de grande efeito para direcionar projetos de edição gênica (ZSÖGÖN *et al.*, 2018).

A viabilidade técnica da caracterização genética de grandes coleções de germoplasma com marcadores moleculares apresenta desafio diferenciado a depender do modo predominante de reprodução da espécie, que por sua vez determina a distribuição da variabilidade genética entre e dentro dos acessos de um BAG. No caso de espécies autógamas, ou seja, que realizam predominantemente autofecundação tais como arroz, feijão e trigo, os acessos são linhagens, e, portanto, é possível assumir com certa segurança que a análise genética de apenas um indivíduo por acesso será adequadamente representativa do acesso. É importante lembrar, entretanto, que mesmo em plantas autógamas dados de uma única planta podem ser errôneos quando a extensão da diversidade intra-acesso é maior do que o esperado por diferentes razões, incluindo um nível mais alto do que o esperado de fecundação cruzada, heterogeneidade fenotípica, mistura de sementes, contaminação de pólen e a presença de plantas fora do padrão (off-types) (ARCA *et al.*, 2021).

Já no caso de espécies alógamas, ou seja, de fecundação cruzada como o milho, salvo linhagens desenvolvidas de forma deliberada, um acesso sabidamente corresponde a uma população com ampla diversidade entre os indivíduos que o compõem. A caracterização de acessos de milho que não são linhagens, como no caso de variedades tradicionais e indígenas, demanda, portanto, a análise de uma amostra representativa de indivíduos para cada acesso visando assim capturar a diversidade existente não apenas entre acessos, mas também dentro de um acesso, tornando o processo potencialmente inviável economicamente, principalmente para grandes coleções (DUBREUIL; CHARCOSSET, 1998; DUBREUIL *et al.*, 1999; ARCA *et al.*, 2021).

Para viabilizar a análise genética de um grupo de indivíduos geneticamente distintos de um acesso ou população a custos reduzidos, a abordagem de análise de “bulks” ou “pools” de indivíduos, também chamada de “alelotipagem”, pode ser utilizada como uma alternativa conveniente para a genotipagem individual permitindo estimar as frequências alélicas do conjunto de indivíduos com acurácia. Em outras palavras, em vez de um dado genotípico para

cada um dos marcadores moleculares analisados, são obtidas estimativas das frequências alélicas dos marcadores. Os vários acessos podem assim ser analisados quanto à diversidade, estrutura e distância genética entre si a partir de suas frequências como se fossem várias populações (SHAW *et al.*; 1998; ARCA *et al.*, 2020).

O sucesso da abordagem de alelotipagem depende evidentemente da acurácia com a qual as frequências alélicas são estimadas. Esta acurácia por sua vez depende em primeira análise do tamanho da amostra de indivíduos analisados para cada acesso, aspecto este presente também caso a análise de representantes fosse realizada individualmente. Entretanto, mais especificamente no caso da análise genética de um bulk de indivíduos com marcadores moleculares dois aspectos adicionais impactam a acurácia das estimativas. O primeiro é o próprio procedimento de montagem do bulk de indivíduos e das suas respectivas amostras de DNA. Caso a contribuição de DNA de cada indivíduo não seja preparada de forma praticamente equimolar, as estimativas poderão ser enviesadas, dominadas por aqueles indivíduos equivocadamente mais representados no bulk (DUBREUIL *et al.* 1999; ARCA *et al.*, 2021).

A segunda fonte de variabilidade nas estimativas é a própria natureza do marcador molecular. No caso de marcadores multialélicos como microsatélites, problemas de competição entre alelos durante a reação de PCR (reação de polimerização em cadeia) são comuns podendo causar sub-representação (“*drop-out*”) de alelos, resultando em vieses importantes das estimativas de suas frequências, além do fato de que alelos raros tenderão a não ser amostrados. Nesse quesito, marcadores bi-alélicos SNPs apresentam uma importante vantagem para a aplicação da alelotipagem. O fato que somente dois alelos são detectados, a competição será limitada a dois alelos, além da técnica de genotipagem em microarranjos (chips), ou mesmo via sequenciamento, não sofrer de problemas de “*drop-out*” tornando a estimativa de ambos significativamente mais robusta (LE HELLARD *et al.*, 2002; ARCA *et al.*, 2021).

A abordagem de alelotipagem foi utilizada inicialmente com marcadores RFLP (ARNHEIM; STRANGE; ERLICH, 1985) e microsatélites (SHAW *et al.*, 1998) e, mais recentemente, com “chips” de SNPs de alta densidade e sequenciamento “*Next generation*” principalmente para a realização de estudos de associação “caso-controle” em seres humanos (SHAM *et al.*, 2002; TEUMER *et al.*, 2013; SCHLÖTTERER *et al.*, 2014) e em menor escala para estudos de populações naturais de peixes (JOHNSTON *et al.*; 2013; OZEROV *et al.*, 2013) e de populações de milho (REBOURG; GOUESNARD; CHARCOSSET, 2001; MIR *et al.*, 2013; ARCA *et al.*, 2021) e arroz (GOUDA *et al.*, 2020). Apesar desta abordagem técnica ter

sido proposta na literatura há mais de 30 anos, não há relato de trabalhos de alelotipagem de coleções de germoplasma de plantas no Brasil e raros são mesmo internacionalmente.

A coleção de milho do Brasil, umas das maiores a nível mundial, conta com mais de 4.000 acessos conservados nos Bancos de Germoplasma da EMBRAPA e no *Centro Internacional de Mejoramiento de Maíz y Trigo* (CIMMYT). Esses acessos, provenientes de coletas em diferentes regiões do país, populações com algum tipo de melhoramento genético e introduções de outros países, são mantidos no Banco Ativo de Germoplasma de Milho (BAG Milho) do Centro Nacional de Pesquisa de Milho e Sorgo (CNPMS-EMBRAPA) (ABADIE *et al.*, 2000; TEIXEIRA *et al.*, 2021). O BAG de Milho conta, também, com a Coleção Nuclear de Germoplasma de Milho, uma amostra representativa do BAG Milho, estabelecida para manter a máxima variabilidade genética do BAG com o mínimo de redundância. Essa coleção, também conservada atualmente no Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN-EMBRAPA), conta com cerca de 300 acessos. Ela foi desenvolvida para permitir uma maior rapidez no acesso e avaliação do germoplasma do milho, sendo a primeira escolha para trabalhos que demandem a utilização do germoplasma, programas de melhoramento ou estudos de natureza básica (ABADIE *et al.*, 2000).

Até o momento, pouco se sabe sobre a diversidade genética molecular efetivamente conservada nessas coleções. Apesar de mais de 80% dos acessos do BAG terem sido caracterizados de acordo com descritores morfológicos da cultura do milho (TEIXEIRA; COSTA, 2010; TEIXEIRA *et al.*, 2019), poucos são os trabalhos de caracterização da diversidade genética desse BAG utilizando marcadores moleculares. Esses estudos de caracterização, entretanto, se restringem a poucos acessos de subgrupos das coleções: milho-dentado (NETTO; SOUZA; OLIVEIRA, 2009) e milho-duro (NETTO *et al.*, 2004), realizados com marcadores AFLP, milho-doce (PINTO; TEIXEIRA; GUIMARÃES, 2016) e milho crioulo (VIEIRA, 2011), conduzidos com marcadores microsatélites. Portanto, os raros trabalhos existentes foram realizados com marcadores moleculares de baixa robustez analítica e transferibilidade, o que impossibilita comparações com bases de dados internacionais existentes que tem utilizado dados de marcadores SNPs (polimorfismos de base individual) ou sequenciamento (KUMAR, B., *et al.* 2022).

No BAG de Milho há, ainda, acessos classificados em raças indígenas e/ou originários de terras indígenas que compõem a coleção indígena desse BAG. Esses acessos se organizam em cinco raças conhecidas: Raça Caingang, Raça Entrelaçado, Raça Lenha, Raça Moroti e Raça Guarani. Entretanto, apesar dessa coleção também ter sido caracterizada para um amplo

conjunto de descritores morfológicos e de produtividade (TEIXEIRA *et al.*, 2021), as variedades indígenas ainda carecem de qualquer estudo genômico para melhor esclarecimento sobre sua origem, estrutura, diversidade genética e domesticação. Além disso, pouco ou nada se sabe sobre a sua divergência em relação às cultivares brasileiras comerciais antigas ou recentes.

Tendo em vista a importância dos recursos genéticos de milho, a disponibilidade de uma coleção estruturada de germoplasma da espécie, e considerando a utilização ainda tímida destes recursos em programas de melhoramento, este trabalho visou contribuir para incrementar o ainda precário conhecimento da diversidade genética em nível molecular do germoplasma de milho conservado no Brasil. Para isso foi testada e utilizada a abordagem analítica de alelotipagem ainda pouco comum, que permitiu a geração de dados genéticos de marcadores SNPs com alta qualidade e portabilidade em todo o genoma para quase 12.000 plantas de 481 acessos, constituindo assim o mais amplo estudo genético de germoplasma de milho realizado no país. Os dados foram utilizados para obter estimativas de distância e estrutura genética dos acessos, as quais foram confrontadas com suas classificações morfológicas tais como tipos de grãos de milho, nível de domesticação e origem do material coletado. Os resultados indicam que os dados genômicos, ao capturar a origem ancestral e histórico dos acessos de forma não enviesada, fornecem informações adicionais muitas vezes inesperadas e importantes para uma melhor tomada de decisão de conservação e uso destes recursos genéticos.

2. OBJETIVOS

2.1 OBJETIVO GERAL

Investigar a diversidade, estrutura e relacionamento genético de um conjunto de acessos da coleção nuclear de milho incluindo cultivares comerciais antigas e modernas e uma coleção de variedades brasileiras indígenas de milho organizadas em raças.

2.2 OBJETIVOS ESPECÍFICOS

1. Avaliar diferentes métodos de preparação de amostras de DNA de acessos de milho em “bulks” visando a seleção de um método eficiente para utilização em larga escala para alelotipagem de acessos de milho;

2. Alelotipar 188 acessos de variedades indígenas e 293 acessos da coleção nuclear de milho da EMBRAPA em bulks para 3.526 SNPs interrogados no *EMBRAPA 65kMultispecies chip*;
3. Realizar análises de dados visando caracterizar a diversidade, estrutura, distância e relacionamento genético entre todos os 481 acessos de milho e seus subgrupos ou raças;
4. Realizar análises conjuntas com dados dos mesmos SNPs extraídos da literatura para acessos de coleções internacionais de germoplasma de milho bem como de populações naturais de teosinto.

3. REVISÃO BIBLIOGRÁFICA

3.1. BOTÂNICA

O milho (*Zea mays* L.) é uma monocotiledônea pertencente à família Gramineae/Poaceae, é uma planta herbácea anual e tipicamente monoica, ou seja, apresenta os dois órgãos reprodutivos, masculino e feminino (HALEVY *et al.*, 1985; MAGALHÃES; DURÃES; PAIVA, 1995; MAGALHAES *et al.*, 2002). As características vegetativas e reprodutivas básicas da planta do milho são resultados da supressão, condensação e multiplicação de diversas partes da anatomia característica das gramíneas. Ademais, essas características do milho podem sofrer alterações decorrentes de interações com fatores ambientais que interferem no controle da ontogênese dessas plantas. No entanto, como resultado geral da seleção natural e da domesticação dessa espécie, a planta do milho é uma eficiente produtora de grãos, comumente robusta, ereta, anual, com altura geralmente maior que dois metros (MAGALHÃES; DURÃES; PAIVA, 1995; MAGALHAES *et al.*, 2002).

O milho apresenta raízes fasciculadas, em que aparecem raízes primárias e seminais, adventícias ou de suporte, com grande crescimento, podendo atingir, em média, até 3 m de comprimento e 0,45 quilos por planta (MAGALHÃES; DURÃES; GOMIDE, 1996; FREITAS, 2019). O crescimento das raízes é diretamente influenciado pela disponibilidade de carboidratos produzidos e armazenados nas partes aéreas da planta (MAGALHÃES; DURÃES; GOMIDE, 1996). Os colmos são esponjosos, em geral sem ramificações laterais e possuem nós e entrenós bem visíveis, funcionam como suporte para as folhas e partes florais, além de servir como órgão de reserva, com armazenamento de sacarose (MAGALHÃES; DURÃES; GOMIDE, 1996; FREITAS, 2019).

As folhas são estreitas, possuem comprimento maior do que a largura e aparecem nos nós de forma alternada, podendo variar entre 5 e 48 folhas, dependendo da variedade e dos fatores ambientais (FREITAS, 2019). Tipicamente monoica, a planta do milho apresenta ainda, além das folhas, inflorescências femininas e masculinas. As inflorescências masculinas estão organizadas em panícula, surgindo no ápice do colmo, enquanto as femininas estão agrupadas em espigas axilares altamente compactadas, dispostas no eixo principal do colmo (HALEVY *et al.*, 1985; FREITAS, 2019). Afetado principalmente pela temperatura, o florescimento ocorre cerca de 50 a 100 dias após a germinação das sementes, que, por sua vez, ocorre em aproximadamente 5 a 6 dias (MAGALHÃES; DURÃES; GOMIDE, 1996).

O milho é uma das plantas existentes na natureza que possui maior eficiência no armazenamento energético, devido à sua grande capacidade de produzir e acumular fotoassimilados. Suas raízes, folhas, colmo e partes florais funcionam como uma eficiente fábrica de energia que será armazenada em grandes quantidades no grão de milho (MAGALHÃES; DURÃES; GOMIDE, 1996; MAGALHAES *et al.*, 2002). De apenas uma semente com um pouco mais de 0,3 g surgirá, em cerca de 9 semanas, uma planta em geral com mais de 2,0 m de altura. Nos meses subsequentes, essa planta produzirá por volta de 600 a 1.000 sementes semelhantes à semente da qual a planta se originou (ALDRICH; SCOTT; LENG, 1982; MAGALHÃES; DURÃES; GOMIDE, 1996).

3.2 PRODUÇÃO

Ao longo das últimas décadas, o milho ultrapassou antigos concorrentes, como o arroz e o trigo, e atingiu o patamar de maior cultura agrícola do mundo, sendo a única a ter superado a marca de 1 bilhão de toneladas (CONTINI *et al.*, 2019). Além da extrema importância mundial em termos de produção, a cultura do milho tem grande valor econômico caracterizado pelos diversos usos, com seus produtos e subprodutos que integram toda uma cadeia alimentícia (DEMÉTRIO *et al.*, 2008; CONTINI *et al.*, 2019; FREITAS, 2019). São estimadas mais de 3.500 utilizações desse cereal, que vai desde a alimentação animal, destinação da maior parte da produção desse grão, até a indústria de alta tecnologia. A cultura do milho possui alta relevância na segurança alimentar, animal e humana, além de fazer parte da produção de diversos produtos como combustíveis, bebidas, polímeros e outros (DEMÉTRIO *et al.*, 2008; CONTINI *et al.*, 2019).

Devido à grande utilização dessa cultura principalmente na alimentação de frangos e suínos, a produção mundial de milho atingiu um aumento de 82% da safra de 2000/01 para a de 2017/18, passando de 591 milhões de toneladas para 1.076 bilhão de toneladas (CONTINI *et al.*, 2019). A produção mundial, liderada pelos Estados Unidos, com 32% (389.7 milhões de toneladas) da produção, seguidos da China com 24% (288.9 milhões de toneladas), é concentrada em poucos países, uma vez que apenas esses dois países são responsáveis por 56% da produção mundial de milho. Somados o Brasil e a União Europeia, que representam 10% e 5% da produção mundial, respectivamente, aos dois maiores produtores mundiais, os quatro são responsáveis por um pouco mais de 70% da produção mundial. Em nível mundial, alguns países têm apresentado expressivo aumento da produção bem acima da média: Argentina, responsável por 4% da produção mundial, Índia (3%), Ucrânia (3%), México (2%), Rússia (1%) e Canadá (1%) (USDA, 2023a).

O milho é um cereal fundamental para a agricultura brasileira, cultivado em todas as regiões do país, sendo os maiores produtores brasileiros os estados do Mato Grosso, Paraná, Goiás, Mato Grosso do Sul e Minas Gerais (CONTINI *et al.*, 2019). Desde a década de 70, a produção de milho no Brasil se expandiu expressivamente, passando de 20 milhões de toneladas para a produção recorde de 131.89 milhões de toneladas em 22.267,4 milhões de hectares na safra de 2022/23 (CONAB, 2023).

O calendário agrícola da produção de milho no Brasil, especificamente, é caracterizado pela divisão em duas épocas de plantio: primeira safra, ou plantio de verão, para o plantio de milho realizado na época tradicional durante o período chuvoso, que varia entre o fim de agosto até os meses de outubro e novembro, e a segunda safra, ou safrinha, para o plantio de milho realizado extemporaneamente de janeiro a março. A safrinha, caracterizada pelo milho de sequeiro, ocorre, principalmente, na região Centro-Oeste e nos estados do Paraná, São Paulo e Minas gerais, quase sempre na sequência do cultivo da soja precoce. Além disso, embora semeada em condições desfavoráveis de clima, a safrinha é conduzida dentro de sistemas de produção que tem sido gradativamente aprimorados e adaptados a essa condição, contribuindo para elevar os rendimentos das lavouras de milho (FILHO; CRUZ; GARCIA, 2021).

3.3 MERCADO GLOBAL

Segundo o Departamento de Agricultura dos Estados Unidos (USDA), as exportações do milho são concentradas por quatro países: Brasil, Estados Unidos, Argentina e Ucrânia.

Juntos, esses quatro países representam mais de 80% das exportações mundiais do cereal. Em 2023, O Brasil superou os Estados Unidos e se tornou o maior exportador de milho do mundo de acordo com dados do USDA referentes à safra de 2022/23. Nessa safra, o Brasil foi responsável por 32% das exportações globais de milho, exportando 50 milhões de toneladas de milho, enquanto os Estados Unidos representaram 23% das exportações, cerca de 42 milhões de toneladas (USDA, 2023b).

O avanço na exportação de milho pelo Brasil é creditado não apenas pela safra recorde do país, mas também à diversificação das rotas logísticas de exportação, aumentando as exportações a partir de portos localizados no norte do país, sobretudo no Pará, que antes se concentravam, principalmente, no estado de São Paulo. Além disso, essa mudança nas posições de liderança ressalta a crescente influência do Brasil no mercado global do milho, que ultrapassou os Estados Unidos também como principal fornecedor de milho da China, após acordo fitossanitário entre os dois países. Ademais, o país vem impulsionando a crescente produção de etanol de milho. A última vez que os Estados Unidos perderam a liderança da exportação do milho foi em 2013, ano em que o país enfrentou perdas significativas na produção ocasionadas por uma seca severa (USDA, 2023b).

Ainda referente à safra 2022/23, a China ultrapassou o Japão, dominando as importações em 2023, tendo adquirido 7,9 milhões de toneladas de milho. Em seguida, o Japão importou 4 milhões de toneladas de janeiro a setembro de 2023, seguido pelo Vietnã (2,6 milhões de toneladas), Coreia do Sul (2,3 milhões de toneladas) e Irã (1,9 milhões de toneladas) (USDA, 2023b).

Apesar da importância do milho como maior cultura agrícola do mundo, o comércio global dessa commodity apresenta, em relação à produção, um percentual baixo, representando somente 14% da produção mundial em 2017/18, enquanto o comércio da soja representou 45,2% da produção mundial no mesmo período, liderando as exportações do setor agropecuário (CONTINI *et al.*, 2019).

3.4 ORGANIZAÇÃO DO GENOMA

Com um tamanho estimado de 2.178 Mb na versão mais recente e de maior qualidade do genoma de telômero a telômero (CHEN *et al.* 2023), o genoma do milho corresponde a aproximadamente 70% do tamanho do genoma humano, e cerca de seis vezes o genoma do arroz (390 Mb) (ZHOU *et al.* 2020) e quase vinte vezes maior que o genoma de *Arabidopsis*

(135 Mb) (LIAN *et al.* 2024), primeira planta sequenciada cujo genoma foi sequenciado em sua totalidade (CHANDLER; BRENDDEL, 2002; MESSING *et al.*, 2004). Ainda que comparável quanto ao tamanho, o genoma do milho tende a ser mais complexo do ponto de vista estrutural que o genoma humano devido à elevada porcentagem de elementos altamente repetitivos (CHANDLER; BRENDDEL, 2002, COE *et al.*, 2002). Uma estimativa inicial sugeriu a presença de 59.000 genes no genoma nuclear do milho, consideravelmente maior do que a quantidade de genes estimada para o arroz (45.000 genes), entretanto com uma densidade média menor de 1 por 40 Kb (MESSING *et al.*, 2004). A análise mais recente demonstrou, entretanto, um número de 42.850 genes, ou seja, essencialmente igual ao número de genes do arroz (CHEN *et al.* 2023).

Os genes do milho corresponde a apenas 20% do genoma (CHANDLER; BRENDDEL, 2002), enquanto grande parte (60 a 80%) do genoma do milho é composto por sequências repetitivas, representadas, principalmente, por retrotransposons, cujos elementos transponíveis são mediados por RNA (FLAVELL *et al.*, 1974; BENNETZEN, 1996; SANMIGUEL *et al.*, 1996; MESSING *et al.*, 2004). Desse modo, os genes estão organizados em ilhas de diferentes tamanhos espalhadas por um mar de retrotransposons altamente conservados e de elevado número de cópias e outras sequências repetitivas (SANMIGUEL *et al.*, 1996; CHANDLER; BRENDDEL, 2002).

Os genes apresentam regiões transcritas com, em média, 4 a 5 Kb e pequenos íntrons, além disso, quando comparados com as sequências repetitivas, os genes ativos apresentam comumente padrão de submetilação (GUIMARÃES, 2002). Os retrotransposons estão organizados aninhados nas regiões intergênicas, possuem natureza intercalada e são intimamente associados aos genes, aparecendo ao lado da maioria desses, possuindo, ainda, elevado número de cópias, variando de 10 a 30.000 cópias (SANMIGUEL *et al.*, 1996).

3.5 RAÇAS DE MILHO NO BRASIL

O sistema racial de classificação do milho foi sugerido originalmente por Paterniani e Goodman (1977). Os autores organizaram as variedades de milho tradicionais brasileiras em quatro grupo de raças: indígenas, comerciais antigas, comerciais recentes e comerciais exóticas. Segundo Paterniani, Nass e Santos (2000), vários tipos de milho já domesticados, cultivados por povos indígenas, foram encontrados pelos europeus quando chegaram à América. O homem ocidental adotou os milhos indígenas cultivados pelos autóctones na América do Sul, que se desenvolveram nas raças comerciais antigas. Posteriormente, milhos de outras localidades da

América Central e América do Norte foram introduzidos, resultando nas raças comerciais recentes. As raças comerciais recentes são decorrentes, principalmente, da introdução de milho norte-americano no Brasil trazido pelos imigrantes americanos no decorrer da Guerra da Secessão, por volta de 1860. Muitos cruzamentos ocorreram com a introdução dos milhos norte-americanos, dando origem às variedades que apresentam germoplasma de raças comerciais antigas em combinação com os milhos introduzidos.

3.5.1 Raças Indígenas

As raças indígenas de milho são aquelas cultivadas por povos indígenas e que não foram cultivadas, pelo menos em larga escala, por outros (PATERNIANI; GOODMAN, 1977). Essa diferenciação é feita uma vez que essas raças estão, provavelmente, em um estado muito semelhante ao do milho cultivado dos tempos pré-colombianos (PATERNIANI; GOODMAN, 1977, COSTA *et al.*, 2022). Dentro do grupo de raças indígenas estão a raça Moroti, com suas sub-raças Moroti Precoces e Moroti Guapi, e as raças Caingang, Lenha e Entrelaçado (BRIEGER *et al.*, 1958; PATERNIANI; GOODMAN, 1977). Paterniani e Goodman (1977) descrevem entre os milhos indígenas, ainda, os de pipoca desenvolvidos principalmente pelos indígenas de etnia Guarani, no Paraguai.

A maioria dos milhos de raças indígenas são farinhentos, muito moles, utilizados sobretudo para moagem e produção de farinha (PATERNIANI; NASS; DOS SANTOS, 2000), o que confirma a preferência por esse tipo de milho pelos povos indígenas de todas as Américas. Na mesma área, milhos de grãos duros e dentados eram preferidos pelo homem ocidental que, pelo que se sabe, nunca cultivaram raças farináceas (PATERNIANI; GOODMAN, 1977).

3.5.2 Raças Comerciais Antigas

As antigas raças comerciais são, de certo modo, indígenas. No entanto, divergem do grupo de raças indígenas uma vez que passaram por mudanças desde o cultivo inicial pelo homem ocidental e, hoje, são amplamente cultivadas, razão pela qual são chamadas de comerciais. Dentro do grupo de raças comerciais antigas estão as raças Cristal, Cristal Sulino, *Canário de Ocho*, Cateto, Cateto Sulino, Cateto Sulino Precoces, Cateto Sulino Grosso e Cateto Nortista (PATERNIANI; GOODMAN, 1977).

3.5.3 Raças Comerciais Recentes

As raças comerciais recentes são raças modernas de milho, todas, provavelmente, com menos de cem anos. Essas raças só foram estabelecidas ou desenvolvidas no Brasil após o período pré-colombiano e foram introduzidas de outras áreas ou obtidas pela hibridação natural de raças introduzidas com as raças anteriormente mencionadas. A maioria das raças são do tipo dentado, apesar do tipo semi-dentado também ser encontrado e, devido ao seu alto rendimento quando comparado ao Cateto local, esse milho dentado ganhou alta aceitação local após sua introdução. Dentro do grupo de raças comerciais recentes estão as raças Dente Riograndense, Dente Paulista, Dente Branco, Semi-Dentado e Cravo (PATERNIANI; GOODMAN, 1977; TEIXEIRA, 2008).

3.5.4 Raças Comerciais Exóticas

As raças comerciais exóticas também foram introduzidas no Brasil a partir de outros países, entretanto, mais recentemente quando comparadas às raças comerciais recentes. As raças classificadas como exóticas são: *Hickory King*, cultivada há muitos anos no país, e *Tusón*, encontrada no estado da Bahia há apenas alguns anos, depois de liberada por estações experimentais locais. Além de *Hickory King* e *Tusón*, outras fontes de germoplasma exótico estão se popularizando recentemente, como as variedades Asteca e Maia, desenvolvidas pelo Instituto Agrônomo de Campinas, e Piramex, desenvolvida pela Escola Superior de Agricultura Luiz de Queiroz (Esalq) a partir de variedades de *Tuxpeño*. As cultivares Pérola Piracicabana, resultado da combinação do Cateto e milhos duros da Colômbia, também desenvolvida pela Esalq, e a cultivar Centralmex, resultado da combinação de Piramex e milhos da América Central, também são representantes de milhos introduzidos que estão ganhando rapidamente ampla popularização (PATERNIANI; GOODMAN, 1977; TEIXEIRA, 2008).

A produção de milho no Brasil aumentou substancialmente a partir da introdução de cultivares exóticas, possibilitando avanços significativos no melhoramento genético dessa cultura. Além disso, a base genética dessa espécie no Brasil é derivada, quase que exclusivamente, de germoplasma exótico introduzido, principalmente dos Estados Unidos e do México (TEIXEIRA, 2008).

3.6 RECURSOS GENÉTICOS

Os recursos genéticos de milho constituem um tesouro imensurável para a humanidade. As estratégias de conservação do germoplasma do milho e o conhecimento sobre a sua variação e utilizações são um esforço de grande importância não apenas para a segurança alimentar, redução da fome e da pobreza, mas também para a sustentabilidade do melhoramento genético da cultura, para o fornecimento de uma base sólida de conhecimentos para as futuras gerações de investigadores e utilizadores tecnológicos e para a mitigação de sua erosão genética (WALTERS, 2004; CGIAR, 2007; PRIYANKA *et al.*, 2021).

Estima-se que haja mais de 135.000 acessos de milho conservados em bancos de germoplasma em todo o mundo, entre eles, um dos maiores é o Centro Internacional de Melhoramento de Milho e Trigo, CIMMYT (*Centro Internacional de Mejoramiento de Maíz y Trigo*), no México, com quase 30.000 acessos. O CIMMYT não apenas desenvolve, mas também distribui linhagens e híbridos de milho melhorados para o mundo inteiro, trabalho do qual já resultaram cerca de 69 variedades de milho. Esse trabalho tem auxiliado na melhoria da segurança alimentar e no aumento de renda e dos meios de subsistência de pequenos agricultores e suas famílias (CGIAR, 2024; CIMMYT, 2024). Outras principais coleções milho, com mais de 10.000 exemplares, estão localizadas na Rússia e nos Estados Unidos. Além dessas principais coleções de germoplasma, nos Estados Unidos há ainda o *Maize Genetics Cooperation Stock Center*, na Universidade de Illinois. Este centro atende à comunidade de pesquisa de milho coletando, mantendo e distribuindo sementes de estoques genéticos de milho, e fornecendo informações sobre estoques de milho e as mutações que eles carregam por meio do *Maize Genetics and Genomics Database* (MaizeGDB) (CGIAR, 2024).

O Banco Ativo de Germoplasma de Milho (BAG Milho) do Brasil é mantido no Centro Nacional de Pesquisa de Milho e Sorgo (CNPMS-EMBRAPA). Com o principal intuito de preservar a agrobiodiversidade do milho, diversas ações como conservação de sementes, multiplicação e regeneração de acessos, introdução de novas fontes de variabilidade, intercâmbio, caracterização, avaliação agronômica, documentação, e promoção do uso das coleções são desenvolvidas nesse banco. O BAG de milho conjuntamente com a coleção base de milho, preservada no Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN-EMBRAPA), somam, atualmente, mais de 4.000 acessos, a maioria mantida como duplicatas em ambas as instituições para aumentar a segurança da preservação. Entre esses acessos, 234 foram classificados em raças indígenas e/ou originários de terras indígenas (TEIXEIRA *et al.*, 2021).

Essas 234 variedades indígenas foram caracterizadas para um amplo conjunto de descritores morfológicos e de produtividade produzindo um importante conjunto de informações para auxiliar no processo de conservação e utilização por programas de melhoramento. Os 234 acessos se organizam em cinco raças conhecidas com os respectivos números de acessos, da seguinte forma: (1) Raça Caingang com 53 acessos; (2) Raça Entrelaçado com 38 acessos; (3) Raça Lenha com 42 acessos; (4) Raça Moroti com 60 acessos e (5) Raça Guarani com cinco acessos. Os 37 acessos que completam o total de 234, embora sejam classificados como sendo de origem indígena, ainda não receberam classificação racial. A grande maioria dos 234 acessos indígenas podem ser atribuídos a diferentes etnias em várias regiões brasileiras principalmente no Mato Grosso, Paraná, Rio Grande do Sul e Amazonas (TEIXEIRA *et al.*, 2021). Essas 234 variedades indígenas ainda carecem de qualquer estudo genômico para melhor esclarecimento sobre sua origem, estrutura, diversidade genética e domesticação. Além disso pouco ou nada se sabe sobre a sua divergência em relação às cultivares brasileiras comerciais antigas ou recentes.

O BAG de Milho conta ainda com a Coleção Nuclear de Germoplasma de Milho, uma amostra representativa da coleção base, estabelecida para manter a máxima variabilidade genética do BAG com o mínimo de redundância. Essa coleção, conservada atualmente no CENARGEN e no CNPMS, que conta com 300 acessos, foi desenvolvida para permitir uma maior rapidez no acesso e avaliação do germoplasma do milho, sendo a primeira escolha para trabalhos que demandem a utilização do germoplasma, programas de melhoramento ou estudos de natureza básica (ABADIE *et al.*, 2000).

3.7 MARCADORES MOLECULARES E ESTUDOS DE DIVERSIDADE GENÉTICA

O sucesso da domesticação, conservação e melhoramento genético de uma espécie está estritamente relacionado ao entendimento da magnitude e organização da sua diversidade genética (GOVINDARAJ; VETRIVENTHAN; SRINIVASAN, 2015). Portanto, informações sobre o grau de diversidade genética, relações e estrutura dentro de um determinado grupo de indivíduos são prementes para contribuir com essas atividades cientificamente importantes e enriquecedoras (SEMAGN *et al.*, 2012; SILVA, K. *et al.*, 2020). Apesar da caracterização morfo-agronômica ser tradicionalmente utilizada como medida de diversidade genética, essa abordagem muitas vezes torna-se limitada por envolver um grande número de descritores, muitos dos quais são altamente sensíveis aos efeitos ambientais (KUMAR, A. *et al.*, 2022).

Como alternativa à caracterização morfo-agronômica, a caracterização molecular permite inferir o grau de variabilidade genética entre indivíduos e populações por meio de marcas genéticas, indiferentes aos efeitos ambientais (KUMAR, A. *et al.*, 2022). Marcadores moleculares são marcadores genéticos baseados na detecção de isoenzimas ou sequências de DNA que revelam polimorfismos, os quais surgem devido a mutações de diferentes tipos (mutações de ponto, inserções, deleções, rearranjos estruturais, transposições etc.) na sequência do genoma, e tornam possível identificar diferenças genéticas entre indivíduos e espécies (KUMAR *et al.*, 2009; FALEIRO, 2011; IDREES; IRSHAD; 2014). Esses marcadores em geral apresentam herança mendeliana, são abundantes e em geral independem das condições do ambiente, salvo alterações epigenéticas, e possibilitam descrições detalhadas da diversidade, ancestralidade e estrutura genética de populações (FERREIRA; GRATTAPAGLIA, 1998; FALEIRO, 2007).

Introduzidos na década de 80, marcadores moleculares provaram ser úteis para os estudos de avaliação de diversidade genética vegetal, uma vez que são potencialmente informativos e mais estáveis que a descrição morfo-agronômica no que se refere à precisão com a qual refletem ancestralidade comum ou divergente. Marcadores têm sido amplamente utilizados para caracterização de germoplasma, mostrando-se úteis para um melhor manejo e conservação dos recursos genéticos e uso mais eficiente de coleções de germoplasma em programas de melhoramento e intercâmbio de germoplasma (CHIORATO *et al.*, 2007; ARCA *et al.*, 2021). Os marcadores auxiliam na estimativa da diversidade genética existente e no reconhecimento do impacto de forças evolutivas tais como endogamia, gargalos genéticos, assinaturas de seleção natural e direcional, além de possibilitar a identificação de relações de parentesco entre os indivíduos. Marcadores permitem ainda o rastreamento da introgressão de genes, a definição da direção de cruzamentos e a seleção de progenitores, ferramentas importantes em programas de melhoramento (PEREIRA; PEREIRA, 2006; DESAI; SINGH, 2015).

Diferentes classes de marcadores moleculares são utilizadas para diferentes aplicações a depender da sua taxa evolutiva, cobertura genômica e nível de resolução demandado. Aplicações incluem estudos populacionais, estudos evolutivos, diversidade genética inter e intraespecífica, identificação de indivíduos, cultivares e novas variantes, testes de parentesco, detecção de mutações, epidemiologia, entre outras (SHARMA *et al.*, 2002; RUDD; SCHOOF; MAYER, 2005; FALEIRO, 2011; IDREES; IRSHAD; 2014). Estudos e pesquisas utilizando marcadores têm gerado importantes informações e dados genéticos, principalmente

relacionadas ao melhoramento de culturas, conservação, caracterização, manutenção e uso de recursos genéticos (FALEIRO, 2011; KORDROSTAMI; RAHIMI, 2015).

O polimorfismo de comprimento de fragmentos de restrição (RFLP) foi o primeiro marcador genético molecular, desenvolvido em 1980 por Bostein *et al.*, usado pela primeira vez na construção de mapas genéticos. A técnica RFLP depende da técnica *Southern blot*, em que o DNA é clivado por enzimas de restrição em diferentes fragmentos para detecção do polimorfismo (KUMAR *et al.*, 2009; KUMAR, A. *et al.*, 2022). Utilizando marcadores RFLP, Gauthier *et al.* (2002) conduziram uma caracterização em nível molecular de uma coleção de 488 populações de milho europeias. A análise de 23 locos RFLP mostrou uma alta riqueza alélica de 11,5 alelos por loco.

O estudo apontou que populações do leste europeu (Polônia, Áustria, Alemanha etc.), quando comparadas com populações do sul da Europa, apresentaram a menor diversidade genética, um menor número de alelos únicos e uma maior porcentagem de locos fixos. A classificação molecular baseada na distância de Rogers baseada em frequências alélicas, permitiu a distinção de três grupos principais altamente consistentes com origens geográficas: (1) cluster da porção do nordeste: populações iniciais ou intermediárias de países do nordeste e dos Bálcãs; (2) cluster da porção sudeste: populações tardias e parcialmente dentadas da Grécia e da Itália; (3) cluster da porção sudoeste: populações iniciais de milho do norte da Espanha, Portugal e dos Pireneus.

A diversidade pareceu maior nas regiões do sul, possivelmente onde foram feitas as primeiras introduções de milho. Além disso, foi observada uma correlação entre frequências alélicas em alguns locos e latitude e/ou longitude, tendências que podem refletir a direção do fluxo gênico entre diferentes raças de milho, como por exemplo, que populações norte-americanas e caribenhas foram introduzidas, respectivamente, no norte e sul da Europa, no passado.

Utilizando o mesmo tipo de marcadores moleculares, Mumm, Hubert e Dudley (1994) avaliaram uma classificação de 148 populações de milho dos EUA resultante da análise de cluster com base em RFLP, para determinar se representava as verdadeiras associações entre os milhos. Duas etapas da análise de cluster resultaram em uma matriz de proximidade contendo estimativas de relacionamento computadas a partir dos dados e um fenograma exibindo grupos na forma de um diagrama de árvore, posteriormente comparados pela estatística F. O agrupamento exibido no fenograma foi ainda comparado ao exibido em fenogramas resultantes

de três análises de cluster adicionais, geradas por diferentes métodos para calcular proximidades.

As populações foram agrupadas de forma semelhante nas quatro classificações, e o nível de correspondência desses agrupamentos ao agrupamento esperado com base nas informações da matriz de proximidade foi semelhante entre as classificações. Portanto, foi sugerido que existe um agrupamento natural dos milhos, geralmente refletido em cada classificação e, desse modo, considerou-se que a classificação representa razoavelmente as verdadeiras associações entre os 148 populações de milho. Marcadores RFLP também foram utilizados para determinar localizações e efeitos da introgressão de segmentos cromossômicos (ARMSTRONG; ROMERO-SEVERSON; HODGES 1992) e para identificação de ligação de um marcador RFLP ao gene *Htl* do milho que confere resistência ao fungo patógeno *Helminthosporium* (BENTOLILA *et al.*, 1991).

Marcadores do tipo RAPD, polimorfismo de DNA amplificado com iniciadores de PCR de sequência aleatória, são um tipo de marcador molecular baseado na amplificação de fragmentos de DNA pela reação em cadeia da polimerase (PCR): utilizando um único oligonucleotídeo iniciador (*primer*) curto, geralmente de dez nucleotídeos, e de sequência aleatória. A sequência amplificada internamente aos iniciadores é desconhecida (CAIXETA *et al.*, 2006; SEMAGN; BJORNSTAD; NDJIONDJOP, 2006; FALEIRO, 2011). Uma vez que o *primer* possui sequência arbitrária, elimina-se a necessidade do conhecimento prévio dos fragmentos a serem amplificados (FERREIRA; GRATTAPAGLIA, 1998; ALI *et al.*, 2004).

Carvalho *et al.* (2004) investigaram as relações genéticas entre 81 acessos de milho (79 raças locais e 2 variedades melhoradas) utilizando marcadores RAPD. Nesse estudo, trinta e dois primers altamente informativos amplificaram 255 marcadores, dos quais 184 (72,2%) eram polimórficos. Com base nos marcadores RAPD, um dendrograma foi construído e a faixa de similaridade genética foi de 0,78 a 0,91. Os dados moleculares agruparam os acessos em dois grupos principais, que foram correlacionados de acordo com as cores do grão. Além disso, pequenos grupos foram associados a características, como a morfologia do grão.

Utilizando marcadores RAPD, Okumus (2007) caracterizou 17 variedades locais de milho. Quatorze primers forneceram bandas polimórficas confiáveis e consistentes, amplificando 125 fragmentos (89%), com uma média de 8,90 fragmentos por primer. As variedades foram, também, caracterizadas com base na matriz de dissimilaridade usando a abordagem de agrupamento baseada em UPGMA e a faixa de similaridade genética foi de 0,05 a 0,88. Foi observada uma relação próxima entre os acessos de endosperma amarelo e amarelo-

alaranjado, com a dissimilaridade entre 0,08 e 0,20. O grupo contendo principalmente acessos de endosperma branco exibiu uma faixa de distância genética, ademais, os acessos isolados, que mostraram a maior distância genética, apresentam origens incertas derivada de uma amostra coletada por fazendeiros proprietários de terras.

Ainda utilizando o mesmo tipo de marcadores, Bruel *et al.* (2007) analisaram a diversidade genética entre 16 linhagens de milho. Foram utilizados 22 primers, resultando na amplificação de 265 fragmentos, dos quais 237 (84,44%) eram polimórficos, além disso, usando a abordagem de agrupamento baseada em UPGMA as associações genéticas obtidas mostraram 5 grupos heteróticos distintos.

Considerado um intermediário entre as metodologias RFLPs e RAPDs, o polimorfismo de comprimento de fragmentos amplificados (AFLP) é baseado na combinação da digestão de DNA por enzimas de restrição e PCR, duas técnicas primárias de análise. Utilizando AFLP, Oliveira *et al.* (2004) investigaram as relações genéticas entre linhagens endogâmicas de milho tropical de duas origens diferentes. Nesse estudo foram utilizadas nove combinações de primers para obtenção de marcadores AFLP, produzindo 638 bandas, 569 das quais foram polimórficas. As similaridades genéticas, baseadas no coeficiente de similaridade de Jaccard, variaram de 0,345 a 0,891, com média de 0,543. Além disso, o dendrograma baseado nas similaridades genéticas e no método UPGMA não resultou em agrupamentos bem definidos. Desse modo, para separar as linhagens em grupos mais precisos, os autores realizaram um procedimento de otimização, identificando, então, 17 grupos, em que foi detectada associação entre o tempo de floração e os pools de germoplasma.

Laborda *et al.* (2005) analisaram 85 linhagens com 569 bandas AFLP e 50 locos microssatélites. Os marcadores utilizados revelaram variabilidade substancial entre as linhagens, com altas taxas de polimorfismo. Zhang, Wan e Pan (2006) utilizaram 63 combinações de primers AFLP para detecção de polimorfismos de duas linhagens endogâmicas, para posterior mapeamento de QTLs de resistência à podridão da espiga por *Fusarium moniliforme*.

Marcadores SSR (*Simple Sequence Repeats*), comumente conhecidos como microssatélites, são pequenas sequências de DNA, com 2 a 5 pb, repetidas em tandem, cuja detecção é feita por meio da amplificação de fragmentos de DNA contendo essas sequências por PCR, utilizando *primers* específicos (de 20 a 30 bases) (FERREIRA; GRATTAPAGLIA, 1998; SEMAGN; BJORNSTAD; NDJIONDJOP, 2006; FALEIRO, 2011). Warburton *et al.* (2002) caracterizaram 7 populações do (CIMMYT) e linhagens endogâmicas utilizando 85

marcadores microssatélites, entretanto, 53 desses marcadores foram considerados os mais discriminatórios para posteriores estudos de *fingerprinting*.

