UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS CORNÉLIO PROCÓPIO
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ANDRESSA CAROLINE PATERA

ANÁLISE COMPARATIVA DE ESTRATÉGIAS DE IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO EM *Gossypium hirsutum*

DISSERTAÇÃO - MESTRADO

CORNÉLIO PROCÓPIO
2023

ANDRESSA CAROLINE PATERA

ANÁLISE COMPARATIVA DE ESTRATÉGIAS DE IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO EM *Gossypium hirsutum*

COMPARATIVE ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISM IDENTIFICATION STRATEGIES IN *Gossypium hirsutum*

Dissertação apresentada como requisito para obtenção do título de Mestre em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Prof. Dr. Douglas Silva Domingues.
Coorientador: Prof. Dr. Alexandre Rossi Paschoal.

CORNÉLIO PROCÓPIO

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

ANDRESSA CAROLINE PATERA

ANÁLISE COMPARATIVA DE ESTRATÉGIAS DE IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO EM GOSSYPIUM HIRSUTUM

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 18 de Agosto de 2023

Dr. Douglas Silva Domingues, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Alexandre Rossi Paschoal, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Liliane Santana Oliveira Kashiwabara, Doutorado - Embrapa Soja

Dr. Luiz Filipe Protasio Pereira, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 18/08/2023.

AGRADECIMENTOS

Agradeço ao meu orientador Prof. Dr. Douglas Silva Domingues e ao meu coorientador Prof. Dr. Alexandre Rossi Paschoal, pela sabedoria com que me guiaram nesta trajetória.

A Tropical Melhoramento e Genética (TMG) pelo aporte a realização desta pesquisa.

Ao Daniel Longhi Fernandes Pedro por todo o apoio com as análises de bioinformática e pelo compartilhamento de conhecimentos.

A Secretaria do Curso, pela cooperação.

Gostaria de deixar registrado também, o meu reconhecimento à minha família e a todos os que por algum motivo contribuíram para a realização desta pesquisa.

RESUMO

PATERA, Andressa Caroline. **Análise Comparativa de Estratégias de Identificação de Polimorfismos de Nucleotídeo Único em *Gossypium hirsutum*.** 2023. Dissertação (Mestrado em Bioinformática). Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2023.

Plataformas de sequenciamento de alto rendimento viabilizam a geração de enorme quantidade de dados de forma extremamente rápida. No entanto, metodologias de sequenciamento são altamente sensíveis a erros, tornando o processo de obtenção de dados altamente dependente de ferramentas de bioinformática. A identificação de variantes alélicas é um importante desafio no processamento de dados de sequenciamento, o qual inclui o alinhamento das sequências com o genoma de referência da espécie alvo. A diferença nas variantes genéticas, obtida através de várias abordagens de identificação de polimorfismos, pode causar impactos diretos no uso desses dados em estudos genéticos. Esses impactos podem ser observados em áreas como o mapeamento associativo e a seleção genômica. O presente estudo teve como objetivo comparar duas abordagens de detecção de variantes (Fast-GBS e BWA/BFCTools) para determinar o seu impacto na identificação de nucleotídeos de polimorfismo único (SNPs) em um painel de 250 genótipos de algodão (*Gossypium hirsutum),* dos quais 72 correspondem a genótipos do banco de germoplasma da TMG (sequências *single-end* obtidas por GBS com sequenciamento *Ion Torrent*) e outros 178 são provenientes de um estudo na literatura (sequências *paired-end* obtidas por sequenciamento Illumina). Os resultados foram comparados através do levantamento do número total de SNPs recuperados, bem como o número de SNPs recuperados por cromossomo. Outras métricas utilizadas foram o SNP-Score (capaz de ponderar o número de ocorrências de SNPs por *pipeline* de chamada de alelos), tempo computacional e análise de componentes principais. O *pipeline* Fast-GBS recuperou um total de 417.975 SNPs para o subconjunto de dados brutos da TMG e 38.685.370 SNPs para o subconjunto de dados brutos da literatura enquanto o *pipeline* BWA/BCFTools recuperou um total de 254.805 SNPs para o subconjunto de dados brutos da TMG e 38.685.377 SNPs para o subconjunto de dados brutos da literatura. Podemos identificar que existem 24.402 SNPs em comum em todos os conjuntos de dados quando o *pipeline* BWA/BCFTools foi utilizado e 15.348 SNPs em comum entre todos os conjuntos de dados quando o *pipeline* Fast-GBS foi utilizado. Ao final das análises, foi possível concluir que o *pipeline* Fast-GBS possui um melhor desempenho computacional e que sequências *paired-end* sofrem pouca influência do *software* utilizado para chamada de alelos, devido à sua elevada precisão. Para sequências *single-end*, o pipeline Fast-GBS obteve melhor desempenho para dados brutos de sequenciamento e o BWA/BCFTools obteve melhor desempenho com dados filtrados. Os resultados obtidos reforçam a necessidade de considerar vários aspectos durante a escolha dos métodos para análise.

Palavras-chave: Sequênciamento de alto rendimento, variantes alélicas, Polimorfismo de nucleotídeo único, ferramentas de bioinformática

ABSTRACT

PATERA, Andressa Caroline. **Comparative Analysis of Single Nucleotide Polymorphism Identification Strategies in *Gossypium hirsutum***. 2023. Dissertation (Master's Degree in Bioinformatics). Federal Technological University of Paraná. Cornélio Procópio, 2023.

High-throughput sequencing platforms make it possible to generate huge amounts of data extremely quickly. However, sequencing methodologies are highly sensitive to errors, making the process of obtaining data highly dependent on bioinformatics tools. The challenge of identifying allelic variants in the processing of sequencing data encompasses the alignment of sequences with the reference genome of the target species. Variations in genetic variants, acquired through diverse approaches to polymorphism identification, may impart direct impacts upon the utilization of such data in genetic studies, including associative mapping and genomic selection.The present study aimed to compare two variant calling approaches (Fast-GBS and BWA/BFCTools) to determine their impact on the identification of single polymorphism nucleotides (SNPs) in a panel of 250 cotton (*Gossypium hirsutum*) genotypes of which 72 correspond to genotypes from TMG's germplasm bank (single-end sequences obtained by GBS with Ion Torrent sequencing) and 178 come from a study in the literature (paired-end sequences obtained by Illumina sequencing). The results were compared by surveying the total number of SNPs recovered, as well as the number of SNPs recovered per chromosome. Other metrics used were the SNP-Score (capable of weighting the number of SNP occurrences per allele calling pipeline), computational time and principal component analysis. The Fast-GBS pipeline retrieved a total of 417,975 SNPs for the TMG raw data subset and 38,685,370 SNPs for the literature raw data subset. We can identify that there are 24,402 SNPs in common across all datasets when the BWA/BCFTools pipeline was used and 15,348 SNPs in common across all datasets when the Fast-GBS pipeline used. At the end of these analyses, it was possible to conclude that the Fast-GBS pipeline has a better computational performance and that paired-end sequences suffer little influence from the software used to call alleles, due to its high precision. For single-end sequences, Fast-GBS pipeline performed better for raw sequencing data and BWA/BCFTools performed better with filtered data. The results obtained reinforce the need to consider several aspects when choosing methods for analysis.

Keywords: High-throughput sequencing, Allelic variants, Single nucleotide polymorphisms (SNPs), Bioinformatics tools

LISTA DE ANEXOS

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS

HTS	Sequenciamento de alto rendimento
SNP	Polimorfismo de Nucleotídeo Único
INDELS	Inserções e Deleções
GWAS	Estudo de Mapeamento Associativo
GP	Predição Genômica
MAS	Seleção Assistida por Marcadores Moleculares
WGS	Sequenciamento de Genoma Completo
Loci	Plural de *locus* – Posição que o gene ocupa no cromossomo
DNA	Ácido Desoxirribonucleico
GBS	Genotipagem por sequenciamento
dNTPs	Deoxinucleotídeos
ddNTPs	Dideoxinucleotídeos
Primers	Iniciadores utilizados na reação de PCR
pb	Pares de Bases
QTL	Lócus de característica quantitativa
Mb	Mega pares de bases (1.000.000 pares de base)
Read	Sequencia inferida de pares de bases correspondente a um fragmento de DNA
PCR	Reação em Cadeia da Polimerase
pH	Potencial hidrogeniônico
pHFET	Transistor de efeito de campo sensível ao pH
RNA	Ácido Ribonucleico
NCBI	Centro Nacional de Informações Biológicas
MB	Megabyte
GB	Gigabyte
Phred	Coeficiente de qualidade, indica a confiabilidade dos resultados de sequenciamento para cada uma das bases que compões uma *read*.
CNV	Variação de Número de Cópias Comum
SAM	Mapa de Alinhamento de Sequencias (do inglês *Sequence Alignment Map*)
BAM	Contra-parte Binária do SAM (Formato binário ordenado)
BCF	Formatação Binária de Variantes

VCF	Formato de Chamada de Variantes (do inglês, *variant call format*)
MAF	Frequência do Alelo Menor (do inglês, *Minor Allele Frequency*)
PCA	Análise de Componentes Principais

SUMÁRIO

1 INTRODUÇÃO

A elevada disponibilidade de dados genotípicos, obtidos por meio da tecnologia de sequenciamento de alto rendimento (em inglês, *high throughput sequencing*, HTS), permite a realização de estudos por diversas abordagens, entre as quais se destacam os estudos de mapeamento associativo (em inglês, *genome-wide association studies,* GWAS) e a predição genômica (em inglês, *genomic prediction,* GP), amplamente utilizados para seleção de indivíduos portadores de alelos determinantes de fenótipos desejáveis no melhoramento genético vegetal. A eficácia destes estudos depende, além dos modelos matemáticos que representem adequadamente o modelo genotípico e de dados fenotípicos de elevada qualidade, da quantidade e da qualidade das variantes genéticas empregadas na análise.

Dentre as abordagens comumente utilizadas para identificação de variantes genéticas estão o sequenciamento de genoma completo (do inglês, *whole genome sequencing,* WGS) e a genotipagem por sequenciamento (do inglês, *genotyping by sequencing,* GBS), que utiliza enzimas de restrição específicas, com a finalidade de reduzir a complexidade do genoma e aumentar a probabilidade de amostragem de regiões gênicas.

Após o sequenciamento, são obtidas elevadas quantidades de sequências curtas (*reads*), geradas por sequenciamento para uma mesma região genômica. Uma tarefa primordial para a identificação de variantes genéticas neste cenário é o alinhamento das sequências curtas com o genoma de referência da espécie alvo. Após o alinhamento, se torna possível determinar a presença de variantes genômicas em relação a uma referência. Vários programas para determinações de variantes genéticas estão atualmente disponíveis, tais como FreeBayes, GATK, Platypus, Samtools/mpileup, SNVer, VarScan, dentre outros. No entanto, existe discordância entre as variantes determinadas por cada algoritmo de chamada de alelos.

A desuniformidade na identificação das variantes genéticas obtidas entre procedimentos de genotipagem e de chamada de variantes levam a diferenças que podem impactar diretamente na qualidade e na quantidade de dados utilizados em estudos genéticos (como por exemplo, no mapeamento associativo e na predição genômica). Fatores como: 1) a abordagem do algoritmo de alinhamento de sequências a um genoma de referência, 2) estratégias de detecção das variantes e 3)

características da própria espécie de estudo, como seus hábitos reprodutivos, origem ancestral, ploidia e diversidade entre genótipos, podem ter impacto nesse processo.

1.1 OBJETIVO GERAL

A presente dissertação teve como objetivo comparar duas estratégias de detecção de variantes para determinar o impacto desta etapa na identificação de SNPs em um painel de genótipos de algodão (*Gossypium hirsutum).*

2 REVISÃO BIBLIOGRÁFICA

2.1 SEQUENCIAMENTO DE ALTO RENDIMENTO (HTS)

As tecnologias de sequenciamento de alto rendimento desempenham papel fundamental na geração de dados de sequências genômicas para diversos organismos. O sequenciamento de Sanger, também conhecido como “método de terminação de cadeia”, é um método para determinar a sequência de nucleotídeos do DNA, que foi desenvolvido pelo ganhador do Prêmio Nobel Frederick Sanger e colaboradores em 1977 [1]. Automatizado na década de 1980, foi a principal técnica de sequenciamento de DNA até meados dos anos 2000.

Uma nova era nas tecnologias de sequenciamento surgiu com o advento do sequenciador 454 da Roche *Life Science* em 2005, abrindo novas perspectivas para a exploração e análise do genoma, garantindo elevado rendimento e menor custo do que as primeiras tecnologias disponíveis no mercado [1].

As novas tecnologias de sequenciamento de alto rendimento permitem a geração de bilhões de dados em paralelo em uma única corrida. Infelizmente estas abordagens ainda são incapazes de ler sequências completas de DNA, sendo limitadas a pequenas sequências de fragmentos, gerando milhões de leituras e elevada demanda computacional para montagem dos genomas [1].

As tecnologias de HTS podem ser divididas em abordagens distintas, como: a) tecnologias de sequenciamento caracterizadas pela necessidade de preparo bibliotecas de fragmentos antes de iniciar o sequenciamento de clones de DNA amplificados (como por exemplo, na tecnologia Illumina) [2]; b) tecnologias de sequenciamento de leituras longas, que são classificadas como tecnologia de sequenciamento de molécula única [3] (como por exemplo, tecnologia PacBio); e c) tecnologias de sequenciamento na qual uma molécula Nanopore produz sequências longas de leituras contíguas [4].

Na tecnologia Illumina o sequenciamento é baseado em síntese, onde fragmentos de DNA são ligados a adaptadores e amplificados em uma superfície sólida (denominada *flow cell*), gerando *clusters*. Em seguida, a leitura da sequência do fragmento de DNA é realizada por meio da detecção de uma molécula de fluorescência específica, atrelada a cada nucleotídeo presente na fita de DNA [5]. O sequenciamento PacBio é uma técnica baseada em sequenciamento por ligação de

pontes, onde os fragmentos de DNA são ligados a adaptadores e amplificados em esferas, gerando longas cadeias de DNA. Nessa técnica, as bases são adicionadas sequencialmente e detectadas por fluorescência, permitindo a leitura da sequência de bases em tempo real [7]. O sequenciamento Nanopore, por sua vez, é baseado na passagem de fragmentos de DNA por nanoporos muito pequenos, gerando alteração da corrente elétrica e permitindo a leitura da sequência de bases [6].

Além das tecnologias de sequenciamento mencionadas acima, outra forma de detecção dos nucleotídeos presentes em uma fita de DNA se dá por alterações no pH da solução durante a incorporação destes nucleotídeos na etapa de síntese, como ocorre na plataforma Ion Torrent, que será detalhada no item 2.2 [8]. Um breve comparativo entre as principais plataformas de HTS é apresentado na tabela 1.

Tabela 1. Comparação das plataformas de HTS: Illumina MiSeq, Illumina HiSeq 2000, Ion Torrent, PacBio e Oxford Nanopore. Adaptado de Vestergaard e colaboradores, 2021 [9].

Parâmetros	Illumina MiSeq	Illumina HiSeq 2000	Ion Torrent PGM	PacBio SMRT	Oxford Nanopore MinION
Comprimento de Leitura	≤300 pb	≤125 pb	≤600 pb	10.000 - 30.000 pb	10.000 - 30.000 pb
Imobilização	*Flow Cell*	*Flow Cell*	Emulsão em *Bead*	N/A	Enzima de Processamento
Amplificação	Amplificação em Ponte	Amplificação em Ponte	PCR em emulsão	N/A	N/A
Taxa de Erro Bruto	0,20%	0,20%	1,00%	10-15%	5-20%
Tempo de Corrida	4-55 h	7-144 h	2-7.5 h	Tempo Real	Tempo Real

A partir das estratégias de sequenciamento mencionadas acima é possível obter sequências *single-end* ou *paired-end*, que representam abordagens diferentes na geração de dados genômicos. O sequenciamento *single-end* envolve a leitura de apenas uma extremidade de cada fragmento de DNA enquanto o sequenciamento *paired-end* resulta na leitura de ambas as extremidades das sequências de cada fragmento. A escolha pela abordagem [27].

