

Desenvolvimento de Marcadores SSR Polimórficos para *Agave* spp.: Uma Cultura Bioenergética Brasileira em Ascensão

Palavras-Chave: sisal, SSR, polimorfismo.

João Vitor Rodrigues Mio¹, Marina Püpke Marone² (coorientadora), Lucas Miguel Carvalho³, Adriele Bárbara de Oliveira¹, Gonçalo Amarante Guimarães Pereira¹, Marcelo Falsarella Carazzolle¹ (orientador)

¹ Instituto de Biologia, Unicamp.

² Leibniz Institute of Plant Genetics and Crop Plant Research – Gatersleben, Germany

³ Ciências da Saúde, Universidade São Francisco.

1. INTRODUÇÃO.

Os agaves são plantas que desempenham um papel importante na produção de fibras de sisal e possuem um potencial significativo na produção de açúcar fermentescível e biomassa, podendo se tornar uma cultura equivalente à cana-de-açúcar em regiões semiáridas, incluindo o Brasil, que possui uma extensão aproximada de 83 milhões de hectares (PROJETO MAPBIOMAS, 2019). Para aproveitar esse potencial, é necessário prospectar indivíduos altamente produtivos. Em 2019, foi realizada uma coleta em regiões semiáridas da Bahia, maior região produtora de fibra de sisal do Brasil, para obter indivíduos para o banco de germoplasma e genotipá-los. Assim, 95 amostras foram selecionadas e sequenciadas por ddRADseq para desenvolver marcadores moleculares e selecionar indivíduos interessantes.

Dentro do âmbito de melhoramento de culturas, os marcadores moleculares são recursos valiosos para seleção de indivíduos com fenótipo desejado, como plantas resistentes a doenças ou estresses abióticos (Bered; Barbosa Neto; Carvalho, 1997). Entre os tipos de marcadores, destacam-se os microssatélites ou SSR (*Simple Sequence Repeat*), que consistem em repetições de 1 a 6 nucleotídeos (Idrees; Irshad, 2014). Esses marcadores são frequentemente encontrados em regiões intergênicas e são importantes para diferenciar grupos de organismos semelhantes (Senan *et al.*, 2014). Nesse contexto, a genotipagem de diversos indivíduos é uma etapa fundamental para orientar programas de melhoramento.

O objetivo deste projeto é desenvolver um pipeline de bioinformática para selecionar marcadores moleculares SSRs polimórficos entre indivíduos de *Agave*. A partir de genomas parciais de dois genótipos de *Agave*, obtivemos sequências de SSR e selecionamos sequências polimórficas entre os 95 indivíduos sequenciado por ddRAD-seq, distribuídos nestes dois genótipos. Com este pipeline, será possível realizar uma validação preliminar *in silico*, seguida da validação em bancada dos marcadores para identificar a variação de tamanho, visando desenvolver um kit de identificação dos acessos de agave.

2. METODOLOGIA.

2.1 Dados utilizados.

Os dados para identificação dos locos SSRs são genomas *draft* (parciais) de *A. sisalana* e do híbrido 11648 (H11648), que possuem regiões gênicas, incluindo promotor e íntrons (dados ainda não publicados). Os dados de genotipagem foram obtidos através da técnica ddRAD-seq para um total de 95 indivíduos, sendo 77 *A. sisalana* e 18 híbrido 11648. Além disso, o pipeline está sendo testado com um outro genoma de H11648 em escala de cromossomos (*reads* longos obtidos da Oxford Nanopore Technologies e *reads* de captura da conformação cromossômica - Hi-C), publicado recentemente (Yang et al 2024).

2.2 Identificação de SSRs.

A identificação dos locos SSR foi feita a partir do conjunto das regiões gênicas de *A. sisalana* e híbrido 11648. Utilizamos o software MISA v.2.1 (Thiel *et al.*, 2003), que classifica os SSR com relação ao tamanho, unidade de repetição e a localização nos contigs da montagem do genoma. Na sequência, o software PRIMER3 v. 0.6 (UNTERGASSER et al., 2012) gerou os pares de primers que flanqueiam as regiões dos SSRs identificadas com informações detalhadas sobre o tamanho do amplicon e temperatura de anelamento.