As sete populações apresentaram diversidade genética significativamente maior dentro de cada população do que entre populações, indicando que as populações são heterogêneas em nível molecular, além disso, agruparam-se conforme o previsto, a partir de dados de pedigree e agrupamento heterótico. As linhagens endogâmicas também apresentaram alto nível de diversidade genética entre si, entretanto os clusters formados baseados em microssatélites não formaram associação com as populações de origem, resultado consistente com o alto nível de diversidade nas populações de origem das linhagens endogâmicas.

Vaz Patto *et al.* (2004) avaliaram a diversidade genética de 54 linhagens endogâmicas de milho do Banco Português de Germoplasma Vegetal (BPGV), representantes da diversidade de germoplasma de milho dentado e flint (duro) português. Foram utilizadas 50 linhagens americanas de elite e, também, algumas linhagens europeias a nível de comparação. Quinze locos microssatélites distribuídos por todo o genoma do milho foram utilizados com base. Nesse estudo, foi detectado um número total de 80 alelos, com um número médio de 5,33 alelos por marcador SSR. Entre as 104 linhagens endogâmicas, o valor médio do conteúdo de informação polimórfica (PIC) foi de 0,557 e a média compartilhada da distância genética foi de 0,6301, demonstrando a existência de grande variabilidade entre as linhagens dessa coleção. A distância genética média entre as linhagens portuguesas (0,6157) foi apenas ligeiramente inferior à distância genética média entre as linhagens estrangeiras representativas (0,6262), mostrando que as linhagens portuguesas mantêm um nível de diversidade genética semelhante às linhagens estrangeiras. Quase todas as linhagens endogâmicas puderam ser distinguidos entre si na análise de agrupamento, e as linhagens portuguesas estavam presentes em todos os grupos formados, confirmando a origem mista dos materiais portugueses consistente com os dados de pedigree.

Utilizando 31 marcadores microssatélites, Adu *et al.* (2019) avaliaram a diversidade de 70 populações de milho branco desenvolvidas no *Savanna Agricultural Research Institute* (CSIR), na África do Sul, quanto à resistência/tolerância à seca e baixo teor de nitrogênio (N) no solo. Foram detectados 288 alelos no germoplasma utilizado, com uma faixa de 4 a 17 alelos por loco e um número médio de alelos de 9,60 alelos por loco. O valor médio do conteúdo de informação polimórfica (PIC) foi de 0,68, variando de 0,32 a 0,85. A heterozigosidade média obtida a partir dos marcadores foi de 0,54 e, além disso, 58,10% do total de marcadores utilizados apresentaram heterozigosidade (0,57 a 0,93) superior ao valor médio de 0,54. A diversidade genética variou de 0,35 a 0,87, com uma valor médio de diversidade genética de

0,71 indicando um alto nível de polimorfismo entre as populações. Foram revelados 5 agrupamentos para as 70 populações, baseados no coeficiente de similaridade de Jaccard e no método UPGMA. O método de Evanno, para avaliação do número mais provável de agrupamentos revelou a existência de duas subpopulações. A subpopulação 1 apresentou maior diferenciação genética ($F_{ST} = 0,18$) entre seus acessos do que a subpopulação 2 ($F_{ST} = 0,013$).

Labate *et al.* (2003) genotiparam 20 locos microssatélites em 461 plantas de milho de matrizes diversas, representantes do germoplasma das raças *Com-Belt Dents*, *Northern Flints*, *Southern Dents* e também de linhagens endogâmicas. O número médio de alelos por loco foi de 6,5 no geral, 3,17 dentro dos acessos e 3,20 dentro das linhagens endogâmicas. A heterozigosidade esperada (H_E) média foi de 0,53 dentro acessos, enquanto dentro das linhagens endogâmicas apresentou o valor médio de 0,61. Os acessos de polinização aberta (*Com-Belt Dents*) apresentaram uma tendência à consanguinidade ($F = 0,09$), e 85% da variação genética foi compartilhada entre eles. Além disso, houve uma forte distinção entre o germoplasma de milhos flint e dentados, porém não foram reveladas evidências de estrutura genética.

Marcadores SSR têm sido amplamente utilizados para a análise de diversidade genética e caracterização de germoplasma, investigando a diversidade genética de raças locais de milho, linhagens endogâmicas e de polinização aberta (KUMAR A. *et al.*, 2022). Esses marcadores foram utilizados, ainda, para comparar e identificar possíveis réplicas por meio do acoplamento de novos acessos ao *Maize Research Institute Zemun Polje* (MRIZP), na Sérvia (ANDJELKOVIC *et al.* 2018). Sharopova *et al.* (2002) desenvolveram 1.051 novos marcadores SSR para milho através de bibliotecas enriquecidas com microssatélites e pela identificação de sequências contendo microssatélites em bancos públicos e privados. Foram utilizadas três populações de mapeamento para gerar um mapa com 978 desses marcadores.

3.7 MARCADORES SNPs

A diversidade genética pode ser estimada a partir de diversos tipos de marcadores moleculares, entretanto, os avanços na tecnologia molecular produziram uma mudança em direção aos marcadores de polimorfismo de nucleotídeo único (SNPs) (HAMBLIN; WARBURTON; BUCKLER, 2007; SEMGAN *et al.*, 2012). Baseados na identificação de mutações e polimorfismos resultantes da alteração de um único nucleotídeo, esses marcadores genéticos são codominantes e bialélicos, possuem alta especificidade por loco e alta abundância genômica, podendo ocorrer tanto em regiões expressas quanto em não expressas, embora sejam

mais prevalentes nas regiões não codificantes do genoma. (SEMAGN *et al.*, 2012, KUMAR, A. *et al.*, 2022). Os marcadores SNPs possuem distribuição variável entre as espécies: no milho, estima-se que um SNP é encontrado a cada 60 a 120 pb do genoma, enquanto em humanos, um a cada 1.000 a 1250 pb (KUMAR, A. *et al.*, 2022).

Devido a sua alta densidade no genoma, somada ao desenvolvimento de tecnologias de genotipagem, esses marcadores abriram novas possibilidades para seu uso, podendo ser utilizados no mapeamento na obtenção de mapas de alta resolução, no melhoramento assistido por marcadores moleculares, além de em estudos de diversidade genética. Ademais, os SNPs são estáveis do ponto de vista evolutivo, possuindo taxa de mutação relativamente baixa e, ainda, possuem grande potencial para análise por meio de técnicas de alto rendimento, menores taxas de erro de genotipagem e são facilmente automatizados, sendo assim importantes marcadores genotípicos, amplamente aceito pelos melhoristas de plantas. (CAIXETA *et al.*, 2006; SEMAGN *et al.*, 2012).

Marcadores SNPs são predominantemente analisados por meio de microarranjos de sondas imobilizadas, comumente denominados “chips” ou “*arrays*”. Chips de DNA fornecem a mais alta qualidade e reprodutibilidade de genotipagem acima de 99% em geral, constituindo hoje o padrão ouro da indústria. Chips de SNPs foram desenvolvidos para o milho desde os primórdios da disponibilização desta tecnologia devido à importância da cultura, com um sistema comercial 50.000 marcadores SNPs (GANAL *et al.* 2011) que logo foi lançado como um produto comercial (Illumina® MaizeSNP50 BeadChip), alcançando sistemas de alta densidade com 600.000 SNPs (UNTERSEER *et al.* 2014) logo em seguida.

Metodologia alternativas e mais econômicas de genotipagem de SNPs surgiram com abordagens baseadas em sequenciamento de nova geração (NGS), chamada genotipagem por sequenciamento (GBS) combinando, simultaneamente, a descoberta de polimorfismo e a genotipagem em um único passo (ELSHIRE *et al.*, 2011; BALOCH *et al.*, 2017). Desse modo, ao invés de inicialmente definir quais são os SNPs que serão genotipados ao longo do genoma, nessa abordagem, primeiro os SNPs são detectados e genotipados, para, posteriormente, serem selecionados para a utilização na análise com base em um conjunto de critérios.

Essa tecnologia além de poder ser aplicada em grupos de amostras, gera uma grande quantidade de marcadores amplamente distribuídos ao longo do genoma (SCHNABLE *et al.* 2009; PARK *et al.*, 2011; ROMAY *et al.* 2013). Entretanto, a qualidade do dado genotípico gerado é substancialmente inferior, com ampla variação na taxa de chamada de dados, com apenas uma fração de SNPs com taxa de chamada acima de 90% e reprodutibilidade altamente

dependente do rigor de análise aplicado aos dados que somente pode ser melhorada resequenciando a mesma amostra com maior cobertura o que elimina a vantagem do baixo custo alegado (BEISSINGER *et al.* 2013). Isto tudo faz com que esta técnica não seja utilizada na indústria, mas sim apenas no ambiente acadêmico.

Romay *et al.* (2013) utilizaram GBS para analisar um total de 4.351 amostras de milho de 2.815 acessos do *National Maize Inbred Seed Bank* dos Estados Unidos com 681.257 marcadores SNPs distribuídos por todo o genoma. Desse conjunto de marcadores, foram selecionados 620.279 SNPs, polimórficos entre as amostras e distribuídos por todo o genoma do milho, com capacidade de detectar alelos raros com altos níveis de confiança. Os dados mostraram uma ampla variação na taxa de chamada de genótipos de apenas 2 a 75% com uma média de 35% que somente foi melhorada para 65% para parte dos acessos após dobrar a cobertura de sequenciamento, aumentando o custo de análise. A análise das distâncias genéticas mostrou estratificação populacional, entretanto, o valor médio do índice de fixação pareada (F_{ST} = 0,06) entre os diferentes grupos de acessos indicou uma diferenciação moderada entre as três principais subpopulações de milho. O espectro de frequência para toda a coleção mostrou que a maioria dos SNPs (68%) possuía frequências do alelo menor (MAF) inferiores a 0,1, com mais da metade de todos os SNPs sendo raros (MAF < 0,05), resultado que sugere que alguns alelos podem ser exclusivos de diferentes subgrupos do germoplasma. O desequilíbrio de ligação (DL), que é uma associação não aleatória entre alelos, em vários locos decaiu muito rapidamente em toda a coleção, mas a extensão do DL se mostrou altamente dependente do grupo específico de germoplasma e da região do genoma. Os autores exploraram, ainda, a possível utilidade dos dados para a realização de estudos de associação genômica ampla, testando duas características mendelianas simples e uma característica complexa.

Kumar, B. *et al.* (2022) utilizaram 60.227 SNPs obtidos via GbS para compreender a diversidade genética, estrutura populacional e desequilíbrio de ligação de 384 linhagens endogâmicas de milho originárias do CIMMYT e de instituições da Índia. A maioria dos valores de parentesco relativo entre pares (58,9%) foram iguais ou próximos de 0, o que sugere um nível mais baixo de parentesco e falta de redundância entre as linhagens. A distância genética entre a maioria dos pares de linhagens (98,3%) variou entre 0,2 e 0,34, enquanto apenas 1,7% dos pares de linhagens diferiram em < 0,20, sugerindo maior variação genética mesmo entre linhagens irmãs que compartilhavam ascendência em comum. A média geral de 17% de heterogeneidade observada indicou a presença de mais endogamia nos genótipos altamente heterogêneos. A diversidade média de nucleotídeos e a frequência de sítios polimórficos por pb

observados foram 0,28 e 0,02, respectivamente. As análises de estrutura populacional baseada em modelos de componentes principais e filogenética, sugeriram a presença de três a seis grupos possíveis, porém sem padrões claros de agrupamento por linhagens centrais, tipos de milho características do grão, maturidade, altura da planta e posicionamento da espiga, entretanto, os genótipos formaram agrupamentos parcialmente com base em seu germoplasma de origem de onde derivaram.

Utilizando 1065 marcadores SNPs genotipados em um beadchip Illumina, Semagn *et al.* (2012) genotiparam 450 linhagens endogâmicas de milho desenvolvidas e/ou amplamente utilizadas pelos programas de melhoramento do CIMMYT no Quênia e no Zimbábue para investigar a estrutura populacional e os padrões de relacionamento do germoplasma. Em mais de 94% das comparações pareadas de todas as 450 linhagens, a distância genética de Roger ficou entre 0,300 e 0,400 e apenas 0,1% caiu dentro de uma distância genética inferior a 0,05, indicando falta de linhagens redundantes entre o germoplasma estudado. Em concordância, os coeficientes de parentesco para 84% dos pares de linhagens foram ≤ 0.500 . As análises de estrutura populacional baseada em modelo, de componentes principais, de cluster e discriminante revelaram a presença de 3 grupos principais que concordam, geralmente, com as informações de pedigree. Os autores recomendaram 644 dos 1.065 para genotipagem de baixa a média densidade em germoplasma de milho tropical usando ensaios uniplex.

Yan *et al.* (2009) utilizaram um total de 1.229 SNPs informativos genotipados em Beadchip Illumina e 1.749 haplótipos organizados a partir de 321 locos para estimar a diversidade genética, estrutura populacional e parentesco familiar de 632 linhagens endogâmicas de uma coleção global de milho altamente diversificada de programas de melhoramento de regiões temperadas, tropicais e subtropicais. A heterozigosidade variou de 0 a 9,9%, com média de 2,5%, resultado do intervalo esperado encontrado em linhagens endogâmicas de milho. O valor médio da proporção total de SNPs que eram polimórficos entre pares (razão de polimorfismo) foi de 26,2%, variando de 0,1 a 47,4%, enquanto a razão polimórfica média de qualquer linhagem para as outras 631 variou de 19,3% a 41,1%. Os haplótipos SNP obtiveram um grande número de alelos, entretanto a maioria é rara na população, com um pouco mais da metade com uma frequência alélica inferior a 0,1. A análise de estrutura populacional identificou subgrupos tropicais e temperados, e relações familiares complexas foram identificadas na coleção global. A distância de decaimento DL diferiu entre cromossomos e variou entre 1 a 10 kb e foi muito maior nas linhagens temperadas do que nas

tropicais e subtropicais, uma vez que as linhas tropicais e subtropicais são mais diversas e contêm mais alelos raros do que as linhas temperadas.

Marcadores SNPs foram utilizados, ainda, para detectar locos associados a características por estudos de associação genômica ampla para cor, volume de expansão de estalo, forma, pericarpo, índice de floração, conteúdo farináceo e proteína (LI *et al.*, 2021) e para avaliação de linhagens endogâmicas de milho tropical sob condições combinadas de calor e seca e também seca terminal, para posterior utilização em estudo de associação genômica ampla para dias até 50% de antese, sedação, cobertura de casca e rendimento de grãos (OSUMAN *et al.*, 2022).

3.8 ALELOTIPAGEM

Muitos problemas referentes a conservação, registro, utilização e distribuição dos recursos genéticos do milho ainda são persistentes: documentações inconsistentes, base de dados que não são facilmente acessíveis, distribuições individuais de germoplasma ineficazes, difícil integração entre os recursos de germoplasma e o melhoramento, caracterizações genéticas fragmentadas, e outros (CGIAR, 2007). Para que os materiais das coleções de recursos genéticos passem a contribuir com essas atividades cientificamente importantes e enriquecedoras, é essencial conhecer a diversidade genética desses repositórios (MCCOUCH *et al.*, 2012; MCCOUCH *et al.*, 2020) uma vez que as informações genotípicas podem servir como dados de passaporte molecular para complementar, corroborar e também corrigir registros de passaportes tradicionais (MASCHER *et al.*, 2019).

Embora existam numerosas coleções de germoplasma de milho em todo o mundo, a maioria não tem orçamentos adequados para a regeneração, intercâmbio ativo, e caracterização de germoplasma (CGIAR, 2007). Esse desafio torna-se ainda maior uma vez que a caracterização de espécies alógamas, como o milho, demanda a análise de uma grande amostragem de indivíduos por acesso devido à alta diversidade existente entre indivíduos dentro de um acesso, tornando o processo potencialmente muito caro (DUBREUIL; CHARCOSSET, 1998; ARCA *et al.*, 2021). Apesar dos avanços na tecnologia molecular, a genotipagem de um grande número de amostras continua a ser muito dispendiosa no contexto da caracterização genética. Uma alternativa conveniente para viabilizar a análise genética, diminuindo os seus custos, é a abordagem de análise genética de amostras agrupadas, denominados comumente de “bulks” ou “pools” de indivíduos. Esta abordagem denominada de “alelotipagem”, difere da genotipagem propriamente dita, pois ao invés de gerar dados genotípicos, são geradas

estimativas de frequências alélicas de uma amostra única agrupada que consolida um conjunto de indivíduos em geral geneticamente distintos, de um acesso ou população (ARCA *et al.*, 2020; ARCA *et al.*, 2021).

Essa abordagem pode ser utilizada como uma estratégia alternativa valiosa para coletar informações sobre frequência alélica de grupo de indivíduos. A alelotipagem foi pioneiramente utilizada por Arnheim, Strange e Erlich (1985) para uma busca por RFLPs informativos dentro de genes do Sistema de Antígenos Leucocitário Humano associados à suscetibilidade a diabetes tipo 1. Vários RFLPs associados à doença foram revelados utilizando essa abordagem econômica e rápida. Para um estudo genético de distúrbios hereditários utilizando marcadores microsatélites, Shaw *et al.* (1998) testaram, em várias amostras de DNA, a acurácia das estimativas de frequência de alelos através do agrupamento de DNA em pools. Adicionalmente avaliaram se os pools podem ser utilizados para estimar frequências de alelos em cromossomos transmitidos e não transmitidos em uma coleção de famílias visando mapeamento de genes por associação alélica. O estudo mostrou que dados quantitativos precisos sobre a frequência alélica, adequados para identificar marcadores para distúrbios complexos, puderam ser identificados através da alelotipagem, reduzindo drasticamente o trabalho e o custo da genotipagem na identificação inicial de locos associados a doenças.

Johnston *et al.* (2013) conduziram a genotipagem e alelotipagem de SNP em amostras históricas de DNA individuais e de DNA agrupado, respectivamente, extraídas de escamas secas do Salmão do Atlântico (*Salmo salar*) armazenadas em temperatura ambiente por até 35 anos. Os autores avaliaram o sucesso da genotipagem, repetibilidade e acurácia da estimativa das frequências alélicas usando um microarranjo de genotipagem de SNP de alta densidade. Em amostras individuais de DNA, o sucesso e a repetibilidade da genotipagem foram muito altos ($> 0,973$ e $> 0,998$, respectivamente). Nas amostras de pools de DNA, a estimativa da frequência do alelo foi altamente repetível (repetibilidade = 0,986) e a frequência do alelo pôde ser estimada com acurácia em mais de 95% das amostras de DNA agrupadas, demonstrando que a alelotipagem fornece uma eficiente alternativa econômica à genotipagem de indivíduos separadamente.

Gouda *et al.* (2020) compararam métodos de amostragem para avaliar a diversidade genética intra e inter-acesso em três espécies de arroz, entre eles, a genotipagem por sequenciamento e o agrupamento de DNA. No geral, as diferentes análises realizadas no estudo revelaram padrões consistentes de polimorfismos de SNPs, heterozigosidade, índices de diversidade intra-acesso e inter-acesso, dissimilaridade genética, estrutura populacional e

diferenciação genética, independentemente dos métodos de amostragem, levando os autores a conclusões consistentemente semelhantes nos 11 conjuntos de dados. Em seres humanos, Schlötterer *et al.* (2014) e Sham *et al.* (2002) levantaram diversos estudos em que a alelotipagem foi utilizada eficazmente, principalmente em estudos de associação “caso-controle”. O agrupamento de DNA foi utilizado para estimar a prevalência de doenças, avaliar exposição a patógenos ou toxinas, rastreamento de infecção por HIV e em uma ampla gama de questões biológicas.

No milho, a alelotipagem de DNA foi utilizada com sucesso para decifrar a diversidade genética de raças locais usando diferentes tipos de marcadores moleculares. O agrupamento de DNA foi utilizado para comparar o potencial de isoenzimas e marcadores RFLPs na investigação da diversidade genética dentro e entre populações naturais de milho (DUBREUIL; CHARCOSSET; 1998), e para estimar frequências alélicas entre linhagens endogâmicas e populações naturais de milho (DUBREUIL *et al.*, 1999). A comparação entre as duas abordagens analíticas, (1) indivíduos e (2) DNA agrupado para a investigação da diversidade molecular dentro e entre populações também foi realizada por Reif *et al.* (2005) utilizando 55 marcadores microssatélites, com o intuito de analisar 5 populações de milho da Europa Central. Nesse estudo, os autores ainda testaram o desequilíbrio de ligação entre pares de locos dentro das populações. Além disso, a alelotipagem foi utilizada em análises históricas dos movimentos pós-domesticação do milho, utilizando marcadores microssatélites (DUBREUIL *et al.*, 2006; MIR *et al.*, 2013).

Recentemente, a utilização de chip de SNPs de milho, combinado com amostras agrupadas em bulks de DNA, tem sido eficientemente utilizada para caracterizar a diversidade genética de raças locais de milho em grande escala genômica (ARCA *et al.*, 2020). Utilizando essa abordagem, Arca *et al.* (2021) usaram o “chip” de análise de SNPs, *Maize 50k Array*, para estimar frequências alélicas de 23.412 SNPs em 156 raças nativas, representando a diversidade de milho americano e europeu. Nesse estudo, a distância de Roger modificada entre as 153 raças locais estimadas a partir de 23.412 SNPs e 17 SSR foi altamente correlacionada, sugerindo que o viés de amostragem é baixo. Essa abordagem fornece estimativas precisas de frequência alélica que são altamente reprodutíveis em laboratórios, além de proteger contra falsa detecção de fixação de alelos em raças locais.

Diaw *et al.* (2021) utilizaram um chip de SNPs de milho, conjuntamente com a abordagem de DNA em bulks, para avaliar a diversidade genética de 194 raças locais de milho cultivadas no Sudoeste da França para elucidar melhor a sua origem. Esse método foi utilizado,

ainda, para detectar assinaturas de seleção nas raças locais de origens geográficas diferentes e comparar a diversidade genética entre raças e linhagens endogâmicas (ARCA *et al.*, 2020) e para separar uma amostra representativa de 130 populações tradicionais de milho europeias em populações do norte e do sul da Europa, quanto à sua variação morfológica e molecular, e definir seis grupos de populações geneticamente próximas, comparáveis às raças europeias (REBOURG; GOUESNARD; CHARCOSSET, 2001).

3.9 ESTUDOS DE ORIGEM E EVOLUÇÃO DO MILHO COM MARCADORES MOLECULARES

Nas últimas décadas, os processos de domesticação e adaptação do milho tem sido objeto de intensa investigação. Estudos genéticos forneceram evidências sólidas de que o milho (*Zea mays* ssp. *mays*) evoluiu do teosinto selvagem de Balsas (*Zea mays* ssp. *parviglumis*) nas terras baixas do México moderno, há cerca de 9.000 anos AP (antes do presente), e adaptado a uma variedade de ambientes em todo o mundo (VAN HEERWAARDEN *et al.*, 2011; LIU; FERNIE; YAN, 2020; XU *et al.*, 2022). Entretanto, é, principalmente, nas terras altas do México, onde a subespécie *parviglumis* não cresce, que as cultivares de milho estritamente relacionadas com o teosinto selvagem de Balsas são encontradas. Portanto, os dados genéticos apontam para uma dispersão do milho domesticado primariamente das terras altas e não a partir da região inicial de domesticação (VAN HEERWAARDEN *et al.*, 2011).

Para resolver esse paradoxo, Van Heerwaarden *et al.* (2011) utilizaram um grande conjunto de dados de SNP de milho e teosinto para reavaliar a ancestralidade genética das populações de milho existentes e estimar os efeitos da introgressão dos parentes selvagens do milho. Como solução, os autores mostraram que a evidência genética anterior de uma possível origem do milho a partir das terras altas do milho moderno é melhor explicada pelo fluxo gênico da subespécie do teosinto anual adaptada às terras baixas mexicanas *Zea mays* ssp. *mexicana* (mexicana). Há evidências de fluxo gênico da mexicana para o milho domesticado. Os autores demonstraram que a mistura com um parente selvagem não ancestral relacionado pode interferir em análises baseadas em comparações diretas com o ancestral conhecido, impactando significativamente na inferência sobre as origens geográficas. Para solucionar o problema da mistura, nesse estudo foram utilizadas frequências ancestrais estimadas sem depender do ancestral selvagem para a inferência. Ao adotarem essa abordagem, a análise da diferenciação das frequências genéticas ancestrais apontou para a posição ancestral do milho das terras baixas

do oeste do México, resultado consistente com os dados arqueobotânicos e a análise genética da domesticação do milho.

Para entender melhor os processos de domesticação e adaptação do milho, Xu *et al.* (2022) conduziram uma pesquisa genômica de 982 linhagens endogâmicas de milho e 190 acessos de teosinto usando mais de 40.000 marcadores SNPs genotipados em chip. A estrutura populacional, a análise de componentes principais e as árvores filogenéticas confirmaram a relação evolutiva entre o milho e o teosinto e determinaram a linhagem evolutiva de todas as espécies dentro do teosinto. Nesse estudo ainda foram realizadas varreduras em busca de assinaturas de seleção, identificando 394 varreduras de domesticação comparando milho selvagem e cultivado e 360 varreduras de adaptação comparando milho tropical e temperado. A identificação de assinaturas de seleção amplia a compreensão dos processos de domesticação e adaptação, beneficiando grandemente o melhoramento genético do milho.

Essas varreduras, ou assinaturas de seleção no genoma ("*selection sweeps*"), têm sido objeto de inúmeros estudos e investigações genômicas a respeito do genoma do milho e seus processos de domesticação e adaptação. Assinaturas de seleção são trechos no genoma onde uma nova mutação benéfica que aumenta sua frequência e se fixa (ou seja, atinge uma frequência de 1) na população, leva à redução ou eliminação da variação genética nas sequências de nucleotídeos próximas à mutação. Na varredura seletiva, a seleção positiva faz com que a nova mutação atinja a fixação tão rapidamente que os alelos vinculados podem "pegar carona" e também se fixar (SMITH; HAIGH, 1974). No caso do milho, esse tipo de investigação levou, a título de exemplo, à identificação e clonagem de genes de maior efeito envolvidos no processo de domesticação como por exemplo o gene TGA1, que controla a formação das cascas nos grãos, e diversos outros genes relacionados com a arquitetura da planta, teor de amido, florescimento e dispersão de sementes (LIU; FERNIE; YAN, 2020), além da identificação de várias outras regiões possivelmente contendo centenas de genes de menor efeito distribuídos por todo o genoma (XU *et al.*, 2022).

Outra questão acerca da domesticação do milho é se o milho é produto de uma ou mais domesticações únicas ou múltiplas do teosinto (DOEBLEY, 2004). Matsuoka *et al.* (2002) estudaram a diversidade de microssatélites em milho e teosinto. Análises filogenéticas baseadas em dados de microssatélites desafiaram a hipótese de múltiplas origens, favorecendo fortemente uma única domesticação, com pouco fluxo gênico subsequente proveniente de *Zea parviglumis*. Entretanto, vários estudos propuseram que as raças locais de milho eram produtos de múltiplas domesticações (KATO, 1984; GALINAT, 1988). Por exemplo, evidências

arqueológicas revelaram que o milho foi apenas parcialmente domesticado no México há cerca de, aproximadamente, 5.300 AP carregando uma mistura de alelos de *Z. parviglumis* e *Z. mays* em locos envolvidos na síndrome de domesticação (RAMOS-MADRIGAL *et al.* 2016). A título de exemplo, Ramos-Madrigal *et al.* (2016) evidenciaram que o variante genético do gene TGA1, responsável por eliminar a casca resistente do teosinto, já estava presente neste período de tempo, enquanto outros locos associados a mudanças na dispersão de sementes e produção de amido, ocorridas durante a domesticação, ainda continham variantes do tipo selvagem.

O estado de domesticação parcial separa esses arqueogenomas do genoma do milho moderno que é totalmente domesticado com um conjunto completo e estável de alelos de domesticação que conferem o fenótipo domesticado. Este milho parcialmente domesticado foi cultivado no México bem depois que o milho se estabeleceu na América do Sul, o que levanta a questão de como que o milho sul-americano passou a possuir o complemento total de traços fixos de domesticação. Estudos recentes envolvendo dados arqueobotânicos e genômicos de teosinto, cultivares indígenas da América do Sul (incluindo o Brasil) e cultivares modernas, indicaram que o milho sul-americano foi levado do centro de domesticação mesoamericano logo após os estágios iniciais de domesticação. O sudoeste da Amazônia teria assim sido um centro secundário de domesticação, propulsionado por múltiplas ondas de dispersão mediada por humanos resultando na biogeografia e diversidade do milho sul-americano moderno (KRISTLER *et al.*, 2018).

Com o intuito de entender a migração do milho de origem Mesoamericana em direção à América do Sul e ao Caribe, Bedoya *et al.* (2007) realizaram a caracterização genéticas de populações de milho da América Latina do banco de germoplasma de milho do CIMMYT utilizando marcadores microsatélites, a fim de estudar a diversidade e a estrutura populacional. Nesse estudo foram identificados três grupos principais de germoplasma de milho: (1) grupo do México e dos Andes do Sul, destaca o intercâmbio pré-colombiano e moderno de germoplasma entre as Américas do Norte e do sul; (2) grupo das terras baixas da Mesoamérica, apoia a hipótese de que dois eventos distintos de migração humana poderiam ter contribuído para o germoplasma do milho caribenho; (3) grupo andino, apresentou introdução precoce do milho nos Andes, com pouca mistura desde então, além de uma zona de intercâmbio regional ativa no passado. As relações de associação encontradas entre e dentro dos grupos dos acessos avaliados evidenciou que a estrutura dos grupos genéticos formados é influenciada por zonas micro e macroambientais distintas, migrações humanas e comércio, assim como pela seleção e conservação de raças locais pelas etnias indígenas.

Utilizando marcadores microssatélites, Vigouroux *et al.* (2008) analisaram a estrutura genética populacional por meio da genotipagem de 964 plantas individuais, representantes de quase todo o conjunto de aproximadamente 350 raças nativas das Américas. Foram detectados quatro grupos principais, representantes de raças (1) das terras altas do México; (2) do norte dos Estados Unidos; (3) das planícies tropicais e (4) dos Andes. A análise filogenética apontou para o sudoeste dos Estados Unidos como trampolim intermediário entre o México e o norte dos EUA. Além disso, o estudo apontou que as raças do sudoeste dos EUA possivelmente têm ascendência mista de flint do norte e de planícies tropicais, enquanto as raças das terras baixas das planícies do centro da América do Sul têm ascendência mista dos Andes e de planícies tropicais. Foram detectados numerosos casos de movimentos de raças pós-colombianos, principalmente dos Estados Unidos para a América do Sul. Dentro dos principais grupos detectados, a maior diversidade genética foi encontrada nas raças das terras altas do México, enquanto a diversidade foi mais baixa nos Andes e no Norte dos EUA, sugerindo que o isolamento pela distância pode ser o principal fator subjacente à diversificação histórica do milho.

O cenário mais comumente aceito para a dispersão do milho na Europa é de uma única introdução seguida de uma lenta adaptação a climas temperados (DUBREUIL *et al.*, 2006). Com o intuito de esclarecer as origens e as modalidades de introdução do milho na Europa, Dubreuil *et al.* (2006) realizaram um extenso levantamento de 275 populações de milho originadas dos continentes americanos e europeus utilizando análise baseada em microssatélites em grupos de indivíduos. Duas fontes principais, uma do Caribe e outra do nordeste da América, que deram origem à maioria das variedades de polinização aberta cultivadas na Europa, foram fortemente sustentadas pelos dados resultantes desse estudo. A análise histórica detalhada foi consistente com dados históricos sobre a introdução do milho do sul da Europa por Colombo, sugerindo que as introduções do milho temperado do nordeste americano devem ter ocorrido no início do século XVI, na época das expedições espanholas ou francesas. Além disso, os autores sugeriram que as variedades de milho cultivadas em latitudes médias na Europa tenham, provavelmente, resultado da hibridização entre as variedades do sul e do nordeste europeu.

Costa *et al.* (2022) caracterizaram 170 acessos de milho provenientes do banco de germoplasma de EMBRAPA Milho e Sorgo e das coleções recentes do Uruguai e dois híbridos comerciais utilizando marcadores SNPs obtidos por genotipagem por sequenciamento, visando avaliar se a dispersão do milho estava associada a tipos de endosperma e etnias indígenas.

Foram identificados quatro grupos genéticos baseados em análise de componentes principais e cinco grupos baseados em análise de cluster, os grupos estruturados de acordo com a predominância dos tipos de endosperma (pipoca, farináceo, flint/semi-flint). De acordo com a análise de componentes principais da variação genética, diferentes padrões de dispersão para cada tipo de endosperma foram sugeridos. Os padrões inter-regionais sugerem dispersão em consonância com a expansão dos grupos indígenas Tupi, Macro-Jê e Arawak. A partir de uma possível origem no sudoeste da Amazônia, surgiram diferentes rotas de dispersão do milho: (1) em direção ao Norte da Amazônia, que continuou em direção à Caatinga e sudeste da Mata Atlântica (farináceo); (2) em direção ao Sul do Brasil, passando pelo Cerrado e Mata Atlântica, chegando à região do Pampa (farináceo); e (3) ao longo da Costa Atlântica, seguindo os movimentos migratórios Tupi e originados de duas expansões distintas: uma (Tupinamba) de norte a sul, e outra (Guarani) na direção oposta, de sul a norte (flint, farináceo e pipoca). Até o momento presente, o único consenso em relação à domesticação do milho é, na verdade, o reconhecimento da sua evolução a partir de alguma forma de teosinto e a intervenção humana como condição indispensável para a sua transformação no milho atual como resultado de uma seleção e evolução artificial progressiva (SILVA, N. *et al.*, 2020).

4. REFERÊNCIAS BIBLIOGRÁFICAS

- ABADIE, T. *et al.* **A coleção nuclear de germoplasma de milho para o Brasil**. Brasília: Embrapa Recursos Genéticos e Biotecnologia, 2000. 37p. (Embrapa Recursos Genéticos e Biotecnologia. Boletim de Pesquisa, 8). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/179927>>. Acesso em: 6 jul. 2024.
- ADU, G. B. *et al.* Genetic characterization and population structure of maize populations using SSR markers. **Annals of Agricultural Sciences**, v. 64, n. 1, p. 47-54, 2019.
- ALDRICH, S.R.; SCOTT, W.O.; LENG, E.R. **Modern corn production**. 2.ed. Champaign: A & L Publication, 1982. 371 p.
- ALI, B. A. *et al.* A review of random amplified Polymorphic DNA (RAPD) markers in fish research. **Fish Biology and Fisheries**, v. 14, n. 1217, p. 443-453, 2004.
- ANDJELKOVIC, V. *et al.* Conserving maize in gene banks: Changes in genetic diversity revealed by morphological and SSR markers. **Chilean journal of agricultural research**, v. 78, n. 1, p. 30-38, 2018.
- ARCA, M. *et al.* Deciphering the genetic diversity of landraces with high-throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array. **Frontiers in Plant Science**, v. 11, p. 568699, 2021.
- ARCA, M. *et al.* Genome-wide SNP genotyping of DNA pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool. **BioRxiv**, 2020.
- ARMSTRONG, C. L.; ROMERO-SEVERSON, J.; HODGES, T. K. Improved tissue culture response of an elite maize inbred through backcross breeding, and identification of chromosomal regions important for regeneration by RFLP analysis. **Theoretical and Applied Genetics**, v. 84, p. 755-762, 1992.
- ARNHEIM, N.; STRANGE, C.; ERLICH, H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II locos. **Proceedings of the National Academy of Sciences**, v. 82, n. 20, p. 6970-6974, 1985.
- BALOCH, F. S. *et al.* A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. **PLoS ONE**, v. 12, n. 1, p. 1–18, 2017.
- BEDOYA, C. A. *et al.* Genetic diversity and population structure of native maize populations in Latin America and the Caribbean. **Plos one**, v. 12, n. 4, p. e0173488, 2017.
- BENNETZEN, J. L. The contributions of retroelements to plant genome organization, function and evolution. **Trends in Microbiology**, v. 4, n. 9, p. 347-353, 1996.
- BENTOLILA, S. *et al.* Identification of an RFLP marker tightly linked to the Ht1 gene in maize. **Theoretical and applied genetics**, v. 82, p. 393-398, 1991.

BRIEGER, F.G. *et al.* **Races of maize in Brazil and other eastern South American countries** 593. NAS-NRC, Washington, DC, 1958. 282 p.

BRUEL, D. C. *et al.* Assessment of genetic diversity in maize inbred lines using RAPD markers. **Crop Breeding and Applied Biotechnology**, v. 7, n. 2, 2007.

CAIXETA, E. T. *et al.* Tipos de marcadores moleculares. *In*: BORÉM, A.; CAIXETA, E. T. (Ed.). **Marcadores moleculares**. Viçosa, MG: Editora UFV, 2006. cap. 1, p. 9–78.

CARVALHO, V. P. *et al.* Genetic diversity among maize (*Zea mays* L.) landraces assessed by RAPD markers. **Genetics and Molecular Biology**, v. 27, p. 228-236, 2004.

CGIAR. CONSULTATIVE GROUP ON INTERNATIONAL AGRICULTURAL RESEARCH. **Global Strategy for the Ex-situ Conservation and Utilization of Maize Germplasm**. Washington : CGIAR, 2007. Disponível em: <<https://www.genebanks.org/resources/publications/>>. Acesso em: 08 jul. 2024.

CGIAR. CONSULTATIVE GROUP ON INTERNATIONAL AGRICULTURAL RESEARCH. **Maize (*Zea mays*) is one of the world's three most import cereals along with wheat and rice**. Washington : CGIAR, 2024. Disponível em: <<https://www.genebanks.org/resources/crops/maize/>>. Acesso em: 08 jul. 2024

CHANDLER, V. L.; BRENDDEL, V. The maize genome sequencing project. **Plant Physiology**, v. 130, n. 4, p. 1594-1597, 2002.

CHEN, J. *et al.* A complete telomere-to-telomere assembly of the maize genome. **Nature genetics**, v. 55, n. 7, p. 1221-1231, 2023.

CHIORATO, A. F. *et al.* Genetic diversity in common bean accessions evaluated by means of morpho-agronomical and RAPD data. **Scientia Agricola**, v. 64, n. 3, p. 256-262, 2007.

CIMMYT. INTERNATIONAL MAIZE AND WHEAT IMPROVEMENT CENTER. **Our Research Programs**. México: CIMMYT, 2024. Disponível em: <https://www.cimmyt.org/?gad_source=1&gclid=CjwKCAjwnei0BhB-EiwAA2xuBtgF2ltwKg1N0wLemGE-1AhIPuXPibV3QpUhDZA0upraVhSrkdVF5BoCzgsQAvD_BwE>. Acesso em: 08 jul. 2024.

COE, E. *et al.* Access to the maize genome: an integrated physical and genetic map. **Plant Physiology**, v. 128, n. 1, p. 9-12, 2002.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **12º Levantamento - Safra 2022/23**. Brasília: Companhia Nacional de Abastecimento. Disponível em: <<https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/21369-12-levantamento-safra-2022-23>>. Acesso em: 10 jul. 2024.

CONTINI, E. *et al.* **Milho: caracterização e desafios tecnológicos**. Brasília, DF: Embrapa; Sete Lagoas: Embrapa Milho e Sorgo, 2019. 45 p. (Desafios do Agronegócio Brasileiro, 2). Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/195075/1/Milho-caracterizacao.pdf>>. Acesso em: 20 jul. 2024.

COSTA, F. M. *et al.* Maize dispersal patterns associated with different types of endosperm and migration of indigenous groups in lowland South America. **Annals of Botany**, v. 129, n. 6, p. 737-751, 2022.

DEMÉTRIO, C. S. *et al.* Desempenho de híbridos de milho submetidos a diferentes espaçamentos e densidades populacionais. **Pesquisa Agropecuária Brasileira**, v. 43, n. 1, p. 1691-1697, 2008.

DESAI, A. R.; SINGH, N. P. Biodiversity, varietal wealth and molecular applications in cashew. In: SAROJ, P. L. (Ed.). **Advances in Cashew Production Technology**. Puttur: Directorate of Cashew Research, 2015. p 10-25

DIW, Y. *et al.* Genetic diversity of maize landraces from the South-West of France. **PLoS One**, v. 16, n. 2, p. e0238334, 2021.

DOEBLEY, J. The genetics of maize evolution. **Annu. Rev. Genet.**, v. 38, n. 1, p. 37-59, 2004.

DUBREUIL, P. *et al.* Evaluation of a DNA pooled-sampling strategy for estimating the RFLP diversity of maize populations. **Plant Molecular Biology Reporter**, v. 17, n. 2, p. 123-138, 1999.

DUBREUIL, P. *et al.* More on the introduction of temperate maize into Europe: large-scale bulk SSR genotyping and new historical elements. 2006.

DUBREUIL, P.; CHARCOSSET, A. Genetic diversity within and among maize populations: a comparison between isozyme and nuclear RFLP locos. **Theoretical and Applied Genetics**, v. 96, p. 577-587, 1998.

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.

FLAVELL, R. B. *et al.* Genome size and the proportion of repeated nucleotide sequence DNA in plants. **Biochemical genetics**, v. 12, p. 257-269, 1974.

FILHO, I. A. P.; CRUZ, J. C.; GARCIA, J. C. Sistemas Diferenciais de Cultivo. **Ageitec - Agência Embrapa de Informação Tecnológica**, Brasília, 08 de dez. de 2021. Disponível em: <<https://www.embrapa.br/en/agencia-de-informacao-tecnologica/cultivos/milho/producao/sistemas-diferenciais-de-cultivo#:~:text=O%20plantaio%0de%20ver%C3%A3o%2C%20ou,ocorre%20no%20in%C3%ADcio%20do%20ano>>. Acesso em: 11 jul. 2024.

FREITAS, I. L. J. **Fitotecnia: Soja e Milho**. Londrina: Educacional, 2019. 192 p. Disponível em: <https://cm-cls-content.s3.amazonaws.com/201901/INTERATIVAS_2_0/FITOTECNIA_SOJA_E_MILHO/U1/LIVRO_UNICO.pdf>. Acesso em: 6 jul. 2024.

FALEIRO, F. G. **Marcadores genético-moleculares aplicados a programas de conservação e uso de recursos genéticos**. Planaltina, DF: Embrapa Cerrados, 2007, 102p.

FALEIRO, F. G. Princípio científico e análises genéticas utilizando marcadores moleculares. In: FALEIRO, F. G; ANDRADE; S. R. M. de; REIS JUNIOR, F. B. dos (Ed.). **Biotechnologia estado arte e aplicações na agropecuária**. Planaltina, DF: Embrapa Cerrados, 2011. p. 31-52.

FERREIRA, M. E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3ª edição. Brasília, DF: Embrapa CENARGEM, 1998. 220 p

GALINAT, W. C. Corn and corn improvement. **Am. Soc. Agron., Crop Sci. Soc. Am., Soil Sci. Soc. Am., Madison, WI**, p. 1-31, 1988.

GANAL, M. W. *et al.* A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. **PloS one**, v. 6, n. 12, p. e28334, 2011.

GAUTHIER, P. *et al.* RFLP diversity and relationships among traditional European maize populations. **Theoretical and Applied Genetics**, v. 105, p. 91-99, 2002.

GOUDA, A. C. *et al.* Comparisons of sampling methods for assessing intra-and inter-accession genetic diversity in three rice species using genotyping by sequencing. **Scientific Reports**, v. 10, n. 1, p. 13995, 2020.