2.2 PLATAFORMA ION TORRENT

Ion Torrent é uma plataforma de HTS baseada no sequenciamento por detecção eletroquímica da síntese de DNA. Consiste na liberação de um hidrogênio (H+) no sítio 3’OH a cada incorporação de um novo nucleotídeo na sequência de DNA.

A metodologia emprega um *chip* semicondutor e substitui a marcação dos nucleotídeos com moléculas fluorescentes, estratégia aplicada em outras tecnologias de sequenciamento [10] (Figura 1).

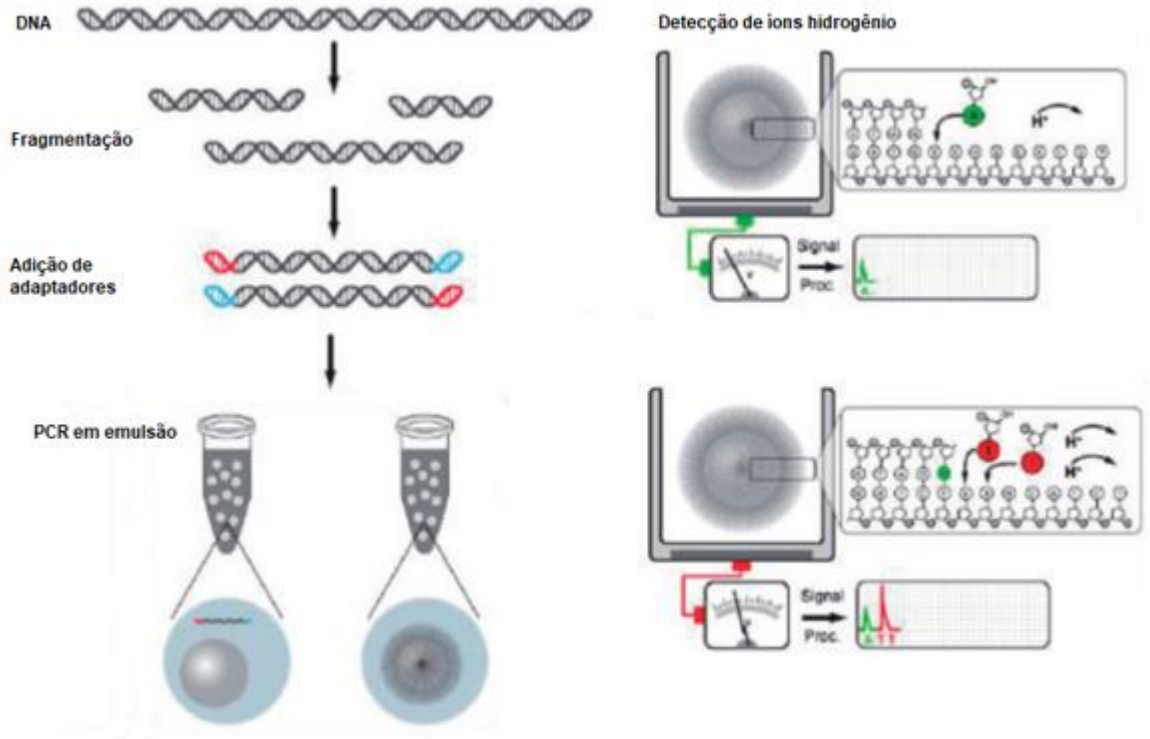


Figura 1. Representação da técnica de sequenciamento Ion Torrent. Adaptado de GOLAN e MEDVEDEV, 2013 [10].

O *chip* semicondutor contém milhões de poços, capazes de capturar informações químicas da molécula de DNA, demonstradas na informação de cada alelo presente na molécula. O processo de sequenciamento é iniciado quando uma molécula de DNA é fragmentada em milhões de pedaços. Estes são anexados por adaptadores em uma esfera metálica e amplificados por PCR em emulsão, aumentando seu número de cópias e cobrindo toda a superfície da esfera metálica. Este processo é reproduzido em todo o *chip*, que contém milhões de esferas com milhões de fragmentos diferentes de DNA. As esferas escoam pelo *chip* e cada uma delas é depositada em um dos poços do *chip*. O *chip* é submetido a adição de cada um dos nucleotídeos presentes no DNA (adenina, timina, citosina e guanina). Sempre que um nucleotídeo é incorporado à sequência de DNA um íon hidrogênio é liberado, a liberação deste íon altera o pH da solução daquele poço, possibilitando a sua detecção [10].

Transistores de H+, conhecidos como pHFET (transistor de efeito de campo sensível ao pH), são utilizados como sensores e organizados em uma matriz de sensores paralelos massivos em um CMOSchip. A concentração local de H+ cria uma voltagem positiva, resultado em uma alteração na corrente que passa pelo transistor. Para o sequenciamento, um sensor é implementado para servir de pHFET na porção inferior de um micro poço. A alteração da voltagem é gravada, indicando que o nucleotídeo foi incorporado. O processo é repetido a cada 15 segundos, ocorrendo em todos os poços do chip de forma sequencial [10].

2.3 GENOTIPAGEM BASEADA EM MÉTODOS DE SEQUENCIAMENTO

Polimorfismos de DNA podem ser utilizados na identificação e seleção de regiões responsáveis por controlar características (genes ou *loci* quantitativos - QTLs), e para fornecer percepções sobre a organização e evolução dos genomas, com possibilidade de aplicação para análise de diversidade em níveis intraespecíficos [11]. Tecnologias de HTS podem ser atreladas a metodologias de redução da complexidade dos genomas para garantir a detecção destes polimorfismos. Em plantas, a genotipagem por sequenciamento (GBS) tem sido a ferramenta principal para identificar e genotipar SNPs utilizando HTS.

Polimorfismos de nucleotídeo único, frequentemente chamados de SNPs, são as variações genéticas mais comumente encontradas nos genomas, caracterizadas pela diferença de um único nucleotídeo em uma sequência de DNA entre indivíduos, causada por mutações. Devido à sua abundância, estas variantes podem ser utilizadas como marcadores genéticos, desde que apresentem associação com uma determinada característica, sendo empregadas em seleção assistida por marcadores (do inglês, *marker-assisted selection,* MAS) [11].

A genotipagem por sequenciamento (GBS) é uma abordagem de genotipagem que se baseia no sequenciamento para descobrir simultaneamente as posições dos nucleotídeos polimórficos (SNPs) dentro de uma coleção de germoplasma. A metodologia é baseada em uma redução de complexidade do genoma, através do uso de enzimas de restrição específicas, possibilitando a inspeção de um subconjunto relativamente pequeno do genoma, ao invés de todos os *loci*. Esta metodologia fornece elevada capacidade de multiplexação de amostras individuais por meio do uso de códigos de barras específicos para cada genótipo [11].

Enzimas de restrição produzem extremidades coesivas de 2 a 3 pares de bases (pb) e não cortam frequentemente na maior fração repetitiva do genoma. A escolha correta pela enzima de restrição é uma etapa crítica, a partir desta escolha regiões repetitivas podem ser evitadas e regiões de baixas cópias podem ser direcionadas com mais eficiência. A porção sequenciada do genoma deve ser altamente consistente dentro de uma população, pois os locais de restrição são geralmente conservados entre as espécies. Isso torna a GBS um protocolo altamente adequado para experimentos que requerem levantamento de elevado número de marcadores dentro de uma população em estudo. Adicionalmente, GBS pode proporcionar maior cobertura de SNPs em regiões ricas em genes. Consequentemente esta abordagem é simples, altamente específica, com ótimo custo efetivo, reprodutível (especialmente se a profundidade do sequenciamento for alta), alta aplicabilidade, permitindo geração de elevada quantidade de polimorfismos [12].

As extremidades geradas pelas enzimas de restrição são complementares a uma extremidade de 3 pb dos adaptadores (código de barras). O desenho do adaptador também permite uma extremidade única ou emparelhada para sequenciamento. Amostras de DNA, *barcodes* e pares de adaptadores comuns são adicionados em uma placa com capacidade para 96 amostras. Na sequência, as amostras são digeridas com enzima de restrição específica e os adaptadores são ligados às extremidades dos fragmentos de DNA genômico. Iniciadores apropriados são adicionados e uma reação de PCR é realizada para aumentar o número de cópias dos fragmentos. Os produtos de PCR são purificados e os tamanhos dos fragmentos da biblioteca resultante são verificados em um analisador de DNA e após remoção dos dímeros e das bibliotecas sem adaptador os fragmentos são sequenciados [12].

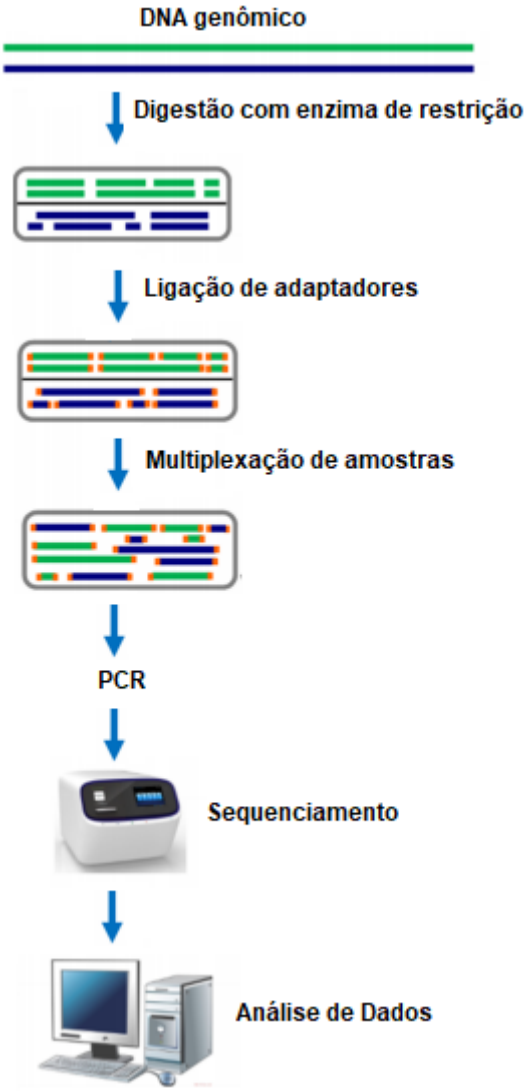


Figura 2. Representação da técnica de sequenciamento por genotipagem (GBS). Adaptado de LI e WANG, 2017 [13].

A análise dos dados de GBS pode se tornar complexa devido à natureza biológica da amostra. Considerando a natureza das amostras, a complexidade se deve ao: 1) número de variantes detectadas; 2) complexidade do genoma; 3) grau de heterozigosidade; 4) a porção de sequências repetitiva ao longo do genoma; 5) nível de polimorfismo e 6) divergência entre as populações. Neste sentido, é necessário considerar alguns fatores como: 1) grau de multiplexação das amostras; 2) número total de leituras por amostra e 3) taxa de erro de sequenciamento. Considerando estes desafios, fluxogramas complexos e bem desenhados de bioinformática são necessários para extrair SNPs das leituras de GBS [11].

Alguns desafios adicionais surgem para genomas poliploides, como o caso de *Gossypium hirsutum*, que possui dois subgenomas homólogos, A e D, resultante da fusão genômica interespecífica (alopoliploide) entre as espécies diploides *Gossypium arboreum* (AA) e *Gossypium raimondii* (DD). O genoma de *G. hirsutum* é notável pela sua complexidade (AADD = 2n = 4x = 52) e pelo tamanho considerável estimado em 2,25 a 2,43 Gb [28].

Genomas poliploides oferecem desafios para genotipagem. Isso ocorre, pois, alelos multi-loci são gerados e existe uma grande dificuldade em distinguir heterozigotos diploides (AB), para um ou ambos os pares de *loci* duplicados que amplificam quatro cópias de alelos (AA/AB ou AB/BB), duplo heterozigoto (AB/AB) ou homozigoto alternativo (AA/BB) (Figura 3). Este fato gera dificuldade para dosagem de alelos. Neste sentido, o uso de genomas de referência no alinhamento de sequências para determinação dos alelos é uma etapa fundamental. Apesar das limitações geradas pela complexidade dos genomas, o GBS têm sido amplamente utilizados para espécies poliploides, especialmente devido à possibilidade de redução da complexidade dos genomas com a aplicação desta técnica [14].

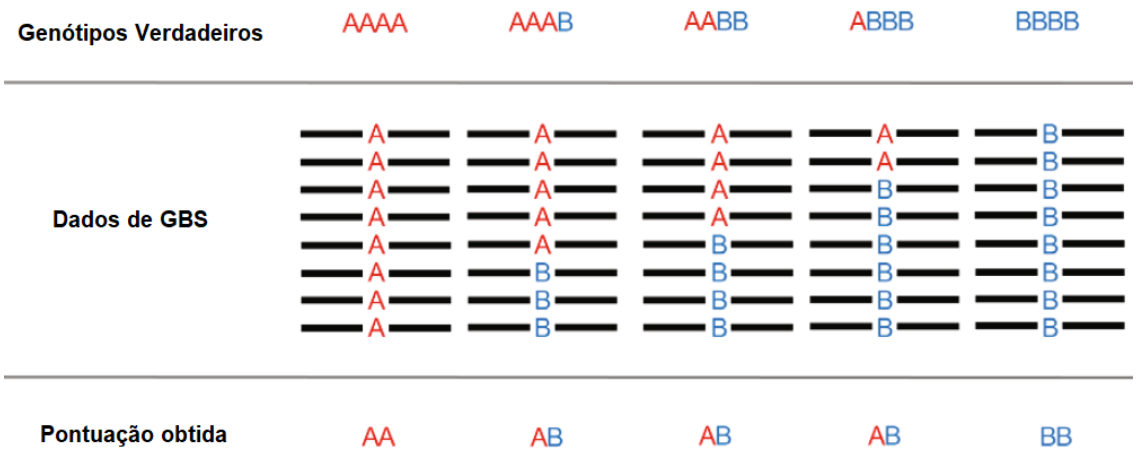


Figura 3. Comparação de genótipos verdadeiros com pontuação de cópia de alelos em um *loci* duplicado, obtidas por GBS. Adaptado de CHNUNG, *et al.* 2017. [14].

2.4 TRANSFORMAÇÃO DE BURROWS-WHEELER (BWA)

As duas maiores etapas de bioinformática para processamento de dados em larga escala de sequências de DNA produzidas pelo HTS correspondem ao alinhamento das sequências curtas com o genoma de referência da espécie alvo e a detecção de variantes genéticas. Além destas, também se torna necessário a

aplicação de filtros para remoção de bases de baixa qualidade e a imputação dos dados [15].

O alinhamento é uma técnica amplamente utilizada para comparar duas ou mais sequências, analisando uma série de caracteres individuais ou padrões de caracteres que estão na mesma ordem em ambas as sequências. Os algoritmos de alinhamento de sequências são comumente utilizados na bioinformática para o alinhamento de sequências de DNA, RNA e proteínas, de forma a identificar regiões de similaridade que possam ser uma consequência de um relacionamento funcional, estrutural ou evolutivo. Além disso, também podem ser utilizados para determinar a divergência entre as sequências, a fim de identificar variações que possam estar associadas a características de interesse econômico no contexto do melhoramento genético vegetal [16].