2.3 Busca por locos polimórficos utilizando dados de ddRAD-seq.

O conjunto de SSRs e os dados de ddRAD-seq estão sendo utilizados no desenvolvimento do pipeline para identificar os locos polimórficos. Os *reads* de ddRAD-seq de cada indivíduo foram mapeados nos genomas *draft* com o software BWA-MEM (Li; Durbin, 2009) e o software Samtools (Danecek *et al.*, 2021) foi utilizado para converter os arquivos SAM em BAM. O software BEDTools (Quinlan; Hall, 2010) foi utilizado para conversão em formato BED e o módulo *subtract* para selecionar somente os *reads* de cada indivíduo que têm sobreposição com as regiões de SSR.

A partir dessas regiões, foi desenvolvido um código em Python que identifica os *reads* alinhados que atravessam regiões de SSRs. Este código gera arquivos com coordenadas para início e fim das regiões que flanqueiam a sobreposição do read com o SSR. Outro código em Python foi desenvolvido para identificar e contar as regiões polimórficas entre os indivíduos. Com base nessas informações, são selecionados os primers com parâmetros adequados para os testes em bancada.

2.4 Busca de novos SSR com estratégia *de novo*.

Devido a utilização de genomas drafts na busca dos SSRs polimórficos, apenas uma parte do total de SSRs acaba sendo obtida. Dessa forma, foi adotada uma estratégia de prospecção de SSRs sem referência (*de novo*), utilizando o pipeline Stacks (Catchen *et al.*, 2013). A ordem dos módulos utilizados foi *ustacks*, *cstacks*, *sstacks*, *tsv2bam* e *gstacks*.

RESULTADOS E DISCUSSÃO.

A partir da abordagem baseada em genoma de referência *draft*, foram encontrados 35.660 e 38.952 locos de SSR para *A.sisalana* e H11648, respectivamente. Os SSRs formados por mononucleotídeos e os compostos foram excluídos da análise (Tabela 1).

Tabela 1. Quantificação de cada tipo de SSR em *A. sisalana* e H11648 a partir dos dados obtidos com o software MISA.

	Dinucleotídeo	Trinucleotídeo	Tetranucleotídeo	Pentanucleotídeo	Hexanucleotídeo	Total
<i>Agave sisalana</i>	21.576	1.267	1.158	251	0	35.660
H11648	23.814	13.523	1.327	285	3	38.952

A partir destas regiões, foi feita a interseção dos SSR com os dados de ddRAD-seq com o software BEDTools. Em seguida, aplicamos o pipeline desenvolvido para identificar os *reads* de ddRAD-seq que atravessam regiões de SSR. Ao final do processo, foram obtidas 4.883 regiões para H11648 e 5.437 para *A. sisalana*. Estabelecemos um critério de cobertura mínima de cinco *reads* para validar um polimorfismo, permitindo-nos considerar uma região como portadora de variação dentro do SSR para um determinado indivíduo. Após este filtro, identificamos um total de 585 regiões para *A. sisalana* e 601 para H11648.

Nesta etapa, foi feita uma escolha manual das regiões com maior polimorfismo entre os indivíduos, ou seja, as regiões que são diferentes entre si. Ao final, foram selecionadas 5 regiões de cada um dos dois genótipos de *Agave* estudados. Com foco na validação experimental deste conjunto de regiões, pares de primers que flanqueiam essas regiões foram desenhados utilizando a plataforma Primer3Plus (versão 3.3.0) com o intuito de gerar produtos de amplificação com diferentes comprimentos. Essa etapa está atualmente sendo desenvolvida em parceria com o Instituto Agrônomo de Campinas, com o uso do genotipador 5200 *Fragment Analyzer Systems*.

Atualmente, o mesmo pipeline está sendo empregado para o genoma completo de H11648, publicado por Yang e colaboradores (2024). A etapa de identificação dos SSRs com o software MISA resultou na geração de 1.127.316 SSRs, quase sete vezes mais marcadores em relação ao primeiro genoma draft, no qual foram identificados apenas 164.758 SSRs. Esse número diminui quando excluimos da análise os SSRs mononucleotídeos e compostos, resultando em 576.622 para o novo genoma e 50.770

para o antigo, uma diferença de aproximadamente dez vezes. Essa discrepância é esperada, considerando o tamanho do genoma e o tipo de sequenciamento utilizado, que facilita a montagem de regiões repetitivas. Para a identificação das interseções de SSRs com os dados de ddRAD-seq, está sendo utilizado o mesmo método com o software BEDtools. Para isso, o cluster de computadores COARACI, disponibilizado pelo Center for Computing in Engineering & Sciences da Unicamp (projeto CEPID FAPESP: 2013/08293-7), é utilizado para processar esses dados.