GOVINDARAJ, M.; VETRIVENTHAN, M.; SRINIVASAN, M. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. **Genetics research international**, v. 2015, n. 1, p. 431487, 2015.

GUIMARÃES, C. T. **Genoma milho: o estado da arte**. Sete Lagoas: Embrapa Milho e Sorgo, 2002. 7 p. Disponível em: < <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/34901/1/Pal-estra-Genoma-milho.pdf>>. Acesso em: 02 jul. 2024.

HALEVY, A. H. *et al.* (Ed.). **CRC handbook of flowering**. Boca Raton^ eFlorida Florida: CRC press, 1985.

HAMBLIN, M. T.; WARBURTON, M. L.; BUCKLER, E. S. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. **PloS one**, v. 2, n. 12, p. e1367, 2007.

HÖLKER, A. C. *et al.* Theoretical and experimental assessment of genome-based prediction in landraces of allogamous crops. **Proceedings of the National Academy of Sciences**, v. 119, n. 18, p. e2121797119, 2022.

IDREES, M.; IRSHAD, M. Molecular Markers in Plants for Analysis of Genetic Diversity: A Review. **European Academic Research**, v. 2, n. 1, p. 1513–1540, 2014.

JOHNSTON, S. E. *et al.* Fish scales and SNP chips: SNP genotyping and allele frequency estimation in individual and pooled DNA from historical samples of Atlantic salmon (*Salmo salar*). **Bmc Genomics**, v. 14, p. 1-13, 2013.

KATO, Y. T. A. Chromosome morphology and the origin of maize and its races. **Evolutionary biology**, v. 17, 1984.

KHOURY, C. K. *et al.* Crop genetic erosion: understanding and responding to loss of crop diversity. **New Phytologist**, v. 233, n. 1, p. 84-118, 2022.

- KISTLER, L. *et al.* Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science*, v. 362, n. 6420, p. 1309-1313, 2018.
- KORDROSTAMI, M.; RAHIMI, M. Molecular markers in plants: Concepts and applications. *Genetics in the Third Millennium*, v. 13, n. 2, p. 4024–4031, 2015.
- KUMAR, A. *et al.* Molecular marker analysis of genetic diversity in maize: A review. *OBM Genetics*, v. 6, n. 1, p. 1-19, 2022.
- KUMAR, B. *et al.* Genetic diversity, population structure and linkage disequilibrium analyses in tropical maize using genotyping by sequencing. *Plants*, v. 11, n. 6, p. 799, 2022.
- KUMAR, P. *et al.* Potential of molecular markers in plant biotechnology. *Plant Omics Journal*, v. 2, n. 4, p. 141-162, 2009.
- LABATE, J. A. *et al.* Molecular and historical aspects of corn belt dent diversity. *Crop Science*, v. 43, n. 1, p. 80-91, 2003.
- LABORDA, P. R. *et al.* Tropical maize germplasm: what can we say about its genetic diversity in the light of molecular markers?. *Theoretical and Applied Genetics*, v. 111, p. 1288-1299, 2005.
- LANGRIDGE, P.; WAUGH, R. Harnessing the potential of germplasm collections. *Nature Genetics*, v. 51, n. 2, p. 200-201, 2019.
- LE HELLARD, S. *et al.* SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic acids research*, v. 30, n. 15, p. e74-e74, 2002.
- LI, J. *et al.* Genome-wide analyses reveal footprints of divergent selection and popping-related traits in CIMMYT's maize inbred lines. *Journal of experimental botany*, v. 72, n. 4, p. 1307-1320, 2021.
- LIAN, Q. *et al.* A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics*, p. 1-10, 2024.
- LIU, J.; FERNIE, A. R.; YAN, J. The past, present, and future of maize improvement: domestication, genomics, and functional genomic routes toward crop enhancement. *Plant Communications*, v. 1, n. 1, 2020.
- MCCOUCH, S. R. *et al.* Genomics of gene banks: a case study in rice. *American journal of botany*, v. 99, n. 2, p. 407-423, 2012.
- MCCOUCH, S. R. *et al.* Mobilizing crop biodiversity. *Molecular plant*, v. 13, n. 10, p. 1341-1344, 2020.
- MAGALHAES, P. C. *et al.* **Fisiologia do milho**. Sete Lagoas: Embrapa Milho e Sorgo, 2002. 23 p. (Embrapa Milho e Sorgo. Circular Técnica, 22). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/486995>>. Acesso em: 6 jul. 2024.

MAGALHÃES, P. C.; DURÃES, F. O. M.; GOMIDE, R. L. Fisiologia da cultura do milho. 1996. In: BENASSI, A. C. *et al.* **Manual técnico para a cultura do milho no Estado do Espírito Santo**. Vitória: EMCAPA, 1996. p. 15-34. Disponível em: < <https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/478294/fisiologia-da-cultura-do-milho>>. Acesso em: 6 jul. 2024.

MAGALHÃES, P. C.; DURÃES, F. O. M.; PAIVA, E. **Fisiologia da planta de milho**. Sete Lagoas: EMBRAPA-CNPMS, 1995. 27 p. (EMBRAPA-CNPMS. Circular técnica, 20). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/475778>>. Acesso em: 6 jul. 2024.

MASCHER, M. *et al.* Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nature genetics*, v. 51, n. 7, p. 1076-1081, 2019.

MATSUOKA, Y. *et al.* A single domestication for maize shown by multilocus microsatellite genotyping. **Proceedings of the National Academy of Sciences**, v. 99, n. 9, p. 6080-6084, 2002.

MESSING, J. *et al.* Sequence composition and genome organization of maize. **Proceedings of the National Academy of Sciences**, v. 101, n. 40, p. 14349-14354, 2004.

MIR, C. *et al.* Out of America: tracing the genetic footprints of the global diffusion of maize. **Theoretical and Applied Genetics**, v. 126, p. 2671-2682, 2013.

MUMM, R. H.; HUBERT, L. J.; DUDLEY, J. W. A classification of 148 US maize inbreds: II. Validation of cluster analysis based on RFLPs. **Crop science**, v. 34, n. 4, p. 852-865, 1994.

NETTO, D. A. M. *et al.* Avaliação agronômica e molecular de acessos da coleção núcleo de milho, subgrupo endosperma duro. **Revista Brasileira de Milho e Sorgo**, v. 3, n. 01, 2004.

NETTO, D. A. M.; SOUZA, I. R. P.; OLIVEIRA, A. C. Validação da Coleção Núcleo de Milho Brasileira Subgrupo Endosperma Dentado. **Revista Brasileira de Milho e Sorgo**, v. 8, n. 03, 2009.

OKUMUS, A. Genetic variation and relationship between Turkish flint maize landraces by RAPD markers. **Am J Agri Biol Sci**, v. 2, n. 2, p. 49-53, 2007.

OLIVEIRA, K. M. *et al.* Evaluating genetic relationships between tropical maize inbred lines by means of AFLP profiling. **Hereditas**, v. 140, n. 1, 2004.

ORTIZ, R. *et al.* Conserving and enhancing maize genetic resources as global public goods—a perspective from CIMMYT. **Crop Science**, v. 50, n. 1, p. 13-28, 2010.

OSUMAN, A. S. *et al.* Genome-wide association study reveals genetic architecture and candidate genes for yield and related traits under terminal drought, combined heat and drought in tropical maize germplasm. **Genes**, v. 13, n. 2, p. 349, 2022.

OZEROV, M. *et al.* Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. **PLoS One**, v. 8, n. 12, p. e82434, 2013.

PARK, J-H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. **Proceedings of the National Academy of Sciences**, v. 108, n. 44, p. 18026-18031, 2011.

PATERNIANI, E.; GOODMAN, M. **Races of maize in Brazil and adjacent areas**. CIMMYT, 1977. 95 p.

PATERNIANI, E.; NASS, L. L.; DOS SANTOS, M. X. O valor dos recursos genéticos de milho para o Brasil. *In*: UDRY, C. V.; DUARTE, W. (Org.) **Uma história brasileira do milho - o valor dos recursos genéticos**. Brasília: Paralelo 15, 2000. p. 11-42.

PEREIRA, M. G.; PEREIRA, T. N. S. Marcadores moleculares no pré-melhoramento de plantas. *In*: BORÉM, A.; CAIXETA, E. T. (Ed.). **Marcadores moleculares**. Viçosa, MG: Editora UFV, 2006. cap. 3, p. 85–106.

PETERSON, G. W. *et al.* Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. **Diversity**, v. 6, n. 4, p. 665–680, 2014.

PINTO, M.; TEIXEIRA, F. F.; GUIMARÃES, C. T. Diversidade Genética de Milho-Doce Utilizando Marcadores Microssatélites. *In*: CONGRESSO NACIONAL DE MILHO E SORGO, 31., 2016, Bento Gonçalves. Milho e sorgo: inovações, mercados e segurança alimentar: **Anais**. Sete Lagoas: Associação Brasileira de Milho e Sorgo, 2016.

PRIYANKA, V. *et al.* Germplasm conservation: instrumental in agricultural biodiversity - a review. **Sustainability**, v. 13, n. 12, p. 6743, 2021.

RAMOS-MADRIGAL, J. *et al.* Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. **Current Biology**, v. 26, n. 23, p. 3195-3201, 2016.

REBOURG, C.; GOUESNARD, B.; CHARCOSSET, A. Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation. **Heredity**, v. 86, n. 5, p. 574-587, 2001.

ROMAY, M. C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. **Genome biology**, v. 14, p. 1-18, 2013.

RUDD, S.; SCHOOF, H.; MAYER, K. PlantMarkers - A database of predicted molecular markers from plants. **Nucleic Acids Research**, v. 33, n. DATABASE ISS., p. 628–632, 2005.

SANMIGUEL, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. **Science**, v. 274, n. 5288, p. 765-768, 1996.

SCHNABLE, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. **science**, v. 326, n. 5956, p. 1112-1115, 2009.

SCHLÖTTERER, C. *et al.* Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. **Nature Reviews Genetics**, v. 15, n. 11, p. 749-763, 2014.

SEMAGN, K. *et al.* Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. **BMC genomics**, v. 13, p. 1-11, 2012.

SEMAGN, K.; BJORNSTAD, A.; NDJIONDJOP, M. N. An overview of molecular markers methods for plants. **African Journal of Biotechnology**, v. 5, n. 25, p. 2540-2568, 2006.

SHAM, P. *et al.* DNA pooling: a tool for large-scale association studies. **Nature Reviews Genetics**, v. 3, n. 11, p. 862-871, 2002.

SHAROPOVA, N. *et al.* Development and mapping of SSR markers for maize. **Plant molecular biology**, v. 48, p. 463-481, 2002.

SHARMA, S. *et al.* Study of interspecific SSR polymorphism among 14 species from Triticum-Aegilops group. **Wheat Information Service**, v. 95, p. 23-28, 2002.

SHAW, S. H. *et al.* Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. **Genome research**, v. 8, n. 2, p. 111-123, 1998.

SILVA, K. J. *et al.* High-density SNP-based genetic diversity and heterotic patterns of tropical maize breeding lines. **Crop Science**, v. 60, n. 2, p. 779-787, 2020.

SILVA, N. C. D. A. *et al.* Relationships among American popcorn and their links with landraces conserved in a microcenter of diversity. **Genetic Resources and Crop Evolution**, v. 67, p. 1733-1753, 2020.

SMITH, J. M.; HAIGH, J. The hitch-hiking effect of a favourable gene. **Genetics Research**, v. 23, n. 1, p. 23-35, 1974.

TANKSLEY, S. D.; MCCOUCH, S. R. Seed banks and molecular maps: unlocking genetic potential from the wild. **science**, v. 277, n. 5329, p. 1063-1066, 1997.

TEIXEIRA, F. F. **Milho cultivado no Brasil e banco de germoplasma - uma forma de classificação da variabilidade genética**. Sete Lagoas: Embrapa Milho e Sorgo, 2008. 11 p. (Embrapa Milho e Sorgo. Comunicado Técnico, 155). Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/CNPMS-2009-09/21382/1/Com_155.pdf>. Acesso em: 20 jul. 2024.

TEIXEIRA, F. F.; COSTA, F. M. **Caracterização de recursos genéticos de milho**. Sete Lagoas: Embrapa Milho e Sorgo, 2010. 10 p. (Embrapa Milho e Sorgo. Comunicado Técnico, 185). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/883797>> Acesso em: 27 jun. 2024.

TEIXEIRA, F. F. *et al.* **Diversidade genética entre acessos do Banco Ativo de Germoplasma de Milho com grãos brancos do tipo dentado**. Sete Lagoas: Embrapa Milho e Sorgo, 2019. 67 p. (Embrapa Milho e Sorgo. Documentos, 233). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1111013>>. Acesso em: 27 jun. 2024.

TEIXEIRA, F. F. *et al.* **Diversidade genética entre acessos do Banco de Germoplasma de Milho de origem indígena**. Sete Lagoas: Embrapa Milho e Sorgo, 2021. 133 p. (Embrapa Milho e Sorgo. Documentos, 259). Disponível em: < <http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1131481>>. Acesso em: 27 jun. 2024.

TEUMER, A. *et al.* Comparison of genotyping using pooled DNA samples (allelotyping) and individual genotyping using the affymetrix genome-wide human SNP array 6.0. **BMC genomics**, v. 14, p. 1-12, 2013.

UNTERSEER, S. *et al.* A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. **BMC genomics**, v. 15, p. 1-15, 2014.

USDA. UNITED STATES DEPARTMENT OF AGRICULTURE. **Production - Corn**. Washington: United States Department of Agriculture. 2023a. Disponível em: <<https://fas.usda.gov/data/production/commodity/0440000>>. Acesso em: 09 jul. 2024.

USDA. UNITED STATES DEPARTMENT OF AGRICULTURE. **Grain and Feed Update**. 2023b. Washington: United States Department of Agriculture. Disponível em: <https://apps.fas.usda.gov/newgainapi/api/Report/DownloadReportByFileName?fileName=Grain%20and%20Feed%20Update_Brasilia_Brazil_BR2023-0028>. Acesso em: 9 jul. 2024.

VAN HEERWAARDEN, J. *et al.* Genetic signals of origin, spread, and introgression in a large sample of maize landraces. **Proceedings of the National Academy of Sciences**, v. 108, n. 3, p. 1088-1092, 2011.

VAZ PATTO, M. C. *et al.* Assessing the genetic diversity of Portuguese maize germplasm using microsatellite markers. **Euphytica**, v. 137, p. 63-72, 2004.

VIEIRA, L. C. **Caracterização de germoplasma de milho crioulo e suas implicações no melhoramento genético**. 2011. 91 p. Tese (Doutorado em Agronomia) - Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2011.

VIGOUROUX, Y. *et al.* Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. **American journal of botany**, v. 95, n. 10, p. 1240-1253, 2008.

WALTERS, C. Principles for preserving germplasm in gene banks. *In*: GUERRANT, E.; HAVENS, K.; MAUNDER, M. (Ed.). **Ex situ plant conservation: supporting species survival in the wild**. Washington: Island Press, 2004 p. 442-453.

WARBURTON, M. L. *et al.* Genetic characterization of CIMMYT inbred maize lines and open pollinated populations using large scale fingerprinting methods. **Crop Science**, v. 42, n. 6, p. 1832-1840, 2002.

XU, G. *et al.* Population genomics of *Zea* species identifies selection signatures during maize domestication and adaptation. **BMC plant biology**, v. 22, n. 1, p. 72, 2022.

YAN, J. *et al.* Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. **PloS one**, v. 4, n. 12, p. e8451, 2009.

ZHANG, F.; WAN, X.; PAN, G.. QTL mapping of *Fusarium moniliforme* ear rot resistance in maize. 1. Map construction with microsatellite and AFLP markers. **Journal of Applied Genetics**, v. 47, p. 9-15, 2006.

ZHOU, Y. *et al.* A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci Data* 7: 113. 2020.

ZSÖGÖN, A. *et al.* De novo domestication of wild tomato using genome editing. **Nature biotechnology**, v. 36, n. 12, p. 1211-1216, 2018.

CAPÍTULO 1 – COMPARAÇÃO DO MÉTODOS DE PREPARAÇÃO DE “BULKS” DE DNA PARA A ESTIMATIVA DE FREQUÊNCIAS ALÉLICAS POR MEIO DA ALELOTIPAGEM DE SNPs EM *ZEAMAYS*

RESUMO

Mesmo com os recentes avanços na tecnologia molecular, a genotipagem de um grande número de indivíduos continua a ser muito dispendiosa no contexto da caracterização genética para a grande maioria coleções de germoplasma. Esse desafio torna-se ainda maior para espécies alógamas, ou seja, de fecundação cruzada como o milho, devido à alta diversidade genética existente dentro de cada acesso. Para essas espécies, um acesso corresponde a uma população com ampla diversidade entre os indivíduos. A caracterização de acessos de milho que não são linhagens, como no caso de variedades tradicionais e indígenas, demanda, portanto, a análise de vários indivíduos por acesso para capturar adequadamente a diversidade existente dentro do acesso, tornando o processo potencialmente muito caro. O agrupamento de indivíduos em “bulks” ou “pools” de amostras populacionais, representa uma alternativa ao permitir a análise de um menor número de amostras físicas, minimizando significativamente o volume de trabalho de genotipagem, reduzindo os custos em até uma ordem de magnitude. Nesse trabalho, três métodos de preparação de bulks de DNA foram avaliados quanto à acurácia das estimativas de frequências alélicas obtidas pela alelotipagem em comparação com as frequências alélicas de referência calculadas a partir da genotipagem de amostras individuais. Os métodos foram: (1) Bulk equimolar de amostras de DNA extraídas de 16 plantas individuais e quantificadas por espectrometria em um equipamento NanoDrop; (2) Idem a (1) porém o DNA quantificado somente para fita dupla utilizando PicoGreen em fluorímetro (3) Extração de DNA em bulk de amostras de tecido foliar pesadas individualmente com precisão de 0,1 grama. As frequências alélicas a 3.526 SNPs estimadas a partir da alelotipagem dos bulks de DNA foram comparadas com as frequências calculadas a partir da genotipagem individual das mesmas 16 plantas. Este procedimento foi realizado para quatro acessos distintos de milho amostrados do BAG Milho. A avaliação da acurácia da estimativa das frequências alélicas foi realizada pelo cálculo do RMSE (Raiz Do Erro Quadrático Médio) que quantifica a diferença entre as estimativas de frequências alélicas estimadas a partir dos bulks e os valores previstos a partir da genotipagem de plantas individuais. O método que forneceu o menor RMSE (Raiz Do Erro Quadrático Médio) foi o método de extração foliar em bulk, com um RMSE médio de 0,063 para os 3.526 SNPs. Além disso, as frequências alélicas estimadas a partir das amostras de bulks de folhas apresentaram elevada correlação com aquelas estimadas a partir de dados genotípicos de plantas individuais, com um coeficiente de correlação médio de $r = 0,984$. A reprodutibilidade entre réplicas biológicas, avaliadas com base em bulks de folhas de 24 indivíduos para 64 acessos distintos de milho utilizando o índice de fixação de Wright indicou estimativas precisas ($F_{ST} = 0,0038$) indicando que o erro padrão devido à amostragem é desprezível. Esse estudo fornece um método de montagem de bulk rápido, econômico e de fácil implementação e mostra que as estimativas das frequências alélicas obtidas pela alelotipagem podem ser altamente acuradas, abrindo assim perspectivas positivas de utilização desta abordagem para a análise genética de amplas coleções de germoplasma de plantas alógamas.

Palavras-chave: *Zea mays* L.; Germoplasma; Alelotipagem; Bulk de DNA; Diversidade Genética; SNPs.

ABSTRACT

Despite the recent advances in molecular technologies, genotyping large numbers of individuals remains very expensive in the context of genetic characterization for the vast majority of germplasm collections. This challenge becomes even greater for allogamous species, i.e., cross-fertilized species such as maize, due to the high genetic diversity within each accession. For these species, an accession corresponds to a population with broad diversity among individuals. The characterization of maize accessions that are not lineages, as in the case of traditional and indigenous varieties, therefore requires the analysis of several individuals per accession to adequately capture the diversity within the accession, making the process potentially very expensive. However, grouping individuals into “bulks” or “pools” of population samples, for example, results in a smaller number of physical samples to be processed, significantly minimizing genotyping labor and reducing costs by at least an order of magnitude. In this study, three methods of DNA bulk assembly were evaluated for their accuracy by comparing the estimates of allele frequencies obtained from pooled samples with the estimate from genotyping of individual samples. The methods were: (1) Equimolar bulk DNA samples extracted from 16 individual plants and quantified by spectrometry in a NanoDrop equipment; (2) Same as (1) but the DNA quantified for double-stranded DNA using PicoGreen in a fluorimeter; (3) Bulk DNA extraction from leaf tissue samples weighed individually with a precision of 0.1 gram. The allele frequencies of 3.526 SNPs estimated from allelotyping of DNA bulks were compared with the frequencies calculated from individual genotyping of the same 16 plants. This procedure was performed for four distinct maize accessions sampled from the Brazilian Maize Germplasm Bank. The accuracy of the allele frequency estimation was assessed by calculating the RMSE (Root Mean Square Error), which quantifies the difference between the allele frequency estimated from the bulks and the predicted values from the genotyping of individual plants. The method that provided the lowest RMSE (Root Mean Square Error) was the bulk leaf extraction method, with an average RMSE of 0.063 across the 3.526 SNPs. Furthermore, the allele frequencies estimated from leaf bulks samples showed a high correlation with those estimated from genotypic data of individual plants, with an average correlation coefficient of $r = 0.984$. The reproducibility between biological replicates, evaluated based on leaf bulks from 24 individuals for 64 distinct maize accessions using the Wright fixation index, indicated precise estimates ($F_{ST} = 0.0038$), indicating that the standard error due to sampling is negligible. This study provides a rapid, economical, and easy-to-implement bulk assembly method and shows that the estimates of allele frequencies obtained by allelotyping can be highly accurate, thus opening positive perspectives for the use of this approach for the genetic analysis of large collections of germplasm of allogamous plants.

Keywords: *Zea mays* L.; Germplasm; Allelotyping; DNA pooling; Genetic Diversity; SNPs.

1. INTRODUÇÃO

A avaliação da diversidade e estrutura genética de acessos de germoplasma de milho é fundamental para a manutenção, regeneração e intercâmbio ativo dessas coleções, bem como para a sustentabilidade do melhoramento genético da cultura, principalmente no desenvolvimento de linhagens endogâmicas melhoradas e na definição de potenciais combinações parentais para a criação de híbridos fenotipicamente superiores e com maior rendimento quando comparados aos parentais (CGIAR, 2007; SEMAGN *et al.*, 2012).

A diversidade genética de coleções de germoplasma de milho tem sido extensivamente estudada utilizando diferentes tipos de marcadores moleculares (ARCA *et al.*, 2021). A título de exemplo, 488 populações de milho europeias (GAUTHIER *et al.*, 2002) e 148 populações de milho americanas (MUMM; HUBERT; DUDLEY, 1994) foram caracterizadas em nível molecular utilizando polimorfismo de comprimento de fragmentos de restrição (RFLP). Marcadores RAPD, polimorfismo de DNA amplificado ao acaso, foram utilizados para caracterizar a diversidade genética de variedades locais de milho duro (CARVALHO *et al.*, 2004), raças locais (OKUMUS, 2007) e linhagens de milho (BRUEL *et al.*, 2007).

Linhagens endogâmicas de milho foram analisadas geneticamente em estudos baseados marcadores AFLP, polimorfismo de comprimento de fragmentos amplificados (AFLP) (LABORDA *et al.*, 2005; ZHANG; WAN; PAN, 2006; OLIVEIRA *et al.*, 2004). Marcadores microsatélites foram amplamente utilizados para investigar a diversidade genética de milhos de origens diversas (LABATE *et al.*, 2003), linhagens endogâmicas (WARBURTON *et al.*, 2002; VAZ PATTO *et al.*, 2004) e populações locais de milho (DUBREUIL *et al.*, 2006; ADU *et al.*, 2019).

Entretanto, os avanços na tecnologia molecular produziram uma mudança em direção ao uso de marcadores de polimorfismo de nucleotídeo único (SNPs) (HAMBLIN; WARBURTON; BUCKLER, 2007; SEMAGN *et al.*, 2012). Nos últimos anos, a abordagem de genotipagem de SNPs, seja baseada em microarranjos ou chips de sondas imobilizadas ou em tecnologias de genotipagem por sequenciamento (GbS), emergiu como uma ferramenta eficiente para muitas aplicações, incluindo estudos de diversidade genética, mapeamento de ligação e de QTLs, melhoramento assistido por marcadores e geração de dados genômicos de alta qualidade para grande número de amostras (KUMAR, B., *et al.* 2022).

A utilização de marcadores SNPs tornou-se muito popular por vários motivos: marcadores SNPs são a classe mais abundante de variação de sequência no genoma, são herdados de forma codominante, possuem baixo custo por ponto de dados, especificidade por

loco, além de serem geneticamente estáveis e apropriados para análises automatizadas de alto rendimento com menores taxas de erro de genotipagem, esta última característica relativa predominantemente a sistemas de microarranjos ou chips de SNPs (SEMAGN *et al.*, 2012; ARCA *et al.*, 2021; KUMAR, B., *et al.* 2022). Estudos de caracterização molecular em milho anteriores ao uso de SNPs, utilizaram principalmente marcadores microssatélites que, embora sejam altamente polimórficos, não são adequados para a caracterização de germoplasma em larga escala, por conta do número limitado de marcadores, imprecisão da genotipagem devido a desvios de tamanho dos alelos e uma distribuição desigual no genoma, apresentando sérias limitações de portabilidade, reprodutibilidade e escalabilidade de dados (MAMMADOV *et al.*, 2012; PRASANNA, 2012; KUMAR, B., *et al.* 2022). Portanto, marcadores SNPs, na impraticabilidade de realizar o sequenciamento completo do genoma, são atualmente o marcador de escolha para análise genética de seres vivos eucariotos em geral e plantas em particular, entre elas o milho (ARCA *et al.*, 2021).

Marcadores SNPs tem sido extensamente utilizados para compreender a diversidade genética, estrutura populacional, estrutura do desequilíbrio de ligação de linhagens endogâmicas (ROMAY *et al.* 2013; BOAKYEWAA ADU *et al.*, 2019; KUMAR, B. *et al.*, 2022), principalmente, e raças locais de milho (SILVA, N. *et al.*, 2020). Além disso, muitos estudos de análise da diversidade genética de linhagens (YAN *et al.*, 2009; VAN INGHELANDT *et al.*, 2010; LU *et al.*, 2011; SEMAGN *et al.*, 2012; VAN HEERWAARDEN; HUFFORD; ROSS-IBARRA, 2012), raças locais (ARTEAGA *et al.*, 2016; ARCA *et al.*, 2020; ARCA *et al.*, 2021; DIAW *et al.* 2021) e, até mesmo, teosinto (AGUIRRE-LIGUORI *et al.*, 2017) têm utilizado “chips” de genotipagem de SNPs de alta densidade recentemente desenvolvidas para o milho, que evoluíram com rapidez de apenas alguns milhares de SNPs (ROMAY *et al.*, 2013), para mais de 50.000 com o Illumina® MaizeSNP50 BeadChip, e até 600.000 SNPs em chips de alta densidade (UNTERSEER *et al.* 2014). Até o presente momento, a tecnologia baseada em “chips” é a plataforma de genotipagem de SNPs “padrão ouro” utilizada na indústria para a análise genética de milho com maior qualidade de dado, maior rendimento e, portanto, mais eficiente do ponto de vista da relação custo/benefício.

Mesmo com os recentes avanços na tecnologia molecular, a genotipagem de um grande número de indivíduos continua a ser muito dispendiosa no contexto da caracterização genética para muitas coleções de germoplasma (ARCA *et al.*, 2021). Esse desafio torna-se ainda maior para espécies alógamas, ou seja, de fecundação cruzada como o milho, devido à alta diversidade nos acessos. Para essas espécies, um acesso corresponde a uma população com ampla

diversidade entre os indivíduos. A caracterização de acessos de milho que não são linhagens, como no caso de variedades tradicionais e indígenas, demanda, portanto, a análise de vários indivíduos por acesso para capturar a diversidade existente entre indivíduos dentro de um acesso, tornando o processo potencialmente muito caro (DUBREUIL; CHARCOSSET, 1998; ARCA *et al.*, 2021).

Como resultado, a abordagem de análise de “bulks” ou “pools” de indivíduos, também chamada de “alelotipagem”, pode ser utilizada como uma alternativa conveniente para a genotipagem individual permitindo estimar as frequências alélicas do conjunto de indivíduos com precisão, reduzindo o trabalho e os custos da genotipagem (SHAW *et al.*; 1998; ARCA *et al.*, 2020, ARCA *et al.*, 2021). Essa abordagem foi utilizada inicialmente com marcadores RFLP (ARNHEIM; STRANGE; ERLICH, 1985) e microssatélites (SHAW *et al.*, 1998) e, mais recentemente, com “chips” de SNPs de alta densidade e sequenciamento “*Next generation*” principalmente para a realização de estudos de associação “caso-controle” em seres humanos (SHAM *et al.*, 2002; TEUMER *et al.*, 2013; SCHLÖTTERER *et al.*, 2014) e em menor escala para estudos de populações naturais de peixes (JOHNSTON *et al.*, 2013; OZEROV *et al.*, 2013) e de populações de milho (REBOURG; GOUESNARD; CHARCOSSET, 2001; MIR *et al.*, 2013; ARCA *et al.*, 2021) e arroz (GOUDA *et al.*, 2020).

Ademais, a alelotipagem foi eficientemente utilizada para detecção de QTLs (LIPKIN *et al.*, 1998; GALLAIS; MOREAU; CHARCOSSET, 2007), caracterização da diversidade genética (DUBREUIL *et al.*, 2006; JOHNSTON *et al.*, 2013; DIAW *et al.* 2021), realização de estudos de associação genômica (SHAM *et al.*, 2002; BAUM *et al.*, 2008) e identificação de assinaturas de seleção (ELFERINK *et al.*, 2012; ARCA *et al.*, 2020; XU *et al.*, 2022).

Portanto, a abordagem de alelotipagem de indivíduos de raças locais de milho com SNPs genotipados em chips pode ser uma alternativa fundamental para a análise da diversidade genética de coleções de germoplasma e uma ferramenta valiosa para promover potenciais avanços no melhoramento genético da espécie, bem como na manutenção, regeneração, intercâmbio ativo e caracterização dessas coleções de germoplasma.

2. MATERIAIS E MÉTODOS

2.1 MATERIAL GENÉTICO

Amostras de 48 sementes de cada um de quatro acessos (WP-21 Composto Cuba; RR-035; RS-101; RS-229) da coleção nuclear do Banco Ativo de Germoplasma de Milho da EMBRAPA foram germinadas em tubetes próprios em estufa, visando a obtenção de 16 plantas a serem utilizadas na montagem dos bulks de amostras DNA dos acessos para a genotipagem de SNPs. As amostras de tecido foliar foram coletadas em aproximadamente 30 a 40 dias e desidratadas em sílica-gel para posterior utilização na extração de DNA genômico total.

2.2 EXTRAÇÃO DE DNA

As extrações foram realizadas a partir de tecido de folhas de plântulas germinadas utilizando o protocolo otimizado desenvolvido no Laboratório de Genética Vegetal (LGV) do Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN-EMBRAPA), baseado na utilização de tampão de lise contendo o detergente CTAB (brometo de cetiltrimetilamônio) com alto teor de sal (INGLIS *et al.*, 2018).

As extrações realizadas baseadas em tampão CTAB consistem em duas etapas de lise: (1) mecânica, em que as amostras são submetidas à agitação em macerador automático, e logo após, (2) química, com tampão de lise contendo CTAB. Esse método consiste ainda nas etapas de separação dos componentes celulares utilizando-se CIA (clorofórmio: álcool isoamílico, 24:1), de precipitação do DNA por meio da adição de isopropanol, e por último, de purificação do DNA com o uso de etanol a 70% para remover resíduos. Os passos do protocolo estão descritos em maior detalhe no Anexo 1.

2.3 EXPERIMENTO DE OTIMIZAÇÃO DE MONTAGEM DE BULKS

Para cada um dos quatro acessos, foram extraídas amostras de DNA de cada uma de 16 plântulas (total de 64). As amostras de DNA de plântulas individuais foram quantificadas com dois métodos: absorvância via NanoDrop™ 2000 (*Thermo Fisher Scientific*) e fluorescência de picogreen utilizando um fluorímetro Qubit® (*Invitrogen*). Foram utilizados três métodos de montagem de bulks (Figura 1) para posterior comparação: (1) Bulk de folhas: montagem de

bulk com discos de folhas de igual tamanho, como em Dubreuil *et al.* (1999) e Arca *et al.* (2020), de cada uma das 16 plântulas, com precisão de peso total de até 0,12 gramas, seguida da maceração dos tecidos de forma conjunta e extração de uma amostra única de DNA em bulk; (2) NanoDrop: montagem de bulk equimolar de DNA quantificado via NanoDrop; (3) Qubit: montagem de bulk equimolar de DNA quantificado via fluorímetro. Para montagem equimolar dos bulks quantificados via NanoDrop e via fluorímetro, foram utilizados 5 ul de cada uma das 16 amostras de DNA individuais e, então, as amostras foram diluídas e ajustadas para uma concentração final de 20ng/ul. Para a montagem da amostra do bulk, portanto, foi reunido 10 ul de cada uma das 16 amostras diluídas, totalizando 3200 ng em 160 ul. Foram montadas três réplicas para o método de montagem de bulk (1) e duas réplicas para cada um dos métodos (2) e (3), além disso, foram montados 4 bulks adicionais reunindo bulks de duas cultivares (RS-101; RS-229), sendo duas réplicas do método (1) e duas do método (2), totalizando 32 bulks montados (Tabela 1).

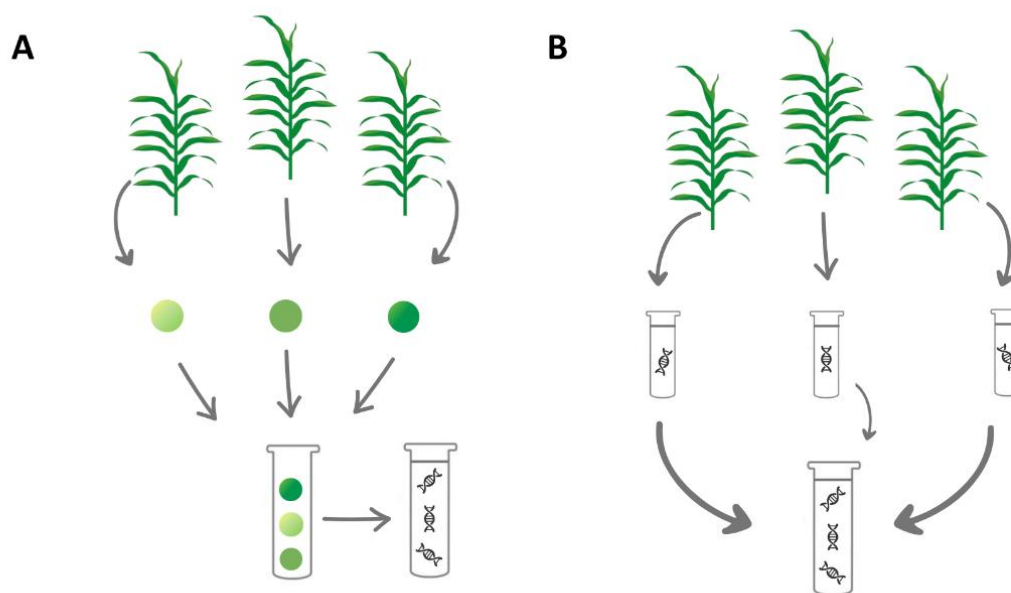


Figura 1 – Fluxograma esquemático da montagem de bulks. a) Método de extração foliar em bulk: o procedimento de mistura começa com amostras de tecido foliar (pré-extração de DNA). b) Métodos de bulk equimolar de DNA quantificado via NanoDrop e via fluorímetro: o procedimento de mistura acontece pipetando cada amostra de DNA para obter uma mistura de DNA equimolar (pós-extração).

Tabela 1 – Esquema experimental dos métodos de preparação das amostras de DNA agrupadas em bulks com as respectivas repetições. BE = bulk de extração a partir de folhas; BN = Bulk de DNA quantificado com Nanodrop; BQ = bulks de DNA quantificado com Qubit; réplicas (numeradas 1, 2 e 3).

Cultivar	Bulk	Amostra	Método
WP-21	WP_21_BE1	Tecido foliar	Extração foliar
	WP_21_BE2	Tecido foliar	Extração foliar
	WP_21_BE3	Tecido foliar	Extração foliar
	WP_21_BN1	DNA	NanoDrop
	WP_21_BN2	DNA	NanoDrop
	WP_21_BQ1	DNA	Qubit
	WP_21_BQ2	DNA	Qubit
RR-035	RR_035_BE1	Tecido foliar	Extração foliar
	RR_035_BE2	Tecido foliar	Extração foliar
	RR_035_BE3	Tecido foliar	Extração foliar
	RR_035_BN1	DNA	NanoDrop
	RR_035_BN2	DNA	NanoDrop
	RR_035_BQ1	DNA	Qubit
	RR_035_BQ2	DNA	Qubit
RS-101	RS_101_BE1	Tecido foliar	Extração foliar
	RS_101_BE2	Tecido foliar	Extração foliar
	RS_101_BE3	Tecido foliar	Extração foliar
	RS_101_BN1	DNA	NanoDrop
	RS_101_BN2	DNA	NanoDrop
	RS_101_BQ1	DNA	Qubit
	RS_101_BQ2	DNA	Qubit
RS-229	RS_229_BE1	Tecido foliar	Extração foliar
	RS_229_BE2	Tecido foliar	Extração foliar
	RS_229_BE3	Tecido foliar	Extração foliar
	RS_229_BN1	DNA	NanoDrop
	RS_229_BN2	DNA	NanoDrop
	RS_229_BQ1	DNA	Qubit
	RS_229_BQ2	DNA	Qubit
RS-101+RS-229	RS_101+RS_229_BE1+BE1	Tecido foliar	Extração foliar
	RS_101+RS_229_BE2+BE2	Tecido foliar	Extração foliar
	RS_101+RS_229_BN1+BN1	DNA	NanoDrop
	RS_101+RS_229_BN2+BN2	DNA	NanoDrop

2.4 GENOTIPAGEM DE SNPs

As amostras individuais de DNA das 64 plântulas e as 32 amostras em bulk foram genotipadas ou alelotipadas, respectivamente, para 3.526 SNPs presentes no *EMBRAPA 65kMultispecies chip*. Esses 3.526 SNPs foram previamente selecionados a partir do Illumina® MaizeSNP50 BeadChip (GANAL *et al.* 2011), no design do chip com base no conteúdo

informativo visando ainda cobrir todos os 10 cromossomos do milho de forma o mais equidistante possível de forma a proporcionar uma cobertura genômica adequada para estudos de diversidade e estrutura genética. As amostras de DNA foram enviadas para genotipagem de forma terceirizada para a Neogen do Brasil (SP).

2.5 ESTIMATIVA DAS FREQUÊNCIAS ALÉLICAS

Todos os SNPs presentes no chip são do tipo A/G, A/C, T/G ou T/C sendo que as bases A ou T correspondem ao código de alelo A e as bases C ou G ao alelo B. Os dados genotípicos foram exportados em formato AB após controle de qualidade padrão da Illumina (2010) a partir dos dados de intensidade de fluorescência utilizando o software GENOMESTUDIO (Illumina Inc., San Diego, CA, USA). As intensidades de fluorescências foram carregadas no software GENOMESTUDIO e os genótipos exportados após controle de qualidade padrão da Illumina (2010). As frequências alélicas a partir das plantas genotipadas individualmente foram calculadas para cada um dos 3.526 SNPs com o software GENALEX 6.5 (PEAKALL; SMOUSE, 2012). As frequências alélicas das amostras de DNA em bulk foram estimadas pela estimativa normalizada do *B Allele Frequency* (BAF) (para o alelo B) e 1-B (para o alelo A) para cada um dos 3.526 SNPs. A BAF foi estimada por meio do módulo específico do software GENOMESTUDIO (Illumina Inc., San Diego, CA, USA) utilizado tipicamente para a análise de CNVs (copy number variants - variantes de número de cópias). BAF é uma medida normalizada da razão de intensidade alélica dos dois alelos SNP (A e B). BAF próximo a 1 ou 0 indica a ausência completa de um dos alelos (AA ou BB), e valores intermediários de BAF correspondem às frequências alélicas dos dois alelos ao SNP.

2.6 FILTRAGEM DE SNP E CONTROLE DE QUALIDADE

A partir dos 3.526 SNPs presentes no *EMBRAPA 65kMultispecies chip* foi aplicado um filtro de taxa de chamada de genótipo, comumente denominada de *call rate* (CR) de 87,5%. A *call rate* para um determinado SNP é definida como a proporção de indivíduos no estudo para os quais os dados do SNP correspondente foram declarados. Portanto, é uma métrica utilizada para eliminar SNPs com grande quantidade de “dados perdidos” (*missing data*) devido a problemas de genotipagem (LAURIE *et al.*, 2010; REED *et al.*, 2015). A Equação 2.1 mostra como se calcula a CR:

$$CR = \frac{x}{y} \quad (2.1)$$

Onde x representa o número de indivíduos com dados válidos por SNP (*nonmissing data*) e y é o número total de indivíduos no estudo/amostra. Como exemplo, considere para um determinado marcador SNP genotipado em 16 indivíduos, sendo que desse total para genótipos foram obtidos para 14 indivíduos. Assim, a CR é calculada pela Equação 2:

$$CR = \frac{14}{16} = 0,875 \quad (2.2)$$

Após a aplicação do filtro de CR nos 3.526 SNPs, foram retidos dados para 3.464 SNPs para o acesso WP-21, 3.430 SNPs para o acesso RR-035, 3.432 SNPs para o acesso RS-101, 3.432 SNPs para o acesso RS-229 e 3.419 SNPs para os acessos BRA88480_RS101 e RS-229 reunidos.

2.7 ANÁLISE DE DADOS

2.7.1 Cálculo de Root-mean-square Error

As estimativas de frequências alélicas obtidas com os três diferentes métodos de montagem de bulks foram comparadas com aquelas determinadas pela genotipagem direta por meio da estimativa da raiz do erro quadrático médio (RMSE, do inglês *Root-mean-square Error*), como em Shaw *et al.* (1998) e Gautier *et al.* (2013). O RMSE foi calculado para cada método de bulk e cada um dos 4 acessos via fórmulas no Excel pela raiz da soma dos quadrados das diferenças entre a frequência estimada pelo “*B allele*” e a calculada a partir dos genótipos das plântulas individuais dividido pelo número total de SNPs. O RMSE é uma métrica utilizada para quantificar as diferenças entre os valores estimados das frequências alélicas por um modelo (no caso os três métodos de montagem de bulks) e os valores esperados, ou seja, as frequências efetivamente estimadas a partir dos genótipos individuais (SHAW *et al.*, 1998), penalizando os erros muito elevados (KARUNASINGHA, 2022). A Equação 2.3 foi utilizada para calcular o RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (2.3)$$

Onde n é o número total de observações, y_i é o valor da variável analisada e $\hat{y}_i$ é o valor previsto para a variável analisada com base em um modelo. Portanto, n é o número total de SNPs, $\hat{y}_i$ é a frequência alélica do alelo B obtida a partir dos dados genotípicos das amostras individuais e y_i é a frequência do alelo B equivalente à BAF do bulk.