Em geral, a maioria das ferramentas de alinhamento de sequências curtas contra uma referência usam o alinhamento global associado a uma etapa prévia de “posicionamento global no genoma”. A indexação, primeira etapa do processo, corresponde a este posicionamento global, após a qual o processo de alinhamento real pode ser iniciado [16].

Na indexação, a maioria dos alinhadores constrói índices para a sequência de referência e/ou para o conjunto de dados de leitura (sequências curtas). Dois algoritmos de indexação principais, baseados em tentativas de prefixo e sufixo (Transformação de Burrows-Wheeler) ou tabelas de *hash* são incorporados na grande maioria dos algoritmos amplamente disponíveis [17]. Uma lista detalhada de estratégias de indexação foi recentemente apresentada por Alser e colaboradores. [15 e 16].

Após a indexação, ocorre um alinhamento global, que possui um sistema de pontuação composto por duas partes: a) matriz de pontuação e b) função de penalidade, que são atribuídas ao aparecimento de INDELs no alinhamento. O alinhamento ótimo entre duas sequências é aquele que maximiza a pontuação de alinhamento, apresentando pontuação máxima dentre todos os alinhamentos produzidos pelo algoritmo para um dado par de sequências, de acordo com um sistema de penalidades pré-estabelecido. O algoritmo de alinhamento global obtém o alinhamento ótimo por meio de uma matriz de similaridades e recuperação de alinhamento ótimo (*backtracking*). O sistema de pontuação associa uma pontuação alta sempre que as sequências possuem elevada similaridade [16].

A tabela 2 apresenta uma breve descrição da metodologia de alinhamento empregada pelo alinhador utilizado neste estudo.

Tabela 2. Metodologia de alinhamento alvo deste estudo.

Software	**Algoritmo**	**Descrição**	**Indexação**
BWA-MEM	Transformada de Burrows-Wheeler	Utiliza índice FMD para obter a indexação e realiza as buscas para ambos os sentidos da sequência, buscando alinhamentos com correspondências exatas máximas (MEM).	Índice FMD

O algoritmo de transformação de Burrows-Wheeler é um dos mais aplicados. A popularidade desta metodologia se deve ao fato de ser uma abordagem computacionalmente rápida, permitindo execução em computadores comuns. Dentre os *softwares* que empregam essa estratégia está o Bowtie2, que utiliza índice FM (do inglês, *full-text minute-space*) para indexar o genoma de referência e lançar uma sequência de consulta, com a finalidade de encontrar vários alinhamentos sem lacunas, que posteriormente serão estendidos [15].

O programa BWA é baseado na pesquisa em sentido contrário, associado à transformação de Burrows-Wheeler, sendo utilizado para alinhar leituras curtas a grandes sequências de referência com eficiência, permitindo lacunas e incompatibilidades. Em termos de otimização de memória a técnica de busca utilizada é semelhante à adotada pelo Bowtie2. BWA-MEM é o algoritmo mais recente desenvolvido pelo BWA para alinhamento de sequências, utilizando uma estrutura de índice FMD, na qual ambas as vertentes direta e reversa da sequência de DNA são indexadas. MEM (do inglês, *maximal exact matches)* é uma correspondência exata entre duas *strings* que não podem ser estendidas em nenhuma direção sem permitir lacunas [15].

2.5 CHAMADA DE ALELOS

Várias etapas de pós-alinhamento são utilizadas com o intuito de facilitar e melhorar a qualidade dos dados obtidos para as etapas subsequentes de análise. Dentre as tarefas mais comumente aplicadas estão a conversão do formato do arquivo de saída, criação de relatórios de processamento de alinhamento (número de leituras alinhadas, número de pares de leituras alinhadas corretamente e estatísticas resumidas), remoção de artefatos e duplicatas de PCR. Pacotes como SAMtools stats, SAMtools flagstat, SAMtools rmdup, UnifiedGenotyper do pacote Genome Analysis Toolkit, Qualimap, Picard CollectMultipleMetrics, GATK BaseRecalibrator e Picard Markduplicates são exemplos de ferramentas que podem ser utilizados para estas tarefas. Além da conversão de arquivos e redução do tamanho do conjunto de dados de saída, estes pacotes atuam em conjunto com os *softwares* utilizados para realizar a chamada de alelos. Estatísticas resumidas fornecidas por essas ferramentas são essenciais para avaliar a qualidade geral e a exatidão do alinhamento [17].

Processamentos adicionais específicos para chamada de alelos são recomendados antes da detecção das variantes genéticas. Uma das etapas mais importantes é a correção do artefato de alinhamento, que pode resultar na incompatibilidade de muitas bases próximas do local de incompatibilidade, podendo ser facilmente confundidas com variantes verdadeiras. Posteriormente é necessário realizar a recalibração da pontuação das estatísticas de qualidade da sequência [17].

Ferramentas como o GATK realizam realinhamento local, sendo projetadas para realinhar leituras nas proximidades de um artefato de alinhamento identificado para minimizar o número de bases incompatíveis, em geral grande quantidade de regiões genômicas requerem este tipo de realinhamento, devido à presença de INDELs no genoma da sequência em análise em relação ao genoma de referência. A ferramenta SAMtools utiliza uma abordagem conjunta para realizar a correção dos artefatos e recalibração. O *software* atribui uma pontuação de qualidade de alinhamento de base (BAQ) para cada base, que é calculada como probabilidade em escala de Phred. Vale reforçar que nem todas as ferramentas fornecem a opção de realinhamento [17].

Para a detecção de uma variante de forma confiável, é recomendada a utilização de alta profundidade e cobertura de sequenciamento, uma vez que elevado

número de leituras alinhadas a cada base é utilizado para diferenciar entre erros de sequenciamento e polimorfismos verdadeiros.

Os algoritmos de chamada de alelos se baseiam em métodos heurísticos ou probabilísticos. Métodos heurísticos chamam variantes com base em várias fontes de informação associadas à estrutura e qualidade dos dados (cobertura, qualidade de bases e frequência do alelo variante) [13]. Depois disso o teste exato e Fisher de contagens de leitura é aplicado e comparado com a distribuição esperada baseada apenas no erro do sequenciamento [4].

Os métodos probabilísticos, por sua vez, fornecem medidas de incerteza estatística para os genótipos, tornando possível monitorar a precisão da chamada dos genótipos. Além disso, informações adicionais de frequência de alelos e padrões de desequilíbrio de ligação podem ser inclusas na análise, adotando o teorema de Bayes. Ocorre um cálculo de probabilidade para cada genótipo possível em cada base (homozigoto para o alelo de referência, homozigoto para o alelo alternativo e heterozigoto). É baseado em pontuações de qualidade e contagem de alelos a partir das leituras no *loci* do SNP. Programas que incorporam este método é o SAMtools, SAMtools mpileup, GATK, Atlas-SNP2, SOAPsnp e SNVer [18]. Na tabela 3 são apresentadas as metodologias empregadas nos *softwares* que serão utilizados neste estudo.

Alguns estudos de *benchmarking* foram realizados na tentativa de identificar a melhor estratégia para combinação de ferramentas de alinhamento de sequências e chamada de variantes em dados gerados por HTS, conforme descrito na tabela 4. Considerando os estudos recentes apresentados na literatura, os *pipelines* escolhidos para a realização deste estudo empregam os *softwares* BCFTolls e Platypus (Tabela 3). O *software* BCFTools foi escolhido por ter sido utilizado para chamada de alelos com dados de sequenciamento Illumina de algodão, por Li e colaboradores (2021) [25]. O *software* Platypus, por sua vez, foi escolhido por ter sido disponibilizado como pipeline Fast-GBS v2.0, por Torkamaneh e colaboradores (2020) [22].

Tabela 3. Metodologias de chamada de alelos alvo deste estudo.

Software	**Descrição**
Bcftools	Utiliza comandos para gerar probabilidades de genótipo para cada posição genômica e posteriormente realiza a chamada real de alelos [25].
Platypus	Aplica local *de novo assembly*, seguido de realinhamento local e estimativa probabilística de haplótipos para identificação de SNPs, INDELs e polimorfismos complexos [24].

Tabela 4. Estudos comparativos de ferramentas de alinhamento e recuperação de polimorfismos

Abordagem de SNP *calling*	**Abordagem de Mapeamento**	**Resumo**	**Cultura/ Organismo**	**Métricas avaliadas**	**Ano**	**Referência**	**Observação**
FreeBayes, GATK, Platypus, Samtools/mpileup, SNVer, VarScan, VarDict	BWA-mem e Bowtie2	Avaliação de 7 ferramentas de chamada de SNPs com o uso de 2 ferramentas de mapeamento	Trigo (Alohexaplóide)	Taxa de mapeamento, acurácia, qualidade de mapeamento, ROC, AUC	2020	[19]	BWA-mem possuiu melhor acurácia e taxa de mapeamento, detectando maior número de variantes do que Bowtie2. Samtools/mpileup com BWA-mem superaram as outras abordagens testadas, seguidas de FreeBayes e GATK.
GATK e Samtools/mpileup	SOAP2, BWA-mem e Bowtie2	Avaliação de 2 ferramentas de chamada de SNPs com o uso de 3 ferramentas de mapeamento	Tomate e dados simulados de sequenciamento para diferentes culturas	Porcentagem de Mapeamento, acurácia de alinhamento, precisão, recall e tempo de processamento	2019	[20]	BWA-mem alinhou mais *reads* com elevada acurácia enquanto Bowtie2 teve a maior precisão em geral. Os softwares de chamadas de alelo afetaram a precisão e o recall de acordo com níveis de cobertura, diversidade e complexidade do genoma.
SAMtools, GATK, CTAT, FreeBayes, MuTect2, Strelka2 e VarScan2	Splace-aware	Avaliação de 7 ferramentas de chamada de SNVs	Dados reais e simulados de RNA-seq de célula única.	NA	2019	[18]	SAMtools apresentou a maior sensibilidade e FreeBayes demonstra bom desempenho em casos de altas frequências de variantes alélicas.

SAMTools, BCFTools, CLC-caller, FreeBayes, GATK, LoFreq, SNVer, VarDict, VarScan	BWA-mem, Bowtie2, CLC-mapper, GEM3, Novoalign	Avaliação de 50 pipelines de SNP *calling*	*Arabidopsis thaliana*	Sensibilidade e especificidade	2020	[21]	Todos os pipelines são adequados para uso em dados de NGS de plantas. As melhores ferramentas para alinhamento foram BWA-MEM e Novoalign e GATK foi o melhor software para chamada de alelos.
Platypus	BWA-mem	Pipeline Fast-GBS (v2.0)	NA	Número de variantes, porcentagem de dados perdidos, porcentagem de hets, tempo, porcentagem de lócus com >50% de heterozigotos	2020	[22]	NA
16 GT, GATK, BCFTools-Single, BCFTools-Multiple, VarScan2-Single, VarScan2-Multiple e FreeBayes	Bowtie2	Avaliação de 7 *pipelines* de SNP *calling*	Frango	Número de SNPs, sensibilidade e especificidade	2022	[26]	BCFTools-multiple foi o melhor *pipeline* para identificação de SNPs em dados de frango

NA: Não Disponível.

3 MATERIAL E MÉTODOS

3.1 OBTENÇÃO DOS DADOS BRUTOS DE SEQUENCIAMENTO DE DNA

Um conjunto de dados brutos de sequenciamento genômico de 250 genótipos de *Gossypium hirsutum* foi selecionado, com base em sua distribuição em dendrograma (conforme descrito no item 3.2), para a análise comparativa de duas estratégias de identificação de polimorfismos do tipo SNP a partir de dados disponibilizados pela TMG (Tropical Melhoramento e Genética SA), da qual foram selecionados 72 acessos e outros 178 acessos obtidos na literatura [25].

3.1.1 DADOS BRUTOS: TMG

Realizamos a extração de DNA de discos foliares de 474 genótipos de algodão do programa de melhoramento genético da TMG (Tropical Melhoramento e Genética). Utilizamos o kit comercial ReliaPrep™ gDNA Miniprep da Promega (referência A2051) para a extração de DNA. As mostras foram submetidas à técnica de GBS, utilizando a enzima de restrição ApeKI, na Universidade de Laval, onde ocorreu o sequenciamento de DNA utilizando a plataforma Ion Torrent, chip 540. Obtivemos os dados brutos de sequenciamento a partir de um arquivo FASTQ disponibilizado pela Universidade de Laval. Dos 474 genótipos, um subconjunto de 72 genótipos foi selecionado para análises posteriores (conforme descrito no item 3.2).

3.1.2 DADOS BRUTOS: LITERATURA

Foram utilizados 178 acessos disponibilizados por Li e colaboradores [25] na literatura, de um total de 1.961 acessos (conforme descrito no item 3.2). Os dados brutos de sequenciamento Illumina (FASTA *forward* e *reverse*) desse subconjunto de dados foram obtidos no NCBI [25], com o código SRA PRJNA576032, através do uso da ferramenta SRAtoolkit V. 3.0.

3.2 ANÁLISE DE DENDROGRAMA

Para as análises de dendrograma dos subconjuntos de dados da TMG e da literatura, foi utilizado o *software* Tassel versão 5.2.80 para gerar as clusterizações do tipo UPGMA (do inglês, *Unweighted pair-group Method using Arithmetic Averages*).

Para o subconjunto de dados da TMG, um total de 474 genótipos estavam disponíveis e destes foram selecionadas um total de 72 amostras. A escolha destes genótipos foi baseada na formação de três grupos na árvore filogenética. Em cada grupo formado na árvore filogenética, 24 amostras foram selecionadas de forma aleatória, totalizando 72 amostras (Anexo 1).

Para o subconjunto de dados da literatura, não foram observadas grandes variações na árvore filogenética, considerando a proximidade entre as amostras deste grupo, foram selecionadas aleatoriamente 178 genótipos do subconjunto total de dados (Anexo 2).

3.3 DETERMINAÇÃO DE POLIMORFISMOS (SNPs)

A determinação dos polimorfismos, etapa que visa identificar e recuperar posições genômicas onde uma ou mais amostras diferem da sequência de referência, foi realizada por duas abordagens: 1) *Pipeline* Fast-GBS v2.0 e 2) *Pipeline* proposto por Li e colaboradores, denominado a partir de agora como “BWA/BCFTools” (2021) [25].

As sequências brutas (*Single-end* para o subconjunto da TMG e *Paired-end* para o subconjunto da literatura) foram filtradas com o auxílio do *software* Trimmomatic versão 0.32, seguindo os critérios de tamanho mínimo de sequência de 75 pb e coeficiente phred de 33. Posteriormente, as sequências filtradas foram alinhadas com o genoma de referência de *Gossypium hirsutum* (acc TM-1_HAU v1.0) [25], utilizando o *software* BWA-MEM (versão 0.7.17.r1188). O alinhamento de cada sequência filtrada ao genoma de referência gerou um arquivo SAM que foi convertido para BAM e ordenado por coordenadas, utilizando o *software* Samtools versão 1.9. Sequências duplicadas foram removidas com a utilização do software Picard MarkDuplicates versão 2.26.0.

A partir desta etapa, as duas abordagens utilizadas neste estudo divergem em sua metodologia e foram especificados separadamente, conforme detalhado nos itens 3.3.1 e 3.3.2.

3.3.1 PIPELINE BWA/BCFTools

Após a remoção das duplicidades de sequência com Picard MarkDuplicates versão 2.26.0, o *software* bcftools e função *mpileup* (versão 1.9) foram utilizados para unificar todos os arquivos BAM e empilhar todas as variantes CNV (c*ommon copy number variations*) encontradas, gerando um arquivo bruto BCF. Após esta etapa, utilizamos o bcftools e função *call* para a chamada das variantes, gerando ao final um arquivo VCF bruto. Com esse VCF bruto, os cromossomos foram filtrados com *scripts* awk, possibilitando a remoção de SNPs mapeados em *scaffolds* e *contigs*. Posteriormente os dados foram imputados com o *software* Beagle versão 5.4.