Uma outra proposta do projeto foi prospectar novos SSRs utilizando a montagem *de novo* dos *reads* de ddRAD-seq com o software Stacks. O objetivo foi identificar novas regiões de SSRs polimórficas que não foram amostradas no pipeline anterior devido ao uso de genomas drafts incompletos de *A. sisalana* e H11648. Após a execução do programa MISA, foram obtidos 6.452 e 4.596 SSRs para *A. sisalana* e H11648, respectivamente (Tabela 2), menos locos SSRs em comparação com análises baseadas em genoma de referência (Tabela 1). Após a comparação entre os arquivos FASTA do genoma draft do H11648 e os dados *de novo*, foi identificado um total de 2.123 novas regiões de consenso montadas pelo Stacks. No entanto, estas sequências possuem um tamanho de 127 pb, tornando inviável projetar primers flanqueadores capazes de amplificar as regiões de SSR.

Tabela 2. Quantificação de cada tipo de SSR em *Agave sisalana* e H11648 a partir dos dados *de novo* obtidos com o software Stacks.

	Dinucleotídeo	Trinucleotídeo	Tetranucleotídeo	Pentanucleotídeo	Hexanucleotídeo	Total
<i>A. sisalana</i>	4.619	1.695	94	26	18	6.452
H11648	3.283	1.263	17	15	18	4.596

CONCLUSÕES.

Neste projeto, estamos desenvolvendo um pipeline de bioinformática para realizar a validação *in silico* de marcadores moleculares do tipo SSR em indivíduos de *A. sisalana* e H11648, para reduzir a complexidade dos testes em bancada. A partir dos dados baseados na montagem genômica, encontramos 601 regiões polimórficas em H11648 e 585 em *A. sisalana* nos genomas *draft*, sendo 10 atualmente em fase de validação em bancada. Resultados a partir do novo genoma de H11648 possibilitarão uma nova escolha de SSR polimórficos, visto o aumento da quantidade de informação. O pipeline desenvolvido permite aceleração da validação de primers que amplificam locos SSRs e redução de custos para os testes em bancada. Esses resultados permitirão a geração de primers que serão utilizados para distinguir indivíduos de *A. sisalana* e H11648, impulsionando o melhoramento genético dessa cultura no Brasil.

BIBLIOGRAFIA.

BERED, F.; BARBOSA NETO, J. F.; CARVALHO, F. I. F. de. Marcadores moleculares e sua aplicação no melhoramento genético de plantas. *Ciência Rural*, [s. l.], v. 27, n. 3, p. 513–520, 1997.

- CATCHEN, J. *et al.* Stacks: an analysis tool set for population genomics. **Molecular Ecology**, [s. l.], v. 22, n. 11, p. 3124–3140, 2013.
- DANECEK, P. *et al.* Twelve years of SAMtools and BCFtools. **GigaScience**, [s. l.], v. 10, n. 2, p. giab008, 2021.
- IDREES, M.; IRSHAD, M. Molecular Markers in Plants for Analysis of Genetic Diversity: A Review. [s. l.], n. 1, p. 28, 2014.
- LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics**, [s. l.], v. 25, n. 14, p. 1754–1760, 2009.
- QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, [s. l.], v. 26, n. 6, p. 841–842, 2010.
- SENAN, S. *et al.* Methods for Development of Microsatellite Markers: An Overview. **Notulae Scientia Biologicae**, [s. l.], v. 6, n. 1, p. 1–13, 2014.
- THIEL, T. *et al.* Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). **Theoretical and Applied Genetics**, [s. l.], v. 106, n. 3, p. 411–422, 2003.
- UNTERGASSER, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Research*, [s. l.], v. 40, n. 15, p. e115–e115, 2012.
- YANG, Z. *et al.* A chromosome-level genome assembly of *Agave hybrid* NO.11648 provides insights into the CAM photosynthesis. **Horticulture Research**, [s. l.], v. 11, n. 2, p. uhad269, 2024.