Quanto menor o RMSE, mais acurada é a estimativa fornecida pelo modelo (WANG, 2007; CHICCO; WARRENS; JURMAN, 2021). Valores de até 5% de RMSE são considerados aceitáveis para o caso em questão, mesmo porque o erro sobre a estimativa da frequência alélica associado com a simples amostragem de sementes de um acesso tende a ser maior do que isso (LE HELLARD *et al.*, 2002; JAWAID; SHAM, 2009; ARCA *et al.*, 2021).

2.7.2 Coeficiente de Correlação de Pearson

A acurácia da estimativa das frequências alélicas a partir dos bulks foi também avaliada com base no coeficiente médio de correlação de Pearson (r) entre as frequências esperadas pela genotipagem individual e as estimadas a partir dos bulks seguindo o procedimento utilizado por Shaw *et al.* (1998), Gruber; Colligan e Wolford (2002) e Bélanger *et al.* (2016). A correlação de Pearson foi estimada com o Excel. A análise de correlação linear entre duas variáveis é amplamente utilizada para dizer o quanto uma variável está relacionada com outra. O coeficiente da correlação de Pearson pode variar entre -1,0 e 1,0, sendo esses valores máximos para correlações negativas e positivas, respectivamente, enquanto 0 indica que não há uma correlação linear entre as variáveis (STANTON, 2001). Desse modo, correlações de frequências alélicas acima de 0,9 são consideradas muito fortes. A Equação 2.4 é utilizada para calcular o coeficiente de correlação de Pearson entre duas variáveis:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2} \quad (2.4)$$

Onde X_i e Y_i são valores individuais das variáveis X e Y , e $\bar{X}$ e $\bar{Y}$ são as médias de X e Y , respectivamente.

2.7.3 Coeficiente de Determinação

Ainda para a avaliação da acurácia das estimativas das frequências alélicas, o coeficiente de determinação (R^2) foi estimado por meio de regressão linear para avaliar a relação estatística entre as frequências alélicas estimadas em bulk (y) e as frequências alélicas estimadas na genotipagem individual (x) através do software R (R Core Team, 2020). O modelo de regressão linear é dado por:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (2.5)$$

Onde Y é a variável dependente (representa o que o modelo tentará prever), β_0 e β_1 são constantes que representam o ponto de interseção e a inclinação, respectivamente, X é a variável independente (explicativa) e ε é erro aleatório.

O ajuste da regressão linear foi dado pela função *lm* com o formato *lm* ($y \sim x$) ou seja, a variável resposta é y e a preditora é x, sempre nessa ordem. Um intervalo de confiança de 95% para respostas individuais (predição) foi obtido através da função *predict*, dado pela equação 2.6:

$$y_0 = \hat{y}_0 \pm t_{1-\alpha/2, n-2} \cdot \sqrt{MSE \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x} \right)} \quad (2.6)$$

Em que y_0 é o valor real do parâmetro previsto, $\hat{y}_0$ é o valor previsto usando a relação de regressão, t é o valor crítico da distribuição T de Student em nível de significância de $1 - \alpha$ e $n-2$ graus de liberdade, n é o tamanho da amostra, $(x_0 - \bar{x})$ é a diferença entre o valor médio de x e x_0 para o qual se está tentando prever y_0 e SS_x é o desvio ao quadrado da amostra de valores x .

Ainda utilizando função *predict*, foi dada a porcentagem de pontos fora do intervalo de confiança de 95% e, então, obtido a porcentagem de pontos dentro dos limites de confiança. O coeficiente de determinação foi obtido através da função *summary* (Multiple R-Squared), dado por:

$$R^2 = r^2 \quad (2.7)$$

Valores próximos de 1 implicam alta relação entre o modelo e os dados, enquanto valores próximos de 0 indicam um poder explicativo baixo do modelo (SAUNDERS; RUSSELL; CRABB, 2012; ALEXANDER; TROPSHA; WINKLER, 2015).

2.8 AVALIAÇÃO DO EFEITO DE AMOSTRAGEM DE PLANTAS NA ESTIMATIVA DE FREQUÊNCIAS ALÉLICAS

O objetivo deste experimento verificar o impacto nas estimativas de frequências alélicas ao utilizar amostras de mesmo tamanho ($n= 22$ a 24) porém compostas por diferentes plantas de um mesmo acesso. Para isso foram preparadas réplicas biológicas na forma de dois bulks independentes amostrando 22 a 24 plantas diferentes para cada bulk para um total de 102 acessos da Coleção Nuclear de Germoplasma de Milho. Esta avaliação foi realizada por meio da estimativa do índice de fixação de Wright (F_{ST}) (WRIGHT, 1978) entre os dois bulks de cada acesso. O índice de fixação é uma medida de diferenciação populacional amplamente utilizado para quantificar a distância genética entre populações (BATHIA *et al.*, 2013). O F_{ST} foi calculado para cada par de réplicas utilizando o software STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000; FALUSH; STEPHENS; PRITCHARD, 2003). Portanto quanto menores forem os valores de F_{ST} entre as estimativas obtidas para os dois bulks menor é o impacto do efeito de amostragem de quais indivíduos são utilizados na montagem de um bulk. Estimativas elevadas de F_{ST} indicariam, por outro lado, um forte impacto do efeito de amostragem de quais plantas são utilizadas na montagem de um bulk.

Considerando dois bulks (A e B) para um determinado acesso, tem-se que o $F_{ST(AB)}$ para cada marcador i foi dado por:

$$F_{ST(AB)i} = 1 - \frac{0.5 \times [p_{Ai} \times (1 - p_{Ai}) + p_{Bi} \times (1 - p_{Bi})]}{0.5 \times (p_{Ai} + p_{Bi}) \times 0.5 \times [(1 - p_{Ai}) + (1 - p_{Bi})]} \quad (2.8)$$

Em que, p_{Ai} e p_{Bi} são as frequências alélicas do i -ésimo marcador para os bulks A e B, respectivamente.

O $F_{ST(AB)}$ final entre o par de bulks foi dado pela média do $F_{ST(AB)i}$ de todos os marcadores SNPs considerados. Assim, $F_{ST(AB)} = \sum_{i=1}^N \frac{F_{ST(AB)i}}{N}$ em que N é o total de marcadores SNPs alelotipados em ambos os bulks. Tipicamente é adotado um critério *ad hoc* no qual valores de F_{ST} entre 0,00 e 0,05 indicam pequena diferenciação genética entre populações

- no caso em questão entre réplicas biológicas do acesso - enquanto valores entre 0,05 e 0,15 e valores 0,15 e 0,25 indicam moderada e alta diferenciação genética, respectivamente (WRIGHT, 1978).

3. RESULTADOS

Dos 3.526 SNPs presentes no *EMBRAPA 65kMultispecies chip*, foram alelotipados com sucesso mais de 3.400 SNPs após aplicação do filtro de *call rate*. Foram alelotipados 3.464 SNPs para o acesso WP-21 (98,2 %), 3.430 SNPs para RR-035 (97,3 %), 3.432 SNPs para RS-101 (97,33%), 3.432 SNPs para RS-229 (97,33%) e 3.419 SNPs para os acessos RS-101 e RS-229 reunidos (97%).

3.1 COMPARAÇÃO DOS MÉTODOS DE MONTAGEM DE BULKS

Um total de 109.967 estimativas de RMSE foram obtidas ao considerar de forma conjunta os 3.430 a 3.464 SNPs e os 32 bulks preparados e alelotipados. A média dos valores de RMSE por SNP individual foi de 0,00494, variando de um mínimo de zero a um máximo de 0,334. Raros (313 em 109.967; 0,28%) foram os casos de SNPs com valores de RMSE acima de 0,10. A mediana dos valores de RMSE foi de 0,00071, percentil 95% igual a 0,023 e percentil 99% igual 0,028 o que indica que 99% dos 109.967 valores de RMSE foram abaixo de 6%. Este resultado demonstra que as estimativas de frequências alélicas obtidas para cada SNP individual têm RMSE plenamente satisfatório, abaixo de 2,5% para 95% dos SNPs e abaixo de 6% ao considerar 99% dos SNPs alelotipados.

Em seguida foi considerado o RMSE da avaliação conjunta de todos os mais de 3.400 SNPs genotipados para cada um dos 32 bulks. De maneira geral os valores de RMSE ao considerar todos os SNPs genotipados foram estimados entre 0,05 e 0,10 e somente para os bulks de DNA quantificado por Qubit para um dos acessos (WP_21) alcançou 0,12. Entre os três métodos de montagem de bulks, o método de extração foliar em massa apresentou o menor erro quadrático médio (RMSE) entre as frequências alélicas estimadas a partir dos bulks e as frequências alélicas calculadas a partir de genótipos de amostras individuais (Tabela 2).

Tabela 2 – Estimativas da raiz do erro quadrático médio (RMSE), correlação (r) e coeficiente de determinação (R^2) entre as frequências alélicas estimadas a partir do bulk e as frequências alélicas calculadas a partir de genótipos de amostras individuais. Estimativas são listadas para: (i) quatro diferentes acessos de milho (WP_21, RR_035, RS_101, RS_229) ou combinações destes; (ii) os três métodos de preparação de bulks (BE = bulk de extração a partir de folhas; BN = Bulk de DNA quantificado com Nanodrop; BQ = bulks de DNA quantificado com Qubit) para cada uma das réplicas (numeradas 1, 2 e 3).

Bulk	SNPs genotipados	RMSE	r	R^2
WP_21_BE1	3524	0,073	0,976	0,9533
WP_21_BE2	3524	0,072	0,977	0,9552
WP_21_BE3	3524	0,076	0,976	0,9517
WP_21_BN1	3524	0,085	0,963	0,9267
WP_21_BN2	3524	0,086	0,962	0,9257
WP_21_BQ1	3524	0,124	0,925	0,8554
WP_21_BQ2	3524	0,121	0,93	0,8642
RR_035_BE1	3521	0,056	0,99	0,9805
RR_035_BE2	3521	0,056	0,99	0,9805
RR_035_BE3	3521	0,055	0,991	0,9812
RR_035_BN1	3521	0,062	0,988	0,9757
RR_035_BN2	3521	0,062	0,988	0,9756
RR_035_BQ1	3521	0,070	0,984	0,9686
RR_035_BQ2	3521	0,071	0,984	0,9678
RS_101_BE1	3524	0,062	0,984	0,9674
RS_101_BE2	3524	0,063	0,983	0,9670
RS_101_BE3	3524	0,062	0,984	0,9675
RS_101_BN1	3524	0,079	0,973	0,9461
RS_101_BN2	3524	0,069	0,979	0,9590
RS_101_BQ1	3524	0,071	0,978	0,9565
RS_101_BQ2	3524	0,078	0,975	0,9499
RS_229_BE1	3522	0,051	0,992	0,9845
RS_229_BE2	3522	0,053	0,992	0,9833
RS_229_BE3	3522	0,052	0,992	0,9843
RS_229_BN1	3522	0,059	0,989	0,9789
RS_229_BN2	3522	0,059	0,990	0,9792
RS_229_BQ1	3522	0,061	0,989	0,9773
RS_229_BQ2	3522	0,062	0,988	0,9771
RS_101+RS_229_BE1+BE1	3522	0,080	0,974	0,9487
RS_101+RS_229_BE2+BE2	3522	0,078	0,975	0,9509
RS_101+RS_229_BN1+BN1	3522	0,078	0,975	0,9514
RS_101+RS_229_BN2+BN2	3522	0,080	0,974	0,9487

De forma consistente, o método de extração foliar em massa resultou na estimativa de frequências alélicas com valores de RMSE mais baixos ou essencialmente equivalentes àqueles obtidos a partir dos métodos de montagem de bulk equimolar de DNA quantificado via

NanoDrop e via fluorímetro Qubit. Correlações entre as frequências alélicas estimadas a partir dos bulks e calculadas na genotipagem individual foram elevadas para praticamente todos os métodos possivelmente como resultado da lei dos grandes números pelo fato destas estimativas serem derivadas de dados para mais de 3.400 SNPs.

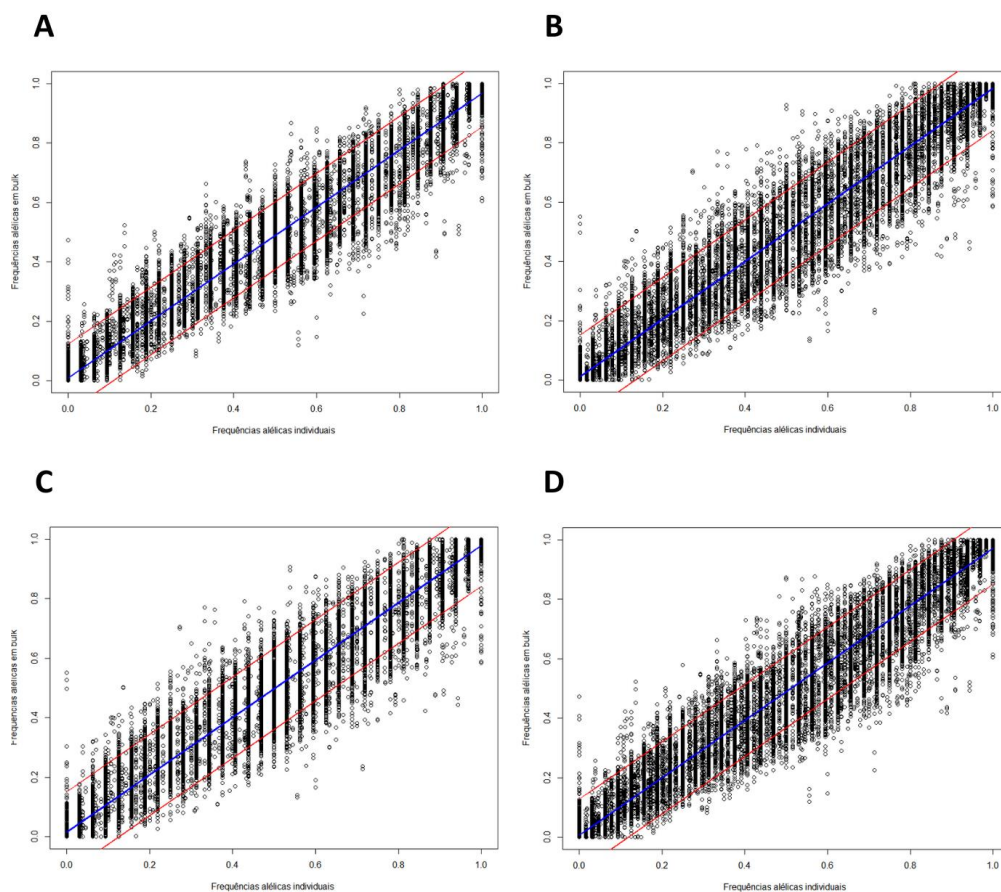
A estimativas das frequências alélicas diretas e agrupadas em todos os três métodos de bulks foram altamente correlacionadas ($r > 0,9$). O método de extração foliar em massa apresentou o maior valor de correlação média ($r = 0,984$) e os valores de r nesse método variaram entre 0,974 (RS_101+RS_229_BE1+BE1) e 0,992 (RS_229_BE1). Para comparação, a correlação média foi menor para os bulks quantificados via fluorímetro ($r = 0,967$), com valores de r variando entre 0,924 (WP_21_BQ1) e 0,988 (RS_229_BQ1). Para os bulks quantificados via Nanodrop, os valores de r variaram entre 0,962 (WP_21_BN2) e 0,989 (RS_229_BN2), com um valor médio de $r = 0,978$.

Consistente com os elevados valores de correlação, os valores dos coeficientes determinação (Figura 2) foram elevados e altamente significativos ($p < 0,001$). Os valores do coeficiente de determinação (R^2) foram acima de 0,9 para todos os métodos, variando entre 0,969 e 0,973 no método de extração foliar em massa para os bulks com e sem os bulks de diferentes cultivares conjuntas, respectivamente. O método de bulk por extração foliar em massa apresentou uma maior porcentagem de pontos dentro dos limites de confiança de 95% (93,54%), essa média aumenta ainda mais (93,83%) quando os bulks envolvendo duas cultivares não foram considerados. A menor porcentagem de pontos dentro do intervalo de confiança (93,17%) foi observado para o método de bulks de DNA quantificado via fluorímetro, resultante de uma maior dispersão de pontos observada com este método. No método de bulk de DNA quantificado via NanoDrop o R^2 variou entre 0,959 e 0,962, com e sem os bulks de cultivares conjuntas, respectivamente, e foi de 0,945 para o método de bulk de DNA quantificado via fluorímetro.

As regressões lineares entre as frequências estimadas a partir dos bulks e calculadas a partir de genótipos individuais foram altamente significativas, e apenas em dois casos (WP_21_BQ1 e WP_21_BQ2) das 32 comparações o coeficiente de determinação (R^2) foi abaixo de 0,9 (Tabela 2). Considerando os três métodos de montagem de bulks, os valores de R^2 variaram entre 0,855 (WP_21_BQ1) e 0,984 (RS_229_BE1). Os valores de R^2 variaram entre 0,948 (RS_101+RS_229_BE1+BE1) e 0,984 (RS_229_BE1) no método de extração foliar em massa, entre 0,925 (WP_21_BN2) e 0,979 (RS_229_BN2) no método de bulks de DNA

quantificado via NanoDrop e, entre 0,855 (WP_21_BQ1) e 0,977 (RS_229_BQ1) no método de bulks de DNA quantificado via Qubit.

Importante observar que de maneira geral, mesmo os bulks de DNA previamente extraído e quantificado resultaram em valores aceitáveis de RMSE com destaque para o fato que o método de menor acurácia foi de bulk preparados a partir de DNA quantificado por Qubit. Os resultados indicam ainda uma variabilidade desprezível para os valores de RMSE e correlação entre as diferentes réplicas de um mesmo método. Assim, embora múltiplas réplicas possam ser úteis para adicionar confiança a um experimento de alelotipagem, em termos efetivos, elas não aumentam substancialmente a acurácia das estimativas de frequências alélicas.



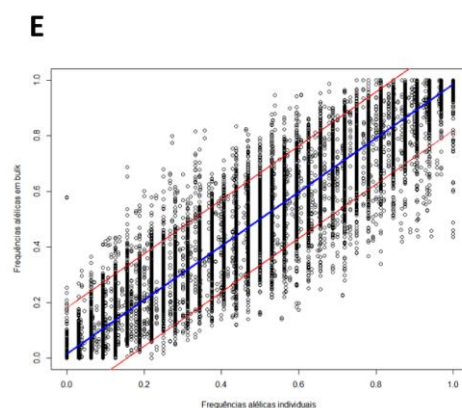


Figura 2 – Gráficos de correlação (*correlation plot*) entre frequências alélicas estimadas por meio de bulks de DNA (eixo Y) e calculadas a partir de dados genotípicos (eixo X) para os mais de 3430 SNPs (pontos no gráfico). Os pontos representam as estimativas de frequência para os SNPs considerados para cada método de preparação de bulk. A linha azul indica a regressão linear e as linhas vermelhas representam o intervalo de confiança a 95% de probabilidade. (A) método de extração foliar em massa para acessos analisados individualmente; (B) método de extração foliar em massa incluindo bulks de duas cultivares; (C) método de bulk de DNA quantificado via NanoDrop para acessos analisados individualmente; (D) método de bulk de DNA quantificado via NanoDrop incluindo bulks de duas cultivares; (E) método de bulk de DNA quantificado via Qubit para acessos analisados individualmente.

3.2 AVALIAÇÃO DO EFEITO DE AMOSTRAGEM DE PLANTAS NA ESTIMATIVA DE FREQUÊNCIAS ALÉLICAS

Os valores de F_{ST} estimados entre as duas réplicas alelotipadas para os mais de 3400 SNPs analisados conjuntamente nos 102 acessos analisados, foram extremamente baixos indicando não haver diferença significativa nas estimativas de frequências alélicas obtidas nos dois bulks independentes, ou seja, entre as duas réplicas biológicas. Valores variaram de 0,000831 a 0,013117 com média de 0,003852, mediana de 0,003648 e percentil 95% igual a 0,0075 (Tabela 3). Este resultado obtido para uma amostra considerável de 102 acessos indica que o efeito de quais 22 a 24 plantas são amostradas ao acaso para a composição de um bulk a ser utilizado para a estimativa de frequências alélicas é essencialmente desprezível.

Tabela 3. Estimativas do índice de fixação (F_{ST}) entre réplicas biológicas na avaliação do efeito de amostragem de plantas na montagem de bulks por meio do método de extração foliar em massa.

Acesso	
AC-005	0,004442
AC-006	0,005054
AC-014	0,004933
AC-022	0,004659
AC-027	0,007393

AC-031	0,004815
AC-036	0,006904
AL-001	0,004696
AL-018	0,005646
AL-020	0,000831
AM-003	0,003119
AM-006	0,001969
AM-007	0,002026
AZTECA	0,001033
BA-003	0,001118
BA-007	0,006097
BA-019	0,003272
BA-027	0,005886
BA-042	0,004422
BA-050	0,001997
BA-064	0,004290
BA-066	0,005364
BA-080	0,001861
BA-094	0,005742
BA-116	0,005735
BA-117	0,005524
BA-122	0,004669
BA-125	0,004828
BA-154	0,003581
BA-166	0,005356
BA-178	0,005558
BA-194	0,005731
BA-197	0,007589
BA-237	0,004667
BR-126-XIII	0,012133
CAMELIA	0,004211
CATETO-SETE-LAGOAS	0,003628
CE-029	0,001104
CE-039	0,003083
CMS-04-AMARILLO-DENTADO	0,003719
CMS-05-SUWAN-DMR	0,002981
CMS-15-POOL-26	0,003416
CMS-36-SINTETICO-CERRADO	0,002710
COMPUESTO-TUXPENNO-VERACRUZIANO	0,002950
MAYA-OPACO-2	0,004021
MG-060	0,002772
MG-069	0,004426
MG-076	0,003135
MG-149-PIPOCA-TIMOTEO	0,003451
MS-046-PIPOCA	0,002121
MS-052	0,003440
MS-061	0,011105

MT-008	0,004217
MT-009	0,004774
MT-033	0,001460
MT-035	0,003832
NODZOB-RONRE-TROPE	0,010495
PA-049	0,005484
PA-090	0,013117
PB-010	0,003713
PE-001	0,001525
PE-002	0,005741
PR-005	0,004223
PR-015	0,003276
PR-025	0,004420
PR-030	0,004438
PR-034	0,003312
PR-039	0,004247
PR-053	0,003474
PR-123-TUNICATA	0,001095
RN-003	0,003667
RO-015	0,004321
RR-003	0,002268
RR-006	0,002494
RR-007	0,004974
RR-012	0,001035
RR-021	0,001070
RR-022	0,007034
RR-028	0,001288
RR-035	0,001570
RR-036	0,002568
RR-047	0,002570
RR-048	0,001172
RR-064	0,000894
RS-009	0,000948
RS-020-PIPOCA	0,004952
RS-101	0,000917
RS-102	0,002317
RS-109	0,001138
RS-229	0,001509
RS-230	0,001421
RS-715	0,001036
SC-005	0,004071
SC-010	0,001416
SC-015	0,002220
SC-042	0,000971
SE-016	0,003500
SP-181	0,002947
TUXPEN0-1	0,003808

TUXPENNO-CARIBE	0,004801
WP-21-COMPOSTO-CUBA	0,001492
WP-2-AMARELO-DENTADO	0,008476
Média:	0,003852

4. DISCUSSÃO

A análise genética de marcadores moleculares de uma grande número de indivíduos para cada um de centenas de acessos ou populações é, ainda, uma tarefa trabalhosa e cara o que tem limitado o trabalho de caracterização de coleções de germoplasma (SHAW *et al.*, 1998; FUTSCHIK; SCHLÖTTERER, 2010). Entretanto, o agrupamento de indivíduos em amostras de bulks populacionais levando a um menor número de amostras físicas a serem processadas, minimiza significativamente os custos e o trabalho do processo de genotipagem (GAUTIER *et al.*, 2013). Embora vários estudos já tenham utilizado a alelotipagem de acessos ou populações em seres humanos bem como diferentes espécies de plantas e animais, poucos relatos existem para plantas bem como no que se refere à avaliação de diferentes métodos de preparação de bulks de DNA, passo central para a otimização do processo.

Os resultados deste estudo indicam que a alelotipagem de amostras em bulk de DNA extraído diretamente de uma amostra agrupada de tecidos foliares de peso equivalente fornece estimativas acuradas de frequências alélicas, ou seja, estimativas muito próximas dos valores considerados de referência. Além disso, o estudo mostrou que a variabilidade entre réplicas biológicas compostas por 22 a 24 indivíduos, o que corresponde a 44 a 48 cromossomos amostrados em uma planta diploide alógama, tem magnitude desprezível o que indica que o efeito de amostragem pelo menos ao se considerar marcadores SNPs bialélicos é irrelevante. A alelotipagem de SNPs se mostrou, portanto, como uma excelente alternativa que permite estudos genômicos a custos e esforço de trabalho reduzidos.

O sucesso da abordagem de alelotipagem de SNPs em um grupo de indivíduos analisados em bulk de DNA depende da acurácia, também denominada de exatidão (proximidade de um valor estimado com um valor de referência), com a qual as frequências alélicas a um marcador SNP bialélico são estimadas (ARCA *et al.*, 2021). Basicamente existem três fontes de erro que afetam essa acurácia. O primeiro é o erro associado à própria amostragem de quais indivíduos são tomados ao acaso para compor o grupo. Este fator em geral é considerado a maior fonte de variabilidade nas estimativas. Entretanto, é importante lembrar que o erro de amostragem não é exclusivo do processo de alelotipagem, mas também presente no procedimento de genotipagem individualizada. Amostras muito reduzidas de indivíduos

tendem a gerar estimativas de frequência alélica com um intervalo de confiança muito amplo em torno delas (GAUTIER *et al.*, 2013; ARCA *et al.*, 2021).

O erro de amostragem é mitigado aumentando o número de indivíduos no bulk e, portanto, aumentando a probabilidade de amostrar alelos de baixa frequência no grupo. É possível calcular a priori o intervalo de confiança (IC) com 95% de probabilidade em torno de uma estimativa de frequência alélica para um SNP bialélico para diferentes números de indivíduos incluídos no bulk e assim decidir qual amostragem melhor compatibiliza custos, trabalho e acurácia. Por exemplo, para um alelo cuja frequência na população é igual a 0,05, usando um bulk de $N=20$ indivíduos, ou seja, $2N = 40$ cromossomos, o IC a 95% vai de 0,0138 a 0,165. Se a amostra for de $N = 50$ indivíduos, $2N = 100$ cromossomos, a acurácia aumenta reduzindo o IC de 0,0215 a 0,112.

Estava além do escopo desse estudo testar o efeito do número de indivíduos utilizados no bulk, ou seja o tamanho do bulk na acurácia das estimativas de frequência de alelos das populações. Foram assim utilizados tamanhos fixos sendo um bulk de 16 indivíduos no experimento de avaliação dos métodos de preparação de bulks e de 22 a 24 indivíduos no experimento de avaliação das réplicas, já em preparação à análise de todo o banco de germoplasma descrito no capítulo 2. No experimento de avaliação de métodos de montagem de bulks, uma amostra de 16 indivíduos por bulk se mostrou satisfatória para obter estimativas de frequências alélicas com alta correlação e baixo RMSE em comparação com as frequências alélicas obtidas a partir da genotipagem individual, apresentando ($r > 0,9$) para todos os 32 bulks avaliados nesse estudo. Correlações equivalentes foram encontradas quando bulks com tamanhos de amostra de 15 indivíduos (GOUDA *et al.*, 2020; ARCA *et al.*, 2021) e 20 indivíduos (BANSAL *et al.*, 2011; GAUTIER *et al.*, 2013) foram utilizados. Este resultado indicou que um tamanho de $n=16$ já seria satisfatório embora pudesse ser aumentado, o que efetivamente foi, passando para $n=22$ ou 24 na etapa seguinte do trabalho.

A fonte de erro derivada do processo de amostragem de quais plantas compõem o bulk foi avaliada quantitativamente pelo índice de fixação de Wright, (F_{ST}), que mede a diferença de frequências alélicas entre populações ou também corresponde à proporção da variância genética total contida em uma subpopulação em relação à variância genética da população total. No caso deste experimento o F_{st} foi utilizado como uma medida da diferença de frequências alélicas entre as duas amostras distintas de um mesmo acesso, ou como a proporção da variância genética total contida em uma amostra em relação à variação total contida no acesso.

Estimativas de F_{ST} foram obtidas na comparação de frequências alélicas entre duas réplicas biológicas, ou seja, bulks preparados com amostras de 22 ou 24 plantas distintas, alelotipadas para cada um de 102 acessos de milho. Das 102 estimativas de F_{ST} obtidas nesta análise, 95% delas apresentaram valores abaixo de 0,0075. Embora existam tentativas de estabelecer intervalos de valores de F_{ST} para considerar diferenciação genética baixa ($< 0,05$), moderada (0,05 a 0,15) ou alta ($> 0,15$) (HARTL; CLARK, 1997), estas tentativas são subjetivas, pois, as estimativas de F_{ST} dependem de diversos fatores. Entre eles estão o tipo e taxa evolutiva do marcador molecular, o tamanho da amostra, se a estimativa é para um SNP individual ou múltiplos SNPs ou ainda se ocorre uma seleção de SNPs utilizados na estimativa, e finalmente o próprio estimador de F_{ST} utilizado (BHATIA *et al.* 2013). De qualquer forma, no âmbito deste estudo os valores de F_{ST} observados predominantemente abaixo de 0,0075 podem seguramente ser considerados muitos baixos do ponto de vista de diferenciação genética entre as réplicas biológicas avaliadas. Adicionalmente o fato que estimativas foram obtidas para 102 acessos independentes fornece suporte adicional à esta conclusão (Tabela 3). Isto indica que pelo menos no que se refere a acessos de milho, o efeito de amostragem de plantas na preparação de bulks compostos por 22 a 24 indivíduos tem efeito essencialmente desprezível nas estimativas de frequências alélicas obtidas pelo método de alelotipagem.

A segunda fonte de variabilidade na estimativa das frequências alélicas é o próprio procedimento de reunião das contribuições individuais de material biológico, seja ele tecido ou DNA, para a composição de uma amostra agrupada em bulk a ser alelotipada. O impacto desta segunda fonte de variabilidade foi avaliado ao comparar três métodos de preparação de bulks. A extração direta da partir de uma amostra composta de tecido foliar de vários indivíduos foi o método mais simples e rápido, porém depende de uma adequada pesagem de tecido e maceração homogênea. Isso foi obtido ao se utilizar discos cortados de folha com dimensões equivalentes e uma secagem completa do tecido foliar antes da maceração de forma a obter uma pulverização completa do tecido. Desta forma todo o tecido foi submetido à extração e purificação de uma mesma forma, o que deve ter contribuído de forma importante para que o DNA de todos os indivíduos estivesse representado na amostra final de DNA genotipado no chip. Os resultados mostraram uma alta correspondência entre as frequências alélicas estimadas pela genotipagem individual e aquelas estimadas em bulk para todos os três métodos de preparação de bulks utilizando as métricas de RMSE e correlação (Tabela 2). Este resultado está em linha com estudos anteriores que mostraram que se a mistura, tanto de tecido antes da extração, quanto das amostras individuais de DNA, for realizada com o zelo inerente, a alelotipagem fornece

dados robustos de frequências alélicas com valores de correlação com frequências de referência acima de 0,9 (LE HELLARD *et al.*, 2002; JAWAID; SHAM, 2009; GOUDA *et al.* 2020, ARCA *et al.*, 2021; SHAW *et al.*, 1998; GRUBER; COLLIGAN; WOLFORD, 2002; GAUTIER *et al.*, 2013; RELLSTAB *et al.*, 2013, BÉLANGER *et al.*, 2016).

Entre os três métodos avaliados, entretanto, o método de extração foliar em massa foi superior aos demais. Ele alcançou a maior acurácia das estimativas de frequências alélicas avaliada pelo menor erro quadrático médio entre as frequências alélicas de referência e frequências alélicas estimadas a partir dos bulks, com uma média de 0,063, quando comparado com os métodos de montagem de bulks com amostras de DNA quantificados via NanoDrop (0,079) e via fluorímetro Qubit (0,082). Os valores do coeficiente de correlação também apresentaram média maior para o método de extração foliar em massa ($r = 0,984$), variando entre 0,974 e 0,992, quando comparados aos valores observados para os bulks preparados via Nanodrop ($r = 0,978$), variando entre 0,962 e 0,989 e Qubit ($r = 0,967$), variando entre 0,924 e 0,988. Ainda assim, correlações com as frequências esperadas, estimadas com os três métodos foram maiores do que aquelas relatadas para pools de DNA previamente extraídos por Van Tassel *et al.* (2008) ($r = 0,67$), Out *et al.* (2009) ($r = 0,86$) e Shaw *et al.* (1998) ($r = 0,95$) e o espectro de variação foi menor do que aquele encontrado por Gautier *et al.* ($r = 0,93 - 0,99$) e Bélanger *et al.* (2016) ($r = 0,91 - 0,98$). As regressões lineares entre as frequências alélicas de referência e aquelas estimadas a partir dos bulks foram altamente significativas ($p < 0,001$) e os valores do coeficiente de determinação também foram maiores para o método de extração foliar em massa ($R^2 = 0,948 - 0,984$) quando comparados com os outros métodos, sendo com NanoDrop ($R^2 = 0,925 - 0,979$) e com Qubit ($R^2 = 0,855 - 0,977$). Valores semelhantes foram descritos por Rellstab *et al.* (2013) ($R^2 = 0,979$), Holt *et al.* (2009) ($R^2 = 0,92 - 0,95$), Druley *et al.* (2009) ($R^2 = 0,96$), Ingman e Gyllensten (2009) ($R^2 = 0,88 - 0,91$) encontrados utilizando pools de DNA.

Além de resultar em estimativas de frequências alélicas mais acuradas, o método de extração foliar em massa reduziu significativamente o número de extrações de DNA em comparação aos demais métodos, uma vez que as amostras de tecido foliar já são montadas em bulk, enquanto nos métodos envolvendo bulks de DNA, estes são extraídos de cada planta individualmente e apenas depois é realizado o agrupamento, pipetando cada amostra para a obtenção de um bulk de DNA equimolar. Portanto, como o custo da extração de DNA de uma grande quantidade de amostras também não é negligenciável (ARCA *et al.*, 2021), o método de extração foliar em massa, em que há a mistura de material com base em sua massa antes de

extração, é altamente vantajoso. Dois estudos anteriores realizados com RFLP (DUBREUIL *et al.* 1999) e SNPs (ARCA *et al.*, 2021), este último em milho também, já haviam demonstrado esta vantagem ao estimar coeficientes de determinação muito altos ($R^2 = 0.99$).

O fato dos dois métodos de preparação de bulks utilizado DNA previamente extraído terem sido inferiores ao método de extração em bulk de folhas, foi de certa forma inesperado. Os dois métodos foram selecionados para o experimento considerando que eles são comumente utilizados e sabidamente tem precisão distinta. O método de quantificação por espectrofotometria com NanoDrop embora amplamente utilizado no mundo todo por sua facilidade, é considerado menos preciso do que com base em fluorescência utilizando Picogreen no fluorímetro Qubit. Alguns problemas importantes do NanoDrop incluem: contaminação de outras substâncias tais como polissacarídeos e proteínas que afetam as leituras de absorbância, baixa sensibilidade em baixas concentrações, heterogeneidade da amostra, variabilidade da técnica por parte do usuário, e a necessidade de calibração e manutenção regulares para garantir resultados precisos; tudo isso pode levar a medições não confiáveis da concentração de DNA.

Em vista disso, recomendações têm sido feitas de se usar fluorimetria como método mais preciso por medir especificamente DNA de fita dupla, (GOLDSSTEIN; HAUSMANN; DOUVDEVANI, 2009; Garcia-Elias *et al.* 2017), moléculas que efetivamente são utilizadas em análises que envolvem amplificação via PCR o que inclui a genotipagem em chips de DNA. O método comumente utilizado para quantificar DNA por fluorimetria é por meio do Picogreen um corante fluorescente que permite a quantificação de DNA fita dupla (dsDNA) e a leitura feita em um fluorímetro Qubit. Embora este método seja, a princípio, mais preciso do que o Nanodrop ele não é livre de problemas uma vez a intensidade de fluorescência é afetada por várias condições experimentais tais como pH, sais e reagentes químicos (BAN; KIM, 2024).

O fato dos dois métodos de preparação de bulks com base em DNA previamente extraído terem sido inferiores ao método de bulk de folhas pode ser explicado por pelo menos dois fatores. O primeiro diz respeito ao fato de que as próprias quantificações de DNA contêm um erro associado devido a possíveis interferências nas leituras de absorbância ou fluorimetria conforme mencionado acima. Além disso este erro acontece em cada uma das várias amostras e, portanto, é cumulativo no bulk final. O segundo fator diz respeito ao erro operacional associado ao grande número de pipetagens necessárias para a normalização das concentrações de DNA de cada uma das amostras e por fim da reunião de quantidades iguais de DNA na amostra em bulk. Em vista destes dois fatores, a preparação de bulks de DNA a partir de uma extração única, desde que as quantidades em peso de tecido sejam bem equivalentes e calibradas

por meio de algum sistema de amostragem de tecido como aquele utilizado neste estudo (discos foliares), espera-se que sempre serão superiores.

A terceira fonte de variabilidade na estimativa de frequências alélicas se refere ao processo de detecção e quantificação relativa dos alelos a um marcador no bulk. Esta fonte de variabilidade depende essencialmente da natureza do marcador, se multialélico ou bialélico e do método molecular de detecção e quantificação da frequência dos alelos. A alelotipagem baseada em microssatélites tem sido, historicamente, a metodologia mais utilizada. Entretanto ela é susceptível a problemas de detecção desigual de alelos de forma variável a diferentes marcadores, uma vez que cada marcador apresenta múltiplos alelos os quais competem durante a amplificação via PCR, além de problemas comuns de “gaguejo” de alelos. Alelos que por algum motivo estocástico não são igualmente amplificados nas etapas iniciais da PCR podem ou não ser detectados ou ser detectados em frequências inferiores às aquelas efetivamente presentes no bulk.

Métodos matemáticos tem sido propostos para mitigar estes problemas (SCHNACK *et al.* 2004) embora com sucesso variável a depender dos marcadores e espécie analisada. No caso de SNPs detectados de forma bi alélica por fluorescência em microarranjos de sondas, como o caso do sistema utilizado neste trabalho, são consideravelmente menos susceptíveis a problemas de não detecção ou detecção parcial. O ensaio Infinium II de chips Illumina, metodologia utilizada neste trabalho, é baseado na hibridização específica do alelo acoplada a uma extensão de base única específica de alelo usando marcação de duas cores com os corantes fluorescentes Cy3 e Cy5 (GUNDERSON *et al.* 2005; STEEMERS *et al.* 2006).

Além disso o ensaio é realizado em dois canais e os dados consistem em dois valores de intensidade (X, Y) para cada SNP, com um canal de intensidade para cada um dos corantes fluorescentes associados aos dois alelos do SNP. Os marcadores SNP estão presentes em alta redundância no chip (dezenas de sondas replicadas para cada SNP) e as intensidades específicas do alelo (X, Y) são, portanto, estimativas resumidas derivadas destas múltiplas sondas replicadas o que contribui substancialmente para uma elevada robustez analítica do ensaio. As intensidades de sinal alelo específicas são por fim normalizadas usando um algoritmo proprietário no software de análise GENOMESTUDIO utilizado para a exportação de dados genotípicos e alelotípicos (BAF). Todos estes fatores contribuem sobremaneira para uma elevada especificidade e acurácia na quantificação dos dois alelos em um bulk, explicando os excelentes resultados obtidos neste trabalho. Vale mencionar que a estimativa de BAF foi anteriormente utilizada, com eficiência, para estimar frequências alélicas a partir de pools de

DNA para espécies de salmão do Atlântico (OZEROV *et al.*, 2013a; OZEROV *et al.*, 2013b; OZEROV *et al.*, 2016), bovinos (STRILLACCI *et al.*, 2014a; STRILLACCI *et al.*, 2014b; GONZÁLEZ-RUIZ *et al.*, 2019), humanos (BAKHSH, 2014) e, ainda, para comparar frequências alélicas e agrupadas em humanos (JANICKI; LIU, 2009). Entretanto, no melhor da nossa avaliação este é o primeiro trabalho com utilização da estimativa de BAF em amostras em bulk para uma espécie de planta.

Neste contexto, vale mencionar ainda para fins comparativos a possibilidade de alelotipagem em bulks de amostras via sequenciamento de alto rendimento (NGS *Next Generation Sequencing*). Este método vem sendo utilizado principalmente para a descoberta, mas também para a alelotipagem, pois permite a detecção e quantificação de frequências alélicas para grandes números de SNPs de forma não enviesada, incluindo SNPs muito raros. Entretanto este método é sujeito a variabilidade na medida que a capacidade de detecção de alelos alternativos a um SNP depende em grande medida da cobertura de sequenciamento utilizada e da taxa de erro do método de sequenciamento, demandando assim ajustes estatísticos complexos das estimativas (FUTSCHIK; SCHLÖTTERER, 2010), ou a utilização de coberturas muito altas de sequenciamento para a obtenção de estimativas acuradas de frequências alélicas. Entretanto nos últimos anos com a evolução de tecnologias que levaram a uma forte redução de custos de sequenciamento e das taxas de erro, aliadas ao desenvolvimento de novos softwares de análise, esta abordagem vem se tornando cada vez melhor com a vantagem de poder ser aplicada em qualquer espécie sem a necessidade do desenvolvimento prévio de chips de genotipagem.

Por fim, a utilização de réplicas de bulks pode também ser utilizada como uma estratégia para auxiliar a mitigar todas as três fontes de variabilidade acima descritas. Entretanto, os valores essencialmente desprezíveis de F_{ST} observados entre as réplicas analisadas para 102 acessos (Tabela 3) indicaram que para a alelotipagem com SNPs em microarranjos, réplicas não contribuem para a obtenção de maior acurácia nas estimativas de frequências alélicas.

Em conclusão, este estudo mostra que as estimativas de frequências alélicas derivadas da alelotipagem a partir de bulks de DNA preparados de forma simples e econômica a partir de uma extração única agrupando tecido foliar de diversos indivíduos podem ter elevada acurácia em relação às frequências de referência estimadas a partir da genotipagem de plantas individuais. Esta abordagem analítica é especialmente interessante para estudos em escala genômica que requerem uma amostragem de grande número de indivíduos de populações ou acessos contendo altos níveis de diversidade genética intra-populacional.