Os seguintes filtros de qualidade foram aplicados aos dados: MAF (*Minor Alelle Frequency*) < 0,05, identificação de variantes bialélicas, máximo de dados faltantes de 20% e heterozigosidade permitida de 50% para obtenção do arquivo VCF final. Além disso, o s*oftware* VCFtools versão 0.1.17 foi utilizado para remoção de *INDELs*

3.3.2 *PIPELINE* FAST-GBS

Após a remoção das duplicidades de sequência com Picard MarkDuplicates versão 2.26.0, o *software Platypus* versão 0.8.1 foi utilizado para chamada de alelos, os cromossomos foram filtrados com *scripts* awk, possibilitando a remoção de SNPs mapeados em *scaffolds* e *contigs.* Posteriormente os dados foram imputados com o *software* Beagle versão 5.4.

Os seguintes filtros de qualidade foram aplicados aos dados: MAF (*Minor Alelle Frequency*) < 0,05, identificação de variantes bialélicas, máximo de dados faltantes de 20% e heterozigosidade permitida de 50% para obtenção do arquivo VCF final. Além disso, o s*oftware* VCFtools versão 0.1.17 foi utilizado para remoção de *INDELs*

3.4 MÉTRICAS DE ANÁLISE DOS RESULTADOS

O comparativo do seguinte conjunto de métricas foi realizado: 1) Número total de SNPs recuperados por *pipeline*; 2) Número de SNPs recuperados por cromossomo (considerando cada *pipeline*) 3) SNP-Score: permitir avaliar as ferramentas por meio de uma métrica de pontuação proposta, rotulada SNP-Score, capaz de identificar e ponderar o número de ocorrências de SNPs por *software* de chamada de alelos. SNP-score é uma métrica (0 a 1) dada por:

$$SNP - score = \frac{QSNP}{QP}$$

, onde QSNP = Quantidade de SNPs identificados por *software* e QP = Quantidade de *software*), 4) Tempo computacional e 5) Estrutura de populações.

4 RESULTADOS E DISCUSSÃO

4.1 RECUPERAÇÃO DE SNPs

Um total de 5 MB de sequências *single-end* foram produzidas no arquivo VCF com a recuperação de SNPs dos 72 acessos do subconjunto de dados da TMG e 14 GB de sequências *paired-end* foram produzidas com a recuperação de SNPs dos 178 acessos do subconjunto de dados da literatura.

O *pipeline* Fast-GBS recuperou um total de 417.975 SNPs para o subconjunto de dados da TMG (dados brutos), sendo estes 216.053 referentes ao subgenoma A e 201.922 referentes ao subgenoma D. Para o mesmo subconjunto de dados, quando foram considerados os dados filtrados, o total de SNPs recuperados foi de 3.726, deste montante 1.872 são referentes ao subgenoma A e outros 1.854 SNPs são referentes ao subgenoma D. A distribuição dos SNPs para cada cromossomo é apresentada na tabela 5.

O *pipeline* BWA/BCFTools recuperou um total de 254.805 SNPs para o subconjunto de dados da TMG (dados brutos), sendo estes 134.092 referentes ao subgenoma A e 120.713 referentes ao subgenoma D. Para o mesmo subconjunto de dados, quando foram considerados os dados filtrados o total de SNPs recuperados foi de 18.548, deste montante 10.020 são referentes ao subgenoma A e outros 8.528 SNPs são referentes ao subgenoma D, a distribuição dos SNPs para cada cromossomo é apresentada na tabela 6.

Com relação ao subconjunto de dados da literatura, o *pipeline* Fast-GBS recuperou um total de 38.685.370 SNPs quando foram considerados os dados brutos, destes, 23.444.751 são representantes do subgenoma A e 15.240.619 são representantes do subgenoma D. Quando foram considerados os dados filtrados, foram obtidos 20.259.083 SNPs, 12.057.297 representantes do subgenoma A e 8.201.786 representantes do subgenoma D. A distribuição dos SNPs para cada cromossomo é apresentada na tabela 7. É possível observar que o subconjunto de dados da literatura resultou em uma maior quantidade de SNPs recuperados pelo *pipeline* Fast-GBS. [23].

O *pipeline* BWA/BCFTools recuperou um total de 38.685.377 SNPs quando foram considerados os dados brutos da literatura, destes, 23.444.753 são representantes do subgenoma A e 15.240.624 são representantes do subgenoma D.

Em comparação, quando foram considerados os dados filtrados, foram obtidos 20.261.590 SNPs, 12.052.421 representantes do subgenoma A e 8.209.169 representantes do subgenoma D. A distribuição dos SNPs para cada cromossomo é apresentada na tabela 8.

Tabela 5. Total de SNPs recuperados por cromossomo no subconjunto de dados da TMG (72 acessos) submetidos ao processo de recuperação de SNPs com *pipeline* Fast-GBS.

***Pipeline* FastGBS | Subconjunto de dados da TMG**							
Raw				**Filtrado**			
Quantidade de SNPs por Cromossomo (Chr)				Quantidade de SNPs por Cromossomo (Chr)			
Chr A01	15.799	Chr D01	13.863	Chr A01	303	Chr D01	144
Chr A02	12.823	Chr D02	14.863	Chr A02	89	Chr D02	194
Chr A03	16.233	Chr D03	10.991	Chr A03	125	Chr D03	65
Chr A04	10.231	Chr D04	13.193	Chr A04	96	Chr D04	79
Chr A05	26.407	Chr D05	22.575	Chr A05	249	Chr D05	431
Chr A06	15.278	Chr D06	14.072	Chr A06	98	Chr D06	73
Chr A07	15.198	Chr D07	13.795	Chr A07	101	Chr D07	74
Chr A08	18.394	Chr D08	16.385	Chr A08	154	Chr D08	249
Chr A09	15.054	Chr D09	14.332	Chr A09	94	Chr D09	108
Chr A10	16.508	Chr D10	16.194	Chr A10	136	Chr D10	107
Chr A11	21.777	Chr D11	20.703	Chr A11	159	Chr D11	118
Chr A12	16.937	Chr D12	16.424	Chr A12	108	Chr D12	118
Chr A13	15.414	Chr D13	14.532	Chr A13	160	Chr D13	94
TOTAL A	216.053	Total D	201.922	TOTAL A	1.872	Total D	1.854
Total Geral		417.975		Total Geral		3.726	

Tabela 6. Total de SNPs recuperados por cromossomo considerando dados brutos e filtrados do subconjunto de dados da TMG (72 acessos) submetidos ao processo de recuperação de SNPs com *pipeline* BWA/BCFTools.

***Pipeline* BWA/BCFTools | Subconjunto de dados da TMG**							
Raw				**Filtrado**			
Quantidade de SNPs por Cromossomo (Chr)				Quantidade de SNPs por Cromossomo			
Chr A01	10.050	Chr D01	8.612	Chr A01	1.152	Chr D01	614
Chr A02	8.447	Chr D02	9.448	Chr A02	592	Chr D02	670
Chr A03	9.667	Chr D03	6.705	Chr A03	723	Chr D03	430
Chr A04	6.726	Chr D04	7.362	Chr A04	584	Chr D04	419
Chr A05	14.883	Chr D05	14.201	Chr A05	1.164	Chr D05	1.494
Chr A06	9.956	Chr D06	8.300	Chr A06	612	Chr D06	507
Chr A07	9.568	Chr D07	8.583	Chr A07	610	Chr D07	629
Chr A08	12.173	Chr D08	10.144	Chr A08	957	Chr D08	968
Chr A09	8.904	Chr D09	8.231	Chr A09	621	Chr D09	551
Chr A10	10.420	Chr D10	9.504	Chr A10	765	Chr D10	586
Chr A11	13.289	Chr D11	11.688	Chr A11	840	Chr D11	589
Chr A12	10.317	Chr D12	9.536	Chr A12	646	Chr D12	594
Chr A13	9.692	Chr D13	8.399	Chr A13	754	Chr D13	477
TOTAL A	134.092	Total D	120.713	TOTAL A	10.020	Total D	8.528
Total Geral		254.805		Total Geral		18.548	

Tabela 7. Total de SNPs recuperados por cromossomo considerando dados brutos e filtrados do subconjunto de dados da literatura (178 acessos) submetidos ao processo de recuperação de SNPs com *pipeline* Fast-GBS.

***Pipeline* FastGBS | Subconjunto de dados da Literatura**							
Raw				**Filtrado**			
Quantidade de SNPs por Cromossomo (Chr)				Quantidade de SNPs por Cromossomo (Chr)			
Chr A01	2.022.667	Chr D01	1.127.267	Chr A01	1.113.325	Chr D01	562.212
Chr A02	1.592.119	Chr D02	1.184.778	Chr A02	804.073	Chr D02	590.321
Chr A03	1.650.112	Chr D03	873.062	Chr A03	851.289	Chr D03	443.520
Chr A04	1.233.520	Chr D04	865.066	Chr A04	617.199	Chr D04	455.584
Chr A05	1.754.455	Chr D05	1.175.527	Chr A05	852.911	Chr D05	604.487
Chr A06	2.403.312	Chr D06	1.166.197	Chr A06	1.242.351	Chr D06	562.555
Chr A07	1.570.538	Chr D07	1.009.008	Chr A07	817.562	Chr D07	496.601
Chr A08	2.172.920	Chr D08	2.050.121	Chr A08	1.001.900	Chr D08	1.408.859
Chr A09	1.421.871	Chr D09	1.073.728	Chr A09	728.775	Chr D09	577.905
Chr A10	1.890.072	Chr D10	1.391.433	Chr A10	975.560	Chr D10	791.886
Chr A11	1.963.719	Chr D11	1.234.859	Chr A11	996.002	Chr D11	627.611
Chr A12	1.577.053	Chr D12	1.068.265	Chr A12	796.376	Chr D12	559.908
Chr A13	2.192.393	Chr D13	1.021.308	Chr A13	1.259.974	Chr D13	520.337
TOTAL A	23.444.751	Total D	15.240.619	TOTAL A	12.057.297	Total D	8.201.786
Total Geral		38.685.370		Total Geral		20.259.083	

Tabela 8. Total de SNPs recuperados por cromossomo considerando dados brutos e filtrados do subconjunto de dados da literatura (178 acessos) submetidos ao processo de recuperação de SNPs com *pipeline* BWA/BCFTools.

***Pipeline* BWA/BCFTools | Subconjunto de dados da Literatura**							
Raw				**Filtrado**			
Quantidade de SNPs por Cromossomo (Chr)				Quantidade de SNPs por Cromossomo (Chr)			
Chr A01	2.022.668	Chr D01	1.127.267	Chr A01	1.113.884	Chr D01	562.172
Chr A02	1.592.120	Chr D02	1.184.778	Chr A02	803.279	Chr D02	590.580
Chr A03	1.650.112	Chr D03	873.063	Chr A03	851.687	Chr D03	443.845
Chr A04	1.233.520	Chr D04	865.069	Chr A04	617.135	Chr D04	455.716
Chr A05	1.754.455	Chr D05	1.175.527	Chr A05	852.298	Chr D05	604.087
Chr A06	2.403.312	Chr D06	1.166.198	Chr A06	1.243.029	Chr D06	565.119
Chr A07	1.570.538	Chr D07	1.009.008	Chr A07	817.901	Chr D07	496.970
Chr A08	2.172.922	Chr D08	2.050.121	Chr A08	1.000.469	Chr D08	1.408.850
Chr A09	1.421.872	Chr D09	1.073.728	Chr A09	726.726	Chr D09	578.910
Chr A10	1.890.072	Chr D10	1.391.433	Chr A10	975.573	Chr D10	792.355
Chr A11	1.963.717	Chr D11	1.234.858	Chr A11	994.004	Chr D11	629.159
Chr A12	1.577.054	Chr D12	1.068.266	Chr A12	796.318	Chr D12	560.328
Chr A13	2.192.391	Chr D13	1.021.308	Chr A13	1.260.118	Chr D13	521.078
TOTAL A	23.444.753	Total D	15.240.624	TOTAL A	12.052.421	Total D	8.209.169
Total Geral		38.685.377		Total Geral		20.261.590	

Uma comparação foi realizada entre os pipelines BWA/BCFTools e Fast-GBS, focando na diferença do número total de SNPs recuperados por cada um. No subconjunto de dados da TMG, o pipeline Fast-GBS identificou uma maior quantidade de SNPs a partir dos dados brutos. No entanto, alguns desses SNPs são removidos

quando os filtros de qualidade são aplicados. O pipeline BWA/BCFTools resulta em mais SNPs após aplicação de filtros de qualidade em comparação ao Fast-GBS neste mesmo subconjunto de dados, conforme demonstrado na Tabela 9. Também é possível observar que para o conjunto de dados da TMG, tanto para dados brutos, quanto para dados filtrados e para ambos os *pipelines* os cromossomos que possuem maior quantidade de SNPs são A05 e D05.

Em relação ao subconjunto de dados da literatura, os dois *pipelines* conseguem recuperar quantidades similares de SNPs. Ao aplicar os filtros de qualidade da sequência, há uma variação no número de SNPs identificados em cada cromossomo. Ainda assim, na maioria dos casos, o pipeline BWA/BCFTools apresenta um desempenho superior ao Fast-GBS, como ilustrado na Tabela 10. Os cromossomos A06 e D08 apresentam maior quantidade de SNPs para dados brutos de ambos os *pipelines*. Após aplicação de filtros de qualidade, os cromossomos com maior número de SNPs passam a ser A13 e D08 para ambos os *pipelines*.

O *pipeline* Fast-GBS demonstrou um desempenho superior ao lidar com dados brutos de sequenciamento no subconjunto de dados da TMG (sequências *single-end* obtidas por meio de GBS em sequenciamento Ion Torrent) em relação ao *pipeline* BWA/BCFTools. Esse *pipeline* é comumente empregado na chamada de variantes em dados de GBS, utilizando um algoritmo baseado em filtros de qualidade e frequência alélica para a identificação de SNPs. Apesar de observarmos a perda de vários SNPs após aplicação de filtros de qualidade, os dados de sequenciamento *single-end* não demonstraram problemas de qualidade de sequenciamento. Podemos inferir que os critérios de filtro aplicados foram muito restritivos, pois foram baseados em sequencias *paired-end*.

Tabela 9. Comparativo do número de SNPs recuperados por cromossomo considerando dados brutos e filtrados do subconjunto de dados da TMG (72 acessos) por ambos os *pipelines*.

Comparativo* entre *Pipeline* BWA/BCFTools x FastGBS | Subconjunto de dados da TMG							
* Diferença entre o número de SNPs identificados por Chr (SNPs *Pipeline* BWA/BCFTools - SNPs *Pipeline* FastGBS)							
Raw					**Filtrado**		
Quantidade de SNPs por Cromossomo (Chr)					Quantidade de SNPs por Cromossomo (Chr)		
Chr A01	-5.749	Chr D01	-5.251		Chr A01	849	Chr D01	470
Chr A02	-4.376	Chr D02	-5.415		Chr A02	503	Chr D02	476
Chr A03	-6.566	Chr D03	-4.286		Chr A03	598	Chr D03	365
Chr A04	-3.505	Chr D04	-5.831		Chr A04	488	Chr D04	340
Chr A05	-11.524	Chr D05	-8.374		Chr A05	915	Chr D05	1.063
Chr A06	-5.322	Chr D06	-5.772		Chr A06	514	Chr D06	434
Chr A07	-5.630	Chr D07	-5.212		Chr A07	509	Chr D07	555
Chr A08	-6.221	Chr D08	-6.241		Chr A08	803	Chr D08	719
Chr A09	-6.150	Chr D09	-6.101		Chr A09	527	Chr D09	443
Chr A10	-6.088	Chr D10	-6.690		Chr A10	629	Chr D10	479
Chr A11	-8.488	Chr D11	-9.015		Chr A11	681	Chr D11	471
Chr A12	-6.620	Chr D12	-6.888		Chr A12	538	Chr D12	476
Chr A13	-5.722	Chr D13	-6.133		Chr A13	594	Chr D13	383

Tabela 10. Comparativo do número de SNPs recuperados por cromossomo considerando dados brutos e filtrados do subconjunto de dados da literatura (178 acessos) por ambos os *pipelines*.

Comparativo* entre *Pipeline* BWA/BCFTools x FastGBS | Subconjunto de dados da literatura							
* Diferença entre o número de SNPs identificados por Chr (SNPs *Pipeline* BWA/BCFTools - SNPs *Pipeline* FastGBS)							
Raw					**Filtrado**		
Quantidade de SNPs por Cromossomo (Chr)					Quantidade de SNPs por Cromossomo (Chr)		
Chr A01	1	Chr D01	0		Chr A01	559	Chr D01	-40
Chr A02	1	Chr D02	0		Chr A02	-794	Chr D02	259
Chr A03	0	Chr D03	1		Chr A03	398	Chr D03	325
Chr A04	0	Chr D04	3		Chr A04	-64	Chr D04	132
Chr A05	0	Chr D05	0		Chr A05	-613	Chr D05	-400
Chr A06	0	Chr D06	1		Chr A06	678	Chr D06	2.564
Chr A07	0	Chr D07	0		Chr A07	339	Chr D07	369
Chr A08	2	Chr D08	0		Chr A08	-1.431	Chr D08	-9
Chr A09	1	Chr D09	0		Chr A09	-2.049	Chr D09	1.005
Chr A10	0	Chr D10	0		Chr A10	13	Chr D10	469
Chr A11	-2	Chr D11	-1		Chr A11	-1.998	Chr D11	1.548
Chr A12	1	Chr D12	1		Chr A12	-58	Chr D12	420
Chr A13	-2	Chr D13	0		Chr A13	144	Chr D13	741

Os resultados também foram comparados por meio de diagrama de Venn. Na figura 4, podemos observar a comparação dos resultados para dados brutos (arquivo

VCF). O conjunto "Literatura" representa o *pipeline* BWA/BCFTools com o subconjunto de dados da literatura, o conjunto "Li" representa os SNPs recuperados por Li et al. (2021) [25] com o conjunto de dados da literatura, e o conjunto "TMG" representa os SNPs recuperados com o pipeline BWA/BCFTools com o subconjunto de dados da TMG.

Podemos identificar que existem 24.402 SNPs em comum em todos os conjuntos de dados quando o pipeline BWA/BCFTools foi utilizado. Além disso, há 15.236 SNPs em comum entre o subconjunto de dados da TMG e da literatura (conjuntos "Literatura" e "TMG"), 3.491 SNPs em comum entre o subconjunto de dados da TMG e os dados públicos utilizados por Li et al. (2021) (conjuntos "Li" e "TMG"), e 9.509.600 SNPs em comum entre o subconjunto da literatura e os dados públicos utilizados por Li et al. (2021) (conjuntos "Literatura" e "Li"). Além disso, existem 28.393.106 SNPs exclusivos do subconjunto de dados da literatura (conjunto "Literatura"), 24.940.043 SNPs exclusivos do conjunto "Li", e 211.310 SNPs exclusivos do subconjunto de dados da TMG (conjunto "TMG") (Figura 4).

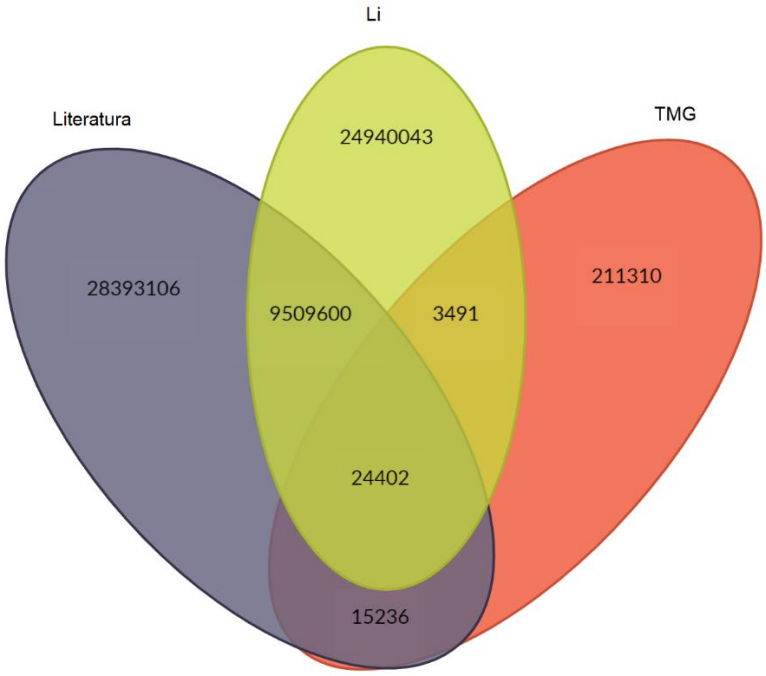


Figura 4. Diagrama de Venn do comparativo de resultados para o *pipeline* BWA/BCFTools em arquivo VCF com dados brutos. Conjunto “Literatura”: SNPs recuperados no subconjunto de dados da literatura. Conjunto “Li”: SNPs recuperados por Li e colaboradores (2021) em dados públicos, disponibilizados no NCBI e Conjunto “TMG”: SNPs recuperados no subconjunto de dados da TMG.

A comparação do arquivo VCF filtrado revela distintas interseções e exclusividades entre os conjuntos de dados. Quando o pipeline BWA/BCFTools foi utilizado, existem 3.648 SNPs em comum em todos os conjuntos de dados (Figura 5). Além disso, encontram-se 6.011 SNPs em comum entre o subconjunto de dados da TMG e os dados públicos utilizados por Li et al. (2021), e 4.656.077 SNPs em comum entre o subconjunto de dados da literatura e os dados públicos de Li et al. (2021).

A análise de interseções também revela várias exclusividades, com 15.518.741 SNPs pertencentes somente ao subconjunto de dados da literatura, 29.811.800 SNPs exclusivos do conjunto "Li", e 8.432 SNPs exclusivos do subconjunto de dados da TMG. Todas essas relações e diferenças estão detalhadas, mostrando a complexa interação entre os conjuntos de dados.

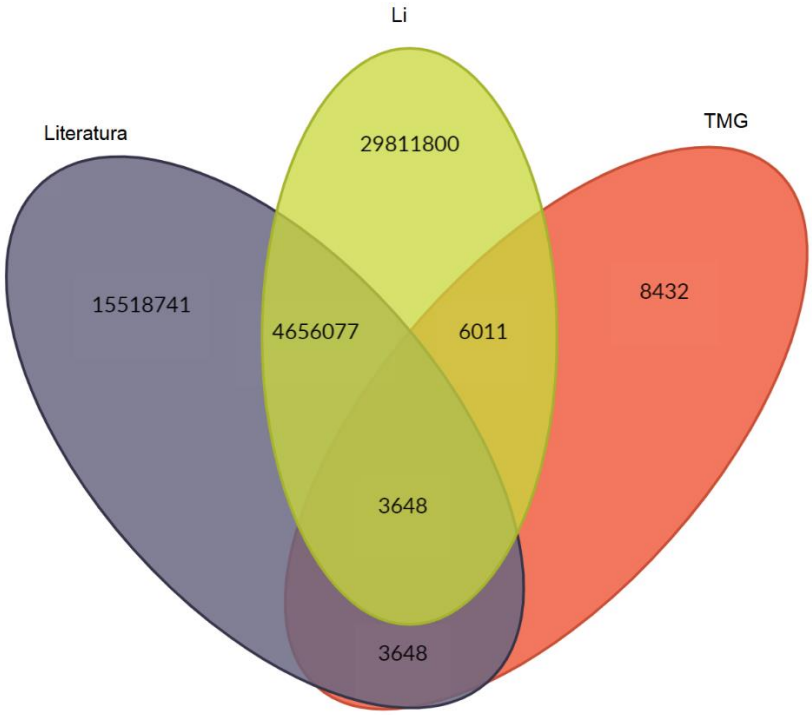


Figura 5. Diagrama de Venn do comparativo de resultados para o *pipeline* BWA/BCFTools em arquivo VCF com dados filtrados. Conjunto Literatura: SNPs recuperados no subconjunto de dados da literatura. Conjunto Li: SNPs recuperados por Li e colaboradores (2021) em dados públicos, disponibilizados no NCBI e Conjunto TMG: SNPs recuperados no subconjunto de dados da TMG.

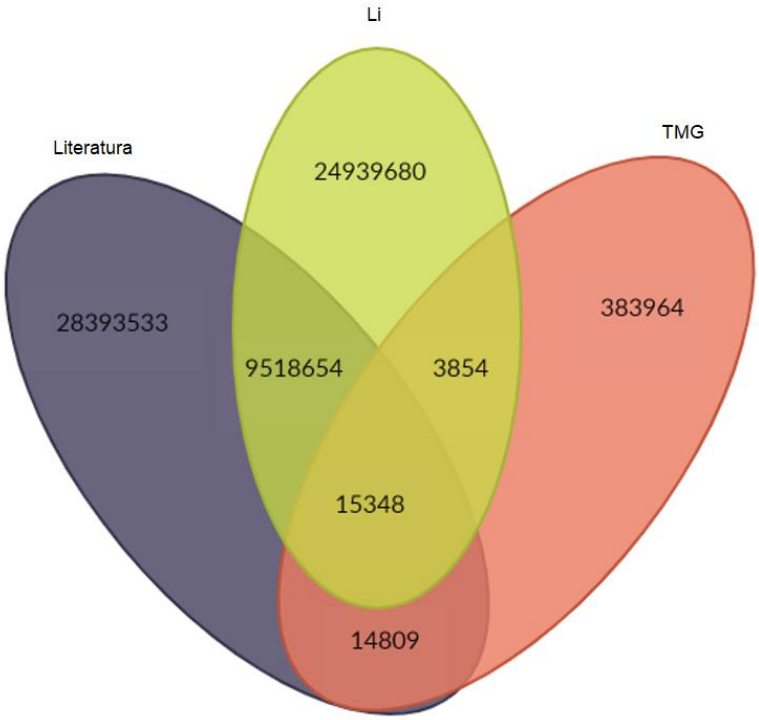


Figura 6. Diagrama de Venn do comparativo de resultados para o *pipeline* Fast-GBS em arquivo VCF com dados brutos. Conjunto Literatura: SNPs recuperados no subconjunto de dados da literatura. Conjunto Li: SNPs recuperados por Li e colaboradores (2021) em dados públicos, disponibilizados no NCBI e Conjunto TMG: SNPs recuperados no subconjunto de dados da TMG.

Em relação aos dados brutos (arquivo VCF) quando o *pipeline* Fast-GBS foi utilizado, foi possível identificar 15.348 SNPs em comum em todos os conjuntos de dados (Figura 6), além de interseções específicas de 14.809 SNPs entre os subconjuntos de dados da TMG e da literatura, 3.854 SNPs entre a TMG e os dados públicos de Li et al. (2021), e 9.518.654 SNPs entre o subconjunto da literatura e os dados públicos de Li et al. (2021) (Figura 6).

Além disso, observamos grandes números de SNPs exclusivos: 28.393.533 no subconjunto de dados da literatura, 24.939.680 no conjunto "Li", e 383.964 no subconjunto de dados da TMG, evidenciando a complexidade e especificidade dos diferentes conjuntos de dados (Figura 6).

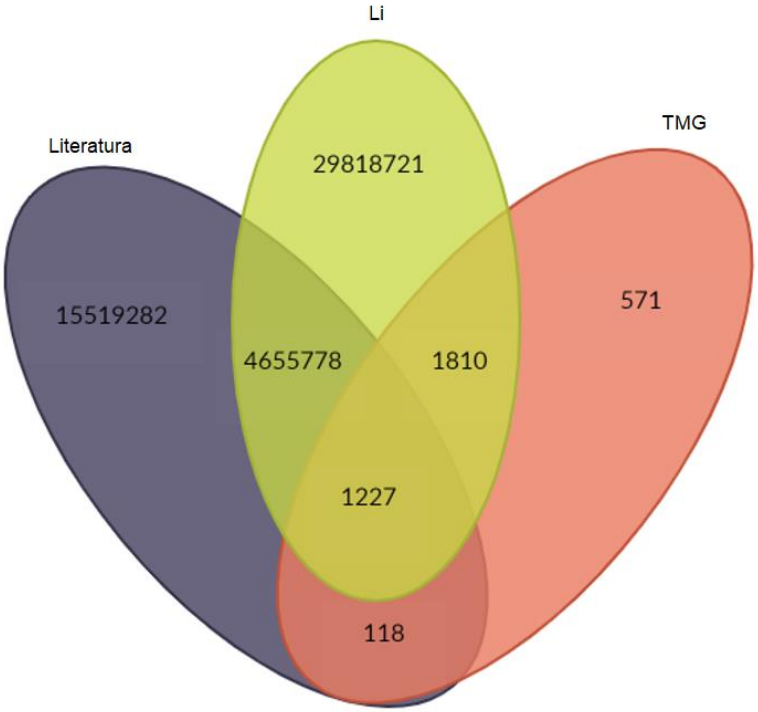


Figura 7. Diagrama de Venn do comparativo de resultados para o *pipeline* Fast-GBS em arquivo VCF com dados filtrados. Conjunto Literatura: SNPs recuperados no subconjunto de dados da literatura. Conjunto Li: SNPs recuperados por Li e colaboradores (2021) em dados públicos, disponibilizados no NCBI e Conjunto TMG: SNPs recuperados no subconjunto de dados da TMG.

O comparativo para o arquivo VCF filtrado, utilizando o mesmo conjunto de dados descrito anteriormente, revela 1.227 SNPs em comum em todos os conjuntos quando o pipeline Fast-GBS foi aplicado (Figura 7). Além disso, a análise identifica interseções específicas: 118 SNPs entre o subconjunto de dados da TMG e da literatura, 1.810 SNPs entre a TMG e os dados públicos de Li et al. (2021), e 4.655.778 SNPs entre o subconjunto da literatura e os dados públicos de Li et al. (2021) (Figura 7). A análise também mostra 15.519.282 SNPs exclusivos do subconjunto de dados da literatura, 29.818.721 SNPs exclusivos do conjunto "Li", e 571 SNPs exclusivos do subconjunto da TMG (Figura 7). Já o comparativo geral entre os pipelines para os diferentes conjuntos de dados, considerando o arquivo VCF bruto, é detalhado com diferentes combinações de pipelines e conjuntos (Figura 8). Assim, o conjunto “Literatura_BWA/BCFTools” representa o pipeline BWA/BCFTools com o subconjunto de dados da literatura, e assim por diante para outros conjuntos (Figura 8).

Deste comparativo, é possível concluir que 23.500 SNPs são comuns entre todos os conjuntos de dados e pipelines; 6.807 SNPs entre

Literatura_BWA/BCFTools, TMG_BWA/BCFTools, e Literatura_Fast-GBS; 16.288 SNPs entre Literatura_BWA/BCFTools, TMG_BWA/BCFTools, e TMG_Fast-GBS; 37.895.899 SNPs entre Literatura_BWA/BCFTools e TMG_Fast-GBS; 94.223 SNPs entre TMG_BWA/BCFTools e Literatura_Fast-GBS; 120.578 SNPs exclusivos do TMG_BWA/BCFTools; e outros 293.395 SNPs exclusivos do Literatura_Fast-GBS, sem qualquer SNP exclusivo para os conjuntos Literatura_BWA/BCFTools e TMG_Fast-GBS.