5. REFERÊNCIAS BIBLIOGRÁFICAS

- ADU, G. B. *et al.* Genetic characterization and population structure of maize populations using SSR markers. **Annals of Agricultural Sciences**, v. 64, n. 1, p. 47-54, 2019.
- AGUIRRE-LIGUORI, J. A. *et al.* Connecting genomic patterns of local adaptation and niche suitability in teosintes. **Molecular Ecology**, v. 26, n. 16, p. 4226-4240, 2017.
- ALEXANDER, D. L.; TROPSHA, A.; WINKLER, D. A. Beware of R²: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. **Journal of chemical information and modeling**, v. 55, n. 7, p. 1316-1322, 2015.
- ARCA, M. *et al.* Deciphering the genetic diversity of landraces with high-throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array. **Frontiers in Plant Science**, v. 11, p. 568699, 2021.
- ARCA, M. *et al.* Genome-wide SNP genotyping of DNA pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool. **BioRxiv**, 2020.
- ARNHEIM, N.; STRANGE, C.; ERLICH, H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II locos. **Proceedings of the National Academy of Sciences**, v. 82, n. 20, p. 6970-6974, 1985.
- ARTEAGA, M. C. *et al.* Genomic variation in recently collected maize landraces from Mexico. **Genomics data**, v. 7, p. 38-45, 2016.
- BAKSH, A. D. **Molecular characterisation of benign and malignant thyroid dysfunction**. Tese de Doutorado – School of Medicine, Cardiff University, Cardiff, 2014.
- BAN, E.; KIM, A. PicoGreen Assay for Nucleic Acid Quantification-Applications, Challenges, and Solutions. **Analytical Biochemistry**, p. 115577, 2024.
- BANSAL, V. *et al.* Efficient and cost effective population resequencing by pooling and in-solution hybridization. **PloS one**, v. 6, n. 3, p. e18353, 2011.
- BHATIA, G. *et al.* Estimating and interpreting FST: the impact of rare variants. **Genome research**, v. 23, n. 9, p. 1514-1521, 2013.
- BAUM, A. E. *et al.* A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. **Molecular psychiatry**, v. 13, n. 2, p. 197-207, 2008.
- BÉLANGER, S. *et al.* Genotyping-by-sequencing on pooled samples and its use in measuring segregation bias during the course of androgenesis in barley. **The plant genome**, v. 9, n. 1, p. plantgenome2014.10.0073, 2016.
- BOAKYEWAA ADU, G. *et al.* Genetic diversity and population structure of early-maturing tropical maize inbred lines using SNP markers. **PloS one**, v. 14, n. 4, p. e0214810, 2019.

BRUEL, D. C. *et al.* Assessment of genetic diversity in maize inbred lines using RAPD markers. **Crop Breeding and Applied Biotechnology**, v. 7, n. 2, 2007.

CARVALHO, V. P. *et al.* Genetic diversity among maize (*Zea mays* L.) landraces assessed by RAPD markers. **Genetics and Molecular Biology**, v. 27, p. 228-236, 2004.

CGIAR. CONSULTATIVE GROUP ON INTERNATIONAL AGRICULTURAL RESEARCH. **Global Strategy for the Ex-situ Conservation and Utilization of Maize Germplasm**. Washington : CGIAR, 2007. Disponível em: <<https://www.genebanks.org/resources/publications/>>. Acesso em: 08 jul. 2024.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. **Peerj computer science**, v. 7, p. e623, 2021.

DIAM, Y. *et al.* Genetic diversity of maize landraces from the South-West of France. **PLoS One**, v. 16, n. 2, p. e0238334, 2021.

DRULEY, T. E. *et al.* Quantification of rare allelic variants from pooled genomic DNA. **Nature methods**, v. 6, n. 4, p. 263-265, 2009.

DUBREUIL, P. *et al.* Evaluation of a DNA pooled-sampling strategy for estimating the RFLP diversity of maize populations. **Plant Molecular Biology Reporter**, v. 17, n. 2, p. 123-138, 1999.

DUBREUIL, P. *et al.* More on the introduction of temperate maize into Europe: large-scale bulk SSR genotyping and new historical elements. 2006.

DUBREUIL, P.; CHARCOSSET, A. Genetic diversity within and among maize populations: a comparison between isozyme and nuclear RFLP locos. **Theoretical and Applied Genetics**, v. 96, p. 577-587, 1998.

ELFERINK, M. G. *et al.* Signatures of selection in the genomes of commercial and non-commercial chicken breeds. **PloS one**, v. 7, n. 2, p. e32720, 2012.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J. K. Inference of population structure using multiloco genotype data: linked locos and correlated allele frequencies. **Genetics**, v. 164, n. 4, p. 1567-1587, 2003.

FUTSCHIK, A.; SCHLÖTTERER, C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. **Genetics**, v. 186, n. 1, p. 207-218, 2010.

GALLAIS, A.; MOREAU, L.; CHARCOSSET, A. Detection of marker-QTL associations by studying change in marker frequencies with selection. **Theoretical and Applied Genetics**, v. 114, n. 4, p. 669-681, 2007.

GANAL, M. W. *et al.* A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. **PloS one**, v. 6, n. 12, p. e28334, 2011.

GARCIA-ELIAS, A. *et al.* Defining quantification methods and optimizing protocols for microarray hybridization of circulating microRNAs. **Scientific reports**, v. 7, n. 1, p. 7725, 2017.

GAUTIER, M. *et al.* Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. **Molecular Ecology**, v. 22, n. 14, p. 3766-3779, 2013.

GAUTHIER, P. *et al.* RFLP diversity and relationships among traditional European maize populations. **Theoretical and Applied Genetics**, v. 105, p. 91-99, 2002.

GONZÁLEZ-RUIZ, S. *et al.* Genome-wide association study in mexican holstein cattle reveals novel quantitative trait loci regions and confirms mapped loci for resistance to bovine tuberculosis. **Animals**, v. 9, n. 9, p. 636, 2019.

GOUDA, A. C. *et al.* Comparisons of sampling methods for assessing intra-and inter-accession genetic diversity in three rice species using genotyping by sequencing. **Scientific Reports**, v. 10, n. 1, p. 13995, 2020.

GOLDSHTEIN, H.; HAUSMANN, M. J.; DOUVDEVANI, A. A rapid direct fluorescent assay for cell-free DNA quantification in biological fluids. **Annals of clinical biochemistry**, v. 46, n. 6, p. 488-494, 2009.

GRUBER, J. D.; COLLIGAN, P. B.; WOLFORD, J. K. Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing. **Human genetics**, v. 110, p. 395-401, 2002.

GUNDERSON, K. L. *et al.* A genome-wide scalable SNP genotyping assay using microarray technology. **Nature genetics**, v. 37, n. 5, p. 549-554, 2005.

HAMBLIN, M. T.; WARBURTON, M. L.; BUCKLER, E. S. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. **PloS one**, v. 2, n. 12, p. e1367, 2007.

HARTL, D. L. **Principles of population genetics**. 3 ed. Sunderland: Sinauer Association Inc, 1997.

HOLT, K. E. *et al.* Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. **Bioinformatics**, v. 25, n. 16, p. 2074-2075, 2009.

INGLIS, P. W. *et al.* Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. **PloS one**, v. 13, n. 10, p. e0206085, 2018.

INGMAN, M.; GYLLENSTEN, U. SNP frequency estimation using massively parallel sequencing of pooled DNA. **European Journal of Human Genetics**, v. 17, n. 3, p. 383-386, 2009.

JANICKI, P.; LIU, J. Accuracy of allele frequency estimates in pool DNA analyzed by high-density Illumina Human 610-Quad microarray. **Internet J Genom Proteomics**, v. 5, 2009.

JAWAID, A.; SHAM, P. Impact and quantification of the sources of error in DNA pooling designs. **Annals of Human Genetics**, v. 73, n. 1, p. 118-124, 2009.

JOHNSTON, S. E. *et al.* Fish scales and SNP chips: SNP genotyping and allele frequency estimation in individual and pooled DNA from historical samples of Atlantic salmon (*Salmo salar*). **Bmc Genomics**, v. 14, p. 1-13, 2013.

KARUNASINGHA, D. S. K. Root mean square error or mean absolute error? Use their ratio as well. **Information Sciences**, v. 585, p. 609-629, 2022.

KUMAR, B. *et al.* Genetic diversity, population structure and linkage disequilibrium analyses in tropical maize using genotyping by sequencing. **Plants**, v. 11, n. 6, p. 799, 2022.

LABATE, J. A. *et al.* Molecular and historical aspects of corn belt dent diversity. **Crop Science**, v. 43, n. 1, p. 80-91, 2003.

LABORDA, P. R. *et al.* Tropical maize germplasm: what can we say about its genetic diversity in the light of molecular markers?. **Theoretical and Applied Genetics**, v. 111, p. 1288-1299, 2005.

LAURIE, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic epidemiology**, v. 34, n. 6, p. 591-602, 2010.

LE HELLARD, S. *et al.* SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. **Nucleic acids research**, v. 30, n. 15, p. e74-e74, 2002.

LIPKIN, E. *et al.* Quantitative trait loco mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. **Genetics**, v. 149, n. 3, p. 1557-1567, 1998.

LU, Y. *et al.* Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapid LD decay in tropical than temperate germplasm in maize. **PloS one**, v. 6, n. 9, p. e24861, 2011.

MAMMADOV, J. *et al.* SNP markers and their impact on plant breeding. **International journal of plant genomics**, v. 2012, n. 1, p. 728398, 2012.

MIR, C. *et al.* Out of America: tracing the genetic footprints of the global diffusion of maize. **Theoretical and Applied Genetics**, v. 126, p. 2671-2682, 2013.

MUMM, R. H.; HUBERT, L. J.; DUDLEY, J. W. A classification of 148 US maize inbreds: II. Validation of cluster analysis based on RFLPs. **Crop science**, v. 34, n. 4, p. 852-865, 1994.

OKUMUS, A. Genetic variation and relationship between Turkish flint maize landraces by RAPD markers. **Am J Agri Biol Sci**, v. 2, n. 2, p. 49-53, 2007.

OLIVEIRA, K. M. *et al.* Evaluating genetic relationships between tropical maize inbred lines by means of AFLP profiling. **Hereditas**, v. 140, n. 1, 2004.

OUT, A. A. *et al.* Deep sequencing to reveal new variants in pooled DNA samples. **Human mutation**, v. 30, n. 12, p. 1703-1712, 2009.

OZEROV, M. *et al.* Cost-effective genome-wide estimation of allele frequencies from pooled DNA in Atlantic salmon (*Salmo salar* L.). **BMC genomics**, v. 14, p. 1-9, 2013a.

OZEROV, M. *et al.* Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. **PLoS One**, v. 8, n. 12, p. e82434, 2013b.

OZEROV, M. Y. *et al.* Genomewide introgressive hybridization patterns in wild Atlantic salmon influenced by inadvertent gene flow from hatchery releases. **Molecular Ecology**, v. 25, n. 6, p. 1275-1293, 2016.

PRASANNA, B. M. Diversity in global maize germplasm: characterization and utilization. **Journal of biosciences**, v. 37, n. 5, p. 843-855, 2012.

PEIRIS, B. L. *et al.* Predicting allele frequencies in DNA pools using high density SNP genotyping data. **Animal genetics**, v. 42, n. 1, p. 113-116, 2011.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945-959, 2000.

R CORE TEAM: A language and environment for statistical computing, R Foundation for Statistical Computing, 2020. Disponível em <<https://www.R-project.org/>>. Acesso em: 03 ago. 2024.

REBOURG, C.; GOUESNARD, B.; CHARCOSSET, A.. Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation. **Heredity**, v. 86, n. 5, p. 574-587, 2001.

REED, E. *et al.* A guide to genome-wide association analysis and post-analytic interrogation. **Statistics in medicine**, v. 34, n. 28, p. 3769-3792, 2015.

RELLSTAB, C. *et al.* Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. **PLoS one**, v. 8, n. 11, p. e80422, 2013.

ROMAY, M. C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. **Genome biology**, v. 14, p. 1-18, 2013.

SAUNDERS, L. J.; RUSSELL, R. A.; CRABB, D. P. The coefficient of determination: what determines a useful R² statistic?. **Investigative ophthalmology & visual science**, v. 53, n. 11, p. 6830-6832, 2012.

SCHLÖTTERER, C. *et al.* Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. **Nature Reviews Genetics**, v. 15, n. 11, p. 749-763, 2014.

SCHNACK, H. G. *et al.* Accurate determination of microsatellite allele frequencies in pooled DNA samples. **European journal of human genetics**, v. 12, n. 11, p. 925-934, 2004.

SEMAGN, K. *et al.* Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. **BMC genomics**, v. 13, p. 1-11, 2012.

SHAM, P. *et al.* DNA pooling: a tool for large-scale association studies. **Nature Reviews Genetics**, v. 3, n. 11, p. 862-871, 2002.

SHAW, S. H. *et al.* Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. **Genome research**, v. 8, n. 2, p. 111-123, 1998.

SILVA, N. C. D. A. *et al.* Relationships among American popcorn and their links with landraces conserved in a microcenter of diversity. **Genetic Resources and Crop Evolution**, v. 67, p. 1733-1753, 2020.

STANTON, J. M. Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. **Journal of Statistics Education**, v. 9, n. 3, 2001.

STEEMERS, F. J. *et al.* Whole-genome genotyping with the single-base extension assay. **Nature methods**, v. 3, n. 1, p. 31-33, 2006.

STRILLACCI, M. G. *et al.* Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. **BMC genetics**, v. 15, p. 1-10, 2014a.

STRILLACCI, M. G. *et al.* Quantitative trait loci mapping for conjugated linoleic acid, vaccenic acid and Δ 9-desaturase in Italian Brown Swiss dairy cattle using selective DNA pooling. **Animal genetics**, v. 45, n. 4, p. 485-499, 2014b.

TEUMER, A. *et al.* Comparison of genotyping using pooled DNA samples (allelotyping) and individual genotyping using the affymetrix genome-wide human SNP array 6.0. **BMC genomics**, v. 14, p. 1-12, 2013.

UNTERSEER, S. *et al.* A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. **BMC genomics**, v. 15, p. 1-15, 2014.

VAN HEERWAARDEN, J.; HUFFORD, M. B.; ROSS-IBARRA, J. Historical genomics of North American maize. **Proceedings of the National Academy of Sciences**, v. 109, n. 31, p. 12420-12425, 2012.

VAN INGHELANDT, D. *et al.* Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. **Theoretical and Applied Genetics**, v. 120, p. 1289-1299, 2010.

VAZ PATTO, M. C. *et al.* Assessing the genetic diversity of Portuguese maize germplasm using microsatellite markers. **Euphytica**, v. 137, p. 63-72, 2004.

VAN TASSELL, C. P. *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. **Nature methods**, v. 5, n. 3, p. 247-252, 2008.

WANG, J. Triadic IBD coefficients and applications to estimating pairwise relatedness. **Genetics Research**, v. 89, n. 3, p. 135-153, 2007.

WARBURTON, M. L. *et al.* Genetic characterization of CIMMYT inbred maize lines and open pollinated populations using large scale fingerprinting methods. **Crop Science**, v. 42, n. 6, p. 1832-1840, 2002.

WRIGHT, S. **Evolution and Genetics of Populations**. Chicago: University of Chicago, 1978. 511p

XU, G. *et al.* Population genomics of *Zea* species identifies selection signatures during maize domestication and adaptation. **BMC plant biology**, v. 22, n. 1, p. 72, 2022.

YAN, J. *et al.* Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. **PloS one**, v. 4, n. 12, p. e8451, 2009.

ZHANG, F.; WAN, X.; PAN, G.. QTL mapping of *Fusarium moniliforme* ear rot resistance in maize. 1. Map construction with microsatellite and AFLP markers. **Journal of Applied Genetics**, v. 47, p. 9-15, 2006.

CAPÍTULO 2 – DIVERSIDADE E ESTRUTURA GENÉTICA ENTRE ACESSOS CLASSIFICADOS EM RAÇAS COMERCIAIS E INDÍGENAS DO BANCO DE GERMOPLASMA DE MILHO

RESUMO

O Banco Ativo de Germoplasma de Milho (BAG Milho) conta com mais de 4.000 acessos preservados em sua coleção de base. A partir de informações fenotípicas e de origem geográfica foi criada a Coleção Nuclear de Milho, uma amostra representativa da coleção de base, com máxima representatividade de variabilidade genética e mínima redundância. Adicionalmente, foi montada uma coleção de acessos classificados como raças indígenas e/ou originários de terras indígenas. Embora mais de 80% dos acessos do BAG tenham sido caracterizados de acordo com descritores morfológicos da cultura do milho, até o momento, pouco se sabe a respeito da diversidade molecular ampla no genoma, dentro e entre os acessos destas coleções. O conhecimento da diversidade, ancestralidade comum e estrutura genética em nível molecular de uma coleção de germoplasma é uma informação importante para validar e complementar as informações fenotípicas e de origem no desenho de estratégias de conservação e uso de recursos genéticos e melhoramento. Com base em um método de preparação de amostras em bulk otimizado anteriormente, neste trabalho pools de DNA de 24 plantas de cada um dos 293 acessos da Coleção Nuclear Brasileira de milho e 188 acessos da coleção de milho indígena foram alelotipados com 3.526 SNPs no *EMBRAPA 65kMultiespécies* Illumina chip. Os acessos estudados provêm de uma ampla distribuição geográfica, contêm grande variabilidade morfológica e incluem representantes das raças brasileiras, comerciais e indígenas de milho. Análises de distância genética e estrutura foram realizadas para os 481 acessos juntamente com dados dos mesmos SNPs alelotipados obtidos da literatura para acessos de milho e Teosinto conservados pelo centro internacional CIMMYT (*Centro Internacional de Mejoramiento de Maíz y Trigo*) representativos de germoplasma tropical, subtropical e temperado. As coleções nuclear e indígena apresentaram diferenciação genética moderada, com um F_{ST} médio de 0,136 revelando um padrão de estrutura populacional no qual acessos da coleção nuclear se sobrepuseram a acessos de origem indígena e germoplasma CIMMYT. Entretanto, a coleção indígena apresentou maior diferenciação genética entre os acessos ($F_{ST} = 0,140$) em comparação à coleção nuclear ($F_{ST} = 0,109$), o que sugere a necessidade de uma atualização da composição da coleção nuclear incluindo mais acessos indígenas e assim ampliar a diversidade genética representada no germoplasma disponível da espécie. O agrupamento por k-means detectou sete grupos, que continham proporções variáveis de acessos da coleção nuclear e indígena. Padrões de estrutura genética baseados na distância genética, conforme definido pelos dados de SNPs, foram encontrados com sobreposição parcial com classificações dos acessos de germoplasma usados no estabelecimento das coleções, tais como tipos de grãos de milho, nível de domesticação e origem do material coletado. No entanto, algumas discrepâncias importantes também foram detectadas onde tais classificações contrastam com os agrupamentos indicados pelos dados de DNA. Esses resultados indicam que informações genômicas não enviesadas derivadas dos dados do SNP fornecem informações adicionais importantes que refletem de forma mais precisa a origem ancestral dos acessos de germoplasma. Propomos que os dados genômicos gerados neste trabalho sejam prontamente integrados aos dados de passaporte disponíveis dos acessos estudados para chegar a uma representação mais precisa do germoplasma de milho conservado nas duas coleções.

Palavras-chave: *Zea mays* L.; Raças Indígenas; Germoplasma; Alelotipagem; Diversidade Genética; SNPs.

ABSTRACT

The Brazilian Maize Germplasm Bank (Maize BAG) has over 4,000 accessions preserved in its base collection. This BAG also has the Brazilian Maize Core Collection, a representative sample of the base collection, and a collection of accessions classified as indigenous landraces and/or originating from indigenous lands. Although over 80% of the BAG accessions have been characterized according to morphological descriptors of the maize plant, to date, little is known about the genome-wide molecular diversity effectively preserved in these collections. Knowledge of the diversity, common ancestry and genetic structure of a germplasm collection at the molecular level is important information to validate and complement phenotypic and origin information in the design of conservation and use strategies for genetic resources and targeted improvement. Using a previously optimized method for preparation of DNA bulks, in this work DNA pools of 24 plants for 293 accessions of the Brazilian Maize Core Collection and 188 accessions of the indigenous maize collection were allelotyped with a set of 3,526 SNPs on the *EMBRAPA 65kMultispecies* Illumina chip. Genetic distance and structure analyses were performed for the 481 accessions together with data from the same allelotyped SNPs obtained from the literature for maize and Teosinte accessions conserved by the international center CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) representative of tropical, subtropical and temperate germplasm. The core and indigenous collections showed moderate genetic differentiation, with an average F_{ST} of 0.136 revealing a pattern of population structure in which accessions from the core collection overlapped with accessions of indigenous origin and CIMMYT germplasm. However, the indigenous collection showed greater genetic differentiation between accessions ($F_{ST} = 0.140$) compared to the nuclear collection ($F_{ST} = 0.109$), which suggests the need to update the composition of the nuclear collection including indigenous accessions and thus expand the genetic diversity represented in the available germplasm of the species. Clustering by k-means detected seven groups, that contained variable proportions of core and indigenous accessions. Patterns of genetic structure based on genetic distance, as defined by SNP data, were found to partially overlap with classifications of the germplasm accessions used in the establishment of the collections, such as maize grain types, domestication level and origin of the collected material. However, some important discrepancies were also detected where such classifications contrasted with the grouping revealed by the DNA data. These results indicate that the unbiased genomic information derived from SNP data provides important additional and alternative information that more precisely reflects the ancestral origin of the germplasm accessions. We propose that the genomic data generated in this work be promptly integrated with the available passport data of the studied accessions to arrive at a more accurate representation of the maize germplasm conserved in the two collections.

Keywords: *Zea mays* L.; Indigenous Landraces; Germplasm; Allelotyping; Genetic Diversity; SNPs.

1. INTRODUÇÃO

O milho (*Zea mays* L.) se tornou, ao longo das últimas décadas, a maior cultura agrícola do mundo, sendo a cultura agrícola para alimentação humana e animal e uso industrial mais amplamente plantada no mundo (YAN *et al.*, 2009; CONTINI *et al.*, 2019; XU *et al.*, 2022). Além da extrema importância mundial em termos de produção, a cultura do milho tem grande valor econômico caracterizado pelos diversos usos, com seus produtos e subprodutos que integram toda uma cadeia alimentícia (DEMÉTRIO *et al.*, 2008; CONTINI *et al.*, 2019; FREITAS, 2019). São estimadas mais de 3.500 utilizações desse cereal, que vai desde a alimentação animal, destinação da maior parte da produção desse grão, até a indústria de alta tecnologia. A cultura do milho possui alta relevância na segurança alimentar, animal e humana, além de fazer parte da produção de diversos produtos como combustíveis, bebidas, polímeros e outros (DEMÉTRIO *et al.*, 2008; CONTINI *et al.*, 2019).

Os recursos genéticos (RG) de milho constituem, portanto, um tesouro imensurável para a humanidade. As estratégias de conservação do germoplasma do milho e o conhecimento sobre a sua variação e utilização são um esforço de grande importância não apenas para a segurança alimentar, redução da fome e da pobreza, mas também para o fornecimento de uma base sólida de conhecimentos para as futuras gerações de investigadores em diferentes áreas da genética, melhoramento e biotecnologia (WALTERS, 2004; CGIAR, 2007; PRIYANKA *et al.*, 2021). Estima-se que haja mais de 135.000 acessos de milho conservados em bancos de germoplasma em todo o mundo. Entretanto, esse recurso é, ainda, em grande parte inexplorado com documentações inconsistentes e, por vezes, incompletas e caracterizações fragmentadas, dificultando a integração dos diferentes bancos de germoplasma, sua utilização eficiente no melhoramento e um intercâmbio mais ativo. É necessário que os RGs existentes deixem de ser apenas um repositório estático de diversidade e passem a contribuir com atividades cientificamente importantes e enriquecedoras (MCCOUCH *et al.*, 2012; MCCOUCH *et al.*, 2020).

O conhecimento da diversidade e estrutura genética de uma coleção de germoplasma é um componente essencial para a concepção de estratégias de conservação de uma espécie, e a potencial utilização dos recursos genéticos no melhoramento (VIGOUROUX, *et al.*, 2008; SEMAGN *et al.*, 2012). A avaliação da extensão da diversidade genética e entendimento de possíveis padrões de estrutura em um banco de germoplasma são ações úteis para a otimização de diversas iniciativas. Permitem a quantificação dos níveis de variabilidade genética existente

entre e dentro possíveis grupos de germoplasma estabelecidos em função de características morfofisiológicas específicas; facilitam a identificação de lacunas, exclusividades ou redundâncias genéticas; permitem a comparação de coleções da mesma cultura em bancos de germoplasma distintos e a detecção de erros ou confusões na identidade genética que podem ter ocorrido em uma coleção de germoplasma (ANGLIN *et al.* 2025). Além disso, o conhecimento dos padrões de diversidade genética de um banco de germoplasma facilitam o acesso aos recursos genéticos; auxiliam no direcionamento da montagem de coleções nucleares (ABADIE *et al.*, 2000); e auxiliam na escolha de quais materiais priorizar para a caracterização fenotípica ou para resgate de variedades tradicionais (TEIXEIRA *et al.*, 2021). Finalmente, quando integrados em programas de melhoramento, os conhecimentos mais detalhados de bancos de germoplasma auxiliam no direcionamento dos primeiros passos do desenvolvimento de populações segregantes e linhagens endogâmicas para o desenvolvimento de cultivares de alto rendimento adaptadas a diferentes ambientes, atendendo à crescente demanda por alimentos em todo o mundo (SILVA *et al.*, 2020).

A coleção de milho do Brasil, uma das maiores a nível mundial, é mantida no Banco Ativo de Germoplasma de Milho (BAG Milho) do Centro Nacional de Pesquisa de Milho e Sorgo (CNPMS-EMBRAPA) e conta com mais de 4.000 acessos conservados em sua coleção base (TEIXEIRA *et al.*, 2021). O BAG de Milho conta ainda com a Coleção Nuclear de Germoplasma de Milho, uma amostra representativa da coleção base, conservada atualmente no Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia (CENARGEN-EMBRAPA) e estabelecida para manter a máxima variabilidade genética do BAG com o mínimo de redundância (ABADIE *et al.*, 2000). Conta, também, com uma coleção de acessos classificados em raças indígenas e/ou originários de terras indígenas que compõem a coleção indígena desse BAG (TEIXEIRA *et al.*, 2021).

Apesar de mais de 80% dos acessos do BAG terem sido caracterizados de acordo com descritores morfológicos da cultura do milho (TEIXEIRA; COSTA, 2010; TEIXEIRA *et al.*, 2019), até o momento, pouco se sabe sobre a diversidade genética molecular efetivamente conservada nessas coleções. Os escassos trabalhos de caracterização da estrutura e diversidade genética desse BAG utilizando marcadores moleculares se restringem a poucos acessos de subgrupos das coleções, além de terem sido conduzidos com marcadores com baixa ou nenhuma transferibilidade, o que impossibilita comparações entre diferentes estudos e muito menos com bases de dados internacionais (KUMAR, B., *et al.* 2022). A coleção indígena, a título de exemplo, carece de qualquer estudo genômico para melhor esclarecimento sobre sua

origem, estrutura, diversidade genética e domesticação. Além disso, pouco ou nada se sabe sobre a sua divergência em relação às cultivares brasileiras comerciais antigas ou recentes.

No capítulo anterior, foi utilizado com sucesso um protocolo envolvendo a preparação de bulks de DNA de vários indivíduos a partir de uma extração de tecido foliar agrupado, seguido da alelotipagem de SNPs para a estimativa de frequências alélicas para 3.526 SNPs de milho. Esse protocolo permitiu estimar com elevada acurácia as frequências alélicas para os SNPs alelotipados reduzindo custos e tempo de trabalho. Neste capítulo, este protocolo foi aplicado em maior escala para investigar a estrutura e diversidade genética de 293 acessos da coleção nuclear do Banco Ativo de Germoplasma de Milho da EMBRAPA e 188 acessos de variedades indígenas. Os dados genéticos dos SNPs foram gerados com o *EMBRAPA 65kMultispecies chip*. Estes 3.526 SNPs foram selecionados a partir de milhares de SNPs validados e utilizados internacionalmente para o milho com o objetivo de permitir a imediata portabilidade e integração de dados gerados com dados públicos da literatura. Esta característica particular é praticamente restrita a SNPs genotipados em microarranjos em função da qualidade dos dados gerados nestas plataformas. Com base nesta portabilidade e interoperabilidade de SNPs, dados genéticos publicados para acessos de bancos de germoplasma internacionais (XU *et al.*, 2022) puderam ser incluídos neste estudo.

Análises de distância genética, agrupamento e estrutura foram realizadas com base nos dados de SNPs para os 481 acessos das coleções brasileiras juntamente com dados dos mesmos SNPs obtidos da literatura para acessos de milho e Teosinto conservados em bancos de germoplasma internacionais. Com base nos dados genéticos levantados, juntamente com classificações morfológicas e geográficas utilizadas na caracterização dos acessos estudados, perguntas específicas foram formuladas e tentativamente respondidas sobre algumas questões relevantes a respeito do germoplasma conservado no Brasil.

2. MATERIAIS E MÉTODOS

2.1 LEVANTAMENTO DE DADOS DE REPRESENTATIVIDADE DE COLEÇÕES DE GERMOPLASMA BRASILEIRAS

Informações da coleção nuclear do BAG de Milho foram obtidas a partir do Boletim de Pesquisa e Desenvolvimento da EMBRAPA publicado por Abadie *et al.* (2000) e averiguadas através do portal da Plataforma Alelo Recursos Genéticos que abriga a base de dados atualizada

das coleções de germoplasma da EMBRAPA. Na descrição da coleção nuclear, os acessos são classificados quanto a diferentes características: origem (coleta, materiais melhorados e introduções); tipo de grão (amiláceo ou farináceo, semi-amiláceo, dentado, semi-dentado, duro, semi-duro pipoca, doce, opaco II, tunicata e ceroso); região ecogeográfica de origem (Sul, Cerrados, Cerrados – Norte, Amazônia, Caatinga e Agreste – litoral). Os materiais melhorados foram ainda classificados em: a) pipoca, b) não-pipoca do Centro Nacional de Pesquisa de Milho e Sorgo da EMBRAPA (CNPMS) e c) não-pipoca de outros programas de melhoramento. As introduções foram classificadas quanto à sua origem: tropical ou temperada. Foi possível ainda obter informações quanto à raça dos acessos e sua localização de origem quando disponíveis na Plataforma Alelo.

Informações sobre os acessos de origem indígena foram obtidas a partir de Teixeira *et al.* (2021). Os acessos dessa coleção foram classificados em raças (Caingang, Entrelaçado, Lenha, Moroti e Guarani) e também quanto ao tipo de grão (dentado, semi-dentado, duro, semi-duro, doce, farináceo, opaco e pipoca), ao tipo de arranjo dos grãos (reto, espiral e entrelaçado) e, por fim, à cor dos grãos (branco, amarelo, alaranjado, vermelho, cinza, preto, e púrpura). Foram ainda obtidas informações sobre localização da coleta dos acessos quando disponível no documento. A partir das informações obtidas, foram estimados os percentuais de acessos enquadrados nas diferentes categorias de cada classificação morfológica, ecogeográfica etc. para entender a representatividade da coleção para cada uma das classificações utilizadas. Estas estimativas foram obtidas separadamente para a coleção nuclear e a coleção indígena.

2.2 MATERIAL VEGETAL E PREPARAÇÃO DE AMOSTRAS DE DNA EM BULK

Amostras de 48 sementes de cada um de 293 acessos da coleção nuclear do Banco Ativo de Germoplasma de Milho da EMBRAPA e 188 acessos de variedades indígenas foram germinadas em tubetes próprios em estufa, visando a obtenção de 22 a 24 plantas a serem utilizadas na montagem dos bulks de amostras dos acessos para a alelotipagem de SNPs. As amostras de tecido foliar foram coletadas após aproximadamente 30 a 40 dias e desidratadas em sílica-gel para posterior utilização na extração de DNA genômico total.

Foram montados bulks com discos de folhas de igual tamanho, conforme descrito no capítulo anterior, seguindo a metodologia em Dubreuil *et al.* (1999) e Arca *et al.* (2020). Múltiplos discos foram coletados de cada uma das 22 a 24 plântulas e pesados de forma que a contribuição de tecido de cada planta fosse equivalente, com precisão de 0,1 gramas seguida da maceração dos tecidos de forma conjunta e extração de uma amostra única de DNA em bulk.

Um total de 293 acessos da coleção nuclear e 188 acessos de variedades indígenas foram estudados. Réplicas biológicas de 102 acessos foram também preparadas para avaliação da reprodutibilidade das estimativas de frequências alélicas por alelotipagem. As réplicas foram mais adequadamente analisadas, entretanto, no âmbito do capítulo anterior de avaliação do protocolo geral de alelotipagem.

As extrações foram realizadas a partir de tecido de folhas de plântulas germinadas utilizando o protocolo otimizado desenvolvido no Laboratório de Genética Vegetal (LGV) da EMBRAPA Recursos Genéticos e Biotecnologia (CENARGEN), baseado na utilização de tampão de lise contendo o detergente CTAB (brometo de cetiltrimetilamônio) com alto teor de sal (INGLIS *et al.*, 2018).

As extrações realizadas baseadas em tampão CTAB consistem em duas etapas de lise: mecânica, em que as amostras são submetidas à agitação em macerador automático, e logo após, química, com tampão de lise contendo CTAB. Esse método consiste ainda nas etapas de separação dos componentes celulares utilizando-se CIA (clorofórmio: álcool isoamílico, 24:1), precipitação do DNA por meio da adição de isopropanol, e por último, purificação do DNA com o uso de etanol a 70% para remover resíduos. Os passos do protocolo estão descritos em maior detalhe no Anexo 1.

2.3 ALELOTIPAGEM DE SNPs

As 583 amostras de DNA em bulk foram alelotipadas para 3.526 SNPs presentes no *EMBRAPA 65kMultispecies* chip. Esses 3.526 SNPs foram previamente selecionados a partir do Illumina® MaizeSNP50 BeadChip (GANAL *et al.* 2011), no design do chip com base no conteúdo informativo visando ainda cobrir todos os 10 cromossomos do milho de forma o mais equidistante possível de forma a proporcionar uma cobertura genômica adequada para estudos de diversidade e estrutura genética. As amostras de DNA foram enviadas para genotipagem de forma terceirizada para a Neogen do Brasil (SP).

2.4 ESTIMATIVA DAS FREQUÊNCIAS ALÉLICAS

As frequências alélicas das amostras de DNA em bulk foram estimadas pela estimativa normalizada do *B Allele Frequency* (BAF) (para o alelo B) e 1-B (para o alelo A) para cada um dos 3.526 SNPs. A BAF foi estimada por meio do módulo específico do software

GENOMESTUDIO (Illumina Inc., San Diego, CA, USA) utilizado tipicamente para a análise de CNVs (variantes de número de cópias). BAF é uma medida normalizada da razão de intensidade alélica dos dois alelos SNP (A e B). BAF próximo a 1 ou 0 indica a ausência completa de um dos alelos (AA ou BB), e valores intermediários de BAF correspondem às frequências alélicas dos dois alelos ao SNP.

2.5 OBTENÇÃO DE DADOS GENOTÍPICOS DE MILHO E TEOSINTO DA LITERATURA

Dados para os mesmos SNPs alelotipados foram obtidos da literatura para 1.172 acessos de milho representativos de germoplasma mundial e Teosinto conservados pelo centro internacional CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) (N=691), China (N=221), Estados Unidos (N= 66), Tailândia (N=3) e Peru (N=1), publicados por Xu *et al.* (2022). Dados de 1.172 acessos incluindo 982 linhagens representativas do germoplasma global tropical, subtropical e temperado e 190 acessos de teosinto gerados com o chip MaizeSNP50 BeadChip (Illumina Inc., San Diego, CA, USA) contendo 56.110 SNPs foram obtidos no sítio do “European Variation Archive” em <https://wwwdev.ebi.ac.uk/eva/?eva-study=PRJEB41335>. Os dados obtidos em formato “top strand” foram convertidos para o formato AB seguindo a regra de conversão (alelos A e T convertidos para o alelo código A; alelos G e C convertidos para o alelo código B) e em seguida filtrados de forma a extrair somente os dados dos 3.526 SNPs em comum com o *EMBRAPA 65kMultispecies* chip. Utilizando a anotação fornecida sobre a origem e classificação do germoplasma genotipado no artigo de Xu et al. (2022), foram geradas seis pseudo-populações para as 982 linhagens de milho com base na classificação de adaptação climática: subtropical, temperado, temperado_subtropical, tropical, tropical_subtropical e tropical_temperado. Para os 190 acessos de teosinto, sete grupos a seguir foram gerados com base na origem geográfica: Central_Balsas, Central_Plateau, Chalco, Eastern_Balsas, Jalisco, Southern_Guerrero e *Z. nicaraguensis*. Estes acessos incluíam um acesso de *Z. luxurians*, três de *Z. diploperennis*, dois de *Z. perennis*, um de *Z. huehuetenangsis*, 96 das subespécies *Zea mays mexicana*, 75 de *Zea mays parviglumis* e 12 acessos das espécie *Z. nicaraguensis*. Em seguida foram calculadas as frequências alélicas para os 3.526 SNPs para cada um destes grupos visando a sua análise conjunta com os dados dos acessos do BAG Milho do Brasil.

2.6 ANÁLISE DE DADOS

2.6.1 Cálculo de matrizes de distância genética entre acessos

Com base nas frequências alélicas dos 481 acessos dos BAG Milho e 13 grupos de linhagens e teosinto totalizando assim 494 acessos estudados foram estimadas as distâncias de Nei (NEI) (NEI, 1972; TAKEZAKI; NEI, 1996), euclidiana (EUC) e estimado o índice de fixação de Wright (F_{ST}) (WRIGHT, 1978) para todas as comparações dois a dois. Uma vez que o índice de fixação é uma medida formal de diferenciação genética entre populações, a matriz de estimativas de F_{ST} foi considerada equivalente a uma matriz de distâncias genética entre os acessos para fins das análises subsequentes.

2.6.2 Agrupamento de acessos com base em clusterização k-means

De posse das matrizes de distância, o método de agrupamento k-means (JAIN, 2010) foi implementado utilizando o pacote nls2 do software R (R Core Team, 2020). O agrupamento k-means foi utilizado para organizar os acessos em grupos com base em sua similaridade genética dada pelo índice de fixação. O algoritmo funciona primeiro escolhendo aleatoriamente alguns pontos de dados (acessos) centrais chamados centróides e cada acesso é então atribuído ao centróide mais próximo, formando um cluster. Depois que todos os acessos são atribuídos a um cluster, os centróides são atualizados encontrando a posição média dos pontos em cada cluster. Esse processo se repete até que os centróides parem de mudar, formando clusters. O objetivo do agrupamento é dividir os pontos de dados em clusters para que pontos de dados semelhantes pertençam ao mesmo grupo. A escolha ótima de grupos foi obtida por um método aproximativo. Esse método foi dado pela relação entre a soma de quadrados entre os acessos sobre a soma total de quadrados para cada número de acessos variando de 2 a 100 com posterior ajuste de uma função logística. O ponto ótimo do número de grupos é então obtido utilizando a segunda derivada, em que seu ponto de mínimo indica a estabilização da relação de soma de quadrados entre todos os grupos.

2.6.3 Análise de estrutura genética: reconstituição de indivíduos a partir das frequências alélicas

A partir dos dados de frequências alélicas de cada acesso foram “reconstituídos” genótipos multiloco de SNPs para cada um de 50 indivíduos diploides por acesso, visando assim

executar análises genéticas de estrutura populacional utilizando o software STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000; FALUSH; STEPHENS; PRITCHARD, 2003). Para isso, a partir dos 3.526 SNPs alelotipados foram inicialmente selecionados SNPs a cada ~ 4 Mb de acordo com sua posição física no genoma, resultando em um painel de 504 SNPs a princípio independentes. A independência foi utilizada de forma que a reconstituição de genótipos pudesse ser feita assumindo a ausência de ligação genética ou equilíbrio de ligação gamética. Para estes 504 SNPs foram então reconstituídos genótipos multiloco para 50 indivíduos diploides para cada um dos 494 acessos, totalizando assim 24.700 indivíduos. Para avaliar a acurácia da representatividade da população de indivíduos reconstituídos para cada acessos, frequências alélicas foram estimadas a partir dos 50 indivíduos gerados e estas frequências comparadas com as frequências originalmente utilizadas na reconstituição. O RMSE estimado para todas as 248.976 comparações (504 SNPs x 494 acessos) foi sempre menor do que $10E-6$ indicando que os 24.700 indivíduos reconstituídos representavam com elevada acurácia os acessos estudados. Estes 24.700 indivíduos foram em seguida submetidos a uma análise de estrutura genética com o software STRUCTURE. Foram utilizadas 10000 iterações com burn-in de 1000 considerando um modelo de mistura testando até 10 grupos com quatro repetições cada e, posteriormente, definidos o número de grupos pelo método do Evanno (EVANNO; REGNAUT; GOUDET, 2005) com o pacote pophelper (FRANCIS, 2017) no software R. Além disso, para cada acesso foi obtida a média dos 50 indivíduos da proporção genômica alocada a cada um dos grupos assinalados pelo STRUCTURE para fins de visualização das proporções ancestrais de cada grupo genético detectado na análise para cada acesso. Para fins de potencial comparação, o agrupamento k-means também foi executado utilizando exclusivamente os 504 SNPs independentes selecionados para a análise de estrutura. Entretanto, os resultados foram equivalentes àqueles obtidos com todos os 3.526 SNPs marcadores e, portanto, omitidos deste trabalho.

2.6.4 Compartimentalização das análises de agrupamento e estrutura de acordo com a origem dos acessos

Ambas as análises de agrupamento k-means e de estrutura foram feitos separadamente para: (i) a coleção nuclear doravante identificada pelo acrônimo BRA; (ii) a coleção de acessos indígenas (IND); (iii) coleções nuclear e indígena conjuntamente (BRA_IND); (iv) coleção nuclear com os dados de acessos internacionais e teosinto (BRA_XU); e (v) coleções nuclear, indígena e acessos internacionais e teosinto (BRA_IND_XU). Uma análise mais detalhada do

agrupamento do conjunto de acessos BRA_IND_XU foi feita no software STRUCTURE de forma a identificar a alocação de acessos das coleções nuclear, indígena e acessos internacionais e teosinto a cada grupo detectado pelo software STRUCTURE. A automação dessas análises e análise de componentes principais para visualização da estrutura populacional foram executadas no software R e os gráficos construídos no pacote ggplot2 (WICKHAM, 2011).

2.7 COMPARAÇÃO DAS CLASSIFICAÇÕES MORFOLÓGICAS E GEOGRÁFICAS COM AGRUPAMENTOS DETECTADOS PELOS DADOS GENÉTICOS

Os dados levantados e os percentuais de acessos enquadrados dentro de cada classificação obtidos anteriormente foram estimados dentro de cada grupo genético identificado com base nos marcadores SNPs pelo método de agrupamento k-means. O objetivo desta análise foi avaliar se os agrupamentos identificados pelos marcadores moleculares poderiam coincidir ou explicar parcial ou integralmente as classificações morfológicas, raciais ou geográficas dos acessos. Com base na matriz de valores de diferenciação genética (F_{ST}) entre os acessos tomados dois a dois foram construídos dendrogramas para permitir a fácil visualização de agrupamentos dos acessos. Utilizando um código de cores, os acessos nos dendrogramas foram identificados quanto às diferentes classificações visando assim fornecer uma visualização das possíveis sobreposições dos agrupamentos genéticos identificados pelos marcadores SNPs e as diferentes classificações morfológicas e ecogeográficas utilizadas.

Códigos de cores foram utilizados para assinalar: coleção de origem (BRA, IND ou XU), raça (Indígenas, Comerciais Antigas, Comerciais Recentes e Teosinto), tipo de grão (amiláceo ou farináceo, semi-amiláceo, dentado, semi-dentado, duro, semi-duro pipoca, doce, opaco II, tunicata e ceroso) e região geográfica de origem (Norte, Nordeste, Centro-Oeste e Sul e Sudeste). Um dendrograma apenas da coleção indígena também foi construído e sobre ele os acessos foram identificados quanto ao seu pertencimento às diferentes raças (Caingang, Entrelaçado, Lenha, Moroti e Guarani). A construção dos dendrogramas baseados em F_{ST} foi feita com o pacote ggtree no software R utilizando o método UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*).

3. RESULTADOS

3.1 REPRESENTATIVIDADE DA COLEÇÃO NUCLEAR QUANTO ÀS DIFERENTES CLASSIFICAÇÕES

A coleção nuclear é composta em sua maioria (79,5%) por acessos originados através de coleta (Figura 1 – A). Dos materiais melhorados (11,3% da coleção), apenas 12,1% são classificados no grupo pipoca, 39,4% são não-pipoca resultantes do programa de melhoramento do CNPMS, enquanto 48,5% são não-pipoca advindos de outros programas de melhoramento (Figura 1 – B). A coleção nuclear conta ainda com 9,2% de materiais introduzidos, dos quais 63% de origem tropical e 37% de origem temperada (Figura 1 – C).

As raças comerciais recentes (Dente, Semi-dentado e Cravo) aparecem com mais frequência (37,9%) na coleção nuclear, sendo desses, 73% de acessos da raça Dente, 26% da raça Semi-dentado e 1% da raça Cravo. A classificação racial (Figura 1 – D) ainda é bastante incipiente, uma vez que 48,8% dos acessos não possuem informações disponíveis sobre a raça (NA, do inglês: *not available*). Dos onze tipos de grãos mencionados por Abadie *et al.* (2000), oito aparecem na coleção nuclear (Figura 1 – E) com o tipo semi-dentado sendo o mais frequente (28%). Mais de 20% dos acessos não têm informações disponíveis sobre o tipo de grão. Vale mencionar que dos 60 acessos sem classificação quanto ao grão, 27 são introduções, portanto todos os acessos desse grupo na coleção, 29 são materiais melhorados e apenas 4 são originados de coleta, evidenciando a lacuna de informação para os acessos não originários de coleta, uma vez que apenas quatro acessos dos materiais melhorados foram classificados como do tipo de grão pipoca.

Os acessos provenientes de coleta foram ainda classificados quanto à sua região geográfica (Figura 1 – F) e região ecogeográfica (Figura 1 – G) brasileira de origem. A região Norte foi a região geográfica que mais contribuiu acessos (33,9%) enquanto a região ecogeográfica com maior contribuição de acessos foi o Cerrado (22,7%), os quais somados aos acessos dos Cerrados da região Norte, ultrapassam mais de 35% dos acessos na coleção. A coleção nuclear possui 22 acessos classificados como sendo de origem indígena, os quais também fazem parte da coleção indígena, mantidos separadamente em cada uma das coleções.

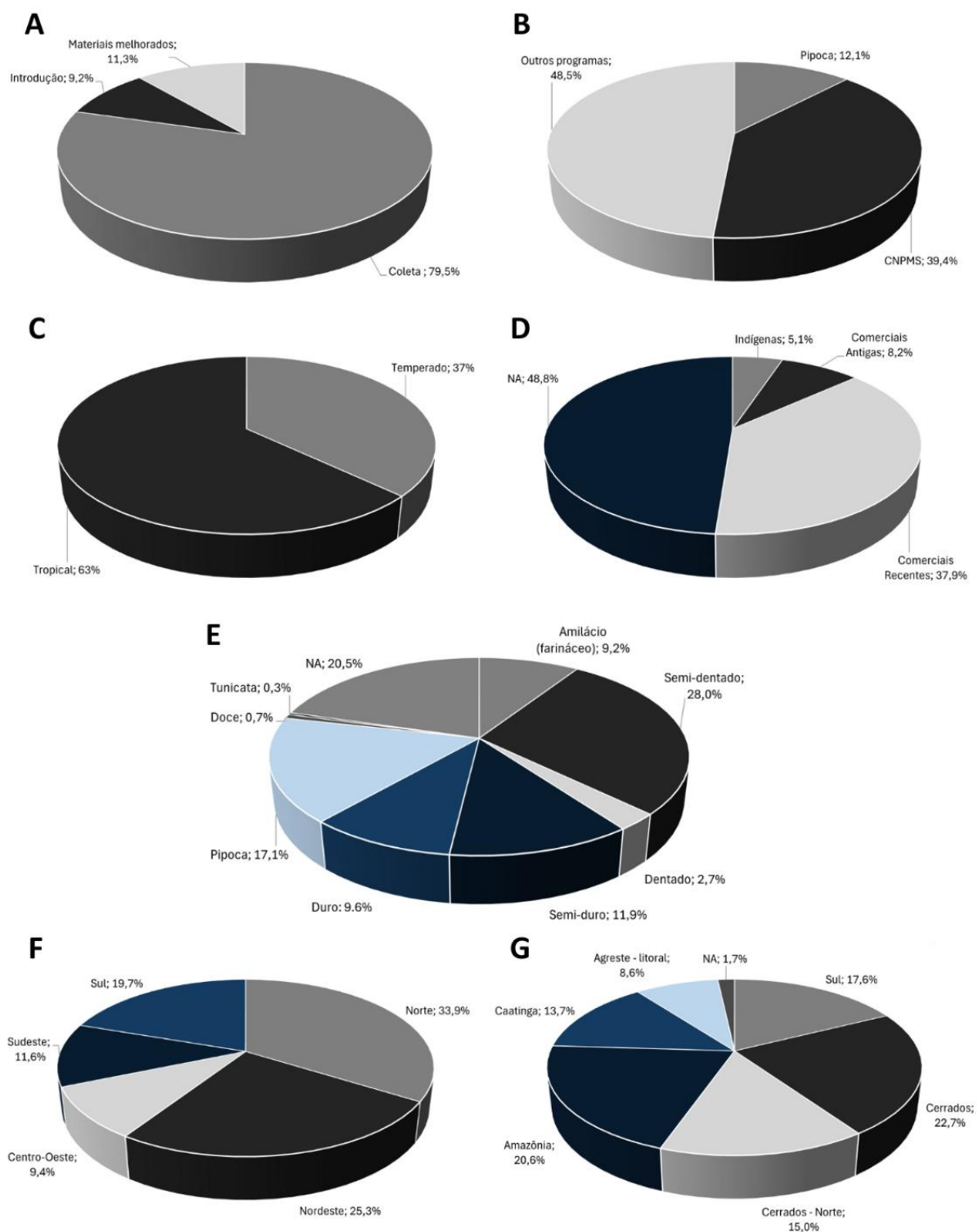


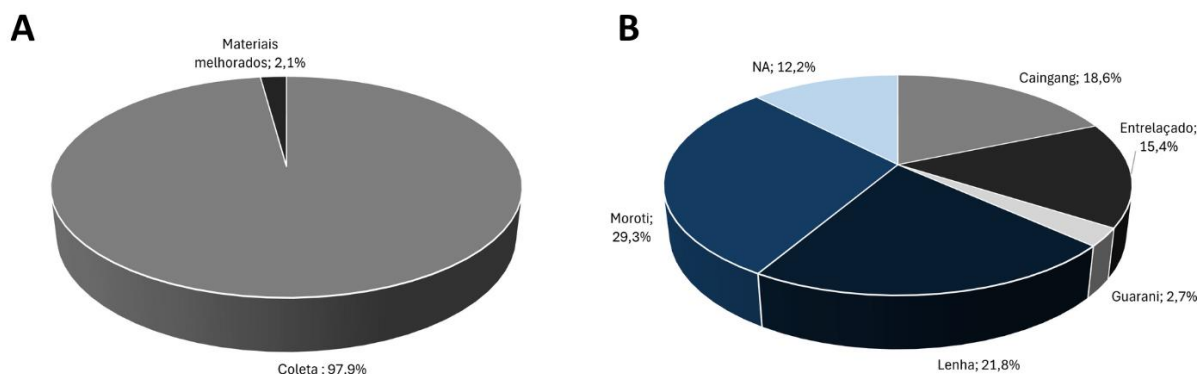
Figura 1 – Composição e classificação dos acessos da coleção nuclear. A. Tipo de origem, B. Classificação dos materiais melhorados, C. Origem dos materiais introduzidos, D. Raça, E. Tipo de grão, F. Região geográfica, G. Região ecogeográfica. NA = não disponível (*not available*).

3.2 ANÁLISE DA COLEÇÃO DE ACESSOS DE ORIGEM INDÍGENA

A coleção de acessos de origem indígena é composta majoritariamente (98%) de acessos provenientes de coleta (Figura 2 – A). Apenas quatro acessos são classificados como sendo materiais melhorados (Brasil 9595, Caingang Composto, Composto Guarani e Composto Indígena). Os quatro acessos fazem parte de programas de melhoramento que não do CNPMS (3 acessos advindos da USP-ESALQ e 1 da CENARGEN). Com exceção da raça Guarani, os acessos são relativamente bem distribuídos entre as diferentes raças indígenas: Guarani – 2,7%, Entrelaçado – 15,4%, Caingang – 18,6%, Lenha – 21,8% e Moroti – 29,3%. Há ainda na coleção 12,3% de acessos sem classificação racial disponível, inseridos no grupo “outras raças ou raça não identificada” (Figura 2 – B).

O tipo de grão que aparece mais comumente dentro dessa coleção é do tipo amiláceo ou farináceo, representando 45,2% dos acessos, entretanto, 33,5% dos acessos ainda carecem de informações quanto a essa classificação (Figura 2 – C). Na coleção indígena, os acessos são também classificados quanto ao tipo de arranjo e a cor dos grãos. O arranjo dos grãos do tipo reto aparece com mais frequência, 51,1% (Figura 2 – D), enquanto grãos brancos representam a coloração mais frequente dentro da coleção, seguida pelos grãos coloridos (17,6%) (Figura 2 – E).

A região Sul foi a região geográfica que mais contribuiu acessos (36,6%) resultantes de coleta, seguida pela região Centro-Oeste (31,7%), que juntas somam mais do que a metade dos acessos (Figura 3 – D). Não foi possível levantar informações sobre a região ecogeográfica de origem dos acessos, uma vez que essa informação não estava disponível nem mesmo para 5% dos acessos da coleção.



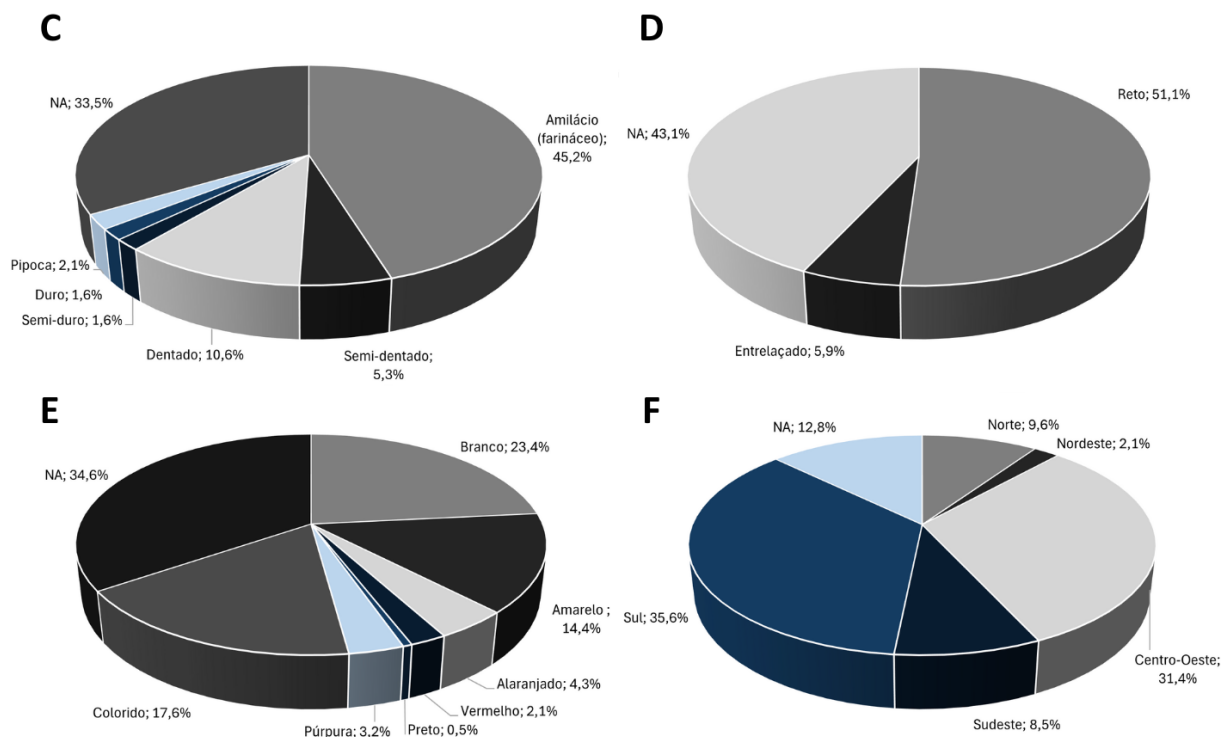


Figura 2 – Composição e classificação dos acessos da coleção indígena. A. Tipo de origem, B. Raça, C. Tipo de grão, D. Arranjo dos grãos. E. Cor dos grãos. F. Região geográfica. NA = não disponível (*not available*).

3.3 ESTRUTURA GENÉTICA DETECTADA PELO AGRUPAMENTO K-MEANS

A análise das diferentes matrizes de distâncias genéticas revelou um padrão de estrutura populacional (Figura 3) em que foi constatada uma sobreposição dos acessos da coleção nuclear com os acessos de origem indígena e os dados da literatura (Figura 3). De maneira geral as diferentes métricas de distância genética utilizados revelaram padrões semelhantes de agrupamento e sobreposição entre as diferentes coleções de germoplasma. Entretanto a distância Euclidiana aparentemente produziu uma dispersão mais ampla dos acessos o que permitiu uma melhor resolução na visualização de sobreposições de acessos individuais.

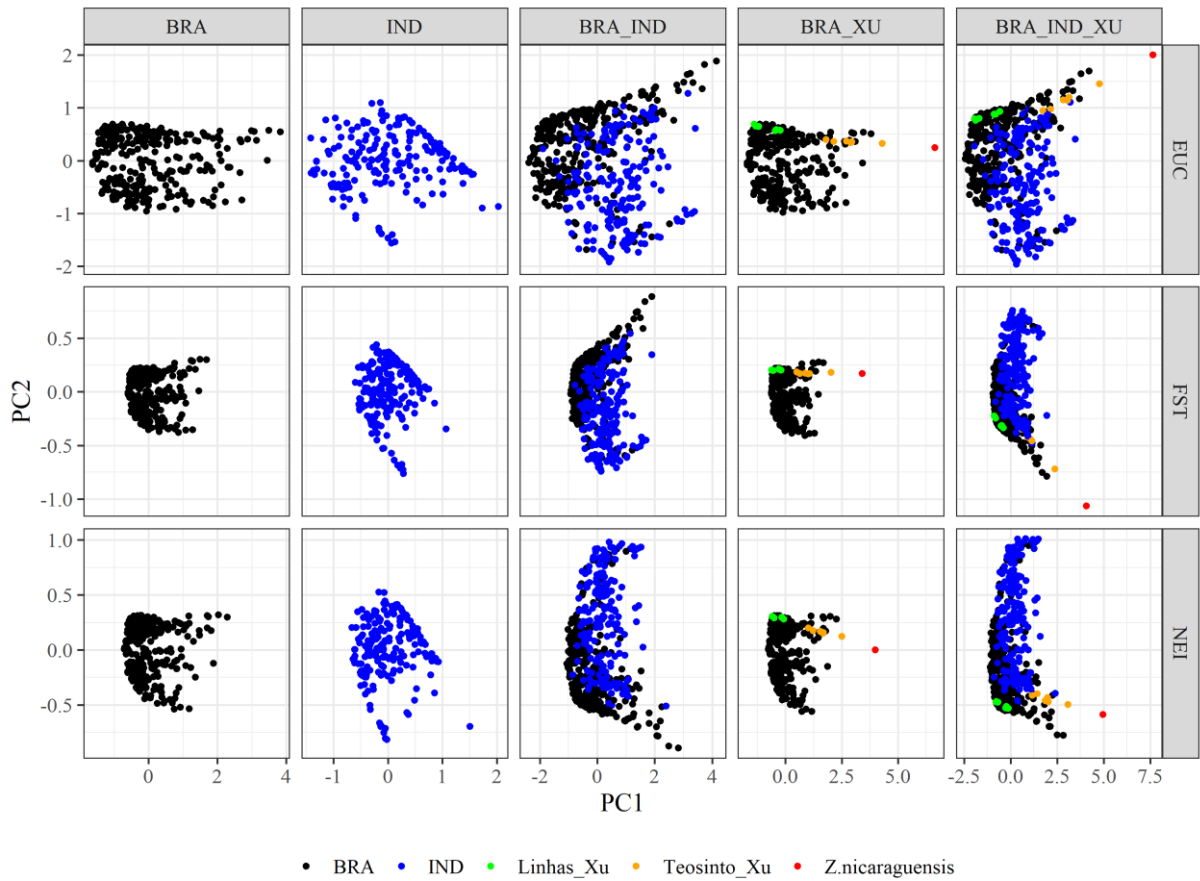


Figura 3 – Estruturas populacionais dos acessos detectadas pela análise de agrupamento k-means com base em diferentes matrizes de distância genética (Euclidiana, F_{ST} e distância de Nei). Análises individualizadas dos acessos da coleção nuclear (BRA), indígenas (IND), conjunta nuclear e indígena (BRA_IND); nuclear e dados de uma coleção de acessos internacionais e teosinto (BRA_XU) e considerando todos os acessos (BRA_IND_XU) para a distância com base em F_{ST} (índice de fixação de Wright) em função dos dois primeiros componentes principais extraídos da respectiva distância.

O agrupamento pelo método k-means utilizando todos os 3.526 SNPs detectou um total de sete grupos (Figura 4).

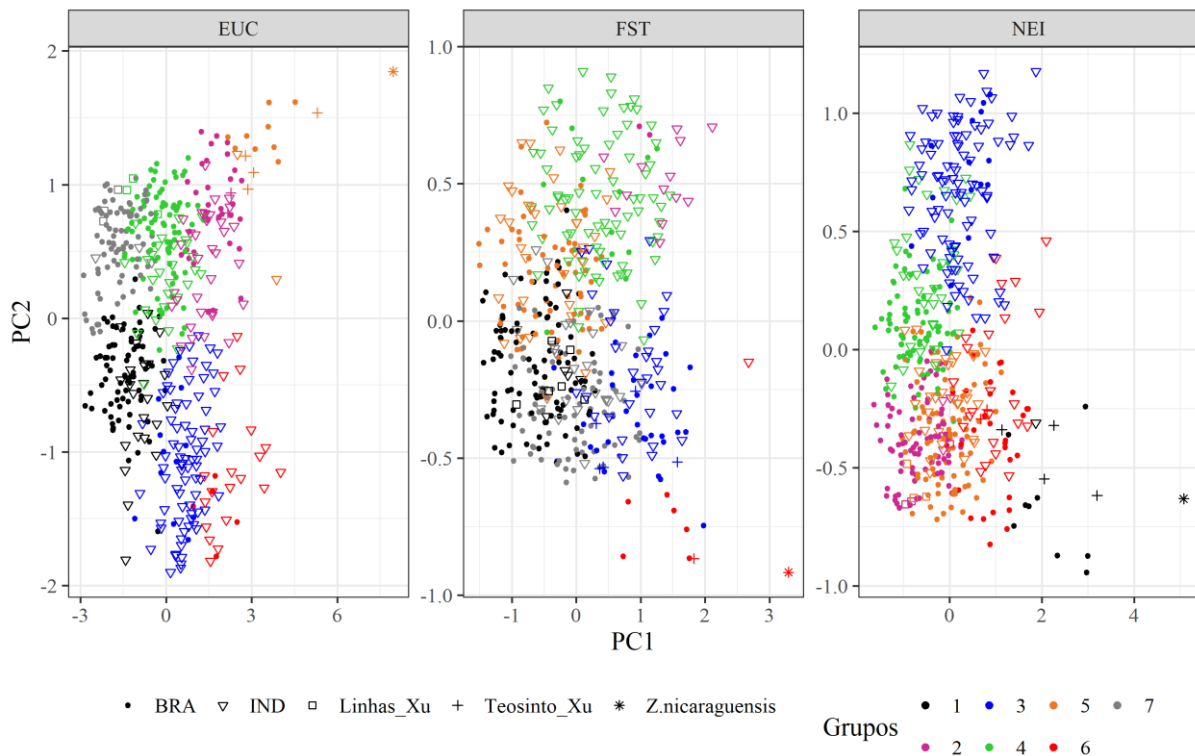


Figura 4 – Diagramas de agrupamento k-means envolvendo todos os acessos (BRA_IND_XU) com base nas diferentes matrizes de distância genética em função dos dois primeiros componentes principais estimados pela respectiva matriz de distância. Os acessos provenientes das diferentes coleções são indicados com diferentes marcadores e os sete grupos detectados pelo agrupamento k-means são indicados por códigos de cores.

Para uma melhor visualização deste resultado foi quantificado o percentual de participação de acessos das diferentes coleções de germoplasma (BRA, IND, Linhas Xu e Teosinto Xu) em cada um dos grupos genéticos detectados pelo k-means exclusivamente com base em F_{ST} já que os resultados com os demais índices de distância foram essencialmente iguais (Tabela 1). Não se observou uma participação exclusiva de nenhuma das coleções nos sete grupos detectados pelos marcadores de DNA. Isto revela que não há uma segregação de alguma das coleções para algum grupo específico, mas sim há uma ampla variabilidade genética dentro de cada coleção. Isto era esperado para a coleção nuclear a qual foi efetivamente pensada e montada com o objetivo de manter ampla variabilidade. Entretanto, não poderia se esperar isso para a coleção indígena. Este resultado demonstra, portanto, que mesmo dentro de uma classificação supostamente indígena, há ampla variabilidade genética com acessos pertencentes a diferentes subgrupos dentro do grupo indígena. Os grupos 1, 5, 6 e 7 envolvem predominantemente acessos da coleção nuclear, enquanto os grupos 2 e 4 envolvem predominantemente acessos indígenas. As linhagens das coleções internacionais participaram exclusivamente do grupo 1 o que é consistente com o fato da grande maioria dos acessos ser

oriundo da coleção nuclear. Já os materiais genéticos de Teosinto, mais distantes a princípio de material genético melhorado, participaram somente dos grupos 3 e 6. Poderia se esperar uma maior participação destes acessos filogeneticamente mais distantes do milho geneticamente melhorado nos grupos 2 e 4 que contam com maior participação de acessos indígenas, mas não foi isso que a análise genômica revelou.

Tabela 1 – Percentual de acessos provenientes das diferentes coleções de germoplasma em cada um dos sete grupos detectados pelo k-means com base na matriz F_{ST} . O acesso *Z. nicaraguensis* foi incluído no grupo de germoplasma de Teosinto para fins de simplificação.

Grupos genéticos pelo k-means	Coleção de germoplasma				Total
	BRA	IND	Linhas Xu	Teosinto Xu	
1	89,91	4,59	5,5	0	100
2	13,33	86,67	0	0	100
3	46,38	46,38	0	7,25	100
4	14,29	85,71	0	0	100
5	66,33	33,67	0	0	100
6	66,67	11,11	0	22,22	100
7	74,76	25,24	0	0	100

Consistente com a análise de agrupamento pelo k-means o dendrograma UPGMA baseado na matriz de F_{ST} (Figura 5) embora tenha apresentado uma separação entre os acessos da coleção nuclear (em vermelho) e indígena (em verde), vários acessos se posicionam de forma dispersa nos dois grupos. É importante lembrar que há 22 acessos indígenas em comum entre as duas coleções, uma vez que há representantes da coleção indígena na coleção nuclear. Acessos iguais entre as duas coleções foram identificados com BRA (coleção nuclear) e IND (coleção indígena) após o nome do acesso, a título de exemplo: MT-009-IND e MT-009-BRA, que aparecem na porção superior do dendrograma. As linhagens da coleção internacional analisadas por Xu *et al.* (2022) mostraram esperada separação entre as linhas tropicais/subtropicais e temperadas/subtemperadas, e também com acessos de teosinto. Apenas o acesso de *Zea nicaraguensis* apareceu claramente separado dos demais acessos da coleção internacional. Os acessos NODZOB-RONRE-TROPE-IND e NODZOB-RONRE-TROPE-BRA não se agrupam juntos como os demais acessos compartilhados entre as coleções, apontando possível erro de identificação do acesso dentro da coleção nuclear, podendo se tratar de um dos demais acessos denominados “Nodzob”, que estão relacionados à etnia Xavante, uma vez que não se trata do mesmo acesso.



Figura 5 – Dendrograma UPGMA construído com base na matriz de base em F_{ST} (índice de fixação de Wright) envolvendo acessos da coleção nuclear, indígena e de bancos de germoplasma internacionais publicados por Xu et al. 2022; BRA = coleção nuclear. IND = coleção indígena. Xu = dados da literatura.

Os sete grupos detectados pelo método k-means não apresentaram também predominância de participação de acessos caracterizados por algum fator morfológico, racial ou geográfico. Entretanto, alguns grupos apresentaram uma participação significativa de acessos coletados em uma região geográfica específica: grupo 1 – região Nordeste (43,8%), grupo 2 – região Norte (73,3%), grupo 3 – região Sul (51,3%) e grupo 4 – região Centro-Oeste (48,8%). Os grupos 2, 4 e 6 apresentaram uma composição de 80% de acessos com grãos do

tipo amiláceo, 68,1% de grãos do tipo amiláceo e 55,6% de grãos do tipo pipoca respectivamente. O grupo 1 apresentou 62,4% de acessos representantes das raças Comerciais Recentes, enquanto os grupos 2 e 4 apresentaram, respectivamente, 53,4% e 85% de acessos representantes de raças Indígenas. A seguir são apresentados os resultados observados ao se fazer uma análise conjunta de cada um dos sete grupos genéticos detectados quanto à participação de acessos caracterizados por fatores morfológicos, geográficos e raciais.

Grupo1:

O grupo 1, contendo 109 acessos, é majoritariamente composto por acessos da coleção nuclear (89,9%), 5,5% por acessos de coleções internacionais e apenas 4,6% de acessos da coleção indígena (Figura 6 – A). Este grupo apresentou a maior sobreposição com uma determinada coleção, no caso a coleção nuclear, e a totalidade das linhagens melhoradas das coleções internacionais. Todas as linhagens publicadas por Xu *et al.* (2022) – subtropical, temperada, temperada/subtropical, tropical, tropical/subtropical e tropical/ temperada – foram agrupadas nesse grupo. Os acessos desse grupo são predominantemente oriundos de coletas (67%) (Figura 6 – B). Dos materiais melhorados (18,3%), 50% são não-pipoca resultantes do programa de melhoramento do CNPMS, enquanto 40% são não-pipoca advindos de outros programas de melhoramento, restando 10% de materiais melhorados do tipo pipoca (Figura 6 – C). O grupo 1 conta ainda com 9,2% de materiais introduzidos, dos quais 90% de origem tropical e apenas 10% de origem temperada (Figura 6 – D). As raças comerciais recentes aparecem com mais frequência (56,9%) (Figura 6 – E) nesse grupo, sendo desses, 80,6% da raça Dentada e 19,4% da raça Semi-dentada. O tipo de grão semi-dentado foi o grão mais frequente (37,6%) entre os acessos (Figura 6 – F). Os acessos do grupo 5 são advindos, principalmente, da região Nordeste (43.8%) (Figura 6 – G), em que mais da metade (59,4%) foram coletados no estado da Bahia. A principal região ecogeográfica detentora de 23,3% dos acessos foi o Cerrado, seguido do Agreste – Litoral (16,4%) (Figura 6 – H).

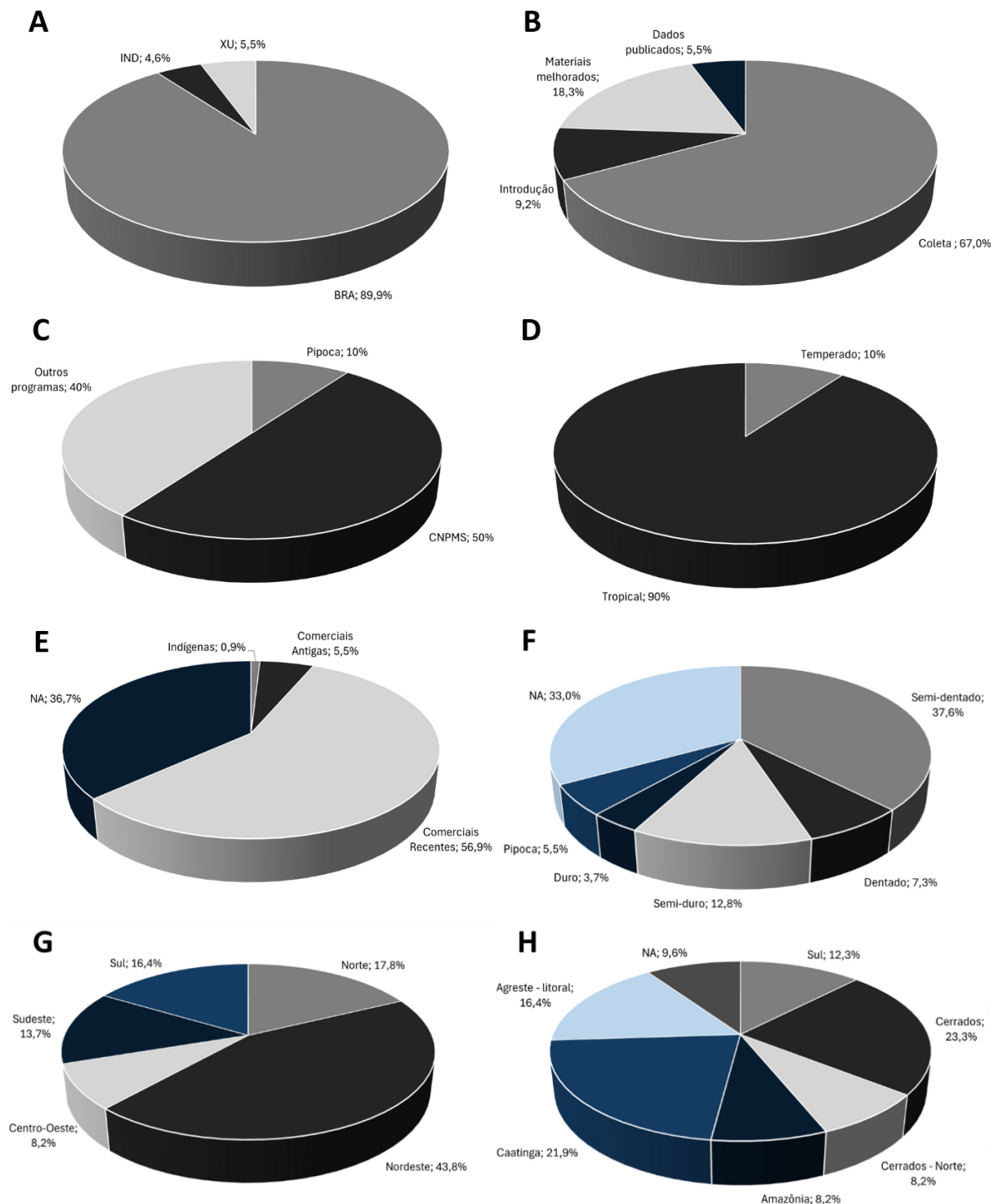
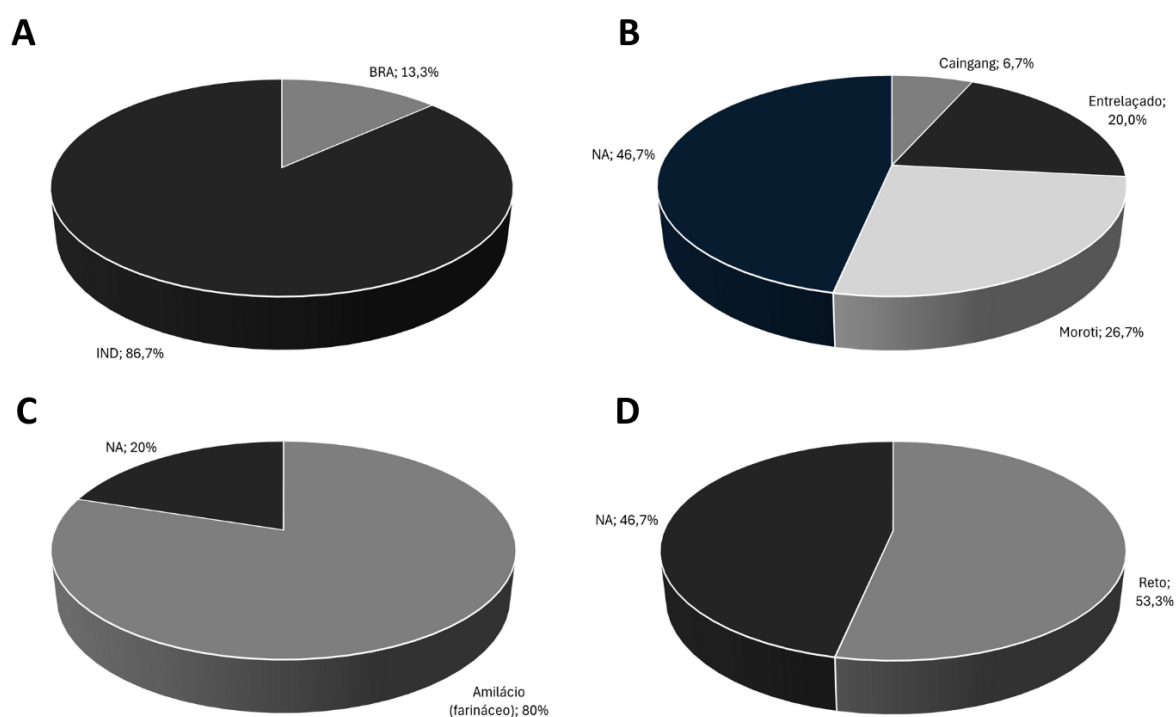


Figura 6 – Distribuição dos 109 acessos alocados no grupo 1 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Origem do acesso; C. Fonte dos materiais melhorados; D. Origem climática dos materiais; E. Classificação do nível de domesticação; F. Tipo de grão; G. Origem geográfica; H. Origem ecogeográfica. NA = sem informação (*not available*).

Grupo 2:

O grupo 2, contendo 15 acessos, é majoritariamente composto por acessos de origem indígena (86,7%) e 13,3% por acessos da coleção nuclear (Figura 7 – A). Entretanto, os dois únicos acessos da coleção nuclear nesse grupo são acessos representantes da coleção indígena dentro da coleção nuclear (AC-045, RO-013) o que aumenta a porcentagem de acessos de origem indígena à 100%, apontando para um grupo com elevada coincidência entre o agrupamento detectado pelos marcadores SNPs e a classificação indígena. Todos os acessos são advindos de coleta. A raça mais frequente nesse grupo é a raça Moroti (26,7%), contudo 46,7% dos acessos não possuem informação de raça disponível (Figura 7 – B). Este resultado sugere que todos estes acessos ainda não classificados racialmente provavelmente pertencem a esta raça. O único tipo de grão presente entre os acessos do grupo 2 foi do tipo amiláceo (80%) (Figura 7 – C) e o único tipo de arranjo dos grãos foi do tipo reto (53,3%) (Figura 7 – D). A cor de grão que aparece com mais frequência é o amarelo, seguido da cor branca (13,3%) e dos grãos coloridos (13,3%) (Figura 7 – E). Os acessos são advindos, principalmente, da região Norte (Figura 7 – F) e a maior parte dos acessos não tinham informações sobre a região ecogeográfica de origem. Apenas os acessos compartilhados entre as duas coleções (AC-045-IND, AC-045-BRA e RO-013-IND e RO-013) foram indicados como integrantes da região ecogeográfica Amazônia. A informação genômica sugere que todos os acessos sem informação de origem ecogeográfica também sejam oriundos da região Norte.



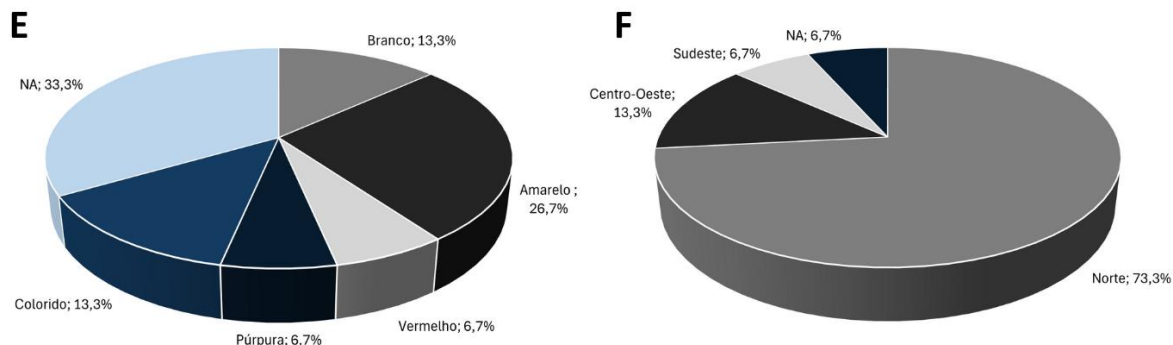


Figura 7 – Distribuição dos 15 acessos alocados no grupo 2 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Raça; C tipo de grão; D. Arranjo dos grãos; E. Cor dos grãos; F. Origem geográfica. NA = sem informação (*not available*).

Grupo 3:

O grupo 3 com 69 acessos apresenta composição mais diversificada quanto à fonte dos acessos: 46,4% pertencem à coleção nuclear, 46,4% à coleção indígena e 7,2% às coleções internacionais (Figura 8 – A). As duas coleções compartilham entre si dois acessos (RS-229 e RS-230). Composto por acessos provenientes principalmente de coleta (78,3%) (Figura 8 – B), esse grupo possui 5,8% de materiais melhorados, dos quais metade desses acessos são não-pipoca advindos de outros programas de melhoramento e a outra metade é igualmente dividida entre os grupos pipoca e não-pipoca resultantes do programa de melhoramento do CNPMS (Figura 8 – C), e, ainda, 8,7% de materiais introduzidos majoritariamente de origem temperada (83,3%) (Figura 8 – D). Dos acessos com a classificação racial disponível (61%) (Figura 8 – E), 47,8% são classificados em raças indígenas, 1,4% são de raças comerciais antigas, 4,3% de raças comerciais recentes e 7,2% são teosintos (*Zea parviglumis* do grupo geográfico Sul de Guerreiro, Chalco, Central Balsas, Leste Balsas e Central Plateau). Informações sobre tipo de grãos estava disponível apenas para a metade dos acessos (Figura 8 – F). O tipo de grão pipoca (20,3%) e amiláceo (15,9%) foram os mais frequentes nesse grupo. A região Sul foi a região geográfica mais representada neste grupo (51,9%) (Figura 8 – G), sendo desses, 37% provenientes do estado do Rio Grande do Sul. Mais da metade dos acessos (59,3%) não possuíam informações sobre a região ecogeográfica de origem (Figura 8 – H), entretanto a região ecogeográfica mais representada entre aqueles com informação foi a região Sul (14,8%), seguida pelo Cerrados - Norte (13%).

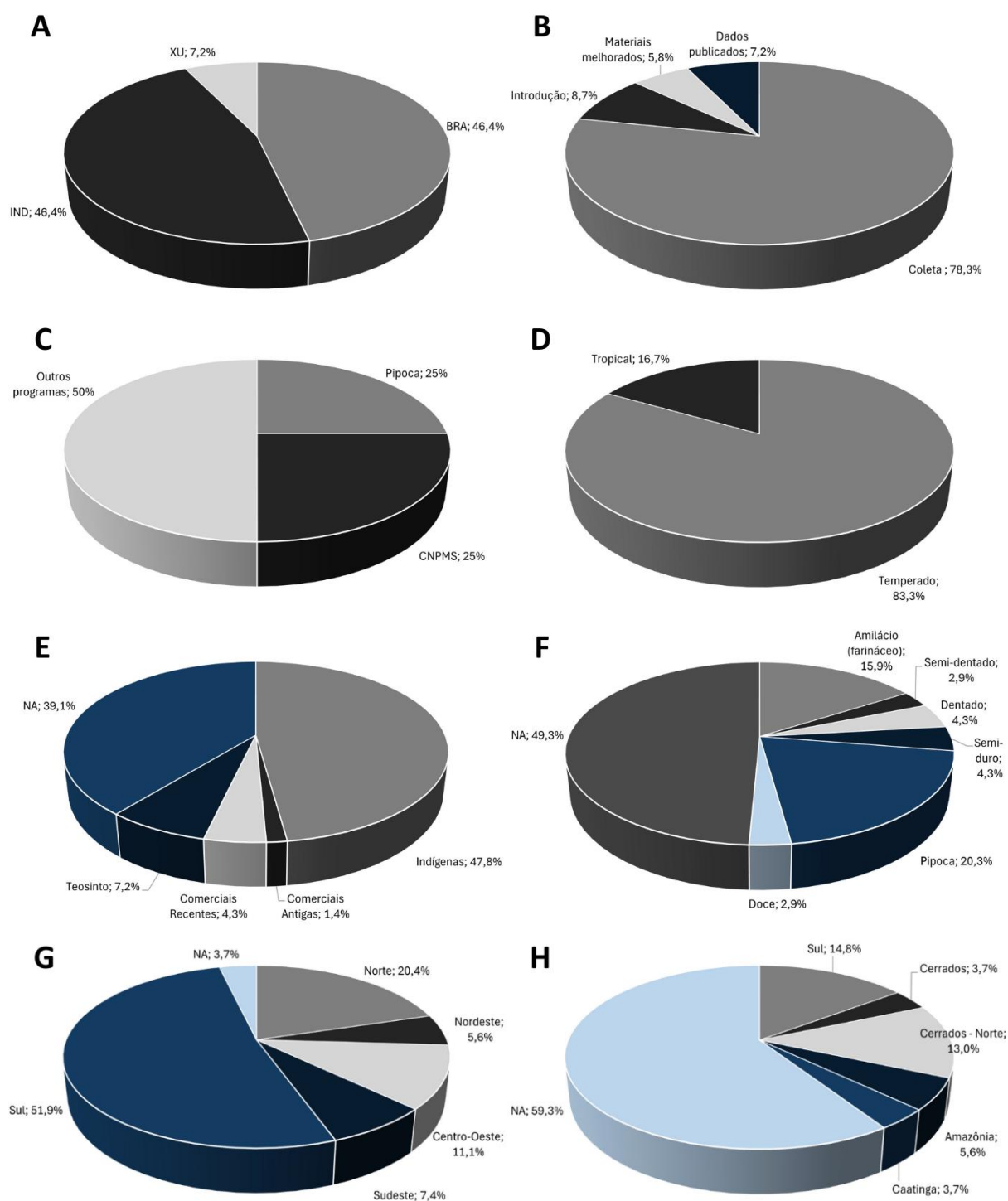
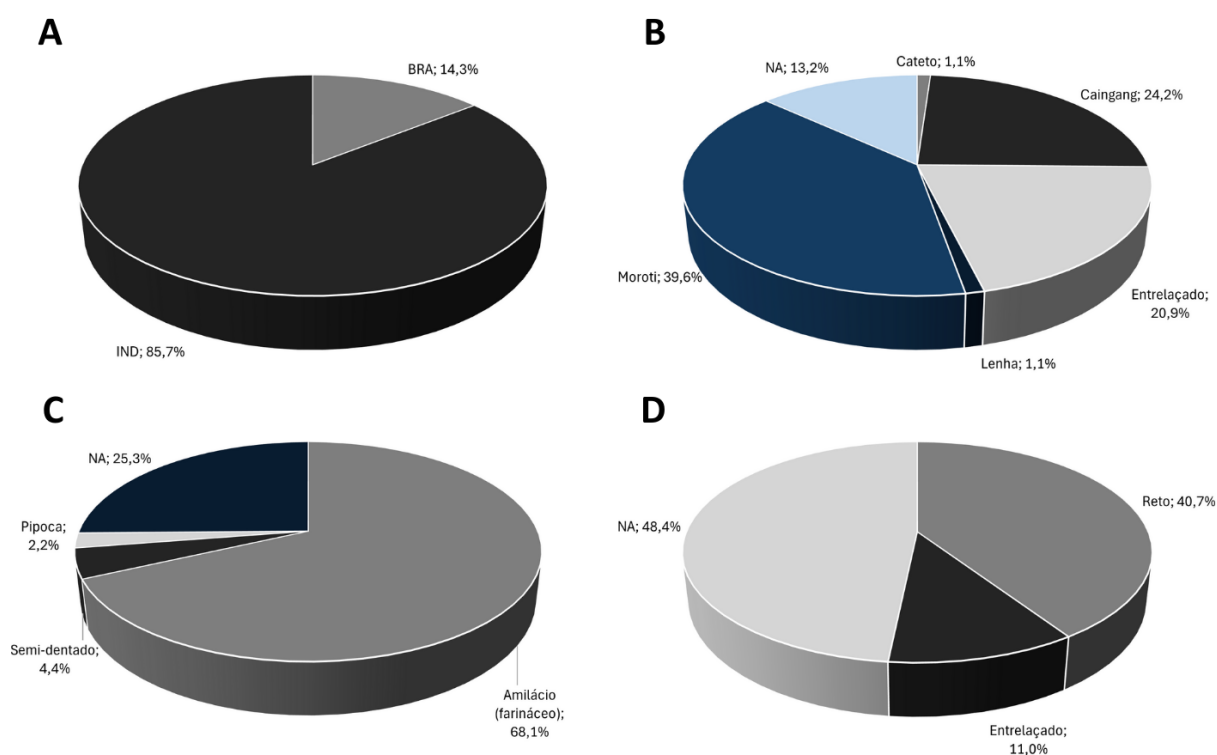


Figura 8 – Distribuição dos 69 acessos alocados no grupo 3 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Origem do acesso; C. Fonte dos materiais melhorados; D. Origem climática dos materiais; E. Classificação do nível de domesticação; F. Tipo de grão; G. Origem geográfica; H. Origem ecogeográfica; NA = sem informação (*not available*).