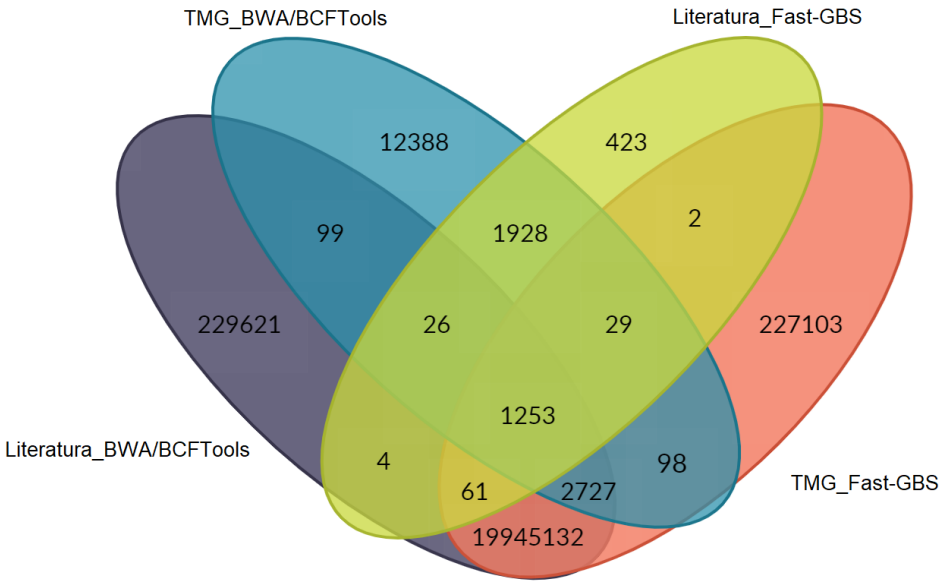


Figura 8. Diagrama de Venn do comparativo de resultados em arquivo VCF com dados brutos. “Literatura_BWA/BCFTools”: SNPs recuperados no subconjunto de dados da literatura com o *pipeline* BWA/BCFTools. “TMG_BWA/BCFTools”: SNPs recuperados no subconjunto de dados da TMG com o *pipeline* BWA/BCFTools. “Literatura_Fast-GBS”: SNPs recuperados no subconjunto de dados da literatura com o *pipeline* Fast-GBS. “TMG_Fast-GBS”: SNPs recuperados no subconjunto de dados da TMG com o *pipeline* Fast-GBS.

A Figura 9 ilustra um comparativo abrangente entre os pipelines para os vários conjuntos de dados, com base no arquivo VCF filtrado, mantendo-se alinhada à descrição de conjuntos de dados já apresentada na Figura 8. Nessa análise, foi possível identificar que existem 1.253 SNPs comuns entre todos os conjuntos de dados e pipelines. Existem 26 SNPs comuns entre Literatura_BWA/BCFTools, TMG_BWA/BCFTools e Literatura_Fast-GBS; 2.727 SNPs entre Literatura_BWA/BCFTools, TMG_BWA/BCFTools e TMG_Fast-GBS; 29 SNPs entre

TMG_BWA/BCFTools, Literatura_Fast-GBS e TMG_Fast-GBS; 99 SNPs entre Literatura_BWA/BCFTools e TMG_BWA/BCFTools; 4 SNPs entre Literatura_BWA/BCFTools e Literatura_Fast-GBS; e 19.945.132 SNPs entre Literatura_BWA/BCFTools e TMG_Fast-GBS. Além disso, há 1.928 SNPs entre TMG_BWA/BCFTools e Literatura_Fast-GBS; 98 SNPs entre TMG_BWA/BCFTools e TMG_Fast-GBS; 61 SNPs entre Literatura_BWA/BCFTools, Literatura_Fast-GBS e TMG_Fast-GBS; e 2 SNPs entre Literatura_Fast-GBS e TMG_Fast-GBS. Enquanto 229.621 SNPs são exclusivos do conjunto Literatura_BWA/BCFTools, 12.388 são exclusivos do conjunto TMG_BWA/BCFTools, 423 SNPs são exclusivos do conjunto Literatura_Fast-GBS, e 227.103 SNPs são exclusivos do conjunto TMG_Fast-GBS.

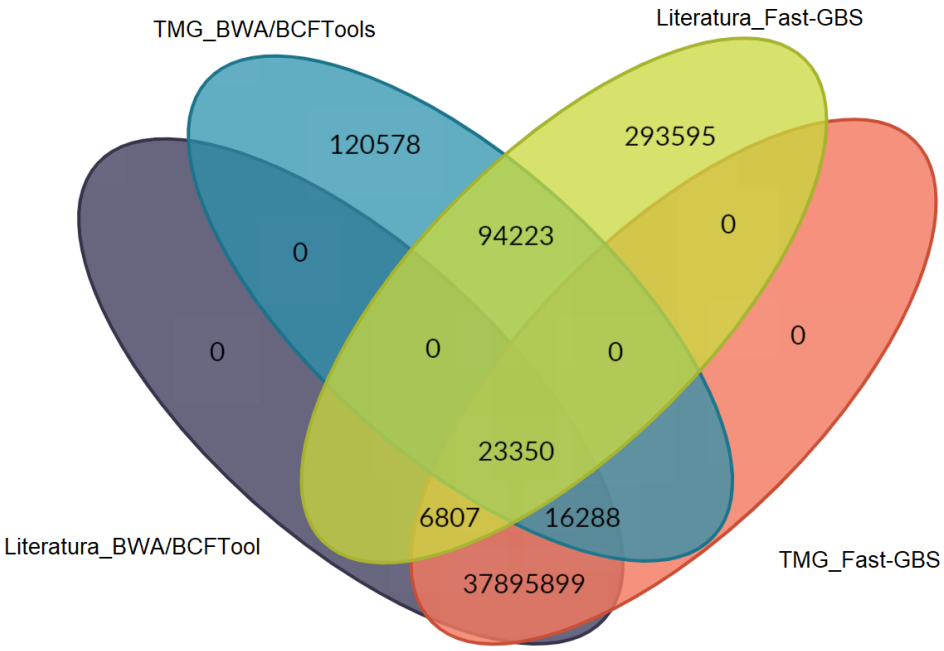


Figura 9. Diagrama de Venn do comparativo de resultados em arquivo VCF com dados filtrados. “Literatura_BWA/BCFTools”: SNPs recuperados no subconjunto de dados da literatura com o *pipeline* BWA/BCFTools. “TMG_BWA/BCFTools”: SNPs recuperados no subconjunto de dados da TMG com o *pipeline* BWA/BCFTools. “Literatura_Fast-GBS”: SNPs recuperados no subconjunto de dados da literatura com o *pipeline* Fast-GBS. “TMG_Fast-GBS”: SNPs recuperados no subconjunto de dados da TMG com o *pipeline* Fast-GBS.

4.2 SNP-SCORE

Os resultados foram comparados por meio do SNP-Score, avaliando as ferramentas através de uma métrica de pontuação proposta, capaz de identificar e ponderar o número de ocorrências de SNPs por software de chamada de alelos. É possível concluir que, para o subconjunto de dados da literatura (178 acessos, Tabela 11), não houve diferenças no comparativo entre SNP-Score para os dados brutos quando foram comparados os pipelines BWA/BCFTools e Fast-GBS; a mesma conclusão vale para os dados filtrados (conforme a Tabela 11). Para o subconjunto de dados da TMG (72 acessos, Tabela 11), a melhor performance de recuperação de SNPs foi observada com o pipeline Fast-GBS para dados brutos de sequenciamento, enquanto o pipeline BWA/BCFTools teve o melhor desempenho com os dados filtrados.

É possível inferir que alguns fatores podem influenciar na redução dos SNPs, como a cobertura e profundidade de sequenciamento de DNA, que podem sofrer maior influência das ferramentas utilizadas para chamada de alelos (Tabela 11 e Anexos 3 ao 10). Além disso, ao aplicar o filtro para MAF de 0,05%, podemos estar perdendo SNPs devido a uma pequena variação de determinados SNPs dentro da população, uma vez que os indivíduos pertencentes ao conjunto de dados da TMG são muito aparentados.

Tabela 11. Métrica SNP-Score para ambos os *pipelines*.

SNP-Score	**Subconjunto de dados da TMG**		**Subconjunto de dados da Literatura**	
Pipeline	**Dado Bruto**	**Dado Filtrado**	**Dado Bruto**	**Dado Filtrado**
BWA/BCFTools	0,00215	0,00016	0,32621	0,17086
Fast-GBS	0,00352	0,00003	0,32621	0,17086

4.3 TEMPO COMPUTACIONAL

As análises foram realizadas em uma máquina virtual Azure Standard E32ads v5, com processador de 32 núcleos e memória RAM de 256 GB. Com relação ao tempo computacional, o *pipeline* BWA/BCFTools resultou em 4:30 horas de análise por acesso quando são considerados dados brutos *paired-end* e 2:30 horas de análise por acesso quando são considerados dados brutos *single-end.* Desta forma, foram

necessárias 765,4 horas para analisar os acessos do subconjunto de dados da literatura (31,9 dias) e 165,6 horas para analisar os acessos do subconjunto de dados da TMG (6,9 dias).

O *pipeline* Fast-GBS resultou em 3:00 horas de análise por acesso quando são considerados dados brutos *paired-end* e 1:40 horas de análise por acesso quando são considerados dados brutos *single-end.* Desta forma, foram necessárias 534 horas para analisar os acessos do subconjunto de dados da literatura (22,2 dias) e 100,8 horas para analisar os acessos do subconjunto de dados da TMG (4,2 dias) por meio do *pipeline* Fast-GBS (Tabela 12).

Tabela 12. Tempo computacional (dias) para análise de ambos os *pipelines* em cada subconjunto de dados utilizados neste estudo.

Pipeline	**Subconjunto de dados da TMG**	**Subconjunto de dados da Literatura**
BWA/BCFTools	6,9 dias	31,9 dias
Fast-GBS	4,2 dias	22,2 dias

4.4 PCA

O total de SNPs obtidos por subconjunto de dados (TMG e literatura) para cada *pipeline* (Fast-GBS e BWA/BCFTools) foi utilizado para construir uma PCA. O objetivo desta análise foi verificar se as diferenças presentes nas estratégias de identificação de variantes utilizadas impactariam na distribuição das amostras na PCA.

Ao realizar o comparativo entre dados brutos e filtrados para ambos os pipelines no subconjunto de dados da literatura (Figura 10, 11, 12 e 13) é possível inferir que 1) existe pouca variação na PCA entre os *pipelines* Fast-GBS e BWA/BCFTools para o conjunto de dados da literatura e 2) Existem SNPs monomórficos que são removidos após aplicação de filtros de qualidade.

A PC1 é capaz de explicar 57% da variância após a filtragem para ambos os *pipelines* no conjunto de dados da literatura, enquanto para os dados da TMG a variância máxima é de 15% para dados brutos com pipeline Fast-GBS.

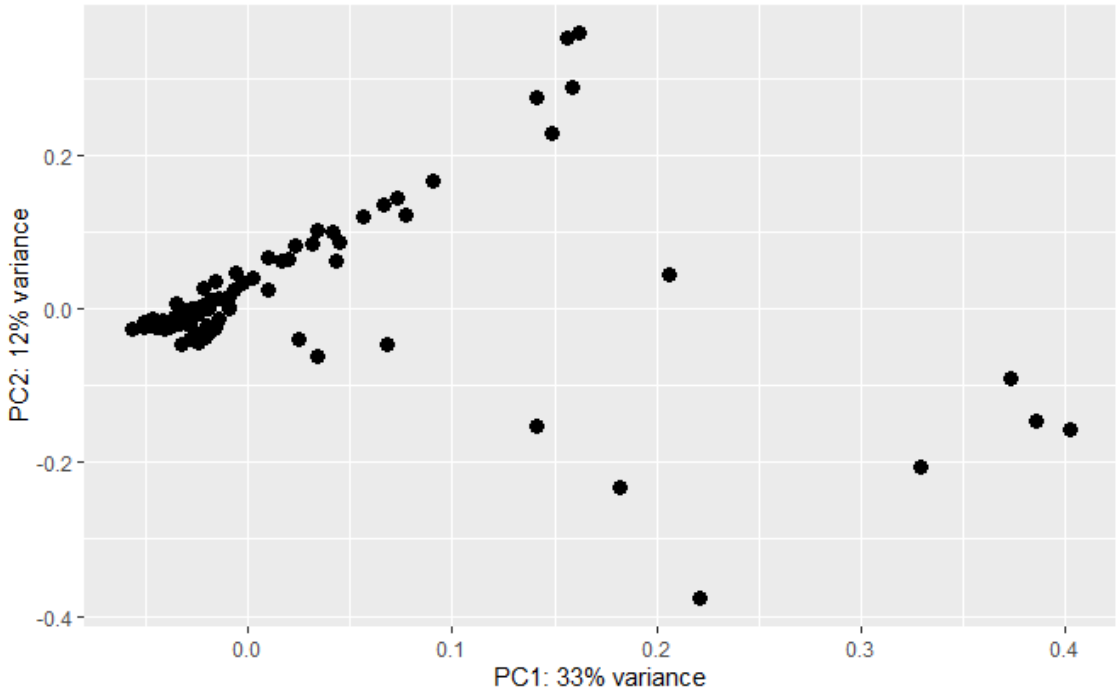


Figura 10. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados brutos da literatura com o *pipeline* BWA/BCFTolls.

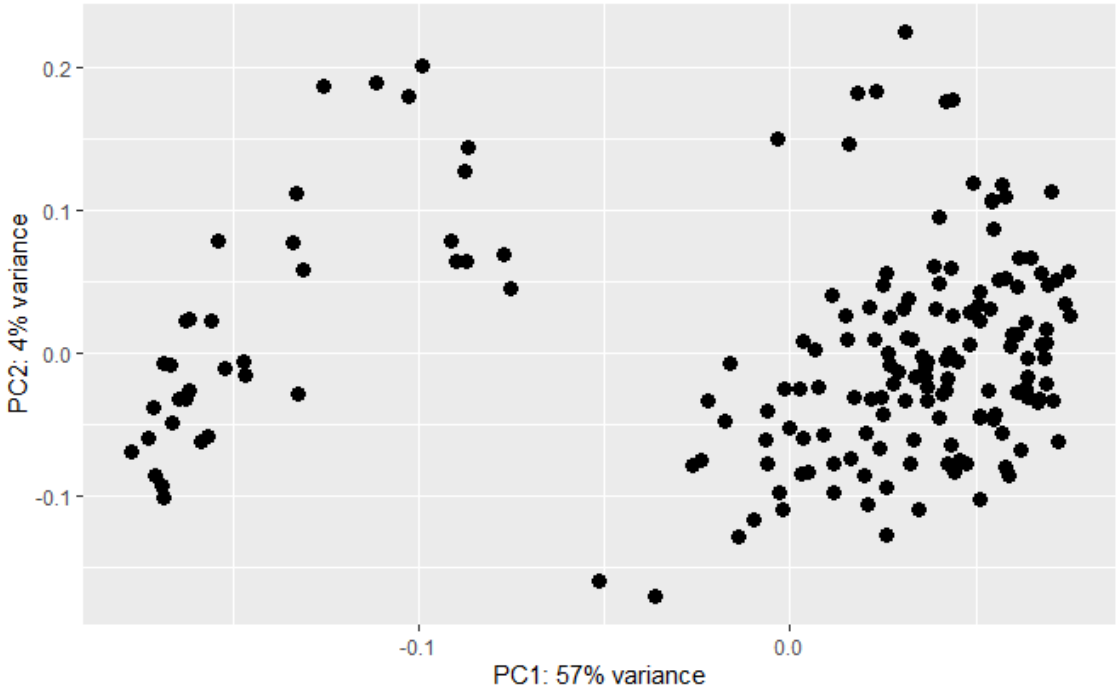


Figura 11. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados filtrados da literatura com o *pipeline* BWA/BCFTolls.