Grupo 4:

O grupo 4, contendo 91 acessos, é majoritariamente composto por acessos de origem indígena (85,7%) e por 14,3% de acessos pertencentes à coleção nuclear (Figura 9 – A). Entretanto, 5 acessos são compartilhados entre as duas coleções (MT-008, MT-0012, MT-033, MT-035 e SP-620 Glicerio de Icatu) o que aumenta a composição final de acessos de origem indígena para 90,1%. Este grupo apresentou a segunda maior sobreposição com uma determinada coleção, no caso a coleção indígena. Os acessos desse grupo são todos oriundos de coleta. No grupo 4 não há acessos representantes das raças comerciais recentes e 98,9% dos acessos são de raças indígenas (Figura 9 – B), com a raça Moroti sendo a raça mais frequente (39,6%), seguida da raça Caingang (24,2%) e logo após a raça Entrelaçado (20,9%). Apenas o acesso MS-052 é classificado como Cateto, representando as raças Comerciais Antigas. Mais da metade dos acessos (68,1%) possuem o tipo de grão amiláceo (Figura 9 – C). O arranjo de grão reto é o mais frequente nesse grupo (40,7%) (Figura 9 – D), assim como a coloração branca (28,6%) (Figura 9 – E). A região Centro-Oeste foi a região geográfica mais representada (48,4%) (Figura 9 – F), sendo desses, 40,6% coletados no estado do Mato Grosso. A maior parte dos acessos (85,7%) não possuíam dados sobre a região ecogeográfica e apenas 11 acessos foram classificados: 3 acessos da região Cerrados, 1 da região Cerrados – Norte e 7 da região da Amazônia).



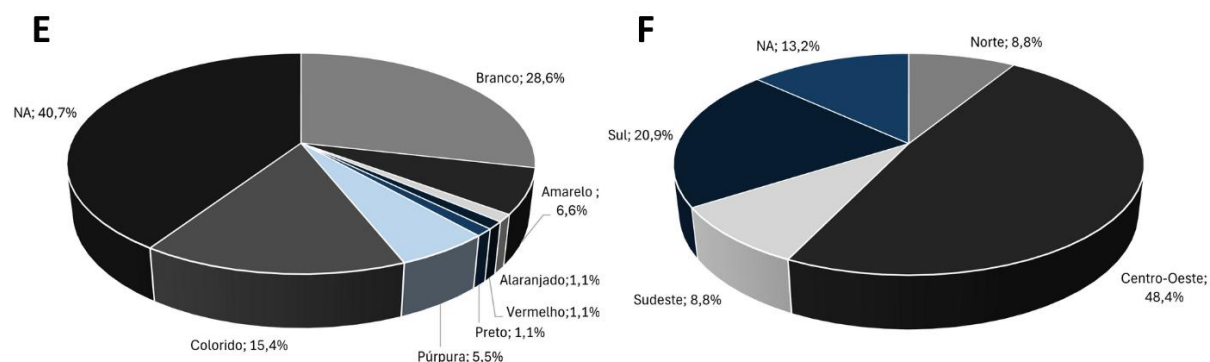
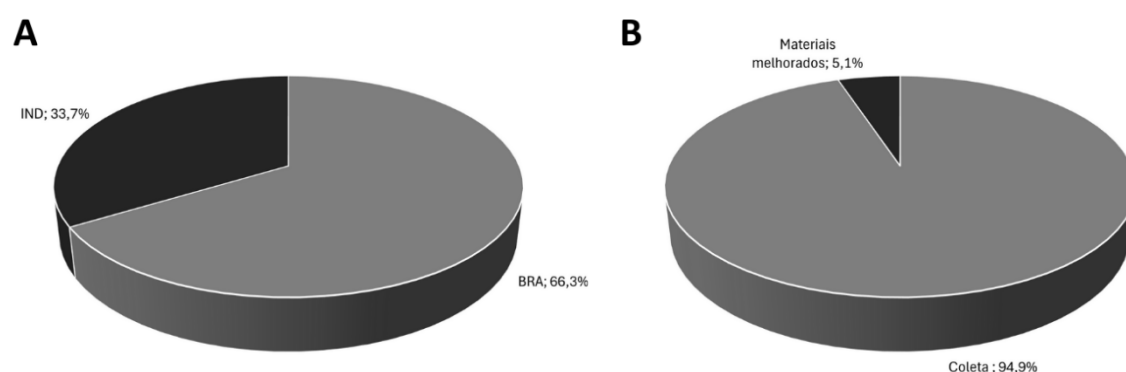


Figura 9 – Distribuição dos 91 acessos alocados no grupo 4 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Raça; C tipo de grão; D. Arranjo dos grãos; E. Cor dos grãos; F. Região geográfica. NA = sem informação (*not available*).

Grupo 5: O grupo 5 contém 98 acessos, sendo 66,3% desses da coleção nuclear e 33,7% acessos de origem indígena (Figura 10 – A). As duas coleções compartilham entre si 5 acessos (GO-002, MS-014, MS-028, PR-054 e RO-002). Esse grupo é composto em sua maioria (94,9%) por acessos provenientes de coleta (Figura 10 – B). O grupo 4 conta ainda com 5,1% de materiais melhorados, todos não-pipoca advindos de outros programas de melhoramento. Dos acessos com a classificação racial disponível (64,3%), mais da metade (35,7%) são classificados em raças indígenas, 12,2% representam as raças comerciais antigas e 16,3% as raças comerciais recentes (Figura 10 – C). Os tipos de grão amiláceo (23,5%) e duro (20,4%) foram os tipos de grão mais frequentes nesse grupo (Figura 10 – D). A região Norte e Nordeste aparecem igualmente representadas (23,7%) (Figura 10 – E). A região ecogeográfica mais representada foi o Cerrado (21,5%), seguida pela região da Amazônia (19,4%) (Figura 10 – F).



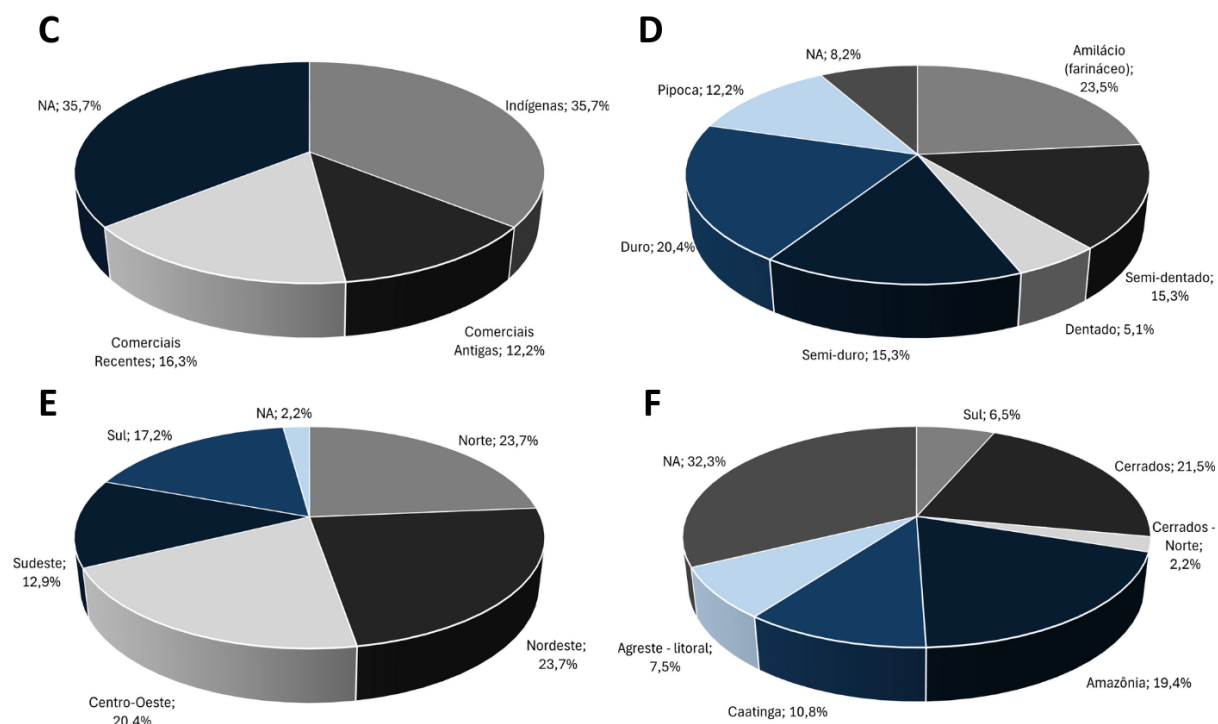


Figura 10 – Distribuição dos 98 acessos alocados no grupo 5 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Origem do acesso; C. Classificação do nível de domesticação; D. Tipo de grão; E. Origem geográfica; F. Origem ecogeográfica. NA = sem informação (*not available*).

Grupo 6:

O grupo 6 contém apenas nove acessos, dos quais seis pertencem à coleção nuclear, dois ao germoplasma de Teosinto (*Z. nicaraguensis* e *Z. parviglumis* do grupo geográfico Jalisco) e apenas um acesso da coleção de origem indígena (Índios Paumari) (Figura 11 – A). Dos seis acessos pertencentes às coleções da EMBRAPA, cinco vieram de coleta (Índios Paumari, Pipoca Gisela, MG-152 Pipoca Magno, RS-020 Pipoca e SC-008), o acesso Denty Flints 393 faz parte das introduções de origem temperada e o acesso Pipoca Amarelo, dos materiais melhorados, classificado como pipoca (Figura 11 – B). Ainda sobre os acessos das coleções, não havia informações disponíveis sobre a classificação racial, o tipo de grão predominante foi o tipo pipoca (Pipoca Amarelo, Pipoca Gisela, MG-152 Pipoca Magno, RS-020 Pipoca e SC-008), e apenas o acesso Índios Paumari foi classificado no tipo de grão duro (Figura 11 – C). Os acessos originados de coleta vieram das regiões Norte (Índios Paumari), Sudeste (Pipoca Gisela e MG-152 Pipoca Magno) e Sul (RS-020 Pipoca e SC-008) (Figura 11 – D). As regiões ecogeográficas encontradas nesse grupo foram Amazônia (Índios Paumari), Cerrados ((Pipoca Gisela e MG-152 Pipoca Magno) e Sul (RS-020 Pipoca e SC-008) (Figura 11 – E). O agrupamento genético de um acesso dos Índios Paumari com germoplasma de Teosinto é

interessante, faz sentido, enquanto a presença de materiais melhorados deste grupo não é esperado.

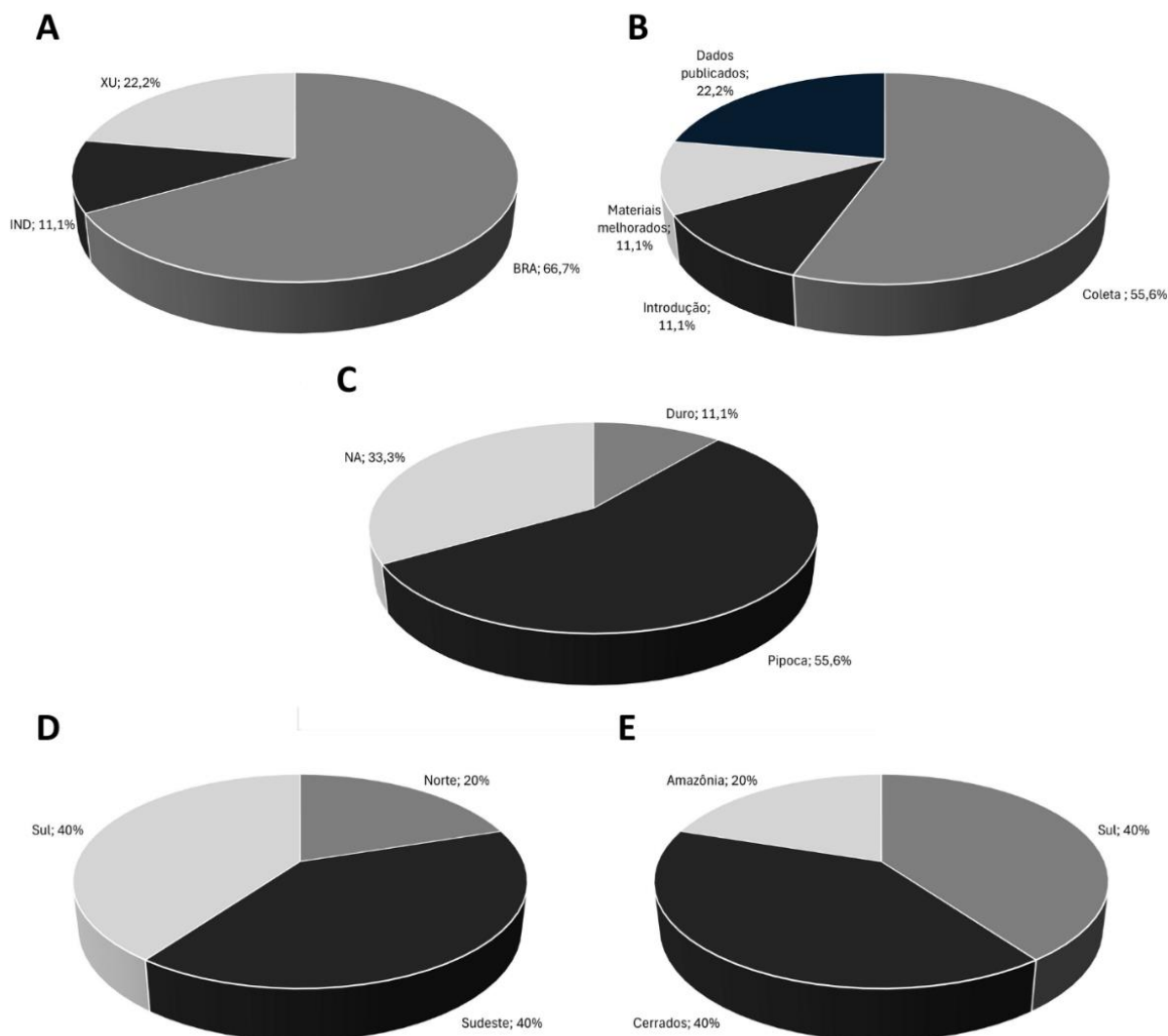
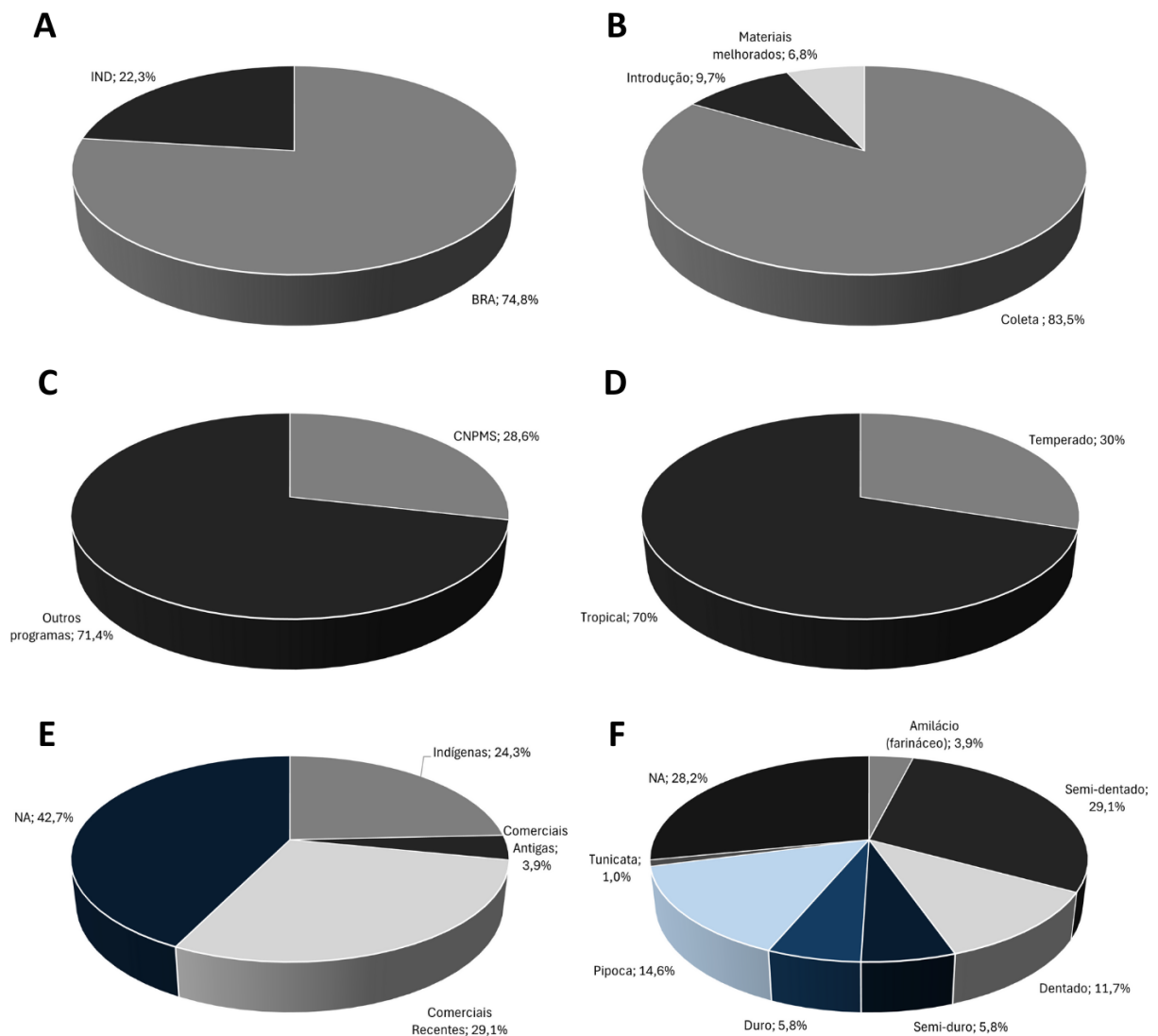


Figura 11 – Distribuição dos nove acessos alocados no grupo 6 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Origem do acesso; C. Tipo de grão; D. Origem geográfica. E. Origem ecogeográfica NA = sem informação (*not available*).

Grupo 7: O grupo 7 contém 103 acessos, sendo 74,8% deles oriundos da coleção nuclear (Figura 12 – A) e em sua maioria (83,5%) provenientes de coleta (Figura 12 – B). Dos materiais melhorados (6,8%), 71,4% são não-pipoca resultantes do programa de melhoramento do CNPMS enquanto 28,6% são não-pipoca advindos de outros programas de melhoramento. Não houve nenhum material melhorado do tipo pipoca nesse grupo (Figura 12 – C). O grupo 7 conta ainda com 9,7% de introduções, dos quais 70% de origem tropical e 30% de origem temperada

(Figura 12 – D). Os acessos com a classificação racial disponível (57,3%), foram bem divididos, principalmente, entre as raças comerciais recentes (29,1%) e as raças indígenas (24,3%) Apenas 3,9% representaram as raças comerciais antigas (Figura 12 – E). Entre as comerciais recentes, o tipo Dentado apareceu com mais frequência (66,7%), e entre as raças indígenas, 72% dos acessos são classificados na raça Lenha. O tipo de grão mais frequente nesse grupo foi o tipo semi-dentado (29,1%), seguido do tipo pipoca (14,6%) e do tipo dentado (11,7%) (Figura 12 – F). A região Sul foi a região geográfica que mais contribuiu acessos (41,9%) (Figura 12 – G) enquanto a região ecogeográfica mais representada foi o Cerrado da região Norte (22,1%) (Figura 12 – H).



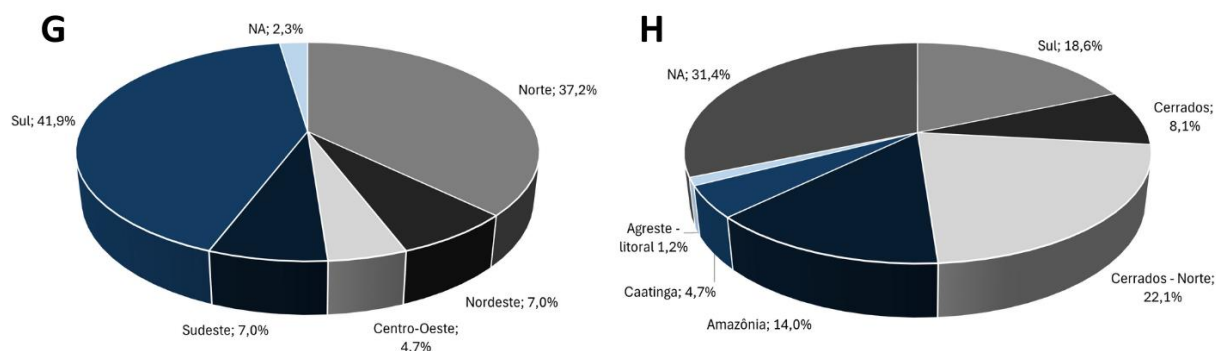


Figura 12 – Distribuição dos 103 acessos alocados no grupo 7 detectado pelo k-means nas diferentes classificações morfológicas, ecogeográficas e de origem. A. Origem de coleções de germoplasma; B. Origem do acesso; C. Fonte dos materiais melhorados; D. Origem climática dos materiais; E. Classificação do nível de domesticação; F. Tipo de grão; G. Origem geográfica; H. Origem ecogeográfica; NA = sem informação (*not available*).

3.4 ANÁLISE DE ESTRUTURA GENÉTICA COM BASE EM GENÓTIPOS INDIVIDUAIS RECONSTITUIDOS

A análise de estrutura genética com o software STRUCTURE realizada com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos, encontrou um número mais provável de $k=2$ grupos genéticos ($p1$ e $p2$) (Figura 13). Em seguida, na análise individualizada de cada um destes dois grupos foram detectados dois subgrupos dentro do grupo $p1$ e três dentro do grupo $p2$ (Figura 14 – A e Figura 14 – B, respectivamente).

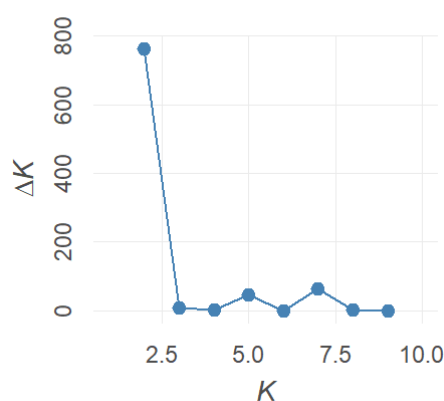


Figura 13 – Definição do número de grupos genéticos com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos. O número de grupos mais provável foi definido como sendo $k=2$ e os dois grupos denominados $p1$ e $p2$.

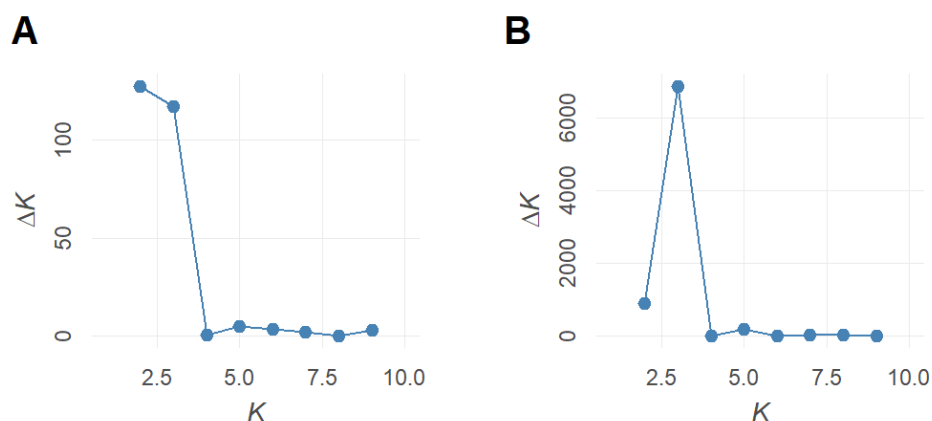


Figura 14 – Definição do número de subgrupos genéticos dentro de cada um dos dois grupos definidos em análise anterior com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos. A. k=2 subgrupos no grupo p1; B. k=3 no grupo p2.

A análise do STRUCTURE foi realizada sobre uma base de dados genotípicos individuais como se tivessem sido genotipados 50 indivíduos para cada um dos 494 acessos. Esta análise permitiu, portanto, estimar a proporção da ancestralidade genômica de cada um dos 24.650 indivíduos seja aos dois grandes grupos genéticos detectados inicialmente na análise, bem como aos subgrupos detectados na análise posterior. Os resultados da proporção genômica alocada a cada um dos grupos para os 24.700 indivíduos é disponível na forma de material suplementar dada a dimensão da tabela. Para fins de visualização deste resultado, a partir dos resultados obtidos para os 50 indivíduos reconstituídos de cada acesso foi calculada a média da proporção genômica de cada um dos grupos para cada acesso. Este resultado é apresentado na Figura 15. Cada barra representa, portanto, um acesso, e as alturas dos segmentos coloridos de cada barra correspondem às proporções médias de ancestralidade alocação aos grupos genéticos detectados pelo STRUCTURE.

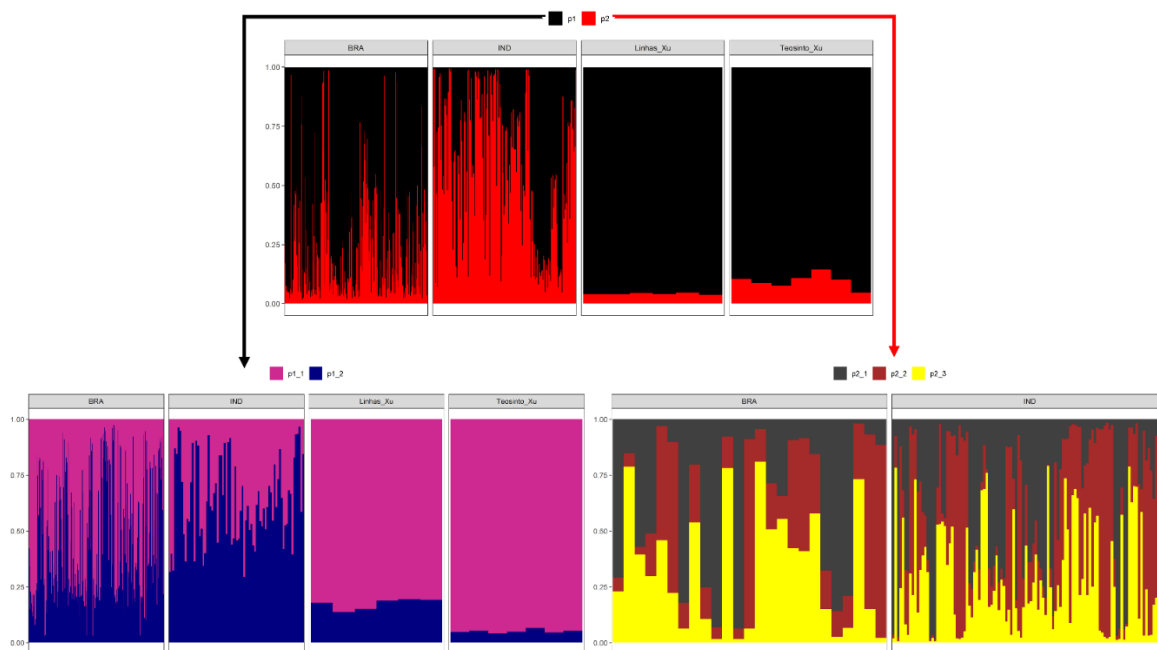


Figura 15 – Proporções médias da ancestralidade genômica dos indivíduos de cada um dos 494 acessos nos grupos genéticos detectados pelo STRUCTURE. Cada barra representa, portanto, um acesso, as diferentes cores representam diferentes grupos genéticos detectados pelo STRUCTURE e as alturas dos segmentos coloridos de cada barra correspondem às proporções médias de ancestralidade alocadas aos diferentes grupos detectados.

O gráfico de barras produzido pela análise do STRUCTURE permite visualizar, a composição genômica de cada acesso, com acessos mais homogêneos nos quais predomina uma das cores (vermelha correspondendo ao grupo p1, ou preta, correspondendo ao grupo p2) enquanto outros acessos apresentam-se mais miscigenados com barras apresentando proporções variáveis das duas cores. Da mesma forma, na análise subsequente é possível visualizar os níveis de miscigenação ou pureza genética de cada um dos acessos sempre com base nas proporções dos tamanhos dos segmentos com cores alternativas em cada barra correspondendo a cada acesso. Os percentuais de ancestralidade genômica podem ser melhor visualizados por acesso na tabela de dados suplementares a esta tese. Este tipo de visualização não seria possível de ser feito com base apenas em frequências alélicas estimadas para cada acesso a partir dos bulks. A de reconstituição de indivíduos realizada neste trabalho abriu, portanto, a possibilidade de avaliar o nível de miscigenação presente em cada acesso.

Da mesma forma que a análise anterior com base nos sete grupos detectados pelo k-means, foi quantificado o percentual de participação de acessos das diferentes coleções de germoplasma (BRA, IND, Linhas Xu e Teosinto Xu) em cada um dos cinco grupos genéticos detectados pelo STRUCTURE. Esta análise também demonstrou que não há uma participação exclusiva de nenhuma das coleções nos cinco grupos detectados pelos marcadores de DNA,

mas sim há apenas predominâncias de acessos oriundos de alguma das coleções (Tabela 2). Os subgrupos p1_1 e p1_2 contam com maior número de acessos oriundos da coleção nuclear. Conforme esperado eles também incluem todas as linhagens de germoplasma internacional, mas de certa forma inesperada incluem também todos os materiais de Teosinto. Já os subgrupos do grupo 2 contém uma predominância de acessos indígenas.

Tabela 2 – Percentual de acessos provenientes das diferentes coleções de germoplasma em cada um dos cinco grupos detectados pela análise de estrutura com o software STRUCTURE realizada com base nos genótipos multiloco de 504 SNPs independentes para 50 indivíduos reconstituídos a partir das frequências alélicas para cada um dos acessos. O acesso *Z. nicaraguensis* foi incluído no grupo de germoplasma de Teosinto para fins de simplificação.

Grupos genéticos pelo STRUCTURE	Coleção de germoplasma				Total
	BRA	IND	Linhas Xu	Teosinto Xu	
p1_1	82,07	10,87	3,26	3,8	100
p1_2	67,63	32,37	0	0	100
p2_1	19,61	80,39	0	0	100
p2_2	14,89	85,11	0	0	100
p2_3	20,51	79,49	0	0	100

Ao compararmos a alocação de acessos aos sete grupos detectados pelo método de agrupamento k-means com a alocação de acessos aos cinco grupos detectados pelo STRUCTURE (Figura 16 – A; Tabela 3) percebe-se que os grupos 1, 4, 5 e 7 detectados pelo k-means praticamente correspondem o grupo p1 (composição predominante de acessos da coleção nuclear BRA) detectado pela análise do STRUCTURE, enquanto que os grupos 3 e 6 do agrupamento k-means correspondem ao grupo p2 (composição predominante de acessos da coleção indígena IND). O grupo 2 do agrupamento k-means com apenas 15 acessos aparece com participações relevantes de acessos das coleções BRA e IND. O mapa de calor utilizado na Tabela 3 facilita esta observação, na medida que quanto mais verde, maior a correspondência entre as alocações de acessos a grupos, produzidas pelos dois métodos de análise.

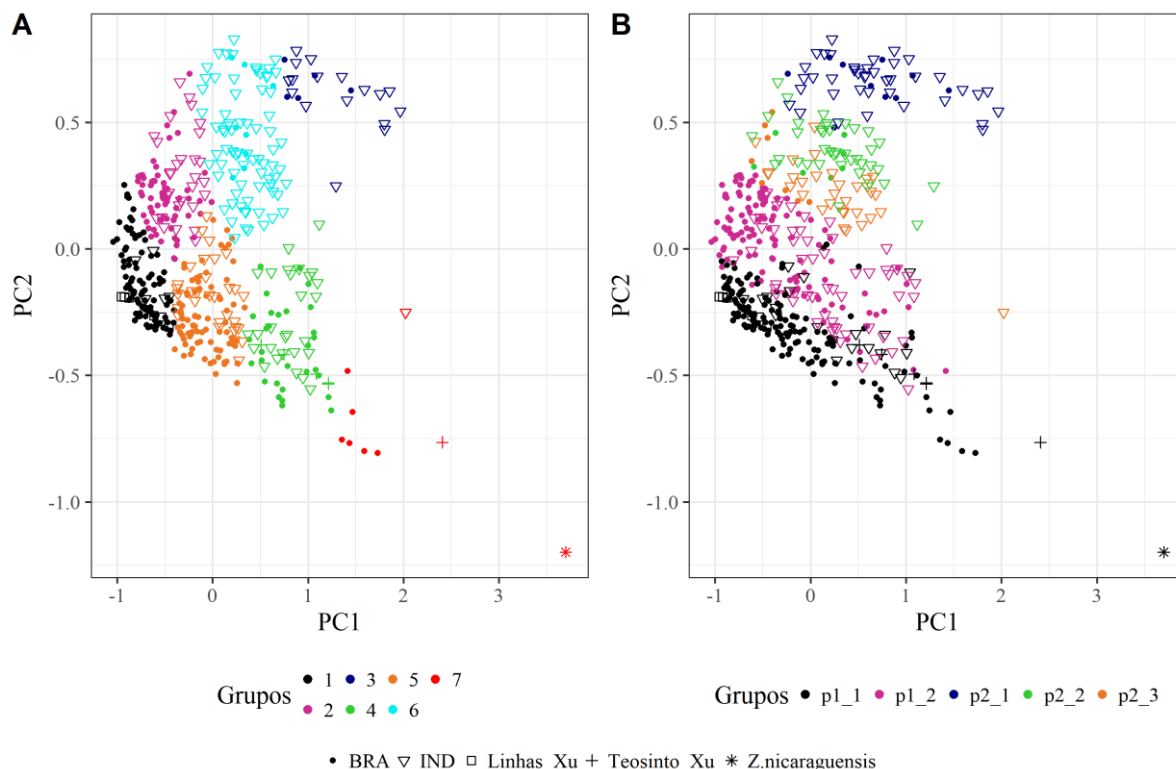


Figura 16 – Comparação entre (A) a dispersão de acessos nos sete grupos genéticos detectados pelo agrupamento k-means (A) e (B) nos cinco grupos genéticos detectados pela análise de estrutura com base em indivíduos simulados com os software STRUCTURE. Os eixos correspondem aos dois primeiros componentes principais da matriz de distância com base no índice de fixação F_{ST} .

Tabela 3 – Mapa de sobreposição da alocação de acessos a grupos genéticos detectados nas análises de estrutura com o software STRUCTURE e agrupamento pelo k-means. Os valores correpondem aos percentuais de acessos alocados a cada um dos sete grupos definidos pelo método k-means, atribuidos aos cinco grupos genéticos detectados na análise de estrutura com o software STRUCTURE. O mapa de calor indica que quanto mais verde, maior o percentual de correspondência entre as alocações de acessos a grupos produzidas pelos dois métodos de análise.

Grupos STRUCTURE	Grupos K-means						
	1	2	3	4	5	6	7
p1_1	70.37	0	0	52.38	59.13	0	77.78
p1_2	29.63	67.39	0	46.03	40.87	2.38	11.11
p2_1	0	2.17	95.65	0	0	32.14	0
p2_2	0	13.04	4.35	1.59	0	39.29	0
p2_3	0	17.39	0	0	0	26.19	11.11
	100	100	100	100	100	100	100

3.5 COMPARAÇÃO DAS CLASSIFICAÇÕES MORFOLÓGICAS E GEOGRÁFICAS COM AGRUPAMENTOS DETECTADOS PELOS DADOS GENÉTICOS

Além das avaliações anteriores dos percentuais de coincidência de alocação de acessos aos grupos genéticos detectados pelos dois métodos analíticos utilizados, as análises a seguir visaram avaliar se os agrupamentos de acessos identificados pelos marcadores moleculares e visualizados nos dendrogramas UPGMA poderiam coincidir ou explicar parcial ou integralmente as classificações morfológicas, raciais ou geográficas dos acessos. Nos dendrogramas a seguir foram utilizados códigos de cores correspondendo às diferentes classificações visando assim tentativamente fornecer uma visualização das possíveis sobreposições dos agrupamentos genéticos identificados pelos marcadores SNPs e as diferentes classificações morfológicas e ecogeográficas utilizadas.

Na análises anteriores referentes à dispersão dos acessos nos gráficos de k-means (ex. Figura 4) foi possível visualizar que embora haja uma separação entre os acessos das coleções nuclear e indígena há significativa sobreposição de acessos de diferentes coleções, corroborada pelas análises dos percentuais de acessos provenientes das diferentes coleções de germoplasma nos diferentes grupos genéticos detectados (Tabelas 1 a 3).

Padrões complexos são também observados agora nos dendrogramas. Em relação à classificação de acessos quanto ao grau de domesticação (Figura 17) há uma separação relativamente clara entre um grupo de raças Indígenas menos domesticadas (em azul) na porção inferior do dendrograma em relação aos acessos Comerciais Recentes (em verde) na porção superior. Contudo, três acessos indígenas são observados entre os materiais comerciais recentes o que levanta hipóteses de que eles possam, na verdade ser materiais comerciais recentes utilizados por aquelas populações. De forma inversa dois materiais classificados como comerciais recentes são observados agrupados com materiais indígenas o que sugere mais provavelmente problemas de classificação. Seja *Z. nicaraguenses*, bem com todos os acessos de Teosinto (em roxo) aparecem em um grupo bastante discreto junto às cultivares indígenas o que é o esperado. Não houve agrupamento significativo das raças Comerciais Antigas as quais se posicionam de forma dispersa no dendrograma.



Figura 17 – Sobreposição da classificação racial dos 494 acessos estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright). NA = *not available*.

Em uma segunda análise, foi sobreposta a classificação racial dos acessos indígenas no agrupamento produzido no dendrograma UPGMA (Figura 18). De maneira geral o dendrograma indica que a classificação racial dos acessos indígenas tem baixa correlação com a proximidade genética dos acessos. A raça Caingang (em vermelho) formou um grupo com acessos advindos do estado do Paraná e, também, com acessos do estado de São Paulo, sugerindo possível intercâmbio de materiais genéticos entre etnias de estados vizinhos. As raças Entrelaçado (em ocre) e Moroti (em rosa) apresentaram acessos intercalados ao longo do dendrograma, advindos principalmente dos estados do Mato Grosso e Rondônia, também estados fronteiriços. Em outras palavras, não se observam agrupamentos claros consistentes com as diferentes raças, com exceção das raças Lenha e Caingang, que apresentaram alguns grupos mais claros, embora mesmo para estas raças vários dos seus acessos se posicionam de forma dispersa em relação aos seus principais grupos formados. Acessos das raças Entrelaçado e Moroti apresentaram uma dispersão sem agrupamentos claros. Estes resultados indicam de forma geral, que a ancestralidade genética comum entre acessos pouco explica a classificação racial adotada. Isto por sua vez pode estar refletindo forte intercâmbio de materiais genéticos entre as tribos de etnia homônima.

A sobreposição de raças com análise genética também é particularmente interessante pois permite, a princípio, atribuir a raça mais provável aos acessos ainda não classificados, partindo da premissa que acessos de uma mesma raça apresentariam ancestralidade comum e, portanto, maior proximidade genética. Por exemplo, os acessos MT-009, MG-059 e BA-168, todos eles não classificados quanto à raça, são geneticamente próximos e se agruparam a pequenas distâncias genética de três acessos da raça Moroti, assim como os acessos KAYABI-1-AMARELO e KAYABI-2-PRETO. Todos estes acessos mais provavelmente pertencem à raça Moroti. Já o acesso I-74-6-PI-KATETE mais provavelmente pertence à raça Guarani.

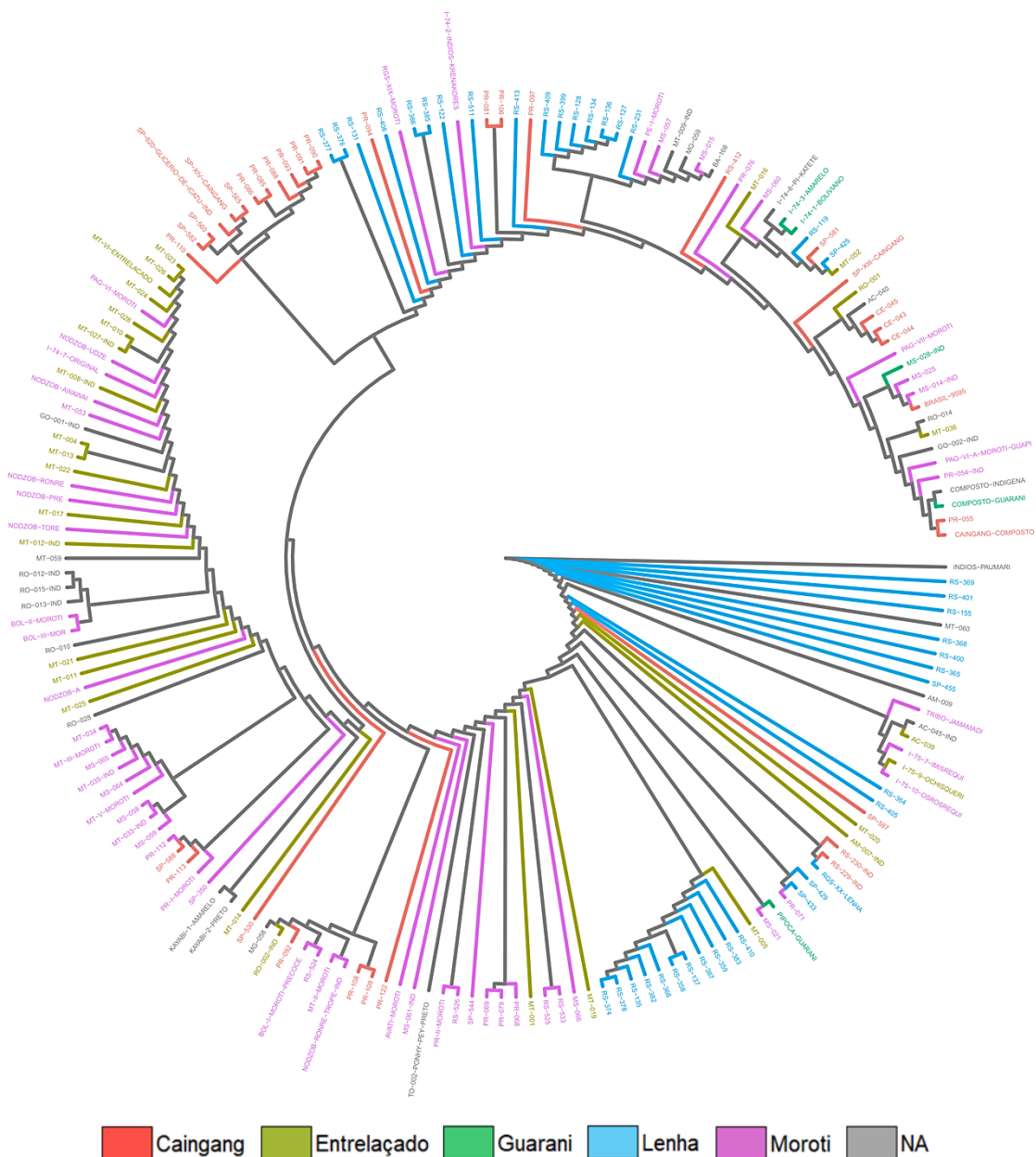


Figura 18 – Sobreposição da classificação racial dos 188 acessos indígenas estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright). NA = *not available*.

Ao sobrepor a informação de origem geográfica sobre o dendrograma, não há agrupamentos específicos coincidentes com regiões geográficas o que indica uma ampla dispersão da diversidade genética do germoplasma de milho por todo país bem como corrobora a expectativa da coleção nuclear representar de forma satisfatória esta diversidade (Figura 19). Contudo, houve a formação de pequenos grupos bem definidos. A região Centro-Oeste (indicada em vermelho) pareceu formar dois grupos: um do estado do Mato Grosso e um do estado do Mato Grosso do Sul. Esses grupos foram intercalados por grupo de acessos da região Norte (em verde) advindos do estado de Rondônia, estado fronteiro ao estado do Mato Grosso, e alguns poucos acessos coletados no estado do Acre, estado vizinho do estado de Rondônia, sugerindo possível intercâmbio de material genético entre esses estados ou migrações de povos indígenas. Acessos da região Norte formaram um pequeno grupo bem definido de acessos do estado de Roraima, sugerindo possível isolamento em relação a outros estados.

Alguns outros grupos não tão bem definidos puderam ser observados: acessos intercalados dos estados de São Paulo e do Paraná, estados vizinhos, na porção inferior do dendrograma, acessos do Rio Grande do Sul na porção medial esquerda. É possível observar que houve uma concentração maior de acessos da região Sul e Centro-Oeste na mesma porção (inferior) do dendrograma onde a maior parte dos acessos advindos da coleção indígena se agrupou. Acessos destas duas regiões geográficas constituem 35,6% e 31,4% da coleção indígena, respectivamente, enquanto acessos da região Nordeste que representam 25,3%, concentraram-se majoritariamente na porção superior do dendrograma contendo mais acessos da coleção nuclear (Figuras 5 e 19). Acessos da região Norte que representam 32,9% dos acessos da coleção nuclear e apenas 9,6% da coleção indígena apareceram de forma mais dispersa ao longo do dendrograma.

A sobreposição da classificação dos acessos quanto ao tipo de grão sobre o dendrograma revelou dois grandes grupos: (1) acessos com grão do tipo amiláceo advindos, principalmente, dos estados do Mato Grosso, Mato Grosso do Sul, Rondônia, São Paulo e Paraná e (2) acessos com o tipo de grão semi-dentado e dentado advindos, majoritariamente, de estados das regiões Norte e Nordeste. Nota-se também um pequeno grupo de acessos com grãos do tipo pipoca. Contudo, não houve uma segregação total entre acessos com o mesmo tipo de grão o que indica a existência de ampla variabilidade genética seja entre bem como dentro as diferentes categorias de tipo de grão utilizadas neste classificação de germoplasma de milho (Figura 20).



Figura 19 – Sobreposição da origem geográfica de coleta dos 494 acessos estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright). NA = *not available*.

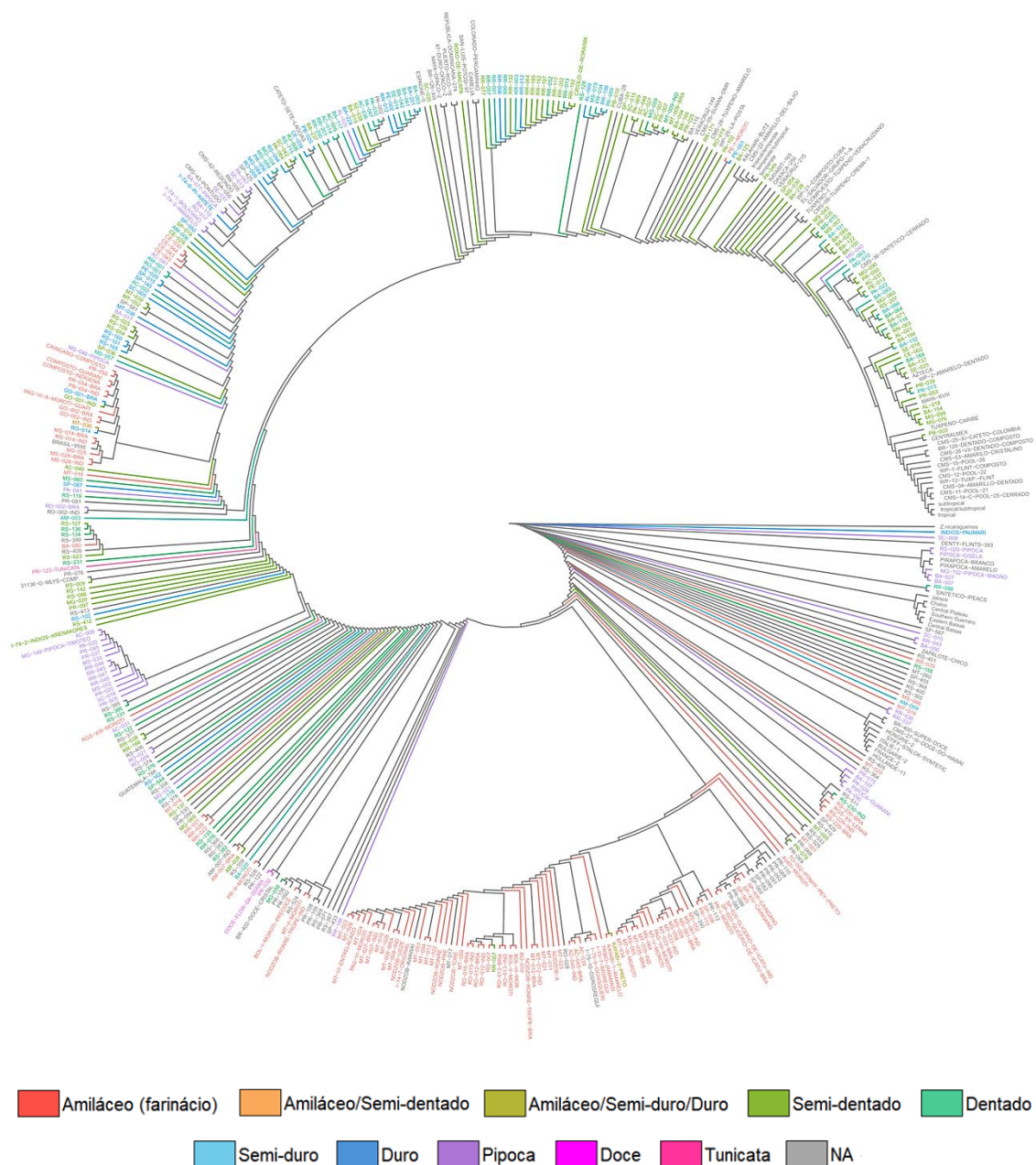


Figura 20 – Sobreposição da classificação quanto ao tipo de grão dos 494 acessos estudados sobre o dendrograma UPGMA construído a partir da matriz de distâncias genéticas entre os acessos com base da matriz de F_{ST} (índice de fixação de Wright). NA = *not available*.

4. DISCUSSÃO

Tendo em vista os resultados descritos quanto aos diferentes agrupamentos detectados com base na análise genética, e as avaliações da sobreposição das classificações de germoplasma tipicamente utilizadas sobre a informação genética gerada, a discussão a seguir é apresentada na forma de perguntas e respostas. Algumas das possíveis perguntas a serem feitas com base nesta análise genética são propostas e tentativamente respondidas com o intuito de aumentar o conhecimento a respeito destas coleções e assim agregar, ainda mais, valor a estes recursos genéticos. Esta discussão não pretende ser exaustiva, uma vez que outras perguntas poderão surgir bem como novas análises e abordagens exploratórias a partir dos dados genéticos gerados nesse trabalho serão possíveis no futuro.

Qual a magnitude da diferenciação genética entre as coleções nuclear e indígena?

As duas coleções apresentaram moderada diferenciação genética, com um F_{ST} médio igual a 0,136. Os acessos das duas coleções apresentaram separação na análise de k-means e análise de estrutura com a definição de grupos genéticos contendo predominantemente acessos de uma coleção ou outra (Figuras 3 e 4) embora haja acessos que violem esta segregação. Muito embora 22 acessos da coleção indígena também estejam contidos na coleção nuclear, a análise revelou que a sobreposição de acessos da coleção nuclear com os acessos de origem indígena pode ultrapassar isso. O agrupamento k-means indicou grupos com predominância de acessos indígenas (2 e 4) e grupos com mais acessos da coleção nuclear (1, 6 e 7), enquanto apenas o grupo 3 mostrou-se com composição mista (Tabela 2), evidenciando a separação entre as coleções. A análise com o software STRUCTURE corroborou este padrão separando os acessos em duas populações p1 e p2, com mais acessos da coleção nuclear e indígenas, respectivamente (Tabela 3).

Qual a magnitude da diferenciação genética entre as raças indígenas? Os 188 acessos da coleção indígena dividem-se em 35 acessos da raça Caingang (18,6%), 29 acessos da raça Entrelaçado (15,4%), 5 acessos da raça Guarani (2,7%), 41 acessos da raça Lenha (21,8%), 55 acessos da raça Moroti (29,3%) e 23 acessos sem classificação racial (12,2%). As raças indígenas apresentaram diferenciação moderada à alta, com um F_{ST} médio variando de $F_{ST} = 0,117$ entre Caingang e Guarani e $F_{ST} = 0,162$ entre Lenha e Moroti.

Caingang. A raça Caingang apresentou maior diferenciação genética em relação à raça Lenha ($F_{ST} = 0,149$) e maior proximidade com a raça Guarani ($F_{ST} = 0,117$). A diferenciação genética também foi moderada quando comparada com as raças Moroti e Entrelaçado com o F_{ST} médio igual a 0,129 e 0,130, respectivamente.

Entrelaçado. A raça Entrelaçado teve alta diferenciação com a raça Lenha ($F_{ST} = 0,162$), e moderada com as demais (Caingang – $F_{ST} = 0,130$; Guarani – $F_{ST} = 0,126$; Moroti – $F_{ST} = 0,125$).

Guarani. A diferenciação da raça Guarani também foi considerada moderada quando comparada às outras raças, variando entre: Caingang – $F_{ST} = 0,117$; Moroti – $F_{ST} = 0,124$; Entrelaçado – $F_{ST} = 0,126$ e Lenha – $F_{ST} = 0,133$).

Lenha. A raça Lenha foi a raça que mostrou maior diferenciação genética quando comparada às demais raças, apresentando alta diferenciação com as raças Entrelaçado ($F_{ST} = 0,162$) e Moroti ($F_{ST} = 0,160$) e diferenciação moderada com as raças Caingang ($F_{ST} = 0,149$) e Guarani ($F_{ST} = 0,133$).

Moroti. A raça Moroti teve alta diferenciação com a raça Lenha ($F_{ST} = 0,160$), e moderada com as demais (Caingang – $F_{ST} = 0,129$; Entrelaçado – $F_{ST} = 0,125$; Guarani – $F_{ST} = 0,124$).

A raça Lenha é a raça com maior diferenciação genética em relação às demais raças. No dendrograma com sobreposição da classificação racial das raças indígenas (Figura 18) pode ser notado uma separação mais visível dessa raça. A distribuição restrita ao estado de Rio Grande do Sul (TEIXEIRA *et al.*, 2021) pode ter ocasionado uma diminuição do intercâmbio de acessos da raça Lenha com outras etnias e, com isso, aumentado a sua diferenciação genética. Apesar dessa grande diferenciação, vale notar que a raça Caingang é a raça com menor diferenciação em relação à raça Lenha, possivelmente devido ao compartilhamento da área de dispersão na região Sul do país. As raças Lenha e Caingang, raças que representam juntas 40,4% da coleção indígena, são as responsáveis pela coleção ser representada, principalmente, pela região Sul, embora fosse esperado que a representação da região Norte, onde é encontrado o maior número de etnias indígenas brasileiras (TEIXEIRA *et al.*, 2021), fosse mais ampla.

Apesar da raça Guarani ter sido estudada em um grupo só com a raça Moroti no trabalho de Teixeira *et al.* (2021) por terem sua origem vinculada à mesma etnia, a raça Guarani apresentou menor diferenciação em relação à raça Caingang do que à raça Moroti, entretanto é importante lembrar o pequeno número de representantes dessa raça. Vale ressaltar que indígenas Caingang viviam lado a lado com indígenas Guarani, vivendo entre as áreas ocupadas pelos

Guaranis do Sul do país e, também, Tupi-Guaranis em uma faixa que se estende desde São Paulo até o Uruguai, embora estas raças de milho apresentem grãos bastante distintos (PATERNIANI; GOODMAN, 1977). Os acessos intercalados entre as raças Entrelaçado e Moroti no dendrograma com sobreposição da informação de raças indígenas (Figura 18) evidencia a menor diferenciação genética entre essas duas raças. Isso pode ser resultado da distribuição dessas raças, uma vez que para a raça Entrelaçado 82% dos acessos foram coletados na região Centro-Oeste, enquanto a raça Moroti apresentou 49% dos acessos advindos da mesma região.

Qual das duas coleções, nuclear e indígena, apresenta maior diversidade genética? A coleção nuclear foi pensada e montada originalmente com o objetivo de manter ampla variabilidade genética. Entretanto esta coleção nuclear foi proposta exclusivamente com base em dados de passaporte envolvendo origem geográfica do acesso, dados morfométricos e demais informações disponíveis. Não havia informações sobre diversidade genética. Os resultados gerados neste trabalho permitem agora agregar dados genéticos de marcadores SNPs em escala genômica. Estes dados fornecem informações quanto à ancestralidade comum ou diferente entre acessos permitindo um enriquecimento importante do conhecimento das coleções. Além disso, as novas informações genéticas abrem a perspectiva de atualização da coleção nuclear pela remoção ou introdução de acessos visando ampliar a representatividade da diversidade genética da espécie. Um resultado importante foi a observação de que a coleção indígena apresentou maior diferenciação genética entre os acessos ($F_{ST} = 0,140$) em comparação à coleção nuclear ($F_{ST} = 0,109$). A coleção indígena apresentou maior dispersão de acessos (Figura 3) na análise de estrutura populacional, assim como foi a coleção que apresentou acessos em todos os sete grupos detectados pelo método k-means. Este resultado demonstra, portanto, que há ampla variabilidade genética dentro da coleção indígena com acessos pertencentes a diferentes subgrupos dentro do grupo indígena. Além disso, com a exceção da raça Guaraní ($F_{ST} = 0,101$), com apenas cinco acessos, todas as demais raças indígenas (Caingang: $F_{ST} = 0,117$; Entrelaçado: $F_{ST} = 0,115$; Lenha: $F_{ST} = 0,141$; Moroti: $F_{ST} = 0,128$) apresentaram maior diferenciação genética do que materiais das raças Comerciais Antigas ($F_{ST} = 0,085$) e Comerciais Recentes ($F_{ST} = 0,109$). É importante lembrar que a raça Lenha, que apresenta maior diversidade entre as raças indígenas não está representada na coleção nuclear. Além disso o acesso Índios Paumari, sem classificação racial, apresentou alta diferenciação genética em relação as demais raças indígenas raça Guaraní, $F_{ST} = 0,229$ com a

raça Lenha, e, ainda, $F_{ST} = 0,212$ com a coleção nuclear. O acesso Índios Paumari também não está representado na coleção nuclear. Estes resultados sugerem a necessidade de uma atualização da composição da coleção nuclear incluindo acessos indígenas e assim ampliar a diversidade genética representada no germoplasma disponível da espécie., variando de $F_{ST} = 0,212$ com a

Qual seria a raça mais provável dos acessos indígenas ainda não classificados quanto à raça? Dos 23 acessos não classificados, um apresentou menor diferenciação genética com a raça Caingang, 12 com a raça Entrelaçado e cinco com a raça Guarani, enquanto quatro acessos apresentaram maior proximidade com a coleção nuclear do que com raças indígenas (Ver anexo 2). Entre esses, 12 acessos apresentaram um F_{ST} médio $\geq 0,1$. O acesso Índios Paumari não apresentou proximidade com nenhuma raça indígena e nem com a coleção nuclear, sempre apresentando alta diferenciação genética. O menor F_{ST} médio apresentado com esse acesso foi de 0,164, com os acessos BA-003 e PA-002, da coleção indígena e da coleção nuclear, respectivamente, contudo essa diferenciação ainda é considerada alta. Foi também analisada sua diferenciação com acessos dos dados publicados por Xu *et al.*, 2022, entretanto a diferenciação foi ainda maior ($F_{ST} = 0,256$). Esse acesso está relacionado aos indígenas que habitam tradicionalmente os lagos e várzeas situados ao longo do rio Purus, no estado do Amazonas (BONILLA, 2005).

Há representantes de todas as raças de milho brasileiras na coleção nuclear? A coleção nuclear apresenta 5,1% de acessos da coleção de raças Indígenas, 8,2% de raças Comerciais Antigas e 37,9% de raças Comerciais Recentes. Há ainda 48,8% da coleção sem classificação racial, evidenciando a lacuna de classificação ainda presente nesta coleção. Entre as raças Indígenas, há a representação de 8,5% dos acessos Caingang considerando o total na coleção indígena, 17% da raça Entrelaçado, 11% da raça Moroti e 20% da raça Guarani. Não há acessos da raça Lenha na coleção nuclear. Vale ressaltar que, apesar de ser uma raça pouco produtiva quando comparada às demais, a raça Lenha pode ser uma importante fonte para o melhoramento quando o objetivo for aumentar o número de fileiras de grãos, além de sua característica de precocidade no florescimento. Além disso, não é mais encontrada em cultivos, evidenciando a importância da inclusão desta raça na coleção nuclear, uma vez que quando cruzadas com raças brasileiras como Dente Paulista e Cateto, Lenha apresenta maiores índices de produtividade (TEIXEIRA *et al.*, 2021).

A coleção nuclear apresenta 23 acessos da raça Cateto e apenas 1 acesso da raça Cristal, raças Comerciais Antigas. Não há nenhum acesso representante da raça Canário de Ocho. A coleção apresenta, ainda, 81 acessos da raça Dente, 29 acessos da raça Semi-dentado e somente 1 acesso da raça Cravo, raças Comerciais Recentes. A coleção nuclear conta também com 33 acessos advindos do programa de melhoramento do CNPMS, além de programas de melhoramento de outras instituições, e 27 acessos decorrentes de introduções, 17 de origem Tropical e 10 de origem Temperada.

As raças Comerciais Antigas são geneticamente mais próximas das raças Indígenas ou das Raças Comerciais Recentes? As antigas raças comerciais são descritas por Paterniani e Goodman (1977) como, de certo modo, indígenas. No entanto, divergem consideravelmente do grupo de raças indígenas uma vez que passaram por mudanças desde o cultivo inicial pelo homem ocidental e, hoje, são amplamente cultivadas, razão pela qual são chamadas de comerciais. Embora seja esperado que estas raças sejam geneticamente mais próximas das raças Indígenas, dado que possuem origem indígena (TEIXEIRA, 2008), os acessos do BAG representantes das raças Comerciais Antigas apresentaram maior diferenciação genética com as raças Indígenas ($F_{ST} = 0,123$) do que com as Comerciais Recentes ($F_{ST} = 0,102$). Isso pode dever-se à ampla utilização, principalmente do milho Cateto, em programas de milho híbrido devido à sua alta heterose em cruzamentos com outras raças (PATERNIANI; NASS; DOS SANTOS, 2000), levando a uma maior proximidade com as raças Comerciais Recentes também utilizadas em programas de melhoramento. O maior distanciamento entre raças Indígenas e raças Comerciais Recentes ($F_{ST} = 0,136$) pode ser elucidado no dendrograma da classificação racial de todas as coleções (Figura 18), onde essas raças aparecem visivelmente separadas entre si, bem como com o grupo de acessos da espécie Teosinto.

As introduções tropicais e temperadas diferenciam-se entre si? As linhagens temperadas e tropicais dos dados de Xu *et al.* (2022) são geneticamente mais próximas das introduções temperadas e tropicais, respectivamente? O germoplasma de milho forma naturalmente dois grupos principais quanto à adaptação climática, tropical e temperado, com base nas características ambientais das áreas de plantio de milho no mundo. Esses grupos foram formados ao longo de séculos após a migração do milho para fora de seu centro tropical de origem, no México (YAN *et al.*, 2009). As introduções temperadas e tropicais apresentaram

moderada diferenciação entre si, com F_{ST} médio de 0,130. Como esperado, as introduções temperadas apresentaram menor diferenciação quando comparadas às linhagens temperadas ($F_{ST} = 0,085$) estudadas por Xu *et al.* (2022) do que com linhagens subtropicais ($F_{ST} = 0,096$) e tropicais ($F_{ST} = 0,101$), enquanto as introduções tropicais foram mais próximas às linhagens tropicais ($F_{ST} = 0,054$) e subtropicais ($F_{ST} = 0,056$) do que às linhagens temperadas ($F_{ST} = 0,751$). Vale apontar que os acessos de introduções tropicais apresentaram menor diferenciação entre si ($F_{ST} = 0,086$) do que os acessos de introduções temperadas ($F_{ST} = 0,126$). Esse padrão não é esperado, uma vez que há muito mais diversidade nas linhas tropicais de milho do que nas linhas temperadas (YAN *et al.*, 2009). Esse padrão pode refletir um efeito de amostragem, uma vez que 52,9% dos acessos tropicais tem a mesma origem, enquanto as introduções temperadas são oriundas de dez países diferentes.

A estrutura genética revelada pela análise genética pode ser explicada por algum fator morfológico, racial ou geográfico de classificação dos acessos das coleções? Nenhuma das classificações morfo-geográficas apresentou uma sobreposição evidente com a estrutura genética detectada pela análise de marcadores moleculares. Entre todas as classificações avaliadas, a que mais demonstrou alguma sobreposição mais evidente foi quanto a raça ao separar acessos de raças indígenas, raças Comerciais Recentes e da espécie de Teosinto (Figura 17). Não houve agrupamento significativo das raças Comerciais Antigas. Divisão parecida pode ser observada nos grupos obtidos pelo k-means, uma vez que o grupo 1 é formado por 62,4% de acessos de raças Comerciais Recentes (Figura 6 – E), os grupos 2, 3 e 6 apresentaram, respectivamente, 53,3%, 47,8% e 85% de acessos de origem indígena (Figura 7 – B, Figura 8 – E, Figura 10 – B) e os acessos de Teosinto foram todos agrupados no grupo 6 (Figura 11 – B).

A classificação racial considerando apenas a coleção indígena também mostrou algum nível de consistência com os agrupamentos genéticos, porém para algumas raças apenas (Figura 18), apontando a formação de grupos da raça Lenha e Caingang e uma maior mistura de acessos das raças Moroti e Entrelaçado. Vale ressaltar que o grupo 2 da análise k-means, com 53,3% de acessos de raças indígenas apresentou, entre eles, 20% de acessos da raça Entrelaçado e 26,7% da raça Moroti (Figura 7 – B), o grupo 3 foi formado por 47,8% de acessos de origem indígena, sendo desses, 54% da raça Lenha (Figura 8 – E).

No dendrograma com sobreposição da informação de origem geográfica também houve a formação de alguns grupos bem definidos como aquele da região Centro-Oeste e Norte, um grupo mais misturado, contudo com mais acessos advindos da região Sul, e por fim, grupos

bem pequenos da região Nordeste (Figura 19). No agrupamento dado pelo k-means foram observados grupos com uma porcentagem maior de determinadas regiões: o grupo 1 apresentou 43,8% de acessos advindos da região Nordeste (Figura 6 – G), o grupo 2, 73,3% da região Norte (Figura 7 – F), o grupo 3, 51,9% da região Sul (Figura 8 – G) e, por fim, o grupo 4 com 48,4% de acessos coletados na região Centro-Oeste (Figura 9 – F).

A sobreposição da informação do tipo de grão revelou dois grandes grupos, sendo um de grãos do tipo amiláceo e outro de grãos semi-dentado e dentado. É possível notar, também, um grupo menor com acessos do tipo de grão pipoca (Figura 20). No grupo 1, os tipos de grãos semi-dentado e dentado representaram 45% dos acessos (Figura 6 – F), no grupo 2, 80% dos grãos foram do tipo pipoca (Figura 7 – C), no grupo 3, 68,1% do tipo amiláceo e no grupo 6, 55,6% do tipo pipoca.

O tipo de endosperma dos grãos está mais ligado à raça ou à região geográfica dos acessos? Goodman (1995) associa os diferentes tipos de endospermas às diferentes etapas do processo de domesticação do milho. É sugerido que o milho mais antigo era a pipoca, com os outros tipos de endosperma, como duro, farináceo, dentado e doce, aparecendo mais tarde. Os endospermas duro e farinhento poderiam representar o segundo e terceiro estágio da domesticação, respectivamente, e por fim, o endosperma dentado seria um tipo de milho mais moderno. Acredita-se que os diferentes tipos de endosperma possam ter se originado independentemente em épocas e locais diferentes (COSTA *et al.*, 2021). As raças representadas nas coleções sempre apresentaram mais de 50% de um mesmo tipo de endosperma. Acessos da raça Caingang, Entrelaçado e Moroti apresentaram 82,3%, 87% e 86% de acessos com grãos do tipo amiláceo, respectivamente. A raça Lenha apresentou 86,6% dos grãos do tipo dentado/semi-dentado, enquanto as raças Comerciais Antigas apresentaram 86,3% de grãos duro/semi-duro, e 92% dos acessos das raças Comerciais Recentes são do tipo dentado/semi-dentado. Vale a pena lembrar que esse dado não está disponível para 100% dos acessos. Ao se investigar o tipo de endosperma dentro das regiões geográficas, apenas a região Centro-Oeste apresentou mais de 50% dos acessos com o mesmo tipo de grão (75% dos acessos com endosperma amiláceo). Acessos das demais regiões do país não apresentaram nenhum padrão predominante de tipo de grão dos acessos. Essa falta de padrão predominante se torna ainda maior ao se considerar o tipo de região ecogeográfica.

5. CONCLUSÃO E PERSPECTIVAS

Em conclusão, as coleções estudadas do Banco Ativo de Germoplasma de Milho da EMBRAPA apresentaram ampla variabilidade genética seja dentro bem como entre as principais categorias das diferentes classificações morfo-geográficas utilizadas na sua caracterização. As coleções nuclear e indígena apresentaram diferenciação genética moderada, com um F_{ST} médio de 0,136. Análises de estrutura genética populacional revelaram a existência de grupos genéticos distintos nos quais observa-se predominância de acessos da coleção nuclear em um e de acessos indígenas no outro. A coleção indígena apresentou maior diferenciação genética entre os acessos em comparação à coleção nuclear o que sugere a necessidade de uma atualização da composição da coleção nuclear ao incluir maior número de acessos indígenas e assim ampliar a diversidade genética representada no germoplasma disponível da espécie. Padrões de estrutura genética, conforme definidos pelos dados do SNP, foram encontrados com sobreposição parcial com classificações dos acessos de germoplasma usados no estabelecimento das coleções, tais como tipos de grãos de milho, nível de domesticação e origem do material coletado. No entanto, algumas discrepâncias importantes também foram detectadas onde tais classificações não correspondiam ao que os dados de DNA mostraram. Esses resultados indicam que informações genômicas não enviesadas derivadas dos dados do SNP fornecem informações adicionais importantes que refletem de forma mais precisa a origem ancestral dos acessos de germoplasma.

Os resultados encontrados nesse trabalho reforçam a importância do BAG de Milho, ampliando o conhecimento dos recursos genéticos conservados dentro das coleções de germoplasma e gerando novas informações sobre sua composição e sobre sua estrutura e diversidade genética. A presente avaliação da diversidade genética permitiu a investigação de similaridades e diferenças entre os acessos dentro das raças bem como de outros descritores e distribuição geográfica. Permitiu também a caracterização dos acessos sem classificação racial, dessa maneira, agregando, ainda mais, valor a esses recursos. Propomos, por fim, que os dados genômicos gerados neste trabalho sejam prontamente integrados aos dados de passaporte disponíveis dos acessos estudados para chegar a uma representação mais precisa do germoplasma de milho conservado nas duas coleções.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- ABADIE, T. *et al.* **A coleção nuclear de germoplasma de milho para o Brasil**. Brasília: Embrapa Recursos Genéticos e Biotecnologia, 2000. 37p. (Embrapa Recursos Genéticos e Biotecnologia. Boletim de Pesquisa, 8). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/179927>>. Acesso em: 6 jul. 2024.
- ANGLIN, N. L. *et al.* Genotyping Genebank Collections: Strategic Approaches and Considerations for Optimal Collection Management. **Plants**, v. 14, n. 2, p. 252, 2025.
- ARCA, M. *et al.* Genome-wide SNP genotyping of DNA pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool. **BioRxiv**, 2020.
- BONILLA, O. Cosmologia e organização social dos Paumari do médio Purus (Amazonas). **Revista de Estudos e Pesquisas**, v. 2, n. 1, p. 7-60, 2005.
- CGIAR. CONSULTATIVE GROUP ON INTERNATIONAL AGRICULTURAL RESEARCH. **Global Strategy for the Ex-situ Conservation and Utilization of Maize Germplasm**. Washington : CGIAR, 2007. Disponível em: <<https://www.genebanks.org/resources/publications/>>. Acesso em: 08 jul. 2024.
- CONTINI, E. *et al.* **Milho: caracterização e desafios tecnológicos**. Brasília, DF: Embrapa; Sete Lagoas: Embrapa Milho e Sorgo, 2019. 45 p. (Desafios do Agronegócio Brasileiro, 2). Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/195075/1/Milho-caracterizacao.pdf>>. Acesso em: 20 jul. 2024.
- COSTA, F. M. *et al.* Maize dispersal patterns associated with different types of endosperm and migration of indigenous groups in lowland South America. **Annals of Botany**, v. 129, n. 6, p. 737-751, 2022.
- DUBREUIL, P. *et al.* Evaluation of a DNA pooled-sampling strategy for estimating the RFLP diversity of maize populations. **Plant Molecular Biology Reporter**, v. 17, n. 2, p. 123-138, 1999.
- DEMÉTRIO, C. S. *et al.* Desempenho de híbridos de milho submetidos a diferentes espaçamentos e densidades populacionais. **Pesquisa Agropecuária Brasileira**, v. 43, n. 1, p. 1691-1697, 2008.
- EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular ecology**, v. 14, n. 8, p. 2611-2620, 2005.
- FALUSH, D.; STEPHENS, M.; PRITCHARD, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. **Genetics**, v. 164, n. 4, p. 1567-1587, 2003.
- FRANCIS, R. pophelper: an R package and web app to analyse and visualize population structure. **Molecular ecology resources**, v. 17, n. 1, p. 27-32, 2017.

FREITAS, I. L. J. **Fitotecnia: Soja e Milho**. Londrina: Educacional, 2019. 192 p. Disponível em: <https://cm-cls-content.s3.amazonaws.com/201901/INTERATIVAS_2_0/FITOTECNIA_SOJA_E_MILHO/U1/LIVRO_UNICO.pdf>. Acesso em: 6 jul. 2024.

GANAL, M. W. *et al.* A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. **PloS one**, v. 6, n. 12, p. e28334, 2011.

GOODMAN, M. Maize. In: SMARTT, J.; SIMMONDS, N. (Ed.). Evolution of crop plants. Logman: London, 1998. p. 192-202.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern recognition letters**, v. 31, n. 8, p. 651-666, 2010.

KUMAR, B. *et al.* Genetic diversity, population structure and linkage disequilibrium analyses in tropical maize using genotyping by sequencing. **Plants**, v. 11, n. 6, p. 799, 2022.

MCCOUCH, S. R. *et al.* Genomics of gene banks: a case study in rice. **American journal of botany**, v. 99, n. 2, p. 407-423, 2012.

MCCOUCH, S. R. *et al.* Mobilizing crop biodiversity. **Molecular plant**, v. 13, n. 10, p. 1341-1344, 2020.

NEI, M. Genetic distance between populations. **The American Naturalist**, v. 106, n. 949, p. 283-292, 1972.

PATERNIANI, E.; GOODMAN, M. **Races of maize in Brazil and adjacent areas**. CIMMYT, 1977. 95 p.

PATERNIANI, E.; NASS, L. L.; DOS SANTOS, M. X. O valor dos recursos genéticos de milho para o Brasil. In: : UDRY, C. V.; DUARTE, W. (Org.) **Uma história brasileira do milho - o valor dos recursos genéticos**. Brasília: Paralelo 15, 2000. p. 11-42.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945-959, 2000.

PRIYANKA, V. *et al.* Germplasm conservation: instrumental in agricultural biodiversity - a review. **Sustainability**, v. 13, n. 12, p. 6743, 2021.

R CORE TEAM: A language and environment for statistical computing, R Foundation for Statistical. **Computing**, 2020. Disponível em <<https://www.R-project.org/>>. Acesso em: 03 ago. 2024.

SEMAGN, K. *et al.* Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. **BMC genomics**, v. 13, p. 1-11, 2012.

SILVA, K. J. *et al.* High-density SNP-based genetic diversity and heterotic patterns of tropical maize breeding lines. **Crop Science**, v. 60, n. 2, p. 779-787, 2020.

TAKEZAKI, N.; NEI, M. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. **Genetics**, v. 144, n. 1, p. 389-399, 1996.

TEIXEIRA, F. F. **Milho cultivado no Brasil e banco de germoplasma - uma forma de classificação da variabilidade genética**. Sete Lagoas: Embrapa Milho e Sorgo, 2008. 11 p. (Embrapa Milho e Sorgo. Comunicado Técnico, 155). Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/CNPMS-2009-09/21382/1/Com_155.pdf>. Acesso em: 20 jul. 2024.

TEIXEIRA, F. F.; COSTA, F. M. **Caracterização de recursos genéticos de milho**. Sete Lagoas: Embrapa Milho e Sorgo, 2010. 10 p. (Embrapa Milho e Sorgo. Comunicado Técnico, 185). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/883797>> Acesso em: 27 jun. 2024.

TEIXEIRA, F. F. *et al.* **Diversidade genética entre acessos do Banco Ativo de Germoplasma de Milho com grãos brancos do tipo dentado**. Sete Lagoas: Embrapa Milho e Sorgo, 2019. 67 p. (Embrapa Milho e Sorgo. Documentos, 233). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1111013>>. Acesso em: 27 jun. 2024.

TEIXEIRA, F. F. *et al.* **Diversidade genética entre acessos do Banco de Germoplasma de Milho de origem indígena**. Sete Lagoas: Embrapa Milho e Sorgo, 2021. 133 p. (Embrapa Milho e Sorgo. Documentos, 259). Disponível em: <<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1131481>>. Acesso em: 27 jun. 2024.

VIGOUROUX, Y. *et al.* Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. **American journal of botany**, v. 95, n. 10, p. 1240-1253, 2008.

WALTERS, C. Principles for preserving germplasm in gene banks. *In*: GUERRANT, E.; HAVENS, K.; MAUNDER, M. (Ed.). **Ex situ plant conservation: supporting species survival in the wild**. Washington: Island Press, 2004 p. 442-453.

WRIGHT, S. **Evolution and Genetics of Populations**. Chicago: University of Chicago, 1978. 511p

XU, G. *et al.* Population genomics of Zea species identifies selection signatures during maize domestication and adaptation. **BMC plant biology**, v. 22, n. 1, p. 72, 2022.

YAN, J. *et al.* Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. **PloS one**, v. 4, n. 12, p. e8451, 2009.

ANEXO 1 - PROTOCOLO DE EXTRAÇÃO DE DNA DE PLANTAS COM 3X CTAB

- 1- Colocar a amostra de tecido em microtubo de 1.5 ml ou 2.2 ml com tampa de rosquear, junto com 8-10 esferas de aço de 2 mm.
- 2- Congelar as amostras junto com o bloco do macerador no -80°C por ~ 30 minutos (material seco de herbário ou liofilizado pode ser macerado na temperatura ambiente).
- 3- Macerar no BIOSPEC (tempo variável de 30 a 90 segundos de acordo com tecido com ou sem virar as posições de tubos).
- 4- Adicionar 2 ml de β -mercaptoetanol para cada 100 ml de tampão de extração CTAB 3X;
- 5- Adicionar 700 μ l de tampão de extração CTAB 3X e 5 μ l de RNase (10 mg/ml) ao pellet de tecido e bater +- 5 segundos no BIOSPEC para misturar o tecido ao tampão.
- 6- Incubar os microtubos no banho-maria a 65°C durante 30 - 45 minutos (não mais do que isso para evitar oxidação excessiva);
- 7- Retirar do banho-maria e deixar esfriar os microtubos por alguns minutos.
- 8- Acrescentar 750 μ l de clorofórmio: álcool isoamílico (24:1) ao sobrenadante e agitar por inversão durante 1 minuto;
- 9- Centrifugar a 10.000 rpm por 15 minutos;
- 10- Recuperar 500-600 μ l de sobrenadante sem perturbar a interface e transferir para outro microtubo;
- 11- Acrescentar 1/10 do volume (40-50 μ l de acetato de sódio 3 M e 2/3 do volume (270 - 330 μ l) de isopropanol gelado e manter a -20°C por, no mínimo, 1 hora ou preferencialmente overnight.
- 12- Centrifugar a 14.000 rpm por 20 minutos.
- 13- Descartar o sobrenadante e lavar o *pellet* com 1 ml de etanol 70% uma vez; centrifugar por 14.000 rpm por 1 minuto.
- 14- Descartar o etanol 70% (cuidando para não perder o pellet) e secar na *speedvac* (*eppendorf dryer*) por 10 minutos.
- 15- Acrescentar 50 a 100 μ l de tampão TE para dissolver a pellet.

ANEXO 2 - COMPARAÇÃO DA DIFERENCIAÇÃO GENÉTICA ENTRE ACESSOS DE ORIGEM INDÍGENA NÃO CLASSIFICADOS RAÇAS COM ACESSOS CLASSIFICADOS EM RAÇAS INDÍGENAS E ACESSOS DA COLEÇÃO NUCLEAR.

Tabela S1 – Comparação da diferenciação genética (F_{ST}) entre acessos de origem indígena não classificados em raças com acessos classificados em raças indígenas e acessos da coleção nuclear. Quanto mais verde, menor é a diferenciação genética entre o acesso e a raça/coleção.

Acesso	Raça/Coleção					
	Caingang	Entrelaçado	Guarani	Lenha	Moroti	BRA
AC-040	0,109	0,113	0,101	0,131	0,116	0,115
AC-045	0,174	0,136	0,178	0,226	0,157	0,193
AM-009	0,176	0,172	0,164	0,188	0,176	0,160
BA-168	0,121	0,129	0,100	0,111	0,128	0,084
COMPOSTO-INDIGENA	0,090	0,092	0,093	0,133	0,095	0,115
GO-001	0,102	0,092	0,097	0,143	0,102	0,120
GO-002	0,106	0,094	0,099	0,138	0,104	0,113
I-74-6-PI-KATETE	0,108	0,121	0,084	0,108	0,119	0,089
INDIOS-PAUMARI	0,208	0,224	0,198	0,229	0,221	0,212
KAYABI-1-AMARELO	0,136	0,123	0,128	0,159	0,131	0,133
KAYABI-2-PRETO	0,142	0,126	0,135	0,167	0,136	0,142
MG-058	0,125	0,136	0,115	0,130	0,135	0,114
MG-059	0,122	0,132	0,103	0,120	0,130	0,095
MT-009	0,129	0,137	0,111	0,124	0,137	0,099
MT-059	0,132	0,106	0,137	0,180	0,121	0,155
MT-060	0,179	0,187	0,155	0,181	0,186	0,161
RO-010	0,119	0,102	0,114	0,154	0,113	0,230
RO-012	0,150	0,116	0,154	0,203	0,134	0,272
RO-013	0,150	0,118	0,155	0,203	0,135	0,171
RO-014	0,108	0,103	0,097	0,129	0,110	0,107
RO-015	0,140	0,108	0,143	0,191	0,125	0,159
RO-028	0,163	0,128	0,166	0,211	0,147	0,182
TO-002-PONHY-PEY-PRETO	0,145	0,134	0,140	0,173	0,140	0,145