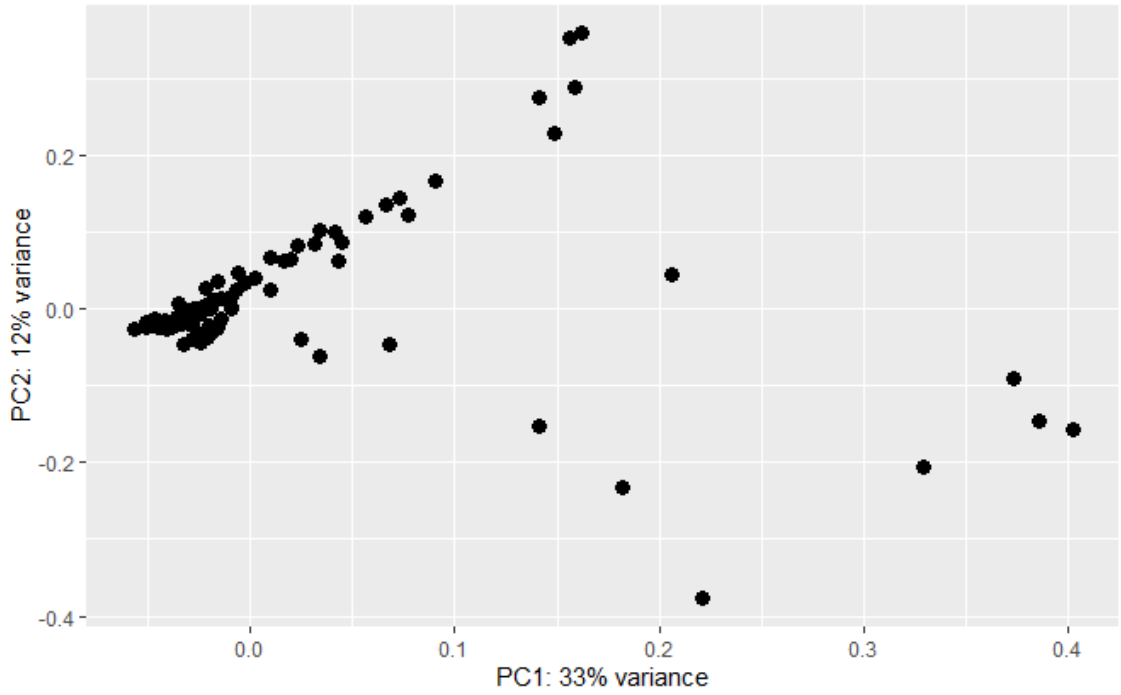


Figura 12. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados brutos da literatura com o *pipeline* Fast-GBS.

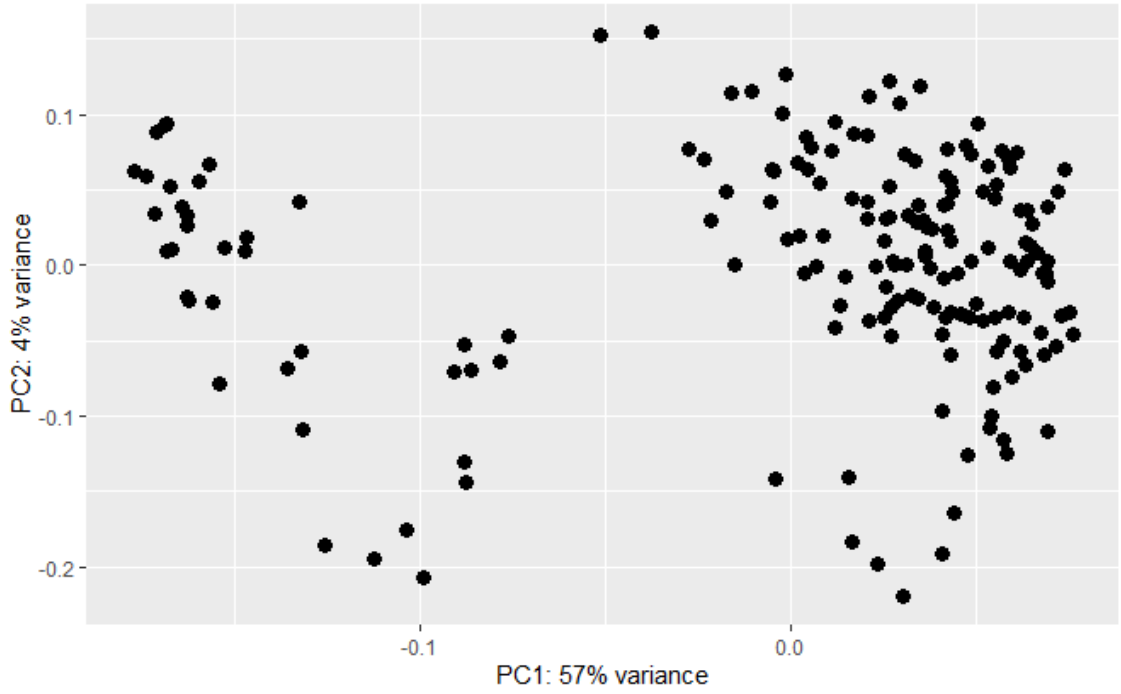


Figura 13. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados filtrados da literatura com o *pipeline* Fast-GBS.

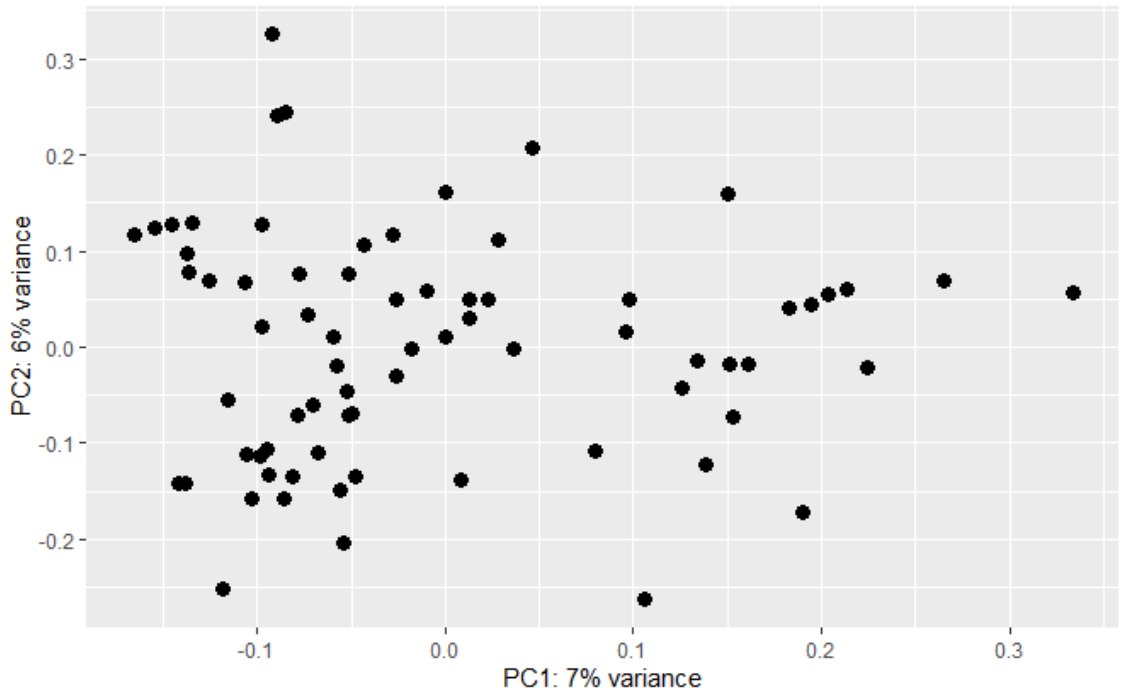


Figura 14. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados brutos da TMG com o *pipeline* BWA/BCFTolls.

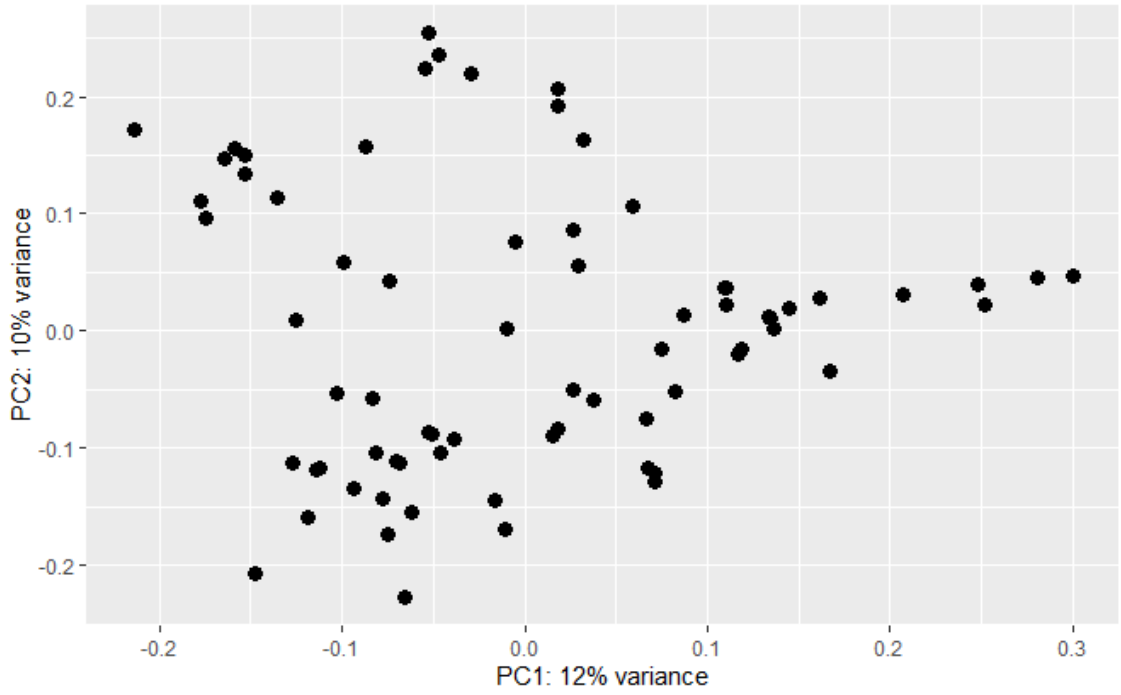


Figura 15. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados filtrados da TMG com o *pipeline* BWA/BCFTolls.

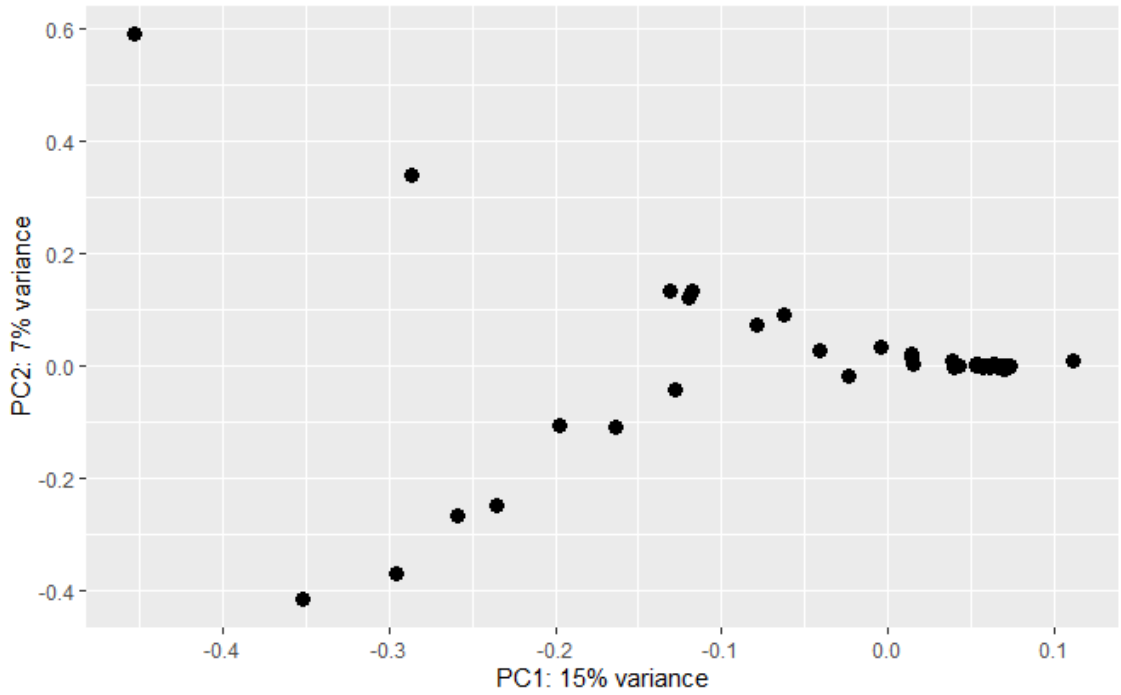


Figura 16. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados brutos da TMG com o *pipeline* Fast-GBS.

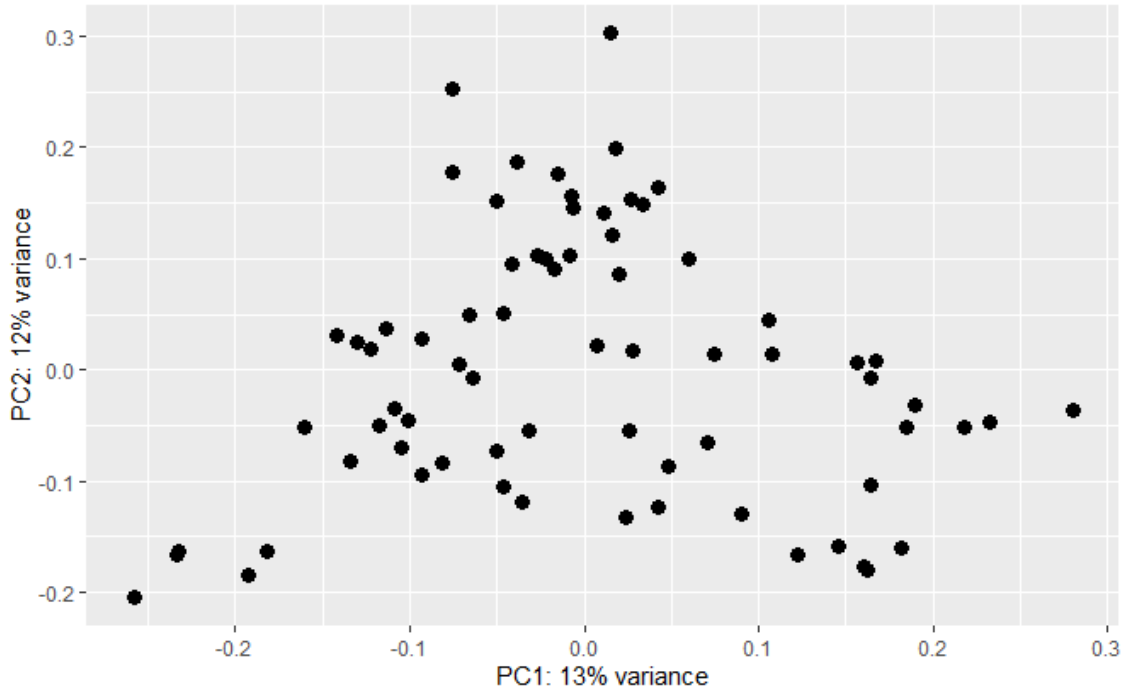


Figura 17. Análise de Componentes Principais (PCA) para SNPs obtidos por meio dos dados filtrados da TMG com o *pipeline* Fast-GBS.

5 CONCLUSÕES

Foi possível concluir que o subconjunto de dados retirado da literatura levou à recuperação de uma quantidade significativa de SNPs através de ambos os *pipelines*. Essa observação foi suportada pela maior cobertura proporcionada pelo sequenciamento *paired-end* (Figuras 10 a 13).

No caso das sequências *paired-end*, elas foram menos influenciadas pelo tipo de *pipeline* utilizado para a chamada de alelos, graças à robustez dos dados. Vale ressaltar que, embora o sequenciamento *paired-end* tenha proporcionado uma maior cobertura, ele também é mais dispendioso financeiramente e exige mais tempo de processamento em comparação ao sequenciamento *single-end*.

Em relação aos dados *single-end*, o pipeline Fast-GBS mostrou-se mais eficaz na recuperação de SNPs em comparação ao BWA-BCFTools, usando os dados brutos de sequenciamento. No entanto, uma grande parte dos SNPs foi descartada quando foram aplicados filtros de qualidade em ambos os *pipelines*. Isso sugeriu que os critérios de filtragem foram bastante rigorosos, já que foram os mesmos utilizados para as sequências *paired-end*.

A análise comparativa, representada pelos diagramas de Venn (Figuras 04 a 07), permitiu concluir que, para sequências brutas *single-end* (sem filtragem), o *pipeline* Fast-GBS proporcionou a melhor discriminação entre os conjuntos de dados. As sequências *paired-end* apresentaram pouca variação no número de SNPs exclusivos de cada conjunto. Em contraste, para dados filtrados, o *pipeline* BWA/BCFTools destacou-se na discriminação de SNPs exclusivos, especialmente com sequências *single-end*.

Finalmente, no que diz respeito ao tempo computacional, o Fast-GBS emergiu como o *pipeline* com o desempenho mais eficiente. Esse fato reforçou a necessidade de considerar várias dimensões na escolha dos métodos de análise para projetos futuros.

6 REFERÊNCIAS

[1] GUPTA, M.; SALGOTRA, R.; CHAUHAN, B. **Next-Generation Sequencing Technologies and Their Implications for Efficient Utilization of Genetic Resources** on Rediscovery of Genetic and Genomic Resources for Future Food Security. Cap 8. Springer. 2020. DOI: 10.1007/978-981-15-0156-2

[2] SCHATZ, M.; DELCHER A.; SALZBERG S. **Assembly of large genomes using second-generation sequencing**. Genome Res 20:1165–1173. 2010

[3] VEZZI, F. **Next generation sequencing revolution challenges: Search, assemble, and validate genomes.** Ph.D, Universita degli Studi di Udine, Italy, 2012.

[4] MICHAEL, L. **Sequencing technologies – the next generation.** Nat Rev Genet 11:31–46. 2010

[5] MARDIS, E. R. **The impact of next-generation sequencing technology on genetics**. Trends in genetics, 24(3), 133-141, 2008.

[6] KOLMOGOROV, M., BICKHART, D. M., BEHSAZ, B., GUREVICH, A., RAYKO, M., SHIN, S. B e RHIE, A. **MetaFlye: Scalable long-read metagenome assembly using repeat graphs**. Nature methods, 17(11), 1103-1110, 2020.

[7] MARDIS, E. R. **DNA sequencing technologies: 2006-2016**. Nature protocols, 12(2), 213-218, 2017.

[8] ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M E LEAMON, J. H. **An integrated semiconductor device enabling non-optical genome sequencing**. Nature, 475(7356), 348-352, 2011.

[9] VESTERGAARD, L.; OLIVEIRA, D.; HOGDALL, C.; HOGDALL, E. **Next Generation Sequencing Technology in the Clinic and Its Challenges**. Cancers, 13, 1751, 2021.

[10] GOLAN, D.; MEDVEDEV, P. **Using state machines to model the Ion Torrent sequencing process and to improve read error rates**. Bioinformatics. 2013.

[11] TORKAMANEH, D.; BOYLE, B.; BELZILE, F. **Efficient genome-wide genotyping and data integration in crop plants.** Springer. 2018. DOI: 10.1007/s00122-018-3056-z

[12] ELSHIRE, R., GLAUBITZ, J., SUN, Q., POLAND, J., KAWAMOTO, K., BUCKLER, E., MITCHELL, S. **A Robust Simple Genotyping –by-Sequencing (GBS) Approach for High Diversity Species.** Plos One 6(5):e19379. 2011. DOI:10.1371/journal.pone.0019379

[13] LI, Y.; WANG, H. **Advances of genotyping-by-sequencing in fisheries and aquaculture. Fish Biol Fisheries**. 2017. DOI: 10.1007/s11160-017-9473-2

[14] CHUNG, Y.; CHOI, S.; JUN, T.; KIM, C. **Genotyping-by-sequencing: a Promising Tool for Plant Genetics Research and Breeding.** Hortic. Environ. Biotechnol. 58 (5):425-431. 2017. DOI: 10.1007/s13580-017-0297-8

[15] YE, H.; MEEHAN, J.; TONG, W.; HONG, H. **Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine**. Pharmaceutics, 523-541. 2015. DOI: 10.3390/pharmaceutics7040523

[16] ALSER, M., ROTMAN, J., DESHPANDE, D., TARASZKA, K., SHI, H., BAYKAL, P., YANG, H., XUE, V., KNYAZEV, S., SINGER, B., BALLIU, B., KOSLICKI, D., SKUMS, P., ZELIKOVSKY, A., ALKAN, C., MUTLU, O., MANGUL, S. **Technology dictates algorithms: recent developments in read alignment.** Genome Biology, 22:249. 2021. DOI: 10.1186/s13059-021-02443-7

[17] MIELEZAREK, M., SZYDA, J. **Review of alignment and SNP calling algorithms for next-generation sequencing data.** J Appl Genetics. 57:71-79. 2016. DOI: 10.1007/s13353-015-0292-7.

[18] LIU, F., ZHANG, Y., LI, L., FAND, Q., GAO, R., ZHANG, Z. **Systematic comparative analysis of single nucleotide variant detection methods from single-cell RNA sequencing data.** Genome Biology. 20:242.2019. DOI: 10.1186/s13059-019-1863-4

[19] YAO, Z., N´DIAYE, A., McCARTNEY, R., HIEBERTR, C., POZNIAK, C., XU, W. **Evaluation of variant calling tools for large plant genome re-sequencing.** BMC Bioinformatics. 21:360. 2020. DOI: 10.1186/s12859-020-03704-1

[20] WU, X., HEFFELFINGER, C., ZGAO, H., DALLAPORTA, S. **Benchmarking variant identification tools for plant diversity discovery**. BMC Genomics. 20: 701. 2019. DOI: 10.1186/s12864-019-6057-7

[21] SCHILBERT, H., REMPEL, A., PUCKER, B. **Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data**. Plants.9:429. 2020. DOI: 10.3390/plants9040439

[22] TORKAMANEH, D.; LAROCHE, J.; BELZILE, F. **Fats-GBS v2.0: an analysis toolkit for genotyping-by-sequencing data.** Techniques. 2020. https://doi.org/10.1139/gen-2020-0077

[23] SHENDURE, J., & JI, H. **Next-generation DNA sequencing**. Nature biotechnology, 26(10), 1135-1145, 2008.

[24] RIMMER, A. PHAN, H. MATHIESON, L. LQBAL, Z. WILKIE, O. MCVEAN, G. LUNTER, G. **Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications**. Nature Genetics, 46, 912-918, 2014.

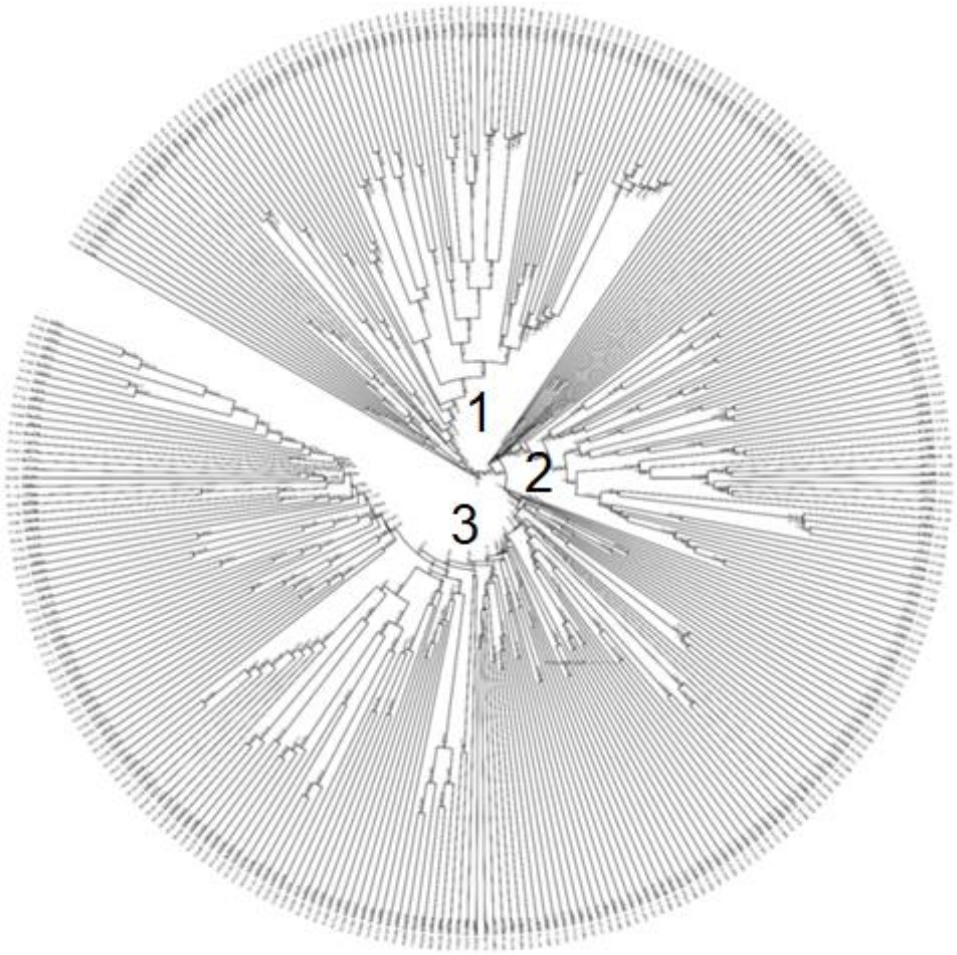
[25] LI, J. YUAN, D. WANG, P. WANG, Q. SUN, M. LIU, Z. SI, H. XU, Z. MA, Y. ZHANG, B. PEI, L. TU, L. ZHU, L. CHEN, L. LINDSEY, K. ZHANG, X. JIN, S, wang, M. **Cotton pan-genome retrieves the lost sequences and genes during domestication and selection.** Genome Biology, 22, 119, 2021.

[26] LIU J, SHEN Q, BAO H. **Comparison of seven SNP calling pipelines for the next- generation sequencing data of chickens**. PLoS ONE 17(1): e0262574. 2022. https://doi.org/10.1371/journal.pone.0262574

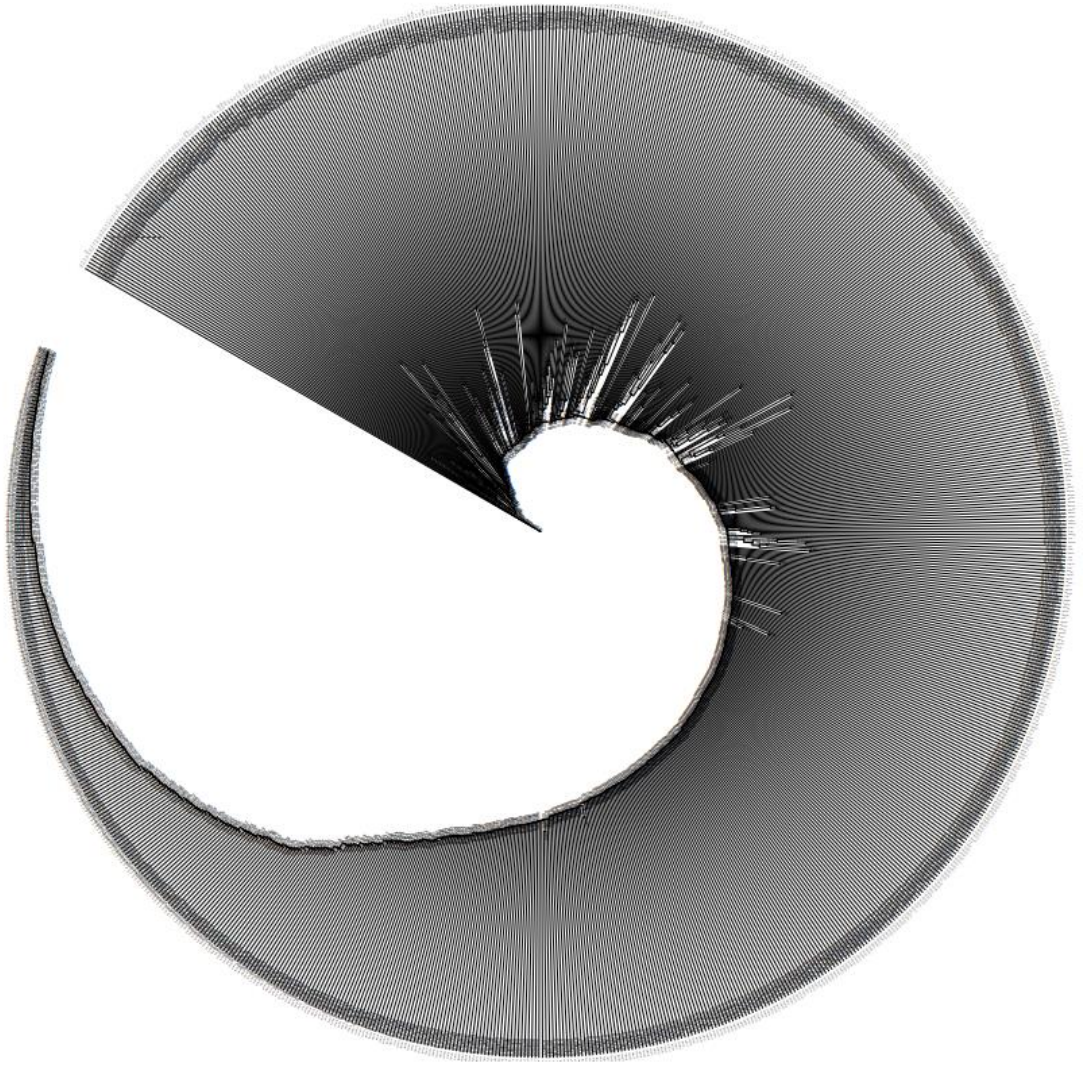
[27] METZKER, M. L. **Sequencing technologies - the next generation**. Nature Reviews Genetics, 11(1), 31-46. 2010 DOI:10.1038/nrg2626

[28] LI F, FAN G, LU G, XIAO G, ZOU C , KOHEL R, MA Z, SHANG H, MA X, WU J, LIANG X, HUANG G, PERCY R, LIU K, YANG W, CHEN W, DU X e colaboradores. **Genome Sequence of cultivated Upland Cotton (*Gossypium hirsutum* TM-1) provides insights into genome Evolution**. Nature Biotechnology, 33, 524-530. 2015.

7 ANEXOS



Anexo 1. Dendograma dos 474 genótipos da TMG.



Anexo 2. Dendograma dos 1.961 genótipos da literatura.